Structure Learning

Lecture 2.

Preambule

Setting

Data : $(X_i, Y_i)_{i=1\dots m}$

Model : $Y_i = f^*(X_i) + \varepsilon_i$ with

→ $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0,1) \perp\!\!\!\perp (X_i)_{i=1..m}$

→ $f^* \in \mathcal{F}$ ← functional class

Empirical risk minimizer :

$$\hat f \in \operatorname*{argmin}_{f\in\mathcal F} \frac{1}{m} \sum_{i=1}^m (Y_i - f(X_i))^2$$

Notation : $d_m(f,g)^2 = \frac{1}{m}\sum_{i=1}^m (f(X_i)-g(X_i))^2$

A simple bound

By definition, for any $f\in\mathcal F$:

$$\frac{1}{m}\sum_i (\overbrace{f^*(X_i)+\varepsilon_i}^{Y_i} - \hat f(X_i))^2 \le \frac{1}{m} \sum_i (f^*(X_i)+\varepsilon_i - f(X_i))^2$$

so

$$d_m(\hat f, f^*)^2 \le d_m(f,f^*)^2 + \frac{2}{m}\sum_{i=1}^m (\hat f(X_i) - f(X_i))\varepsilon_i$$

So for $f=f^*$:

$$d_m(\hat f,f^*)^2 \le \frac{2}{m} \sum_{i=1}^m (\hat f(X_i)-f^*(X_i))\varepsilon_i$$

Assume for simplicity that

$$\{f-f^* : f\in\mathcal F\} \subset \{\lambda g : \to \lambda\ge 0,\ \to g\in\mathcal G\}$$

with $\frac{1}{m}\sum_{i=1}^m g(X_i)^2 = 1$ for all $g\in\mathcal G$.

Then, we have $\hat f - f^* = d_m(\hat f,f^*)\hat g$, with $\hat g\in\mathcal G$
and

$$d_m(\hat f,f^*)^2 \le 2\, d_m(\hat f, f^*) \sup_{g\in\mathcal G} \frac{1}{m}\sum_{i=1}^m g(X_i)\varepsilon_i$$

Hence

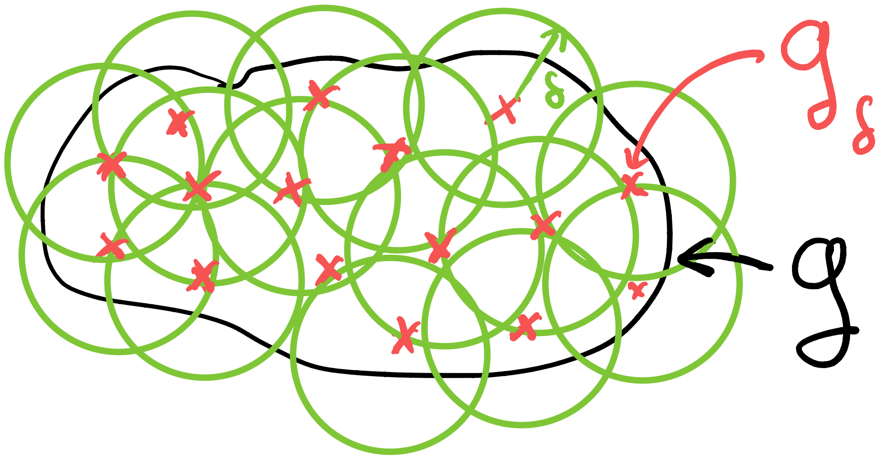
$$d_m(\hat f,f^*) \le 2 \underbrace{\sup_{g\in\mathcal G} \underbrace{\frac{1}{m} \sum_{i=1}^m g(X_i)\,\varepsilon_i}_{\sim \mathcal N(0,1/m)}}_{=: \mathcal S(\mathcal G)}$$

How large is $\mathcal S(\mathcal G)$? problem: if $\mathcal G$ is infinite, union bound is useless...

We can provide a simple upper-bound based on covering numbers.

