Minimax Lower Bounds

① Minimax risk

- Statistical setting:
  - $(\mathbb{P}_f)_{f\in\mathcal{F}}$: a set of distributions on a measurable space $(\mathcal{Y},\mathcal{A})$
  - d: a distance on $\mathcal{F}$
  - risk: for any estimator $\hat{f}:\mathcal{Y}\to\mathcal{F}$, we consider the risk $R(\hat{f}) := \mathbb{E}_f\left[d(\hat{f}(Y),f)^q\right]$ for some $q>0$.

Ex: $\mathcal{F}=\mathbb{R}^n$, $\mathbb{P}_f=\mathcal{N}(f,\sigma^2 I_n)$, d = Euclidean distance, $q=2$

$\rightsquigarrow$ $R(\hat{f})=\mathbb{E}_f\left[\|\hat{f}(Y)-f\|^2\right]$

- Best estimator $\hat{f}$?

⚠ For all $f\in\mathcal{F}$, we have

$$\min_{\hat{f}:\mathcal{Y}\xrightarrow{\text{meas.}}\mathcal{F}} \mathbb{E}_f\left[d(\hat{f}(Y),f)^q\right]=0 \qquad (\text{reached for } \hat{f}(Y)=f)$$

$\rightsquigarrow$ no sense

💡 we want $\hat{f}$ to be good on the whole class $\mathcal{F}$

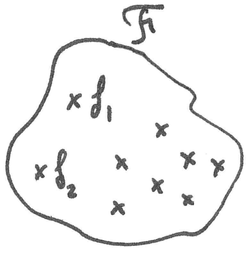
- Minimax risk:

$$R^*(\mathcal{F}) := \min_{\hat{f}:\mathcal{Y}\xrightarrow[\text{meas.}]{}\mathcal{F}} \ \max_{f\in\mathcal{F}} \mathbb{E}_f\left[d(\hat{f}(Y),f)^q\right]$$

- **our goal:** proving some lower bounds on $R^*(\mathcal{F})$.
- **Useful?:** If
  - $\to$ we prove $R^*(\mathcal{F}) \geq$ lower bound
  - $\to$ and find $\hat{f}$ such that $R(\hat{f}) \approx$ lower bound

Then, $\hat{f}$ performs almost as well as the best estimator in terms of minimax risk.

- **Recipe:**
  - discretization of $\mathcal{F}$: $\max\limits_{f \in \mathcal{F}} \geq \max\limits_{f \in \{f_1, \dots, f_N\}}$
  - use lower bounds from information theory.



② A recipe for proving lower bounds (in 3 steps)

Step 1: a key lemma from information theory.

Kullback-Leibler divergence: For any $\mathbb{P} \ll \mathbb{Q}$, then

$$KL(\mathbb{P}, \mathbb{Q}) := \int \log \frac{d\mathbb{P}}{d\mathbb{Q}} \, d\mathbb{P} \geq 0$$

Ex: Gaussian distribution $\mathbb{P}_f = \mathcal{N}(f, \sigma^2 I_n)$

$$\begin{aligned} KL(\mathbb{P}_f, \mathbb{P}_g) &= \int_{x \in \mathbb{R}^n} \log \frac{e^{-\|x-f\|^2/2\sigma^2}}{e^{-\|x-g\|^2/2\sigma^2}} \, d\mathbb{P}_f(x) \\ &= \int_{x \in \mathbb{R}^n} \frac{1}{2\sigma^2} \left(\|f-g\|^2 + 2\langle x-f, f-g \rangle \right) d\mathbb{P}_f(x) \\ &= \frac{\|f-g\|^2}{2\sigma^2} \end{aligned}$$

Fano's Lemma

For any $\mathbb{P}_1, .., \mathbb{P}_N, \mathbb{Q}$ probability distribution on $\mathcal{Y}$, such that $\mathbb{P}_j \ll \mathbb{Q}$, for $j=1,..,N$, we have

$$\min_{\hat{J}:\mathcal{Y}\to\{1,..,N\}} \frac{1}{N}\sum_{j=1}^{N} \mathbb{P}_j\left[\hat{J}(Y)\neq j\right] \geq 1 - \frac{1+\frac{1}{N}\sum_{j=1}^{N} KL(\mathbb{P}_j,\mathbb{Q})}{\log(N)}$$

