Lecture 12: Global Fano's method

Lecturer: Yanjun Han

May 5, 2021

Announcements

- HW2 due tonight
- HW3 (last HW) released today

Today's plan

Applying Fano's method without manually constructing the hypotheses or upper bounding mutual information:

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

Covering and packing

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

Definition (covering)

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

Definition (packing)

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

Covering and packing number

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

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

A basic property

Lemma

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

- proof of first inequality:

  $M \geq N+1 \Rightarrow$ at least two points in 2ε-packing belongs to the same ε-ball in the covering

  $\Rightarrow d(x_i, x_j) \leq d(x_i, c) + d(x_j, c) \leq 2\varepsilon$

- proof of second inequality:

  any maximal ε-packing is an ε-covering.

  if not, assume that $x \in A$ s.t. $d(x, x_i) > \varepsilon$

  $\Rightarrow$ add x to the packing is still ε-packing

The volume bound

Lemma

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

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

where B is the unit ball under $\|\cdot\|$.

$A + B = \{a + b : a \in A, b \in B\}$

$A \subseteq \bigcup_{i=1}^{N} B(x_i, \varepsilon)$

$\bigcup_{i=1}^{M} B(x_i, \frac{\varepsilon}{2}) \subseteq A + \frac{\varepsilon B}{2}$

The special case where $A = rB$:

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

so $\log M(\varepsilon) \asymp \log N(\varepsilon) \asymp d \log(1 + r/\varepsilon)$.

More results and properties

Theorem (ℓ_q-covering/packing entropy of ℓ_p ball)

For $0 < p < q \leq \infty$ and dimension d,

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

Theorem (duality)

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

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

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

Example: Gaussian mean estimation under L_p loss

- model: $X_1, \cdots, X_n \sim \mathcal{N}(\theta, I_d)$ with unknown $\theta \in \mathbb{R}^d$
- target: estimate θ under general L_p loss $L(\theta, T) = \|T - \theta\|_p$, $p \in [1, \infty]$

Claim

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

Applying packing to GLM when $p > 2$

- construct hypotheses in $\{\theta : \|\theta_2\|_2 \le r\}$
- mutual information bound:

$$I(\theta; X) \le \mathbb{E}[D_{\mathsf{KL}}(\mathcal{N}(\theta, I_p)^{\otimes n} \| \mathcal{N}(0, I_p)^{\otimes n})] \le \frac{nr^2}{2}$$

- Fano's inequality: for any $\delta > 0$,

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

- choice of parameters (δ, r):
  - for $2 < p < \infty$, choose $r \asymp \sqrt{d/n}, \delta \asymp d^{1/p}/\sqrt{n}$
  - for $p = \infty$, choose $r \asymp \delta \asymp \sqrt{(\log d)/n}$

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

- question: if $\theta \in \Theta$ almost surely, could we find an upper bound on $I(\theta; X)$?
- in information theory, this is the channel capacity

Definition (KL covering number)

For $\varepsilon > 0$, let $N_{\mathrm{KL}}(\Theta, \varepsilon)$ be the smallest integer n such that there exist distributions $Q_1, \cdots, Q_n$ on $\mathcal{X}$ such that

$$\sup_{\theta \in \Theta} \min_{i \in [n]} D_{\mathrm{KL}}(P_\theta \| Q_i) \leq \varepsilon^2.$$

Theorem

$$I(\theta; X^n) \leq \inf_{\varepsilon > 0} (n\varepsilon^2 + \log N_{\mathrm{KL}}(\varepsilon)).$$

Proof of the capacity upper bound

- first step: golden rule of mutual information

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

- second step: lower bound sum by maximum

$$\log \frac{P_\theta^{\otimes n}(x^n)}{N^{-1} \sum_{i=1}^{N} Q_i^{\otimes n}(x^n)} \leq \min_{i \in [N]} \log \frac{P_\theta^{\otimes n}(x^n)}{Q_i^{\otimes n}(x^n)} + \log N$$

- concluding step: use the covering property

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

$\leq \varepsilon^2$

Global Fano's method

steps of global Fano's method:

- fix some $\delta > 0$ and $\Theta_0 \subseteq \Theta$, find a δ-packing of Θ_0 under the following (pseudo-)metric

