Lec 9: Advanced Fano's method

Yanjun Han

Covering and packing

Let (X, d) be a metric space, and $A \subseteq X$ be a compact set.

Defn (covering) $\{x_1, \cdots, x_n\} \subseteq X$ is an ε-covering (or ε-net) of A if

$$A \subseteq \bigcup_{i=1}^{n} B(x_i ; \varepsilon), \quad \text{with } B(x ; \varepsilon) = \{y \in X : d(x,y) \le \varepsilon\}.$$

Defn (packing) $\{a_1, \cdots, a_n\} \subseteq A$ is an ε-packing of A if

$$\min_{i \neq j} d(a_i, a_j) > \varepsilon.$$

Defn (covering and packing numbers)

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

$$M(A, d, \varepsilon) = \max\{m : \exists\ \varepsilon\text{-packing of } A \text{ of size } m\}$$

Basic relationship. $M(A, d, 2\varepsilon) \overset{①}{\le} N(A, d, \varepsilon) \overset{②}{\le} M(A, d, \varepsilon)$

(In other words, up to a multiplicative factor of 2 on ε, it's equivalent to consider covering or packing numbers.)

Pf. ①: If $M(A, d, 2\varepsilon) \ge N(A, d, \varepsilon) + 1$, by pigeonhole principle,

$\exists$ two points x, x' in a (2ε)-packing belong to the same ball $B(y ; \varepsilon)$ in an ε-covering

$\Rightarrow d(x, x') \le d(x, y) + d(x', y) \le 2\varepsilon$, a contradiction to (2ε)-packing.

②: If $a_1, \cdots, a_m$ is a maximal ε-packing of A, it must also be an ε-covering:

if not, then $\exists\ a \in A$ s.t. $d(a, a_i) > \varepsilon\ \forall i \in [m]$

$\Rightarrow \{a_1, \cdots, a_m, a\}$ is a larger ε-packing, a contradiction. $\square$

Bounding the covering/packing number.

1. Volume bound: let $\|\cdot\|$ be any norm on $\mathbb{R}^d$, $B = \{x : \|x\| \le 1\}$ be the unit ball.

Then

$$\left(\frac{1}{\varepsilon}\right)^d \frac{\mathrm{Vol}(A)}{\mathrm{Vol}(B)} \overset{①}{\le} N(A, \|\cdot\|, \varepsilon) \le M(A, \|\cdot\|, \varepsilon) \overset{②}{\le} \left(\frac{2}{\varepsilon}\right)^d \frac{\mathrm{Vol}(A + \frac{\varepsilon}{2}B)}{\mathrm{Vol}(B)}.$$

Pf. ①: As $A \subseteq \bigcup_{i=1}^{n} B(x_i; \varepsilon)$ for an ε-covering $\{x_1, \cdots, x_n\}$,

$$\mathrm{Vol}(A) \le \sum_{i=1}^{n} \mathrm{Vol}(B(x_i;\varepsilon)) = n\varepsilon^d \cdot \mathrm{Vol}(B) \Rightarrow n \ge \left(\frac{1}{\varepsilon}\right)^d \frac{\mathrm{Vol}(A)}{\mathrm{Vol}(B)}.$$

②: As $\bigcup_{i=1}^{m} B(a_i; \frac{\varepsilon}{2}) \subseteq A + \frac{\varepsilon}{2}B$, and the sets are disjoint under an ε-packing $\{a_1, \cdots, a_m\}$,

$$\mathrm{Vol}(A+\tfrac{\varepsilon}{2}B) \ge \sum_{i=1}^{m} \mathrm{Vol}(B(a_i;\tfrac{\varepsilon}{2})) = m\left(\frac{\varepsilon}{2}\right)^d \mathrm{Vol}(B) \Rightarrow m \le \left(\frac{2}{\varepsilon}\right)^d \frac{\mathrm{Vol}(A+\frac{\varepsilon}{2}B)}{\mathrm{Vol}(B)}.$$

Example 1.1. If $A = \{x : \|x\| \le 1\}$ is the unit ball under the same norm, then

$$\left(\frac{1}{\varepsilon}\right)^d \le N(A, \|\cdot\|, \varepsilon) \le \left(1+\frac{2}{\varepsilon}\right)^d \le \left(\frac{3}{\varepsilon}\right)^d \quad \text{for all } 0<\varepsilon\le 1.$$

Example 1.2 (Gilbert-Varshamov bound). If $A = \{0,1\}^d$ and $d_H(x,x') = \sum_{i=1}^{d} 1(x_i \ne x_i')$ is the Hamming distance, then for $1 \le r \le d-1$:

$$\frac{2^d}{\sum_{i=0}^{r} \binom{d}{i}} \le M(A, d_H, r) \le \frac{2^d}{\sum_{i=0}^{\lfloor r/2 \rfloor} \binom{d}{i}}.$$

If $r = pd$ with $d \to \infty$, then by Stirling's approximation,

$$2^{d(1-h(p)+o(1))} \le M(\{0,1\}^d, d_H, pd) \le 2^{d(1-h(\frac{p}{2})+o(1))}.$$

$\left(h(p) = p\log_2 \frac{1}{p} + (1-p)\log_2 \frac{1}{1-p} \right)$

Pf. Left inequality: $M(r) \ge N(r)$ and $\{0,1\}^d \subseteq \bigcup_{i=1}^{n} B(x_i; r)$.

