Apprentissage supervisé

$\vec{\beta} \in \mathbb{R}^{D+1}$

$\vec{x} \in \mathbb{R}^{D}$

→ 1er modèle : régression linéaire

$$h_\beta(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_D x_D$$

$y(x) \in \mathbb{R}$

$\tilde{x} = [1, \vec{x}] \in \mathbb{R}^{D+1}$

β_j : coefficients de régression

Ensemble de données d'entraînement $\mathcal{D} = \{ x^{(i)}, t^{(i)} \}_{i=1}^{N}$

fonction de coût $\mathcal{L}(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(h_\beta(x^{(i)}) - t^{(i)} \right)^2$

→ Apprentissage : Minimisation de $\mathcal{L}(\beta)$

→ 2 approches : Descente de gradient

β_{init} : $\beta \leftarrow \beta - \eta \operatorname{grad}_{\beta} L(\beta)$

Résolution des équations normales

$$\rightarrow (\underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}})\vec{\beta} = \underline{\underline{\tilde{X}}}^T \vec{t} \Rightarrow \vec{\beta}_{OLS} = (\underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}})^{-1} \underline{\underline{\tilde{X}}}^T \vec{t}$$

$$\underline{\underline{X}} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_D^{(1)} \\ \vdots & \vdots & & & \\ 1 & x_1^{(N)} & \cdots & & x_D^{(N)} \end{bmatrix} \rightarrow \begin{bmatrix} x_1 & x_2 \\ x_1' & x_2' \end{bmatrix}$$

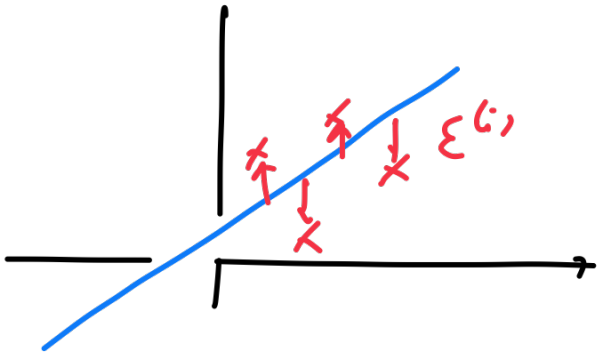
$$\begin{bmatrix} x_1^2 + (x_1')^2 & x_1 x_2 + x_1' x_2' \\ x_1 x_2 + x_1' x_2' & x_2^2 + (x_2')^2 \end{bmatrix} = \begin{bmatrix} x_1 & x_1' \\ x_2 & x_2' \end{bmatrix} \begin{bmatrix} x_1 & x_2 \\ x_1' & x_2' \end{bmatrix}$$

→ Aujourd'hui

→ Motivations statistiques

→ Cas de $\underline{\underline{X}}^T \underline{\underline{X}}$ non inversible

(présence de dépendances au niveau des caractéristiques)



Motivation statistique

Hypothèse #1 $t^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D + \varepsilon^{(i)}$ $\quad \varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$

Hypothèse #2 $t^{(i)}$ i.i.d

(indépendants et identiquement distribués)

$$\mathcal{N}(0, \sigma^2) \sim \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

$$t^{(i)} \sim \mathcal{N}(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}, \sigma^2)$$

$$\sim \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}))^2}{2\sigma^2}\right)$$

$$p(\{t^{(i)}\}_{i=1}^N | \{x^{(i)}\}_{i=1}^N, \beta) \sim \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t^{(i)} - h_\beta(x^{(i)}))^2}{2\sigma^2}\right)$$