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

- fix some $\varepsilon > 0$, find the KL covering of Θ_0
- apply Fano's method:

$$R_n^\star \geq \frac{\delta}{2}\left(1 - \frac{\log N_{\mathrm{KL}}(\Theta_0, \varepsilon) + n\varepsilon^2 + \log 2}{\log M(\Theta_0, d, \delta)}\right)$$

- choose the parameters appropriately to make the above bound as large as possible

Example I: nonparametric density estimation

- model: $X_1, \cdots, X_n \sim f$ with f supported on $[0,1]^d$ and Hölder smooth with smoothness parameter $s \in [0, \infty)$
- target: estimate the density f under L_p loss, with $p \in [1, \infty)$
- claim:

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

Theorem (Kolmogorov–Tikhomirov)

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

Application of global Fano

- KL covering: assuming $f \geq 1/2$ everywhere (denoted by $\mathcal{F}_{\mathrm{L}}$), then

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

  so $\log N_{\mathrm{KL}}(\mathcal{H}_d^s \cap \mathcal{F}_{\mathrm{L}}, \varepsilon) \asymp \log N(\mathcal{H}_d^s \cap \mathcal{F}_{\mathrm{L}}, \|\cdot\|_2, \varepsilon)$
- L_p packing: compute $\log M(\mathcal{H}_d^s \cap \mathcal{F}_{\mathrm{L}}, \|\cdot\|_p, \delta)$
- Fano's inequality:

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

- choice of parameters: $\varepsilon \asymp \delta \asymp n^{-\frac{s}{2s+2d}}$

Example II: isotonic regression

- model: $X_1, \cdots, X_n \sim P_X$ with bounded density on $[0,1]$, and $Y_i \sim \mathcal{N}(f(X_i), 1)$
- shape constraint: $f : [0,1] \to [0,1]$ is non-decreasing
- target: estimate f under L_p loss, i.e. $L(f, T) = \|f - T\|_p$ with $p \in [1, \infty)$

Claim

$$R^\star_{n,p} \asymp n^{-1/3}$$

Covering entropy of monotone functions

- KL covering: as P_X upper bounded from above,

$$D_{\mathrm{KL}}(P_f \| P_{f'}) = \frac{1}{2}\|f - f'\|^2_{L_2(P_X)} \lesssim \|f - f'\|_2^2$$

  so $\log N_{\mathrm{KL}}(\mathcal{F}_{\mathrm{M}}, \varepsilon) \lesssim \log N(\mathcal{F}_{\mathrm{M}}, \|\cdot\|_2, \varepsilon)$
- L_p packing: compute $\log M(\mathcal{F}_{\mathrm{M}}, \|\cdot\|_p, \delta)$
- global Fano's method:

$$R^\star_{n,p} \gtrsim \delta\left(1 - \frac{\log N(\mathcal{F}_{\mathrm{M}}, \|\cdot\|_2, \varepsilon) + n\varepsilon^2 + \log 2}{\log M(\mathcal{F}_{\mathrm{M}}, \|\cdot\|_p, \delta)}\right)$$

Theorem

$$\log N(\mathcal{F}_{\mathrm{M}}, \|\cdot\|_p, \varepsilon) \asymp \frac{1}{\varepsilon}$$

Example III: convex regression

- model: $X_1, \cdots, X_n \sim P_X$ with bounded density on $[0,1]^d$, and $Y_i \sim \mathcal{N}(f(X_i), 1)$
- shape constraint: $f : [0,1]^d \to [0,1]$ is convex
- target: estimate f under L_p loss, i.e. $L(f, T) = \|f - T\|_p$ with $p \in [1, \infty)$

Claim

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

Covering entropy of convex functions

- KL covering: as P_X upper bounded from above,

$$D_{\mathrm{KL}}(P_f \| P_{f'}) = \frac{1}{2}\|f - f'\|^2_{L_2(P_X)} \lesssim \|f - f'\|_2^2$$

  so $\log N_{\mathrm{KL}}(\mathcal{F}_C, \varepsilon) \lesssim \log N(\mathcal{F}_C, \|\cdot\|_2, \varepsilon)$
- L_p packing: compute $\log M(\mathcal{F}_C, \|\cdot\|_p, \delta)$
- global Fano's method:

$$R^\star_{n,d,p} \gtrsim \delta\left(1 - \frac{\log N(\mathcal{F}_C, \|\cdot\|_2, \varepsilon) + n\varepsilon^2 + \log 2}{\log M(\mathcal{F}_C, \|\cdot\|_p, \delta)}\right)$$

