Minimax method for evaluation

"Have you try your best?" This kind of question is often philosophical and mathematical, which would constantly bother you if you wonder the posibility to improve again.

In statistics, for a given problem, we usually need to impoes a suitable model frownwork and solve it by any well-defined statistics of observed data, which leads to the same problem:

"Are you giving the best statistics?"

To evaluate this, it's crucial to know what is "the best".

Example 1

Problem: the mean of population μ

Model: $\mu \in \mathbb{R}$

Observed Data: $X_1, \cdots, X_n \overset{iid}{\sim} (V(\mu, \sigma^2)$

Statistics: $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$

"Is $\bar{X}_n$ the best?"

By Lehmann-Scheffe Theorem, we know for $\forall$ unbias $\hat{\mu}$, i.e. $\mathbb{E}\hat{\mu}=\mu$, we have

$$\mathrm{Var}_\mu \bar{X}_n \le \mathrm{Var}_\mu \hat{\mu},$$

which means that $\bar{X}_n$ is optimal in some sense.

What if we know $\mu \le 0$, indeed, $\min\{0, \bar{X}_n\}$ is always better than $\bar{X}_n$ since it utilizes the prior information of μ, regardless of the fact that

$$\mathbb{E}\min\{0, \bar{X}_n\} = \mathbb{E}\,\bar{X}_n \mathbb{1}(\bar{X}_n \le 0) < \mathbb{E}\,\bar{X}_n,$$

which means that the unbiasness may not be appropriate to be used in this situation.

Note that

$$\mathbb{E}\,|\min\{0, \bar{X}_n\} - \mu| \quad (\mu \le 0)$$
$$= \mathbb{E}\,|\min\{-\mu, \bar{X}_n - \mu\}| \le \mathbb{E}\,|\bar{X}_n - \mu|,$$

which implies that $\mathbb{E}\,|\cdot - \mu|$ may be intuitional correct for evaluation of $\hat{\mu}$.

Decision Theory

Decision space: $\mathcal{H}$, $\theta_0 \in \mathcal{H}$

Observed data: $X_1, \cdots, X_n \sim (\mathcal{X}^n, P_\theta)$

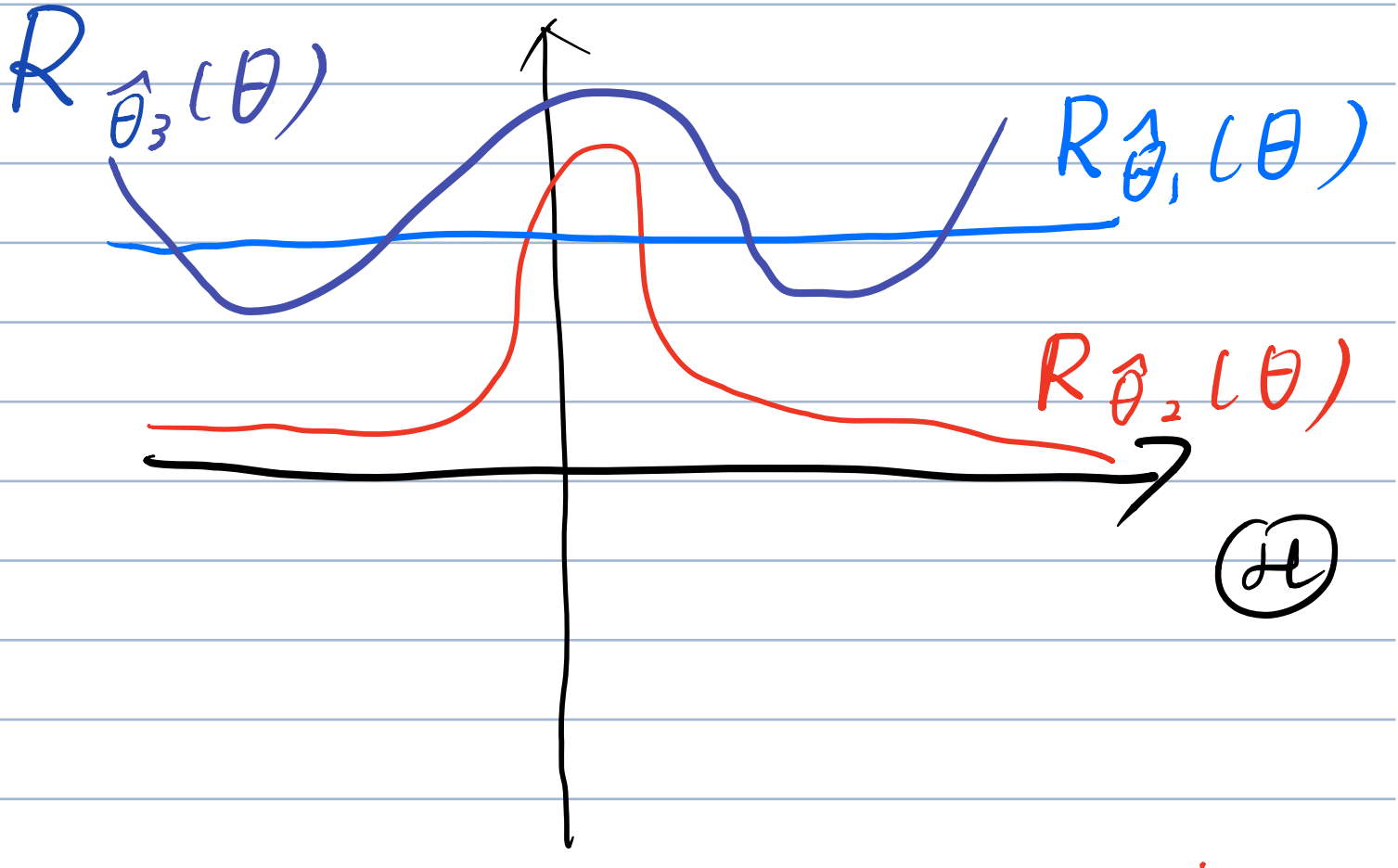
Target: θ_0 (density estimation)

Decision: $\hat{\theta} : \mathcal{X}^n \to \mathcal{H}$

Loss function: $L : \mathcal{H} \times \mathcal{H} \to \mathbb{R}^+$

Risk function:

$$R_{\hat{\theta}}(\theta_0) = \mathbb{E}_{\theta_0} L(\hat{\theta}, \theta_0)$$



"Is $\hat{\theta}_1$ better than $\hat{\theta}_2$?"

