MML ⑨

Chapter 2 = Optimality in Statistic : Minimax lower bounds

Intro :

Chapter 1, we prove that there is a statistical method (a kernel density estimator) such that $\forall$ f density on $[0,1]$, twice differentiable,

$$\mathbb{E}\left[\int_0^1 \left(\hat{f}(x) - f(x)\right)^2 dx\right] = O(n^{-4/5})$$

($\hat{f}$: our estimator, f : truth)

Now, we will prove that, in a sense, we can't do better. (therefore, $\hat{f}$ is optimal)

I) Useful tools from Information Theory

① Some definitions :

- Let P and Q be two probability measures on the same measurable space $(\Omega, \mathcal{F})$.
We say that "P is absolutely continuous with respect to Q" (We write $P \ll Q$) iff
$\forall A \in \mathcal{F}$, $Q(A) = 0 \Rightarrow P(A) = 0$

(w.r.t.)

Radon-Nikodym's theorem :
If $P \ll Q$, then there exists a non negative function $f : \Omega \to \mathbb{R}_+$ such that
$\forall A \in \mathcal{F}$, $P(A) = \int_A f \, dQ$

Convention: $0 \cdot \log(0) = 0$ s.t. = such that

Furthermore, f is unique up to Q a.s. equality.

f is called "the Radon Nikodym derivative of P w.r.t. Q" or "the density of P w.r.t. Q"

We write: $f = \frac{dP}{dQ}$

One important situation:

Assume $\exists \mu$ positive measure s.t.

$$\begin{cases} P \ll \mu \\ Q \ll \mu \end{cases} \longrightarrow f = \frac{dP}{d\mu} \text{ and } g = \frac{dQ}{d\mu}$$

If $P \ll Q$, then $\frac{dP}{dQ} = \frac{f}{g}$

$$\left(\frac{dP}{dQ} = \frac{dP/d\mu}{dQ/d\mu} = \frac{f}{g} \right)$$

- the proof is easy exercise

② Divergences

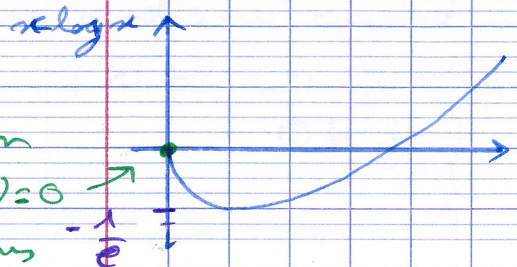
- **Kullback - Leibler divergence** (also called "relative entropy")

If P and Q are two proba. on $(\Omega, \mathcal{F})$

$$KL(P, Q) = \begin{cases} \int_\Omega \frac{dP}{dQ} \log\left(\frac{dP}{dQ}\right) dQ & \text{if } P \ll Q \\ +\infty & \text{otherwise} \end{cases}$$

↑ the KL divergence between P and Q

N.B.: Why is it well defined?



$$\frac{d}{dx}(x \log(x)) = \log x + 1$$

MML 10

$f_- = \max(-f(x), 0)$

$f = \frac{dP}{dQ}$ if $P \ll Q$

positive part / negative part

$$\int_\Omega f \log f \, dQ \overset{\text{def}}{=} \int_\Omega (f \log f)_+ \, dQ - \int_\Omega (f \log f)_- \, dQ$$

→ always well defined in $\mathbb{R} \cup \{+\infty\}$

is always finite because $(f \log f)_- \leq \frac{1}{e}$

Proposition: $KL(P, Q) \geq 0$

Proof: the function $x \in [0, +\infty[\mapsto \underbrace{x \log x}_{\varphi(x)}$ is convex so that by Jensen's inequality

$$\int f \log f \, dQ = \int \varphi(f) \, dQ \geq \varphi\left(\underbrace{\int f \, dQ}_{=1}\right) = 0$$

Exercise $KL(P, Q) = 0 \Leftrightarrow P = Q$

How to compute it in pratice?

If μ s.t. $P, Q \ll \mu$ $f = \frac{dP}{d\mu}$, $g = \frac{dQ}{d\mu}$

and $P \ll Q$, then $\frac{dP}{dQ} = \frac{f}{g}$

so $KL(P, Q) = \int \frac{f}{g} \log\left(\frac{f}{g}\right) g \, d\mu$

Correct if $g(w) > 0 \; \forall w \in \Omega$ but also makes sense more generally provided $P \ll Q$

Esc: $(\Omega, \mathcal{F}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$

$P = \mathcal{N}(m_1, \sigma^2)$ $Q = \mathcal{N}(m_2, \sigma^2)$

ν Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$

(N.B. $P \ll Q$ $Q(A) = \int_A g \, d\nu = 0 \Rightarrow \nu(A) = 0 \Rightarrow P(A) = 0$)

