Chapter 5 : Minimaxity

- Suppose we observe $w \in \mathcal{W}$ drawn from a distr. P belonging to a statistical model $\mathcal{P}$.
  - ↳ We'll be most interested in setting $P = Q^n$ with Q^n denoting the n-fold product measure of distr Q.

- Based on data realization w, we take action $a \in \mathcal{A}$
  - ↳ We'll consider a collection of decision rules $T : \mathcal{W} \to \mathcal{A}$

- The quality of an action a is judged by loss func. $L : \mathcal{A} \times \mathcal{P} \to \mathbb{R}$.

- The quality of a decision rule, the risk which corresponds to the expected loss :

$$R(T, P) = \int_{\mathcal{W}} L(T(w), P)\, d\mathcal{P}(w)$$

Ex 1 : Point estimation

The objective is estimate $\Psi(P) \in \mathbb{R}$.

↳ E.g. : $P = Q^n$, and $\Psi(P) = \int x\, dQ(x)$

If $L(a, P) = \{a - \Psi(P)\}^2$ is the squared error loss

then $R(T, P) = \int \{T(w) - \Psi(P)\}^2 \, dP(w)$

is the mean-squared error.

Ex 2: Estimating a regression function

- We observe n iid copies of $(X, Y) \sim Q$ with support on $\mathcal{X} \times \mathbb{R}$ and we wish to estimate the func. $f_Q : x \longmapsto \mathbb{E}_Q(Y \mid X = x)$

- The action space is a collection of funcs. mapping $\mathcal{X}$ to $\mathbb{R}$.

- We may consider an integrated squared error loss: $L(a, P) = \int \{a(x) - f_Q(x)\}^2 \, d\nu(x)$

  The corresponding risk is known as the mean integrated squared error.

Recall from 581, performance of decision rule can be judged by its maximal risk:

$$\sup_{P \in \mathcal{P}} R(T, P)$$

A minimax estimator minimizes the maximal risk.

Challenge: Deriving a minimax estimator is often an intractable problem.

Minimax rate optimality

Because we often can't derive the minimax estimator, we'll instead look to find estimator that achieves minimax optimal rate.

For now, consider the case $P = Q^n$

For each n, let $\mathcal{T}_n$ denote the collection of allowed decision rules.

In many problems, the risk is always non-negative, and

$$\inf_{T \in \mathcal{T}_n} \sup_{Q \in \mathcal{Q}} R(T, Q^n) \xrightarrow{n \to \infty} 0$$

Focus on these decision problems here.

A seq. of decision rules $\{T_n\}_{n=1}^{\infty}$ is called minimax rate optimal if the following:

$$(\star) \quad \liminf_{n \to \infty} \frac{\inf_{T \in \mathcal{T}_n} \sup_{Q \in \mathcal{Q}} R(T, Q^n)}{\sup_{Q \in \mathcal{Q}} R(T_n, Q^n)} > 0$$

$$\frac{\inf_{T \in \mathcal{T}_n} \sup_{Q} R(T, Q^n)}{\sup_{Q} R(T_n, Q_n)} \leq \frac{\sup_{Q} R(T_n, Q^n)}{\sup_{Q} R(T_n, Q^n)} = 1$$

If the numerator is of the form $C_1 n^{-\alpha}$ and denominator is of the form $C_2 n^{-\beta}$ and

(★) holds iff $\alpha = \beta$.

Problem: We can't typically derive the form of the numerator or denominator of (★)!

Sol'n: Find bounds on them.

Later in this course, we'll derive upper bounds on the maximal risk of particular estimators $\{T_n\}_{n=1}^{\infty}$.

↳ That is, we'll find $\{U_n\}_{n=1}^{\infty}$ s.t.

$$\sup_{Q} R(T_n, Q^n) \leq U_n$$

In this chapter, we'll derive lower bounds on the minimax risk, of the form

$$\inf_{T \in \mathcal{T}_n} \sup_{Q} R(T, Q^n) \geq L_n$$

Once we have $\{L_n\}$, $\{U_n\}$,

$$(\star)\quad \frac{\inf_T \sup_Q R(T, Q^n)}{\sup_Q R(T_n, Q^n)} \geq \frac{L_n}{U_n} > 0$$

(asymp.)

