

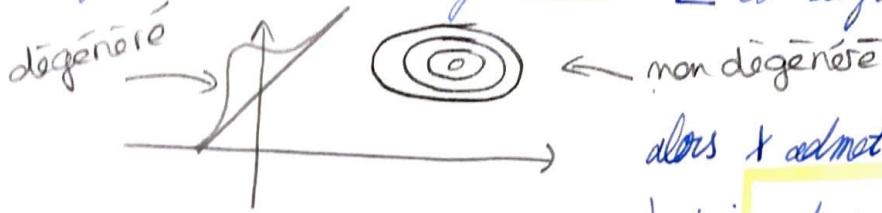
# Classification

On considère un couple de rv \$(X, Y)\$ à valeurs dans \$\mathbb{R}^d \times [\underline{z}, \bar{z}]\$

**Rappel:** Vecteur gaussien.

\$X\$ est un vecteur gaussien si \$a^T X\$ est gaussien  
 $a \in \mathbb{R}^k$  constant p.s.

On définit \$\mu = E[X]\$ et \$\Sigma = \text{Var}(X)\$ où \$\Sigma\$ est symétrique définie  
\$X\$ est gaussien et non dégénéré si \$\Sigma\$ est définie positive (i.e. \$\lambda\_i > 0\$)



alors \$X\$ admet une densité par rapport à  
*à*  $f(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|^{\frac{1}{2}}}} e^{-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}}$

Estimation:  $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$  est un estimateur non biaisé de \$\mu\$ (ETV)

$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)(x_i - \bar{x}_n)^T$  estimateur biaisé de \$\Sigma\$ (ETV)

Si on a \$k\$ échantillons: \$(x\_1, \dots, x\_k)\$ iid copies de \$X\$, alors  
\$(x\_1^T, x\_2^T, \dots, x\_n^T)\$ iid \$\sim \mathcal{N}(\mu, \Sigma)\$  
\$(x\_1^{(k)}, \dots, x\_m^{(k)})\$ iid \$\sim \mathcal{N}(\mu\_k, \Sigma)\$

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i^{(j)} \quad \text{et} \quad \hat{\Sigma} = \frac{1}{\sum_{j=1}^k n_j - k} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_i^{(j)} - \hat{\mu}_j)(x_i^{(j)} - \hat{\mu}_j)^T$$

est un estimateur non biaisé de \$\Sigma\$.

On cherche \$f^\*\$ mesurable tq \$f^\* \in \arg \min\_{f: \mathbb{R} \rightarrow [\underline{z}, \bar{z}]} \text{IP}(Y + f(X))

## Lemma 4 (avis D. Brin)

Classifieur de Bayes :  $g^*(x) \in \arg\max_{1 \leq j \leq c} P(Y=j | X=x)$

Dans le cas  $c=2$  et  $Y$  à valeurs dans  $\{-1, 1\}$ .

$$g^*: x \in \mathbb{R}^d \rightarrow \begin{cases} 1 & \text{si } P(Y=1 | X=x) > P(Y=-1 | X=x) \\ -1 & \text{sinon.} \end{cases}$$

## 2 approches :

1) Plugin (Rota) :  $\hat{g}_n(x) = \begin{cases} 1 & \text{si } \hat{P}(Y=1 | X=x) > \hat{P}(Y=-1 | X=x) \\ -1 & \text{sinon} \end{cases}$

2) Optimisation :  $\hat{g}_n(x) \in \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{Y_i \neq g(x)}$

## R | LDA | LQA

$Y$  à valeurs dans  $\{1, \dots, C\}$ .

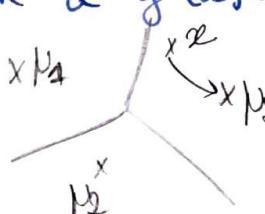
- 1)  $\forall j \in \{1, \dots, C\}$  :  $X | Y=j \sim \mathcal{N}(\mu_j, \Sigma_j)$  avec  $\mu_j \in \mathbb{R}^d$ ,  $\Sigma_j \in \mathbb{R}^{d \times d}$  DP.
- 2)  $\forall j \in \{1, \dots, C\}$  :  $P(Y=j) = \pi_j \in [0; 1[$ .

## Proposition :

Le classifieur de Bayes  $g^*(x) \in \arg\max_{1 \leq j \leq C} P(Y=j | X=x)$ ,

$$\in \arg\max_{1 \leq j \leq C} \log \pi_j - \frac{1}{2} \log |\Sigma_j| - \frac{1}{2} (x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j)$$

cas particulier :  $\pi_1 = \dots = \pi_C$  et  $\Sigma_1 = \dots = \Sigma_C$ , on a  $g^*(x) \in \arg\max_{1 \leq j \leq C} \|x - \mu_j\|^2$ .  
est le classifieur de + perte représentant :



## Démonstration

$$X|Y=j \sim \mathcal{N}(\mu_j, \Sigma_j)$$

$$\text{Formule de Bayes : } \Pr\{Y=j|X=x\} = \frac{\Pr(Y=j) \Pr(X=x|Y=j)}{\Pr(X=x)}$$

$\Pr(X=x)$  — densité marginale

$$\log \Pr(Y=j|X=x) = \text{cste} + \log(\pi_j) - \frac{1}{2} \log \Sigma_j - \frac{1}{2} (x - \mu_j) \Sigma_j^{-1} (x - \mu_j)^T$$

Sait

①  $Y$  à valeurs dans  $\{-1, 1\}$ ,  $g^*(x) = \begin{cases} 1 & \text{si } \Pr(Y=1|X=x) > \Pr(Y=-1|X=x) \\ -1 & \text{sinon} \end{cases}$

$$\textcircled{2} \quad \Sigma_1 = \Sigma_{-1} = \Sigma$$

## Proposition

Sous les hypothèses 1 et 2, le classifieur de Bayes peut s'exprimer :

$$g^*(x) = \text{sign}(w^T x + b), \text{ avec } w = \Sigma^{-1} (\mu_1 - \mu_{-1})$$

$$= \begin{cases} 1 & \text{si } w^T x + b > 0 \\ -1 & \text{si } w^T x + b \leq 0 \end{cases} \quad b = \log \frac{\pi_1}{1-\pi_1} + \frac{1}{2} (\mu_1^T \Sigma^{-1} \mu_1 - \mu_{-1}^T \Sigma \mu_{-1})$$

il g<sup>\*</sup> est un hyperplan

$$\textcircled{1} \quad \uparrow^w \quad \Pr(\mu_1, \Sigma) \quad \Pr(\mu_2, \Sigma)$$

(-1)

L'Analyse linéaire discriminante (LDA) est une implémentation de ce résultat.

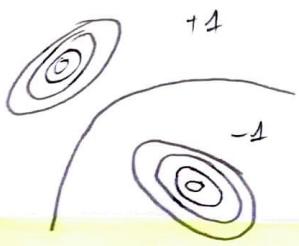
## Proposition

Dans le cas général et sans l'hypothèse 1, le classifieur de Bayes est :

$$g^*(x) = \text{sign}\left(\frac{1}{2} x^T Q x + w^T x + b\right) \quad \text{avec } Q = \Sigma_{-1}^{-1} - \Sigma_1^{-1}$$

$$w = \Sigma_1^{-1} \mu_1 - \Sigma_{-1}^{-1} \mu_{-1}$$

$$b = \log \frac{\pi_1}{1-\pi_1} + \frac{1}{2} (\mu_1^T \Sigma_{-1}^{-1} \mu_1 - \mu_{-1}^T \Sigma_1^{-1} \mu_1)$$



→ Analyse quadratique discriminante

## Remarque d'implémentation (LDA)

$$w = \sum (\mu_i - \mu_{-i}) \text{ avec } \sum = \text{Var}(X|Y=1) = \text{Var}(X|Y=-1) -$$

$w \propto \text{Var}(X)^{-1} (\mu_i - \mu_{-i})$

done  $\frac{\text{E}(X|Y) - \mu_{-i}}{\Delta \mu} \sim \text{Beta}(\alpha \mu_i)$

$\text{Var}(X) = \text{E}[\text{Var}(X|Y)] + \text{Var}(\text{E}[X|Y])$

$\text{E}[X|Y] = \begin{cases} \mu_i & Y=1 \\ \mu_{-i} & Y=-1 \end{cases}$

$\text{Var}(\alpha X + b) = \alpha^2 \text{Var}(X)$

$= \pi_1 \sum + (1-\pi_1) \sum + \underbrace{\pi_1 (1-\pi_1)}_{\alpha} \Delta \mu \Delta \mu^T$  avec  $\Delta \mu = \mu_i - \mu_{-i}$

$= \sum + \alpha \Delta \mu \Delta \mu^T$

$\text{Var}(X) \sum^{-1} = \text{Id} + \alpha \Delta \mu \Delta \mu^T \sum^{-1}$  poly

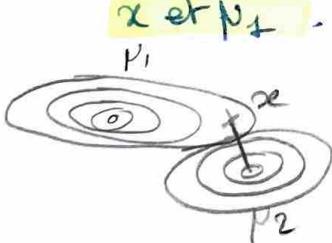
$\text{Var}(X) \sum^{-1} \Delta \mu = \Delta \mu + \alpha \Delta \mu \Delta \mu^T \sum^{-1} \Delta \mu = (1 + \alpha \Delta \mu^T \sum^{-1} \Delta \mu) \Delta \mu$

Donc  $\sum^{-1} \Delta \mu \propto \text{Var}(X) \Delta \mu$  scalaire

$$\pi_1 = \pi_{-1} = \frac{1}{2}, g^*(x) = \text{sign}(w^T x + b), w = \sum^{-1} \Delta \mu$$

$$b = \frac{1}{2} (\mu_{-1}^T \sum^{-1} \mu_{-1} - \mu_1^T \sum^{-1} \mu_1)$$

$$g^*(x) = 1 \iff \underbrace{(x - \mu_1)^T \sum^{-1} (x - \mu_1)}_{\text{distance de Mahalanobis entre } x \text{ et } \mu_1} < \sqrt{(x - \mu_{-1})^T \sum^{-1} (x - \mu_{-1})}.$$