Remark: a classical choice for $\mathbb{Q}$ is $\mathbb{Q} = \frac{1}{N}\sum_{j=1}^{N}\mathbb{P}_j$

Proof:

- We first observe that

$$\min_{\hat{J}:\mathcal{Y}\to\{1,..,N\}} \frac{1}{N}\sum_{j=1}^{N} \mathbb{P}_j\left[\hat{J}(Y)\neq j\right] = 1 - \underbrace{\max_{\hat{J}:\mathcal{Y}\to\{1,..,N\}} \frac{1}{N}\sum_{j=1}^{N} \mathbb{P}_j\left[\hat{J}(Y)= j\right]}_{\text{to be upper-bounded}}$$

Lemma: explicit formula

$$\max_{\hat{J}:\mathcal{Y}\to\{1,..,N\}} \frac{1}{N}\sum_{j=1}^{N} \mathbb{P}_j\left[\hat{J}(Y)= j\right] = \frac{1}{N}\mathbb{E}_{\mathbb{Q}}\left[\max_{j=1,..,N} \frac{d\mathbb{P}_j}{d\mathbb{Q}}(Y)\right]$$

<u>Proof of formula:</u>

- $$\sum_{j=1}^{N} \mathbb{P}_j\left[\hat{J}(Y)=j\right] = \sum_{j=1}^{N} \int_{\mathcal{Y}} \mathbb{1}_{\hat{J}(y)=j} \underbrace{\frac{d\mathbb{P}_j}{dQ}(y)}_{\leq \max\limits_{k=1,\dots,N} \frac{d\mathbb{P}_k}{dQ}(y)} dQ(y)$$

$$\leq \int_{\mathcal{Y}} \underbrace{\sum_{j=1}^{N} \mathbb{1}_{\hat{J}(y)=j}}_{=1} \times \max_{k=1,\dots,N} \frac{d\mathbb{P}_k}{dQ}(y) \, dQ(y)$$

$$= \mathbb{E}_Q\left[\max_{k=1,\dots,N} \frac{d\mathbb{P}_k}{dQ}(Y)\right]$$

- In addition, the inequality is an equality for the MLE

$$\hat{J}(y) \in \operatorname*{argmax}_{k=1,\dots,N} \frac{d\mathbb{P}_k}{dQ}(y)$$

$\square$

- We can upper bound $\mathbb{E}_Q\left[\max\limits_{j=1,\dots,N} \frac{d\mathbb{P}_j}{dQ}(Y)\right]$ with a lemma from Lecture 1

<u>Lemma</u> (Lecture 1)

For any $Z_1, \dots, Z_N$ random variables with value in an interval $I \subset \mathbb{R}$, and any $\varphi : I \to \mathbb{R}^+$ convex, we have

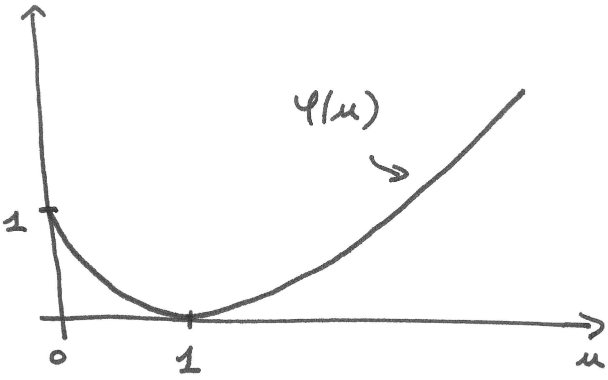
$$\varphi\left(\mathbb{E}\left[\max_{j=1,\dots,N} Z_j\right]\right) \leq \sum_{j=1}^{N} \mathbb{E}\left[\varphi(Z_j)\right]$$

