# Minimax risk theory

# 1. Problem Setting

Data : $\underbrace{X_1, X_2, \cdots, X_n}$ ~ $P$ , $P \in \mathcal{P}$ family

$$\hat{\theta}(X_1, X_2, \cdots, X_n) \xrightarrow{estimate} \theta(P)$$

Wondering :

$$R_n = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{X_i \sim P} [d(\hat{\theta}, \theta(P))]$$

"minimax risk"

(For lower regret bound, we should firstly find a $d(\cdot, \cdot)$ that lower bounds Reg.)

$d$ satisfies the "triangular inequality".

Example 1: $\mathcal{P} = \{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$

want to know $R_n = \inf_{\hat{\theta}} \sup_{\theta} \mathbb{E}[(\hat{\theta} - \theta)^2]$

why? $\Longrightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i$

Note: What we really care about is the problem hardness than the estimator.

minimax risk is useful to represent the hardness even for non-parametric family where we donot even have a close-form solution.

---

To ~~know~~ $R_n$ is not necessary.
$\downarrow$
bound as $L_n \le R_n \le U_n$ is necessary.

(best result: $U_n = O(L_n)$)

---

Define 2 metrics:

(i) $KL(P,Q) = \int_P p \log \frac{p}{q}$ or $= \sum_i P_i \log \frac{P_i}{q_i}$

Property 1: $KL(P^n, Q^n) = n \cdot KL(P,Q)$ for iid

Property 1': $KL(P^T, Q^T) = \sum_{t=1}^{T} KL(P[X_t | X^{t-1}], Q[X_t | X^{t-1}])$ for martingale

called "Chain rule of Relative entropy"

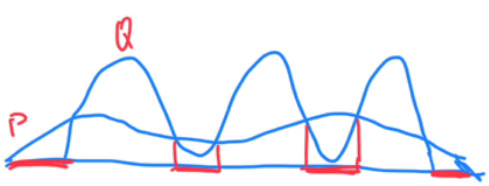
(ii) $TV(P,Q) = \sup_{A} |P(A) - Q(A)| = \frac{1}{2}\int |P-q|$

$A$ event

Homework

Proof: $\int |P-q| = \underbrace{\int_{P(x)\ge q(x)} (P(x)-q(x))\,dx}_{C}$

$+ \underbrace{\int_{P(x)<q(x)} (q(x)-P(x))\,dx}_{C'}$




Notice that $C - C' = \int (P(x)-q(x))\,dx$

$$+\int_{c}(P(x)-q(x))dx$$
$$=\int(P(x)-q(x))dx$$
$$=1-1$$
$$=0$$

$\Rightarrow C=C'$

Also, $\sup_A |P(A)-Q(A)| = \max\{C, C'\}$

since C is the sum of all positive items, and C' is the sum of all negative items.

$$\Rightarrow \int |p-q| = 2\sup_A |P(A)-Q(A)|$$

---

Property 2: $TV(P,Q) \leq \sqrt{KL(P,Q)}$

---

Recall: $R_n = \inf_{\hat\theta} \sup_{P\in\mathcal{P}} \mathbb{E}_P[d(\hat\theta, \theta(P))]$ hard to calculate

△ Reduction ①: "Finite Covering"

discretize distribution family $\mathcal{P}$

Pick a finite set $M=\{P_1, P_2, \ldots, P_N\}\subset\mathcal{P}$

Then:

$$R_n = \inf_{\hat\theta} \sup_{P\in\mathcal{P}} \mathbb{E}_P[d(\hat\theta, \theta(P))]$$

$$\geq \inf_{\hat{\theta}} \sup_{P_j \in M} \mathbb{E}_{P_j}[d(\hat{\theta}, \theta_j)] \quad , \theta_j = \theta(P_j)$$

Remark: if $M$ can be a good representative of family $\mathcal{P}$, then this lower bound would approach $R_n$. Both "Representable" and "Indistinguishable"

Define: $S = \min_{j \neq k} d(\theta_j, \theta_k)$, then

Reduction ②: by Markov

$$P_j\left[d(\hat{\theta}, \theta_j) > \frac{S}{2}\right] \leq \frac{2}{S} \cdot \mathbb{E}_{P_j}[d(\hat{\theta}, \theta_j)] \quad , \forall j \in [N]$$

