Mathématiques du Machine Learning M2 MAPI3

2024-2025

Exercice n° 3 :

1) Il s'agit d'un problème de classification.

2) $\hat{f} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{f(x_i) \neq y_i}$

3) $\forall x \in \mathcal{X}, c \in \{-1, 1\}$ $\underbrace{f_X(x)}$

$$\mathbb{P}(X = x, Y = c) = \mathbb{P}(X = x)\,\mathbb{P}(Y = c \mid X = x)$$

$$\text{de plus } \mathbb{P}(X = x, Y = c) = \underbrace{\mathbb{P}(Y = c)}_{\pi_c}\,\underbrace{\mathbb{P}(X = x \mid Y = c)}_{f_{X|Y=c}(x)}$$

$$\text{donc } \forall x, \forall c, \underbrace{f_X(x)}_{>0}\,\mathbb{P}(Y = c \mid X = x) = \pi_c f_{X|Y=c}(x)$$

$$\text{d'où : } \mathbb{P}(Y = c \mid X = x) = \frac{\pi_c f_{X|Y=c}(x)}{f_X(x)}$$

or, d'après la formule des probabilités totales,

$$f_X(x) = \pi_1 f_{X|Y=1}(x) + \pi_{-1} f_{X|Y=-1}(x)$$

d'où, $\forall x \in \mathbb{R}$ $\mathbb{P}(Y=c|X=x) = \dfrac{\pi_c f_{X|Y=c}}{\pi_1 f_{X|Y=1}(x) + \pi_{-1} f_{X|Y=-1}(x)}$

4) $\mathbb{P}_{\hat{\pi}_1}(Y_1,\dots,Y_n) \prod_{i=1}^{n} \hat{\pi}_1^{\mathbb{1}_{Y_i=1}} (1-\hat{\pi}_1)^{1-\mathbb{1}_{Y_i=1}}$

donc en prenant le log :

$$\log \mathbb{P}_{\hat{\pi}_1}(Y_1,\dots,Y_n) = \sum_{i=1}^{n} \left(\mathbb{1}_{Y_i=1} \log(\hat{\pi}_1) + (1-\mathbb{1}_{Y_i=1}) \log(1-\hat{\pi}_1) \right)$$

$$\frac{\partial}{\partial \hat{\pi}_1} \log \mathbb{P}_{\hat{\pi}_1}(Y_1,\dots,Y_n) = \sum_{i=1}^{n} \left(\frac{\mathbb{1}_{Y_i=1}}{\hat{\pi}_1} - \frac{1-\mathbb{1}_{Y_i=1}}{1-\hat{\pi}_1} \right)$$

$$\frac{\partial}{\partial \hat{\pi}_1} \log \mathbb{P}_{\hat{\pi}_1}(Y_1,\dots,Y_n) = 0 \quad \text{ssi}$$

$$\sum_{i=1}^{n} \left(\frac{\mathbb{1}_{Y_i=1}}{\hat{\pi}_1} - \frac{1-\mathbb{1}_{Y_i=1}}{1-\hat{\pi}_1} \right) = 0$$

ssi $\dfrac{1-\hat{\pi}_1}{\hat{\pi}_1} \left(\sum_{i=1}^{n} \mathbb{1}_{Y_i=1} \right) = n - \sum_{i=1}^{n} \mathbb{1}_{Y_i=1}$

ipp $\quad \sum_{i=1}^{\hat{n}} \mathbb{1}_{Y_i=1} - \hat{\pi}_1 \sum_{i=1}^{\hat{n}} \mathbb{1}_{Y_i=1} = n\hat{\pi}_1 - \hat{\pi}_1 \sum_{i=1}^{\hat{n}} \mathbb{1}_{Y_i=1}$

ipp $\quad \boxed{\hat{\pi}_1 = \frac{1}{n} \sum_{i=1}^{\hat{n}} \mathbb{1}_{Y_i=1}}$

de plus, $\frac{\partial^2}{\partial \hat{\pi}_1^2} \log \mathcal{P}_{\hat{\pi}_1}(Y_1, \dots, Y_n) = \sum_{i=1}^{\hat{n}} \left(-\frac{\mathbb{1}_{Y_i=1}}{\hat{\pi}_1^2} - \frac{1-\mathbb{1}_{Y_i=1}}{(1-\hat{\pi}_1)^2} \right) \leq 0$

Ainsi la fonction à optimiser étant concave, le point précédent est bien le maximum.

