Lec 6: Statistical decision theory & classical asymptotics

Yanjun Han

Statistical decision theory.

Statistical model: a family of distributions $(P_\theta)_{\theta\in\Theta}$

(parametric: $\dim(\Theta)<\infty$; nonparametric: $\dim(\Theta)=\infty$;

semiparametric: $\Theta=\Theta_1\times\Theta_2$ with $\dim(\Theta_1)<\infty$, $\dim(\Theta_2)=\infty$)

Observation: $X\sim P_\theta$, with an unknown $\theta\in\Theta$.

Decision rule/estimator: a (possibly random) map $\hat\theta:\mathcal{X}\to\mathcal{A}$ (called "action space")

Loss: a given function $L:\Theta\times\mathcal{A}\to\mathbb{R}_+$.

Risk (expected loss): The risk of an estimator $\hat\theta$ under L is

$$r(\hat\theta;\theta)=\underbrace{\mathbb{E}_{X\sim P_\theta}}_{\text{usually abbreviated as } \mathbb{E}_\theta}[L(\theta,\hat\theta(X))].$$

Although originally proposed by Wald for statistical estimation, this framework is also general enough to encapsulate many other scenarios.

Example (Density estimation). $X_1,\cdots,X_n\overset{\text{i.i.d.}}{\sim}f$, so $\theta=f$, $P_\theta=f^{\otimes n}$.

Different losses capture different goals, such as

Density at a point: $L_1(f,a)=|a-f(0)|$

Global estimation: $L_2(f,a)=\int|f(x)-a(x)|^2\,dx$

Functional estimation: $L_3(f,a)=|a-\int h(f(x))dx|$.

Example (Linear regression). $X_1,\cdots,X_n$ either fixed or random design

$P_{Y|X}$ satisfies $\mathbb{E}[Y|X]=\langle\theta,X\rangle$

Losses include:

Estimation error: $L_1(\theta,\hat\theta)=\|\hat\theta-\theta\|^2$

Prediction error: $L_2(\theta,\hat\theta)=\mathbb{E}_{X\sim P_X}[(\langle\theta,X\rangle-\langle\hat\theta,X\rangle)^2]$.

Example (learning theory) $(X_1, Y_1), \cdots, (X_n, Y_n) \sim P_{XY}$

Loss to capture excess risk w.r.t. a given function class $\mathcal{F}$:

$$L(P_{XY}, \hat{f}) = \mathbb{E}_{P_{XY}}[(Y - \hat{f}(X))^2] - \inf_{f \in \mathcal{F}} \mathbb{E}_{P_{XY}}[(Y - f(X))^2].$$

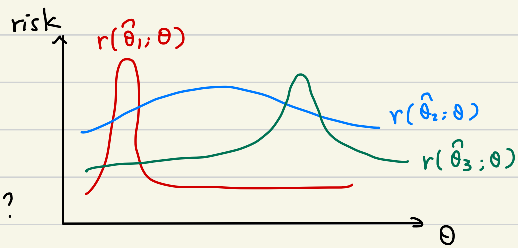
Example (optimization) Parameter: function f to be minimized

Action: a query strategy $x_{t+1} = \phi(x^t, y^t)$

Observation: queries x^T and answers y^T (e.g. $y_t = f(x_t) + \varepsilon_t$)

Loss: $L(f, x_{T+1}) = f(x_{T+1}) - \min f$.

Comparison of estimators



For an estimator $\hat{\theta}$, recall that its risk $r(\hat{\theta}; \theta)$ is a function of θ.

How to compare two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$?

① Option I: $\hat{\theta}_2$ is inferior to $\hat{\theta}_1$ if $r(\hat{\theta}_2; \theta) \geq r(\hat{\theta}_1; \theta)$ for every $\theta \in \Theta$,
and $r(\hat{\theta}_2; \theta) > r(\hat{\theta}_1; \theta)$ for some θ.

- In this case, $\hat{\theta}_2$ is called inadmissible
- However, admissibility is a weak notion: even $\hat{\theta} \equiv \theta_0$ is admissible

② Option II: given a probability distribution $\pi(\theta)$ on Θ, look at the weighted average $r_\pi(\hat{\theta}) = \int \pi(\theta) r(\hat{\theta}; \theta) d\theta$

most common

- π is called the prior
- the minimizer of $\hat{\theta} \mapsto r_\pi(\hat{\theta})$ is called the Bayes estimator under π.

③ Option III: look at the worst-case risk $r^*(\hat{\theta}) = \max_\theta r(\hat{\theta}; \theta)$

- the minimizer of $\hat{\theta} \mapsto r^*(\hat{\theta})$ is called the minimax estimator.

Define $\quad r_\pi^* = \inf_{\hat\theta} r_\pi(\hat\theta) = \inf_{\hat\theta} \mathbb{E}_{\theta\sim\pi}[r(\hat\theta;\theta)]$ (Bayes risk)

$\quad r^* = \inf_{\hat\theta} r^*(\hat\theta) = \inf_{\hat\theta}\sup_{\theta} r(\hat\theta;\theta)$ (minimax risk)

we have:

Thm. $r^* \geq r_\pi^* \quad \forall \pi$

$r^* = \sup_\pi r_\pi^*$ under regularity conditions

(minimax theorem, and the maximizer π^* is called the least favorable prior)

Pf. $\sup_\theta r(\hat\theta;\theta) \geq \mathbb{E}_{\theta\sim\pi}[r(\hat\theta;\theta)]$ (max ≥ average) $\Longrightarrow r^* \geq r_\pi^*$.