$[N] = \{1, 2, \cdots, N\}$

$$\Rightarrow \mathbb{E}_{P_j}[d(\hat{\theta}, \theta_j)] \geq \frac{S}{2} \cdot P_j\left[d(\hat{\theta}, \theta_j) > \frac{S}{2}\right]$$

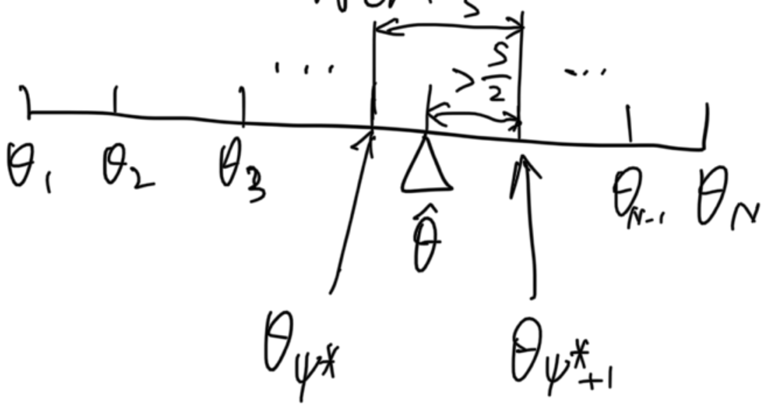
$$\Rightarrow R_n \geq \inf_{\hat{\theta}} \sup_{P_j \in M} \mathbb{E}_{P_j}[d(\hat{\theta}, \theta_j)]$$

$$\geq \frac{S}{2} \cdot \inf_{\hat{\theta}} \sup_{P_j \in M} P_j\left[d(\hat{\theta}, \theta_j) > \frac{S}{2}\right]$$

Still hard to compute, we need further lower bound

How to lower bound

$$\sup_{P_j \in \mathcal{M}} P_j[d(\hat{\theta}, \theta_j) > \tfrac{s}{2}] \quad ?$$




Denote

$$\psi^*(\hat{\theta}) = \arg\min_j d(\hat{\theta}, \theta_j)$$

Then: for any $j \neq \psi^*$, we have

$$d(\hat{\theta}, \theta_j) \geq \frac{s}{2}$$

Since

$$\begin{aligned} s &\leq d(\theta_j, \theta_{\psi^*}) \\ &\leq d(\theta_j, \hat{\theta}) + d(\hat{\theta}, \theta_{\psi^*}) \\ &\leq d(\theta_j, \hat{\theta}) + d(\hat{\theta}, \theta_j) \\ &= 2d(\hat{\theta}, \theta_j) \end{aligned}$$

$\Rightarrow$ **Reduction ③**:

$$\begin{aligned} P_j[d(\hat{\theta}, \theta_j) \geq \tfrac{s}{2}] &\geq P_j[\psi^*(\hat{\theta}) \neq j] \\ &\geq \inf_{\psi(\hat{\theta})} P_j[\psi(\hat{\theta}) \neq j] \\ &\geq \inf_{\psi(X_1, X_2, \cdots, X_n)} P_j[\psi \neq j], \quad \forall \psi: X^n \to [N] \end{aligned}$$

Remark: Notice that $\psi^*(\hat{\theta})$ is sth you choose, related to $\hat{\theta}$ but $j$ is chosen by the

..., j is chosen by the sup, or, say, by the nature.

$$\Longrightarrow R_n \geq \frac{s}{2} \cdot \inf_{\hat{\theta}} \sup_{P_j \in M} P_j[d(\hat{\theta}, \theta_j) \geq \frac{s}{2}]$$