Hence, to show $(\star)$, it suffices to show

$$\liminf_{n\to\infty} \frac{L_n}{U_n} > 0$$

Lower bounds $\{L_n\}_{n=1}^{\infty}$

$$\inf_T \sup_P R(T, P) \geq L$$

Recall from 581: For any prior Π on $\mathcal{P}$, the Bayes risk of a decision rule T is given by

$$r(T, \Pi) = \int R(T, P)\, d\Pi(P)$$

Lemma: For any decision prob. with rule T,

$$\sup_P R(T, P) \geq \sup_\Pi r(T, \Pi)$$

Proof: $\forall \Pi$, sup over $\mathcal{P}$ is lower bounded by

$$\sup_P R(T, P) \geq r(T, \Pi)$$

Taking $\sup_\Pi$ gives the result.

Thm: For any decision prob.,

$$\inf_T \sup_P R(T,P) \geq \sup_\pi \inf_T r(T,\pi)$$

minimax risk — Bayes risk under LFP.

Proof: The max-min ineq. yields

$$\inf_T \sup_\pi r(T,\pi) \geq \sup_\pi \inf_T r(T,\pi)$$

Combine this w/ the preceding Lem.

$$\inf_T \sup_P R(T,P) \geq \sup_\pi \inf_T r(T,\pi)$$

□

Note: The above implies that, for any prior π,

$$\inf_T \sup_P R(T,P) \geq \inf_T r(T,\pi)$$

Method #1 for deriving minimax lower bounds:
Le Cam's method.

Thm (Le Cam): Let R be a risk func. defined according to some non-negative loss L.
For any $P_1, P_2 \in \mathcal{P}$,

$$\inf_T \sup_P R(T, P) \geq \frac{1}{2} d(P_1, P_2) \|P_1 \wedge P_2\|_1$$

$$\geq \frac{1}{4} d(P_1, P_2) \exp[-KL(P_1, P_2)]$$

where

"discrepancy" $d(P_1, P_2) = \inf_{a \in \mathbb{R}} [L(a, P_1) + L(a, P_2)]$

"testing affinity" $\|P_1 \wedge P_2\|_1 = \int \min\left\{\frac{dP_1}{d\nu}(w), \frac{dP_2}{d\nu}(w)\right\} d\nu(w)$

$$KL(P_1, P_2) = \begin{cases} \int \log\left[\frac{dP_1}{dP_2}(w)\right] dP_1(w) & P_1 \ll P_2 \\ +\infty & \text{o.w.} \end{cases}$$

Remark: Understanding discrepency.

We'll look at the discrepency in 2 cases.

a) Point estimation w/ squared error loss.

We want to estimate $\Psi(P) \in \mathbb{R}$ and the loss is $L(a, P) = \{a - \Psi(P)\}^2$

A little calculus shows that

$$d(P_1, P_2) = \inf_{a \in \mathbb{R}} [L(a, P_1) + L(a, P_2)]$$

$$= \frac{1}{2}\left[\Phi(P_1) - \Phi(P_2)\right]^2$$

b) Estimating a func w./ integrated squared error loss. Say $P = Q^n$, the action space is a convex subset of the space of funcs. mapping from $\mathcal{X}$ to $\mathbb{R}$, and

$$L(a, P) = \int [a(x) - f_a(x)]^2 \, d\nu(x)$$

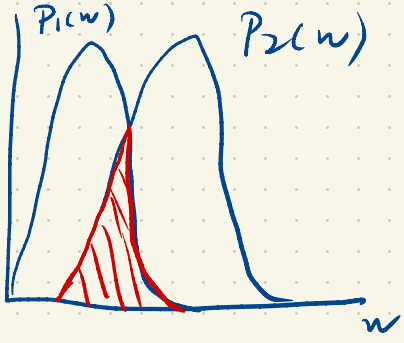
In this case,

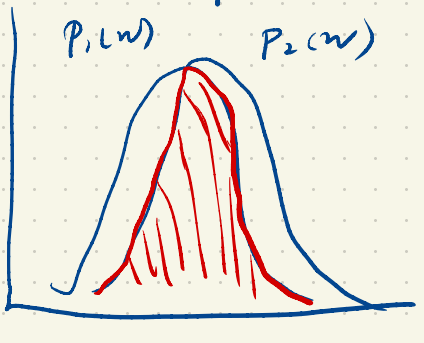
$$d(P_1, P_2) = \frac{1}{2}\int [f_{Q_1}(x) - f_{Q_2}(x)]^2 d\nu(x)$$

where $P_1 = Q_1^n$, $P_2 = Q_2^n$.