$b_2 \rightarrow \text{donne } \mu_2$

$\pi_1 \rightarrow \text{donne } \mu_1 \text{ car prend en cpt } \Sigma$

$$\| \sum^{-1/2} x - \sum^{-1/2} \mu_1 \|_2^2$$



$\Rightarrow$  norme euclidienne ds un nouvel espace

## Détail pour la régression linéaire

$X|Y = j \sim N(\mu_j, \Sigma) \Rightarrow$  classifieur linéaire

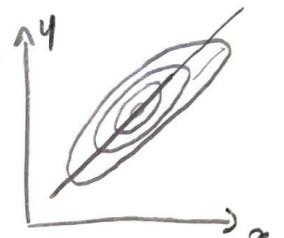
Régression linéaire:  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  fonctionnelle qui minimise  $E_{\mathbb{R}}[ (Y - f(x))^2 ]$

$$f^*(x) = E[Y | X=x]$$

### Proposition.

Soit  $(X_i, Y)$  un couple avec  $\bar{x}$  valeur dans  $\mathbb{R}^d \times \mathbb{R}$  t.q

$$\left( \begin{array}{c} X \\ Y \end{array} \right) \sim \mathcal{D}\mathcal{P} \left( \frac{N}{m}, \left( \begin{array}{c|cc} \Sigma & \mathbb{R}^d \\ \hline \mathbb{R}^d & \sigma^2 \end{array} \right) \right)$$



Si  $w = \Sigma^{-1} \ell$  et  $\sigma^2 = \sigma^2 - \ell^T \Sigma^{-1} \ell$ , alors  $\forall x \in \mathbb{R}^d$ ,  $Y|X=x \sim \mathcal{D}\mathcal{P}(m + w^T(x - \mu), \sigma'^2)$

Donc  $E[Y|X=x] = (\mu - w^T \mu) + w^T x$

### Démonstration

$\mu = w = 0$  sans perte de généralité, Dq  $Y|X=x \sim \mathcal{D}\mathcal{P}(w^T x, \sigma'^2)$ .

$Y - w^T X | X \sim \mathcal{D}\mathcal{P}(0, \sigma'^2)$ .

$$\left( \begin{array}{c} X \\ Y - w^T X \end{array} \right) = \left( \begin{array}{cc} I_d & 0 \\ -w^T & 1 \end{array} \right) \left( \begin{array}{c} X \\ Y \end{array} \right) \sim \mathcal{D}\mathcal{P}(0, \left( \begin{array}{c|c} \Sigma & 0 \\ \hline 0 & \sigma'^2 \end{array} \right))$$

D'où  $Y - w^T X \sim \mathcal{D}(0, \sigma'^2)$  et  $X \perp Y - w^T X$ .

Donc  $Y - w^T X | X=x$  est gaussien de paramètres :

$$\triangleright E[Y - w^T X | X=x] = E[Y - w^T X] = 0$$

$$\hookrightarrow E[Y | X=x] = w^T x$$

$$\triangleright \text{Var}[Y - w^T X | X=x] = \text{Var}(Y - w^T X) = \sigma'^2$$

$$\text{Var}(Y | X=x) = \sigma'^2$$

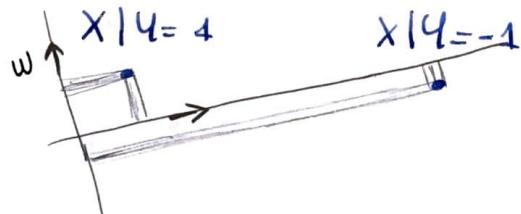
Finallement, on a  $Y | X=x \sim \mathcal{D}(w^T x, \sigma'^2)$ .

## B1 Analyse linéaire discriminante de Fisher

Génova 07

Soit  $(x,y)$  à valeurs dans  $\mathbb{R}^d \times \{\pm 1\}$

$$\text{Var}(E[X|Y]) = \text{SpSp}$$



Quotient de Rayleigh:  $\forall w \neq 0$

$$\pi(w) = \frac{\text{Var}(E[w^T X|Y])}{E[\text{Var}(w^T X|Y)]}$$

interclasse

intraclasse

Adopter un moment d'ordre 2

Proposition

$$\mu_1 = E[X|Y=1]$$

$$\mu_{-1} = E[X|Y=-1]$$

$$\nu = E[X]$$

$$\Sigma_1 = \text{Var}(X|Y=1)$$

$$\Sigma_{-1} = \text{Var}(X|Y=-1)$$

$$\Sigma = \text{Var}(X)$$

$$\begin{aligned} \forall w \neq 0, \quad \pi(w) &= \pi_1(1-\pi_1) \frac{(w^T(\mu_1 - \mu_{-1}))^2}{w^T(\pi_1 \Sigma_1 + (1-\pi_1) \Sigma_{-1})w} \\ &= \pi_1(1-\pi_1) \frac{w^T \Delta w}{w^T \Sigma w} \end{aligned}$$

$$\arg \max_{w \neq 0} \pi(w) = \underset{(\text{range})}{\text{Vect.}} \left\{ \Sigma^{-1}(\mu_1 - \mu_{-1}) \right\} \text{dy}$$

Remarque Si  $\Sigma_1 = \Sigma_{-1}$ , alors  $\Sigma^{-1}(\mu_1 - \mu_{-1})$  est la direction que LDA (mais on ne connaît pas le sens!).

Démonstration

$$\begin{aligned} \forall w \neq 0, \quad \pi(w) &= \pi_1(1-\pi_1) \frac{w^T \Delta w}{w^T \Sigma w}, \quad \Delta = (\mu_1 - \mu_{-1})(\mu_1 - \mu_{-1})^T \\ \boxed{\forall \lambda \neq 0, \quad \pi(\lambda w) = \pi(w)}. \end{aligned}$$

On peut se restreindre à maximiser  $\pi(w)$   
 $w \in \mathbb{R}^d, \|w\|=1$

Observations:  $\nabla r(w)$ ,  $\|w\|=1$

a)  $r(w) \geq 0$

$$\text{b)} r(w) = \pi_1(1-\pi_1) \frac{w^T \pi w}{\|w\|^2} - \frac{\|w\|^2}{w^T \Sigma w}$$

$$\leq \pi_1(1-\pi_1) \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$$

$$\leq \pi_1(1-\pi_1) \frac{\|\mu_1 - \mu_2\|^2}{\lambda_{\min}(\Sigma)}$$

$\Sigma$  supposée inversible  $\Rightarrow \lambda \neq 0$

c)  $r$  est différentiable en  $\mathcal{S}$

$r$  est  $C^1$  sur la sphère unité qui est compacte. Donc  $r$  admet et atteint son maximum et son minimum sur cette sphère.

Recherche des pts critiques.



$$\nabla r|_{\mathcal{S}(w)} = 0 \Leftrightarrow P \nabla r(w) = 0, \text{ Projecteur } \perp \text{ sur } \text{Vect}\{w\}^\perp$$

$$\Leftrightarrow \nabla r(w) = 0$$

$$\mathcal{S} = \{w \mid \|w\|=1\}$$

$$\nabla r(w)^T w = 0$$

$$\nabla w \neq 0: r(w) = \pi_1(1-\pi_1) \frac{w^T \pi w}{w^T \Sigma w}$$

$$\nabla r(w) = \frac{2(w^T \Sigma w) \pi w - 2w^T \pi w \Sigma w}{(w^T \Sigma w)^2} = 0$$

$$\Leftrightarrow (w^T \Sigma w) \pi w = (w^T \pi w) \Sigma w$$

$$(w^T \Sigma w) \underbrace{\Delta p}_{\text{scalaire}} \underbrace{s p^T w}_{} = (w^T \underbrace{\Delta p}_{\text{scalaire}} \underbrace{s p^T w}_{}) \Sigma w \text{ en notant } \Delta p = \mu_1 - \mu_2$$

$$(\Delta p^T w)(w^T \Sigma w) \Delta p = (w^T \pi w) \Sigma w = (w^T \Delta N)^2 \Sigma w$$

1)  $w \perp \Delta p \Leftrightarrow \nabla r(w) = 0 \quad (w^T \Delta p = 0)$

2) si  $w$  n'est pas  $\perp \Delta p$ , alors  $w \propto \Sigma^{-1} \Delta p := w_0$

$$r(w_0) > 0$$

Donc  $r$  atteint son maximum pour Vect $\{ \Sigma^{-1} \Delta p \} \setminus \{0\}$

□

Thm 8.

$g^*$  est un classifieur de Bayes car  $g^*(x) > 0 \iff \Pr[Y=1|X=x] > \Pr[Y=-1|X=x]$ .

En sommant :

$$\forall (x, y) \in \mathbb{R}^d, \Pr[Y=y|X=x] = \frac{1}{1 + e^{-y(w^T x + b^*)}}$$

??

Sans perte de généralité  $b^* = 0$ , on pose :

$$\psi := (x, y, w) \mapsto \log(1 + e^{-y w^T x})$$

Alors  $\forall w \in \mathbb{R}^d$  : (TCDL)

$$\star 0 \leq \mathbb{E}[\psi(X, Y, w)] \leq \mathbb{E}[1 - Y w^T X] + 1 = \mathbb{E}[1 w^T X] + 1 < +\infty \text{ car } X \in L^1$$

\*  $\psi$  différentiable en  $w$ .

$$\star \nabla_w \psi(x, y, w) = -Y \frac{e^{-Y w^T X}}{1 + e^{-Y w^T X}} X \text{ et } \|\nabla_w \psi(x, y, w)\|_2 \leq \left\| \frac{e^{-Y w^T X}}{1 + e^{-Y w^T X}} X \right\|_2 \in L^1$$

Integral Leibniz

$$\begin{cases} \text{Si } X(t) \in L^1 \\ t \mapsto X(t) \text{ différentiable,} \\ \|X(t)\| \leq C \in L^1 \end{cases} \quad \left\{ \begin{array}{l} \text{Alors } \nabla \mathbb{E}[X(t)] = \mathbb{E}[\nabla X(t)] \\ \nabla \mathbb{E}[X(t)] \rightarrow \min w^* \end{array} \right.$$

$$\text{D'où } \nabla_w \mathbb{E}[\psi(X, Y, \cdot)](w) = \mathbb{E}[\nabla_w \psi(X, Y, w)] = \mathbb{E}[\mathbb{E}[\nabla_w \psi(X, Y, w)|X]]$$

$$\text{De plus, } \mathbb{E}[\nabla_w \psi(X, Y, w)|X] = \Pr[Y=1|X] \nabla_w \psi(X, 1, w) +$$

$$\Pr[Y=-1|X] \nabla_w \psi(X, -1, w)$$

$$= \frac{1}{1 + e^{-w^T X}} \left( -\frac{e^{-w^T X}}{1 + e^{-w^T X}} X \right) + \frac{1}{1 + e^{w^T X}} \left( \frac{e^{w^T X}}{1 + e^{w^T X}} X \right)$$

$$= \frac{1}{1 + e^{-w^T X}} \left( -\frac{e^{-w^T X}}{1 + e^{-w^T X}} X \right) + \frac{e^{-w^T X}}{1 + e^{-w^T X}} \left( \frac{1}{1 + e^{-w^T X}} X \right)$$

$$\text{Donc } \mathbb{E}[\nabla_w \psi(X, Y, w)|X] = 0 \text{ et } \nabla_w \mathbb{E}[\psi(X, Y, \cdot)](w^*) = 0$$