For the other direction, recall that a randomized estimator $\hat\theta$ is a probability distribution $p(\cdot|X)$ over actions, we have

$$\sup_\pi r_\pi^* = \sup_\pi \inf_p \mathbb{E}_{\theta\sim\pi}\mathbb{E}_X \mathbb{E}_{a\sim p(\cdot|X)} L(\theta,a) \quad \text{(affine in both } \pi \text{ and } p)$$

$$= \inf_p \sup_\pi \mathbb{E}_{\theta\sim\pi}\mathbb{E}_X \mathbb{E}_{a\sim p(\cdot|X)} L(\theta,a) \quad \text{(by Sion's minimax theorem)}$$

$$= \inf_p \sup_\theta \mathbb{E}_X \mathbb{E}_{a\sim p(\cdot|X)} L(\theta,a) = r^* \qquad \square$$

Finding the Bayes estimator is statistically easy: the prior $\pi(\theta)$ induces a joint distribution $\pi(\theta)p_\theta(x)$ on (θ, X), which therefore admits the posterior

$$\pi(\theta|X) \propto \pi(\theta)p_\theta(X).$$

Then the Bayes estimator is the barycenter of $\pi(\theta|X)$ under L, i.e.

$$\hat\theta_\pi(X) = \arg\min_a \mathbb{E}_{\theta\sim\pi(\cdot|X)}[L(\theta,a)].$$

However, the Bayes estimator can be computationally hard.

Finding the minimax estimator can be statistically hard, and is only feasible for a few examples (see later). Therefore, one is often interested in asymptotically minimax estimators (second part of lecture) or rate-optimal results, i.e. find $\hat\theta$ s.t. $\quad r^*(\hat\theta) \leq C r^*$ for some constant C (next few lectures)

Example (Binomial) Let $X \sim B(n, \theta)$ and $L(\theta, a) = (\theta - a)^2$.

To find the least favorable prior, try $\pi(\theta) \propto \theta^{b-1}(1-\theta)^{b-1}$ (Beta(b,b))

then posterior is $\pi(\theta | X) \propto \pi(\theta) \cdot \theta^X (1-x)^{n-X} = \theta^{b+X-1}(1-\theta)^{b+n-X-1}$ (Beta$(b+X, b+n-X)$)

and the Bayes estimator is

$$\hat{\theta}(X) = \mathbb{E}_\pi[\theta | X] = \frac{X+b}{n+2b}.$$

The risk function of $\hat{\theta}$ is

$$r(\hat{\theta}; \theta) = \mathbb{E}_\theta (\hat{\theta} - \theta)^2 = \text{Bias}^2 + \text{Var}$$

$$= \left(\frac{n\theta + b}{n+2b} - \theta \right)^2 + \frac{n\theta(1-\theta)}{(n+2b)^2}$$

$$= \frac{1}{(n+2b)^2} \left[b^2 + (n - 4b^2)\theta(1-\theta) \right].$$

By choosing $b = \frac{\sqrt{n}}{2}$, we have

$$r(\hat{\theta}; \theta) \equiv \frac{1}{4(\sqrt{n}+1)^2}.$$

Therefore, $\hat{\theta} = \frac{X + \frac{\sqrt{n}}{2}}{n + \sqrt{n}}$ attains the worst-case risk $r^*(\hat{\theta}) = \frac{1}{4(\sqrt{n}+1)^2}$, and

$$r^* \le r^*(\hat{\theta}) = r_\pi(\hat{\theta}) = r_\pi^* \le r^* \implies r^* = \frac{1}{4(\sqrt{n}+1)^2}. \qquad \boxed{\square}$$

Example (GLM). Let $X \sim N(\theta, I_n)$, $L(\theta, a) = \rho(\theta - a)$ where $\rho: \mathbb{R}^n \to \mathbb{R}_+$ is a continuous and bowl-shaped loss (i.e. $\rho(x) = \rho(-x)$ and ρ is quasi-convex).

Claim: $\hat{\theta} = X$ is the minimax estimator, with risk $r^* = \mathbb{E}[\rho(Z)]$, $Z \sim N(0, I_n)$

Pf: Try prior $\pi = N(0, \tau^2 I_n)$, then

posterior $\pi(\theta | X) \propto \exp\left(-\frac{\|\theta\|^2}{2\tau^2} - \frac{\|X - \theta\|^2}{2}\right)$ is $N\left(\frac{\tau^2 X}{1+\tau^2}, \frac{\tau^2 I_n}{1+\tau^2}\right)$.

So $$r^* \ge r_\pi^* = \mathbb{E}_X \left[\min_{a \in \mathbb{R}^n} \mathbb{E}_{\theta \sim N\left(\frac{\tau^2 X}{1+\tau^2}, \frac{\tau^2 I_d}{1+\tau^2}\right)} \rho(\theta - a) \right]$$

$$= \mathbb{E}_X \left[\rho\left(\sqrt{\frac{\tau^2}{1+\tau^2}} Z \right) \right].$$ (by Anderson's lemma below)

Let $\tau \to \infty$ gives $r^* \ge \mathbb{E}[\rho(Z)]$. $\boxed{\square}$

Lemma (Anderson) If $X \sim N(0, \Sigma)$ and ρ is bowl-shaped, then

$$\min_{a \in \mathbb{R}^n} \mathbb{E}[\rho(X+a)] = \mathbb{E}[\rho(X)].$$

Pf. Let $K_c = \{x : \rho(x) \le c\}$. Since ρ is bowl-shaped, K_c is convex, and $K_c = -K_c$.