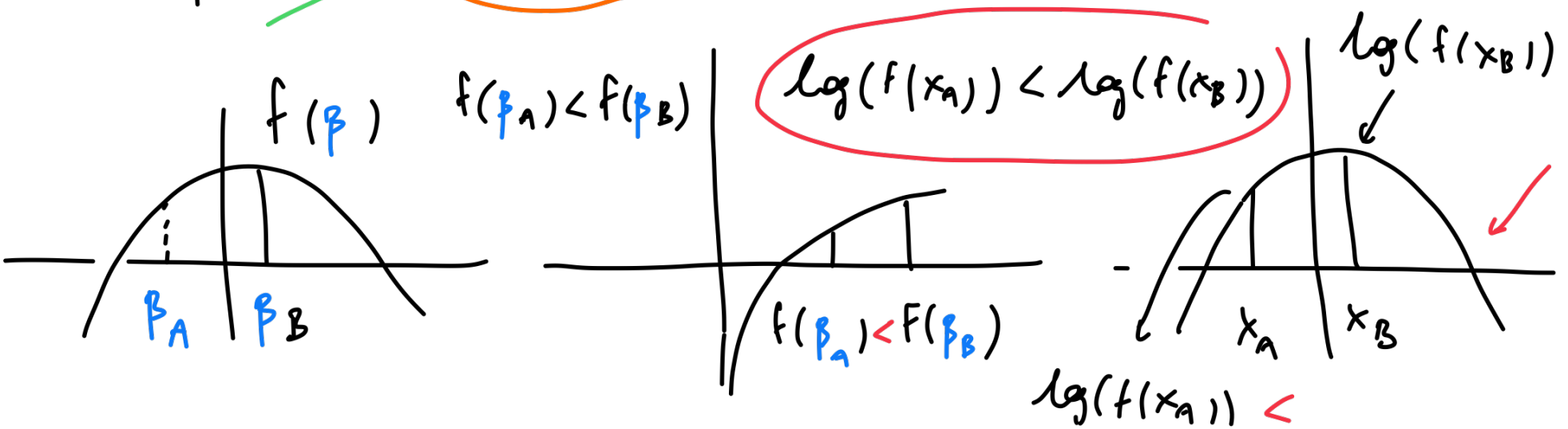
$$\rightarrow \hat{\beta}_{MLE} = \operatorname{argmax} \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}))^2}{2\sigma^2}\right)$$

$\hat{\beta}_{MLE} \rightarrow$ Estimateur de Maximum de vraisemblance

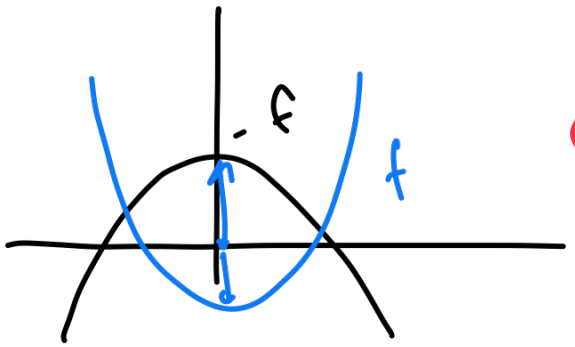
$$\hat{\beta}_{MLE} = \operatorname{argmax} \log \left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}\right)\right)^2}{2\sigma^2} \right) \right)$$

$$= \operatorname{argmax} \sum_{i=1}^{N} \left[\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{\left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}\right)\right)^2}{2\sigma^2} \right]$$

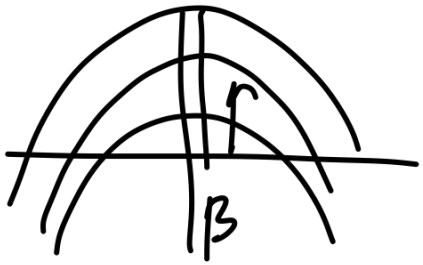
$$= \operatorname*{argmax}_{\beta} \; N \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(t^{(i)} - h_\beta(x^{(i)})\right)^2$$



$$\beta_{MLE} = \underbrace{\operatorname{argmax}}_{\beta} -\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_D x_D^{(i)}\right)\right)^2$$



$$\beta_{MLE} = \operatorname{argmin}_{\beta} \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(t^{(i)} - h_\beta(x^{(i)})\right)^2$$



$$\operatorname*{argmax}_{\beta} f(\beta) + c$$

$$\operatorname{argmax} f(\beta)$$

Cas de $\tilde{\underline{\underline{X}}}^T \tilde{\underline{\underline{X}}}$ non inversible $\quad \tilde{\underline{\underline{X}}} = \begin{bmatrix} 1 & x_1^{(2)} & \cdots & x_D^{(2)} \\ \vdots & \vdots & & \vdots \\ 1 & x_1^{(N)} & \cdots & x_D^{(N)} \end{bmatrix}$

