Infotheory for Statistics and Learning

Lecture 9

- Minimax bounds[1]
  - From estimation to testing
  - Packing and metric entropy
  - Fano's method
  - Yang-Barron method

[1] based on notes by J. Duchi and Y. Wu and book by M. Wainwright

Generalized Framework of Statistical Decision Problem

- $\mathcal{P}$ denotes class of distributions defined on sample space $\mathcal{X}$.
- $\theta : \mathcal{P} \to \Theta$ denotes function that maps distribution P on $\theta(P)$
  - A generalized framework, since $\theta(P)$ might not uniquely determine P (i.e. $\theta(P_1) = \theta(P_2)$ iff $P_1 = P_2$). Previously, θ parametrized the family of distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$.
- **IID data:** $x^n = (x_1, \ldots, x_n)$ are n iid observations $X_i \sim P$
- **Estimator:** measurable function $\hat{\theta} : \mathcal{X}^n \to \Theta$

Minimax risk: Let $\rho : \Theta \times \Theta \to \mathbb{R}_+$ be a metric and $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ a non-decreasing function (e.g. $\rho(\theta, \theta')) = |\theta - \theta'|$ and $\Phi(t) = t^2$). The minimax risk[2] $\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho)$ is defined as

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\Phi(\rho(\hat{\theta}(X^n), \theta(P))) \right]$$

[2] Notation $\theta(\mathcal{P})$ means we consider $\theta(P)$ for $P \in \mathcal{P}$; $\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho)$ corresponds to R^* previously and will be abbreviated with $\mathfrak{M}_n$.

From estimation to testing

Key idea: *Reduce* estimation problem to testing problem which allows to lower bound estimation risk by testing error probability!

Construction of hypothesis testing problem:

1. Let $\{P_v\}_{v\in\mathcal{V}}$ denote finite set of distributions $P_v \in \mathcal{P}$ for all $v \in \mathcal{V}$ with finite index set $\mathcal{V}$.
   - Induced $\{\theta(P_v)\}_{v\in\mathcal{V}}$ parameter set is called a 2δ-**packing** if

$$\rho(\theta(P_v), \theta(P_{v'})) > 2\delta \qquad \forall v \neq v'$$

2. Assume RV V uniformly distributed over $\mathcal{V}$ that chooses P_v if $V = v$; samples $x^n = (x_1, \ldots, x_n)$ are then iid drawn $X_i \sim P_v$
3. Let $\Psi : \mathcal{X}^n \to \mathcal{V}$ denote an arbitrary but fixed test function to guess v given x^n with error probability $\mathbb{P}[\Psi(X^n) \neq V]$.

Theorem

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\Psi} \mathbb{P}[\Psi(X^n) \neq V] \qquad (1)$$

Proof

- For arbitrary but fixed P, θ and $\hat{\theta}$ we have

$$\mathbb{E}\left[\Phi(\rho(\hat{\theta},\theta))\right] \geq \mathbb{E}\left[\Phi(\delta)\,\mathbb{1}\{\rho(\hat{\theta},\theta) \geq \delta\}\right] = \Phi(\delta)\,\mathbb{P}\left[\rho(\hat{\theta},\theta) \geq \delta\right]$$

- For testing fct $\Psi(\hat{\theta}) = \arg\min_{v\in\mathcal{V}}\{\rho(\hat{\theta},\theta_v)\}$ with $\theta_v = \theta(P_v)$
  - if $\rho(\hat{\theta},\theta_v) < \delta$, then $\Psi(\hat{\theta}) = v$ since Δ-ineq & 2δ-packing implies $\rho(\hat{\theta},\theta_{v'}) \geq \rho(\theta_v,\theta_{v'}) - \rho(\hat{\theta},\theta_v) > 2\delta - \delta = \delta$, $\forall v' \neq v$
  - equivalently $\Psi(\hat{\theta}) \neq v$ implies $\rho(\hat{\theta},\theta_v) \geq \rho$ so that we have $\mathbb{P}[\rho(\hat{\theta},\theta_v) \geq \delta | V = v] \geq \mathbb{P}[\Psi(\hat{\theta}) \neq v | V = v]$

