Lecture 10: Testing multiple hypotheses, Fano and Assouad

Lecturer: Yanjun Han

April 28, 2021

Today's plan

Test multiple hypotheses:

$$\inf_{T} \sup_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, T)] \geq \inf_{T} \sup_{\theta \in \{\theta_1, \cdots, \theta_m\}} \mathbb{E}_\theta[L(\theta, T)]$$

- pairwise separated hypotheses: Fano's inequality and derivation
- cube-type hypotheses: Assouad's lemma
- mostly separated hypotheses: generalized Fano's inequality
- examples

Fano's inequality

Theorem (Fano's inequality)

Fix any $\theta_1, \cdots, \theta_m \in \Theta$. Suppose that the following separation condition holds:

$$\min_{i \neq j} \min_{a \in \mathcal{A}} L(\theta_i, a) + L(\theta_j, a) \geq \Delta > 0.$$

$U = i \longrightarrow P_{X|U=i} = P_{\theta_i}$

$\{1, 2, \cdots, m\}$

Then

$$\inf_T \max_{\theta \in \{\theta_1, \cdots, \theta_m\}} \mathbb{E}_\theta[L(\theta, T(X))] \geq \frac{\Delta}{2} \left(1 - \frac{I(U; X) + \log 2}{\log m} \right),$$

$I(U; X) \leq \frac{1}{2} \log m$

where $U \sim \mathrm{Unif}(\{\theta_1, \cdots, \theta_m\})$, $P_{X|U=\theta_i} = P_{\theta_i}$.

Mutual information:

$$\begin{aligned} I(X; Y) &= D_{\mathrm{KL}}(P_{XY} \| P_X \times P_Y) \\ &= \min_{Q_Y} D_{\mathrm{KL}}(P_{Y|X} \| Q_Y \mid P_X) = \mathbb{E}_X[D_{\mathrm{KL}}(P_{Y|X} \| Q_Y)] \end{aligned}$$

First step: from estimation to testing

Lemma

Fix any $\theta_1, \cdots, \theta_m \in \Theta$. Suppose that the following separation condition holds:

$$\min_{i \neq j} \min_{a \in \mathcal{A}} L(\theta_i, a) + L(\theta_j, a) \geq \Delta > 0.$$

Then

$$\inf_T \max_{\theta \in \{\theta_1, \cdots, \theta_m\}} \mathbb{E}_\theta[L(\theta, T(X))] \geq \frac{\Delta}{2} \cdot \inf_{\Psi: \mathcal{X} \to [m]} \frac{1}{m} \sum_{i=1}^m P_{\theta_i}(\Psi \neq i)$$

$H_1: \theta = \theta_1, \quad \cdots, \quad H_m: \theta = \theta_m$

$x \in \mathcal{X} \qquad \psi(x) \in [m]$

estimator $T \implies \psi(x) = \arg\min_{i \in [m]} L(\theta_i, T(x))$

$L(\theta_{\psi(x)}, T(x)) \leq L(\theta_i, T(x))$

$L(\theta_i, T(x)) \geq \frac{L(\theta_{\psi(x)}, T(x)) + L(\theta_i, T(x))}{2} \geq \frac{\Delta}{2} \cdot \mathbb{1}(\psi(x) \neq i)$

$\text{LHS} \geq \inf_T \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{\theta_i}[L(\theta_i, T)]$

$\geq \frac{\Delta}{2} \inf_\psi \frac{1}{m} \sum_{i=1}^m P_{\theta_i}(\psi \neq i)$

Lower bound of test error: first proof

- the optimal test:

$$\Psi^\star(x) = \arg\max_{i\in[m]} P_{\theta_i}(x)$$

$$\frac{1}{m}\sum_{i=1}^{m} P_{\theta_i}(\psi \neq i) = 1 - \frac{1}{m}\sum_{i=1}^{m} P_{\theta_i}(\psi = i)$$

$$= 1 - \sum_{x\in X} \frac{1}{m}\sum_{i=1}^{m} P_{\theta_i}(x)\cdot 1(\underbrace{\psi(x)}=i)$$

