Lecture 4: Statistical decision theory: model distance and equivalence

Lecturer: Yanjun Han

April 7, 2021

Announcement

HW1 is released:

- covers lecture 1–4
- due two weeks later (April 21, 11:59 PM)
- submit via Gradescope (entry code: 3YD8J7)
- students enrolled for letter grade are required to complete homeworks
- other students are also encouraged to attempt homeworks

Today's plan

The idea of reduction:

- general setting of statistical decision theory
- deficiency and Le Cam's distance
- examples of asymptotic equivalence
- application I: Hájek–Le Cam theory (lecture 5)
- application II: statistical-computational tradeoff (lecture 6)

Statistical decision theory

$L(\theta, a)$

$$\theta \in \Theta \xrightarrow{\quad X \sim P_\theta \quad} X \in \mathcal{X} \xrightarrow{\quad a \sim \delta(\cdot \mid X) \quad} a \in \mathcal{A}$$

deterministic case: $\hat{\theta} = T(X)$

- Θ: parameter space
- $\mathcal{X}$: observation space
- $\mathcal{A}$: action space
- loss function $L : \Theta \times \mathcal{A} \to \mathbb{R}_+$

Definition (Risk)

The risk of the decision rule δ under loss function L and the true parameter θ is defined as

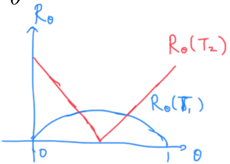
in det. case: $R_\theta(\delta) = \mathbb{E}_{X \sim P_\theta}[L(\theta, T(X))]$.

$$R_\theta(\delta) = \mathbb{E}_{a \sim \delta(\cdot|X)} \mathbb{E}_{X \sim P_\theta}[L(\theta, a)]$$

$p_\theta(x, a) = p_\theta(x)\,\delta(a|x)$

Risk comparison: Bayes and minimax

For two decision rules δ_1, δ_2, typically $R_\theta(\delta_1) \geq R_\theta(\delta_2)$ for some θ, and $R_\theta(\delta_1) \leq R_\theta(\delta_2)$ for other θ



$X_1, \cdots, X_n \overset{i.i.d}{\sim} \mathrm{Bern}(\theta)$

$\Theta = \mathcal{A} = [0,1]$

$T_1(X_1, \cdots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$

$T_2(X_1, \cdots, X_n) \equiv \frac{1}{2}$

- minimax criterion: compare $\max_{\theta \in \Theta} R_\theta(\delta)$
- Bayes criterion: fix a prior π on θ, compare $\mathbb{E}_{\theta \sim \pi}[R_\theta(\delta)]$

Exercise

Bayes estimator is easy to find in principle:

$$T(x) \in \arg\min_{a \in \mathcal{A}} \mathbb{E}_{\theta \sim \pi(\cdot|x)}[L(\theta, a)]$$

$\pi(\theta, x) = \pi(\theta) p_\theta(x)$

Example I: linear regression

$$\theta \in \mathbb{R}^p \longrightarrow (x_1, y_1), \cdots, (x_n, y_n) \longrightarrow \widehat{\theta} \in \mathbb{R}^p$$

$$x_1, \cdots, x_n \sim P_X$$

$$y_i \mid x_i \sim \mathcal{N}(x_i^\top \theta, 1)$$

- estimation error: $L_1(\theta, \widehat{\theta}) = \|\theta - \widehat{\theta}\|_2^2$
- prediction error: $L_2(\theta, \widehat{\theta}) = \mathbb{E}_{(x,y)\sim P_\theta}[(y - x^\top \widehat{\theta})^2]$

Example II: density estimation

$$f \in \mathcal{F} \longrightarrow x_1, \cdots, x_n \sim f \longrightarrow T$$

- loss at a point: $L_1(f, T) = |T - f(0)|$
- global loss: $L_2(f, T) = \int |T(x) - f(x)| dx$
- functional estimation: $L_3(f, T) = |T - \|f\|_2|$