$$\sup_{P\in\mathcal{P}} \mathbb{E}\left[\Phi(\rho(\hat{\theta},\theta(P)))\right] \geq \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \Phi(\delta)\mathbb{P}\left[\rho(\hat{\theta},\theta_v) \geq \delta \middle| V = v\right]$$

$$\geq \Phi(\delta)\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathbb{P}\left[\Psi(\hat{\theta}) \neq v \middle| V = v\right] \geq \Phi(\delta)\inf_{\Psi}\mathbb{P}\left[\Psi(\hat{\theta}) \neq V\right]$$

- Result follows taking the infimum over all estimators $\hat{\theta}$. $\square$

Remaining challenge and outlook

$$\mathfrak{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\Psi} \mathbb{P}[\Psi(X^n) \neq V]$$

Remaining challenges for minimax lower bound:

1. Find a good 2δ-packing
   - larger δ results in larger factor $\Phi(\delta)$
2. Find a good lower bound on the error probability
   - packing with uniform error probability seems desirable

Outlook

- Packing: metric entropy and packing numbers
- Fano's method: $|\mathcal{V}| \geq 2$ and multiple hypothesis test
- Le Cam's method: $|\mathcal{V}| = 2$ and binary hypothesis test (lect 10)
- Assouad's method:$|\mathcal{V}| = 2^d$ and multiple binary hypothesis tests (lect 10)

Covering - Metric entropy

Q: How many balls of radius δ are needed to cover the space Θ?

Definition

The set $\{\theta_1, \ldots, \theta_N\}$ is a δ-cover of the non-empty set Θ with respect to metric ρ if for any point $\theta \in \Theta$ there exists a $v \in \{1, \ldots, N\}$ such that $\rho(\theta, \theta_v) \leq \delta$. The δ-covering number is

$$N(\delta, \Theta, \rho) = \inf\{N \in \mathbb{N} : \exists\ \delta\text{-cover}\{\theta_1, \ldots, \theta_N\} \text{ of } \Theta\}.$$

Then the metric entropy of Θ is defined as $\log N(\delta, \theta, \rho)$.

Example: Unit cubes in $\mathbb{R}$: $N(\delta, [-1,1], |\cdot|) \leq \frac{1}{\delta} + 1$

- Let Θ be interval $[-1, 1] \subset \mathbb{R}$ and metric $\rho(\theta, \theta') = |\theta - \theta'|$.
- Divide interval in $L = \lfloor \frac{1}{\delta} \rfloor + 1$ sub-intervals with center-points $\theta_i = -1 + 2(i-1)\delta$ for all $i = 1, \ldots, L$ gives result.
- HW: $N(\delta, [-1,1]^d, \|\cdot\|_\infty) \leq (\frac{1}{\delta} + 1)^d$ for unit cubes in $\mathbb{R}^d$.

Packing

Q: How many balls of radius δ can be disjointly placed in space Θ?

Definition

The set $\{\theta_1, \dots, \theta_M\}$ is a δ-packing of the non-empty set Θ with respect to metric ρ if for all distinct v, $v' \in \{1, 2, \dots, M\}$ such that $\rho(\theta_v, \theta_{v'}) > \delta$. The δ-packing number is

$$M(\delta, \Theta, \rho) = \sup\{M \in \mathbb{N} : \exists\ \delta\text{-packing}\{\theta_1, \dots, \theta_M\} \text{ of } \Theta\}.$$

Lemma

$$M(2\delta, \Theta, \rho) \leq N(\delta, \Theta, \rho) \leq M(\delta, \Theta, \rho)$$

Proof: HW!

- *Example:* Unit cubes in $\mathbb{R}$: $M(2\delta, [-1, 1], |\cdot|) \geq \lfloor \frac{1}{\delta} \rfloor$ since $|\theta_i - \theta_j| \geq 2\delta > \delta$ for all $i \neq j$ (θ_i defined as before).