Remark: Understanding the testing affinity,

Statistically, we can show that given a draw from P_j, $j = \{1, 2\}$ is unknown, the difficulty of identifying the value of j is quantified by the testing affinity.





$$\|P_1 \wedge P_2\| = \int \min\{P_1, P_2\}\, d\nu$$

- Consider a test $T : \mathcal{W} \to \{1, 2\}$

We'll quantify performance of the decision rule T via misclassification error:

$$\mathcal{E}(T) = \mathbb{E}_{P_1}[\mathbb{1}\{T(w) = 2\}] + \mathbb{E}_{P_2}[\mathbb{1}\{T(w) = 1\}]$$

Observe that

$$\mathcal{E}(T) = \int [P_1(w)\,\mathbb{1}\{T(w) = 2\} + P_2(w)\,\mathbb{1}\{T(w) = 1\}]\, d\nu(w)$$

$$\geq \int [\min\{P_1(w), P_2(w)\} [\mathbb{1}\{T(w) = 2\} + \mathbb{1}\{T(w) = 1\}]]\, d\nu(w)$$

$$= \int \min(P_1, P_2)\, d\nu$$

$$= \|P_1 \wedge P_2\|_1$$

The lower bound is achieved by

$$T^*(w) = \operatorname*{argmax}_j P_j(w)$$

Remark: $\|P_1 \wedge P_2\|_1 = 1 - TV(P_1, P_2)$,

where $TV(P_1, P_2) = \sup_A |P_1(A) - P_2(A)|$

Proof of main Thm: (Le Cam)

Let π denote a uniform prior over $\{P_1, P_2\}$, that is $\pi(P_1) = \pi(P_2) = 1/2$

For any rule T,

$$r(T, \pi) = \sum_{j=1}^{2} \int L[T(w), P_j] P_j(w) \, d\nu(w) \pi(P_j)$$

$$= \frac{1}{2} \int \left[\sum_{j=1}^{2} L(T(w), P_j) P_j(w) \right] d\nu(w)$$

$$\geq \frac{1}{2} \int \min\{P_1(w), P_2(w)\} \left[\sum_{j=1}^{2} L(T(w), P_j) \right] d\nu(w)$$

$$\geq \frac{1}{2} \int \min\{P_1(w), P_2(w)\} \left[\inf_a \sum_{j=1}^{2} L(a, P_j) \right] d\nu(w)$$

$$= \frac{1}{2} d(P_1, P_2) \|P_1 \wedge P_2\|_1$$

Why we also need other lower bounding techniques?

- In certain cases, Le Cam will provide a rate-optimal lower bound
  - ↳ In fact, this is the case for prob. 2 HW1, which concerns estimation of a smooth density at a point x_0.
- In other cases, Le Cam can't give a rate-optimal LB.
  - ↳ This is the case when the goal is to estimate a smooth regression func in terms of the mean integrated squared error.

Thm (Fano's method):

Let R be a risk func. defined according to some non-negative loss L. For $N \geq 3$, $P_1, P_2, \ldots, P_N \in \mathcal{P}$ and define

$$\eta := \min_{j \neq k} d(P_j, P_k) = \min_{j \neq k} \inf_a \left[L(a, P_j) + L(a, P_k) \right]$$

$$\bar{P} = \frac{1}{N} \sum_{j=1}^{N} P_j \quad \text{(uniform mixture measure over } P_1, \ldots, P_N\text{)}$$

$$\inf_T \sup_P R(T, P) \geq \frac{\eta}{2} \left[1 - \frac{\log 2 + \frac{1}{N} \sum_{j=1}^{N} KL(P_j, \bar{P})}{\log N} \right]$$

$$\geq \frac{\eta}{2} \left[1 - \frac{\log 2 + \max_{j \neq k} KL(P_j, P_k)}{\log N} \right]$$

$$= \frac{\eta}{2} \left[1 - \frac{\log 2 + K}{\log N} \right],$$

where $K = \max_{j \neq k} KL(P_j, P_k)$

Remark: The bound increases when:

↳ N increases and η, K remain the same.