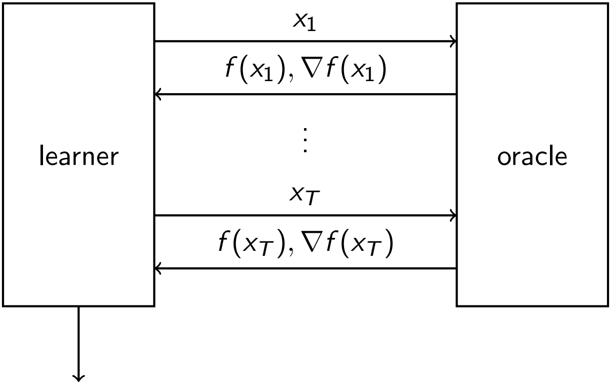
Example III: learning theory

$$P_{XY} \longrightarrow (x_1, y_1), \cdots, (x_n, y_n) \sim P_{XY} \longrightarrow f \in \mathcal{F}$$

- excess risk:

$$L(P_{XY}, f) = \mathbb{E}_{(x,y)\sim P_{XY}}[L_0(f(x), y)] - \min_{f^\star \in \mathcal{F}} \mathbb{E}_{(x,y)\sim P_{XY}}[L_0(f^\star(x), y)]$$

Example IV: optimization



$\widehat{x} \approx \operatorname{argmin}_{x^*} f(x^*)$

- suboptimality gap:

$$L(f, \widehat{x}) = f(\widehat{x}) - \min_{x^\star} f(x^\star)$$

Comparison of models

Motivating question

For two statistical models $\mathcal{M} = (\mathcal{X}, (P_\theta)_{\theta \in \Theta})$ and $\mathcal{N} = (\mathcal{Y}, (Q_\theta)_{\theta \in \Theta})$, when can we say model $\mathcal{M}$ is stronger than model $\mathcal{N}$? How can we translate a solution to model $\mathcal{N}$ to a solution to model $\mathcal{M}$?

Which model is stronger?

- $\mathcal{X} = \mathcal{Y} = \Theta = \mathbb{R}$, $P_\theta = \mathcal{N}(\theta, 0.1), Q_\theta = \mathcal{N}(\theta, 1)$

  P_θ better the Q_θ (P_θ can simulate Q_θ)

- $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, $\Theta = [-1, 1]$, $P_\theta = \text{Uniform}(\{\theta - 0.1, \theta + 0.1\})$, $Q_\theta = \text{Uniform}(\{\theta - 1, \theta + 1\})$

  Q_θ better the P_θ (Q_θ can generate θ, thus P_θ)

Deficiency

Definition (Deficiency; Le Cam (1964))

For two statistical models $\mathcal{M} = (\mathcal{X}, (P_\theta)_{\theta\in\Theta})$ and $\mathcal{N} = (\mathcal{Y}, (Q_\theta)_{\theta\in\Theta})$, we say $\mathcal{M}$ is ε-deficient relative to $\mathcal{N}$ if

- for any finite subset $\Theta_0 \subseteq \Theta$;
- for any finite action space $\mathcal{A}$;
- for any loss function $L : \Theta \times \mathcal{A} \to [0, 1]$;
- for any decision rule $\delta_\mathcal{N}$ for model $\mathcal{N}$;

there exists a decision rule $\delta_\mathcal{M}$ for model $\mathcal{M}$ such that

$$R_\theta(\delta_\mathcal{M}) \le R_\theta(\delta_\mathcal{N}) + \varepsilon, \quad \forall\theta \in \Theta_0.$$

$$\sup_{\Theta_0} \sup_{L} \sup_{\delta_N} \inf_{\delta_M} \sup_{\theta} \left(R_\theta(\delta_M) - R_\theta(\delta_N) \right) \le \varepsilon.$$

hardness result for model $\mathcal{M} \Longrightarrow$ hardness result for model $\mathcal{N}$

Randomization of statistical models

Definition (Randomization)

For two statistical models $\mathcal{M} = (\mathcal{X}, (P_\theta)_{\theta \in \Theta})$ and $\mathcal{N} = (\mathcal{Y}, (Q_\theta)_{\theta \in \Theta})$, we say $\mathcal{N}$ is a randomization of $\mathcal{M}$ if there exists a stochastic kernel $\mathsf{K} : \mathcal{X} \to \mathcal{Y}$ such that $Q_\theta = \mathsf{K} P_\theta$ for all $\theta \in \Theta$, i.e.