$$\geq \frac{s}{2} \cdot \inf_{\psi} \sup_{P_j \in M} P_j[\psi \neq j]$$

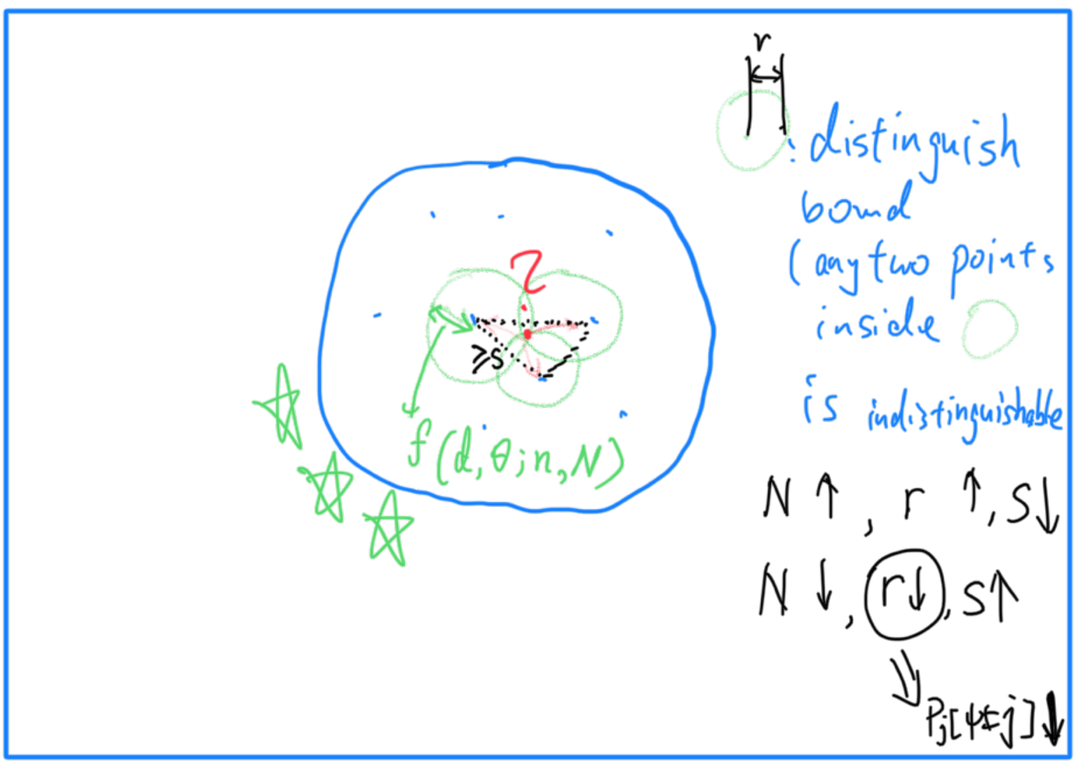
Remember:

$s = \min_{j \neq k} d(\theta_j, \theta_k)$

compete

a minimax game

Illustration :




In the following, we focus on :

$R_n = \inf \sup E[d(\hat{\theta}, \theta(P))] \geq \frac{s}{2} \inf \max P_j[\psi \neq j]$

$R_n = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\hat{\theta}, \theta(P))] \geq \frac{S}{2} \inf_{\psi} \max_{P_j \in M} P_j[\psi \neq j]$

want them to be as large

Recall that $S = \min_{j \neq k} d(\theta_j, \theta_k)$ is related to $M$

$M$ too dense $\to$ selection error $\uparrow$ (hard to distinguish between similar $P_j$ and $P_k$)
$\to$ $S \downarrow$

$M$ too sparse $\to$ $S \uparrow$
$\to$ selection error $\downarrow$

$\to$ How to characterize the "distinguishability"?

i.e., how to calculate $\inf_{\psi} \max_{j} P_j[\psi \neq j]$?

---

3 Standard methods:

- Le Cam
- Fano
- Tsybakov

1. Le Cam: distinguish between 2 guys:

$$M = \{P_0, P_1\}$$

Theorem 4 (Le Cam): for any $P_0, P_1 \in \mathcal{P}$, we have

$$R_n = \inf_{\hat{\theta}} \sup \mathbb{E}[d(\hat{\theta}, \theta(P))] \geq \frac{S}{8} \cdot e^{-n \cdot KL(P_0, P_1)}$$

$\hat{\theta} \quad P \in \mathcal{P}$

(Here $S = d(\theta(P_0), \theta(P_1))$)

**Corollary**: if $KL(P_0, P_1) \le \log \frac{2}{n}$, then

$$R_n = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}[d(\hat{\theta}, \theta(P))] \ge \frac{S}{16}$$

Notice that LHS is nothing to do with $P_0$ and $P_1$. Therefore, just find 2 distributions in $\mathcal{P}$ that is $\log \frac{2}{n}$ apart, and then they are hard to distinguish, and this in return ensures the lower bound inequality to be true.