↳ K decreases and N, η remain the same

↳ η increases and N, K remain the same

Moreover, adding a dist. to a set $\{P_1, P_2, \dots, P_N\}$

- $N++$
- $\eta--$
- $K++$

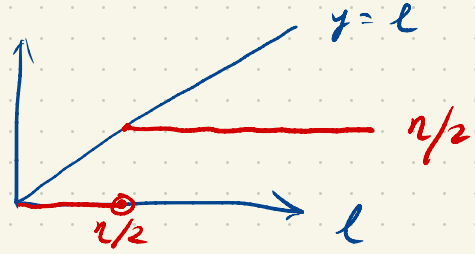
Proof: (Fano)

Let Π denote a uniform prior on $P_1, \dots, P_N$. For any rule T, we have that

$$r(T, \Pi) = \frac{1}{N} \sum_{j=1}^{N} \int L(T(w), P_j) p_j(w) \, d\nu(w)$$

$$(2) \qquad \geq \frac{\eta}{2N} \sum_{j=1}^{N} \int \mathbb{1}\{L(T(w), P_j) \geq \tfrac{\eta}{2}\} p_j(w) \, d\nu(w)$$

where we've used that $\ell \geq \frac{\eta}{2} \mathbb{1}\{\ell \geq \frac{\eta}{2}\}$ for all ℓ, y.



Continuing Eqn. (2),

$$(3)\quad r(T,\pi) \geq \frac{\eta}{2}\left[1 - \frac{1}{N}\int \sum_{j=1}^{N} \mathbb{1}\left\{L(T(w), P_j) < \frac{\eta}{2}\right\} P_j(w)\, dr(w)\right]$$

$$\leq \max_j P_j(w)$$

For each w,

$$\sum_{j=1}^{N} \mathbb{1}\left\{L(T(w), P_j) < \frac{\eta}{2}\right\} P_j(w)$$

$$\leq \left[\max_j P_j(w)\right] \sum_{j=1}^{N} \mathbb{1}\left\{L(T(w), P_j) < \frac{\eta}{2}\right\}$$

It suffices to show ≤ 1.

$$(\leq)\quad \left[\max_j P_j(w)\right] \sup_a \sum_{j=1}^{N} \mathbb{1}\left\{L(a, P_j) < \frac{\eta}{2}\right\}$$

NTS ≤ 1

To show

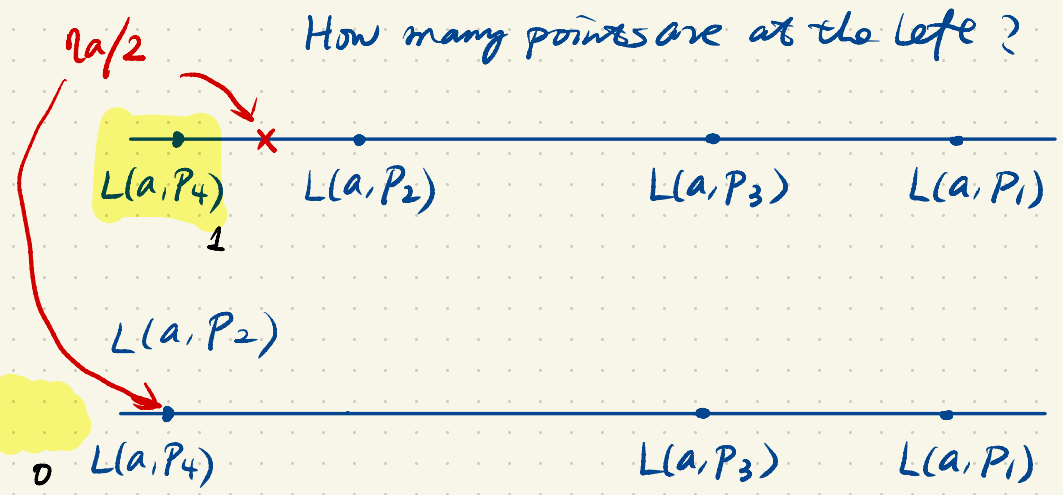
$$(\star)\quad \sup_a \sum_{j=1}^{N} \mathbb{1}\{L(a,P_j) < \tfrac{\eta}{2}\} \le 1.$$