Traiter la dépendance

Possibilité #1 : Orthogonalisation successive (a.k.a Gram Schmidt)

→ Commencer avec z_0 (= première colonne de $\tilde{\underline{\underline{X}}}$)

For $j = 1, \dots, D$

Calculer les coefficients $\hat{\gamma}_{lj} = \dfrac{\langle z_l, x_j \rangle}{\langle z_l, z_l \rangle}$

for $l = 0, \dots, j-1$

→ Et définir le nouveau z_j comme

$$z_j = c_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} z_k$$

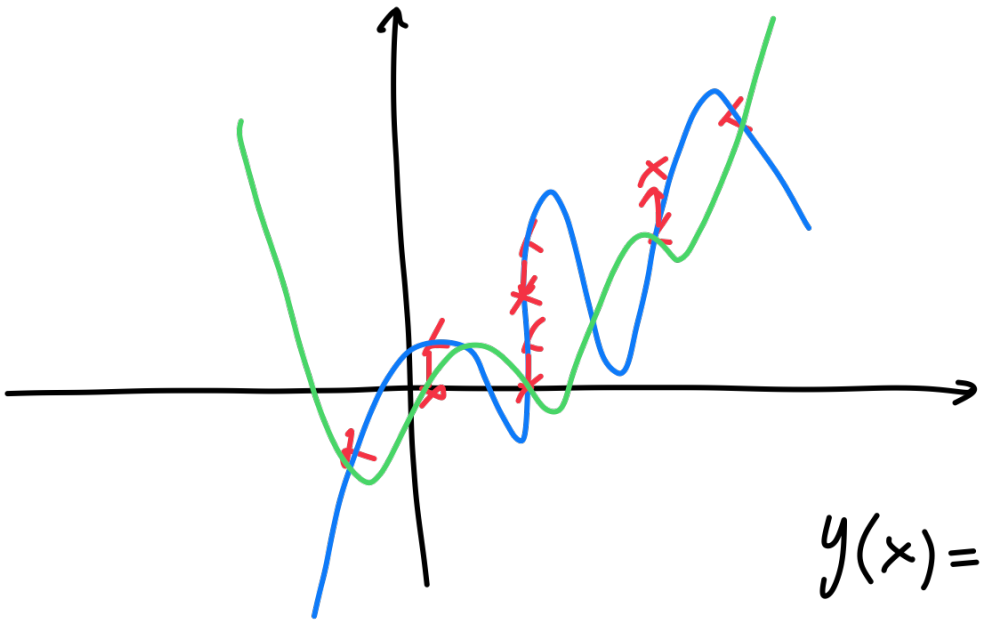
$$\rightarrow z_{D+2} = c_{D+1} - \sum_{k=0}^{D} \hat{\gamma}_{kj} z_k$$

$\vec{z}_{D+2}$ est la seule colonne représentant c_{D+1}

$$\beta_{D+2} = \frac{\langle z_{D+1}, \vec{t} \rangle}{\langle z_{D+1}, z_{D+2} \rangle}$$

→ Si z_{D+1} est petit
(caractéristiques dépendantes)