↳ ind. of Θ

$$Q_\theta(y) = \sum_{x \in \mathcal{X}} P_\theta(x) \mathsf{K}(y \mid x), \quad \forall y \in \mathcal{Y}, \theta \in \Theta.$$

$\mathcal{N}$ is a randomization of $\mathcal{M}$ $\Longrightarrow$ $\mathcal{M}$ is 0-deficient relative to $\mathcal{N}$

$$\begin{tikzcd} \mathcal{X} \arrow[r, "K"] \arrow[rr, bend right, "\delta_{\mathcal{M}} = \delta_{\mathcal{N}} \circ K"'] & \mathcal{Y} \arrow[r, "\delta_{\mathcal{N}}"] & \mathcal{A} \end{tikzcd}$$

Equivalence of deficiency and randomization

Theorem

Model $\mathcal{M}$ is ε-deficient relative to $\mathcal{N}$ if and only if there exists a stochastic kernel $\mathsf{K} : \mathcal{X} \to \mathcal{Y}$ such that

$$\sup_{\theta \in \Theta} \| Q_\theta - \mathsf{K} P_\theta \|_{\mathsf{TV}} \leq \varepsilon.$$

showing deficiency results $\iff$ showing randomization results

Easy direction: randomization $\Rightarrow$ deficiency

- choose any action space $\mathcal{A}$ and loss L
- fix any decision rule $\delta_{\mathcal{N}}$ for model $\mathcal{N}$
- choose $\delta_{\mathcal{M}} = \delta_{\mathcal{N}} \circ \mathsf{K}$

$$
\begin{aligned}
& R_\theta(\delta_{\mathcal{M}}) - R_\theta(\delta_{\mathcal{N}}) \\
& = \mathbb{E}_{a \sim \delta_{\mathcal{M}}(\cdot|X)} \mathbb{E}_{X \sim P_\theta}[L(\theta, a)] - \mathbb{E}_{a \sim \delta_{\mathcal{N}}(\cdot|Y)} \mathbb{E}_{Y \sim Q_\theta}[L(\theta, a)] \\
\text{def. of } \delta_{\mathcal{M}} \quad & = \mathbb{E}_{a \sim \delta_{\mathcal{N}}(\cdot|Y)} \mathbb{E}_{Y \sim \mathsf{K} P_\theta}[L(\theta, a)] - \mathbb{E}_{a \sim \delta_{\mathcal{N}}(\cdot|Y)} \mathbb{E}_{Y \sim Q_\theta}[L(\theta, a)] \\
& = \mathbb{E}_{a \sim \delta_{\mathcal{N}}(\cdot|Y)} (\mathbb{E}_{Y \sim \underline{\mathsf{K} P_\theta}}[L(\theta, a)] - \mathbb{E}_{Y \sim \underline{Q_\theta}}[L(\theta, a)]) \\
& \leq \|Q_\theta - \mathsf{K} P_\theta\|_{\mathrm{TV}}
\end{aligned}
$$

$\hookrightarrow \|P - Q\|_{TV} = \sup_{0 \le f \le 1} \mathbb{E}_P[f] - \mathbb{E}_Q[f]$

- for simplicity assume that Θ is a finite set
- deficiency: for every $\delta_{\mathcal{N}}$,

$$\underbrace{\sup_{L} \sup_{\pi}}_{\substack{\pi \to \text{prior} \\ = \sup_{L \cdot \pi}}} \inf_{\delta_{\mathcal{M}}} \underbrace{\mathbb{E}_{\theta\sim\pi}\left(\mathbb{E}_{a\sim\delta_{\mathcal{M}}(\cdot|X)}\mathbb{E}_{X\sim P_\theta} - \mathbb{E}_{a\sim\delta_{\mathcal{N}}(\cdot|Y)}\mathbb{E}_{Y\sim Q_\theta}\right)[L(\theta,a)]}_{\substack{\text{linear in } \delta_{\mathcal{M}} \\ \text{linear in } \{\pi(\theta)L(\theta,a)\}_{\theta\in\Theta,\, a\in A}}} \le \varepsilon$$