Example: Covering of a parametric function family

- Consider function class $\mathcal{P} = \{f_\theta : [0,1] \to \mathbb{R} : \theta \in [0,1]\}$ with $f_\theta(x) = 1 - \mathrm{e}^{-\theta x}$ and norm $\|f - g\|_\infty = \sup_{x \in [0,1]} |f(x) - g(x)|$.

$$1 + \left\lfloor \frac{1 - 1/\mathrm{e}}{2\delta} \right\rfloor \leq N(\delta, \mathcal{P}, \|\cdot\|_\infty) \leq \frac{1}{2\delta} + 2$$

- *Upper bound:* Set $T = \lfloor \frac{1}{2\delta} \rfloor$ and $\theta_k = 2\delta k$, for $k = 0, 1, \dots, T$ and $\theta_{T+1} = 1$. Then $\{f_{\theta_0}, \dots, f_{\theta_{T+1}}\}$ forms a δ-cover of $\mathcal{P}$.
  - For any $f_\theta \in \mathcal{P}$ there exists θ_k such that $|\theta_k - \theta| \leq \delta$

  $\Rightarrow$ $\|f_{\theta_k} - f_\theta\|_\infty = \max_{x \in [0,1]} |\mathrm{e}^{-\theta_k x} - \mathrm{e}^{-\theta x}| \leq |\theta_k - \theta| \leq \delta$

  $\Rightarrow$ $N(\delta, \mathcal{P}, \|\cdot\|_\infty) \leq T + 2 \leq \frac{1}{2\delta} + 2$
- *Lower bound:* Construct a packing as follows; set $\theta_0 = 0$ and $\theta_k = -\log(1 - \delta k)$ for k as long as $-\log(1 - \delta k) \leq 1$
  - i.e., $k \leq T$ for $\frac{1}{\mathrm{e}} = 1 - \delta T$. Note that $T \geq \lfloor \frac{1-1/\mathrm{e}}{\delta} \rfloor$.
  - We have $\|f_{\theta_s} - f_{\theta_t}\|_\infty \geq |f_{\theta_s}(1) - f_{\theta_t}(1)| \geq \delta \ \forall s \neq t$

  $\Rightarrow$ $M(\delta, \mathcal{P}, \|\cdot\|_\infty) \geq \lfloor \frac{1-1/\mathrm{e}}{\delta} \rfloor + 1$ so that by the previous Lemma $N(\delta, \mathcal{P}, \|\cdot\|_\infty) \geq M(2\delta, \mathcal{P}, \|\cdot\|_\infty) \geq \lfloor \frac{1-1/\mathrm{e}}{2\delta} \rfloor + 1$.

Volume ratios and metric entropy

Obviously, the volume of the set Θ governs the metric entropy. This can be made more precise if the set Θ is a unit ℓ_q-ball:

$$\mathbb{B}_q^d = \{x \in \mathbb{R}^d \,:\, \|x\|_q \leq 1\}$$

Lemma

$$\left(\frac{1}{\delta}\right)^d \;\leq\; N(\delta, \mathbb{B}_q^d, \|\cdot\|_q) \;\leq\; \left(1 + \frac{2}{\delta}\right)^d$$

- Thus, the metric entropy $\log N(\delta, \mathbb{B}_q^d, \|\cdot\|_q)$ scales linearly with the dimension d and logarithmically with $1/\delta$.
- $\left(\frac{1}{\delta}\right)^d \frac{\mathrm{vol}(\mathbb{B}_q^d)}{\mathrm{vol}(\mathbb{B}_p^d)} \;\leq\; N(\delta, \mathbb{B}_q^d, \|\cdot\|_p\} \;\leq\; \frac{\mathrm{vol}(\frac{2}{\delta}\mathbb{B}_q^d + \mathbb{B}_p^d)}{\mathrm{vol}(\mathbb{B}_p^d)}$ actually holds.