Comme  $\psi$  est convexe en  $w$ ,  $w^*$  est un minimiseur de  $\mathbb{E}[\psi(X, Y, \cdot)]$ . □

## Preuve Prop 9

On construit  $f^*$  point par point. Pour tout  $x \in \mathbb{R}^d$ , soit  $\varphi: a \mapsto \underset{\in \mathbb{R}}{\text{I.E.}}$  Preuve / Partie

$\ell$  différentiable et convexe, tjs au-dessus de sa tangente,  $a \in \mathbb{R} \mapsto \ell(a) + \ell'(a)(a - \ell(a)) = \eta(x)\ell(a) + (1-\eta(x))\ell'$   
 $\ell'(a) \leq 0$  car  $\ell \downarrow$  sur  $\mathbb{R}^+$  et décroissante).

$$\lim_{a \rightarrow +\infty} \ell(a) = +\infty$$

Comme  $\eta(x) \in (0,1)$ ,  $\ell \geq 0$ ,  $\varphi$  est coercive,  $\lim_{a \rightarrow +\infty} \varphi(a) = \lim_{a \rightarrow -\infty} \varphi(a) = +\infty$

Donc  $\varphi$  admet un minimum  $\in \mathbb{R}$ . On choisit  $f^*(x)$  tel que  $\varphi(a)$

Pour la fonction  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  mesurable, par construction de  $f^{**}$ ,

$$\text{I.E.}[\ell(Yf(x)) | X=x] \geq \text{I.E.}[\ell(Yf^*(x)) | X=x]. \text{ Par linéarité,}$$

$$\text{I.E.}[\ell(Yf(x))] = \text{I.E.}[\text{I.E.}[\ell(Yf(x)) | X]] \geq \text{I.E.}[\text{I.E.}[\ell(Yf^*(x)) | X=x]].$$

Reste à prouver que  $g^* := \text{sign } f^*(x)$  est un classifieur de Bayes.

Pour tout  $x \in \mathbb{R}^d$ ,  $f^*(x)$  minimum d'une fct  $\varphi$  convexe et différentiable,

$$\text{on a: } \varphi'(\ell(f^*(x))) = \eta(x)\ell'(f^*(x)) - (1-\eta(x))\ell'(-f^*(x)) = 0 \text{ i.e. } \frac{\eta(x)}{1-\eta(x)} = \frac{\ell'(-f^*(x))}{\ell'(f^*(x))} \neq 0.$$

Alors,  $g^*(x) = 1 \Leftrightarrow f^*(x) > 0$ .

$$\begin{aligned} -\ell'(x) &< \ell'(f^*(x)) \\ \ell'(-f^*(x)) &< \ell'(f^*(x)) \quad \text{I.E. } \ell' \text{ sur } \mathbb{R}^+ \text{ croissante} \\ \frac{\ell'(-f^*(x))}{\ell'(f^*(x))} &> 1 \quad \ell'(f^*(x)) < 0 \end{aligned}$$

$$\frac{\eta(x)}{1-\eta(x)} > 1$$

$$\eta(x) > 1 - \eta(x) \Leftrightarrow \eta(x) > \frac{1}{2} \Leftrightarrow \text{P}\{Y=1 | X=x\} > \text{P}\{Y=-1 | X=x\}$$

Et donc  $g^*$  est un classifieur de Bayes.

Nouvelle Propriété p.21

Pour tout  $t \in \mathbb{N} - \{t\}$  et  $(w, g) \in \mathbb{R}_+ \times \mathcal{C}$ .

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n e^{-y_i(f_{t-1}(x_i) + w g(x_i))} &= \frac{e^{-w}}{n} \sum_{i=1}^n e^{-y_i f_{t-1}(x_i)} \mathbb{1}_{y_i = g(x_i)} \\ &\quad + \frac{e^w}{n} \sum_{i=1}^m e^{-y_i f_{t-1}(x_i)} \mathbb{1}_{y_i \neq g(x_i)} \\ &= \frac{e^{-w}}{n} \sum_{i=1}^m e^{-y_i f_{t-1}(x_i)} + \\ &\quad \underbrace{\frac{e^w - e^{-w}}{n} \sum_{i=1}^m e^{-y_i f_{t-1}(x_i)} \mathbb{1}_{y_i \neq g(x_i)}}_{\text{argmin } g \in \mathcal{C} \text{ de } \frac{1}{n} \sum_{i=1}^m Q_t(i) \mathbb{1}_{y_i \neq g(x_i)}} \end{aligned}$$

$$\Rightarrow \text{argmin}_{g \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^m Q_t(i) \mathbb{1}_{y_i \neq g(x_i)} \quad \textcircled{1}$$

$\Downarrow w \Rightarrow$  on pose  $\tilde{g} \in \text{argmin}_{g \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^m Q_t(i) \mathbb{1}_{y_i \neq g(x_i)}$

Posons,  $\varphi: w \mapsto \frac{1}{n} \sum_{i=1}^m e^{-y_i(f_{t-1}(x_i) + w g(x_i))}$  est différentiable et

Si  $\tilde{g} \neq 0$ ,  $\varphi$  est décroissante si  $w < \frac{1}{2} \log \left( \frac{1-\tilde{\epsilon}_t}{\tilde{\epsilon}_t} \right)$  où  $\tilde{\epsilon}_t = \sum_{i=1}^m Q_t(i) \mathbb{1}_{y_i \neq g(x_i)}$

Donc  $\varphi$  minimisée sur  $\mathbb{R}$  en  $\tilde{w} = \frac{1}{2} \log \left( \frac{1-\tilde{\epsilon}_t}{\tilde{\epsilon}_t} \right)$ . mino au +∞ et ≥ 0 par hypothèse.

De plus,  $\tilde{g}$  est donc  $\tilde{w}$  est un minimum aussi.

En regroupant les termes, pour tout  $(w, g) \in \mathbb{R}_+ \times \mathcal{C}$ ,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^m e^{-y_i(f_{t-1}(x_i) + w g(x_i))} &\geq \frac{1}{n} \sum_{i=1}^m e^{-y_i(f_{t-1}(x_i) + w \tilde{g}(x_i))} \\ &\geq \frac{1}{n} \sum_{i=1}^m e^{-y_i(f_{t-1}(x_i) + \tilde{w} \tilde{g}(x_i))} \end{aligned}$$

Donc on peut considérer,

$$g_t \in \text{argmin}_{g \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^m Q_t(i) \mathbb{1}_{g(x_i) \neq y_i} \quad \text{et}$$

$$w_t = \frac{1}{2} \log \left( \frac{1-\tilde{\epsilon}_t}{\tilde{\epsilon}_t} \right)$$

# Perceptron

$$\partial_{\text{Per}}(U) = \frac{1}{n} \sum_{j=1}^n e^{-w_j y_j g_t(x_j)} / z_j$$

$$= \frac{1}{n} \frac{e^{-\sum_{j=1}^n w_j y_j g_t(x_j)}}{\sum_{j=1}^n z_j} = \frac{e^{-\sum_{j=1}^n w_j y_j f_{t-1}(x_j)}}{n \sum_{j=1}^n z_j}$$

We minimise our function strictly with respect to  $w$ .

$$0 = \sum_{i=1}^m y_i g_t(x_i) e^{-y_i (f_{t-1}(x_i) + w_t g_t(x_i))}$$

$$= \sum_{i=1}^m e^{-y_i f_{t-1}(x_i)} e^{-w_t} - \sum_{i=1}^m e^{-y_i f_{t-1}(x_i)} e^{w_t}$$

$$= \sum_{\substack{i < j < m \\ y_i g_t(x_i) = 1}} e^{-y_i f_{t-1}(x_i)} - e^{2w_t} \sum_{\substack{i < j < m \\ y_i g_t(x_i) = -1}} e^{-y_i f_{t-1}(x_i)}$$

$$w_t = \frac{1}{2} \log \left( \frac{\sum_{y_i g_t(x_i) = 1} e^{-y_i f_{t-1}(x_i)}}{\sum_{y_i g_t(x_i) = -1} e^{-y_i f_{t-1}(x_i)}} \right)$$

$$= \frac{1}{2} \log \left( \frac{\sum_{i=1}^m e^{-y_i f_{t-1}(x_i)} \text{ if } y_i = g_t(x_i)}{\sum_{i=1}^m e^{-y_i f_{t-1}(x_i)} \text{ if } y_i \neq g_t(x_i)} \right)$$

$$= \frac{1}{2} \log \left( \frac{\sum_{i=1}^m \frac{1}{z_i} \prod_{j \neq i} D_t(j) \text{ if } y_i = g_t(x_i)}{\sum_{i=1}^m \frac{1}{z_i} \prod_{j \neq i} D_t(j) \text{ if } y_i \neq g_t(x_i)} \right)$$

$$= \frac{1}{2} \log \left( \frac{\sum_{j=1}^m D_t(j) \text{ if } y_j = g_t(x_j)}{\sum_{j=1}^m D_t(j) \text{ if } y_j \neq g_t(x_j)} \right)$$

$$= \frac{1}{2} \log \left( \frac{1 - \epsilon_t}{\epsilon_t} \right)$$

$$z_t = \sum_{i=1}^n D_t(i) e^{-w_t y_i g_t(x_i)} = \sum_{\substack{i < j < m \\ y_i g_t(x_i) = 1}} D_t(i) e^{-w_t} + \sum_{\substack{i < j < m \\ y_i g_t(x_i) = -1}} D_t(i) e^{w_t}$$

$$= (1 - \epsilon_t) e^{-w_t} + \epsilon_t e^{w_t}$$

$$= (1 - \epsilon_t) \sqrt{\frac{\epsilon_t}{1 - \epsilon_t}} + \epsilon_t \sqrt{\frac{1 - \epsilon_t}{\epsilon_t}}$$

$$= 2 \sqrt{\epsilon_t (1 - \epsilon_t)}$$

## Preuve Proposition 18

$$\forall i \in \{1, \dots, m\}, y_i (\langle w_n^*, x_i \rangle_{\ell_2} + b_n^*) \geq 1 > 0$$

Comme  $\exists i \neq j \in \{1, \dots, m\}$ ,  $y_i \neq y_j$ , donc  $w_n^* \neq 0$ , Alors  $(w_n^*, b_n^*)$  admissible pour (P5)

En résulte, on voit que  $\min_{1 \leq i \leq m} y_i (\langle w_n^*, x_i \rangle_{\ell_2} + b_n^*) \geq 1$  donc  $\nu(w_n^*, b_n^*) = \frac{1}{\|w_n^*\|_{\ell_2}}$

Alors, pour tout  $(v, a) \in \mathbb{R}^d$  tel que  $\forall i \in \{1, \dots, m\}, y_i (v^T x_i + a) \geq 0$ .