- the optimal test error:

$$\inf_{\Psi:\mathcal{X}\to[m]} \frac{1}{m}\sum_{i=1}^{m} P_{\theta_i}(\Psi \neq i) = 1 - \underline{\frac{1}{m}\sum_{x\in\mathcal{X}} \max_{i\in[m]} P_{\theta_i}(x)}$$

- our target:

$I(U;X)$

$\shortparallel$

$\overline{P} = \frac{1}{m}\sum_i P_{\theta_i}$

$$\underline{\frac{1}{m}\sum_{x\in\mathcal{X}} \max_{i\in[m]} P_{\theta_i}(x) \le \frac{1}{\log m}\left(\frac{1}{m}\sum_{i=1}^{m} D_{\mathrm{KL}}(P_{\theta_i}\|\overline{P}) + \log 2\right)}$$

Lower bound of test error: first proof (cont'd)

- an equivalent target:

$$\sum_{x\in\mathcal{X}} \overline{P}(x)\cdot \overset{\max_i}{\frac{P_{\theta_i}(x)}{\overline{P}(x)}} \le \sum_{x\in\mathcal{X}} \overline{P}(x)\cdot \underline{\frac{1}{\log m}\left(\sum_{i=1}^{m} \frac{P_{\theta_i}(x)}{\overline{P}(x)} \log \frac{P_{\theta_i}(x)}{\overline{P}(x)} + m\log 2\right)}$$

- suffices to show that for non-negative $x_1, \cdots, x_m$ with $\sum_{i=1}^m x_i = m$, it holds that

$$\max_{i\in[m]} x_i \le \frac{1}{\log m}\left(\sum_{i=1}^{m} x_i \log x_i + m\log 2\right)$$

$$t \triangleq \max_i x_i = x_1$$

$$\begin{aligned}
\sum_i x_i \log x_i &= t\log t + \sum_{i=2}^{m} x_i \log x_i \\
&\ge t\log t + (m-t)\log\frac{m-t}{m-1} \\
&= \underbrace{t\log t + (m-t)\log(m-t)}_{\ge m\log\frac{m}{2}} - \underbrace{(m-t)\log(m-1)}_{\le (m-t)\log m} \ge t\log m - m\log 2
\end{aligned}$$

Lower bound of test error: second proof

- for each $x \in \mathcal{X}$, define the following kernel K_x:

$$
\begin{array}{rccl}
 & [m] & \longrightarrow & \{0,1\} \\
K_x : & \underline{i} & \longmapsto & \underline{\mathbb{1}(\Psi(x) = i)}
\end{array}
$$

- induced distribution mapping:

$$
\begin{aligned}
P_U &\longmapsto \mathrm{Bern}(1/m) \\
P_{U|X=x} &\longmapsto \mathrm{Bern}(p_x) = \mathrm{Bern}(\underline{P_{U|X=x}(U = \Psi(x))})
\end{aligned}
$$

$\mathbb{E}[p_x] = 1 - \underline{\mathbb{P}(U \neq \Psi(X))}$

- data-processing inequality:

$$
D_{\mathrm{KL}}(P_{U|X=x} \| P_U) \geq D_{\mathrm{KL}}(\mathrm{Bern}(p_x) \| \mathrm{Bern}(1/m))
$$

$$
\begin{aligned}
&= p_x \log(m p_x) + (1-p_x) \log \frac{1-p_x}{1-\frac{1}{m}} \\
&= p_x \log m + \underbrace{p_x \log p_x + (1-p_x) \log(1-p_x)}_{\geq -\log 2} - \underbrace{(1-p_x)\log\left(1-\frac{1}{m}\right)}_{<0} \\
&\geq p_x \log m - \log 2
\end{aligned}
$$