- swap the inf and sup:

$$\inf_{\delta_{\mathcal{M}}} \underbrace{\sup_{L} \sup_{\pi} \mathbb{E}_{\theta\sim\pi}\left(\mathbb{E}_{a\sim\delta_{\mathcal{M}}(\cdot|X)}\mathbb{E}_{X\sim P_\theta} - \mathbb{E}_{a\sim\delta_{\mathcal{N}}(\cdot|Y)}\mathbb{E}_{Y\sim Q_\theta}\right)[L(\theta,a)]}_{=\|\delta_{\mathcal{M}}P_\theta - \delta_{\mathcal{N}}Q_\theta\|_{\mathrm{TV}}} \le \varepsilon$$

Choose $A = \Theta$ and $\delta_{\mathcal{N}} = \text{identity}$.

Le Cam's distance

Definition (Le Cam's distance)

For two statistical models $\mathcal{M} = (\mathcal{X}, (P_\theta)_{\theta\in\Theta})$ and $\mathcal{N} = (\mathcal{Y}, (Q_\theta)_{\theta\in\Theta})$, Le Cam's distance $\Delta(\mathcal{M}, \mathcal{N})$ is defined to be the smallest $\varepsilon \geq 0$ such that

- $\mathcal{M}$ is ε-deficient to $\mathcal{N}$;
- $\mathcal{N}$ is ε-deficient to $\mathcal{M}$.

- a pseudo-metric between models (symmetric, triangle inequality)
- alternative characterization via randomization:

$$\Delta(\mathcal{M}, \mathcal{N}) = \max\left\{ \inf_{\mathsf{K}:\mathcal{X}\to\mathcal{Y}} \underbrace{\sup_{\theta\in\Theta} \|\mathsf{K}P_\theta - Q_\theta\|_{\mathsf{TV}}}_{\text{convex in K}}, \right.$$

$$\left. \inf_{\mathsf{L}:\mathcal{Y}\to\mathcal{X}} \sup_{\theta\in\Theta} \|\mathsf{L}Q_\theta - P_\theta\|_{\mathsf{TV}} \right\}$$

- a convex program, but ... infinite card. of Θ and evaluation of TV.

Asymptotic equivalence

For two sequences of statistical models $\mathcal{M}_n = (\mathcal{X}_n, (P_{n,\theta})_{\theta \in \Theta_n})$ and $\mathcal{N}_n = (\mathcal{Y}_n, (Q_{n,\theta})_{\theta \in \Theta_n})$ with $\Theta_n \uparrow \Theta$, we say they are asymptotically equivalent if and only if

$$\Delta(\mathcal{M}_n, \mathcal{N}_n) \to 0, \qquad \text{as } n \to \infty.$$

Upcoming examples of (asymptotic) equivalence:

- reduction by sufficiency
- multinomial vs. Poissonized model
- density estimation, regression, and Gaussian white noise model
- localized regular model and Gaussian location model (next lecture)

Example I: sufficiency

- statistical model $\mathcal{M} = (\mathcal{X}, (P_\theta)_{\theta\in\Theta})$, with $X \sim P_\theta$
- let $T = T(X)$ be a function of X
- T-induced model $\mathcal{N} = (\mathcal{T}, (P_\theta \circ T^{-1})_{\theta\in\Theta})$
- when do we have $\Delta(\mathcal{M}, \mathcal{N}) = 0$?

Theorem

$\Delta(\mathcal{M}, \mathcal{N}) = 0$ if and only if T is a sufficient statistic, i.e. $\theta - T - X$ forms a Markov chain.