it will be useful to define $\eta_a = \min[L(a,P_j) + L(a,P_k)]$

Note: $\eta = \inf_a \eta_a \quad \therefore \quad \eta \le \eta_a \ \forall a.$

Fix an action a. We'll show that

$$\sum_{j=1}^{N'} \mathbb{1}\{L(a,P_j) < \tfrac{\eta_a}{2}\} \le 1$$



Also, because $\eta \le \eta_a$,

$$\sum_{j=1}^{n} \mathbb{1}\{L(a,P_j) < \tfrac{\eta}{2}\} \le \sum_{j=1}^{n} \mathbb{1}\{L(a,P_j) < \tfrac{\eta_a}{2}\}$$

Taking a sup over all actions on the left

Plug into (3),

$$r(T,\pi) \geq \frac{\eta}{2}\left[1 - \frac{1}{N}\int \left[\max_j P_j(w)\right] dv(w)\right] \quad (4)$$

On slide 23, we show that

$$\max_j P_j(w) \leq \frac{N\bar{p}(w)\log 2 + \sum_j P_j(w)\log\frac{P_j(w)}{\bar{p}(w)}}{\log N}$$

$$\Rightarrow \int \max_j P_j(w)\, dv(w) \leq \frac{N\log 2 + \sum_{j=1}^{\tilde{N}} KL(P_j, \bar{P})}{\log N}$$

$$\Rightarrow \frac{1}{N}\int \left[\max_j P_j(w)\right] dv(w) \leq \frac{\log 2 + \frac{1}{N}\sum_{j=1}^{N} KL(P_j, \bar{P})}{\log N}$$

Plug in (4),

$$r(T,\pi) \geq \frac{\eta}{2}\left[1 - \frac{\log 2 + \frac{1}{N}\sum_{j=1}^{N} KL(P_j, \bar{P})}{\log N}\right]$$

Ex: Minimax lower bound for estimating a smooth regression func.

- We observe $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} Q \in \mathcal{Q}$
- For each Q, the marginal diso. of X is Unif$[0,1]$.
- Moreover,

$$Y \mid X = x \sim N(f_Q(x), 1)$$

where f_Q is smooth - in particular, $f_Q \in \mathcal{F}(\beta, L)$, where $\mathcal{F}(\beta, L)$ is the following Hölder class:

$$\mathcal{F}(\beta, L) := \{ f : f \text{ is } l\text{-times differentiable and } |f^{(l)}(x_1) - f^{(l)}(x_2)| \le L|x_1 - x_2|^{\beta - l} \text{ for all } x_1, x_2 \in [0,1] \}$$

where l is the greatest integer that is strictly less than β.

fixed nonneg. consts.

Generalized Lipschitz Class

Note: A suff. cond. for a func. f to belong to $\mathcal{F}(\beta, L)$ is that f be $(l+1)$-times differentiable and

$$\sup_x |f^{(l+1)}(x)| \le L$$

Our objective is to estimate

$$f_Q(x) = \mathbb{E}_Q[Y \mid X = x]$$

↳ Performance will be quantified by MISE, arising from the loss

$$L(a, \overbrace{P}^{=Q^n}) = \int_0^1 [a(x) - f_Q(x)]^2 \, dx$$

- We'll derive a minimax lower bound via Fano method.

↳ To do this, we must select a seq./collection $P_0 := \{P_1, \dots, P_M\}$ of dists. in our model.

↳ We're letting $P_j = Q_j^n$.