5) Remarquons que $\forall i, \mathbb{1}_{Y_i=1} \sim \mathcal{B}(\pi_1)$, loi de Bernoulli,

de plus que $\mathbb{1}_{Y_1=1}, \dots, \mathbb{1}_{Y_n=1}$ sont indépendants.

Alors
- $\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{\hat{n}} \mathbb{1}_{Y_i=1}\right) = \frac{1}{n} \sum_{i=1}^{\hat{n}} \mathbb{E}\left(\mathbb{1}_{Y_i=1}\right) = \pi_1 \Rightarrow$ sans biais
- $\mathbb{V}ar\left(\frac{1}{n} \sum_{i=1}^{\hat{n}} \mathbb{1}_{Y_i=1}\right) = \frac{1}{n^2} \sum_{i=1}^{\hat{n}} \mathbb{V}ar\left(\mathbb{1}_{Y_i=1}\right) = \frac{\pi_1(1-\pi_1)}{n}$

Nous allons appliquer l'inégalité de Hoeffding aux

$$V_1, Z_1 = 1_{V_1 \dots}, \dots, b, 1 \text{ et } a, b$$

et t, ε on aura $\varepsilon > 0$.

d'où $P(|\hat{f}_{n,h}(x) - f(x)| > \varepsilon) \leq 2\exp(-2n\varepsilon^2)$

Ainsi, cet estimateur converge exponentiellement vite vers la

Se rappeler

1) $\forall x, f_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$

≥ 0 par hypothèse

donc $f_n(x) \geq 0$.

- $\int_{\mathbb{R}^d} f_n(x)\,dx = \int_{\mathbb{R}^d} \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) dx$

$$= \frac{1}{nh^d} \sum_{i=1}^{n} \int_{\mathbb{R}^d} K\left(\frac{x - X_i}{h}\right) dx$$

Dans chaque terme de la somme, on effectue le changement de variable $u = \frac{x - X_i}{h}$

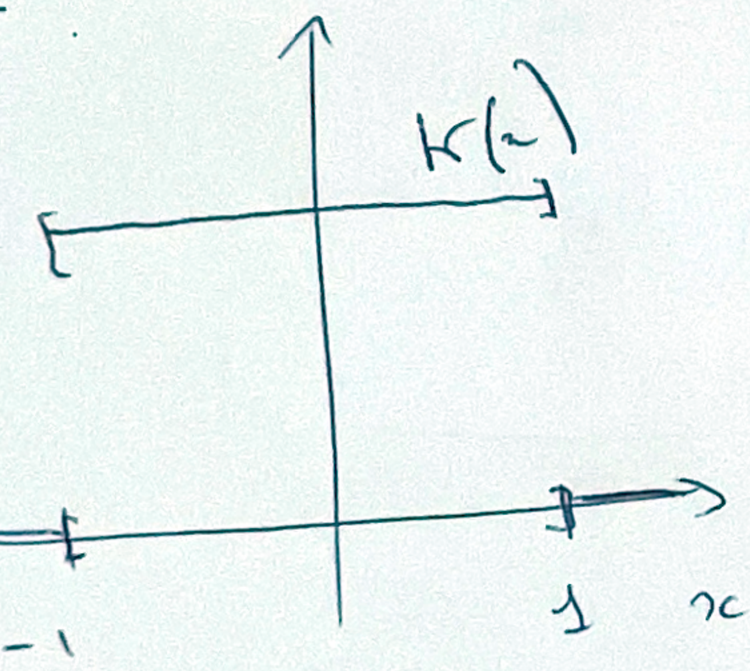
$$\text{d'où } \frac{1}{n} \sum_{i=1}^{n} \int_{\mathbb{R}^d} K(u)\,du = \dots$$

$$\int_{\mathbb{R}^d} \hat{f}_h(x)\,dx = \frac{1}{n h^d} \sum_{i=1}^{n} \int_{\mathbb{R}^d} K\left(\frac{x - X_i}{h}\right) \underbrace{h^d}_{\text{Jacobien du changement de variable.}} dx_i$$