Right inequality: $\bigcup_{i=1}^{m} B(a_i; \lfloor \frac{r}{2} \rfloor) \subseteq \{0,1\}^d$ and balls are disjoint. ∎

2. Sudakov minoration: Let $w(A) = \mathbb{E} \sup_{a\in A} \langle a, Z\rangle$, $Z \sim N(0, I_d)$ be the Gaussian width of A, then

$$w(A) \ge C \sup_{\varepsilon>0} \varepsilon \sqrt{\log M(A, \|\cdot\|_2, \varepsilon)}.$$

(Two results needed to prove it:

① Slepian's lemma: let X, Y be centered Gaussians in $\mathbb{R}^d$ with

$$\mathbb{E}[(Y_i - Y_j)^2] \le \mathbb{E}[(X_i - X_j)^2] \quad \forall i,j \in [d].$$

Then $\quad \mathbb{E}[\max Y_i] \le \mathbb{E}[\max X_i]$.

② Maximum of n Gaussians: let $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(0,1)$, then

$$\mathbb{E}[\max X_i] = (1+o(1)) \cdot \sqrt{2\log n}.$$

)

Pf of Sudakov minoration. Let $\{a_1, \cdots, a_m\}$ be an optimal ε-packing of A.

Define $X_i = \langle a_i, Z\rangle$ (with $Z \sim N(0, I_d)$), and $Y_1, \cdots, Y_m \overset{\text{i.i.d.}}{\sim} N(0, \frac{\varepsilon^2}{2})$.

Then $\mathbb{E}[(Y_i - Y_j)^2] = \varepsilon^2 \le \|a_i - a_j\|_2^2 = \mathbb{E}[(X_i - X_j)^2]$

$$\Rightarrow w(A) \ge \mathbb{E}[\max_{i=1}^m X_i] \ge \mathbb{E}[\max_{i=1}^m Y_i] = \frac{\varepsilon}{\sqrt{2}} \cdot (1 + o(1)) \sqrt{2\log m}.$$

Example 1.3. When $A = B_1 = \{x : \|x\|_1 \le 1\}$, then

$$w(A) = \mathbb{E} \sup_{\|x\|_1 \le 1} \langle x, Z\rangle = \mathbb{E}\|Z\|_\infty \le \sqrt{2\log d}$$

$$\Rightarrow \log M(B_1, \|\cdot\|_2, \varepsilon) = O\left(\frac{\log d}{\varepsilon^2}\right).$$

In fact, it holds that

$$\log M(B_1, \|\cdot\|_2, \varepsilon) \asymp \begin{cases} d\left(1 + \log \frac{1}{\varepsilon^2 d}\right) & \text{if } \varepsilon \le \frac{1}{\sqrt{d}} \quad \text{(volume bound is tight)} \\ \frac{1 + \log(\varepsilon^2 d)}{\varepsilon^2} & \text{if } \frac{1}{\sqrt{d}} < \varepsilon < 1 \quad \text{(Sudakov minoration nearly achieves the tight upper bound)} \end{cases}$$

3. Maurey's empirical method: let $(H, \langle \cdot, \cdot \rangle)$ be an inner product space, and $T \subseteq H$ be a finite set. Then

$$N(\mathrm{conv}(T), \|\cdot\|, \varepsilon) \le \binom{|T| + \lceil \frac{r^2}{\varepsilon^2} \rceil - 2}{\lceil \frac{r^2}{\varepsilon^2} \rceil - 1}, \quad 0 < \varepsilon \le r,$$

with $r = \inf_{y \in H} \sup_{x \in T} \|x - y\|$ is the radius of T.

Pf. We use a probabilistic argument. Let $T = \{t_1, \cdots, t_m\}$, and $c \in H$ satisfy $r = \max_{i=1}^m \|t_i - c\|$.

Then for any $x \in \mathrm{conv}(T)$, we have $x = \sum_{i=1}^m x_i t_i$ for some $x_i \ge 0$, $\sum_i x_i = 1$.

Let Z be an H-valued RV s.t. $\mathbb{P}(Z = t_i) = x_i$ $1 \le i \le m$, then $x = \mathbb{E}[Z]$.