- For any $w \in \{0,1\}^m$, we define

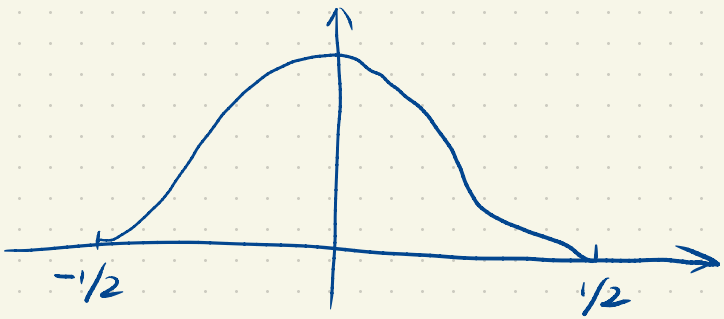
$$f_w(x) = \sum_{j=1}^{m} w_j \phi_j(x),$$

where here, for a fixed const. $h > 0$, we have $m \in [8, \frac{1}{h} - 1]$ and

$$\phi_j(x) = L h^\beta K\left(\frac{x - j/m+1}{h}\right)$$

Here, for suff. small $a > 0$, K is defined as

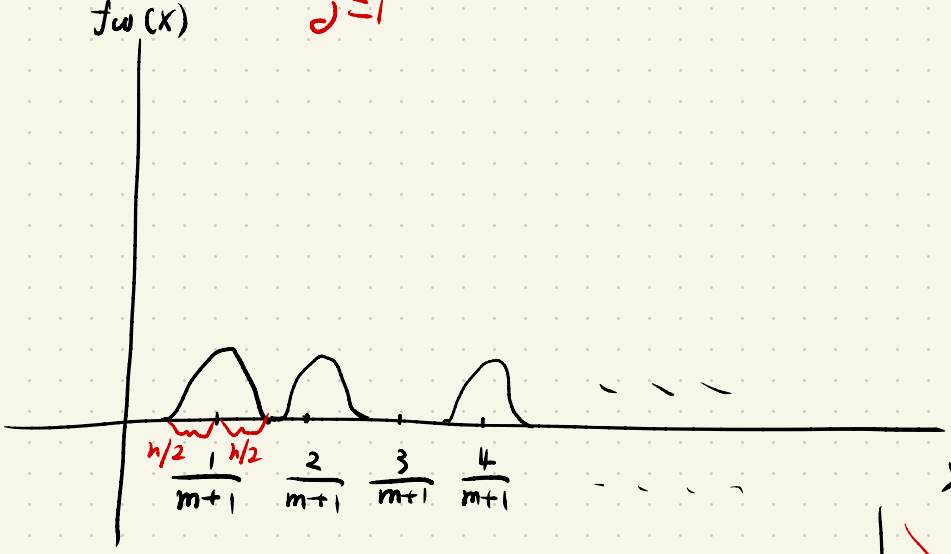
$$K(x) = a \exp\left(-\frac{1}{1-4x^2}\right) \mathbb{1}\{|x| < 1/2\}$$



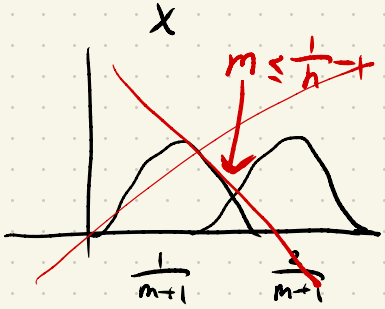
The quantity a is chosen that $f_w \in \mathcal{F}(\beta, L)$ for all $w \in [0,1]^m$.

Recall

$$f_w(x) = \sum_{j=1}^{m} w_j \phi_j(x)$$



$w_1 = 1$
$w_2 = 1$
$w_3 = 0$
$w_4 = 1$
⋮



If, for some $\tilde{\Omega} \subseteq \{0,1\}^m$, we applied Fano's method based on the collection of dists. $\{P_w := Q_w : w \in \tilde{\Omega}\}$, where $Q_w \in \mathcal{Q}$ has regression f_w, then we'd find that

$$\inf_T \sup_P R(T,P) \geq \frac{\min_{w \neq v \in \tilde{\Omega}} d(P_w, P_v)}{2} \left[1 - \frac{\log 2 + \max_{w \neq v} KL(P_w, P_v)}{\log |\tilde{\Omega}|}\right]$$

$$d(P_w, P_v) = \frac{1}{2}\int [f_w(x) - f_v(x)]^2 dx$$

Also, from slide 35,

$$KL(P_w, P_v) = \frac{n}{2}\int_0^1 [f_w(x) - f_v(x)]^2 dx$$

Since none bump func. overlap,

$$\phi_j(x)\phi_k(x) = 0 \quad \text{for } \forall x.$$

Only diagonal terms left.