- We choose

$$\varphi(u) = u \log u - u + 1$$

for $u \geq 0$.

and $Z_j = \frac{d\mathbb{P}_j}{d\mathbb{Q}}(Y)$



- $$\mathbb{E}_{\mathbb{Q}}\left[\varphi\left(\frac{d\mathbb{P}_j}{d\mathbb{Q}}(Y)\right)\right] = \int_{\mathcal{Y}} \log\left(\frac{d\mathbb{P}_j}{d\mathbb{Q}}(y)\right) \underbrace{\frac{d\mathbb{P}_j}{d\mathbb{Q}}(y)\, d\mathbb{Q}(y)}_{d\mathbb{P}_j(y)} - \underbrace{\int_{\mathcal{Y}} \frac{d\mathbb{P}_j}{d\mathbb{Q}}(y)\, d\mathbb{Q}(y)}_{=1} + 1$$

$$= KL(\mathbb{P}_j, \mathbb{Q})$$

So

$$\varphi\Bigg(\underbrace{\mathbb{E}_{\mathbb{Q}}\left[\max_{j=1,\dots,N} \frac{d\mathbb{P}_j}{d\mathbb{Q}}(Y)\right]}_{=: Nu}\Bigg) \leq \sum_{j=1}^{N} KL(\mathbb{P}_j, \mathbb{Q})$$

- $$\varphi(Nu) = Nu(\log N + \log(u)) - Nu + 1$$

$$= Nu \log N + N(\underbrace{u \log u - u + 1}_{\varphi(u) \geq 0}) - (N-1)$$

$$\geq Nu \log N - N$$

So replacing u by its value :

$$\log(N) \times \mathbb{E}_{\mathbb{Q}}\left[\max_{j=1,\dots,N} \frac{d\mathbb{P}_j}{d\mathbb{Q}}(Y)\right] \leq N + \sum_{j=1}^{N} KL(\mathbb{P}_j, \mathbb{Q}).$$

Conclusion:

$$\min_{\hat{\jmath}: \mathcal{Y} \to \{1,\dots,N\}} \frac{1}{N} \sum_{j=1}^{N} \mathbb{P}_j[\hat{\jmath}(Y) \neq j] = 1 - \frac{1}{N} \mathbb{E}_{\mathbb{Q}}\left[\max_{j=1,\dots,N} \frac{d\mathbb{P}_j}{d\mathbb{Q}}(Y)\right]$$

$$\geq 1 - \frac{1}{\log(N)}\left(1 + \frac{1}{N} \sum_{j=1}^{N} KL(\mathbb{P}_j, \mathbb{Q})\right)$$

$\square$

Step 2: From Fano's lemma to a lower bound over a finite set $\{f_1, \dots, f_N\}$

- For any $\hat{f} : \mathcal{Y} \to \mathcal{F}$ measurable, we define

$$\hat{J}(y) \in \underset{j=1,\dots,N}{\operatorname{argmin}}\ d(\hat{f}(y), f_j)$$

- We have $\forall j$:

$$\begin{aligned}
\min_{i \neq k} d(f_i, f_k)\, \mathbb{1}_{\hat{J}(y) \neq j} &\leq d(f_j, f_{\hat{J}(y)}) \\
&\leq d(f_j, \hat{f}(y)) + d(\hat{f}(y), f_{\hat{J}(Y)}) \\
\text{definition of } \hat{J} \rightarrow\ &\leq 2\, d(f_j, \hat{f}(y)) .
\end{aligned}$$

So, for any $\hat{f}$:

$$\frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_{f_j}\left[d(f_j, \hat{f}(Y))^q \right] \geq \frac{1}{2^q} \min_{i \neq k} d(f_i, f_k)^q \times \frac{1}{N} \sum_{j=1}^{N} \mathbb{P}_{f_j}\left[\hat{J}(Y) \neq j \right]$$

$$\geq \frac{1}{2^q} \min_{i \neq k} d(f_i, f_k)^q \times \min_{\hat{J} : \mathcal{Y} \to \{1,\dots,N\}} \frac{1}{N} \sum_{j=1}^{N} \mathbb{P}_{f_j}\left[\hat{J}(Y) \neq j \right] .$$