"Is $\hat{\theta}_3$ admissible?"

① Average case: If we have some measures of Θ: π,

$$\int R_{\hat{\theta}}(\theta)\,\pi(d\theta)$$

could be used in comparison, where π is called the prior of Θ.

For the blessing of prior, the optimal $\hat{\theta}_0$:

$$\operatorname{argmin}_{\hat{\theta}} \int R_{\hat{\theta}}(\theta)\,\pi(d\theta)$$

exists under some mild conditions, which is the functionals of posterior distribution: $\pi(\theta \mid X_{1:n})$. For this reason, optimal decision is often called Bayes procedure.

② The worst case:

$$\sup_{\theta \in \Theta} R_{\hat{\theta}}(\theta) \quad (1)$$

we call the $\hat{\theta}_0$ minimizes (1) the minimax optimal decision. (admissible?)

Minimal Optimality

$$\mathcal{M}(\mathcal{H}) := \inf_{\hat{\theta}} \sup_{\theta \in \mathcal{H}} R_{\hat{\theta}}(\theta)$$

How to know $\hat{\theta}_0$ is minimax optimal?

① Upper bound of $R_{\hat{\theta}_0}(\theta)$:

$$U_{\hat{\theta}_0}(\mathcal{H})$$

② Lower bound of $\mathcal{M}(\mathcal{H})$:

$$L(\mathcal{H})$$

If $U_{\hat{\theta}_0}(\mathcal{H}) = L(\mathcal{H})$, then

$$\sup_{\theta \in \mathcal{H}} R_{\hat{\theta}_0}(\mathcal{H}) \leq U_{\hat{\theta}_0}(\mathcal{H})$$

$$= L(\mathcal{H}) \leq \mathcal{M}(\mathcal{H})$$

$\Rightarrow \hat{\theta}_0$ is minimax optimal.

Somehow, "$U_{\hat{\theta}_0}(\mathcal{H}) = L(\mathcal{H})$" is very strong, we can ease it:

$$U_{\hat{\theta}_0}(\mathcal{H}) \simeq L(\mathcal{H}) \quad (2)$$

Remark: To achieve ②, $U_{\hat{\theta}_0}(\Theta)$ should be small enough, while $H(\Theta)$ should be large enough.

Example 2

In example 1, it implies that $M(\Theta) = \sigma^2/n$, if we constraint the decision $\hat{\mu}$ for μ should be unbias, where

$$\frac{\sigma^2}{n} = \sup_{\mu \in \Theta} \mathbb{E}(\bar{X}_n - \theta)^2 .$$

General Reduction Scheme

① Let $L = \Phi \circ d$, where $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$ is an increasing function and d is a semi-metric of Θ, for a decision $\hat{\theta}_0$ of θ, if

$$\sup_{\theta \in \Theta} R_{\hat{\theta}_0}(\theta) = \sup_{\theta \in \Theta} \mathbb{E}\Phi(d(\hat{\theta}_0, \theta)) \leq \Phi(\delta),$$

simutaneously, note that

$$R_{\hat{\theta}}(\theta) = \mathbb{E}_\theta \Phi(d(\hat{\theta}, \theta))$$

$$\geq \Phi(s)\, P_\theta(d(\hat{\theta}, \theta) > s)$$

if

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_\theta(d(\theta, \hat{\theta}) > s) > 0$$

$$\Rightarrow \mathcal{M}(\Theta) \gtrsim \Phi(s)$$

$$\Rightarrow \mathcal{M}(\Theta) \asymp \Phi(s)$$ and $\hat{\theta}_0$ is minimax optimal.

So what we need to do is to give a sharpen upper bound: $\Phi(s)$, and then prove that

$$\textcircled{3}$$

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_\theta(d(\theta, \hat{\theta}) > s) > 0$$

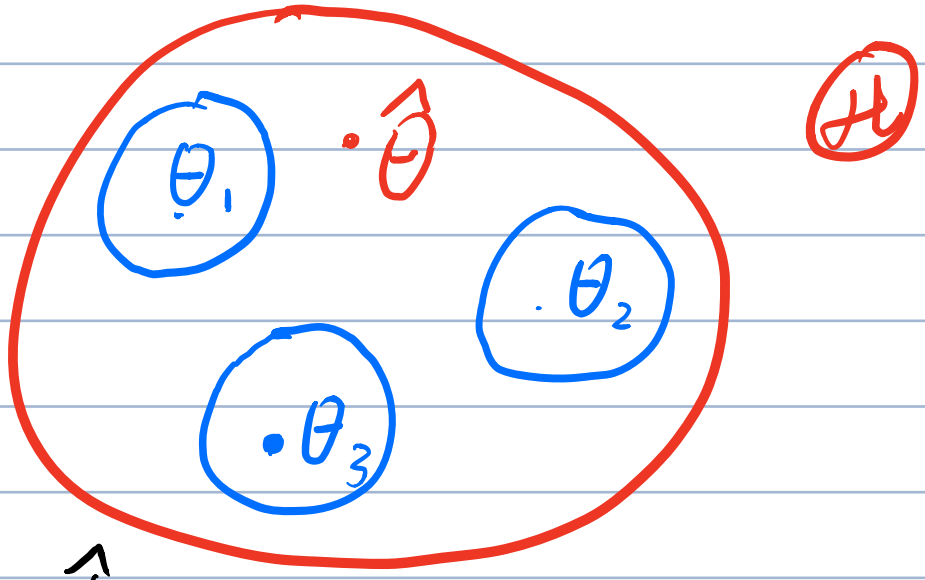
Remark:

$$\textcircled{3} \leq \frac{\mathcal{M}(\Theta)}{\Phi(s)} \leq \frac{\sup_{\theta \in \Theta} \mathbb{E}\, \Phi(d(\hat{\theta}_0, \theta))}{\Phi(s)} \leq C.$$