Sait  $r = \min_{1 \leq i \leq m} y_i (v^T x_i + a) = \min_i |v^T x_i + a|$ , alors on :

- $r=0$  et  $\nu(v, a) = \frac{r}{\|v\|_{\ell_2}} = 0 \leq \nu(w_n^*, b_n^*)$  ou

- $r \neq 0$  et en notant  $(w, b) = \left(\frac{r}{\sigma}, \frac{a}{\sigma}\right)$  , on a  $\forall i \in \{1, \dots, m\}$

Et donc  $(w, b)$  admissible pour (P4). Depuis,  $y_i (w^T x_i + b) \geq 1$ ,

$$\nu(v, a) = \frac{r}{\|v\|_{\ell_2}} = \frac{1}{\|w\|_{\ell_2}} \leq \frac{1}{\|w_n^*\|_{\ell_2}} = \nu(w_n^*, b_n^*)$$

Ce qui prouve que  $(w_n^*, b_n^*)$  est une solution pour (P5).

RKHS

$(X_1, Y_1) \dots (X_n, Y_n) \in \mathbb{R}^d \times \{-1, 1\}$   
iid

Régression logistique:  $\underset{\omega}{\operatorname{argmin}} \frac{1}{n} \sum \log(1 + e^{-Y_i(\omega^\top X_i + b)})$

SVR linéaire:  $\underset{\omega}{\operatorname{argmin}} \frac{1}{n} \sum \max(0, 1 - Y_i f(X_i)) + \frac{d}{2} \|w\|_2^2$

$$f = \omega^\top x + b$$

$f \in \mathcal{H}$  RKHS.

RKHS définit par son noyau:

$$k: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R} \quad (\mathcal{G}, \langle \cdot, \cdot \rangle_{\mathcal{G}}) \quad \phi: \mathbb{R}^d \rightarrow \mathbb{R}$$

$$k(x, x') = \langle \phi(x), \phi(x') \rangle$$

cas linéaire  $\phi: \text{Id}$

cas polynomial  $\phi: \text{polynomiale}$

cas exponentiel:  $k(x, x') = e^{-\gamma \|x - x'\|_2^2}$   $\phi?$  X

Définition

RKHS  $\mathcal{H}$  de noyau  $k: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ , est l'unique espace tq.

$$\star k(\cdot, z) \in \mathcal{H}$$

$$\star \forall f \in \mathcal{H}, \forall x \in \mathbb{R}^d: f(x) = \langle f, k(\cdot, z) \rangle$$

$$\mathcal{H} = \{x \mapsto \langle w, \phi(x) \rangle, w \in \mathcal{G}\}.$$

$$= \left\{ \sum_{i=1}^{+\infty} \alpha_i k(x_i, z), (x_i)_i \in \mathbb{R}^d, (\alpha_i)_i \subset \mathbb{R}^d, \sum_{i=1}^{+\infty} \alpha_i^2 k(x_i, z_i) < +\infty \right\}.$$

Cas exponentiel:

$$f \in \mathcal{H}; f(x) = \sum_{i=1}^{+\infty} \alpha_i e^{-r \|x - x_i\|^2} . \text{ fonctions de décision non linéaires}$$

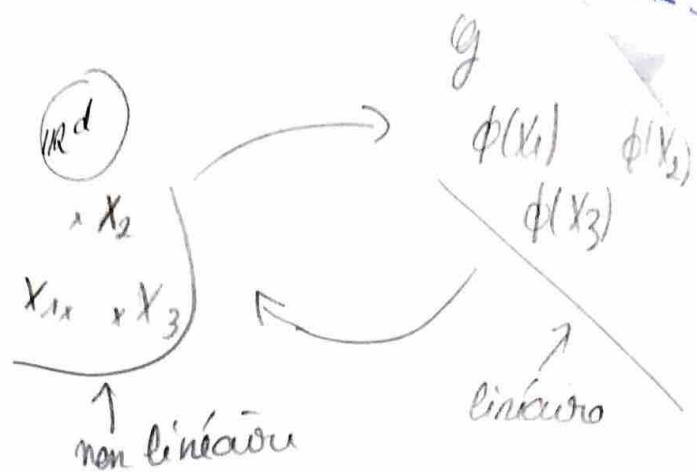
## Astuce de noyan

$$(x_i, y_i) \quad \forall i \in [m]$$

$$k: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R} \Leftrightarrow (\mathcal{G}, \langle \cdot, \cdot \rangle_{\mathcal{G}})$$

$$\mathcal{H} = \{x \mapsto \langle w, \phi(x) \rangle_{\mathcal{G}}\}$$

$$\{\phi(x_i), y_i\} \quad \forall i \in [m]$$



## SVM linéaire

Classifieur:  $x \mapsto \text{sign}(\langle w, \phi(x) \rangle_{\mathcal{G}} + b)$

$$\begin{aligned} & \text{minimiser}_{w \in \mathcal{G}, b \in \mathbb{R}} \frac{1}{2} \|w\|_{\mathcal{G}}^2 + C \sum_{i=1}^n (1 - y_i (\langle w, \phi(x_i) \rangle_{\mathcal{G}} + b))_+ \\ & \quad \quad \quad C \sum \xi_i \end{aligned}$$

$$\text{Sc}, \forall i \in [1, n], \xi_i \geq 0$$

$$\text{On pose } f = \langle w, \phi(\cdot) \rangle_{\mathcal{G}} \quad y_i (\langle w, \phi(x_i) \rangle_{\mathcal{G}} + b) \geq 1 - \xi_i$$

$$\begin{aligned} & \text{minimiser}_{f \in \mathcal{H}, b \in \mathbb{R}, \xi \in \mathbb{R}^n} \inf_{w \in \mathcal{G}} \left\{ \frac{1}{2} \|w\|_{\mathcal{G}}^2 \right\} + C \sum_{i=1}^n \xi_i = \frac{1}{2} \|f\|_{\mathcal{H}}^2 + C \sum_{i=1}^n \xi_i \\ & \quad \quad \quad = \frac{1}{2} \|f\|_{\mathcal{H}}^2 \quad \text{st} \end{aligned}$$

st

$$\xi_i \geq 0$$

$$y_i (f(x_i) + b) \geq 1 - \xi_i$$

$$\text{et } \mathcal{H} = \left\{ \sum_{i=1}^{+\infty} \alpha_i k(\cdot, x_i) \right\}$$

## Définition du représentant

$\mathcal{H}$ : noyau et  $\mathcal{H}$  son RKHS.

$\phi(x_i, y_i)$  fusion avec  $(x_i, y_i)$  à valeurs dans  $\mathbb{R}^d \times \mathbb{R}$

$\psi: \mathcal{H} \rightarrow \mathbb{R}$  croissante

minimiser  $\psi(\|\mathbf{h}\|_{\mathcal{H}}) + \sum_{\substack{h \in \mathcal{H} \\ b \in \mathbb{R}}} c(y_1, \dots, y_n, h(x_1) + b, \dots, h(x_m) + b)$

Si le pb d'optimisation a au moins une solution  $(\bar{h}, \bar{b}^*) \in \mathcal{H} \times \mathbb{R}$ , alors:  
on peut construire, fini (pas besoin de  $\phi$ )

$$\begin{aligned} h^* &= \sum_{i=1}^n \alpha_i k(\cdot, x_i) \quad \text{connu} \\ H &= \left\{ \sum_{i=1}^n \alpha_i k(\cdot, x_i) \right\} \quad \text{inconnu} \end{aligned}$$

SVM: minimiser  $\frac{1}{2} \|\mathbf{h}\|_{\mathcal{H}}^2 + C \sum_{i=1}^n (1 - y_i(h(x_i) + b))_+$

$$\begin{aligned} h &= \left\{ \sum_{i=1}^n \alpha_i k(\cdot, x_i), \alpha \in \mathbb{R}^n \right\} \\ \|\mathbf{h}\|_{\mathcal{H}}^2 &= \langle h, h \rangle_{\mathcal{H}} = \sum_{1 \leq i, j \leq n} \underbrace{\alpha_i \alpha_j \langle k(\cdot, x_i), k(\cdot, x_j) \rangle}_{\mathcal{H}} \end{aligned}$$

$$h(x_i) = \sum_{i=1}^n \alpha_i k(x_i, x_i) = (\mathbf{h} \alpha)_i$$

$$\begin{aligned} \|\mathbf{h}\|_{\mathcal{H}}^2 &= \langle h, h \rangle_{\mathcal{H}} = \sum_{1 \leq i, j \leq n} \alpha_i \alpha_j \underbrace{\langle k(\cdot, x_i), k(\cdot, x_j) \rangle}_{\mathcal{H}} \\ &= \alpha^T K \alpha, \text{ avec } K = (k(x_i, x_j))_{1 \leq i, j \leq n}. \end{aligned}$$



Def RKHS:  
 $1/k(\cdot, x) \in \mathcal{H}$   
 $2/f(\alpha) = \langle f, k(\cdot, x) \rangle$

1)  $(\phi(x_i, y_i))_{1 \leq i \leq n} \rightarrow$  estimateur dans l'espace de redescription  $\mathcal{G}$ .

2) Construction et évaluation de l'estimateur est indépendant de  $\phi \rightarrow$  on n'utilise que  $\mathcal{H}$ !