$P, Q \ll \nu$ with densities:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m_1)^2}{2\sigma^2}\right)$$

$$g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m_2)^2}{2\sigma^2}\right)$$

$$KL(P,Q) = \int_{-\infty}^{+\infty} \frac{f(x)}{g(x)} \log\left(\frac{f(x)}{g(x)}\right) g(x)\,dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{(x-m_1)^2}{2\sigma^2}\right) \log\left(\exp\left(-\frac{(x-m_1)^2 - (x-m_2)^2}{2\sigma^2}\right)\right) dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{(x-m_1)^2}{2\sigma^2}\right) \frac{(x-m_2)^2 - (x-m_1)^2}{2\sigma^2}\, dx$$

$$(m_1-m_2)^2 + 2(x_1-m_1)(m_1-m_2)$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{(x-m_1)^2}{2\sigma^2}\right) \frac{(m_1-m_2)^2}{2\sigma^2}\, dx$$

$$+ \frac{2m_1 - m_2}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{(x-m_1)^2}{2\sigma^2}\right) \frac{(x_1-m_1)}{2\sigma^2}\, dx$$

$$= \mathbb{E}(X - \mathbb{E}(X)) = 0 \quad \text{when } X \sim P$$

Conclusion:

$$KL(\mathcal{N}(m_1, \sigma^2), \mathcal{N}(m_2, \sigma^2)) = \frac{(m_1-m_2)^2}{2\sigma^2}$$

Here $KL(P,Q)$ is large if m_1 is far from m_2. → sort of distance

by ♡

MML (11)

In general $(P, Q) \mapsto KL(P, Q)$ is not a distance (it can even be non-symmetric, or it may not verify the triangle inequality). BUT we should think of it as a rough notion of distance → called a "divergence"

Ex 2: $(\Omega, \mathcal{F}) = (\{0,1\}, \mathcal{P}(\{0,1\}))$, $\mathcal{P}(\{0,1\}) = 2^{\{0,1\}}$

$P = Ber(p)$, $p \in [0,1]$

$Q = Ber(q)$, $q \in [0,1]$

ν: the counting measure $= \delta_0 + \delta_1$

$$\frac{dP}{d\nu}(x) = f(x) = \begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$$

$$g(x) = \frac{dQ}{d\nu}(x) = \begin{cases} 1-q & \text{if } x = 0 \\ q & \text{if } x = 1 \end{cases}$$

Therefore

$$KL(P, Q) = \int_\Omega f \log \frac{f}{g} \, d\nu \qquad kl(p,q)$$

by ♡

$$KL(Ber(p), Ber(q)) = p \log \frac{p}{q} + (1-p) \log \left(\frac{1-p}{1-q} \right)$$

true for all $p, q \in [0,1]$ except maybe when ($q = 0$ and $p > 0$) or ($q = 1$ and $p < 1$) in which case the $KL = +\infty$ but the formula is also correct because it's also $= +\infty$

Fix $q \in (0,1)$ How does $p \mapsto kl(p,q)$ looks like?

$$\frac{\partial}{\partial p} kl(p,q) = \log \left(\frac{p}{q} \right) + 1 - \left(\log \left(\frac{1-p}{1-q} \right) + 1 \right)$$

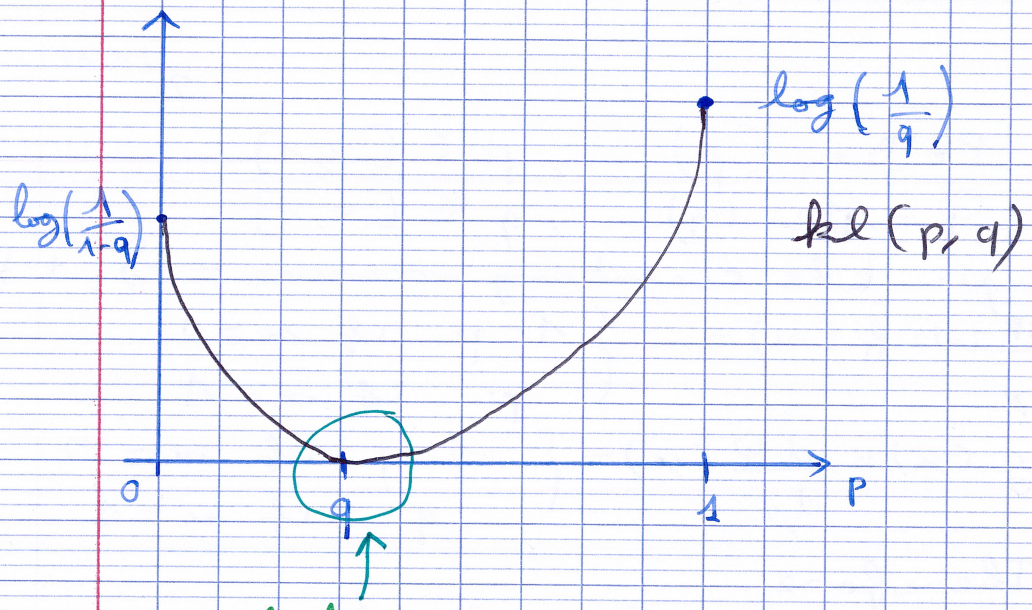
$$= \log \left(\frac{p/(1-p)}{q/(1-q)} \right)$$