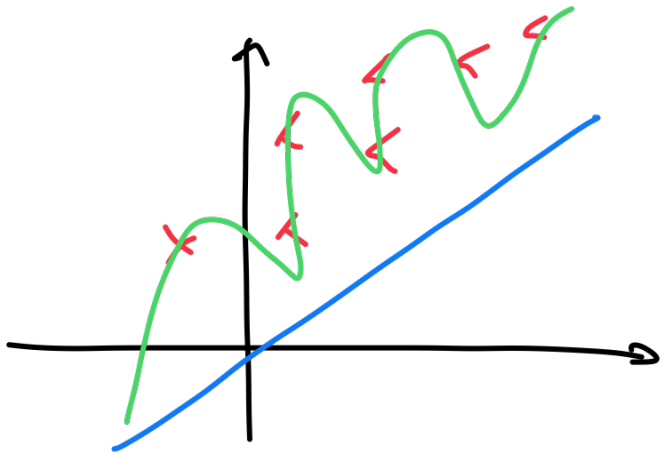
$\beta_{D+1} >>>$ ⇒ estimation instable



$$y(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots +$$

$$\underset{OLS}{\beta} = (\tilde{\underline{\underline{X}}}^T \tilde{\underline{\underline{X}}})^{-2} \tilde{\underline{\underline{X}}}^T \vec{\underline{t}}$$

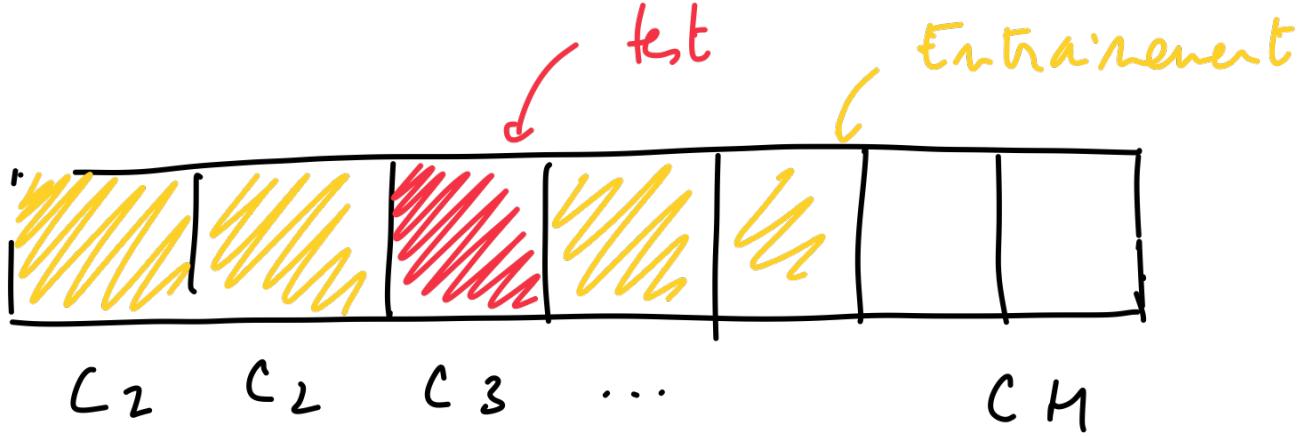
→ Alternative #2 : Sélection du meilleur sous ensemble



Nombre de sous ensembles de taille K choisis parmi un ensemble de taille D

$$\binom{D}{K}$$

→ Evaluation des $\binom{D}{k}$ modèles de taille k par validation croisée



M Compartiments

$C_1, \dots, C_M$

$$|C_i| = \frac{N}{M}$$

Etape #2

$$\rightarrow \ell_{CV}(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - h_\beta^{\mathcal{D}\setminus C(i)}(x^{(i)}) \right)^2$$ où $\mathcal{D}\setminus C(i)$ correspond à l'ensemble des données sans le compartiment contenant la donnée i

Etape #1 → Selectionner sur

Son ensemble de caractéristique $h_\beta^{\mathcal{D}\setminus C(i)}$ → le modèle entraîné sur $\mathcal{D}\setminus C(i)$

→ pour chaque compartiment C_i

$$h_\beta = \underset{\beta}{\arg\min} \frac{M}{(M-1)N} \sum_{i \in \mathcal{D}\setminus C_i} \left(t^{(i)} - h_\beta^{\mathcal{D}\setminus C_i}(x^{(i)}) \right)^2$$