(2) From estimation to testing

For given $\delta > 0$, to prove (2), take 2δ-packing centers $\{\theta_i\}_{i=1:M}$ of $\mathcal{H}$ then: (If they exists!)

$$\{d(\hat{\theta}, \theta_i) < \delta\} \subset \bigcap_{j \neq i} \{d(\hat{\theta}, \theta_j) \geq \delta\}$$



define $\hat{\phi} \in \{1, \dots, M\}$:

$$\hat{\phi} := \phi(\hat{\theta}) = \operatorname{argmin}_j d(\hat{\theta}, \theta_j)$$

$$\Rightarrow \{d(\hat{\theta}, \theta_i) < \delta\} \subset \{\hat{\phi} = i\}$$

$$\Rightarrow$$

$$P_{\theta_i}(d(\hat{\theta}, \theta_i) \geq \delta) \geq P_{\theta_i}(\hat{\phi} \neq i)$$

$\Rightarrow$

$$\sup_{\theta\in\Theta} P_\theta(d(\hat{\theta},\theta)>\delta)$$
$$\geq \sup_i P_{\theta_i}(d(\hat{\theta},\theta_i)>\delta)$$
$$\geq \sup_i P_{\theta_i}(\hat{\phi}\neq i)$$
$$\geq \frac{1}{M}\sum_{i=1}^{M} P_{\theta_i}(\hat{\phi}\neq i)$$

$$\Rightarrow \inf_{\hat{\theta}}\sup_{\theta\in\Theta} P_\theta(d(\hat{\theta},\theta)>\delta)$$
$$\geq \inf_{\hat{\phi}} \frac{1}{M}\sum_{i=1}^{M} P_{\theta_i}(\hat{\phi}\neq i),$$

where $\hat{\phi}$ is any statistics valued in $\{1,\cdots,M\}$. (fixed given data)

Then given $\delta>0$, we just know to prove that

$$\exists \{\theta_i\}_{i=1:M},\ d(\theta_i,\theta_j)\geq 2\delta$$

s.t.

$$\inf_{\hat{\phi}} \frac{1}{M}\sum_{i=1}^{M} P_{\theta_i}(\hat{\phi}\neq i) > 0.$$

Le Cam's method

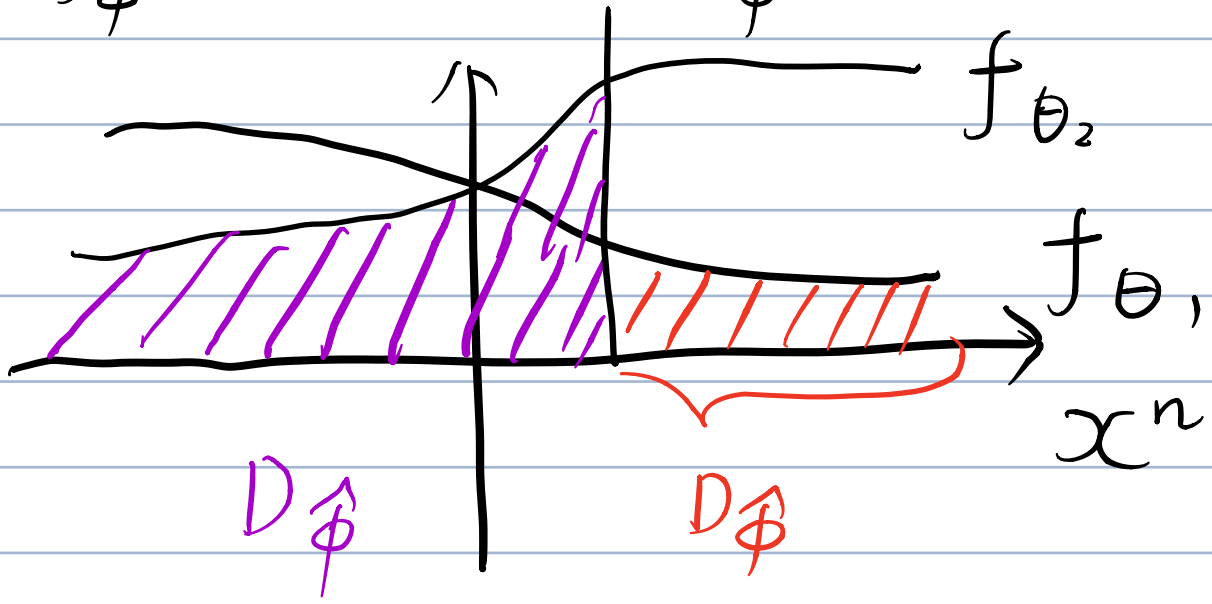
If $M=2$, we need to prove

$$\inf_{\hat{\phi}} \frac{1}{2}\left(P_{\theta_1}(\hat{\phi} \neq 1) + P_{\theta_2}(\hat{\phi} \neq 2) \right) \geq 0$$

We assume that $P_{\theta_i} \ll \nu$, and $D_{\hat{\phi}} = \{x^n \in \mathcal{X}^n : \hat{\phi}(x^n) \neq 1\}$, then

$$P_{\theta_1}(D_{\hat{\phi}}) + P_{\theta_2}(D_{\hat{\phi}}^c)$$

$$= \int_{D_{\hat{\phi}}} f_{\theta_1} d\nu + \int_{D_{\hat{\phi}}^c} f_{\theta_2} d\nu$$



$$\geq \int \min\{f_{\theta_1}, f_{\theta_2}\} d\nu,$$

"$=$" achieve iff $D_{\hat{\phi}} = \{f_{\theta_1} \leq f_{\theta_2}\}$.