$$\begin{cases} = +\infty \text{ when } p \to 1 \\ = -\infty \text{ when } p \to 0 \\ = 0 \text{ at } p = q \end{cases}$$

In particular derivative

$$\frac{\partial^2}{\partial p^2} kl(p,q) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)} \quad : \text{fct convex}$$

so that $\frac{\partial^2}{\partial p^2} kl(p,q)\Big|_{p=q} = \frac{1}{q(1-q)}$



looks like $\frac{(p-q)^2}{2q(1-q)}$ as $p \to q$ by Taylor expansion

locally quadratic

Exercise: Fix $p \in (0,1)$, plot again q

(3) Useful properties of KL

- **Prop: Joint convexity of KL**

  Let $(\Omega, \mathcal{F})$ be any measurable space

  Then, the function

  $(P, Q) \in \mathcal{M}_1^+(\Omega, \mathcal{F})^2 \mapsto KL(P, Q) \in [0, +\infty]$

  is convex i.e. $\forall (P_1, P_2), (Q_1, Q_2), \lambda \in (0,1)$

$\mathcal{M}_1^+(\Omega)$ = set of probability measure

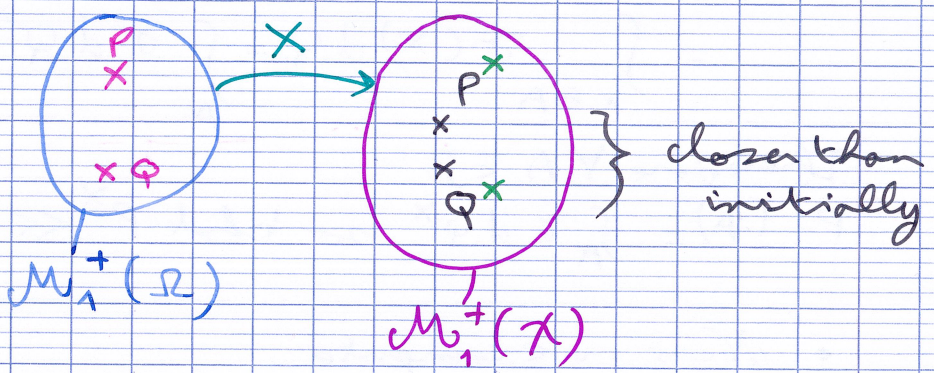
MML 12

$$KL((1-\lambda)(P_1, Q_1) + \lambda(P_2, Q_2)) \leq (1-\lambda) KL(P_1, Q_1) + \lambda KL(P_2, Q_2)$$

Proof: exercise!

- **Data processing inequality** ("contraction of entropy")

- **Theorem:** $(\Omega, \mathcal{F})$, $(\mathcal{X}, \mathcal{B})$ 2 m. spaces. $X: \Omega \to \mathcal{X}$ a random variable. Then $KL(P^X, Q^X) \leq KL(P, Q)$ where P^X and Q^X are the distributions or laws of X w.r.t. P (also "induced" or "push-forward" measures) $P^X(A) = P(X \in A)$



if you transform your data, you make your statistic more difficult, you "loose" info, more difficult to distinguish the 2 distrib

Proof:

if $KL(P, Q) = +\infty$, then obvious

So we can assume that $KL(P, Q) < +\infty$ therefore $P \ll Q$

- Now, we show that $P^X \ll Q^X$ and $\frac{dP^X}{dQ^X}(x) = \gamma(x)$

where $\gamma(x) = \mathbb{E}_Q\left[\frac{dP}{dQ} \mid X = x\right]$

this can be probability $= 0$

i.e. γ is the Q-a.s. unique function s.t.

$$\mathbb{E}_Q\left[\frac{dP}{dQ} \mid X\right] = \gamma(X) \quad Q\text{-a.s.}$$