$\mathbb{E}[\mathrm{LHS}] = I(X;U)$, $\mathbb{E}[\mathrm{RHS}] = (1-\mathbb{P}(U \neq \Psi(X))) \cdot \log m - \log 2$

Another lower bound of test error

Theorem (tree-based lower bound)

Let $T = ([m], E)$ be any undirected tree on the vertex set $[m]$. Then

$$\inf_{\Psi:\mathcal{X}\to[m]} \frac{1}{m}\sum_{i=1}^{m} P_{\theta_i}(\Psi \neq i) \geq \frac{1}{m}\sum_{(i,j)\in E}\left(1 - \|P_{\theta_i} - P_{\theta_j}\|_{\mathrm{TV}}\right)$$

$\sup_T$

A simple technical lemma (HW3):

Lemma

For any real numbers $x_1, \cdots, x_m$ and any tree $T = ([m], E)$, it holds that

$$\sum_{i=1}^{m} x_i - \max_{i\in[m]} x_i \geq \sum_{(i,j)\in E} \min\{x_i, x_j\}.$$

Assouad's lemma: cube-type separation

Theorem (Assouad's lemma)

Fix any $\{\theta_v\}_{v\in\{\pm1\}^p} \subseteq \Theta$. Suppose that the following separation condition holds:

$$\min_{a\in\mathcal{A}} L(\theta_v, a) + L(\theta_{v'}, a) \geq \Delta \cdot \underbrace{d_{\mathrm{H}}(v, v')}_{=\sum_{j=1}^{p} \mathbb{1}(v_j \neq v_j')}, \quad \forall v, v' \in \{\pm1\}^p.$$

Then

$$\inf_T \max_{\theta\in\{\theta_v\}_{v\in\{\pm1\}^p}} \mathbb{E}_\theta[L(\theta, T(X))] \geq \frac{\Delta}{2}\sum_{j=1}^{p}\left(1 - \|P_{j,+} - P_{j,-}\|_{\mathrm{TV}}\right),$$

$$P_{j,+} = \frac{1}{2^{p-1}} \sum_{v: v_j = 1} P_{\theta_v}$$

$$P_{j,-} = \frac{1}{2^{p-1}} \sum_{v: v_j = -1} P_{\theta_v}$$

x: $\hat{v}(x) = \arg\min_v L(\theta_v, T(x))$

$$L(\theta_v, T) \geq \frac{L(\theta_v, T) + L(\theta_{\hat{v}(x)}, T)}{2} \geq \frac{\Delta}{2} \cdot d_H(v, \hat{v}(x))$$

$$= \frac{\Delta}{2} \sum_{j=1}^{p} \underline{\mathbb{1}(v_j \neq \hat{v}_j(x))}.$$

Corollaries of Assouad's lemma

Corollary 1

$$\inf_T \max_{\theta\in\Theta} \mathbb{E}_\theta[L(\theta, T(X))] \geq \frac{p\Delta}{2}\left(1 - \max_{d_{\mathrm{H}}(v,v')=1} \|P_{\theta_v} - P_{\theta_{v'}}\|_{\mathrm{TV}}\right),$$

Corollary 2

$$\inf_T \max_{\theta\in\Theta} \mathbb{E}_\theta[L(\theta, T(X))] \geq \frac{p\Delta}{2}\left(1 - \sqrt{\mathbb{E}_v \mathbb{E}_j \|P_{\theta_v} - P_{\theta_{v\oplus j}}\|_{\mathrm{TV}}^2}\right),$$

$v \sim \mathrm{Unif}(\{\pm 1\}^p)$

$j \sim \mathrm{Unif}([p])$

Generalized Fano's inequality

Theorem (Generalized Fano's inequality)

Let π be any prior distribution on Θ, and

$$p_\Delta \triangleq \sup_{a \in \mathcal{A}} \pi(\{\theta \in \Theta : L(\theta, a) \le \Delta\}).$$

Then for $U \sim \pi$, we have

$$\inf_T \max_{\theta \in \Theta} \mathbb{E}_\theta[L(\theta, T(X))] \ge \Delta \cdot \left(1 - \frac{I(U; X) + \log 2}{\log(1/p_\Delta)}\right).$$

K_x : $\quad \Theta \longrightarrow \{0,1\}$

$\quad \theta \longmapsto \mathbb{1}(L(\theta, T(x)) \le \Delta)$.