$$\Rightarrow \inf_{\phi} P_{\theta_1}(D_\phi) + P_{\theta_2}(D_\phi^c)$$

$$\geq \int \min\{f_{\theta_1}, f_{\theta_2}\}$$

$$= 1 - \frac{1}{2}\int |f_{\theta_1} - f_{\theta_2}|$$

$$= 1 - \|P_{\theta_1} - P_{\theta_2}\|_{TV}, \text{ where}$$

$$\|P_{\theta_1} - P_{\theta_2}\| = \sup_A |P_{\theta_1}(A) - P_{\theta_2}(A)|.$$

Lemma 1 (Neyman-Pearson)

$$\inf_{\phi} P_{\theta_1}(\hat{\phi} \neq 1) + P_{\theta_2}(\hat{\phi} \neq 2)$$
$$\geq 1 - \|P_{\theta_1} - P_{\theta_2}\|_{TV},$$

where "=" achieve iff likelihood ratio

$$\hat{\phi} = \begin{cases} 1, & \{f_{\theta_1} \geq f_{\theta_2}\} \\ 2, & \text{else} \end{cases}.$$

Remark: Indeed, $d(\theta_1, \theta_2) \geq 2\delta$ and $1 - \|P_{\theta_1} - P_{\theta_2}\|_{TV} > 0$ should be achieved both, which is a trade-off for suitable $\theta_1, \theta_2 \in \Theta$).

Hellinger distance

Let P, Q be two probability measures of χ^n ($P, Q \ll v$),

$$H(P,Q) := \left(\int (\sqrt{p} - \sqrt{q})^2\right)^{\frac{1}{2}}$$

$$= \left(2 - 2\int \sqrt{pq}\right)^{\frac{1}{2}}$$

Lemma 2

$$\frac{(H(P,Q))^2}{2} \le \|P - Q\|_{TV} \le H(P,Q)$$

Pf: $\frac{(H(P,Q))^2}{2} = 1 - \int \sqrt{pq}$

$$\le 1 - \int \min\{p, q\} = \|P - Q\|_{TV}$$

$$H^2(P,Q) = 2 - 2\int \sqrt{pq}$$

$$= 2 - \int \sqrt{\max\{p,q\}\,\min\{p,q\}}$$

$$\ge 2 - \left(\int \max\{p,q\}\right)^{\frac{1}{2}} \left(\int \min\{p,q\}\right)^{\frac{1}{2}}$$

$$= 2 - 2\left(1 + \|P-Q\|_{TV}\right)^{\frac{1}{2}} \left(1 - \|P-Q\|_{TV}\right)^{\frac{1}{2}}$$

$$= 2 - 2\left(1 - \|P - Q\|_{TV}^2\right)^{\frac{1}{2}}$$

$$\geq \|P - Q\|_{TV} \quad \text{since}$$

$$2 - 2(1-x)^{\frac{1}{2}} \geq x, \quad 0 \leq x \leq 1$$

□

Kullback-Leibler divergence

$$KL(P\|Q) = \int p \log \frac{p}{q}$$

Lemma 2

$$KL(P\|Q) \geq (H(P,Q))^2$$

Pf:

$$KL(P\|Q) = \int p \log \frac{p}{q}$$

$$= -2 \int p \log\left(\sqrt{\frac{q}{p}} - 1 + 1\right)$$

$$\geq -2 \int p \left(\sqrt{\frac{q}{p}} - 1\right)$$

$$= 2 - 2\int \sqrt{pq}.$$

□

Decoupling Property

Now we mark $P^n = P \otimes \cdots \otimes P$ and P is a probability measure of X.

Pro $H^2(P^n, Q^n) \le n H^2(P, Q)$

$KL(P^n \| Q^n) = n KL(P \| Q)$

Pf:

$$\frac{1}{2} H^2(P^n, Q^n) = 1 - \int \sqrt{p^n q^n}$$

$$= 1 - \left(\int \sqrt{pq}\right)^2 = 1 - \left(1 - \frac{1}{2} H^2(P, Q)\right)^n$$

$$\le \frac{1}{2} n H^2(P, Q) \quad (1 - (1-x)^n \le nx)$$

$$KL(P^n \| Q^n) = \int p^n \log \frac{p^n}{q^n}$$

$$= \int \sum_{i=1}^{n} \int p(x_1, \cdots, x_n) \log \frac{p(x_i)}{q(x_i)} \, dx_i$$

$$= n \; KL(P \| Q)$$

□

Example 3

In example 1, we know

$$KL(P_{\mu_1}, P_{\mu_2}) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2},$$

then

$$1 - \|P_{\mu_1} - P_{\mu_2}\|_{TV} \geq 1 - \frac{n(\mu_1 - \mu_2)^2}{2\sigma^2},$$

then we should choose δ s.t.

$$1 - \frac{(\mu_1 - \mu_2)}{2\sigma^2/n} > 0 \quad \text{and}$$

$$|\mu_1 - \mu_2| \geq 2\delta,$$

$$\Rightarrow \quad \delta \lesssim \sqrt{\frac{\sigma^2}{n}}$$

$$\Rightarrow \quad \mathcal{M}(\Theta) \geq \Phi\left(\sqrt{\frac{\sigma^2}{n}}\right)$$

If $\Phi(t) = t^2$, then

$$\sup_{\mu} \mathbb{E}(\bar{X}_n - \mu)^2 = \frac{\sigma^2}{n}$$

$$\Rightarrow \quad \bar{X}_n \text{ is minimal optimal.}$$

Example 4

$X_1, \dots, X_n \overset{iid}{\sim} U[\theta, \theta+1]$, we use $\min\{X_i\}$ to estimate θ:

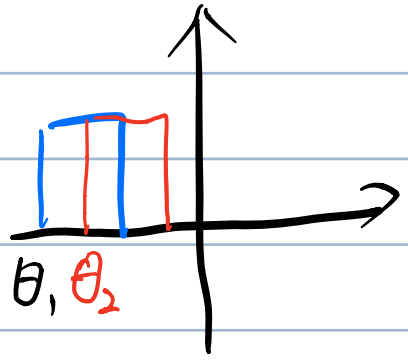
$$\mathbb{E}\,|\min\{X_i\} - \theta|$$
$$= \mathbb{E}\,\min\{Y_i\},\ Y_i \overset{iid}{\sim} U[0,1]$$
$$= \int \mathbb{P}(\min\{Y_i\} > t)\,dt$$
$$= \frac{1}{n+1}$$

And $\mathcal{H}^2(U[\theta_1, \theta_1+1], U[\theta_2, \theta_2+1])$

$$= \begin{cases} 2|\theta_1 - \theta_2|, & \text{if } |\theta_1 - \theta_2| \le 1 \\ 2, & \text{else} \end{cases}$$
$$= 2\min\{|\theta_1 - \theta_2|, 1\}$$

$$\Rightarrow \mathcal{H}^2(P_{\theta_1}, P_{\theta_2}) \le n \min\{|\theta_1 - \theta_2|, 1\}$$

$$\Rightarrow \delta = \frac{1}{2n}$$



$$\Rightarrow M(\theta) \gtrsim \mathbb{E}\left(\frac{1}{2n}\right)$$

$\Rightarrow \min\{X_i\}$ is minimax optimal.

Le Cam for functional

Compared with the traditional way to evaluate a point estimator, such as, UMVUE (Uniformly minimal variance unbias estimator) or Cramér-Rao Lower bound, minimax method is less powerful and even boring for the parametric model. But the situation gets little different when we discuss the point estimator for non-parametric model, since the likelihood of data has an "ambiguous" form of parameter (principle of sufficiency and likelihood is hard to be used) and the procedure of model selection makes the evaluation getting hard, while the minimax aspect mostly focuses on the geometry structure of the parameter space, which is applicable for the evaluation of wide range of point estimator.

Now we let $\mathcal{F} := \mathbb{H}$, which are some density function spaces.

Similarly, we can construct some non-parametric statistics for $f \in \mathcal{F}$: $\hat{f}_n$ and evaluate it by

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f \Phi(d(\hat{f}_n, f)),$$

not that far, we firstly focus on some functional of $\mathcal{F}$ (mean, median, or evaluation functional): θ, and given a statistic for $\theta(f)$: $\hat{\theta}$, define

$$\inf_{\hat{\theta}} \sup_{f \in \mathcal{F}} \mathbb{E}_f \Phi(d(\hat{\theta}, \theta(f)))$$
$$:= \mathcal{M}(\theta(\mathcal{F})).$$

The part gets herder

By Le Cam, if $\exists f, g \in \mathcal{F}$, s.t.

$$d(\theta(f), \theta(g)) \geq 2\delta,$$

for given $\delta > 0$, then

$$\mathcal{M}(\theta(\mathcal{F})) \geq \Phi(\delta)$$

if

$$1 - \| P_f^n - P_g^n \|_{TV} > 0.$$

Note that

$$\|P_f^n - P_g^n\|^2 \le \mathcal{H}^2(P_f^n, P_g^n) \le n\mathcal{H}^2(P_f, P_g),$$

If we know that

$$d(\theta(f), \theta(g)) \ge C\mathcal{H}(P_f, P_g) \quad (4)$$

Then if $\exists f, g \in \mathcal{F}$, s.t.

$$\mathcal{H}^2(P_f, P_g) = \frac{1}{4n}$$

$$\Rightarrow M(\theta(\mathcal{F})) \gtrsim \Phi\left(\frac{C}{4\sqrt{n}}\right)$$

If (4) is hard to achieve, we can define

$$w_\theta(\varepsilon) = \sup_{\substack{f,g\in\mathcal{F} \\ \mathcal{H}(f,g)\le\varepsilon}} \|\theta(f) - \theta(f)\|$$

$$\Rightarrow M(\theta(\mathcal{F})) \ge \Phi\left(\frac{1}{2} w\left(\frac{1}{2\sqrt{n}}\right)\right).$$

Example 5 (Minimax rate of pointwise estimation of Lipschitz density)

$$\mathcal{F}_L = \{f \in C[0,1],\ |f(x)-f(y)| \le L|x-y|\},$$
$$\int f(x)\,dx = 1 \quad (\text{with } \|\cdot\|_{\sup})$$

There is a famous estimator for this kind of density: $\hat{f}_{n,\lambda}$, kernel density estimation, and

$$\sup_{f \in \mathcal{F}_L} \mathbb{E}\left(\hat{f}_{n,\lambda}(x_0) - f(x_0)\right)^2 \le n^{-\frac{2}{3}}, \quad x_0 \in [0,1],$$

if we choose a suitable window width λ ($\lambda = \lambda_n$, the optimal λ depends on n).

Define $\theta(f) = f(x_0)$, $f \in \mathcal{F}_L$.

Now we let $L = 1$, choose $f \in \mathcal{F}_1$: $f \equiv 1$, and $\phi \in \mathcal{F}_1$, $\int \phi = 0$, then

$$g := f + \phi \in \mathcal{F}_1,$$

$$H^2(f,g) = 2 - 2\int \sqrt{1+\phi(x)}\,dx.$$

Let $\|f-g\|_{sup} = \|\phi\|_{sup} = \delta$.