Let $Z_1, \cdots, Z_n$ be i.i.d. copies of Z, and $\bar{Z} = \frac{1}{n+1}(c + \sum_{i=1}^n Z_i)$. Then

$$\mathbb{E}[\|\bar{Z} - x\|^2] = \frac{1}{(n+1)^2}\Big(\underbrace{\|c - x\|^2}_{\substack{\le r^2 \text{ by} \\ \text{convexity}}} + n \underbrace{\mathbb{E}[\|Z - x\|^2]}_{\substack{\le \mathbb{E}[\|Z - c\|^2] \\ \le r^2 \text{ as } \mathbb{E}Z = x.}}\Big) \le \frac{r^2}{n+1}.$$

Consequently, if $n = \lceil \frac{r^2}{\varepsilon^2} \rceil - 1$, $\exists$ realization of $\bar{Z}$ s.t. $\|x - \bar{Z}\| \le \varepsilon$. Meanwhile,

$$\bar{Z} \in \left\{ \frac{1}{n+1}\left(c + \sum_{i=1}^m n_i t_i\right) : \begin{array}{l} n_i \ge 0 \\ \sum_i n_i = n \end{array} \right\} \text{ with cardinality } \binom{n+m-1}{n}.$$

Example 1.3 cont'd. $B_1 = \mathrm{conv}(\{\pm e_1, \cdots, \pm e_d\})$, with radius 1.

By Maurey's empirical method,

$$\log N(B_1, \|\cdot\|_2, \varepsilon) \le \log \binom{2d + \lceil \frac{1}{\varepsilon^2} \rceil - 2}{\lceil \frac{1}{\varepsilon^2} \rceil - 1} = O\left(\frac{1 + \log(\varepsilon^2 d)}{\varepsilon^2}\right) \text{ if } \frac{1}{\sqrt{d}} < \varepsilon < 1.$$

4. More results without proof.

① For $0 < p < q \le \infty$, and $B_p := \{x \in \mathbb{R}^d : \|x\|_p \le 1\}$, then

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

② Let $N(A, B)$ be the smallest translations of B that cover A. There exist universal constant $\alpha, \beta > 0$ such that, for any symmetric convex body A,

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

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

③ Let $H^s = \{f \in C^s([0,1]) : \|f^{(s)}\|_\infty \le 1\}$, then

$$\log N(H^s, \|\cdot\|_p, \varepsilon) \asymp_p \varepsilon^{-\frac{1}{s}} \quad \text{for any } 1 \le p \le \infty.$$

④ Let $\mathcal{F}_M = \{f : [0,1] \to [0,1],\ f \text{ is non-decreasing}\}$, then

$$\log N(\mathcal{F}_M, \|\cdot\|_p, \varepsilon) \asymp_p \frac{1}{\varepsilon} \quad \text{for any } 1 \le p < \infty.$$

⑤ Let $\mathcal{F}_C = \{f : [0,1] \to [0,1],\ f \text{ is convex}\}$, then

$$\log N(\mathcal{F}_C, \|\cdot\|_p, \varepsilon) \asymp_p \frac{1}{\sqrt{\varepsilon}} \quad \text{for any } 1 \le p < \infty.$$

Global Fano's method

Recall the steps of Fano's inequality:

① Find a pairwise separated set $\{\theta_1, \cdots, \theta_m\} \subseteq \Theta$ s.t. for all $i \ne j$,

$$\min_{a \in \mathcal{A}} L(\theta_i, a) + L(\theta_j, a) \ge \Delta$$

② Try to upper bound $I(\theta; X)$ with $\theta \sim \mathrm{Unif}(\{\theta_1, \cdots, \theta_m\})$ and $X \mid \theta \sim P_\theta$

③ If $I(\theta; X) < \frac{1}{2} \log m$, then the minimax risk satisfies $r^* = \Omega(\Delta)$.

Step ① is packing: if there is a metric $d(\theta,\theta')$ satisfies

$$\min_{a\in A} L(\theta,a)+L(\theta',a) \geq h(d(\theta,\theta')) \quad \text{for an increasing function } h:\mathbb{R}_+\to\mathbb{R}_+,$$

then a δ-packing $\{\theta_1,\dots,\theta_m\}$ of Θ under d satisfies the separation condition with $\Delta = h(\delta)$.

Q: Is there a general upper bound of $I(\theta;X)$, or more often, $I(\theta;X^n)$?

A: This is possible using a covering of $(P_\theta)_{\theta\in\Theta}$ under KL!

Def. For a family $\mathcal{P}$ of distributions and $\varepsilon>0$. let $N_{KL}(\mathcal{P},\varepsilon)$ be the smallest integer n s.t. $\exists$ distributions $Q_1,\dots,Q_n$ (not necessarily in $\mathcal{P}$) satisfying

$$\sup_{P\in\mathcal{P}} \min_{i\in[n]} D_{KL}(P\,\|\,Q_i) \leq \varepsilon^2.$$

(D_{KL} is not a metric; Q_i in second argument)

Thm. (Entropic upper bound of $I(\theta;X^n)$). Let $\theta\sim\pi$ with $\mathrm{supp}(\pi)=\Theta_0$, and $X^n|\theta \sim P_\theta^{\otimes n}$.