Then
$$\begin{aligned}
\mathbb{E}[\rho(X+a)] &= \int_0^\infty \mathbb{P}(\rho(X+a) > c)\, dc \\
&= \int_0^\infty (1 - \mathbb{P}(X+a \in K_c))\, dc \\
&\ge \int_0^\infty (1 - \mathbb{P}(X \in K_c))\, dc \quad \text{(see below)} \\
&= \mathbb{E}[\rho(X)],
\end{aligned}$$

where
$$\begin{aligned}
\mathbb{P}(X \in K_c) &= \mathbb{P}\left(X \in \tfrac{1}{2}(K_c + a) + \tfrac{1}{2}(K_c - a)\right) \quad \left(\tfrac{K_c}{2} + \tfrac{K_c}{2} = K_c \text{ by convexity}\right) \\
&\ge \sqrt{\mathbb{P}(X \in K_c + a)\, \mathbb{P}(X \in K_c - a)} \quad (X \text{ has a log-concave distribution}) \\
&= \sqrt{\mathbb{P}(X \in K_c + a)\, \mathbb{P}(X \in -K_c - a)} \quad (K_c = -K_c) \\
&= \mathbb{P}(X \in K_c + a) \quad (\text{distribution of } X \text{ is symmetric around } 0) \ \boxed{\text{Q.E.D.}}
\end{aligned}$$

Hájek–Le Cam classical asymptotics: $X_1, \cdots, X_n \sim P_\theta$ with $n \to \infty$.

Regular models: differentiable in quadratic mean (QMD)

Def (QMD): A statistical model $(P_\theta)_{\theta \in \Theta}$ is called to be QMD at θ if there exists a score function $s_\theta(x)$ s.t.

$$\int \left[\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h^T s_\theta \sqrt{p_\theta}\right]^2 d\mu = o(\|h\|^2),$$

where μ is any dominating measure for (P_θ), and $p_\theta = \frac{dP_\theta}{d\mu}$.

Note: ① When $h \mapsto \sqrt{p_{\theta+h}(x)}$ is differentiable everywhere, then

$$s_\theta(x) = \frac{2}{\sqrt{p_\theta(x)}} \frac{\partial}{\partial\theta} \sqrt{p_\theta(x)} = \frac{\frac{\partial}{\partial\theta} p_\theta(x)}{p_\theta(x)} = \frac{\partial}{\partial\theta} \log p_\theta(x).$$

② Since $\int [\sqrt{p_{\theta+h}} - \sqrt{p_\theta}]^2 d\mu = H^2(P_{\theta+h}, P_\theta) \le 2$, QMD implies that the Fisher information $I(\theta) := \mathbb{E}_\theta[s_\theta s_\theta^T]$ exists.

<u>History of asymptotic theorems</u>: Fisher's program:

① The MLE $\hat{\theta}_n$ satisfies $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, I(\theta)^{-1})$, where $I(\theta)$ is the Fisher information matrix of $(P_\theta)_{\theta \in \Theta}$.

② For any other sequence of estimators $\{T_n\}$ with
$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \Sigma_\theta), \quad \forall \theta \in \Theta,$$
then $\Sigma_\theta \succeq I(\theta)^{-1}$.

(In other words, the MLE attains the asymptotically smallest variance).

While ① is true under mild regularity conditions, ② is unfortunately not true as witnessed by Hodges' estimator (1951).

<u>Hodges' estimator</u>. Let $X_1, \cdots, X_n \overset{i.i.d.}{\sim} N(\theta, 1)$, construct
$$\hat{\theta}_n = \begin{cases} \overline{X}_n & \text{if } |\overline{X}_n| \geq n^{-1/4}, \\ 0 & \text{if } |\overline{X}_n| < n^{-1/4}. \end{cases}$$

It's easy to show that $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \begin{cases} N(0,1) & \text{if } \theta \neq 0 \\ 0 & \text{if } \theta = 0 \end{cases}$

so ② in Fisher's program doesn't hold when $\theta = 0$.

Hodges' example shows that cautions need to be taken when defining the "optimality" of the MLE or inverse Fisher information. It then took statisticians ~20 years to find the right definitions, through the following angles:

1) Hodges' estimator is not "regular" (restricting the class of estimators)
2) the set of violations has Lebesgue measure 0 ("superefficiency" occurs rarely)
3) the performance of Hodges' estimator is bad when $\theta \approx n^{-1/4}$ (a large asymptotic local risk)

A collection of asymptotic theorems

Convolution Thm. Let (P_θ) be QMD. If $\sqrt{n}(T_n - \psi(\theta)) \xrightarrow{d} L_\theta$ under $P_\theta^{\otimes n}$ and $\{T_n\}$ is regular in the sense that

$$\sqrt{n}\left(T_n - \psi(\theta + \tfrac{h}{\sqrt{n}})\right) \xrightarrow{d} L_\theta \quad \text{under } P^{\otimes n}_{\theta+\frac{h}{\sqrt{n}}}, \quad \forall h \in \mathbb{R}^d.$$