$P_\theta \longmapsto \mathrm{Bern}(p_x)$, $\quad p_x \le p_\Delta$

$P_{\theta|x} \longmapsto \mathrm{Bern}(q_x)$, $\quad \mathbb{E}[q_x] = \mathbb{P}_{\theta|x}(L(\theta, T(x)) \le \Delta)$.

Example I: Gaussian location model

- model: $X \sim \mathcal{N}(\theta, \sigma^2 I_p)$ with unknown $\theta \in \mathbb{R}^p$
- target: estimate θ under quadratic loss $L(\theta, T) = \|T - \theta\|_2^2$
- claim: $R^\star_{p,\sigma} = \Theta(p\sigma^2)$
- failure of two-point method:

$$R^\star_{p,\sigma} \geq \underbrace{\frac{\|\theta_0 - \theta_1\|_2^2}{2}\left(1 - \Phi\left(\frac{\|\theta_0 - \theta_1\|_2}{2\sigma}\right)\right)}_{\text{ind. of } p}$$

Proof I: Fano's inequality

- choice of $\{\theta_1, \cdots, \theta_m\}$: some maximal ℓ_2-packing of $\{\pm\delta\}^p$ with radius $\Omega(\delta\sqrt{p})$ and $m = 2^{\Omega(p)}$ (Gilbert–Varshamov)

  $\|\theta_i - \theta_j\|_2 \geq c \cdot \delta\sqrt{p}$

- separation condition: $\Delta = \Omega(p\delta^2)$
- mutual information:

$$I(U;X) \leq \mathbb{E}_U[D_{\mathrm{KL}}(\mathcal{N}(\theta_U, \sigma^2 I_p) \| \mathcal{N}(0, \sigma^2 I_p))] \leq \frac{p\delta^2}{2\sigma^2}$$

- Fano's inequality:

$$R^\star_{p,\sigma} = \Omega\left(p\delta^2 \cdot \left(1 - \frac{p\delta^2/(2\sigma^2) + \log 2}{\Omega(p)}\right)\right)$$

- choice of δ: $\delta \asymp \sigma$

Proof II: Assoud's lemma

- cube-type hypotheses: $\theta_v = \delta v$, with $v \in \{\pm 1\}^p$
- separation condition: $\Delta \asymp \delta^2$

$$\|\delta v - \delta v'\|_2^2 = 4\delta^2 d_{\mathrm{H}}(v, v')$$

- neighboring TV distance:

$$\max_{d_H(v,v')=1} \|P_{\theta_v} - P_{\theta_{v'}}\|_{\mathrm{TV}} = 2\Phi\left(\frac{\delta}{\sigma}\right) - 1$$

- Assouad's lemma:

$$R^\star_{p,\sigma} = \Omega(p\delta^2 \cdot (1 - \Phi(\delta/\sigma)))$$

- choice of δ: $\delta \asymp \sigma$

Proof III: Generalized Fano's inequality

- let π be the uniform distribution on $\{\pm\delta\}^p$
- upper bound of p_Δ: if $\Delta = p\delta^2/3$, then

$$p_\Delta = \exp(-\Omega(p))$$

- mutual information: for $U \sim \pi$, it holds that

$$I(U;X) \le \frac{p\delta^2}{2\sigma^2}$$

- generalized Fano's inequality:

$$R^\star_{p,\sigma} = \Omega\left(p\delta^2 \cdot \left(1 - \frac{p\delta^2/(2\sigma^2) + \log 2}{\Omega(p)}\right)\right)$$