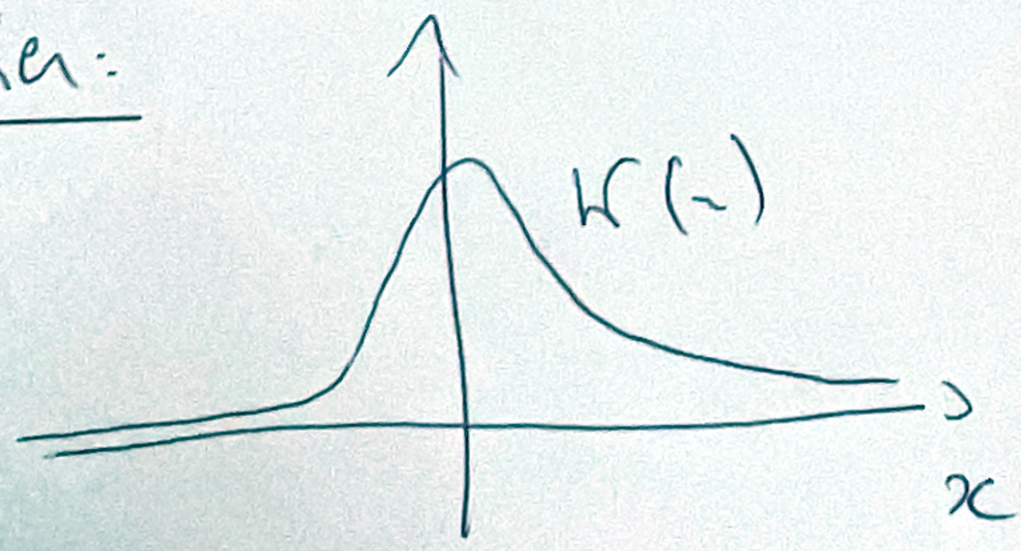
$$= \frac{1}{n} \sum_{i=1}^{n} \underbrace{\int_{\mathbb{R}^d} K}_{=1}$$

$$= 1 .$$

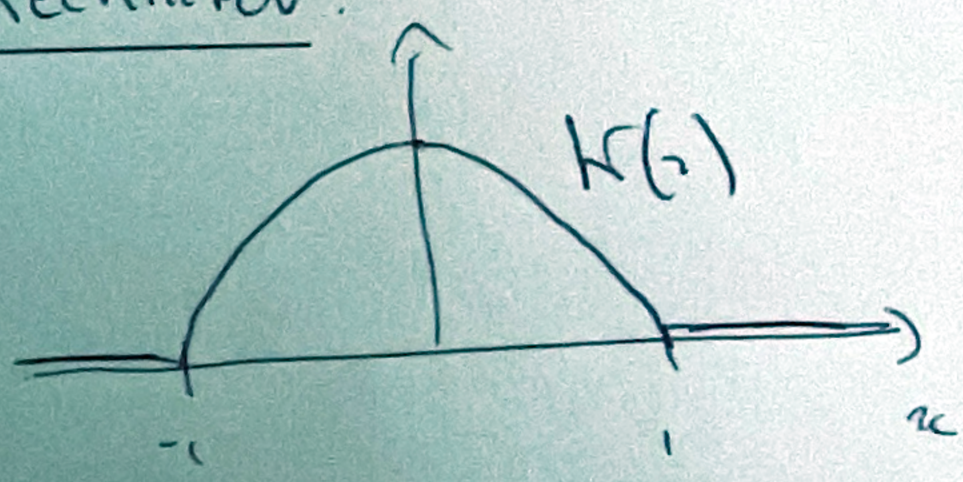
8) Uniforme :



Gaussien :



Épanechnikov :