Then $\exists$ a probability measure M_θ s.t.

$$L_\theta = N(0, \nabla\psi(\theta)^T I(\theta)^{-1} \nabla\psi(\theta)) * M_\theta, \qquad \forall \theta$$

where $*$ denotes the convolution $(\mu * \nu(A) = \int \mu(dx)\nu(A-x))$

(convolution makes the distribution more "noisy")

Almost everywhere convolution thm. Under all above conditions except for the regularity of $\{T_n\}$, then

$$L_\theta = N(0, \nabla\psi(\theta)^T I(\theta)^{-1} \nabla\psi(\theta)) * M_\theta \quad \text{for Lebesgue almost every } \theta.$$

Local asymptotic minimax (LAM) thm. For every continuous and bowl-shaped loss ρ, and any sequence of estimators $\{T_n\}$,

$$\lim_{c\to\infty} \liminf_{n\to\infty} \sup_{\|h\|\le c} \mathbb{E}_{\theta+\frac{h}{\sqrt{n}}}\left[\rho\left(\sqrt{n}\left(T_n - \psi(\theta+\tfrac{h}{\sqrt{n}})\right)\right)\right] \ge \mathbb{E}[\rho(Z)],$$

with $Z \sim N(0, \nabla\psi(\theta)^T I(\theta)^{-1} \nabla\psi(\theta))$.

(this is a lower bound on the minimax risk of the local family $(P_{\theta+\frac{h}{\sqrt{n}}})_{\|h\|\le c}$, under the loss $L(\theta, a) = \rho(\sqrt{n}(a - \psi(\theta)))$.)

The proofs rely on the asymptotic equivalence between models $(P_{\theta+\frac{h}{\sqrt{n}}})_{\|h\|\le c}$ and the GLM $(N(h, I(\theta)^{-1}))_{\|h\|\le c}$; see special topic of this lecture.

A special case of LAM via Bayesian Crámer-Rao

Bayesian CR in 1D (van-Trees inequality)

Let $\theta \in [a,b]$, and $\pi(\cdot)$ be a differentiable prior density on $[a,b]$ with $\pi(a) = \pi(b) = 0$, and $J(\pi) = \int_a^b \frac{\pi'(\theta)^2}{\pi(\theta)} d\theta < \infty$. Then for any $\hat{\theta}$,

$$\mathbb{E}_\pi \mathbb{E}_\theta[(\hat{\theta}-\theta)^2] \geq \frac{1}{\mathbb{E}_\pi[I(\theta)] + J(\pi)}.$$

(Compare with the usual CR $\mathbb{E}_\theta[(\hat{\theta}-\theta)^2] \geq \frac{1}{I(\theta)}$ for unbiased $\hat{\theta}$)

Pf. $\mathbb{E}_\pi \mathbb{E}_\theta\left[(\hat{\theta}-\theta)\,\partial_\theta(\log \pi(\theta)\, p_\theta(x))\right]$

$$= \int_X \int_a^b (\hat{\theta}-\theta)\,\partial_\theta(\pi(\theta) p_\theta(x))\, d\theta\, \mu(dx)$$

$$= \int_X \int_a^b \pi(\theta) p_\theta(x)\, d\theta\, \mu(dx) \qquad \text{(integration by parts)}$$

$$= 1.$$

Then BCR follows from Cauchy-Schwarz and

$$\mathbb{E}_\pi \mathbb{E}_\theta\left[\partial_\theta(\log \pi(\theta) p_\theta(x))^2\right] = \mathbb{E}_\pi\left[\left(\frac{\pi'(\theta)}{\pi(\theta)}\right)^2\right] + \mathbb{E}_\pi \mathbb{E}_\theta\left[\left(\frac{p_\theta'(x)}{p_\theta(x)}\right)^2\right]$$

$$+ \underbrace{2\,\mathbb{E}_\pi \mathbb{E}_\theta\left[\frac{\pi'(\theta)}{\pi(\theta)} \frac{\partial_\theta p_\theta(x)}{p_\theta(x)}\right]}_{=0 \text{ assuming } \int \mu(dx)\partial_\theta p_\theta(x) = \partial_\theta \int \mu(dx) p_\theta(x) = 0.}$$

$$= J(\pi) + \mathbb{E}_\pi[I(\theta)]. \qquad \boxed{\text{B}}$$

Multivariate BCR. Let $\pi = \prod_{i=1}^d \pi_i$ be a differentiable prior density on $\prod_{i=1}^d [a_i, b_i]$ vanishing on the boundary, and $J(\pi) = \mathrm{diag}(J(\pi_1), \cdots, J(\pi_d))$. Then for any $\hat{\theta}$,

$$\mathbb{E}_\pi \mathbb{E}_\theta[\|\hat{\theta}-\theta\|^2] \geq \mathrm{Tr}\left[(\mathbb{E}_\pi[I(\theta)] + J(\pi))^{-1}\right].$$

Pf. Similar to the 1-D proof, can show $\forall k = 1, \cdots, d$,

$$\mathbb{E}_\pi \mathbb{E}_\theta \left[(\hat{\theta}_k - \theta_k) \nabla_\theta \log(\pi(\theta) p_\theta(X)) \right] = e_k \quad (k\text{-th basis vector}).$$

Let $\Sigma = \mathbb{E}[\nabla_\theta \log(\pi(\theta)p_\theta(X)) \nabla_\theta \log(\pi(\theta)p_\theta(X))^T] = \mathbb{E}_\pi[I(\theta)] + J(\pi)$, by Cauchy-Schwarz we have

$$\mathbb{E}_\pi \mathbb{E}_\theta \left[(\hat{\theta}_k - \theta_k)^2 \right] \geq \sup_{u \neq 0} \frac{\langle u, e_k \rangle^2}{u^T \Sigma u} = (\Sigma^{-1})_{kk}. \qquad \boxed{\mathbb{Z}}$$

Deriving LAM from BCR when $\psi(\theta) = \theta$, $\rho(x) = \|x\|^2$.