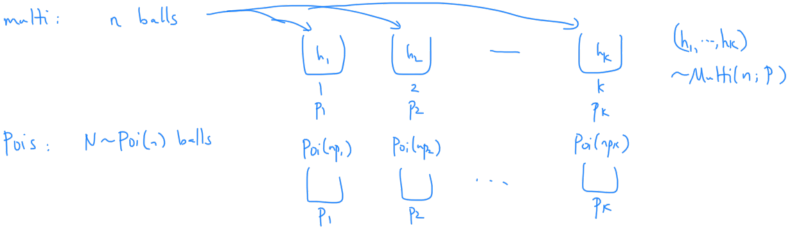
$P_{X|T}$ ind. of θ

rand. from X to T ✓

from T to X

Example II: Poissonization

- n: sample size
- k: support size
- parameter set $\Theta_k = \{P = (p_1, \cdots, p_k) \in \mathbb{R}_+^k : p_1 + \cdots + p_k = 1\}$
- multinomial model $\mathcal{M}_{n,k}$: draw n iid samples from $P \in \Theta_k$
- Poissonized model $\mathcal{P}_{n,k}$: draw $N \sim \mathrm{Poi}(n)$ iid samples from $P \in \Theta_k$



Theorem

For any fix k, we have $\lim_{n\to\infty} \Delta(\mathcal{M}_{n,k}, \mathcal{P}_{n,k}) = 0$.

Randomization from $\mathcal{M}_{n,k}$ to $\mathcal{P}_{n,k}$

- let $(X_1, \cdots, X_n)$ be iid observations from $\mathcal{M}_{n,k}$;
- draw $N \sim \text{Poi}(n)$;
- if $N \leq n$, output $(X_1, \cdots, X_N)$;
- if $N > n$, generate $m \triangleq N - n$ fake samples $(Y_1, \cdots, Y_m)$ from the empirical distribution of $(X_1, \cdots, X_n)$, then output $(X_1, \cdots, X_n, Y_1, \cdots, Y_m)$.

$$(X_1, \cdots, X_n, \underbrace{Y_1, \cdots, Y_m})$$

sample w. replacement

Analysis

$$\mathbb{E}_m\{\| P_{(X^n, Y^m)} - P_{X^{n+m}} \|_{TV}\}$$

- let P_n be the empirical distribution of $(X_1, \cdots, X_n)$
- upper bound of $\mathbb{E}[\chi^2(P_n, P)]$:

$$nP_n(i) \sim B(n, p_i)$$

$$\chi^2(P_n, P) = \sum_{i=1}^{k} \frac{(P_n(i) - p_i)^2}{p_i} \longrightarrow \mathbb{E}[\cdot] = \sum_{i=1}^{k} \frac{\frac{1}{n} p_i (1-p_i)}{p_i} = \frac{1}{n} \sum_{i=1}^{k} (1 - p_i) = \frac{k-1}{n}.$$

- upper bound of $\mathbb{E}[\|P_n^{\otimes m} - P^{\otimes m}\|_{\mathrm{TV}}]$

$$TV \le \sqrt{\tfrac{1}{2} KL} \le \sqrt{\tfrac{1}{2} \log(1+\chi^2)} \le \sqrt{\tfrac{1}{2}\chi^2}$$

$$\| P_n^{\otimes m} - P^{\otimes m} \|_{TV} \le \sqrt{\tfrac{1}{2} D_{KL}(P_n^{\otimes m} \| P^{\otimes m})} = \sqrt{\tfrac{m}{2} D_{KL}(P_n \| P)} \le \sqrt{\tfrac{m}{2} \chi^2(P_n, P)}$$

$$\mathbb{E}[\cdot] \le \sqrt{\tfrac{m}{2} \mathbb{E}[\chi^2]} \le \sqrt{\frac{m(k-1)}{2n}}.$$

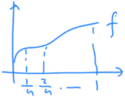
- final expectation w.r.t. $m = (N - n)_+$:

$$\mathbb{E}[\sqrt{m}] \le \mathbb{E}[m^2]^{1/4} \le n^{1/4} \Longrightarrow \mathbb{E}[TV] \le \sqrt{\frac{k-1}{2n^{1/2}}} = O\left(\frac{k^{1/2}}{n^{1/4}}\right) \xrightarrow{n \to \infty} 0$$

- HW1: show the tight bound $\Delta(\mathcal{M}_{n,k}, \mathcal{P}_{n,k}) = \Theta(\min\{1, \sqrt{k/n}\})$
  - asymptotic equivalence breaks down for large k