Let $\delta > 0$, and let $\mathcal{G}_\delta \subset \mathcal{G}$ be a δ-covering of $\mathcal{G}$:

$$\forall g \in \mathcal{G},\ \exists g' \in \mathcal{G}_\delta \text{ s.t. } d_m(g,g') \leq \delta$$



Then, we have:

$$\frac{1}{m}\sum_{i=1}^{m} g(x_i)\,\varepsilon_i = \frac{1}{m}\sum_{i=1}^{m} \underbrace{g'(x_i)}_{\in \mathcal{G}_\delta}\varepsilon_i + \underbrace{\frac{1}{m}\sum_{i=1}^{m}\big(g(x_i)-g'(x_i)\big)\varepsilon_i}_{\leq d_m(g,g')\cdot\|\varepsilon\|_m}$$

So choosing $g' \in \mathcal{G}_\delta$ s.t. $d_m(g,g') \leq \delta$ we get

$$\mathcal{S}(\mathcal{G}) \leq \delta\|\varepsilon\|_m + \sup_{g \in \mathcal{G}_\delta} \underbrace{\frac{1}{m}\sum_{i=1}^{m} g(x_i)\,\varepsilon_i}_{\sim \mathcal{N}(0,1/m) \text{ conditionally to } x_i.}$$

So

$$\mathbb{E}_\varepsilon[\mathcal{S}(\mathcal{G})] \leq \delta + \sqrt{\frac{2}{m}\log \mathrm{Card}(\mathcal{G}_\delta)}$$

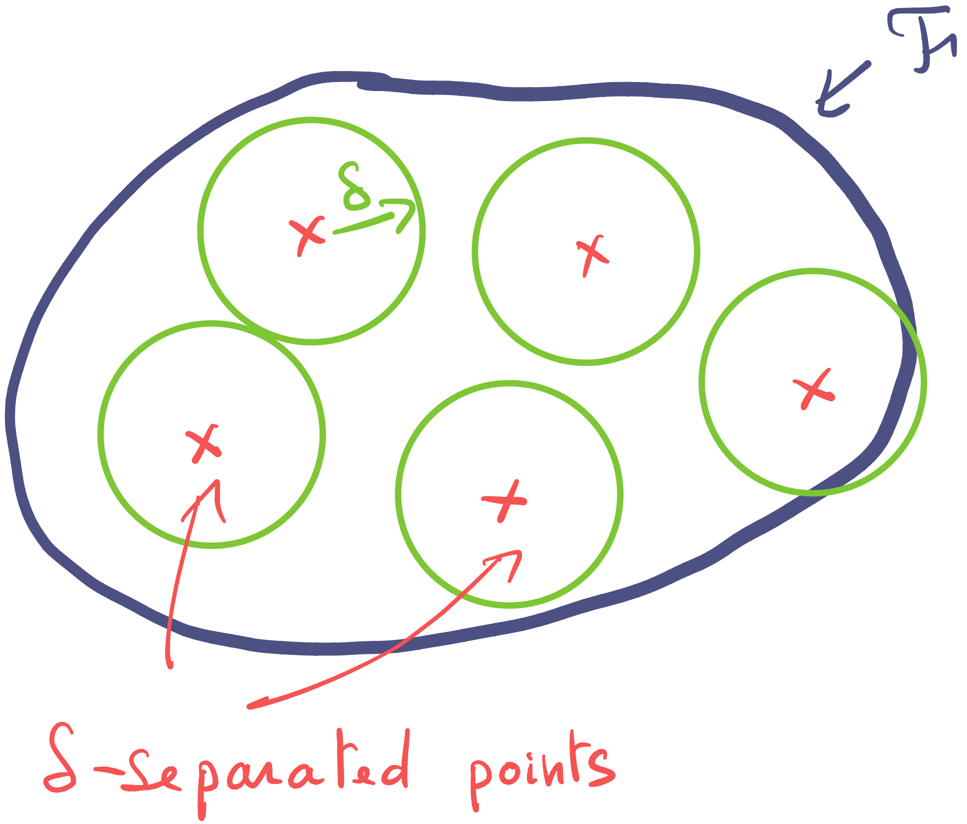
For example, if $\log \mathrm{Card}(\mathcal{G}_\delta) \simeq 1/\delta^\alpha$:

$$\mathbb{E}_\varepsilon[d_m(\hat{f}, f^*)] \leq c \inf_{\delta>0}\left(\delta + \frac{1}{\sqrt{m\delta^\alpha}}\right)$$

$$\delta = \frac{1}{m^{\frac{1}{2+\alpha}}} \rightsquigarrow \leq \frac{c'}{m^{\frac{1}{2+\alpha}}}$$

$\rightsquigarrow$ the massiveness of $\mathcal{G}$ (value α) governs the convergence rate.