First, note that if $\pi(\theta) = \frac{2}{b-a} \cos^2\left(\frac{\pi}{2} \cdot \frac{2\theta - (a+b)}{b-a}\right)$, then $\pi(a) = \pi(b) = 0$, and

$$J(\pi) = \int_a^b \frac{8\pi^2}{(b-a)^3} \sin^2\left(\frac{\pi}{2} \frac{2\theta - (a+b)}{b-a}\right) d\theta = \frac{4\pi^2}{(b-a)^2}.$$

(Exercise: show that this choice of π minimizes the value of $J(\pi)$.)

Next, choosing the above π_i on $\left[\theta_i - \frac{c}{\sqrt{n}}, \theta_i + \frac{c}{\sqrt{n}}\right]$, BCR gives

$$\begin{aligned}
&\inf_{\hat{\theta}} \sup_{\|h\|_\infty \leq c} \mathbb{E}_{\theta + \frac{h}{\sqrt{n}}} \left[\|\hat{\theta} - (\theta + \tfrac{h}{\sqrt{n}})\|^2 \right] \\
&\geq \inf_{\hat{\theta}} \mathbb{E}_\pi \mathbb{E}_{\theta + \frac{h}{\sqrt{n}}} \left[\|\hat{\theta} - (\theta + \tfrac{h}{\sqrt{n}})\|^2 \right] \\
&\geq \mathrm{Tr}\left[\left(n \cdot \mathbb{E}_\pi[I(\theta)] + \frac{n\pi^2}{c^2} I \right)^{-1} \right] \quad \left(\text{Fisher info. for } X_1, \cdots, X_n \sim P_\theta \text{ is } n \cdot I(\theta)\right) \\
&= \frac{1 + o(1)}{n} \mathrm{Tr}(I(\theta)^{-1}) \quad \text{as } n \to \infty \text{ and } c \to \infty,
\end{aligned}$$

assuming that $\theta \mapsto I(\theta)$ is continuous at θ.

Application of LAM: Since the global minimax risk is always lower bounded by the local minimax risk, so LAM gives asymptotic lower bounds on r_n^*.

Example (Binomial). Revisit the previous example $X \sim B(n,\theta)$. Then

$$r_n^* = \inf_{\hat\theta} \sup_{\theta\in[0,1]} \mathbb{E}_\theta[(\hat\theta-\theta)^2]$$

$$\geq \inf_{\hat\theta} \sup_{|h|\leq c_n} \mathbb{E}_{\frac12+\frac{h}{\sqrt n}}\left[\left(\hat\theta-\left(\tfrac12+\tfrac{h}{\sqrt n}\right)\right)^2\right] \quad (c_n\to\infty \text{ as } n\to\infty)$$

$$\geq \frac{1-o_n(1)}{nI(\frac12)} = \frac{1-o_n(1)}{4n}.$$

This is consistent with the exact expression of $r_n^* = \frac{1}{4(\sqrt n+1)^2}$.

Example (nonparametric entropy estimation) Let $X_1,\cdots,X_n \overset{\text{i.i.d.}}{\sim} f$, a density on $[0,1]$. The target is to estimate the differential entropy $h(f)=\int_0^1 -f(x)\log f(x)\,dx$ under the squared loss.

Challenge: This is not a finite-dimensional model, so LAM doesn't directly apply

Solution: Consider a one-parameter subfamily $(f_0+tg)_{|t|\leq\varepsilon}$, then

$$I(0)=\int_0^1 \frac{g(x)^2}{f_0(x)}dx, \qquad \frac{d}{dt}h(f_0+tg)\Big|_{t=0} = -\int_0^1 (1+\log f_0(x))\,g(x)\,dx.$$

LAM applied to this subfamily at $t=0$ gives

$$r_n^* \geq \frac{1-o_n(1)}{n}\left(\int_0^1 \frac{g(x)^2}{f_0(x)}dx\right)^{-1}\left(\int_0^1(1+\log f_0(x))\,g(x)\,dx\right)^2 =: \frac{1-o_n(1)}{n}V(f_0,g)$$

We can maximize this lower bound w.r.t. g. Since $\int g=0$ (as f_0+tg is a density), Cauchy-Schwarz gives

$$V(f_0,g) = \left(\int_0^1\frac{g(x)^2}{f_0(x)}dx\right)^{-1}\left(\int_0^1(\log f_0(x)+h(f_0))\,g(x)\,dx\right)^2$$

$$\leq \int_0^1 f_0(x)(\log f_0(x)+h(f_0))^2\,dx = \int_0^1 f_0(x)\log^2 f_0(x)\,dx - h(f_0)^2,$$

where equality holds when $g(x)=f_0(x)(\log f_0(x)+h(f_0))$.