Theorem

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

domain $[0,1]^d$ ball $\Rightarrow \varepsilon^{-(d-1)}$

Example IV: sparse linear regression/prediction

- model: design matrix $X \in \mathbb{R}^{n \times d}$, response $Y \sim \mathcal{N}(X\theta, I_n)$ with unknown θ
- assumption: $\|\theta\|_q \leq R$ for some $q \in (0, 1)$
- target: estimation error $L_{\text{est}}(\theta, T) = \|T - \theta\|_p$ with $p \in [1, \infty]$, or prediction error $L_{\text{pre}}(\theta, T) = \|X(T - \theta)\|_2/\sqrt{n}$

Claim

Under appropriate conditions,

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

$$R^\star_{n,d,q,R}(\text{prediction}) \asymp R^{q/2} \left(\frac{\log d}{n}\right)^{\frac{2-q}{4}}$$

Estimation error: mild assumption on X

- L_p-packing:

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

- KL covering: $D_{\mathrm{KL}}(P_\theta \| P_{\theta'}) = \|X(\theta - \theta')\|_2^2/2$
- approximation theory gives that

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

  provided that $|||X|||_{1\to 2} = \max_{j\in[d]} \|X_j\|_2 = O(\sqrt{n})$
- global Fano's method:

$$R^\star_{n,d,p,q,R}(\text{estimation}) \gtrsim \delta \left(1 - \frac{\log N(B_q(R), \|\cdot\|_2, \varepsilon) + n\varepsilon^2 + \log 2}{\log M(B_q(R), \|\cdot\|_p, \delta)}\right)$$

Prediction error: strong assumption on X

- same KL covering
- now packing becomes

$$\log M(X \cdot B_q(R), \|\cdot\|_p, \delta)$$

but $(\mathbb{R}^n, \|\cdot\|_p)$ not a Hilbert space, and we need lower bound

- strong assumption: assume that

$$\|X(\theta - \theta')\|_2 \geq \kappa \cdot \sqrt{n}\|\theta - \theta'\|_2 - \text{lower order terms}$$

for all $\theta, \theta' \in B_q(R)$ - some restricted eigenvalue (RE) condition

- consequently,

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

References: covering and packing

- Carsten Schtt. "Entropy numbers of diagonal operators between symmetric Banach spaces." *Journal of approximation theory* 40, no. 2 (1984): 121–128.
- O. Guedon and A. E. Litvak. "Euclidean projections of p-convex body." In *Geometric aspects of functional analysis*, pages 95–108. Springer-Verlag, 2000.
- Shiri Artstein, Vitali Milman, and Stanisaw J. Szarek. "Duality of metric entropy." *Annals of mathematics* (2004): 1313-1328.
- Richard M. Dudley, Hiroshi Kunita, and Franois Ledrappier. *Ecole dEte de Probabilites de Saint-Flour XII*, 1982. (Section 7.3)

References: upper bounds using different coverings

- Lucien Birgé. "Approximation dans les espaces métriques et théorie de l'estimation." *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 65, no. 2 (1983): 181-237. (In French; could also refer to Chapter 18 of `http://www.stat.yale.edu/~yw562/teaching/it-stats.pdf`)
- Yuhong Yang and Andrew Barron. "Information-theoretic determination of minimax rates of convergence." *Annals of Statistics* (1999): 1564-1599.

References: global Fano examples

- Qiyang Han and Jon A. Wellner. "Multivariate convex regression: global risk bounds and adaptation." *arXiv preprint* arXiv:1601.06844 (2016).
- Gil Kur, Yuval Dagan, and Alexander Rakhlin. "Optimality of maximum likelihood for log-concave density estimation and bounded convex regression." *arXiv preprint* arXiv:1903.05315 (2019).
- Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls." *IEEE transactions on information theory* 57, no. 10 (2011): 6976-6994.

Next lecture: compression-based arguments in convex optimization