Hence, the corollary from Fano's lemma

Corollary 3.4

For any $\{f_1, \dots, f_N\} \subset \mathcal{F}$ and $\mathbb{Q} \gg \mathbb{P}_{f_j}$ $j = 1, \dots, N$

$$\min_{\hat{f} : \mathcal{Y} \to \mathcal{F}} \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_{f_j}\left[d(f_j, \hat{f}(Y))^q \right]$$

$$\geq \frac{1}{2^q} \min_{i \neq k} d(f_i, f_k)^q \times \left(1 - \frac{1 + \frac{1}{N} \sum_{j=1}^{N} KL(\mathbb{P}_{f_j}, \mathbb{Q})}{\log(N)} \right)$$

Step 3: finding a good discretization

For any $\{f_1, \dots, f_N\} \subset \mathcal{F}$:

$$R^*(\mathcal{F}) := \min_{\hat{f}: \mathcal{Y} \to \mathcal{F}} \max_{f \in \mathcal{F}} \mathbb{E}_f\left[d(\hat{f}(Y), f)^q\right]$$

$$\geq \min_{\hat{f}: \mathcal{Y} \to \mathcal{F}} \max_{j=1,\dots,N} \mathbb{E}_{f_j}\left[d(\hat{f}(Y), f_j)^q\right]$$

$$\geq \min_{\hat{f}: \mathcal{Y} \to \mathcal{F}} \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_{f_j}\left[d(\hat{f}(Y), f_j)^q\right]$$

$$\underset{\text{Cor. 3.4}}{\geq} \frac{1}{2^q} \min_{i \neq k} d(f_i, f_k)^q \times \left(1 - \frac{1 + \frac{1}{N}\sum_{j=1}^{N} KL(\mathbb{P}_{f_j}, \mathbb{Q})}{\log(N)}\right)$$

All the art is to find a good discretization $\{f_1, \dots, f_N\}$.

balance between
- $\min_{i \neq k} d(f_i, f_k)$ as large as possible
- $\dfrac{1 + \frac{1}{N}\sum_{j=1}^{N} KL(\mathbb{P}_{f_j}, \mathbb{Q})}{\log(N)}$ smaller than 1

recipe: find $f_1, \dots, f_N$ with

$$\frac{1 + \frac{1}{N}\sum_{j=1}^{N} KL(\mathbb{P}_{f_j}, \mathbb{Q})}{\log(N)} \leq \frac{1}{2} \quad \text{and} \quad d(f_i, f_k) \text{ as large as possible.}$$

Remark: there is a variant of Fano's lemma (based on Birgé's Lemma) which is sometimes more handy. It leads to the next variant of Corollary 3.4

Corollary 3.6:

For any $\{f_1, \ldots, f_N\} \subset \mathcal{F}$ such that

$$\max_{j \neq k} KL(\mathbb{P}_{f_j}, \mathbb{P}_{f_k}) \leq \frac{2e}{2e+1} \log(N) \qquad (*)$$

We have

$$\min_{\hat{f}: \mathcal{Y} \to \mathcal{F}} \max_{j=1,\ldots,N} \mathbb{E}_{f_j}\left[d(\hat{f}(Y), f_j)^q \right] \geq \frac{1}{2^q(2e+1)} \min_{j \neq k} d(f_j, f_k)^q .$$

Proof: With the notations of Fano's lemma, the events $A_j = \{\hat{J}(Y) = j\}$ are disjoint so:
Theorem B.13 ensures that

$$\min_{j=1,\ldots,N} \mathbb{P}_j[\hat{J}(Y) = j] \leq \frac{2e}{2e+1} \vee \max_{j \neq k} \frac{KL(\mathbb{P}_{f_j}, \mathbb{P}_{f_k})}{\log(N)}$$

$$\overset{(*)}{\leq} \frac{2e}{2e+1}$$

Hence we get the variant of Fano's lemma: when $(*)$ holds

$$\min_{\hat{J}: \mathcal{Y} \to \{1,\ldots,N\}} \max_{j=1,\ldots,N} \mathbb{P}_{f_j}[\hat{J}(Y) \neq j] \geq \frac{1}{2e+1} .$$