Therefore, $$r_n^* \geq \frac{1-o_n(1)}{n} \sup_{f_0} \left(\int_0^1 f_0(x) \log^2 f_0(x)\, dx - h(f_0)^2 \right)$$

Pros and cons for asymptotic theorems:

- Pro 1: plug-and-play bound for essentially all statistical models
- Pro 2: exact constant for the asymptotic risk
- Con 1: bounds are asymptotic, assuming $n \to \infty$ while d fixed
- Con 2: bounds are for asymptotic variance, while for high-dimensional scenarios bias can be the dominating factor.

This is the reason to study techniques for non-asymptotic lower bounds in the next few lectures.

Special topic: Le Cam's distance between statistical models

(Ref: Liese and Miescke, "statistical decision theory". Springer. 2008)

For two models $(P_\theta)_{\theta\in\Theta}$ and $(Q_\theta)_{\theta\in\Theta}$ with the same parameter set Θ, how to compare the strengths between them?

(Throughout let's assume that Θ is a finite set)

Defn (deficiency) A model $\mathcal{M} = (P_\theta)_{\theta\in\Theta}$ is called ε-deficient w.r.t. $\mathcal{N} = (Q_\theta)_{\theta\in\Theta}$ if $\forall$ finite decision space $\mathcal{A}$;

$\forall$ bounded loss $L(\theta, a) \in [0,1]$;

$\forall$ (randomized) estimator $\hat\theta_{\mathcal{N}}$ under $\mathcal{N}$,

$\exists$ estimator $\hat\theta_{\mathcal{M}}$ under $\mathcal{M}$ s.t.

$$r(\hat\theta_{\mathcal{M}};\theta) \leq r(\hat\theta_{\mathcal{N}};\theta) + \varepsilon, \quad \forall \theta\in\Theta.$$

Thm (Randomization criterion) The following are equivalent:

① $\mathcal{M}$ is ε-deficient w.r.t $\mathcal{N}$;

② for every finite action set $\mathcal{A}$, bounded loss $L(\theta, a) \in [0, 1]$, and prior π on Θ, the Bayes risks satisfy $r_\pi^*(\mathcal{M}) \le r_\pi^*(\mathcal{N}) + \varepsilon$;

③ there exists a kernel K from $\mathcal{X}$ to $\mathcal{Y}$ s.t. $TV(KP_\theta, Q_\theta) \le \varepsilon$, $\forall \theta \in \Theta$.

$(KP_\theta(y) = \sum_x P_\theta(x) K(y|x))$

Pf. ① $\Rightarrow$ ②: trivial.

③ $\Rightarrow$ ①: upon observing X under $\mathcal{M}$, apply the kernel K to simulate Y and apply the estimator $\hat{\theta}_N(y)$

② $\Rightarrow$ ③: let $\mathcal{A} = \Theta$ and $\hat{\theta}_N(y) = y$:

$$\sup_{0 \le L \le 1} \sup_\pi \inf_K \mathbb{E}_{\theta \sim \pi}\left[\mathbb{E}_{X \sim P_\theta} \mathbb{E}_{a \sim K(\cdot|X)} - \mathbb{E}_{a \sim Q_\theta}\right][L(\theta, a)] \le \varepsilon.$$

This objective is linear in $K(\cdot|X)$ and $\{\pi(\theta) L(\theta, a)\}_{\theta \in \Theta, a \in \mathcal{A}}$, by minimax thm,

$$\inf_K \underbrace{\sup_{0 \le L \le 1} \sup_\pi \mathbb{E}_{\theta \sim \pi}\left[\mathbb{E}_{X \sim P_\theta} \mathbb{E}_{a \sim p(\cdot|X)} - \mathbb{E}_{a \sim Q_\theta}\right][L(\theta, a)]}_{= \max_{\theta \in \Theta} TV(KP_\theta, Q_\theta)} \le \varepsilon.$$

$\boxed{\text{☑}}$

Defn (Le Cam's distance) For finite models $\mathcal{M} = (P_\theta)_{\theta \in \Theta}$ and $\mathcal{N} = (Q_\theta)_{\theta \in \Theta}$, define Le Cam's distance as

$$\Delta(\mathcal{M}, \mathcal{N}) = \min\{\varepsilon : \mathcal{M} \text{ is } \varepsilon\text{-deficient to } \mathcal{N},\ \mathcal{N} \text{ is } \varepsilon\text{-deficient to } \mathcal{M}\}.$$

Example (sufficiency) For $\mathcal{M} = (P_\theta)_{\theta \in \Theta}$ and a function $T = T(X)$, define the T-induced model $\mathcal{N} = (T \# P_\theta)_{\theta \in \Theta}$. By randomization criterion,

$\Delta(\mathcal{M}, \mathcal{N}) = 0 \iff \mathcal{M}$ and $\mathcal{N}$ are mutual randomizations

$\iff$ both $\theta - X - T$ and $\theta - T - X$ are Markov chains

$\iff T$ is a sufficient statistic for X.