Indeed $\forall A \in \mathcal{B}$,

$$P^X(A) = P(X \in A) = \int_{X \in A} \frac{dP}{dQ}\, dQ$$

$$= \mathbb{E}_Q\left[\frac{dP}{dQ}\, \mathbb{1}_{X \in A}\right] \quad \text{(total expectation)}$$

$$= \mathbb{E}_Q\left[\mathbb{E}_Q\left[\frac{dP}{dQ}\, \mathbb{1}_{X \in A} \mid X\right]\right]$$

$$= \mathbb{E}_Q\left[\mathbb{1}_{X \in A}\, \gamma(X)\right]$$

$$= \int_{X \in A} \gamma(X)\, dQ = \int_A \gamma\, dQ^X$$

Therefore $\forall A \in \mathcal{B}$

$P^X(A) = \int_A \gamma\, dQ^X$ so that $P^X \ll Q^X$ and $\frac{dP^X}{dQ^X} = \gamma$

- Now, we compute

$$KL(P^X, Q^X) = \int_\Lambda \frac{dP^X}{dQ^X} \log\left(\frac{dP^X}{dQ^X}\right) dQ^X$$

$$= \int_\Lambda \gamma \log \gamma\, dQ^X$$

$$= \int_\Omega \underbrace{\gamma(X) \log(\gamma(X))}_{\varphi(X) \text{ where } \varphi(t) = t \log t}\, dQ$$

$$= \mathbb{E}_Q\left[\varphi\left(\underbrace{\mathbb{E}_Q\left[\frac{dP}{dQ} \mid X\right]}_{\gamma(X)}\right)\right]$$

MML (13)

A_i events: measurable
partition: $\bigcup A_i = \Omega$
$A_i \cap A_j = \emptyset_{i \neq j}$

$$\leq \mathbb{E}_Q\left[\mathbb{E}_Q\left[\varphi\left(\frac{dP}{dQ}\right) \mid \mathcal{X}\right]\right]$$

by conditional Jensen's inequality

$$= \mathbb{E}_Q\left[\varphi\left(\frac{dP}{dQ}\right)\right] = KL(P,Q) \quad \checkmark$$

④ Pinsker's and Fano's inequalities

Theorem: Pinsker's inequality

Let $P, Q \in \mathcal{M}_1^+(\Omega)$

Then $\|P - Q\|_{TV} \leq \sqrt{\frac{KL(P,Q)}{2}}$

where $\|P - Q\|_{TV} = \sup_{A \in \mathcal{A}} |P(A) - Q(A)|$

"the total variation distance between P and Q" (this is a distance)

exercise

Theorem: Fano's inequality

Let $P_1, P_2, \dots P_N \in \mathcal{M}_1^+(\Omega)$, $Q_1 \dots Q_n \in \mathcal{M}_1^+(\Omega)$, $N \geq 2$

$A_1, \dots A_N$ (events in Ω

Then

$$\frac{1}{N}\sum_{i=1}^{N} P_i(A_i) \leq \frac{\frac{1}{N}\sum_{i=1}^{N} KL(P_i, Q_i) + \log 2}{\log\left(\frac{1}{\frac{1}{N}\sum_{i=1}^{N} Q_i(A_i)}\right)}$$

In particular, if $Q_i = Q$ (all equal) and if A_i form a partition of Ω

$$\frac{1}{N}\sum_{i=1}^{N} P_i(A_i) \leq \frac{\frac{1}{N}\sum_{i=1}^{N} KL(P_i, Q) + \log 2}{\log N}$$

• Proof of Pinsker's inequality:

• Let $A \in \mathcal{F}$: Next we prove that

$$|P(A) - Q(A)| \leq \sqrt{\frac{KL(P,Q)}{2}}$$

We set $p = P(A)$ and $q = Q(A)$

By the data processing inequality

$$KL\left(P^{\mathbb{1}_A}, Q^{\mathbb{1}_A}\right) \leq KL(P,Q)$$

$(X = \mathbb{1}_A : \Omega \to \{0,1\})$: $P^{\mathbb{1}_A} = Ber(P(A)) = Ber(p)$

therefore, $kl(p,q) \leq KL(P,Q)$

To conclude the proof we only need to show that $2(p-q)^2 \leq kl(p,q)$

Function study:

$$\frac{\partial}{\partial p} kl(p,q)\Big|_{p=q} = 0 \qquad kl(q,q) = 0$$

$$\frac{\partial^2}{\partial p^2} kl(p,q) = \frac{1}{p(1-p)}$$

Taylor expansion : $\exists \hat{p}$ between p and q s.t.