Proof of Lemma

- **Lower bound:** Let $\{v_1 \dots v_N\}$ be a δ-cover of $\mathbb{B}_q^d$, then

$$\mathrm{vol}(\mathbb{B}_q^d) \leq \sum_{i=1}^{N} \mathrm{vol}(\delta\mathbb{B}_q^d + v_i) = N\mathrm{vol}(\delta\mathbb{B}_q^d) = N\mathrm{vol}(\mathbb{B}_q^d)\delta^d$$

- **Upper bound:** Let $\mathcal{V}$ be $\delta/2$-packing with maximal cardinality. Then $N(\delta, \mathbb{B}_q^d, \|\cdot\|_q^d) \geq M(\delta, \mathbb{B}_q^d, \|\cdot\|_q^d) = |\mathcal{V}|$. The balls $\{\frac{\delta}{2}\mathbb{B}_q^d + v_i\}_{i=1}^M$ are all disjoint and are contained in $\mathbb{B}_q^d + \frac{\delta}{2}\mathbb{B}_q^d$.

$$\begin{aligned}\sum_{i=1}^{M} \mathrm{vol}(\tfrac{\delta}{2}\mathbb{B}_q^d + v_i) &= M\left(\tfrac{\delta}{2}\right)^d \mathrm{vol}(\mathbb{B}_q^d) \\ &\leq \mathrm{vol}\left(\mathbb{B}_q^d + \tfrac{\delta}{2}\mathbb{B}_q^d\right) = \left(1 + \tfrac{\delta}{2}\right)^d \mathrm{vol}(\mathbb{B}_q^d)\end{aligned}$$

- Divide both sides by $\mathrm{vol}(\mathbb{B}_q^d)$ give the bounds. $\square$

Fano's method - Testing on packings with $|\mathcal{V}| > 2$

- Fano's inequality: For any Markov chain $V - X - \hat{V}$ we have

$$h_2(\mathbb{P}(V \neq \hat{V})) + \mathbb{P}(V \neq \hat{V}) \log(|\mathcal{V}| - 1) \geq H(V|\hat{V})$$

  - For V uniformly distributed Fano's ineq implies

$$\mathbb{P}(V \neq \hat{V}) \geq 1 - \frac{I(V;X) + \log 2}{\log(|\mathcal{V}|)} \tag{2}$$

    since $h_2(\mathbb{P}(V \neq \hat{V})) \leq \log 2$, $\log(|\mathcal{V}| - 1) \leq \log(|\mathcal{V}|) = H(V)$, $H(V|\hat{V}) = H(V) - I(V;\hat{V})$ and $I(V;\hat{V}) \leq I(V;X)$ due to the data processing inequality.

- Fano's method: Let $\{\theta(P_v)\}_{v \in \mathcal{V}}$ be a 2δ-packing. Assume V is uniformly distributed over $\mathcal{V}$ and data $X \sim P_v$ for $V = v$.

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\Psi} \mathbb{P}(\Psi(X) \neq V) \overset{(2)}{\geq} \Phi(\delta) \left[1 - \frac{I(V;X) + \log 2}{\log(|\mathcal{V}|)}\right]$$

Discussion on Fano's method

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \left[1 - \frac{I(V;X) + \log 2}{\log(|\mathcal{V}|)}\right]$$

- With decreasing δ,
  - $\Phi(\delta)$ decreases (lower bound becomes worse), and
  - the minimum between different P_v becomes smaller which makes the hypothesis testing more challenging (error prob increases). $H(V|X)$ and $|\mathcal{V}|$ will increase so that $I(V;X)/\log(|\mathcal{V}|)$ decreases (lower bound becomes better).
  - Practical attempt on how to deal with tradeoff: Pick 2δ-packing with δ as small as possible but keeping the mutual information sufficiently small, e.g such that $\frac{I(V;X)+\log 2}{\log(|\mathcal{V}|)} \leq \frac{1}{2}$.
- Mixture representation of mutual information $I(V;X)$
  - mixture distribution: $\bar{P} = \sum_v \pi(v) P_v$ ($= P_X$ marginal)

$$I(V;X) = D(P_{XV}||P_X P_V) = \sum_v \pi(v) D(\underbrace{P_{X|V=v}}_{=P_v}||\underbrace{P_X}_{=\bar{P}})$$