In addition, if $\phi(0) = 0$, then

$$\frac{1}{2}H^2(f,g) = \int \sqrt{1+\phi(t)}\Big|_x^0\,dx$$

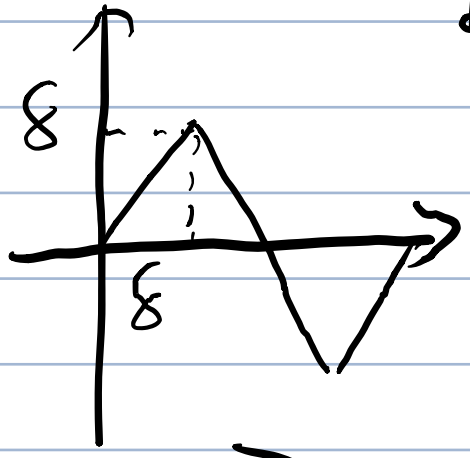
$$\le \int C_1\phi(x) + C_2\phi^2(x)\,dx,$$

(where $C_1 = u'(0)$, $C_2 = \sup_x u''(x)$, $u(s) = \sqrt{1+s}$.)

$$\lesssim \int \phi^2(x) \le \int_0^\delta t^2\,dt \asymp \delta^3$$

Let $\delta^3 \asymp \frac{1}{n}$

$$\Rightarrow w_\theta\left(\frac{1}{2\sqrt{n}}\right) \gtrsim n^{-\frac{1}{3}}$$



$$\Rightarrow M(\varepsilon, \Theta(F)) \gtrsim \Phi(n^{-\frac{1}{3}}).$$

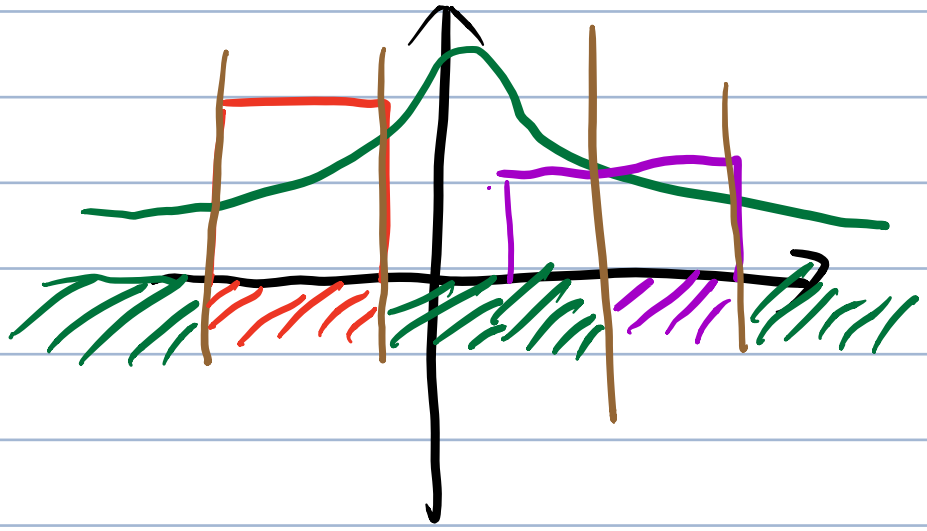
Fano's method

For $M \neq 2$, we need to evaluate the error of multiple testing, i.e.

$$\frac{1}{M}\sum_{i=1}^{M} P_{\theta_i}(\hat{\phi} \neq i)$$

$$= \frac{1}{M}\sum_{i=1}^{M} \int_{D_{\hat{\phi}}^{i}} f_{\theta_i}$$

$$\geq \frac{1}{M}\int \min_i \{ f_{\theta_i} \}$$



An interesting problem is that "Would the optimal error get smaller and smaller if you make more and more comparisons"?

To prove this intuition, we need to introducing some elements of Information Theory.

Entropy

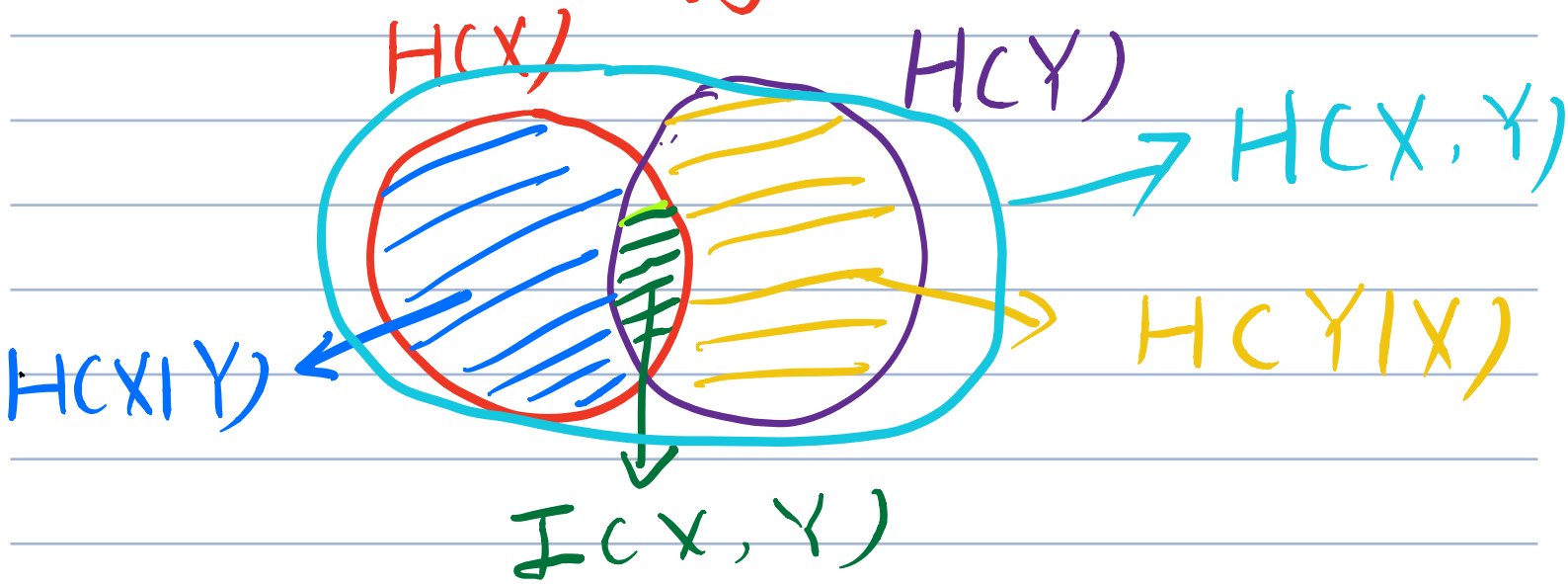
X, Y are two random elements (no need the same), which joint distribution exists.