$$kl(p,q) = kl(q,q) + \frac{\partial}{\partial p} kl(p,q)\Big|_{p=q} \times (p-q) + \frac{1}{2} \frac{\partial^2}{\partial p^2} kl(p,q)\Big|_{p=\hat{p}} \times (p-q)^2$$

Taylor Lagrange

To be rigorous we need to check p or $q = 1, 0$ case

Need C^2 : true because $p \mapsto kl(p,q) \in C^\infty(]0,1[)$

so $kl(p,q) = \frac{(p-q)^2}{2} \underbrace{\frac{1}{\hat{p}(1-\hat{p})}}_{\geq 4} \geq 2(p-q)^2$ ■

• Proof of Fano's inequality:

Let $p_i = P_i(A_i)$ and $q_i = Q_i(A_i)$

again by data processing inequality

MML

14

$$\underbrace{KL(P_i^{\mathbb{1}_{A_i}}, Q_i^{\mathbb{1}_{A_i}})}_{= kl(p_i, q_i)} \leq KL(P_i, Q_i)$$

Besides, by convexity of $(p,q) \in [0,1] \mapsto kl(p,q) \in [0,+\infty]$ we have that

$$kl\Big(\underbrace{\frac{1}{N}\sum_{i=1}^N p_i, \frac{1}{N}\sum_{i=1}^N q_i}_{\frac{1}{N}\sum_{i=1}^N (p_i, q_i)}\Big) \leq \frac{1}{N}\sum_{i=1}^N kl(p_i, q_i)$$

Combining the two bounds above we get

$$kl\left(\frac{1}{N}\sum_{i=1}^N p_i, \frac{1}{N}\sum_{i=1}^N q_i\right) \leq \frac{1}{N}\sum_{i=1}^N KL(P_i, Q_i) \quad (*)$$

Now we set $\bar{p} = \frac{1}{N}\sum_{i=1}^N p_i$ and $\bar{q} = \frac{1}{N}\sum_{i=1}^N q_i$

To conclude the proof, we show that $\bar{p} \leq \frac{\overline{KL} + \log 2}{\log(1/\bar{q})}$ with $\overline{KL} = \frac{1}{N}\sum_{i=1}^N KL(P_i, q_i)$

But $(*)$ yields $\overline{KL} \geq kl(\bar{p}, \bar{q})$

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

$$= \underbrace{\bar{p}\log\bar{p} + (1-\bar{p})\log(1-\bar{p})}_{\geq -\log 2} + \bar{p}\log\frac{1}{\bar{q}} + \underbrace{(1-\bar{p})\log\left(\frac{1}{1-\bar{q}}\right)}_{\geq 0}$$

(exercise: $\forall x \in [0,1]$, $x\log x + (1-x)\log(1-x) \geq -\log 2$)

$$\Rightarrow \overline{KL} \geq \bar{p}\log\frac{1}{\bar{q}} - \log 2$$

$$\text{i.e. } \bar{p} \leq \frac{\overline{KL} + \log 2}{\log(1/\bar{q})}$$

2 very useful inequalities to prove lower bounds in statistics.

II) A gentle start: a minimax lower bound in dimension 1

Model 1 We observe $X_1, \dots X_n \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2)$ where $\theta \in \mathbb{R}$ is unknown

Upper bound on the MSE: Can we find $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ such that $\mathbb{E}_\theta[(\hat{\theta} - \theta)^2]$ is small?

Yes: $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i$ satisfies (bias $= 0$)

$$\forall \theta \in \mathbb{R}, \quad \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] = \mathrm{Var}_\theta(\hat{\theta}) = \frac{\mathrm{Var}_\theta(X_1)}{n} = \frac{\sigma^2}{n}$$

Therefore $\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \leq \frac{\sigma^2}{n}$

- $\inf_{\hat{\theta}}$: infimum over all possible estimator
- $\sup_{\theta \in \mathbb{R}}$: worst case for estimator $\hat{\theta}$

"minimax risk" in model 1

Corresponds to the best possible worst case performance

confidence interval: $\left[\hat{\theta} - q\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\theta} + q\frac{\hat{\sigma}}{\sqrt{n}}\right]$

Q: is this bound improvable?

A: no. We will prove:

Theorem: If $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2)$, $\theta \in \mathbb{R}$

$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \geq \frac{c\sigma^2}{n}$ for some absolute $c > 0$

Game theory

$\inf_{\hat{\theta}}$: statistician

$\sup_{\theta}$: adversary

We will prove that $c = 1/4$ works

Note: in this case $c = 1$ is the optimal cst.