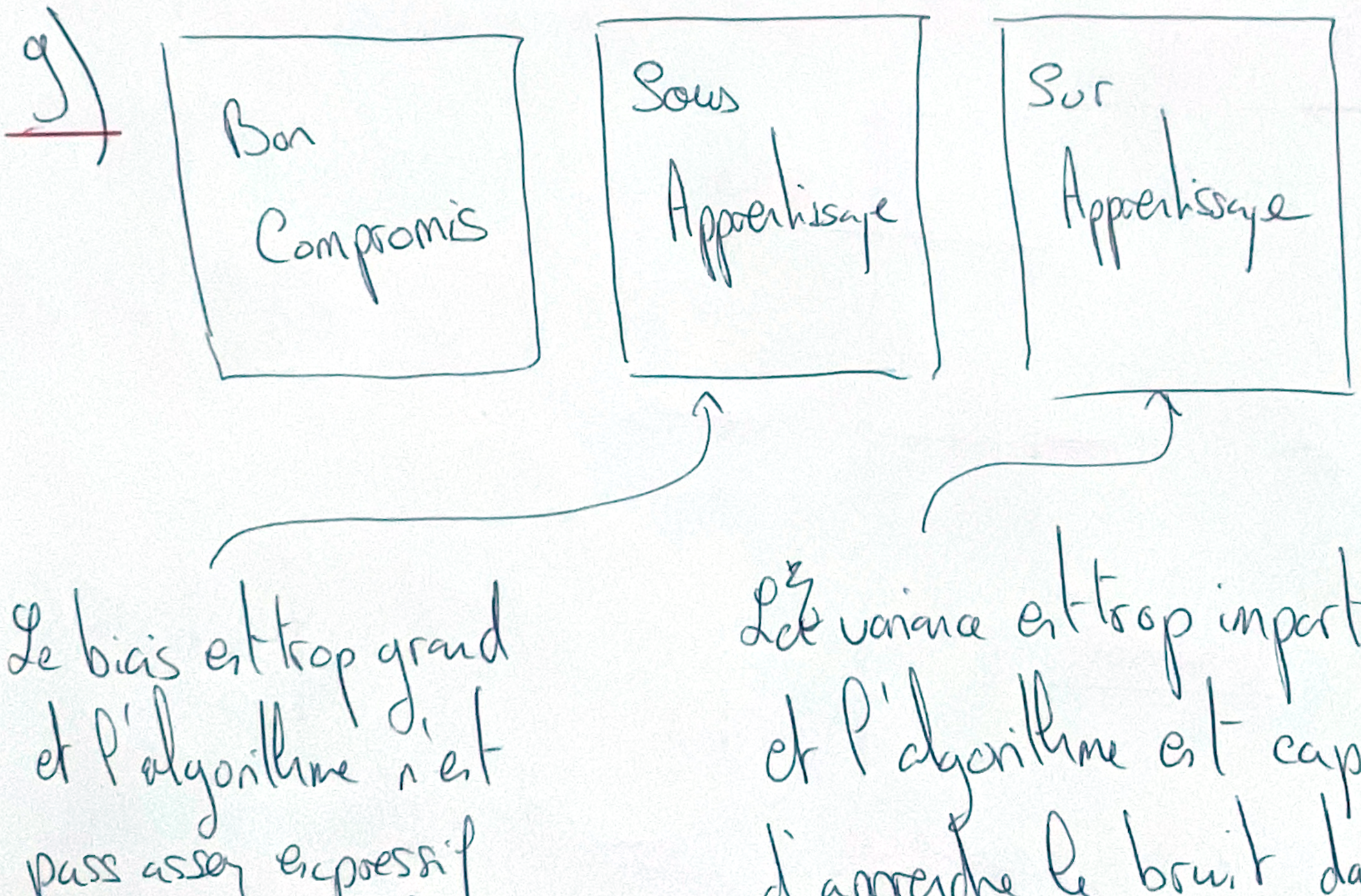
- Lorsque k est grand :

Localement, beaucoup d'information est agrégée, il y a peu de variance mais un fort biais.

- Lorsque k est petit :

Un point n'ajoute de la "masse" que très localement, il y a peu de biais mais une forte variance.

9)



Le biais est trop grand et l'algorithme n'est pas assez expressif par apprendre la "vraie" fonction de classification

Le variance est trop importante et l'algorithme est capable d'apprendre le bruit dans les données.

D

Exercice 2 :

1) Soit $n \geq 1$, $x_1, \dots, x_n \in X$, $\alpha_1, \dots, \alpha_n \in \mathbb{R}$,

$$\sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) = \sum_{i,j} \alpha_i \alpha_j \langle \phi(x_i), \phi(x_j) \rangle \quad \text{(par définition)}$$

$$= \left\langle \sum_i \alpha_i \phi(x_i), \sum_j \alpha_j \phi(x_j) \right\rangle \quad \text{(bilinéarité)}.$$

$$= \left\| \sum_i \alpha_i \phi(x_i) \right\|^2 \quad \text{(définition de la norme)}$$

$$\geq 0 \quad \text{(car carré toujours positif)}.$$

2) D'après le <u>théorème d'Aronszajn</u>, si k est un noyau symétrique positif, alors il existe un espace de Hilbert et $\phi : X \to H$ tq