Démonstration .

$$\min_{\begin{array}{l} h \in H \\ b \in \mathbb{R} \end{array}} \mathcal{V}(Uh \|_H) + L(\underbrace{h(x_i) + b}_{\text{Sign}(h(x_i) + b)} \dots)$$

argme  
Li. Ch. 6/1

On suppose que  $\exists (\bar{h}, \bar{b}^*) \in H \times \mathbb{R}$  solutions du problème.

On note  $V = \text{Vect}\{h(\cdot, x_1), \dots, h(\cdot, x_n)\}$  un sous espace de  $H$ .

$$H = V \oplus V^\perp$$

$\exists (h^*, h_\perp) \in \bar{h} = h^* + h_\perp$  avec  $h^* \in V$  et  $h_\perp \in V^\perp$ .

$$\text{Pythagore : } \|h\|_H^2 = \|h^*\|_H^2 + \|h_\perp\|_H^2$$

$$\begin{aligned} & \forall (h, b) \in H \times \mathbb{R} : \mathcal{V}(Uh \|_H) + L(h(x_i) + b) \xrightarrow{\begin{array}{l} h^*(x_i) + h_\perp(x_i) = h^*(x_i) \\ + \langle h_\perp, h(\cdot, x_i) \rangle = 0 \end{array}} \\ & \geq \mathcal{V}(Uh^* \|_H) + L(\bar{h}(x_i) + \bar{b}^*) \text{ car } (\bar{h}, \bar{b}^*) \text{ est solution} \\ & \geq \mathcal{V}(Uh^* \|_H) + L(h^*(x_i) + b) \text{ car } \forall b. \end{aligned}$$

Donc  $(h^*, \bar{b}^*)$  est solution.

$$h^* \in V \iff \exists \alpha \in \mathbb{R}^n : h^* = \sum_{i=1}^n \alpha_i h(\cdot, x_i)$$

Li.  $V$  est strictement croissante :  $(\bar{h}, \bar{b}^*), (h^*, \bar{b}^*)$  sont solutions, donc :

$$\mathcal{V}(Uh \|_H) + L(\bar{h}(x_i) + \bar{b}^*) = \mathcal{V}(Uh^* \|_H) + L(h^*(x_i) + \bar{b}^*)$$

$$\begin{aligned} \Rightarrow \mathcal{V}(Uh \|_H) &= \mathcal{V}(Uh^* \|_H) \Rightarrow \|h\|_H = \|h^*\|_H \\ &\quad \xrightarrow{L(h(x_i) + \bar{b}^*)} \\ &\quad \Rightarrow h_\perp = 0 \\ &\quad \Rightarrow h = h^* \in V. \end{aligned}$$

□

range

Si  $(\alpha, b) \mapsto V(\alpha h(x)) + L(h(x) + b -)$  est strictement convexe, et si  $V$  est strictement croissante, alors si une solution existe, elle est unique et s'écrit  $\sum_{i=1}^n \alpha_i h(\cdot, x_i) = h^*$ .

⚠  $\alpha \in \mathbb{R}^n$  n'est pas nécessairement unique.

$$VAM^2 = \alpha^T h \alpha, \quad K = (h(x_i, x_j))_{1 \leq i, j \leq n}.$$

si  $\text{rg}(K) < n$ .  
pas unique.

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \alpha^T h \alpha + c \sum_{i=1}^n [1 - y_i ((h \alpha)_i + b)]_+ \quad K = (h(x_i, x_j))_{1 \leq i, j \leq n}$$

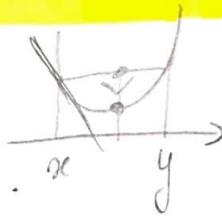
st  $\{\gamma_i\}_{i=1}^n \geq 0$ .

$$y_i ((h \alpha)_i + b) \geq 1 - \gamma_i$$

## Dualité Lagrangienne.

### • Définition

$f: \mathbb{R}^d \rightarrow \mathbb{R}$  convexe si  $\forall x, y \in \mathbb{R}^d, \forall t \in (0, 1),$



$$f(tx + (1-t)y) \leq t f(x) + (1-t) f(y).$$

### Proposition

Si  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  est différentiable, alors  $f$  est convexe :

$$\forall y \in \mathbb{R}^d : f(y) \geq f(x) + \nabla f(x)^T (y - x).$$

### Démonstration

$$\nabla f(x)^T (y - x) = \lim_{\substack{t \rightarrow 0 \\ t \in \mathbb{C}^d}} \frac{f(x + t(y - x)) - f(x)}{t}.$$

$$= \lim_{\substack{t \rightarrow 0 \\ t \in \mathbb{R}, t \neq 0}} \frac{f(ty + (1-t)x) - f(x)}{t} \xrightarrow{f(x) = 0} f(y) - f(x)$$

minimun  
local

## Théorème : Règle de Fermat

$f: \mathbb{R}^d \rightarrow \mathbb{R}$  conv et différentiable, alors :

$$x^* \in \underset{\text{yeux}}{\operatorname{arg\min}} f(y) \Leftrightarrow \nabla f(x^*) = 0$$

Démonstration.

$\Leftarrow$  Si  $\nabla f(x^*) = 0$ , alors  $\forall y \in \mathbb{R}^d$ ,  $f(y) \geq f(x^*) + \nabla f(x^*)^T (y - x^*)$   
 Donc  $x^*$  minimise de  $f$ .

$\Rightarrow$  Si  $\nabla f(x^*) \neq 0$ . Donc  $\|\nabla f(x^*)\| > 0$ ,  $\nabla f(x^*)^T \nabla f(x^*) > 0$

$$\lim_{\substack{t \rightarrow 0 \\ t < 0}} \frac{f(x^* + t \nabla f(x^*)) - f(x^*)}{t} > 0$$

$$\exists t < 0 \text{ tq } \frac{f(x^* + t \nabla f(x^*)) - f(x^*)}{t} > 0$$

$$\Rightarrow f(x^* + t \nabla f(x^*)) - f(x^*) < 0$$

$$\underbrace{f(x^* + t \nabla f(x^*))}_{\bar{x}} < f(x^*)$$

Donc  $x^*$  n'est pas un minimum de  $f$ .

Remarque

Tout minimum local d'une fonction  $f$  conv est aussi minimum global.

minimize  $f(x)$

$$(P) \quad \text{st} \quad \begin{aligned} g_i(x) &\leq \lambda_i \quad i=1 \dots n \\ h_j(x) &= 0 \quad \forall j \in 1 \dots m \end{aligned}$$

$f, g_i, h_j$  sont croissantes. On note  $\mathcal{C} = \{x \in \mathbb{R}^d \mid g_i(x) \leq 0, h_j(x) = 0 \quad \forall i, j\}$ .

$$f(x) + X_C(x) \quad \text{et} \quad X_C(x) = \begin{cases} 0 & \text{si } x \in \mathcal{C} \\ +\infty & \text{si } x \notin \mathcal{C} \end{cases}$$

→ pts intérieurs  
→ formulation variationnelle

## Définition

On appelle lagrangien du pb d'optimisation,

$$L(x, \lambda, \nu) = f(x) + \sum_i g_i(x) + \sum_j \nu_j h_j(x).$$

## Propriétés

$$1. \forall x \in \mathbb{R}^d, \sup_{\lambda \geq 0, \nu} L(x, \lambda, \nu) = \begin{cases} +\infty & \text{si } x \notin \mathcal{C} \\ f(x) & \text{sinon} \\ f(x) + X_C(x) \end{cases}$$

$$2. \inf_{x \in \mathcal{C}} f(x) = \underbrace{\inf_{x \in \mathbb{R}^d} \sup_{\lambda \geq 0, \nu} L(x, \lambda, \nu)}_{\inf_{x \in \mathcal{C}} f(x) = \sup_{\lambda \geq 0, \nu} D(\lambda, \nu)} = \sup_{\lambda \geq 0, \nu} \inf_{x \in \mathcal{C}} L(x, \lambda, \nu)$$

## Définition

La fonction dual du pb d'optimisation (P) est.

$$D(\lambda, \nu) = \inf_{x \in \mathcal{C}} L(x, \lambda, \nu).$$

→ pas unicité  
→ dépend de la formulation du problème

## Proposition Dualité faible

$$\sup_{\lambda \geq 0, \nu \in \mathbb{R}^m} D(\lambda, \nu) \leq \inf_{x \in \mathcal{C}} f(x);$$

primal.

$$\inf_{\lambda \geq 0, \nu} \sup_{x \in \mathcal{C}} L$$

si  $\forall (\lambda, \nu) \in \mathbb{R}_+^m \times \mathbb{R}^m, \forall x \in \mathcal{C}, D(\lambda, \nu) \leq f(x)$

Problème d'optimisation

Dual.

## Décap Astuce du noyau

$k: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  SPP,  $\phi: \mathbb{R}^d \rightarrow \mathcal{G}$  (pas unicité)

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{G}}$$

$$\{\phi(x_i), y_i\}_{i=1}^n$$

1/ Construction de l'estimateur

2/ Évaluation de l'estimateur.