Hint: think about the bayesian risk

Proof: Idea: cf confidence intervals $\left[\hat{\theta} - q\frac{\hat{\sigma}}{\sqrt{n}}, \hat{\theta} + q\frac{\hat{\sigma}}{\sqrt{n}}\right]$ the precision is $\approx 1/\sqrt{n}$

so the statistician cannot distinguish between θ and θ' if $|\theta - \theta'| \lesssim 1/\sqrt{n}$. In this case the error is $(\hat{\theta} - \theta)^2 \gtrsim \left(\frac{1}{\sqrt{n}}\right)^2 = \frac{1}{n}$

- More formally:

Let $\hat{\theta}$ be any estimator, we show next that

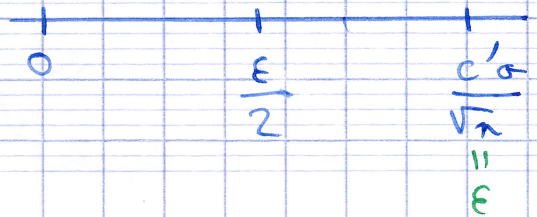
$\sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \geq c\frac{\sigma}{\sqrt{n}}$

To that end, note that

$\sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \geq \max_{\theta \in \{0, \frac{c'\sigma}{\sqrt{n}}\}} \mathbb{E}_\theta[(\hat{\theta} - \theta)^2]$

$\frac{c'\sigma}{\sqrt{n}} = \epsilon$

c' to be chosen later



projection of $\hat{\theta}$ on $\{0, \varepsilon\}$
point in $\{0, \varepsilon\}$ closer to $\hat{\theta}$

We set $\tilde{\theta} \in \underset{\theta \in \{0, \varepsilon\}}{\operatorname{argmin}} |\hat{\theta} - \theta|$

Remark: if $\tilde{\theta} \neq 0$, then $|\hat{\theta} - \theta| \geq \frac{\varepsilon}{2}$
$\theta \in \{0, \varepsilon\}$

Therefore $\forall \theta \in \{0, \varepsilon\}$

$$\mathbb{E}_\theta[(\tilde{\theta} - \theta)^2] \geq \mathbb{E}_\theta[(\tilde{\theta} - \theta)^2 \mathbb{1}_{\hat{\theta} \neq 0}]$$
$$\geq \left(\frac{\varepsilon}{2}\right)^2 \mathbb{P}_\theta(\hat{\theta} \neq \theta)$$
$$= \frac{\varepsilon^2}{4}\left(1 - \mathbb{P}_\theta(\tilde{\theta} = \theta)\right)$$

Thus, $\max_{\theta \in \{0, \varepsilon\}} \mathbb{E}_\theta[(\tilde{\theta} - \theta)^2] \geq \frac{\varepsilon^2}{4}\left(1 - \min_{\theta \in \{0, \varepsilon\}} \mathbb{P}_\theta(\tilde{\theta} = \theta)\right)$

Important message: We have reduce the estimation problem to a testing problem with 2 hypothesis:
$\begin{cases} H_0 : \theta = 0 \\ H_1 : \theta = \varepsilon \end{cases}$

Next we prove that $\min_{\theta \in \{0, \varepsilon\}} \mathbb{P}_\theta(\tilde{\theta} = \theta)$ is small

- $\theta \in \{0, \varepsilon\}$: worst case
- $\mathbb{P}_\theta(\tilde{\theta} = \theta)$: probability of identifying θ

In other words, we will show that, whatever $\hat{\theta}$ and thus $\tilde{\theta}$, $\tilde{\theta}$ must fail to identify $\theta = 0$ or $\theta = \varepsilon$

Let us prove it: by Pinsker inequality

$\forall \theta \in \{0, \varepsilon\}$, $|\mathbb{P}_\theta(\tilde{\theta} = \theta) - \mathbb{P}_0(\tilde{\theta} = \theta)| \leq \|\mathbb{P}_\theta - \mathbb{P}_0\|_{TV}$

$$\leq \sqrt{\frac{KL(\mathbb{P}_\theta, \mathbb{P}_0)}{2}} \quad \text{(Pinsker)}$$

MML 16

where $\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2)^{\otimes n}$

$\mathbb{P}_0 = \mathcal{N}(0, \sigma^2)^{\otimes n}$

$KL(\mathbb{P}_\theta, \mathbb{P}_0) = KL(\mathcal{N}(\theta, \sigma^2)^{\otimes n}, \mathcal{N}(0, \sigma^2)^{\otimes n})$

$= n\, KL(\mathcal{N}(\theta, \sigma^2), \mathcal{N}(0, \sigma^2))$

because:

Lemma:

$$KL\left(\bigotimes_{i=1}^n P_i, \bigotimes_{i=1}^n Q_i\right) = \sum_{i=1}^n KL(P_i, Q_i)$$