Then

$$I(\theta;X^n) \leq \inf_{\varepsilon>0}\left(n\varepsilon^2+\log N_{KL}((P_\theta)_{\theta\in\Theta_0},\varepsilon)\right).$$

Pf. Recall the "golden formula" in Lec 7:

$$I(\theta;X^n) = \min_{Q_{X^n}} \mathbb{E}_{\theta\sim\pi}\left[D_{KL}(P_\theta^{\otimes n}\,\|\,Q_{X^n})\right].$$

For an ε-covering of $(P_\theta)_{\theta\in\Theta_0}$ $Q_1,\dots,Q_N$ with $N=N_{KL}((P_\theta)_{\theta\in\Theta_0},\varepsilon)$, choose $Q_{X^n}=\frac{1}{N}\sum_{i=1}^N Q_i^{\otimes n}$. Then for $\theta\sim\pi$,

$$D_{KL}\left(P_\theta^{\otimes n}\,\Big\|\,\frac{1}{N}\sum_{i=1}^N Q_i^{\otimes n}\right) = \mathbb{E}_{P_\theta^{\otimes n}}\left[\log\frac{P_\theta^{\otimes n}}{\frac{1}{N}\sum_{i=1}^N Q_i^{\otimes n}}\right]$$

$$\leq \mathbb{E}_{P_\theta^{\otimes n}}\left[\min_{i\in[N]}\log\frac{P_\theta^{\otimes n}}{Q_i^{\otimes n}}+\log N\right] \quad \left(\sum_i x_i \geq \max_i x_i\right)$$

$$\leq \min_{i\in[N]} \mathbb{E}_{P_\theta^{\otimes n}}\left[\log\frac{P_\theta^{\otimes n}}{Q_i^{\otimes n}}\right]+\log N$$

$$= \min_{i\in[N]} nD_{KL}(P_\theta\,\|\,Q_i)+\log N$$

$$\leq n\varepsilon^2+\log N \quad \text{a.s. for } \theta\in\Theta_0. \qquad \square$$

Diagram of global Fano's method: for hyperparameters $\Theta_0 \subseteq \Theta$, $\varepsilon, \delta > 0$:

① Find a metric $d(\theta, \theta')$ satisfying $\min_a L(\theta, a) + L(\theta', a) \geq h(d(\theta, \theta'))$ for an increasing non-negative function h, and find a δ-packing of Θ_0 under d.

② Find an ε-covering of $(P_\theta)_{\theta\in\Theta_0}$ under KL.

③ Apply Fano's method to conclude that

$$r^* \geq \frac{h(\delta)}{2}\left(1 - \frac{\log N_{KL}((P_\theta)_{\theta\in\Theta_0}, \varepsilon) + n\varepsilon^2 + \log 2}{\log M(\Theta_0, d, \delta)}\right).$$

Optimize over $(\Theta_0, \delta, \varepsilon)$ to make the lower bound as large as possible.

Example 2.1 (GLM). $X_1, \cdots, X_n \overset{i.i.d.}{\sim} N(\theta, I_d)$ with unknown $\theta \in \mathbb{R}^d$.

Target:

$$\inf_{\hat\theta} \sup_{\theta} \mathbb{E}_\theta [\|\hat\theta - \theta\|_p] \asymp_p \begin{cases} \frac{d^{1/p}}{\sqrt{n}} & 2 < p < \infty \\ \sqrt{\frac{\log d}{n}} & p = \infty \end{cases}.$$

Pf of lower bound. Choose $\Theta_0 = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq r\}$, then for any $\varepsilon, \delta > 0$, global Fano gives

$$r^* \gtrsim \delta\left(1 - \frac{\log N_{KL}(\{N(\theta, I_d)\}_{\theta\in\Theta_0}, \varepsilon) + n\varepsilon^2 + \log 2}{\log M(\Theta_0, \|\cdot\|_p, \delta)}\right)$$

$$= \delta\left(1 - \frac{\log N(\Theta_0, \|\cdot\|_2, \sqrt{2}\varepsilon) + n\varepsilon^2 + \log 2}{\log M(\Theta_0, \|\cdot\|_p, \delta)}\right). \quad \left(D_{KL}(N(\theta, I_d)\|N(\theta', I_d)) = \tfrac{1}{2}\|\theta - \theta'\|_2^2\right)$$

Choice of ε: we choose $\varepsilon = \frac{r}{\sqrt{2}}$, so that $\log N = \log 1 = 0$.

Choice of r/δ: for $p \in (2, \infty)$, choose $\frac{\delta}{r} = d^{\frac{1}{p} - \frac{1}{2}}$, so that

$$\log M(\Theta_0, \|\cdot\|_p, \delta) \asymp d.$$

For $p = \infty$, choose $\frac{\delta}{r} \asymp 1$, so that $\log M(\Theta, \|\cdot\|_\infty, \delta) \asymp \log d$.

Choice of r: now we have

$$r^* \gtrsim \begin{cases} r d^{\frac{1}{p} - \frac{1}{2}}\left(1 - \frac{C_1 n r^2 + \log 2}{C_2 d}\right) & \text{if } p \in (2, \infty), \\ r\left(1 - \frac{C_1 n r^2 + \log 2}{C_2 \log d}\right) & \text{if } p = \infty. \end{cases}$$

So $r = \begin{cases} \sqrt{\frac{d}{n}} & \text{for } p \in (2, \infty) \\ \sqrt{\frac{\log d}{n}} & \text{for } p = \infty \end{cases}$ gives $r^* \gtrsim \begin{cases} \frac{d^{1/p}}{\sqrt{n}} & 2 < p < \infty, \\ \sqrt{\frac{\log d}{n}} & p = \infty. \end{cases}$

Example 2.2 (nonparametric density estimation) $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} f$ on $[0,1]$ with $\|f^{(s)}\|_\infty \le 1$.
(i.e. the function space is H^s)

Target:
$$\inf_{\hat f} \sup_{f \in F} \mathbb{E}_f [\| \hat f(X^n) - f \|_p] \asymp_p n^{-\frac{s}{2s+1}}, \quad p \in [1, \infty).$$

Pf of lower bound. Consider a subset $H_0^s \subseteq H^s$: $H_0^s = \{ f \in H^s : f \ge \frac{1}{2} \text{ on } [0,1] \}$.