(Factorization Thm: T is sufficient $\iff P_\theta(x) = g(x) h(\theta, T)$ for some g, h)

For a sequence of models $(\mathcal{M}_n)_{n\geq 1}$ and $(\mathcal{N}_n)_{n\geq 1}$, how to show asymptotic equivalence

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

Defn (standard model) Let $\mathcal{M} = \{P_1, \cdots, P_m\}$ be a finite model, and $\overline{P} := \frac{1}{m}\sum_{i=1}^m P_i$.

Then $T(X) = \left(\frac{P_1}{\overline{P}}(X), \cdots, \frac{P_m}{\overline{P}}(X)\right)$ is sufficient and lies on $\Delta_m := \{u\in\mathbb{R}_+^m : 1^T u = m\}$.

(applying factorization thm to $p_i(x) = \overline{p}(x) T_i(x)$)

So $\mathcal{M}$ is equivalent to the T-induced model $\mathcal{N} = \{\mu_1, \cdots, \mu_m\}$ with $\frac{\mu_i(T)}{\mu(T)} = T_i$,

where μ is the distribution of T under $\overline{P}$, known as the standard distribution.

($\mathbb{E}_{\mu_i}[f(T)] = \mathbb{E}_{P_i}[f(T(X))] = \mathbb{E}_{\overline{P}}[\frac{P_i}{\overline{P}} f(T(X))] = \mathbb{E}_{\mu}[T_i f(T)]$)

Implication: standard model unifies all statistical models of size m to standard distributions μ on Δ_m.

Thm. If $\mu_n \xrightarrow{d} \mu$, then $\Delta(\mathcal{M}_n, \mathcal{M}) \to 0$.

Pf. By ② in the randomization criterion, suffices to check

$$\sup_{A,\pi,L} \left| r_\pi^*(\mathcal{M}_n) - r_\pi^*(\mathcal{M}) \right| \xrightarrow{n\to\infty} 0.$$

In a standard model,
$$\begin{aligned} r_\pi^*(\mathcal{M}) &= \inf_{\hat\theta} \sum_{i=1}^m \pi_i \mathbb{E}_{\mu_i}[L(i, \hat\theta(T))] \\ &= \inf_{\hat\theta} \mathbb{E}_\mu\left[\sum_{i=1}^m \pi_i T_i L(i, \hat\theta(T))\right] \\ &= \mathbb{E}_\mu\left[\inf_{c\in C} \langle c, T\rangle\right], \quad C := \mathrm{conv}\left(\{(\pi_i L(i,a))_{i=1}^m\}_{a\in A}\right). \end{aligned}$$

Since $f(T) = \inf_{c\in C}\langle c, T\rangle$ is bounded by m and 1-Lip under $\|\cdot\|_1$,

$$\sup_{A,\pi,L} \left| r_\pi^*(\mathcal{M}_n) - r_\pi^*(\mathcal{M}) \right| \leq \sup_{\substack{\|f\|_\infty \leq m \\ |f(x)-f(y)|\leq \|x-y\|_1}} \left| \mathbb{E}_\mu f - \mathbb{E}_{\mu_n} f \right| \to 0.$$

(Dudley's metric metrizes weak convergence) $\square$

Now we're ready to present the main result.

Thm. Let $\mathcal{M}_n = \{P_{1,n}, \cdots, P_{m,n}\}$, $n \geq 1$, and $\mathcal{M} = \{P_1, \cdots, P_m\}$.

Let $L_n = \left(\frac{P_{2,n}}{P_{1,n}}, \cdots, \frac{P_{m,n}}{P_{1,n}}\right)$ and $L = \left(\frac{P_2}{P_1}, \cdots, \frac{P_m}{P_1}\right)$.

Suppose $\mathcal{M}$ is "homogeneous", i.e. P_i and P_j are mutually absolutely continuous.

Then $\quad \mathrm{Law}(L_n \mid P_{1,n}) \xrightarrow{d} \mathrm{Law}(L \mid P_1) \Rightarrow \Delta(\mathcal{M}_n, \mathcal{M}) \to 0$.

(In other words, weak convergence of likelihood ratios implies asymptotic equivalence)

Pf. Suffice to show that the standard distributions $\mu_n \xrightarrow{d} \mu$.

Also, note that $\mathrm{Law}(L_n \mid P_{1,n})$ is unchanged when moving to the standard model.

By compactness of $\Delta_m = \{u \in \mathbb{R}_+^m : 1^T u = m\}$ and Prokhorov's Thm, it suffices to show that if $\mu_{n_k} \xrightarrow{d} \nu$ along some subsequence, then $\nu = \mu$.

For $s = (s_2, \cdots, s_m)$ with $s_i > 0$ and $\sum_{i=2}^m s_i < 1$, then $f_s(L) = \prod_{i=2}^m L_i^{s_i}$ is a continuous function of L. In addition, for $s_1 = 1 - \sum_{i=2}^m s_i \in (0,1)$,

$$\mathbb{E}_{P_1}\left[f_s(L)^{\frac{1}{1-s_1}}\right] = \mathbb{E}_{P_1}\left[L_2^{\frac{s_2}{1-s_1}} \cdots L_m^{\frac{s_m}{1-s_m}}\right] \overset{\text{Hölder}}{\leq} \prod_{i=2}^m \mathbb{E}_{P_1}[L_i]^{\frac{s_i}{1-s_1}} \leq 1,$$

so the sequence of RVs $f_s(L_n)$ is uniformly integrable. Therefore, by weak convergence,

$$\mathbb{E}_\mu[T_1^{s_1} T_2^{s_2} \cdots T_n^{s_n}] = \mathbb{E}_{P_1}[f_s(L)] = \lim_{n\to\infty} \mathbb{E}_{P_{1,n}}[f_s(L_n)].$$