---

**Proof of Le Cam's Theorem**: (maybe show later?)

$$R_n \ge \frac{S}{2} \cdot \inf_{\psi} \max\{P_0(\psi = 1), P_1(\psi = 0)\}$$

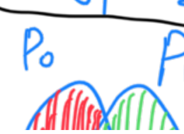
here $\psi: X \to \{0, 1\}$

$$\ge \frac{S}{2} \inf_{\psi} \frac{P_0(\psi = 1) + P_1(\psi = 0)}{2}$$

"Type 1 error"

**Neyman-Pearson's Test**

$$\psi^*(x) = \begin{cases} 0 & \text{if } P_0(x) \ge P_1(x) \\ 1 & \text{if } P_0(x) < P_1(x) \end{cases}$$



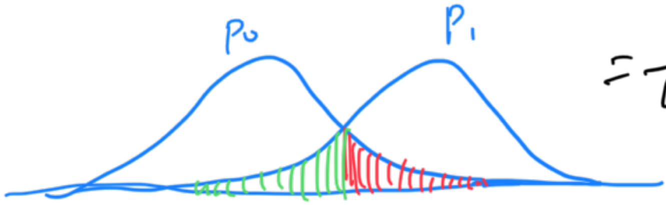
Lemma: for any $\psi$, we have

$$P_0(\psi=1)+P_1(\psi=0) \geq P_0(\psi^*=1)+P_1(\psi^*=0)$$

[Optimality of Neyman-Pearson's Test)

(No proof)

$$\geq \frac{S}{2}\cdot\frac{1}{2}\left(P_0(\psi^*=1)+P_1(\psi^*=0)\right)$$




$$= \frac{S}{4}\cdot\left(\int_{P_1(x)\geq P_0(x)} P_0(x)\,dx + \int_{P_1(x)<P_0(x)} P_1(x)\,dx\right)$$

$$= \frac{S}{4}\int \min\{P_0(x), P_1(x)\}\,dx$$

"Le Cam's Lemma:"

$$\inf_{\hat\theta}\sup_{P\in\mathcal{P}} \mathbb{E}_P[d(\hat\theta,\theta(P))] \geq \frac{S}{4}\cdot\int \min\{P_0(x),P_1(x)\}\,dx$$

Lemma 7: for any $P$ and $Q$, we have:

$$\int \min\{P(x),q(x)\}\,dx \geq \frac{1}{2}e^{-KL(P,Q)}$$

(No proof) (Hint: by Jenson's Inequality

With Lemma 7 we prove Thm 4

Lemma 1,

---

Example of Le Cam:

$$Y = m(X_i) + \varepsilon, \quad X_i \sim \mathrm{Unif}(0,1), \quad \varepsilon_i \sim N(0,1)$$

Let $m(\cdot) \in \mathcal{M} = \{ m : |m(y) - m(x)| \le L \cdot |x - y|, \ x, y \in [0,1] \}$

(L-lipschitz family)

We want to bound the minimax risk of

$$\theta := m(0)$$

loss function: $d(\theta_0, \theta_1) = |\theta_0 - \theta_1|$

Distributions: $P(x, y) = \phi(x)\, p(y|x)$ $\sim N(m(X), 1)$

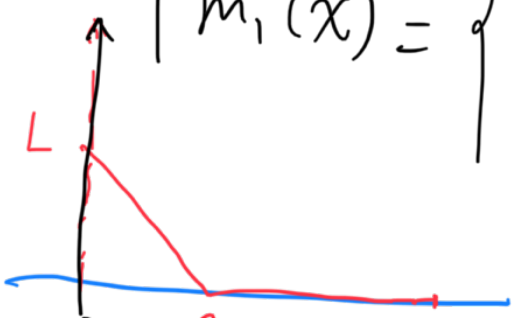
uniform

---

Now, in order to apply Le Cam's method, we have to choose 2 different $m(\cdot)$ functions from lipschitz family.

$$\begin{cases} m_0(x) = 0 \\ m_1(x) = \begin{cases} L(\varepsilon - x) & x \in [0, \varepsilon] \\ 0 & x \in (\varepsilon, 1] \end{cases} \end{cases}$$

, with $\varepsilon$ to be determined later



0 $\varepsilon$ 1 $m_0$

$$\Rightarrow KL(P_0,P_1) = \iint p_0(x,y)\log\frac{p_0(x,y)}{p_1(x,y)}\,dxdy$$

$$= \int_0^{\varepsilon} KL(N(0,1), N(m_1(x),1))\,dx$$

$$= \int_0^{\varepsilon} \frac{L^2(\varepsilon-x)^2}{2}\,dx$$

$$= \frac{L^2\varepsilon^3}{6} \longleftarrow \frac{\log 2}{n}$$

Note:

$$KL(N(\mu_0,\sigma^2), N(\mu_1,\sigma^2)) = \frac{(\mu_1-\mu_0)^2}{2\sigma^2}$$

Let $\varepsilon = \left(\frac{c}{n}\right)^{\frac{1}{3}}$, and we have:

$|m_1(0) - m_2(0)| \geq \left(\frac{c}{n}\right)^{\frac{1}{3}}$ according to Le Cam's lemma, i.e.,

$$\inf_{\hat{m}(0)} \sup_{m\in\mathcal{M}} \mathbb{E}[|\hat{m}(0) - m(0)|] \geq \left(\frac{C}{n}\right)^{\frac{1}{3}}$$

, which is a lower error bound.

Note: This is also an upper bound. $\Rightarrow$ minimax risk.

Note.2: for high-dimensional case: $\dots = \Omega\left(\left(\frac{1}{n}\right)^{\frac{1}{d+2}}\right)$

Recall: how we constructed $m_1(x)$?

for Lipschitz

"Bump function"

for high-order smooth

Intuition: for most of the part, they are the same, but for one ("some" in $F_{\infty}$) place that is "controllable different".

---

But: if we would like to lower bound:

$$\mathbb{E}\left[\int (\hat{m}(x) - m(x))^2 \, dx\right]$$

then a Le Cam method would not have a tight bound. We will use more advanced methods.

---

Recap: $$R_n = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}[d(\hat{\theta}, \theta(P))]$$

$$\geq \frac{\delta}{2} \cdot \inf_{\psi} \max_{j} P_j(\psi \neq j)$$

We want $L_n \leq R_n \leq U_n$

## 2. Fano's Method

Theorem 11. (Fano's Inequality)

$$\max_j P_j(\psi \neq j) \geq \frac{1}{N} \sum_{j=1}^{N} P_j(\psi \neq j)$$

$$\geq 1 - \frac{n\beta + \log 2}{\log N}, \quad \text{with}$$

(No Proof)

$$N > 2, \quad \beta = \max_{j \neq k} KL(P_j, P_k)$$

With Fano's Inequality, we have

$$R_n \geq \frac{S}{2} \cdot \left[ 1 - \frac{n\beta + \log 2}{\log N} \right] \geq \frac{S}{4}, \text{ if } \beta \leq \frac{\log N}{4n}$$

— "Fano Minimax Bound" (Corollary 13)

---

## 3. Tsybakov Bound

Theorem 14. $\underbrace{\text{For } N \geq 3,\ P_j \ll P_0}_{\text{Given } X_1, X_2, \cdots, X_n \sim P \in \mathcal{P}}$ if $\frac{1}{N} \sum_{j=1}^{N} KL(P_j, P_0) \leq \frac{\log N}{16n}$,

then:

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\hat{\theta}, \theta(P))] \geq \frac{S}{16}$$

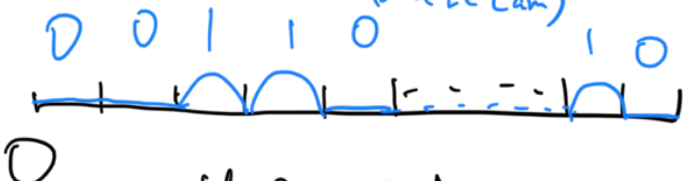
where $S = \min_{0 \le j < k \le N} d(\theta(P_j), \theta(P_k))$

---

Example 2: Lower bound $\int_0^1 (\hat{m}(x) - m(x))^2 dx$,

where $Y_i = m(X_i) + \varepsilon_i \quad m \in \{m_0, m_1, \cdots m_N\}$.

"One bump (and Le Cam) is not sufficiently valid".




$u$ segments

$\Rightarrow \begin{cases} 1 \to \text{add a bump} \\ 0 \to \text{add no bump} \end{cases}$

An $m_i \longleftrightarrow$ a $\{0,1\}^u$ sequences $=: \mathbb{1}_i$

$\Rightarrow N = 2^u$

Define $d(m_j, m_k) = \int (m_j(x) - m_k(x))^2 dx$

$= a(u)^2 \cdot \underbrace{\| \mathbb{1}_j - \mathbb{1}_k \|_1}_{\text{"Hamming distance"}}$

one bump

Of course, Le Cam does not work this time.