Then for $f, g \in H_0^s$, $D_{KL}(f \| g) \le \chi^2(f \| g) \le 2 \| f - g \|_2^2$

$$\Rightarrow N_{KL}(H_0^s, \varepsilon) \le N(H_0^s, \|\cdot\|_2, \tfrac{\varepsilon}{\sqrt{2}}) \le N(H^s, \|\cdot\|_2, \tfrac{\varepsilon}{\sqrt{2}}).$$

By global Fano, for any $\varepsilon, \delta > 0$,
$$r^* \gtrsim \delta \left(1 - \frac{\log N_{KL}(H_0^s, \varepsilon) + n\varepsilon^2 + \log 2}{\log M(H_0^s, \|\cdot\|_p, \delta)} \right)$$
$$\gtrsim \delta \left(1 - \frac{c_1 \varepsilon^{-\frac{1}{s}} + n\varepsilon^2 + \log 2}{c_2 \delta^{-1/s}} \right). \quad \text{(by metric entropy bounds for } H_0^s \text{)}$$

Choosing $\varepsilon \asymp \delta \asymp n^{-\frac{s}{2s+1}}$ gives $r^* = \Omega(n^{-\frac{s}{2s+1}})$. ☺

Example 2.3 (Isotonic regression) $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} P_X$, where (the known or unknown) P_X has a bounded density on $[0,1]$. Conditioned on X^n, $Y_i \overset{\text{ind}}{\sim} N(f(X_i), 1)$ with
$$f \in F_M = \{ f : [0,1] \to [0,1],\ f \text{ is increasing} \}.$$

Target:
$$\inf_{\hat f} \sup_{f \in F_M} \mathbb{E}_f [\| \hat f - f \|_p] \asymp_p n^{-\frac{1}{3}}, \quad \text{for all } p \in [1, \infty).$$

Pf of lower bound. Since P_X has a bounded density,
$$D_{KL}(P_f \| P_{f'}) = \frac{1}{2} \| f - f' \|_{L^2(P_X)}^2 = O(1) \cdot \| f - f' \|_2^2$$
$$\Rightarrow N_{KL}((P_f)_{f \in F_M}, \varepsilon) \le N(F_M, \|\cdot\|_2, \tfrac{\varepsilon}{O(1)}).$$

By global Fano:
$$r^* \gtrsim \delta \left(1 - \frac{\log N(F_M, \|\cdot\|_2, \frac{\varepsilon}{O(1)}) + n\varepsilon^2 + \log 2}{\log M(F_M, \|\cdot\|_p, \delta)} \right)$$
$$\gtrsim \delta \left(1 - \frac{\frac{c_1}{\varepsilon} + n\varepsilon^2 + \log 2}{1/\delta} \right). \quad \left(\log N(F_M, \|\cdot\|_p, \varepsilon) \asymp_p \frac{1}{\varepsilon} \right)$$

Choosing $\varepsilon \asymp n^{-1/3}$ and $\delta \asymp n^{-1/3}$, we obtain $r^* = \Omega(n^{-1/3})$. ☺

Example 2.4 (Convex regression) Same setting as Example 2.3, but with $\mathcal{F}_m$ replaced by

$$\mathcal{F}_c = \{ f : [0,1] \to [0,1],\ f \text{ is convex} \}.$$

Target:

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_c} \mathbb{E}_f [\| \hat{f} - f \|_p] \asymp_p n^{-\frac{2}{5}}, \quad p \in [1, \infty).$$

Pf of lower bound, similar to Example 2.3, now with $\log N(\mathcal{F}_c, \|\cdot\|_p, \varepsilon) \asymp_p \frac{1}{\sqrt{\varepsilon}}$. $\square$

Example 2.5 (Sparse linear regression) $Y \sim N(X\theta, I_n)$ with fixed design $X \in \mathbb{R}^{n \times d}$, where all singular values of X are $O(\sqrt{n})$. The unknown parameter $\theta \in \mathbb{R}^d$ is sparse:

$$\|\theta\|_q \le R \quad \text{for some } q \in (0,1).$$

Target:

$$\inf_{\hat{\theta}} \sup_{\|\theta\|_q \le R} \mathbb{E}_\theta [\| \hat{\theta} - \theta \|_p] \asymp_{p,q} R^{\frac{q}{p}} \left(\frac{\log d}{n} \right)^{\frac{p-q}{2p}} \quad \text{for small enough } R < f(n,d).$$

Pf of lower bound.