$I(X;Y) := KL(P_{X,Y} \| P_X P_Y)$, which quantifies the dependence of X and Y.
(Mutual Information)

$H(X) := -\int p_X \log p_X$, which quantifies the information (randomness) of X.
(Shannon entropy)

$H(X|Y=y) := H(Z), \quad Z \sim X|Y=y.$

$H(X|Y) = \int H(X|Y=y)\, P_Y(dy).$
(Conditional entropy)



Remark:

① If X is a random element valued in $\mathcal{X}$, $|\mathcal{X}| < \infty$, then

$$H(X,U) = H(U) + H(X|U),$$

(where $U = F^{-1}(X) \sim \text{Unif}\{1,\dots,|\mathcal{X}|\}$)

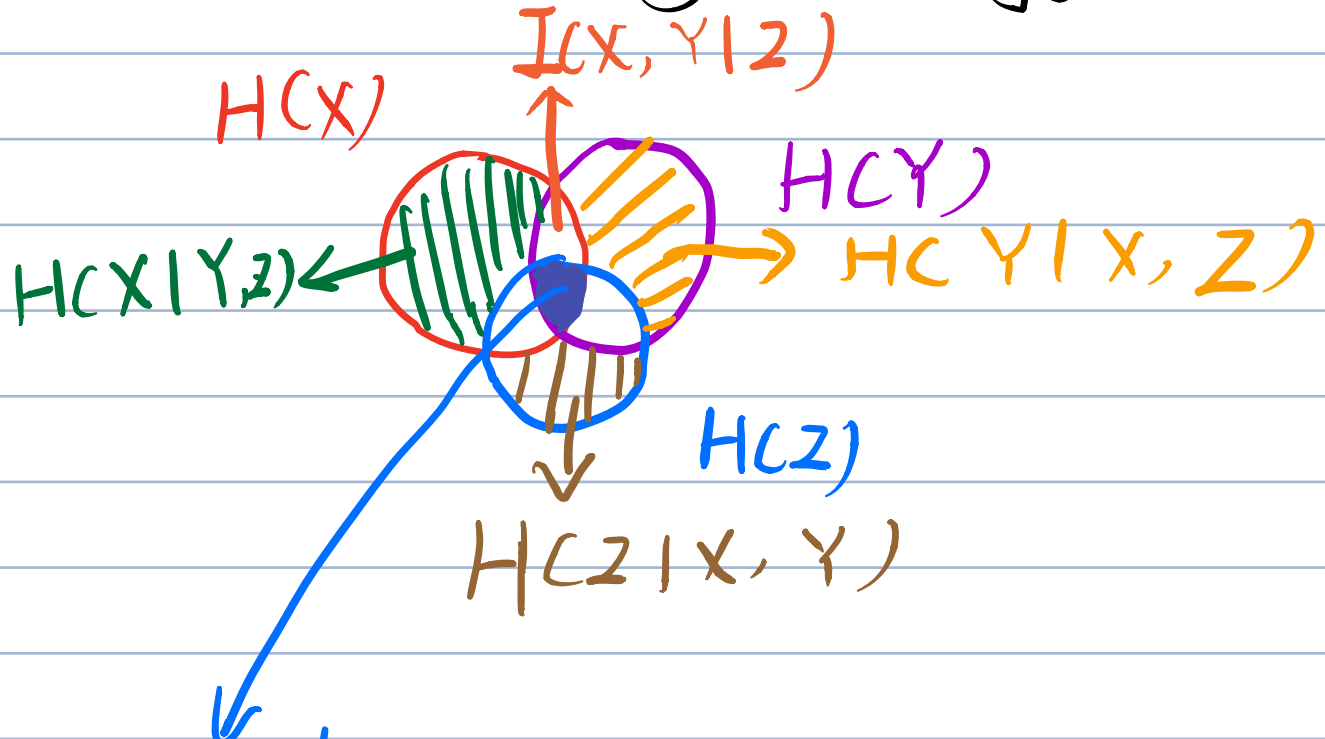
$$= H(U) = \log|\mathcal{X}|$$

$$\Rightarrow H(X) \le H(X,U) = \log|\mathcal{X}|.$$

$\Rightarrow$ the uniform distribution is the most random element.

② We can define $H(X,Y,Z)$ similarly:

$$H(X,Y,Z) = -\int p_{X,Y,Z} \log p_{X,Y,Z}.$$



Not well-defined! And $I(X,Y|Z)$ may be larger than $I(X,Y)$.

Return to our problem, let $U \sim Unif\{1,\dots,M\}$, $Z|U=i \sim P_{\theta_i}$, then

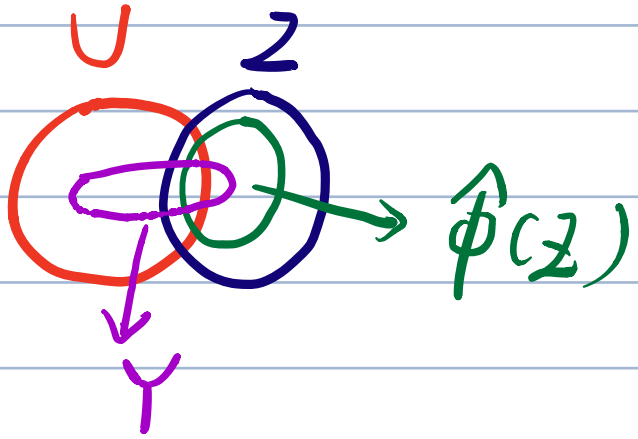
$\frac{1}{M}\sum_{i=1}^{M} P_{\theta_i}(\hat{\phi} \neq i) = \mathbb{P}(\hat{\phi}(Z) \neq U) := p$,

Let $Y = \mathbb{1}(\hat{\phi}(Z) \neq U)$, then

$Y \sim B(1, p)$.