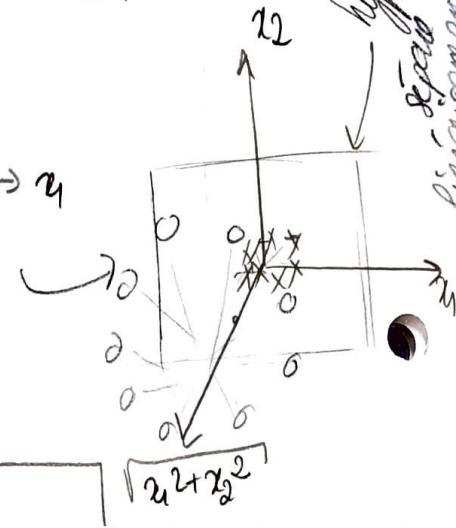
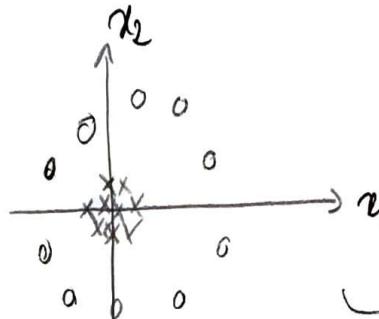
$$\phi \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ \sqrt{x_1^2 + x_2^2} \end{pmatrix}$$

Définition RKHS :

Hilberts de  $k$ ,

$$\forall k(\cdot, x) \in \mathcal{H}, \forall x \in \mathbb{R}^d$$

$$\text{et } \forall f \in \mathcal{H}, f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}, \forall x \in \mathbb{R}^d.$$



$$\left\{ \begin{array}{l} \text{minimiser } f(x) \\ \text{st } g_i(x) \leq 0 \quad \forall i \in \{1, \dots, n\} \\ h_j(x) = 0 \quad \forall j \in \{1, \dots, m\}. \end{array} \right. \quad f, g_i, h_j \text{ convexes et différentiables.}$$

↪ Langrangien :  $L(x, d, v) = f(x) + \sum_i d_i g_i(x) + \sum_j v_j h_j(x)$ .

$$\forall x \in \mathcal{E}, f(x) = \sup L(x, d, v)$$

$$\inf_{x \in \mathcal{E}} \sup_{d, v} L(x, d, v) = \inf_{x \in \mathcal{E}} f(x).$$

$$\begin{aligned} \forall (\lambda, \nu) \text{ admissibles}, \forall x \in \mathcal{E}. D(\lambda, \nu) &= \inf_x L(x, \lambda, \nu) \leq L(x, \lambda, \nu) \\ &= f(x) + \underbrace{\sum_i \lambda_i g_i(x)}_{\leq 0} + \underbrace{\sum_j \nu_j h_j(x)}_{\geq 0} = 0 \end{aligned}$$

Le problème d'optimisation est un problème convexe.

- \*  $f$  est convexe.
- \*  $g$  est convexe.  $\left\{ \mathcal{E} \text{ est un ensemble convexe} \right.$
- \*  $h$  est affine.  $\Delta$  condition forte mais indispensable

$\rightarrow$  Si  $\|x\|_2 = \|x\|_2 = 1 \rightarrow$  cercle unité qui n'est pas convexe!

### Dualité forte

Si  $\exists x \in \mathbb{R}^d \setminus \{0\} \mid \forall i \quad g_i(x) < 0, \forall j \quad h_j(x) = 0$  et si le problème d'optimisation est convexe, alors :

$$1/ \sup_{\lambda \geq 0, \nu} D(\lambda, \nu) = \inf_{x \in \mathcal{E}} f(x).$$

contraintes de qualification de Slater.

$$2/ \exists (\lambda^*, \nu^*) \in \mathbb{R}_+^m \times \mathbb{R}^m \mid \sup_{\lambda \geq 0, \nu} D(\lambda, \nu) = D(\lambda^*, \nu^*).$$

$$\Rightarrow \inf_{x \in \mathcal{E}} f(x) = D(\lambda^*, \nu^*)$$

Primal

Dual

$$\inf_{x \in \mathcal{E}} f(x)$$

$$\max_{\lambda, \nu} D(\lambda, \nu)$$

st  $\lambda \geq 0$ .

### Conditions de Karush-Kuhn-Tucker

Si le pb d'optimisation est convexe, si toutes les fonctions sont dérivables et si les contraintes de qualification de Slater sont vérifiées, alors :

$x \in \mathbb{R}^d$   
 $(\lambda, \nu) \in \mathbb{R}_+^m \times \mathbb{R}^m$  sont respectivement solutions de (P) et (D) si

1/ Faisabilité primaire :  $x \in \mathcal{E}$

2/ Faisabilité duale :  $\lambda \in \mathbb{R}_+^n$

3/ Complémentarité :  $\forall i \in \{1, \dots, m\} : \lambda_i = 0 \Leftrightarrow g_i(x) = 0$

4/ Stationnarité :  $\nabla_x L(x, \lambda, \nu) = 0$

Preuve

$$L(x, d, v) = f(x) + \sum_i \bar{d}_i g_i(x) + \sum_j v_j h_j(x)$$

$\Rightarrow$   $x$  est solution du primal,  $(d, v)$  solution du dual.

1/  $x \in \mathcal{E}$  car  $x$  solution du primal

2/  $d \geq 0$  car  $d$  solution du dual

3/ si il y a dualité forte, donc  $D(d, v) = f(x)$  (conditions de Slater)

$$\begin{aligned} \inf_{x'} L(x', d, v) &= \inf_x L(x, d, v) \\ &= f(x) = D(d, v) \end{aligned}$$

Donc  $f(x) = f(x) + \underbrace{\sum_i \bar{d}_i g_i(x)}_{\geq 0} + \underbrace{\sum_j v_j h_j(x)}_{\geq 0} = 0$

$$\lambda_i \geq 0, g_i(x) \geq 0 \Rightarrow \forall i \in \{1, \dots, m\} \quad \bar{d}_i g_i(x) = 0$$

4/  $L(x, d, v) = \inf_{x'} L(x', d, v)$

$L$  convexe et différentiable donc puisque  $x$  réalise l'infimum, par la règle de Fermat  $\nabla_x L(x, d, v) = 0$ .

$\Leftrightarrow$  Soient  $(x, d, v)$  vérifiant les 4 conditions KKT. Alors :

1/ Par dualité forte,  $\inf_{x \in \mathcal{E}} f(x) = \sup_{d, v} D(d, v)$

Par dualité faible,

$$\sup_{d', v'} D(d', v') \leq \inf_{x'} f(x') \leq f(x). \quad (1.2)$$

3/  $x \in \mathcal{E}$ ,  $\bar{d}_i g_i(x) = 0 \quad \forall i \in \{1, \dots, m\}, \forall j h_j < 0 \Rightarrow v_j h_j = 0$

$$L(x, d, v) = f(x) + \sum_i \bar{d}_i g_i(x) + \sum_j v_j h_j(x). \quad (1.3)$$

$$= f(x).$$

4/  $\nabla_x L(x, d, v) = 0 \Rightarrow L(x, d, v) = \inf_{x'} L(x', d, v) \text{ car } L \text{ convexe.}$

$$= D(d, v). \quad (2h)$$

Donc  $f(x) = D(d, v) = \sup_{v', d'} D(d', v') = \inf_{x' \in \mathcal{E}} f(x')$ . □

## Application aux SSI

18

Soit  $\mathbf{h}$  un noyau  $\rightarrow H$  un RKHS.

$(x_i, y_i)_{1 \leq i \leq m}$  iid  $\mathcal{C}_\infty$ .

$$\begin{aligned} & \min_{h \in H, b \in \mathbb{R}, \xi \in \mathbb{R}^m} \|h\|_H^2 + C \sum_{i=1}^m \xi_i \\ & \text{s.t. } \xi_i \geq 0 \quad \forall i \in 1..m \quad \beta_i \geq 0 \end{aligned}$$

$$y_i(h(x_i) + b) \geq 1 - \xi_i \quad \forall i \in 1..m \quad \alpha_i \geq 0$$

$$\text{Lagrangien:} \quad \mathcal{L}(h, b, \xi, \alpha, \beta) = \frac{1}{2} \|h\|_H^2 + C \sum_{i=1}^m \xi_i - \underbrace{\sum_{i=1}^m \alpha_i (y_i(h(x_i) + b) - 1 + \xi_i)}_{-h^T \alpha + \beta^T \xi} - \underbrace{\sum_{i=1}^m \beta_i \xi_i}_{\beta^T \xi}$$

### Fonction Duale

$$D(\alpha, \beta) = \inf_{h, b, \xi} \mathcal{L}(h, b, \xi, \alpha, \beta) \quad \forall \alpha \geq 0, \beta \geq 0$$

$$\nabla_h \mathcal{L}(h, -) = h - \sum_{i=1}^m \alpha_i y_i k(., x_i) = 0$$

$$h = \sum_{i=1}^m \alpha_i y_i k(., x_i)$$

$$\frac{\partial \mathcal{L}(-, b, -)}{\partial b} = \alpha^T y = 0$$

$$\nabla_\xi (-, \xi, -) = -\alpha - \beta + C \mathbf{1} = 0$$

$$\begin{aligned} & \left( \sum_i \alpha_i y_i k(., x_i), \sum_j \alpha_j y_j k(., x_j) \right) \\ & = \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \\ & \Rightarrow \beta = C - \alpha \geq 0 \end{aligned}$$

(remplacer les primaires)  $\downarrow$

$$\begin{aligned} D(\alpha, \beta) &= \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j) - \underbrace{\sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j)}_{Q_{ij}} + \mathbf{1}^T \alpha \\ &= -\frac{1}{2} \alpha^T Q \alpha + \mathbf{1}^T \alpha, \quad \text{avec } Q = (y_i y_j^T k(x_i, x_j))_{1 \leq i, j \leq m} \end{aligned}$$

$$(1) \quad \min_{x \in \mathbb{R}^n} -\frac{1}{2} \alpha^T Q \alpha - \mathbf{1}^T \alpha$$

$$\text{s.t. } \begin{cases} 0 \leq \alpha_i \leq C & \forall i \in 1..m \\ \mathbf{1}^T \alpha = 0 \end{cases}$$

$\rightarrow n$  variables vs  $2n + d$

$$\rightarrow \boxed{x}$$

# Algorithme SMO (1998, Sequential Minimal Optimization)

Répéter :

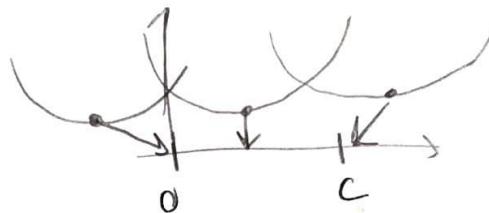
1/ Choisir  $\alpha_i$  qui viole KKT

2/  $\alpha_j$  "au hasard"

3/ Résoudre le problème p/l à  $\alpha_i$  et  $\alpha_j$

$$\alpha_j = \alpha_j - \sum_{l \neq i} \alpha_l \alpha_j$$

pb à + dimension:  
fct obj quadratique  
logisc.



Classification:

$$g(x) = \text{sign}(f(x) + b)$$

$(\hat{f}, \hat{b})$  solution de (P)

$\hat{x}$  solution de (D)

Conditions KKT pour les SOT:

Soit  $(\hat{f}, \hat{b})$  solution de  $\min_{\substack{f \in \mathcal{H} \\ b \in \mathbb{R} \\ \xi \in \mathbb{R}^n}} \frac{1}{2} \|f\|_F^2 + C \|\xi\|^2$   
st  $y_i (f(x_i) + b) \geq 1 - \xi_i \quad \forall i$

Alors le pb d'optimisation dual a une solution  $\hat{\alpha} (\hat{\beta} = C - \hat{\alpha})$  et :

1/  $\hat{f} = \sum_{i=1}^n \hat{\alpha}_i y_i k(\cdot, x_i)$ .