Note: if we choose all $N = 2^u$ functions,

then $S = \min_{j \neq k} d(m_j, m_k) = a_{(u)}^2$ is small

$\Rightarrow R_n \ge \frac{S}{16n}$ is not sufficient to make sense.

Therefore, we have to choose a subset of all $\{0,1\}^u$, where each pair is far enough from each other and total number is large enough as well (or otherwise they are "easy to be distinguish" and again lead to a trivial lower risk bound).

Lemma 15: Varshamov - Gilbert's Lemma:

For $N \geq 8$, There exists a set

$$\{w^0, w^1, \dots, w^N\} \subset \{0,1\}^u \text{ such that}$$

$(w^0 = [0,0,\cdots,0]^T \in \mathbb{R}^u)$

(i) "Constant Proportion Difference Property"

i.e. $d(w^j, w^k) \geq \frac{u}{8}$

(ii) "Exponential Number Hypothesis Property"

i.e. $N \geq 2^{\frac{u}{8}}$.

Proof: Random Sample and Hoeffding Inequality

---

Hoeffding: $\Pr[|S - E[S]| \geq t] \leq \dots \frac{2t^2}{\dots}$

$$\Pr\left[\,|S_n - \mathbb{E}[S_n]| \geq t\,\right] \leq 2\exp\left(-\frac{2t^2}{\sum(b_i-a_i)^2}\right),\ t>0$$

$$\text{or } \Pr\left[\,S_n - \mathbb{E}[S_n] \geq t\,\right] \leq \exp\left(-\frac{2t^2}{\sum(b_i-a_i)^2}\right),\ t>0$$

$$\Rightarrow \Pr\left[d(w^i,w^j) < \frac{u}{8}\right] \leq \exp\left(-\frac{2\cdot\left(\frac{3u}{8}\right)^2}{u}\right)$$

$$= \exp\left(-\frac{9}{32}u\right),\ \forall i,j\in[N]$$

$$\Rightarrow \Pr\left[d(w^i,w^j) \geq \frac{u}{8},\ \forall i,j\in[N]\right] \geq 1 - C_N^2\cdot\exp\left(-\frac{9}{32}u\right)$$

$$\geq 1 - \frac{\left(2^{\frac{u}{8}}\right)^2}{2}\cdot\frac{1}{e^{\frac{9}{32}u}}$$

$$\geq 1 - \frac{1}{e^{\frac{u}{32}}}$$

Total number of combinations:

$$\left(2^u\right)^N$$

$$\Rightarrow \#\ \text{satisfying combination} \geq \left(2^u\right)^N\cdot\left(1-\frac{1}{e^{\frac{u}{32}}}\right) > 1$$

$$\Rightarrow \exists \text{ a combination satisfying } \begin{cases} d(w_i,w_j) \geq \frac{u}{8} \\ N \geq 2^{\frac{u}{8}} \end{cases}$$

$\square$

---

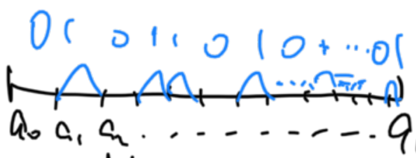
With this V-G theorem, we may construct a set M with $N = 2^{\frac{u}{8}}$ and $s = \frac{u}{8}$.

△ Example 2 continue:

$$Y_i = m(X_i) + \varepsilon_i$$

$$\mathcal{M} = \{ m : | m^{(\cdot)}(x) - m^{(\cdot)}(y) | \le L \cdot |x-y| \}$$

( $\beta$-order-smoothness family)




$u$ segments in total

$$d(m_i, m_j) = \sqrt{\int (m_i - m_j)^2}$$

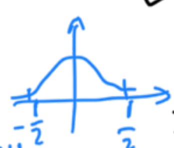
new definition for $\|m_i - m_j\|_2$

$$= \sqrt{\left(\int (\text{one bump})^2\right)} \cdot \sqrt{\|\mathbb{1}_i - \mathbb{1}_j\|_1}$$