- choice of δ: $\delta \asymp \sigma$

Example II: sparse linear regression

- model: $Y \sim \mathcal{N}(X\theta, \sigma^2 I_n)$ with fixed design matrix $X \in \mathbb{R}^{n\times p}$, and unknown sparse vector $\theta \in \mathbb{R}^p$ with $\|\theta\|_0 \le s$
- target: estimate θ under quadratic loss $L(\theta, T) = \|\theta - T\|_2^2$

Theorem (Candès and Davenport, 2013)

$$R^\star_{n,p,s,\sigma} = \Omega\left(\underline{\frac{sp\sigma^2 \log(ep/s)}{\|X\|_\mathrm{F}^2}}\right)$$

$\|X\|_F = \sqrt{\mathrm{Tr}(XX^T)}$

$X = I_p$: $R^* = \Omega(s\sigma^2 \log(\frac{ep}{s}))$ sparse mean est.

$X_{i,j} \sim N(0, \frac{1}{n})$: $R^* = \cdots$ compressed sensing

Proof via generalized Fano's inequality

- let π be the uniform distribution on $\{\delta, -\delta, 0\}^p$ with at most s non-zero components
- upper bound of p_Δ: if $\Delta = s\delta^2/12$, then

$$\log(1/p_\Delta) = \Omega(s \log(ep/s))$$

- upper bound of mutual information:

$$I(U; Y) \le \mathbb{E}[D_{\mathrm{KL}}(\mathcal{N}(XU, \sigma^2 I_n) \| \mathcal{N}(0, \sigma^2 I_n))]$$

$$= \frac{\mathbb{E}[\|XU\|_2^2]}{2\sigma^2} = \frac{\mathrm{Tr}(X\,\mathbb{E}[UU^T]\,X^T)}{2\sigma^2} = \frac{s}{2p\sigma^2} \cdot \|X\|_F^2$$

$$\mathbb{E}[UU^T] = \frac{s}{p}\delta^2 \cdot I_p$$

- generalized Fano's inequality:

$$R^\star_{n,p,s,\sigma} = \Omega\left(s\delta^2\left(1 - \frac{s\delta^2\|X\|_F^2/(2p\sigma^2) + \log 2}{\Omega(s\log(ep/s))}\right)\right)$$

Example III: multi-armed bandit

- K arms associated with unknown mean reward $(\mu_1, \cdots, \mu_K) \in [0,1]^K$
- learner pulls arm $\pi_t \in [K]$ at time t
- learner observes a Gaussian random reward $r_t \sim \mathcal{N}(\mu_{\pi_t}, 1)$
- learner's cumulative regret:

$$R_T(\pi) = T \max_{i \in [K]} \mu_i - \sum_{t=1}^{T} \mu_{\pi_t}$$

- claim: $R^\star_{K,T} = \Omega(\sqrt{KT})$

Proof via tree-based lower bound

- construction of reward distribution:

$$
\begin{aligned}
\mu_1 &= (\delta, 0, 0, \cdots, 0) \\
\mu_2 &= (\delta, 2\delta, 0, \cdots, 0) \\
&\vdots \\
\mu_K &= (\delta, 0, 0, \cdots, 2\delta)
\end{aligned}
$$

- separation condition: $\Delta = \delta T$
- lower bound applied to a star tree $([K], \{(1,2), (1,3), \cdots, (1,K)\})$:

$$
R^{\star}_{K,T} \geq \frac{\delta T}{2} \cdot \frac{1}{K} \sum_{i=2}^{K} (1 - \|P^T_{\mu_1} - P^T_{\mu_i}\|_{\mathrm{TV}})
$$

$$
= \frac{\delta T}{4K} \sum_{i=2}^{K} \exp(-D_{KL}(P^T_{\mu_1} \| P^T_{\mu_i}))
$$

$$
= \frac{\delta T}{4K} \sum_{i=2}^{K} \exp\left(-\frac{(2\delta)^2}{2} \cdot \mathbb{E}_1[T_i]\right)
$$

$$
\geq \frac{\delta T}{4K} \cdot (K-1) \exp\left(-2\delta^2 \cdot \underbrace{\frac{1}{K-1} \sum_{i=2}^{K} \mathbb{E}_1[T_i]}\right)
$$