Local Fano method (aka *generalized*[3] *Fano method*)

- V is uniformly distributed, i.e. $\pi(v) = \frac{1}{|\mathcal{V}|}$
- Since $-\log(\cdot)$ is convex, Jensen's inequality implies

$$I(V;X) = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} D(P_v||\bar{P}) \leq \frac{1}{|\mathcal{V}|^2}\sum_{v,v'\in\mathcal{V}} D(P_v||P_{v'})$$

- Local packing is a 2δ-packing $\{P_v\}_{v\in\mathcal{V}}$, i.e. we have $\rho(\theta(P_v),\theta(P_{v'})) \geq 2\delta$ for all $v \neq v'$, that additionally satisfies

$$D(P_v||P_{v'}) \leq \kappa^2\delta^2 \quad \forall v,v' \in \mathcal{V} \quad \text{for some } \kappa > 0.$$

- Local Fano method. Find a local packing which additionally satisfies $\log|\mathcal{V}| \geq 2(\kappa^2\delta^2 + \log 2)$, then

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi\circ\rho) \geq \frac{1}{2}\Phi(\delta)$$

- Remaining difficulty is to construct such a packing.

[3] Commonly used name is misleading since approach is based on a weak bound on the mutual information and not a generalization.

Yang-Barron method

- What to do if we cannot construct a concrete local packing?
  - **Idea:** Upper bound $I(V;X)$ that holds for any packing!

Lemma (Yang-Barron method)

Let $N_{KL}(\epsilon, \mathcal{P})$ denote the ϵ-covering of $\mathcal{P}$ using the square-root of the KL-divergence as metric (while it is not a metric), then

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

Bound can be then used in Fano's method given a suitable δ.

- Aim $\frac{I(V;X)+\log(2)}{\log|\mathcal{V}|} \leq \frac{1}{2}$ and $\log|\mathcal{V}| \leq \log M(2\delta, \Theta(\mathcal{P}), \rho)$ results in condition to be satisfied for a choice of (ϵ, δ): $\log M(2\delta, \Theta(\mathcal{P}), \rho) \geq 2(\epsilon^2 + \log N_{KL}(\epsilon, \mathcal{P}) + \log 2)$.
- Practical approach: First choose ϵ_n such that $\epsilon_n \geq N_{KL}(\epsilon_n, \mathcal{P})$, then choose largest $\delta_n > 0$ such that $\log M(2\delta, \Theta(\mathcal{P}), \rho) \geq 4\epsilon_n^2 + 2\log 2$.

Proof Yang-Barron method

- We have $\bar{P} = \frac{1}{|\mathcal{V}|}\sum_v \ P_v$, then for any $P \in \mathcal{P}$ we have[4]

$$I(V;X) = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} D(P_v||\bar{P}) \leq \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} D(P_v||P) \leq \max_{v\in\mathcal{V}} D(P_v||P)$$

- Let $\{P_{\kappa_1},\dots,P_{\kappa_N}\}$ be an ϵ-covering of $\mathcal{P}$ using $\sqrt{D(\cdot||\cdot)}$, i.e. for each P_v there exists P_{κ_n} such that $D(P_v||P_{\kappa_n}) \leq \epsilon^2$.
- Set $P = \frac{1}{N}\sum_{i=1}^N P_{\kappa_i}$, then

$$\begin{aligned} D(P_v||P) &= E_{P_v}\left[\log\frac{\mathrm{d}P_v}{\frac{1}{N}\sum_{i=1}^N \mathrm{d}P_{\kappa_i}}\right] \leq E_{P_v}\left[\log\frac{\mathrm{d}P_v}{\frac{1}{N}\mathrm{d}P_{\kappa_n}}\right] \\ &\leq D(P_v||P_{\kappa_n}) + \log N \leq \epsilon^2 + \log N \end{aligned}$$

- The result follows since the previous holds for all $v \in \mathcal{V}$ and $\epsilon > 0$. $\square$

[4] Reminder, $\bar{P}$ is minimizer of $\min_{P\in\mathcal{P}} \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} D(P_v||P)$.