$U \longrightarrow Z \rightarrow \hat{\phi}(Z)$

Information transfers though the chain (Markov)



$$\Rightarrow H(U,Y|Z) = H(U|Z)$$

$$= H(U|Y,Z) + H(Y|Z)$$

$$= p\,H(U|Y=1,Z) + H(Y|Z)$$

$$\leq p\log(M-1) + H(Y)$$

$$= p\log(M-1) - p\log p - (1-p)\log(1-p)$$

$$\leq p\log(M-1) + \log 2$$

$$\Rightarrow p \geq \frac{H(U|Z) - \log 2}{\log(M-1)}$$

$$\Rightarrow p \geq \frac{\log M - I(U;Z) - \log 2}{\log(M-1)}$$

$$\geq 1 - \frac{I(U;Z) + \log 2}{\log M}.$$

(Fano)

If $\exists \{\theta_i\}_{i=1:M} \subset \Theta$ for given $\delta > 0$ s.t. $d(\theta_i, \theta_j) \geq 2\delta$, and

$$1 - \frac{I(U;Z) + \log 2}{\log M} > 0, \text{ then}$$

$$M(\Theta) \gtrsim \Phi(\delta), \text{ where}$$

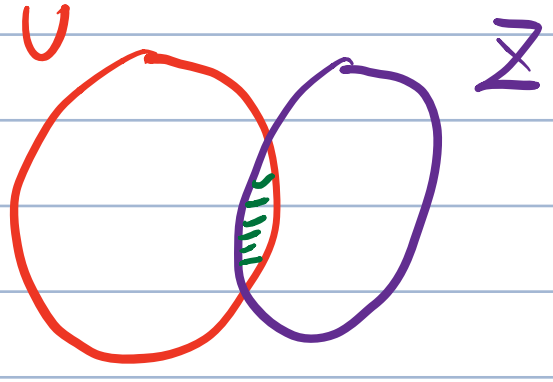
$U \sim \text{Unif}\{1, \cdots, M\}$, $Z|U=i \sim P_{\theta_i}$.

Remark:

① $\log M = H(U) = I(U;U)$, as $\delta \to 0^+$, if

$$\frac{I(U;Z)}{I(U;U)} \to 0^+,$$

then Fano holds.



② $I(U;Z) = KL(P_{U,Z} \| P_U P_Z)$

$= KL(P_{Z|U} P_U \| P_Z P_U)$

$= \int P_{Z|U} P_U \log \frac{P_{Z|U}}{P_Z}$

$= \frac{1}{M} \sum_{i=1}^{M} P_{\theta_i} \log \frac{P_{\theta_i}}{P_Z}$

$= \frac{1}{M} \sum_{i=1}^{M} KL(P_{\theta_i} \| \frac{1}{M} \sum_{k=1}^{M} P_{\theta_k})$

$\leq \frac{1}{M^2} \sum_{i,k} KL(P_{\theta_i} \| P_{\theta_k})$

Example 6 (Minimax risk for linear regression)

$Y = X\beta_0 + \varepsilon$ X is fixed, and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2 I_n)$, we want to evaluate

$$\mathcal{M}(\Theta) := \inf_{\hat{\beta}} \sup_{\beta \in \Theta} \mathbb{E} \|\hat{\mu} - \mu\|_n^2, \quad \mu := X\beta$$

where $\Theta = Im(X)$. (Identifiable)

Let $r = \dim(\Theta) = rank(X)$, then

$$P(2\delta; B_\Theta(0;d)) \asymp \left(\frac{d}{\delta}\right)^r,$$

and

$$KL(P_{\beta_1} \| P_{\beta_2}) = \frac{\|X(\beta_1 - \beta_2)\|_2^2}{2\sigma^2} =: \frac{\|\mu_1 - \mu_2\|_n^2}{2\sigma^2} \leq \frac{n\delta^2}{\sigma^2}$$

To ensure $0 < \frac{\delta^2 16^2}{r \log d/\delta} < 1$

Let $d = 2\delta$, $\delta = \sqrt{\frac{r}{n}}\sigma$, then
$M(\mathcal{H}) = \Phi(c\sqrt{\frac{r}{n}}\sigma)$.

Example 7 (Minimax risk for sparse mean)

(Gilbert - Varshamov)

$\mathcal{H} := \{\theta \in \{0,1\}^p, \|\theta\|_0 = s\}$, if $1 \le s \le p/8$, then there exists $\{\theta_i\}_{i=1:M} \in \mathcal{H}$ s.t.

$$\log M \gtrsim s \log(1 + \frac{p}{2s})$$

and

$$\|\theta_i - \theta_j\|_0 \ge s/2, \quad i \ne j.$$

Pf: Let $X_i \overset{iid}{\sim} \mathrm{Unif}(\mathcal{H})$, $i = 1, \dots, M$

we want to prove

$$\mathbb{P}(\bigcap_{i \ne j} \|X_i - X_j\|_0 \ge s/2) > 0 \text{ if}$$

$$\log M \gtrsim s \log(1 + \frac{p}{2s})$$

□

Now we have

$Y = \mu + \varepsilon$, $\beta_0 \in \mathbb{R}^p$ and $\|\mu\|_0 \le s$, $\varepsilon \sim N(0, \sigma^2 I_n)$, similarly:

$$KL(P_{\mu_1}^n, \| P_{\mu_2}^n) = n\|\mu_1 - \mu_2\|^2 / 2\sigma^2$$

In order to ensure, let $\|\mu_i - \mu_j\| \ge 2\delta$, in order to ensure

$$0 < \frac{n(2\delta)^2/2\sigma^2}{\log M} < 1,$$

Let $\delta \asymp \sqrt{s \log(1 + \frac{p}{s})/n} \cdot \sigma$, indeed $\{\mu_i\}$ s.t. $\|\mu_i - \mu_j\| \geq 2\delta$ exists since

Let $\mu_i = \theta_i \delta / s$, then

$$\|\mu_i - \mu_j\| = \frac{\delta}{s} \|\theta_i - \theta_j\|_0 \geq \delta.$$