We can actually prove the following: Let $N(\mathcal{F}, d_n, \delta)$ be the maximum number of δ-separated points in $(\mathcal{F}, d_n)$



Assume that

$$\log N(\mathcal{F}, d_n, \delta) \overset{[\delta \to 0]}{\simeq} \frac{1}{\delta^\alpha}$$

Then

$$\inf_{\hat{f}} \max_{f^* \in \mathcal{F}} \mathbb{E}_{f^*}\left[d_n(\hat{f}, f^*)^2\right] \geq \frac{c}{n^{2/2+\alpha}}$$

[exercise 3.6.4.].

Examples: for $\beta \in (0,1]$, $\sigma^2 = 1$

① β-Hölder functions in $\mathbb{R}^p$

Set $\mathcal{F} = \{f : [0,1]^p \to \mathbb{R} : |f(x) - f(y)| \leq |x-y|^\beta\}$.

Then

$$\inf_{\hat{f}} \max_{f^* \in \mathcal{F}} \mathbb{E}_{f^*}\left[d(\hat{f}, f^*)^2\right] \overset{n \to \infty}{\simeq} \frac{1}{n^{\frac{2}{2+p/\beta}}}$$

② Single index model

Set $\mathcal{F} = \left\{ f(x) = h(\langle x, \omega \rangle) : \begin{array}{l} \cdot \|\omega\| = 1 \\ \cdot h : \mathbb{R} \to \mathbb{R} \\ \quad \beta\text{-Hölder} \end{array} \right\}$. Then,

h, ω: unknown

$$\inf_{\hat{f}} \sup_{f^* \in \mathcal{F}} \mathbb{E}_{f^*}\left[d_n(\hat{f}, f^*)^2 \right] \underset{n \to \infty}{\asymp} \frac{1}{n^{2/(2 + 1/\beta)}}$$

Benefit of learning the direction ω

Learning structures in the linear model

The linear model is ubiquitous in data analysis. Predictions often take the form

$$\hat{f}(x) = \langle \hat{\beta}, \Phi(x) \rangle$$

features (learnt or chosen)

or L^2-basis, etc...

$\leadsto$ we focus on the linear model

$$f^*(x) = \langle \underbrace{\beta^*}_{\in \mathbb{R}^p,\ \text{unknown}}, x \rangle$$

. Model: $y_i = \langle \beta^*, x_i \rangle + \varepsilon_i$, $i=1,\dots,n$

with $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

. Notation:

$$Y := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}; \quad f^* := \begin{bmatrix} f^*(x_1) \\ \vdots \\ f^*(x_n) \end{bmatrix}; \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$\langle \beta^*, x_1 \rangle$

and $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}$

$$\Rightarrow \quad Y = X\beta^* + \varepsilon = f^* + \varepsilon.$$

. Hidden structure: we assume that

$$\underbrace{|\beta^*|_0 := \operatorname{card}\{j : \beta_j^* \neq 0\} \text{ is small}}_{\text{Coordinate sparse assumption.}}$$

How can we benefit from this assumption?

We set $S^* := \operatorname{supp}(\beta^*) \leftarrow$ unknown

linear span $\rightarrow \bar{S}^* = \{X\beta : \operatorname{supp}(\beta) \subset S^*\} \subset \mathbb{R}^n$

■ **If S^* was known:** Then, we could solve

$$\hat{\beta}^{(S)} := \underset{\mathrm{supp}(\beta) \subset S}{\arg\min} \|Y - X\beta\|^2$$

for $S = S^*$. ⇝ back to low dimensional regression.