2/  $\hat{f}(x_i) + b > 1 \Rightarrow \hat{\alpha}_i = 0$ .  
—————  $\Rightarrow \hat{\alpha}_i = C$ .

$$0 < \hat{\alpha}_i < C, \hat{f}(x_i) + b = 1$$

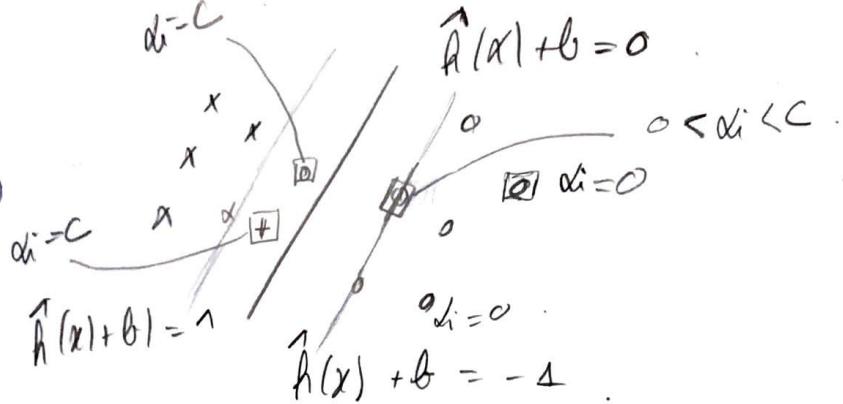
3/  $\hat{\alpha}_i = 0 \Rightarrow \hat{f}(x_i) + b > 1$ .  
 $\hat{\alpha}_i = C \Rightarrow \dots < 1$ .

$$0 < \hat{\alpha}_i < C, \hat{f}(x_i) + b = 1$$

4/  $\hat{b} = y_i - \hat{f}(x_i) \quad \forall i \in \{1..n\} : 0 < \hat{\alpha}_i < C$ .

$\hat{b} \in$  Intervalle (cf notes de cours)

Preuve



$c$  = compromis entre  
work marge et erreur  
 $c \geq 1$ , on s'autorise -  
de points mal classés

$$\hat{R}(w) = \sum \alpha_i y_i R(\cdot, x_i)$$

avec les points mal classés au classe avec tolérance  $< 1$

Première

$$(P) \min_{\substack{w \in \mathbb{R}^m \\ b \in \mathbb{R}}} \frac{1}{2} \|w\|_2^2 + c \sum_{i=1}^n \xi_i$$

$$y_i(\hat{h}(x_i) + b) \geq 1 - \xi_i \quad \xi_i \geq 0$$

$$\frac{1}{2} \|w\|_2^2 + c \sum_{i=1}^n \xi_i$$

(D) a une solution  $(\hat{\alpha}, \hat{\beta})$ ,  $\hat{\beta} = cm - \hat{\alpha}$

Pb d'optimisation convexe

$$\begin{cases} \hat{\alpha}_i = 0 \\ \hat{\alpha}_i = 1 \\ \hat{\alpha}_i = 0 \end{cases} \quad \begin{cases} \xi_i > 0 \\ y_i(\hat{h}(x_i) + \hat{\beta}) > 1 - \xi_i \end{cases}$$

Slater ok  $\Rightarrow$  dualité forte

### DUALITÉ

1/ Faisabilité primaire & duale -  $(\hat{h}, \hat{b}, \hat{\alpha}, \hat{\beta})$

2/ Stationnarité

$$\nabla_{\hat{\alpha}} L(\hat{h}, \hat{b}, \hat{\alpha}, \hat{\beta}) = 0 \Leftrightarrow \hat{h} = \sum \hat{\alpha}_i y_i R(\cdot, x_i)$$

$$3/ \forall i \in 1..n \quad \hat{\alpha}_i = 0 \Leftrightarrow y_i(\hat{h}(x_i) + \hat{b}) = 1 - \xi_i$$

$$\hat{\beta}_i = 0 \quad (\hat{\alpha}_i = 0) \Leftrightarrow \xi_i = 0$$

$$a. y_i(\hat{h}(x_i) + \hat{b}) > 1 \quad (\Rightarrow \hat{\alpha}_i = 0)$$

$$\Rightarrow \xi_i > 0. \text{ ou } y_i(\hat{h}(x_i) + \hat{b}) > 1 - \xi_i \Rightarrow \hat{\alpha}_i = 0. \text{ ok}$$

$$b. y_i(\hat{h}(x_i) + \hat{b}) < 1 \Rightarrow \xi_i > 0 \Rightarrow \hat{\alpha}_i = c. \text{ pb}$$

$$c. \hat{\alpha}_i = 0 \Rightarrow \hat{\beta}_i > 0 \Rightarrow \xi_i = 0 \Rightarrow y_i(\hat{h}(x_i) + \hat{b}) \geq 1 - \xi_i = 1 \quad * \text{ va bien}$$

$$d. \hat{\alpha}_i = c \Rightarrow y_i(\hat{h}(x_i) + \hat{b}) = 1 - \xi_i \Rightarrow y_i(\hat{h}(x_i) + \hat{b}) < 1 \text{ car } \xi_i > 0 \text{ pb}$$

$$0 \leq \hat{y}_i < c \Rightarrow y_i(\hat{h}(x_i) + \hat{b}) = 1.$$

$$\text{Donc } \hat{b} = y_i - \hat{h}(x_i)$$

Vision Statistique  $\lambda \geq 0$

+ facile pr optim

$$(\hat{h}, \hat{b}) \in \underset{\substack{h \in \mathbb{H} \\ b \in \mathbb{R}}}{\operatorname{argmin}} \frac{1}{2} \|h\|_H^2 + \underbrace{\frac{1}{n} \sum_{i=1}^n (1 - y_i(h(x_i) + b))_+}_{\text{eq. b}}. \quad (\text{P1})$$

(P2)

$$\text{vers AS : } \underset{\substack{h \in \mathbb{H} \\ b \in \mathbb{R}}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (1 - y_i(h(x_i) + b))_+ \quad \mathcal{F} = \{h \in \mathbb{H}, \|h\|_H \leq c\} \\ c > 0.$$

+ facile pr stat

$$c'' \propto \frac{1}{\lambda}$$

### Proposition

- \* Si  $(\hat{h}, \hat{b})$  solution de (P1)  $\lambda \geq 0$ , alors  $\exists c \in \mathbb{R}_+ \setminus (\hat{h}, \hat{b})$  et solution de (P2).
- \* Si  $(h^*, b^*)$  solution de (P2),  $c > 0$ , alors  $\exists \lambda \geq 0$  tq  $(h^*, b^*)$  solution de (P1).

### Preuve

$$(\text{P2}) \quad \begin{cases} \min_{\substack{h \in \mathbb{H} \\ b \in \mathbb{R}}} & l(h, b) \\ & \|h\|_H^2 \leq c^2, \quad c > 0. \end{cases} \quad \lambda \geq 0.$$

↑ Puisque  $c > 0$ , les qualifications de Slater sont vérifiées pour  $b = 0$ .  
Pb optimisation convexe. Il y a donc dualité forte :

$$\inf_{\substack{h \in \mathbb{H} \\ b \in \mathbb{R}}} l(h, b) = \sup_{\lambda \geq 0} D(\lambda) = D(\lambda^*) \\ \text{pour } \lambda \geq 0 \quad = \inf_{h, b} L(h, b, \lambda^*) = \inf_{h, b} l(h, b) + \lambda^* \|h\|_H^2 - \lambda^* c^2.$$

On appelle  $(h^*, b^*)$  solution de (P2).

$$\forall (h, b) \in \mathbb{H} \times \mathbb{R}, \quad l(h, b) + \lambda^* \|h\|_H^2 \leq l(h^*, b^*) \leq \inf_{\substack{h' \in \mathbb{H} \\ b' \in \mathbb{R} \\ \|h'\|_H \leq c}} l(h', b')$$

$\|h\|_H^2 \leq c^2$  car  $h^*$  solution de (P2)

$$= \inf_{h, b} l(h, b) + \lambda^* \|h\|_H^2 - \lambda^* c^2 \\ \leq l(h^*, b^*) + \lambda^* \|h\|_H^2 - \lambda^* c^2.$$

$$\hat{u}(h^*, b^*) \in \underset{\substack{h \in H \\ b \in \mathbb{R}}}{\text{argmin}} \ell(h, b) + \underbrace{\frac{d}{2} \|h\|_H^2}_{\geq 0}$$

↓ Sait  $(\hat{h}, \hat{b}) \in \underset{h, b}{\text{argmin}} \ell(h, b) + \frac{d}{2} \|h\|_H^2$  solution de (P1) avec  $d \geq 0$

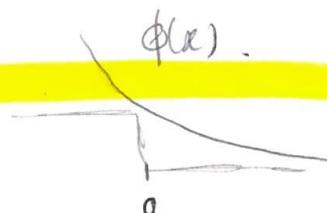
On prend  $c = \|h\|_H$ . Alors  $\forall (h, b) \in H \times \mathbb{R}$  tq  $\|h\|_H \leq c$ .

$$\begin{aligned} \ell(\hat{h}, \hat{b}) &= \ell(h^*, b^*) + \underbrace{\frac{d}{2} \|h\|_H^2 - \frac{d}{2} \|h\|_H^2}_{\leq 0} \\ &\leq \ell(h^*, b^*) + \underbrace{\frac{d}{2} \|h\|_H^2}_{\leq d c^2} - \underbrace{\frac{d}{2} c^2}_{\leq 0} \\ &\leq \ell(h^*, b^*) \quad \text{car } \|h\|_H^2 \leq c^2 \end{aligned}$$

faire TD2 + exerc

Donc  $(\hat{h}, \hat{b})$  est solution de (P2)

### Géométrie Statistique



(Bouch p.47) Rette  $\phi : \mathbb{R} \rightarrow \mathbb{R}_+$  tq  $\forall x \in \mathbb{R} \quad \phi(x)$

et si  $\sup_{y \in \mathbb{R}} |\phi(y)| < B$  ps,  $S = \{h \in H \mid \|h\|_H \leq c\}$

Alors  $\hat{g}_n \in \underset{f \in \mathcal{F}}{\text{argmin}} \sum_{i=1}^n \phi(Y_i f(x_i))$ ,  $D_n = (X_i, Y_i)_{1 \leq i \leq n}$ , alors, avec  $\rho \geq 1-\delta$ ,

$$+ B \sqrt{\frac{\log(1/\delta)}{2n}}$$

### SDM

- $\phi(x) = (1-x)_+$ , 1-lipschitzienne.