each bump $B_j(x)$

$$= \frac{L}{u^\beta} \cdot K\left(\frac{x - a_j}{\frac{1}{u}}\right)$$

$K : [-\frac{1}{2}, \frac{1}{2}] \to \mathbb{R}^+$




$\beta$-order smooth

$$\ge \frac{c}{u^{\beta + \frac{1}{2}}} \cdot \sqrt{\frac{u}{8}} .$$

$$= C_1 \cdot u^{-\beta}$$

Also, $KL(P_i, P_j) = C_2 \cdot u^{-2\beta}$

We omit the calculation

If we want to use Fano's method

we have to make $KL(P_i, P_j) \le \frac{\log N}{4n}$

$$= \frac{\log 2^{\frac{u}{8}}}{4n}$$

$$= \frac{u}{32n}$$

$$\Longrightarrow \quad C_2 \cdot u^{-2\beta} \le \frac{u}{32n}$$

$$\Rightarrow \quad u = (32 C_2 \cdot n)^{\frac{1}{2\beta+1}}$$

$$\Rightarrow \quad R_n \geq \frac{s}{4} = \frac{1}{4} \cdot C_1 \cdot u^{-\beta}$$

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

---

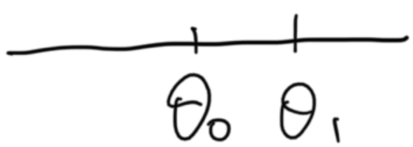
Note: for $d$-dimension setting, we have $R_n \geq \Omega\left(n^{-\frac{\beta}{2\beta+d}}\right)$ in similar way.
This matches the realized upper bound.

---

Example 4: (not to discuss)

Parametric model : $R_n = \Theta\left(\frac{1}{\sqrt{n}}\right)$.

i.e., $\inf_{\hat{\theta}} \sup_{\theta} \mathbb{E}|\hat{\theta} - \theta| \geq \frac{1}{\sqrt{n}}$




, $\theta_1 = \theta_0 + \frac{1}{\sqrt{n}}$

then $R_n \geq \frac{1}{\sqrt{n}} e^{-n KL(P_{\theta_0}, P_{\theta_1})}$ by Le Cam

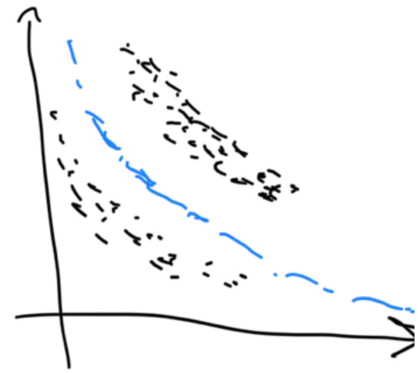
Recap: $KL(P_0, P_1) \simeq \frac{1}{n} \cdot I(\theta)$

Fisher Information

Example 5: (not to taught)

semi-supervised learning

e.g. no-label classification



Example 6: $X_1, X_2, \cdots, X_n \overset{iid}{\sim} P$ (if time permitted)

Usually: $R_n = \Theta\left(n^{-\frac{2\beta}{2\beta+1}}\right)$ (for square loss)

but:

Observe: $Z_i = X_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$, but no access to $X_i$

Estimate the density of $X_i$

$$\Rightarrow R_n = \Theta\left(\left(\frac{1}{\log n}\right)^{\beta}\right), \text{ much slower!}$$

Similarly, if we'd like to estimate $m(\cdot)$ from

$Y_i = m(X_i) + \varepsilon_i$ but only observe

$Z_i = X_i + \delta_i \in N(0, \sigma^2)$, then the $R_n$ is also logarithmic. i.e. we basically cannot estimate $m(\cdot)$ ...

estimate $m$

Very surprising!

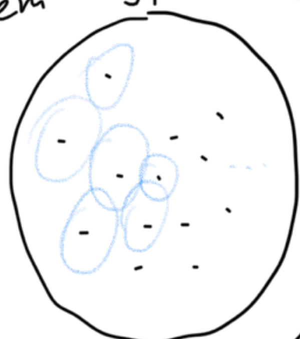
(Note: if we only want $Z_i \to Y_i$ prediction, then it is still fine, but it's not sth that fits $m(\cdot)$ well.)

Difference between prediction and function estimating.

---

In general,

problem hypothesis family



Distinguishability

concentration → likelihood $\xrightarrow{\text{argmax}}$ Upper bound

discretization → covering $\xrightarrow{\text{count covering number}}$ Lower bound $N(\varepsilon)$

" Le Cam equation "

$\log N(\varepsilon) = n\varepsilon^2 \Rightarrow \varepsilon_n$ is exactly the minimax

"Entropy"

under some certain condition.