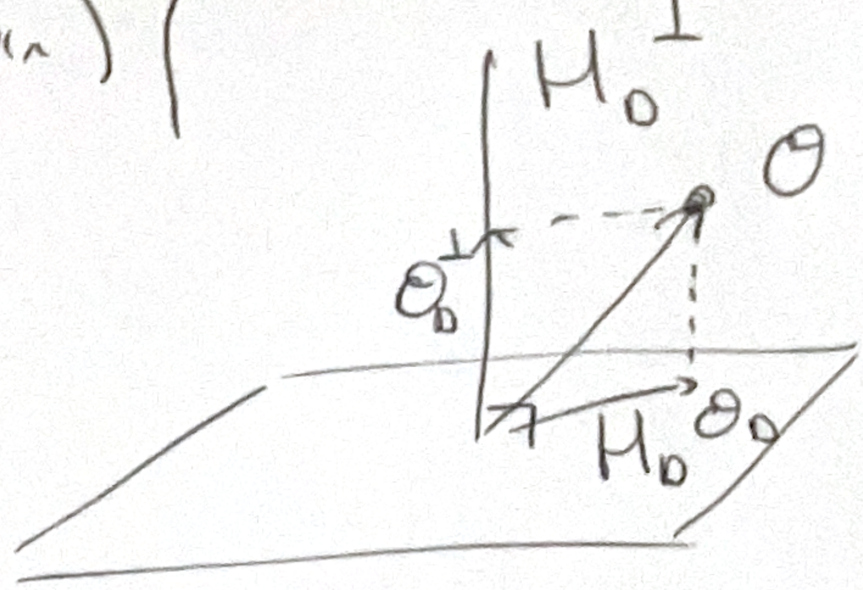
$$\forall x_1, x_2 \in X, \; k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle.$$

Donc, si on a un noyau pas un produit scalaire et une forme ϕ n'est pas restrictif, il y a équivalence entre les deux.

3) Notons $H_D = \text{Vect}\{\phi(x_1), \dots, \phi(x_n)\}$

Soit $\theta \in H$, $\theta = \theta_D + \theta_D^\perp$

$\theta_D \in H_D$, $\theta_D^\perp \in H_D^\perp$



Notons $J_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y_i \langle \theta, \phi(x_i)\rangle}\right) + \lambda\|\theta\|^2$

Alors $J_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y_i \langle \theta_D + \theta_D^\perp, \phi(x_i)\rangle}\right) + \lambda\|\theta_D + \theta_D^\perp\|^2$

$= \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-y_i \underbrace{\left(\langle \theta_D, \phi(x_i)\rangle + \langle \theta_D^\perp, \phi(x_i)\rangle\right)}_{\text{par Pinocchio}}}\right) + \lambda\|\theta_D\|^2 + \lambda\|\theta_D^\perp\|^2$ par Pythagore.