- $K$  est borné i.e  $\exists K \mid \forall x \in \mathbb{R}^n, \|x\|_2 \leq K$

Alors on a  $B = 1 + K$ .

$$\text{Ent}(\hat{f}(X^n)) \leq \frac{OK}{2n}$$

Preuve :

$B = \text{Cauchy-Schwartz}$

$$R_n \leq \frac{C}{m} \sqrt{\sum_{i=1}^m R(x_i, x_i)}$$

$$\leq \frac{CK}{\sqrt{m}}$$

⚠ A faire (Final)

To Do

08/10/21

Regression  
Classification

## A detour to nonparametric regression

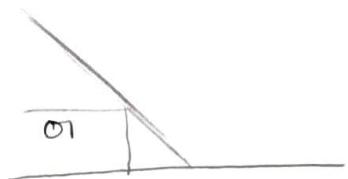
(18)

Régression  $(X, Y)$

Classification ESN + RNNs.

•  $Y$  à valeurs dans  $\mathbb{R}$ ,

• Perte de régression :



Régression des moindres carrés :

$$\text{Modèle : } Y = f^*(X) + \varepsilon \quad \Rightarrow \quad \mathbb{E}(Y|X) = f^*(X)$$

$f^*$  mesurable tq  $f^*(x) \in \mathbb{L}^2$

$$\mathbb{E}(\varepsilon|X) = 0$$

$\rightarrow f \mapsto \mathbb{E}[(Y - f(X))^2]$  est minimisée par  $f^*$  parmi toutes les fonctions mesurables tq  $f(x) \in \mathbb{L}^2$ .

$$\begin{aligned} \text{Soit } f \text{ tq } f(x) \in \mathbb{L}^2, \forall (a, y) \in \mathbb{R}^d \times \mathbb{R}, & (y - f(x))^2 - (y - f^*(x))^2 \\ &= (2y - f(x) - f^*(x))(f^*(x) - f(x)) . \end{aligned}$$

$$\begin{aligned} & \mathbb{E}[(y - f(x))^2 - (y - f^*(x))^2 | X] \\ &= (2\mathbb{E}[Y|X] - f(x) - f^*(x))f^*(x) - f(x) = (f^*(x) - f(x))^2 \geq 0 \end{aligned}$$

$$\mathbb{E}[(y - f(x))^2] \geq \mathbb{E}[(y - f^*(x))^2]$$

$$\underset{\substack{f \in \mathbb{L}^2 \\ f \neq f^*}}{\text{find}} \underset{\substack{f \in \mathbb{L}^2 \\ f \neq f^*}}{\text{argmin}} \frac{1}{n} \sum (y_i - f(x_i))^2 + \frac{\lambda}{2} \|f\|_{\mathbb{L}^2}^2 \quad \text{avec } \mathbb{P} = \{f \in \mathbb{H}, \|f\|_{\mathbb{H}} \leq 1, \omega_0\}$$

Ridge kernel Regression

$$\begin{cases} Y = f^*(x) + \varepsilon \\ (\text{P}) \quad \varepsilon > 0 \quad |Y| \geq \frac{1}{2} \\ f^* \in \mathcal{H} \end{cases}$$

derrière  
exo

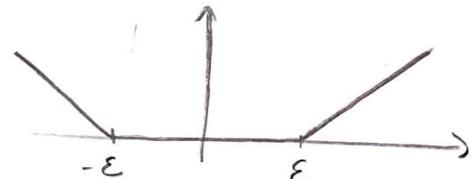
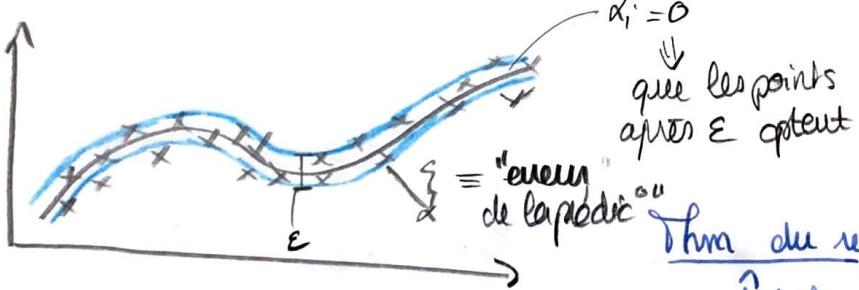
$f^*$  minimise le risque  $f \mapsto E[|Y - f(x)|]$   $\forall f : f(x) \in \mathbb{R}^2$

$$\frac{1}{2} \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n |\gamma_i - f(x_i)|$$

+ redoute  
- d'importance aux valeurs  
absentantes

### Support vector regression

$$\hat{f}_n \in \underset{\substack{f \in \mathcal{H}, \\ b \in \mathbb{R}}}{\operatorname{argmin}} \frac{1}{2} \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n \ell_\varepsilon(Y_i - f(x_i) + b), \quad \ell_\varepsilon(x) = (|x| - \varepsilon)_+$$



Thm du représentant :

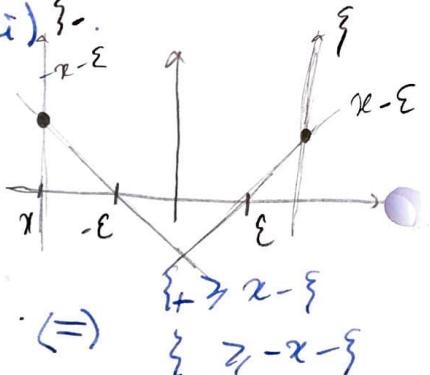
$$\hat{f}_n(x) = \sum \alpha_i k(x_i, x)$$

$$\forall x \in \mathbb{R}, \quad \ell_\varepsilon(x) = (|x| - \varepsilon)_+ = \inf_{\xi_+, \xi_- \geq 0} \{ \xi_+ + \xi_- \}$$

formulation  
variationnelle

→ 2n nouvelles variables

$$\left\{ \begin{array}{l} -\xi_- - \varepsilon \leq x \leq \xi_+ + \varepsilon \\ \xi_+ + \xi_- = 0 \end{array} \right. \quad (=)$$



$$\frac{1}{m} \sum \xi_i^+ + \xi_i^-$$

st

Théorème

$\phi: L_\phi$  - Lipschitzienne

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \phi(y_i - f(x_i)) \quad , \quad \mathcal{F} = \{f \in \mathcal{H}, \|f\|_{\mathcal{H}} \leq c\}$$

$(x_i, y_i)$  iid

$$\sup_{f \in \mathcal{F}} \phi(y - f(x)) < B \text{ ps et } Y \subseteq \mathbb{R}^d.$$

$$\text{avec } 1/p \geq 1-\delta, \mathbb{E}[\phi(y_i - \hat{f}_n(x_i))] | \mathcal{D}_n \leq \inf_{f \in \mathcal{F}} \mathbb{E}[\phi(y - f(x))] +$$

$$8L_\phi(\mathbb{E}[R_n(\mathcal{F}(X))]) + C \sqrt{\frac{2\log(1/\delta)}{n}} + 2B \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Partie 8.2,

$\phi = \ell_\epsilon$  1-lipschitzienne.

$$\sup_{x \in \mathbb{R}^d} \ell_\epsilon(x) \leq K^2, \quad (\text{d})>0.$$

$$|Y| \leq C \text{ ps}$$

$$\mathbb{E}[\ell_\epsilon(Y - \hat{f}_n(x)) | \mathcal{D}_n] - \inf_{f \in \mathcal{F}} \mathbb{E}[\ell_\epsilon(Y - f(x))]$$

$$\leq 8 \frac{\epsilon K}{\sqrt{n}} + 8C \sqrt{\frac{2\log 2}{n}} + 2(C + CK) \sqrt{\frac{\log(1/\delta)}{2n}}$$

$$\sup_{f \in \mathcal{F}} \ell_\epsilon(Y - f(x)) \leq ?$$

$$\forall f \in \mathcal{F}, \|f\|_H \leq C.$$

$$\ell_\epsilon(Y - f(x)) \leq |Y - f(x)| \leq |Y| + |f(x)| \leq C + \underbrace{\langle f, k_\epsilon(x) \rangle_H}_{\text{es: } \|f\|_H \leq \|k_\epsilon(x)\|_H} + \underbrace{\langle C \sqrt{\|k_\epsilon(x)\|_H}, k_\epsilon(x) \rangle_H}_{\leq C \sqrt{\|k_\epsilon(x)\|_H}} + \underbrace{\langle C \sqrt{\|k_\epsilon(x)\|_H}, k_\epsilon(x) \rangle_H}_{\leq C \sqrt{\|k_\epsilon(x)\|_H}} \leq C + CK.$$

$$\leq \frac{C}{\|f\|_H} + CK.$$

$$R_n[\mathcal{F}(X^n)] = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right| / \|\mathcal{D}_n\|$$

$\underbrace{\langle f, k(\cdot, x_i) \rangle}_{\mathcal{H}}$

$$\sqrt{\frac{1}{n} \sum_{i=1}^n \sigma_i \langle f, k(\cdot, x_i) \rangle}$$

$$\sqrt{\frac{1}{n} \|f\|_{\mathcal{H}} \sqrt{\sum_{i,j} \sigma_i \sigma_j k(x_i, x_j)}}$$

$$\leq \frac{c}{n} \mathbb{E} \left[ \sqrt{\sum_{i,j} \sigma_i \sigma_j k(x_i, x_j)} \mid \mathcal{D}_n \right] \quad \text{Jensen, F. concave}$$

$$\leq \frac{c}{n} \sqrt{\mathbb{E} \left[ \sum_{i,j} \sigma_i \sigma_j k(x_i, x_j) \right]}$$

$$\leq \frac{c}{n} \sqrt{\sum_{i,j} k(x_i, x_j)} \leq \frac{cK}{\sqrt{n}}$$

$\sigma_i$  iid s.t.  $\mathbb{E}[f]$ .