1. L_p-packing of $B_q(R) = \{ \theta \in \mathbb{R}^d : \|\theta\|_q \le R \}$:

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

2. KL covering of $\mathcal{P} = \{ N(X\theta, I_n) : \|\theta\|_q \le R \}$:

$$D_{KL}(N(X\theta, I_n) \| N(X\theta', I_n)) = \frac{1}{2} \| X(\theta - \theta') \|_2^2 = O(n) \cdot \| \theta - \theta' \|_2^2$$

$$\Longrightarrow \log N_{KL}(\mathcal{P}, \varepsilon) \le \log N\left(B_q(R), \|\cdot\|_2, \frac{\varepsilon}{O(\sqrt{n})}\right)$$

$$\asymp \left(\frac{\sqrt{n} R}{\varepsilon} \right)^{\frac{2q}{2-q}} \log d \quad \text{if } \varepsilon \gg R\sqrt{n}\, d^{\frac{1}{2} - \frac{1}{q}}.$$

Now choosing

$$\varepsilon \asymp n^{\frac{q}{4}} R^{\frac{q}{2}} (\log d)^{\frac{2-q}{4}}$$

$$\delta \asymp R^{\frac{q}{p}} \left(\frac{\log d}{n} \right)^{\frac{p-q}{2p}},$$

then

$$\log M(\delta) \asymp R^q n^{\frac{q}{2}} (\log d)^{1 - \frac{q}{2}},$$

$$\log N_{KL}(\varepsilon) \asymp \varepsilon^2 \asymp R^q n^{\frac{q}{2}} (\log d)^{1 - \frac{q}{2}},$$

and global Fano gives the result. $\square$

Special topic: generalized Fano with χ^2-informativity.

Since the proof of Fano is simply DPI, replacing KL by other f-divergences also leads to meaningful Bayes risk lower bounds.

Thm. For $\theta\sim\pi$, it holds that

$$\mathbb{P}(L(\theta,X)\ge\Delta)\ \ge\ 1-p_\Delta-\sqrt{p_\Delta\cdot I_{\chi^2}(\theta;X)},$$

where $p_\Delta=\sup_{a}\pi(L(\theta,a)<\Delta)$ is the small-ball probability, and

$$I_{\chi^2}(\theta;X)=\inf_{Q_X}\chi^2(P_{\theta X}\,\|\,\pi_\theta Q_X)=\inf_{Q_X}\mathbb{E}_{\theta\sim\pi}[\chi^2(P_{X|\theta}\,\|\,Q_X)]$$

is the χ^2-informativity.

Pf. Apply DPI to:

$$P_{\theta X},\ \pi_\theta Q_X \xrightarrow{(\theta,X)\mapsto \mathbb{1}(L(\theta,X)\ge\Delta)} \mathrm{Bern}(\mathbb{P}(L(\theta,X)\ge\Delta)),\ \mathrm{Bern}(\ge 1-p_\Delta)$$

We get:

$$\chi^2(P_{\theta X}\,\|\,\pi_\theta Q_X)\ \ge\ \chi^2(\mathrm{Bern}(\mathbb{P}(L(\theta,X)\ge\Delta))\,\|\,\mathrm{Bern}(\ge 1-p_\Delta))$$
$$\ge \frac{(\mathbb{P}(L(\theta,X)\ge\Delta)-(1-p_\Delta))^2}{p_\Delta(1-p_\Delta)}\quad \text{if } \mathbb{P}(L(\theta,X)\ge\Delta)\le 1-p_\Delta.$$

Taking inf over Q_X and rearranging gives the result. ☒

Similarly, we have an entropic upper bound of $I_{\chi^2}(\theta;X)$ based on χ^2-covering.

Thm. Let $\mathcal{P}=(P_\theta)_{\theta\in\Theta}$ and $\mathrm{supp}(\pi)\subseteq\Theta$. Then for $\theta\sim\pi$,

$$I_{\chi^2}(\theta;X)+1\ \le\ \inf_{\varepsilon>0}\,(1+\varepsilon^2)N_{\chi^2}(\mathcal{P},\varepsilon),$$

where

$$N_{\chi^2}(\mathcal{P},\varepsilon)=\min\Big\{n:\ \min_{Q_1,\cdots,Q_n}\sup_{P\in\mathcal{P}}\min_{i\in[n]}\chi^2(P\,\|\,Q_i)\le\varepsilon^2\Big\}.$$

Pf. Exercise.

Example 3.1 (Gaussian model with uniform prior). Let $X\sim N(\theta, I_d)$ with $\theta\sim \overbrace{\mathrm{Unif}(B_2(R))}^{=:\pi}$.