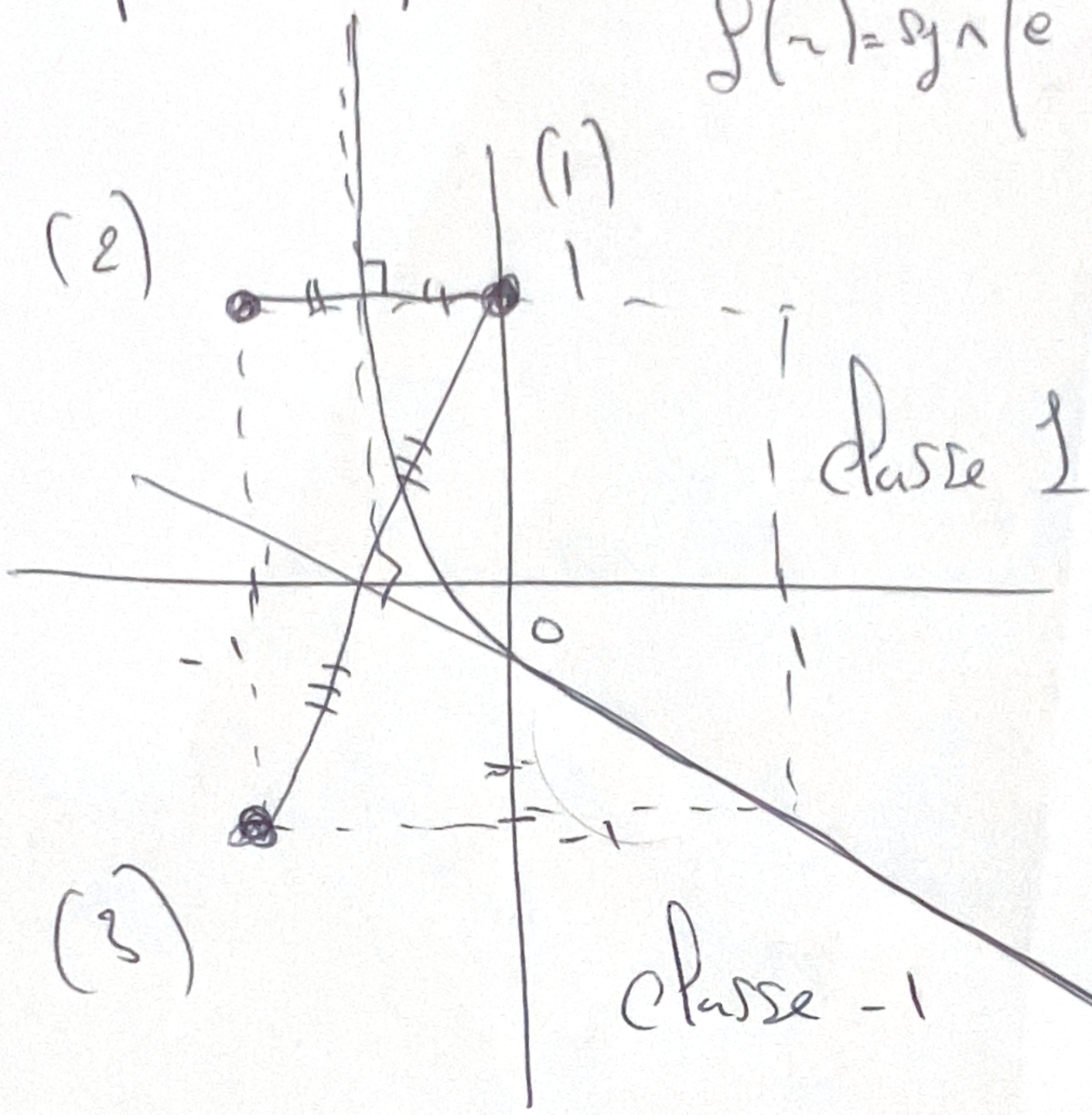
or, $\forall i, \langle \theta_D^\perp, \phi(x_i)\rangle = 0$ car $\theta_D^\perp \in \text{Vect}(\phi(x_1), \dots, \phi(x_n))^\perp$.

donc $\boxed{J_n(\theta) = J_n(\theta) + \lambda\|\theta_D^\perp\|^2}$

Une solution optimale (si elle existe) a donc une composante dans l'orthogonal de H_D nulle. On peut se restreindre à chercher dans H_D.