$$\int [f_w(x) - f_v(x)]^2 dx$$

$$= \sum_{j=1}^{m} [w_j - v_j]^2 \int \phi_j(x)^2 dx$$

$$= \sum_{j=1}^{m} [w_j - v_j]^2 L^2 h^{2\beta+1} \underbrace{\int K(u)^2 du}_{c_2}$$

(bounded compact)

$$= c_2 L^2 h^{2\beta+1} \underbrace{\sum_{j=1}^{m} [w_j - v_j]^2}_{=: H(w,v)}$$

Hamming distance

Plug in and $c_3 = \frac{c_2 L^2}{2}$, we find

$$d(P_w, P_v) = c_3 h^{2\beta+1} H(w, v)$$

and

$$KL(P_w, P_v) = c_3 n h^{2\beta+1} H(w, v)$$

Also, since $H(w, v) \le m$, we have the further lower bound

$$KL(P_w, P_v) \le c_3 n h^{2\beta+1} m$$

$$\max_{w \neq v} KL(P_w, P_v) \le c_3 n h^{2\beta+1} m$$

Plug in above,

$$\inf_T \sup_P R(T,P) \geq \frac{c_3 h^{2\beta+1}}{2} \min_{w \neq v} H(w,v) \left[1 - \frac{\log 2 + c_3 n h^{2\beta+1} m}{\log |\tilde{\Omega}|} \right]$$

Following, think about two choices of the index set $\tilde{\Omega} \subseteq \{0,1\}^m$

Choice 1:

What if $\tilde{\Omega} = \{0,1\}^m$? (Entire thing)

In this case,

$$\min_{w \neq v} H(w,v) = 1$$

$$\log |\tilde{\Omega}| = m \log 2$$

Plug in, we have

$$\inf_T \sup_P R(T,P) \geq \frac{c_3 h^{2\beta+1}}{2} \left(1 - \frac{\log 2 + c_3 n h^{2\beta+1} m}{m \log 2} \right)$$

Two obs

1. If $n h^{2\beta+1}$ is large, then RHS is negative
   ↳ Not useful at all
   Hence, we need $h = O(n^{-\frac{1}{2\beta+1}})$

2. In the best case, the RHS is no more than $c_3 h^{2\beta+1}/2$

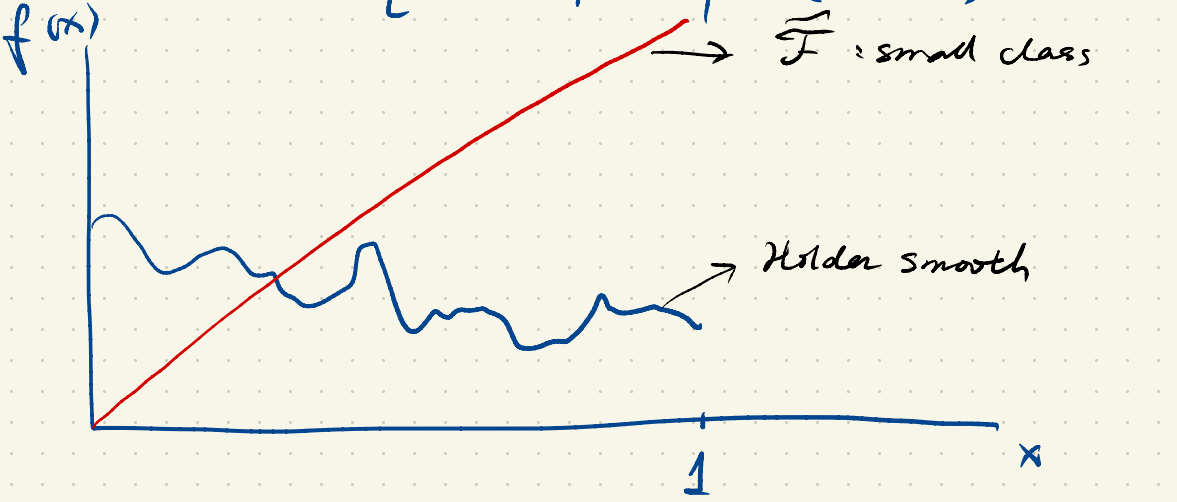
Combining these two, we see at best, the RHS would give a lower bound on order of $1/n$.