On the other hand, as $\mu_{n_k} \xrightarrow{d} \nu$, $\mathbb{E}_{P_{1,n_k}}[f_s(L_{n_k})] = \mathbb{E}_{\mu_{n_k}}[T_1^{s_1} \cdots T_n^{s_n}] \longrightarrow \mathbb{E}_\nu[T_1^{s_1} \cdots T_n^{s_n}]$, so

$$\mathbb{E}_\mu[T_1^{s_1} \cdots T_n^{s_n}] = \mathbb{E}_\nu[T_1^{s_1} \cdots T_n^{s_n}], \quad \forall s_1, \cdots, s_m > 0, \ \sum_{i=1}^m s_i = 1.$$

By uniqueness results for MGFs, this implies that $\tilde{\mu} = \tilde{\nu}$, where $\tilde{\mu}$ represents the restriction of μ to $\Delta_m^\circ = \{x \in \mathbb{R}^m : x_i > 0, \ 1^T x = m\}$, i.e. $\tilde{\mu}(A) = \mu(A \cap \Delta_m^\circ)$.

Since $\mathcal{M}$ is homogeneous, we have $\tilde{\mu} = \mu$, and $\tilde{\nu}(\Delta_m^\circ) = \tilde{\mu}(\Delta_m^\circ) = \mu(\Delta_m^\circ) = 1$.

Since ν is a probability measure, $\tilde{\nu} = \nu$. Therefore, $\mu = \nu$. $\blacksquare$

Finally, we show that if $(P_\theta)_{\theta\in\Theta}$ is QMD, then for any finite set I,

$$\mathcal{M}_n = \left\{ P^{\otimes n}_{\theta_0+\frac{h}{\sqrt{n}}} \right\}_{h\in I} \text{ is asymptotic equivalent to } \mathcal{M} = \left\{ N(h, I(\theta_0)^{-1}) \right\}_{h\in I}.$$

This is called <u>local asymptotic normality</u>.

Pf. Check the likelihood ratio. In the limiting Gaussian model,

$$\log \frac{N(h, I(\theta_0)^{-1})}{N(0, I(\theta_0)^{-1})}(z) = h^T I(\theta_0) z - \frac{1}{2} h^T I(\theta_0) h, \quad \text{with } I(\theta_0) z \sim N(0, I(\theta_0))$$

For the product model, let $W_{ni} = 2\left(\sqrt{\frac{p_{\theta_0+h/\sqrt{n}}}{p_{\theta_0}}}(x_i) - 1\right)$, then

$$\log \frac{p^{\otimes n}_{\theta_0+\frac{h}{\sqrt{n}}}}{p^{\otimes n}_{\theta_0}}(x^n) = 2\sum_{i=1}^n \log\left(1 + \frac{1}{2} W_{ni}\right) = \sum_{i=1}^n W_{ni} - \frac{1}{4}\sum_{i=1}^n W_{ni}^2 + \sum_{i=1}^n o(W_{ni}^2).$$

By QMD, $\mathbb{E}_{P_{\theta_0}}\left[\left(W_{ni} - \frac{1}{\sqrt{n}} h^T s_{\theta_0}(X_i)\right)^2\right] = o\left(\frac{1}{n}\right)$, thus

$$\mathrm{Var}_{P^{\otimes n}_{\theta_0}}\left(\sum_{i=1}^n W_{ni} - \frac{1}{\sqrt{n}}\sum_{i=1}^n h^T s_{\theta_0}(X_i)\right) \le n \cdot o\left(\frac{1}{n}\right) = o(1).$$

and $$\mathbb{E}\sum_{i=1}^n W_{ni} = -n\int\left(\sqrt{p_{\theta_0+\frac{h}{\sqrt{n}}}} - \sqrt{p_{\theta_0}}\right)^2 d\mu \to -\frac{1}{4} h^T \mathbb{E}[s_{\theta_0} s_{\theta_0}^T] h = -\frac{1}{4} h^T I(\theta_0) h.$$

Moreover,

$$\sum_{i=1}^n W_{ni}^2 = \sum_{i=1}^n \left(\frac{1}{\sqrt{n}} h^T s_{\theta_0}(X_i)\right)^2 + o_p(1) = \frac{1}{n}\sum_{i=1}^n h^T s_{\theta_0}(X_i) s_{\theta_0}(X_i)^T h + o_p(1)$$

$$\xrightarrow{P} h^T I(\theta_0) h \quad \text{by LLN.}$$

Therefore,

$$\log \frac{P^{\otimes n}_{\theta_0+\frac{h}{\sqrt{n}}}}{P^{\otimes n}_{\theta_0}}(x^n) = h^T \underbrace{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n s_\theta(X_i)\right)}_{\xrightarrow{d} N(0, I(\theta_0)) \text{ by CLT.}} - \frac{1}{2} h^T I(\theta_0) h + o_p(1) \qquad \square$$

Combining Anderson's lemma and the limiting Gaussian model above, and extending the previous definitions to general models by taking the supremum over all finite submodels, we arrive at the local asymptotic minimax theorem.