Then $\hat{f}^{(S^*)} = X\hat{\beta}^{(S^*)} = \mathrm{Proj}_{\bar{S}^*} Y$

■ **When S^* is unknown:**

- We can try to compare $\hat{\beta}^{(S)}$ for different guessed support $S \in \mathcal{S}$

↑ Collection of guessed supports

Reminder:
$$d_n(f, g)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - g(x_i)\right)^2$$

Risk:

$$R(\hat{f}^{(S)}) := \mathbb{E}\Big[\overbrace{d_n(\hat{f}^{(S)}, f^*)^2}^{\frac{1}{n}\|\mathrm{Proj}_{\bar{S}}(f^* + \varepsilon) - f^*\|^2}\Big]$$

Pythagore

$$= \frac{1}{n}\mathbb{E}\left[\|\mathrm{Proj}_{\bar{S}}\,\varepsilon\|^2\right] + \frac{1}{n}\mathbb{E}\left[\|f^* - \mathrm{Proj}_{\bar{S}} f^*\|^2\right]$$

$$= \underbrace{\frac{\sigma^2}{n}\underbrace{\mathrm{Tr}(\mathrm{Proj}_{\bar{S}})}_{= \dim(\bar{S})}}_{\text{variance term}} + \underbrace{\frac{1}{n}\|f^* - \mathrm{Proj}_{\bar{S}} f^*\|^2}_{\text{bias term}}$$

Best S? oracle choice

$$S_0 = \underset{S \in \mathcal{S}}{\arg\min}\Big\{\underbrace{\frac{1}{n}\|f^* - \mathrm{Proj}_{\bar{S}} f^*\|^2}_{\text{unknown !!}} + \frac{\sigma^2}{n}\dim(\bar{S})\Big\}$$

1) estimate $R(\hat{f}^{(S)})$ by some $\hat{R}(\hat{f}^{(S)})$

2) Choose $\hat{f}^{(\hat{S})}$ with

$$\hat{S} \in \operatorname*{argmin}_{S \in \mathcal{S}} \hat{R}(\hat{f}^{(S)})$$

Questions:

- which $\hat{R}(\hat{f}^{(S)})$?
- which performance? (Can we bypass the curse of dimensionality?)

We can try to estimate the bias $\| f^* - \mathrm{Proj}_{\bar{S}} f^* \|^2$ with $\| Y - \underbrace{\mathrm{Proj}_{\bar{S}} Y}_{= \hat{f}^{(S)}} \|^2$.

Let us analyse it!

$$\mathbb{E}\Big[\|\underbrace{Y}_{Y = f^* + \varepsilon} - \underbrace{\mathrm{Proj}_{\bar S} Y}_{\hat f^{(S)}}\|^2\Big] = \underbrace{\|f^* - \mathrm{Proj}_{\bar S} f^*\|^2}_{} + \underbrace{\mathbb{E}\Big[\|\varepsilon - \mathrm{Proj}_{\bar S}\varepsilon\|^2\Big]}_{= (n - \dim(\bar S))\sigma^2} + 2\,\underbrace{\mathbb{E}\Big[\langle f^* - \mathrm{Proj}_{\bar S} f^*, \varepsilon - \mathrm{Proj}_{\bar S}\varepsilon\rangle\Big]}_{=0}$$

So we have the unbiased estimation of the risk

$$\hat R_{AIC}(\hat f^{(S)}) := \frac{1}{n}\|Y - \hat f^{(S)}\|^2 + \underbrace{\frac{2}{n}\sigma^2 \dim(\bar S)}_{\text{correction term}} \underbrace{(-\sigma^2)}_{\text{can be dropped for computing } \hat S.}$$

⚠ Problem

$\min_{S \in \mathcal{S}} \|Y - \mathrm{Proj}_{\bar S} Y\|^2 + 2\sigma^2 \dim(\bar S)$ can very much

deviate from $\min_{S \in \mathcal{S}} \mathbb{E}\Big[\|Y - \mathrm{Proj}_{\bar S} Y\|^2\Big] + 2\sigma^2 \dim(\bar S)$

when $\mathcal{S}$ is large e.g. $\mathcal{S} = \{S : S \subset \{1, \dots, p\}\}$.