$\sum_{i=2}^{K} T_i \leq T.$

$\delta^2 \asymp \frac{K}{T}$

Example IV: Gaussian mixture estimation

- model: $X_1, \cdots, X_n \sim f$ with f being an unknown Gaussian mixture, i.e. $f = g * \mathcal{N}(0,1)$
- target: estimate f under L_2 loss $L(f,T) = \|f - T\|_2^2$

Theorem (Kim, 2014)

$$R_n^\star = \Theta\left(\frac{(\log n)^{1/2}}{n}\right)$$

Upper bound idea: find a kernel K with $\widehat{K}(\omega) = 1(|\omega| \le 2\sqrt{\log n})$, apply the estimator

$$f_n = \mathbb{P}_n * K = \frac{1}{n}\sum_{i=1}^n K(\cdot - x_i)$$

$\hat{f} = \hat{g} \cdot \frac{1}{\sqrt{2\pi}} e^{-\omega^2/2}$

Construction of cube-type hypotheses

- idea: for $v \in \{\pm 1\}^p$, construct

$$f_v = \left(g_0 + \delta \cdot \sum_{i=1}^{p} v_i g_i \right) * \phi$$

- orthogonality condition:

$$\int_{\mathbb{R}} (g_i * \phi)(x) \cdot (g_j * \phi)(x) dx = 1(i = j),$$

or equivalently,

$$\int_{\mathbb{R}} \widehat{g}_i(\omega) \widehat{g}_j(\omega) \phi(\omega)^2 d\omega = 1(i = j)$$

Choice of g_i

- using orthogonality property of Hermite polynomials and $\phi(\omega)^4 \propto \phi(2\omega)$, one choice would be

$$\widehat{g}_i(\omega) \propto H_{2i-1}(2\omega) \cdot \phi(\omega)$$

- an explicit expression of g_i:

$$g_i(x) = \sqrt{2}(2\pi)^{3/4}\sqrt{\frac{3^{2i-1}}{(2i-1)!}} \cdot \phi(x) H_{2i-1}\left(\frac{2x}{\sqrt{3}}\right)$$

- in particular, $\|g_i\|_\infty \asymp 3^i$, so $p = O(\log n)$

Indistinguishability condition

- neighboring χ^2-divergence:

$$\chi_i^2 \asymp n\delta^2 \cdot \int_{\mathbb{R}} \frac{(g_i * \phi)(x)^2}{g_0 * \phi(x)} dx$$

- choosing $g_0 * \phi$ be the density of $\mathcal{N}(0, \sigma^2)$, then

$$\chi_i^2 = O\left(n\delta^2\sigma\right), \quad \text{if } \sigma = \Omega(\sqrt{p})$$

- final statement of Assouad's lemma:

$$R_n^\star = \Omega(p\delta^2 \cdot (1 - O(n\delta^2\sigma)))$$

  provided that $p = O(\log n)$ and $\sigma = \Omega(\sqrt{p})$
- choice of parameters: $p \asymp \log n, \sigma \asymp \sqrt{\log n}, \delta^2 \asymp 1/(n\sqrt{\log n})$

References

- Bin Yu, "Assouad, Fano, and Le Cam." *Festschrift for Lucien Le Cam*. Springer, New York, NY, 1997. 423–435.
- Xi Chen, Adityanand Guntuboyina, and Yuchen Zhang. "On Bayes risk lower bounds." *The Journal of Machine Learning Research* 17.1 (2016): 7687–7744.
- Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou, "Batched multi-armed bandit problem." *Advances in Neural Information Processing Systems*, 2019.
- Arlene K. H. Kim. "Minimax bounds for estimation of normal mixtures." *Bernoulli* 20.4 (2014): 1802–1818.

Next lecture: more classical examples