Reminder

Theorem

$$I(X;Y) = \min_{Q} D(P_{Y|X}||Q|P_X)$$

Proof:

$$\begin{aligned} I(X;Y) &= D(P_{Y|X}||P_Y|P_X) = E \log \frac{P_{Y|X}}{Q} \frac{Q}{P_Y} \\ &= D(P_{Y|X}||Q|P_X) - D(P_Y||Q) \leq D(P_{Y|X}||Q|P_X) \end{aligned}$$

since $D(P_Y||Q) \geq 0$. $\square$

Non-parametric problem: Density estimation

- Given $X_1, \ldots, X_n \overset{iid}{\sim} p_\theta \in \mathcal{P}$ for some $\theta \in \Theta$ and estimate $\hat{p} = \hat{p}(\cdot|x_1, \ldots, x_n)$
- Consider KL divergence $D(p_\theta||\hat{p})$ as loss fct and average risk

$$E_{p_\theta} D(p_\theta||\hat{p}) = \int D(p_\theta||\hat{p}(\cdot|X^n = x^n)) p_\theta^{\otimes n}(\mathrm{d}x^n)$$

- An upper bound for minimax risk:

Theorem (Yang-Barron)

$$\inf_{\hat{p}} \sup_{\theta \in \Theta} E_{p_\theta} D(p_\theta||\hat{p}) \leq \inf_{\epsilon > 0} \frac{1}{n} \log N_{KL}(\epsilon, \mathcal{P}) + \epsilon$$

Proof

- Choose estimator $\hat{p}(\cdot|x^n) = \frac{1}{n}\sum_{i=1}^{n} p_{X_i|X^{i-1}}(\cdot|x^{i-1})$ with $p_{X_i|X^{i-1}}(x_i|x^{i-1}) = \frac{\int \prod_{j=1}^{i} p_\kappa(x_j)\pi(\mathrm{d}\kappa)}{\int \prod_{j=1}^{i-1} p_\kappa(x_j)\pi(\mathrm{d}\kappa)}$.
  - Note, prior $\pi(\kappa)$ is used for the definition of the estimator only.
- Due to convexity (a), chain rule of KL divergence (b) we have

$$E_{p_\theta} D(p_\theta \| \hat{p}) = E_{p_\theta} D\Big(p_\theta \Big\| \frac{1}{n}\sum_{i=1}^{n} p_{X_i|X^{i-1}}\Big)$$

$$\stackrel{(a)}{\leq} \frac{1}{n}\sum_{i=1}^{n} E_{p_\theta} D(p_\theta \| p_{X_i|X_{i-1}}) \stackrel{(b)}{=} \frac{1}{n} E_{p_\theta} D(p_\theta^{\otimes n} \| p_{X^n})$$

- Fix $\epsilon > 0$, let $\{p_{\kappa_1}, \dots p_{\kappa_N}\}$ be an optimal ϵ-covering of $\mathcal{P}$

$$E_{p_\theta} D(p_\theta^{\otimes n} \| p_{X^n}) \leq E_{p_\theta} D\Big(p_\theta^{\otimes n} \Big\| \frac{1}{N}\sum_{i=1}^{N} p_{\kappa_i}^{\otimes n}\Big) = E \log\Big[\frac{p_\theta^{\otimes n}}{\frac{1}{N}\sum_{i=1}^{N} p_{\kappa_i}^{\otimes n}}\Big]$$

$$\leq E \log\Big[\frac{p_\theta^{\otimes n}}{\frac{1}{N} p_{\kappa_k}^{\otimes n}}\Big] \leq \log N + n\epsilon \text{ since } \exists k : D(p_\theta \| p_{\kappa_k}) \leq \epsilon \qquad \square$$