Conclusion: same lines as proof of Corollary 3.4. $\square$

3 Minimax risk for coordinate sparse regression

- We consider here $\mathcal{F}_D = \{ f = X\beta : |\beta|_0 \leq D \}$,

  $\mathbb{P}_f = \mathcal{N}(f, \sigma^2 I_n)$, $d(f_1, f_2) = \| f_1 - f_2 \|$ and $q = 2$.

- we have seen that $KL(\mathbb{P}_{f_1}, \mathbb{P}_{f_2}) = \dfrac{\| f_1 - f_2 \|^2}{2\sigma^2}$

- Restricted isometry constants: for $D_{max} \leq p/2$

$$\underline{c}_X := \inf_{|\beta|_0 \leq 2D_{max}} \frac{\|X\beta\|}{\|\beta\|} \leq \sup_{|\beta|_0 \leq 2D_{max}} \frac{\|X\beta\|}{\|\beta\|} =: \bar{c}_X$$

- Theorem 3.5

  - Let us fix $D_{max} \leq p/5$.
  - For any $D \leq D_{max}$, we have

$$R^*(\mathcal{F}_D) \geq \frac{e}{4(2e+1)^2} \left(\frac{\underline{c}_X}{\bar{c}_X} \right)^2 \times D \log\left(\frac{p}{5D} \right) \times \sigma^2$$

Proof: The recipe is to

$\to$ find $f_1, \dots, f_N \in \mathcal{F}_D$, well spread and fulfilling $(*)$

$\to$ apply corollary 3.6

Lemma 3.7:

For any $D \leq p/5$, there exists $\{\beta_1, \dots, \beta_N\} \subset \{\beta \in \{0,1\}^p : |\beta|_0 = D\}$ such that

i) $|\beta_j - \beta_k|_0 \geq D, \quad \forall j \neq k$

ii) $\log N \geq \frac{D}{2} \log \frac{p}{5D}$

proof: exercise 3.6.2 $\square$

- We choose $f_j = r X \beta_j$, $j = 1, \dots, N$, with $\overset{\text{scaling}}{\overset{\downarrow}{r}}$ such that $(*)$ holds:

$$\max_{j \neq k} KL(\mathbb{P}_{f_j}, \mathbb{P}_{f_k}) = \max_{j \neq k} \frac{r^2 \|X(\beta_j - \beta_k)\|^2}{2\sigma^2}$$

$$\leq \frac{r^2}{2\sigma^2} \bar{c}_X^2 \max_{j \neq k} \underbrace{\|\beta_j - \beta_k\|^2}_{\leq |\beta_j - \beta_k|_0 \leq 2D}$$

$$\overset{?}{\leq} \frac{2e}{2e+1} \log N$$

OK for $r^2 = \frac{\sigma^2}{\bar{c}_X^2 D} \times \frac{2e}{2e+1} \log N$

- In addition:

$$\|f_j - f_k\|^2 = r^2 \|X(\beta_j - \beta_k)\|^2$$

$$\geq r^2 \underline{c}_X^2 \underbrace{\|\beta_j - \beta_k\|^2}_{= |\beta_j - \beta_k|_0 \underset{i)}{\geq} D} \geq r^2 D \underline{c}_X^2$$

$$\geq \left(\frac{\underline{c}_X}{\bar{c}_X}\right)^2 \sigma^2 \times \frac{2e}{2e+1} \log N \underset{ii)}{\geq} \left(\frac{\underline{c}_X}{\bar{c}_X}\right)^2 \sigma^2 \frac{e}{2e+1} \times D \log \frac{p}{5D}$$

- Applying Corollary 3.6 gives Theorem 3.5 $\square$