Target: $$r_\pi^* = \inf_{\hat\theta} \mathbb{E}_{\theta\sim\pi}[\|\hat\theta-\theta\|_2^2] \asymp d \quad \text{if } R=\Omega(\sqrt{d}).$$

Failure of mutual information. For $\Delta\in(0,R)$, the small ball prob. is

$$P_\Delta = \sup_a \pi(\|\theta-a\|_2^2 \le \Delta^2) = \left(\frac{\Delta}{R}\right)^d.$$

For the mutual info, the entropic upper bound gives

$$\begin{aligned} I(\theta; X) &\le \inf_{\varepsilon>0}\left(\log N(B_2(R), \|\cdot\|_2, \varepsilon) + \varepsilon^2\right) \\ &\le \inf_{\varepsilon>0}\left(d\log\frac{3R}{\varepsilon} + \varepsilon^2\right) \sim d\log\frac{R}{\sqrt{d}} \quad \text{if } R\gg\sqrt{d} \end{aligned}$$

Therefore, Fano gives that

$$r_\pi^* \gtrsim \sup_{\Delta>0} \Delta^2\cdot\left(1-\frac{d\log\frac{R}{\sqrt{d}}+\log 2}{d\log\frac{R}{\Delta}}\right).$$

Usually we make

$$d\log\frac{R}{\sqrt{d}} = (1-\rho)\cdot d\log\frac{R}{\Delta} \Rightarrow \Delta = d^{\frac{1}{2(1-\rho)}} R^{-\frac{\rho}{1-\rho}}, \text{ for some constant } \rho>0$$

$$\Rightarrow r_\pi^* = \Omega\left(d^{\frac{1}{1-\rho}} R^{-\frac{2\rho}{1-\rho}}\right) = \Omega\left(d\cdot\left(\frac{d}{R^2}\right)^{\frac{\rho}{1-\rho}}\right) \text{ is weaker than } \Omega(d).$$

Pf using χ^2-info. The entropic upper bound gives

$$\begin{aligned} I_{\chi^2}(\theta; X)+1 &\le \inf_{\varepsilon>0} (1+\varepsilon^2)\, N\left(B_2(R), \|\cdot\|_2, \sqrt{\log(1+\varepsilon^2)}\right) \quad \left(\chi^2(N(\theta,I)\,\|\,N(\theta',I)) = e^{\|\theta-\theta'\|_2^2}-1\right) \\ &\le \inf_{\varepsilon>0} (1+\varepsilon^2)\left(\frac{3R}{\sqrt{\log(1+\varepsilon^2)}}\right)^d = \exp\left(d\log\frac{O(1)\cdot R}{\sqrt{d}}\right) \quad \text{for } R>C\sqrt{d}, \end{aligned}$$

by choosing $1+\varepsilon^2 = e^d$.

Therefore, generalized Fano gives

$$r_\pi^* \gtrsim \sup_{\Delta>0} \Delta^2\left(1-\left(\frac{\Delta}{R}\right)^d - \underbrace{\sqrt{\left(\frac{\Delta}{R}\right)^d\cdot\exp\left(d\log\frac{O(1)\cdot R}{\sqrt{d}}\right)}}_{\le\frac{1}{2} \text{ for } \Delta=c\sqrt{d} \text{ with a small constant } c}\right)$$

$$= \Omega(d).$$

$\square$

Example 3.2 (ridge bandits). $r_t \sim N(f(\langle \theta^*, a_t \rangle), 1)$ for $\theta^* \sim \mathrm{Unif}(S^{d-1})$,

f: known link function $[-1,1] \to \mathbb{R}$, increasing, and $f(0)=0$.

Target: Define a recursive sequence with a large constant $C>0$:

$$\varepsilon_1 = C\sqrt{\frac{\log(1/\delta)}{d}}, \quad \varepsilon_{t+1}^2 = \varepsilon_t^2 + \frac{C}{d} g(\varepsilon_t)^2 \quad (g(x) := \max\{|f(x)|, |f(-x)|\})$$

Then for any interactive learner,

$$\mathbb{P}(|\langle \theta^*, a_s \rangle| \le \varepsilon_s \text{ for all } 1 \le s \le t) \ge 1 - t\delta.$$

Remark: ① The sequence $\{\varepsilon_t\}$ is a pointwise upper bound on the learning trajectory of any algorithm

② The growth $\varepsilon_{t+1}^2 - \varepsilon_t^2$ increases with t: interactive learning becomes faster and faster!

Intuition. Let $I_t = I(H_t; \theta^*) := I(a^t, r^t; \theta^*)$. Then

$$\begin{aligned} I_{t+1} - I_t &= I(\theta^*; r_{t+1} \mid H_t, a_{t+1}) \\ &\le \mathbb{E}[D_{KL}(N(f(\langle \theta^*, a_{t+1} \rangle), 1) \| N(0,1))] \quad (\text{Golden formula}) \\ &= \frac{1}{2}\mathbb{E}[f(\langle \theta^*, a_{t+1} \rangle)^2]. \end{aligned}$$

We aim to upper bound this information gain. A key observation is that,

$$I(\theta^*; a_{t+1}) \le I(\theta^*; H_t) = I_t,$$

so a_{t+1} is "constrained" in information, and we expect $\langle \theta^*, a_{t+1} \rangle$ to be small.

The intuition is that:

$$I(\theta^*; a) \le d\varepsilon^2 \implies |\langle \theta^*, a \rangle| \le \varepsilon \text{ w.h.p.} \qquad (*)$$

If $(*)$ were true, we'll get the recursion by the correspondence $I_t \asymp d\varepsilon_t^2$.