So $KL(\mathbb{P}_\theta, \mathbb{P}_0) = \frac{n(\theta - 0)^2}{2\sigma^2}$

Combining with (2)

$$\forall \theta \in \{0, \varepsilon\}, \quad \mathbb{P}_\theta(\hat{\theta} = \theta) \leq \mathbb{P}_0(\hat{\theta} = \theta) + \sqrt{\frac{n\varepsilon^2}{2\sigma^2}\frac{1}{2}}$$

We take the average over $\theta \in \{0, \varepsilon\}$

$$\frac{1}{2}\sum_{\theta \in \{0,\varepsilon\}} \mathbb{P}_\theta(\hat{\theta} = \theta) \leq \frac{1}{2}\underbrace{\sum_{\theta \in \{0,\varepsilon\}} \mathbb{P}_0(\hat{\theta} = \theta)}_{=1} + \sqrt{\frac{n\varepsilon^2}{4 \cdot 2\sigma^2}}$$

Finally,

$$\max_{\theta \in \{0,\varepsilon\}} \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \geq \frac{\varepsilon^4}{4}\left(1 - \frac{1}{2} - \sqrt{\frac{n\varepsilon^2}{4\sigma^2}}\right)$$

Now we take $\varepsilon = \frac{c'\sigma}{\sqrt{n}}$

$$\geq \frac{c'^2\sigma^2}{4n}\left(\frac{1}{2} - \sqrt{\frac{c'}{4}}\right)$$

$$\geq \frac{\sigma^2}{256\, n} \quad \text{with } c' = 1/4$$

absolute cst : i.e. doesn't depend on θ

III) Another parametric minimax lower bound, in higher dimensions

Model 2: $X_1, \dots X_n \overset{iid}{\sim} \mathcal{N}(\theta, \sigma^2 I_d)$

$\theta \in \mathbb{R}^d$ unknown

Upper bound?

Yes: $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i$ satisfies

$$\mathbb{E}_\theta\left[\|\hat{\theta} - \theta\|_2^2\right] = \frac{d\sigma^2}{n}$$

so $\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_\theta\left[\|\hat{\theta} - \theta\|_2^2\right] \leq \frac{d\sigma^2}{n}$

We prove this is optimal up to a constant factor

Theorem: In model 2,

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_\theta\left[\|\hat{\theta} - \theta\|_2^2\right] \geq \frac{c d \sigma^2}{n}$$

for some absolute constant $c > 0$

Proof: Same idea as in dimension $d = 1$ but we use Fano's inequality which is key to get the d factor in the lower bound

Let $\hat{\theta}$ be any estimator, $\varepsilon > 0$, $(\varepsilon \asymp \sigma/\sqrt{n})$

$$\sup_{\theta \in \mathbb{R}^d} \mathbb{E}_\theta\left[\|\hat{\theta} - \theta\|_2^2\right] \geq \max_{\theta \in (H)} \mathbb{E}_\theta\left[\|\hat{\theta} - \theta\|_2^2\right]$$

Define the projection $\widehat{\theta} \in \operatorname{argmin}_{\theta \in (H)} \|\hat{\theta} - \theta\|_2$

where $(H) \subseteq \{0, \varepsilon\}^d$

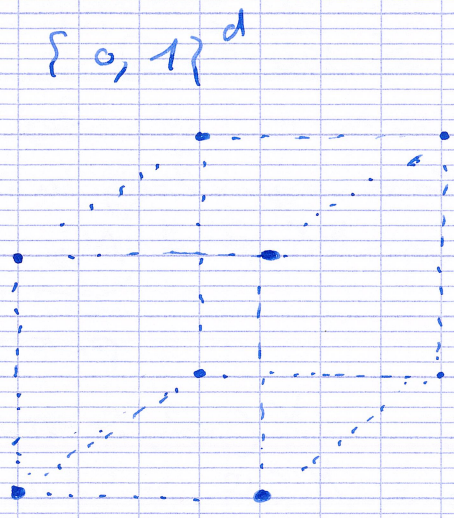
MML (17)

Lemma: Varshamov - Gilbert's lemma

Let $d \geq 6$, there exists $\Gamma \subseteq \{0,1\}^d$ such that

- $|\Gamma| \geq e^{d/8}$
- $\forall x \neq y \in \Gamma, \underbrace{\sum_{i=1}^{d} \mathbb{1}_{\{x_i \neq y_i\}}}_{\text{number of } \neq \text{ coordinates}} > \frac{d}{4}$