From 581, you might expect this lower bound is loose.

In particular, consider the parametric prob.

$f_\alpha \in \tilde{\mathcal{F}}$

where $\tilde{\mathcal{F}} := \{x \mapsto \beta x : \beta \in [0, L]\}$



From last quarter, you know MLE satisfies

$$\sqrt{n}(\hat{\beta} - \beta_0) \rightsquigarrow N(0, \theta^2)$$

Hence,

$$\frac{n(\hat{\beta} - \beta_0)^2}{\theta^2} \rightsquigarrow \chi^2(1)$$

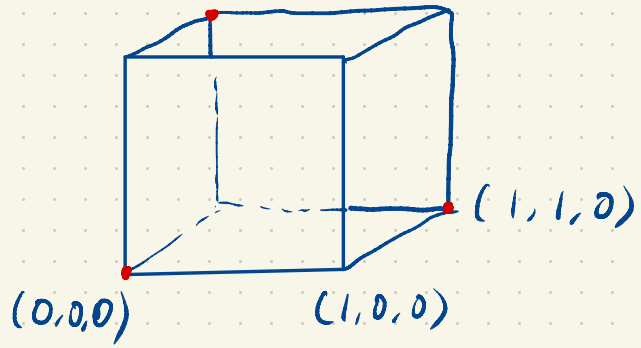
and $(\hat{\beta} - \beta_0)^2 = O_p(1/n)$

Hence, it's reasonable to expect that

$$\mathbb{E}_{\theta_0^n}\left[\int (\hat{\beta}_x - \beta_0 x)^2 dx\right] = O(1/n)$$

Choice 2:

Question: Is there a subset $\tilde{\Omega}$ of $\{0,1\}^m$ for which $|\tilde{\Omega}|$ large and Hamming distance of distinct elements is also large?



Answer: Yes!

Varshamov-Gilbert Lemma:

If $m \geq 8$, then there exists a subset Ω of $\{0,1\}^m$ s.t. $|\Omega|$ is large $\geq 2^{m/8}$ and

$$\min_{w \neq v \in \Omega} H(w,v) \geq m/8$$

Choice 2: $\widehat{\Omega} = \Omega$.

In this case, our lower bound gives that

$$\inf_T \sup_P R(T,P) \geq \frac{c_3 h^{2\beta+1} m}{16}\left(1 - \frac{\log 2 + c_3 n h^{2\beta+1} m}{m \log 2/8}\right)$$

$$= \frac{c_3 h^{2\beta+1} m}{16}\left(1 - \frac{8}{m} - \frac{8 c_3 n h^{2\beta+1}}{\log 2}\right)$$

Choose largest m possible.

$\left(m = \left\lfloor \frac{1}{h} - 1 \right\rfloor\right)$ We'll approximate this lower bound $m = 1/h$ yielding

$$LB \approx \frac{c_3 h^{2\beta}}{16}\left(1 - 8h - \frac{8 c_3 n h^{2\beta+1}}{\log 2}\right) \quad (\star)$$

- $(\star)$ non-negative?

We now choose h. For $(\star) \geq 0$,

We need $h \leq c_4 n^{-\frac{1}{2\beta+1}}$ for some c_4.

Letting $h = \frac{c_4 n^{-\frac{1}{2\beta+1}}}{2}$, shows that there exists const c_5 s.t.

$$\inf_T \sup_P R(T,P) \geq c_5 n^{\frac{-2\beta}{2\beta+1}}$$

For example, if $\beta = 2$, we derived $n^{-4/5}$ rate lower bound on the minimax risk.

This bound is tight (in terms of rates)!