However, mutual info is not strong enough to ensure $(*)$: Fano only gives

$$\mathbb{P}(|\langle \theta^*, a \rangle| \le \varepsilon) \ge 1 - \underbrace{\frac{I(\theta^*; a) + \log 2}{c\, d\varepsilon^2}}_{\text{not small enough to apply union bound!}}$$

Pf using χ^2-info. Let $E_t = \bigcap_{s\le t} \{ |\langle \theta^*, a_s \rangle| \le \varepsilon_s \}$. Define a slight variant of χ^2-info as

$$I_{\chi^2}(X;Y|E) = \inf_{Q_Y} \chi^2(P_{XY|E} \,\|\, P_X Q_Y).$$

then we can still get

$$\mathbb{P}(|\langle \theta^*, a\rangle| \le \varepsilon \mid E) \ge 1 - c_1 \underbrace{e^{-c_0 d \varepsilon^2}}_{\substack{\text{for fixed } a, \\ \mathbb{P}(|\langle \theta^*, a\rangle| \le \varepsilon) \le e^{-c_0 d\varepsilon^2}}} \sqrt{I_{\chi^2}(\theta^*; a|E)+1}.$$

The crux of the proof is to establish the following recursion:

$$I_{\chi^2}(\theta^*; H_t | E_T) + 1 \le \frac{e^{g(\varepsilon_t)^2}}{P(E_T|E_{T-1})^2} \left(I_{\chi^2}(\theta^*; H_{t-1}|E_{T-1}) + 1 \right). \qquad (*)$$

If $(*)$ holds, then $I_{\chi^2}(\theta^*; H_t | E_T) \le \prod_{s=1}^{t} \frac{e^{g(\varepsilon_s)^2}}{P(E_s|E_{s-1})^2} = \frac{1}{P(E_t)^2} \exp\left(\sum_{s\le t} g(\varepsilon_s)^2 \right)$,

so

$$\begin{aligned}
\mathbb{P}(E_{t+1}|E_T) &\ge 1 - c_1 e^{-c_0 d \varepsilon_{t+1}^2} \sqrt{I_{\chi^2}(\theta^*; a_{t+1}|E_t)+1} \\
&\ge 1 - c_1 e^{-c_0 d\varepsilon_{t+1}^2} \sqrt{I_{\chi^2}(\theta^*; H_t | E_T) + 1} \qquad (\text{DPI}) \\
&\ge 1 - \frac{C_1}{P(E_t)} \exp\Big(\underbrace{-c_0 d\, \varepsilon_{t+1}^2 + \tfrac{1}{2} \sum_{s\le t} g(\varepsilon_s)^2}_{\text{recursion ensures that } \le -c_0 d \varepsilon_i^2 \le -c' \log(\frac{1}{\delta})} \Big)
\end{aligned}$$

$$\Rightarrow \quad \mathbb{P}(E_{t+1}) = P(E_t)\cdot \mathbb{P}(E_{t+1}|E_T) \ge P(E_t) - \delta. \qquad \boxed{\text{Q.E.D.}}$$

Pf of $(*)$.

$$\begin{aligned}
I_{\chi^2}(\theta^*; H_t|E_T) + 1 &= \inf_{Q_{H_t}} \int \frac{\mathbb{P}(\theta^*, H_t | E_T)^2}{\pi(\theta^*)\, Q_{H_t}(H_t)} \, d\theta^* da^t dr^t \\
&\le \inf_{Q_{H_{t-1}}} \int \frac{\left[\frac{1(E_t)}{P(E_t)} \pi(\theta^*) \prod_{s=1}^{t} \left(P_s(a_s|H_{s-1}) \varphi(r_s - f(\langle\theta^*, a_s\rangle)) \right) \right]^2}{\pi(\theta^*)\, Q_{H_{t-1}}(H_{t-1}) \cdot P_t(a_t|H_{t-1}) \varphi(r_t)} d\theta^* da^T dr^T \\
&= \inf_{Q_{H_{t-1}}} \int \frac{\Big[\underbrace{\frac{1(E_t)}{P(E_T)}}_{\le \frac{1(E_{t-1})}{P(E_{t-1})} \cdot \frac{1}{P(E_T|E_{t-1})}} \pi(\theta^*) \prod_{s=1}^{t-1} \left(P_s(a_s|H_{s-1}) \varphi(r_s - f(\langle\theta^*, a_s\rangle)) \right) \Big]^2}{\pi(\theta^*)\, Q_{H_{t-1}}(H_{t-1})} \cdot P_t(a_t|H_{t-1}) e^{\overbrace{f(\langle\theta^*, a_t\rangle)^2}^{\le g(\varepsilon_t)^2 \text{ on } E_T}} d\theta^* da^t dr^{t-1} \\
&\le \frac{\exp(g(\varepsilon_t)^2)}{P(E_T|E_{T-1})^2} \left(I_{\chi^2}(\theta^*; H_{t-1}|E_{t-1}) + 1 \right) \qquad \boxed{\text{Q.E.D.}}
\end{aligned}$$