$|\Gamma| = \#\Gamma$



Proof: It uses a probabilistic argument! (Hoeffding's inequality) → cf exercise

We set $\mathcal{H} = \{\varepsilon\Gamma \subseteq \{0,\varepsilon\}^d\}$

and $|\mathcal{H}| = |\Gamma| = e^{d/8}$

$$\forall x, y \in \mathcal{H} \quad \|x-y\|_2^2 = \sum_{i=1}^{d} \underbrace{(x_i - y_i)^2}_{= \varepsilon \mathbb{1}_{x_i \neq y_i}} > \varepsilon^2 \frac{d}{4}$$

From now on, we assume $d \geq 6$

for d between 2 and 5 easy because $d \mapsto \inf \sup(-)$ is increasing.

$$\max_{\Theta \in \mathcal{H}} \mathbb{E}_\Theta\left[\|\hat{\Theta} - \Theta\|_2^2\right] \geq \max_{\Theta \in \mathcal{H}} \mathbb{E}_\Theta\left[\|\hat{\Theta} - \Theta\|_2^2 \, \mathbb{1}_{\hat{\Theta} \neq \Theta}\right]$$

where $\hat{\Theta}$ def before

Remark: if $\Theta \in \mathcal{H}$ and $\tilde{\Theta} \neq \Theta$ then $\|\tilde{\Theta} - \Theta\|_2 \geq \frac{1}{2}\sqrt{\frac{\varepsilon^2 d}{4}}$ (use the triangular inequality)

$$\max_{\theta \in \Theta} \mathbb{E}_\theta\left[\|\hat{\theta} - \theta\|_2^2\right] \geq \max_{\theta \in \Theta} \mathbb{E}_\theta\left[\left(\frac{1}{2}\sqrt{\frac{\varepsilon^2 d}{4}}\right)^2 \mathbb{1}_{\hat{\theta} \neq \theta}\right]$$

$$= \frac{d\varepsilon^2}{16}\left(1 - \min_{\theta \in \Theta} \mathbb{P}_\theta(\hat{\theta} = \theta)\right)$$

We finish the proof by upper bounding.

$$\min_{\theta \in \Theta} \mathbb{P}_\theta(\hat{\theta} = \theta) \leq \frac{1}{|\Theta|} \sum_{\theta \in \Theta} \mathbb{P}_\theta(\hat{\theta} = \theta)$$

$$\leq \frac{\frac{1}{|\Theta|}\sum_{\theta \in \Theta} KL(\mathbb{P}_\theta, \mathbb{Q}) + \log 2}{\log |\Theta|} \quad (3)$$

because of Fano's inequality used for any distribution $\mathbb{Q}$ on $\mathbb{R}^n$ and $\underbrace{A_\theta = \{\hat{\theta} = \theta\}, \theta \in \Theta}_{\text{partition of } \Omega}$

We choose $\mathbb{Q} = \mathbb{P}_{\vec{0}} = \mathcal{N}(\vec{0}, \sigma^2 Id)^{\otimes n}$

recall that $\mathbb{P}_\theta = \mathcal{N}(\theta, \sigma^2 Id)^{\otimes n}$

$KL(\mathbb{P}_\theta, \mathbb{Q}) = n\, KL(\mathcal{N}(\theta, \sigma^2 Id), \mathcal{N}(\vec{0}, \sigma^2 Id))$ (as a product)

$$= n \frac{\|\theta - \vec{0}\|_2^2}{2\sigma^2}$$

$$\leq \frac{n d \varepsilon^2}{2\sigma^2} \quad \text{because } \theta \in \{0, \varepsilon\}^d$$

$$\log |\Theta| \geq \log\left(e^{d/8}\right) = \frac{d}{8}$$

(3) thus yields,

$$\min_{\theta \in \Theta} \mathbb{P}_\theta(\hat{\theta} = \theta) \leq \left(\frac{n d \varepsilon^2}{2\sigma^2} + \log 2\right) / \frac{d}{8}$$

$$= \frac{4 n \varepsilon^2}{\sigma^2} + \frac{8 \log 2}{d} \leq 4c' + \frac{8\log 2}{6}$$

$\frac{8 \log 2}{d} \leq \frac{8 \log 2}{6} < 1$, We choose $\varepsilon = \frac{\sigma}{\sqrt{n}} c'$

such that $\frac{4n\varepsilon^2}{\sigma^2} = 4c'^2$

Combining everything, we get

$$\max_{\theta \in \Theta} \mathbb{E}_\theta \left[\|\hat{\theta} - \theta\|_2^2 \right] \geq \frac{d\varepsilon^2}{16} \left(1 - 4c'^2 - \frac{8\log 2}{6} \right)$$

$$\gtrsim \frac{d\sigma^2}{n}$$

for c' small enough

↑ there is a c_k