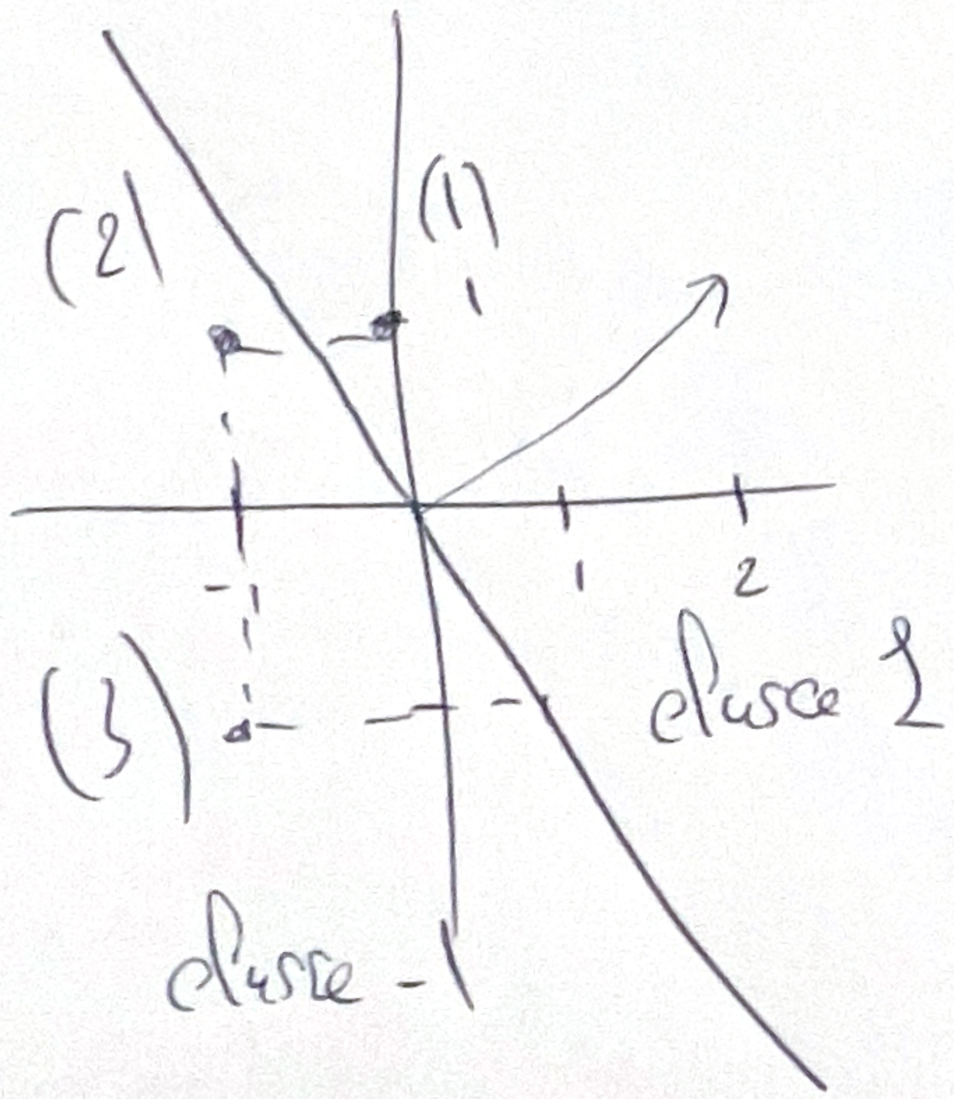
Sur H_3, le problème est <u>parfaitement convexe</u> (grâce au terme $\lambda \|\theta\|^2$). Il admet donc une unique solution.

a) (noyau exponentiel)



$$\rho(x) = \mathrm{sgn}\left(e^{-\|x-\binom{0}{1}\|^2} - e^{-\|x-\binom{-1}{-1}\|^2}\right)$$

b) (noyau linéaire)



$$\rho(x) = \mathrm{sgn}\left(\langle x, \binom{0}{1}\rangle - \langle x, \binom{-1}{1}\rangle - \langle x, \binom{-1}{-1}\rangle\right)$$

$$= \mathrm{sgn}\left(\langle x \xi_2, \binom{2}{1}\rangle\right)$$

5) Changement de variables $\Theta = \alpha_1 \phi(x_1) + \dots + \alpha_n \phi(x_n)$

$$J_n(\alpha_1, \dots, \alpha_n) = \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i \left(\sum_j \alpha_j \phi(x_j), \phi(x_i) \right)}\right) + \frac{\lambda}{2} \left\| \sum_j \alpha_j \phi(x_j) \right\|^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i \sum_j \alpha_j (\phi(x_j), \phi(x_i))}\right) + \frac{\lambda}{2} \left(\sum_i \alpha_i \phi(x_i), \sum_j \alpha_j \phi(x_j) \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i (K\alpha)_i}\right) + \frac{\lambda}{2} \alpha^T K \alpha$$

où $K = \left(k(x_i, x_j) \right)_{i,j}$.

6) Il est possible de résoudre le problème par descente de gradient (stochastique).

Il faut être capable de calculer

$$\nabla_\alpha \left(\frac{\lambda}{2} \alpha^T K \alpha \right) \text{ et } \nabla_\alpha \left(\log\left(1 + e^{-y_i (K\alpha)_i}\right) \right) \forall i$$

ce qui donne

- $\nabla_\alpha \left(\frac{\lambda}{2} \alpha^T K \alpha \right) = \lambda K \alpha$
- $\nabla_\alpha \left(\log\left(1 + e^{-y_i (K\alpha)_i}\right) \right) = \frac{-e^{-y_i (K\alpha)_i}}{1 + e^{-y_i (K\alpha)_i}} y_i \begin{pmatrix} (\phi(x_1), \phi(x_i)) \\ \vdots \\ (\phi(x_n), \phi(x_i)) \end{pmatrix}$