Indeed,

$$\|Y - \mathrm{Proj}_{\bar S} Y\|^2 = \underbrace{\|f^* - \mathrm{Proj}_{\bar S} f^*\|^2}_{\text{what we want to estimate}} + \underbrace{\|\varepsilon\|^2}_{\text{does not depend on } S} - \underbrace{\|\mathrm{Proj}_{\bar S}\varepsilon\|^2}_{\text{depends on } S} + \underbrace{\text{cross-product}}_{\text{let's forget it here}}$$

We have $\mathbb{E}\left[\|\mathrm{Proj}_{\bar S}\varepsilon\|^2\right] = \dim(\bar S)\,\sigma^2$. Setting $d_S := \dim(\bar S)$, we have

$$\mathbb{P}\left[\|\mathrm{Proj}_{\bar S}\varepsilon\|^2 \leq \sigma^2\left(\sqrt{d_S}+\sqrt{2L}\right)^2\right] \geq 1-e^{-L}$$

quite tight

In particular, we have

$$\mathbb{P}\left[\min_{|S|=d} -\|\mathrm{Proj}_{\bar S}\varepsilon\|^2 \geq -\sigma^2\left(\sqrt{d}+\sqrt{2\log C_p^d + 2L}\right)\right] \geq 1-e^{-L}$$

union bound

since $\log C_p^d \leq d\log\left(\frac{ep}{d}\right)$, we must correct $-\|\mathrm{Proj}_{\bar S}\varepsilon\|^2$

not by $d_S\sigma^2$ but by $\left(\sqrt{d_S}+\sqrt{2d_S\log\frac{ep}{d_S}}\right)^2\sigma^2 \asymp 2d_S\log\left(\frac{ep}{d_S}\right)\sigma^2$

d_S small $\to$ $\simeq 2d_S\log(p)\,\sigma^2$

Conclusion: above analysis suggests to select

$$\hat{S} \in \underset{S}{\operatorname{argmin}} \|Y - \hat{f}^{(S)}\|^2 + \operatorname{pen}(S)\,\sigma^2$$

where $\operatorname{pen}(S) = K\, d_S \log\left(\frac{ep}{d_S}\right)$

for some $K \gtrsim 2$.

Theorem: $\exists\, C_K, C'_K > 0$, such that

$$R(\hat{f}^{(\hat{S})}) \leq C_K \min_S \left\{ R(\hat{f}^{(S)}) + \frac{\sigma^2}{n} \operatorname{pen}(S) \right\}$$

$$\leq C'_K \underbrace{\log\left(\frac{p}{d_{S^*}}\right)}_{:(} \underbrace{R(\hat{f}^{(S^*)})}_{= d_{S^*} \frac{\sigma^2}{n}}$$

for $S = S^*$

Proof: See Theorem 2.2 □

Can we avoid the $\log p$ factor?

Minimax estimation:

We can prove that

$$\min_{\hat{f}} \sup_{X \in \mathbb{R}^{n \times p}} \sup_{|S^*| = d} R(\hat{f}, f^*) \asymp \frac{\sigma^2}{n} d \log \frac{p}{d}$$

so the log factor is unavoidable.

Yet, the risk of $\hat{f}^{(\hat{S})}$ is much smaller than the risk of vanilla least-square, if

$$|S^*| \log \frac{p}{|S^*|} \ll p \quad (\leq n) :$$

$$R(\hat{f}^{(\hat{S})}) \leq \frac{\sigma^2}{n} |S^*| \log \frac{p}{|S^*|} \ll \frac{\sigma^2}{n} p = R(\hat{f}^{LS})$$

if $\operatorname{rank}(X) = p$

⇝ We have been able to take benefit of structure to get a better estimator

The complexity for computing $\hat{S}$ is $\geq 2^p$ in general

$\rightsquigarrow$ prohibitive in practice !

<u>Except when</u> :

- $\rightarrow$ the column of X are orthogonal (simple thresholding : exercise 2.8.1)
- $\rightarrow$ $f(x)$ piecewise constant $(n=p)$, then complexity $= O(p^3)$ with dynamic programming (exercise 2.8.4)

To be continued...