Example III: nonparametric estimation

nonparametric regression: $y_i \sim \mathcal{N}(f(i/n), \sigma^2), i \in [n]$



$$\Updownarrow$$

Gaussian white noise model: $dY_t = f(t)dt + \frac{\sigma}{\sqrt{n}} dB_t$

$(Y_t)_{t\in[0,1]}$

$$\Updownarrow$$

Poissonized density estimation: $N(A) \sim \text{Poi}\left(n \int_A g(t)dt\right)$

$$\Updownarrow$$

density estimation: $X_1, \cdots, X_n \sim g$

- constraint: smoothness parameter $> 1/2$
- correspondence: $f(x) = \sqrt{g(x)}, \sigma = 1/2$, density bounded away from zero

Smoothness class

Definition (Hölder ball)

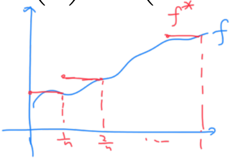
The Hölder ball $\mathcal{H}^s(L)$ with smoothness parameter $s > 0$ is the class of all functions f supported on $[0, 1]$ such that

$$\sup_{x \neq y} \frac{|f^{(m)}(x) - f^{(m)}(y)|}{|x - y|^\alpha} \leq L,$$

where $s = m + \alpha, m \in \mathbb{N}, \alpha \in (0, 1]$.

regression $\Leftrightarrow$ white noise

$$f^\star(t) = \sum_{i=1}^{n} f\left(\frac{i}{n}\right) \cdot \mathbb{1}\left(\frac{i-1}{n} < t \leq \frac{i}{n}\right)$$



$(Y_t^*)_{t\in[0,1]}$ based on f^*

$\Updownarrow$

regression model

$$D_{\mathrm{KL}}(P_{Y^\star_{[0,1]}} \| P_{Y_{[0,1]}}) = \frac{n}{2\sigma^2}\int_0^1 (f(t) - f^\star(t))^2 dt$$

$$\leq \frac{n}{2\sigma^2} \cdot L^2 n^{-2s'} \quad (s' = \min(s,1))$$

$$= \frac{L^2}{2\sigma^2} \cdot n^{1-2s'} \longrightarrow 0 \quad \text{if } s > \tfrac{1}{2}.$$

white noise ⇔ density estimation

High-level idea:

- assume a piecewise constant density with bandwidth A/n
- sufficient statistic in Poissonized density estimation:

$$Y_i \sim \mathrm{Poi}(A \cdot f(t_i)), \quad i = 1, \cdots, n/A$$

- sufficient statistic in Gaussian white noise model:

$$Z_i \sim \mathcal{N}(\sqrt{A \cdot f(t_i)}, 1/4), \quad i = 1, \cdots, n/A$$

- variance-stabilizing transformation:

delta method:

$$\mathrm{Var}(\sqrt{\mathrm{Poi}(\lambda)}) \approx \left(\frac{1}{2\sqrt{\lambda}}\right)^2 \cdot \mathrm{Var}(\mathrm{Poi}(\lambda)) = \frac{1}{4}$$

$$\sqrt{\mathrm{Poi}(\lambda)} \approx \mathcal{N}(\sqrt{\lambda}, 1/4), \quad \lambda \to \infty$$

- details far more complicated and rely on multi-resolutional analysis...

References

- Lucien M. Le Cam. "Sufficiency and approximate sufficiency." *The Annals of Mathematical Statistics* (1964): 1419 – 1455. ← deficiency paper
- Lucien M. Le Cam. "Asymptotic methods in statistical theory." Springer, New York, 1986. ← a hard book
- Lawrence D. Brown, Andrew V. Carter, Mark G. Low, and Cun-Hui Zhang. "Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift." *The Annals of Statistics* 32.5 (2004): 2074 – 2097. ← equivalence result
- Kolyan Ray and Johannes Schmidt-Hieber. "Asymptotic nonequivalence of density estimation and Gaussian white noise for small densities." *Ann. Inst. H. Poincar Probab. Statist.* 55(4), 2195 – 2208, November 2019. ← non-equivalence result.

Next lecture: classical asymptotics and Hájek–Le Cam theory