

# Mathématiques du Machine Learning 12 MAPJ3

2024 - 2025

Exercice n° 2:

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

2)  $\hat{f} \in \underset{f \in F}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \|f(x_i) - y_i\|$

3)  $\forall x \in X, c \in \{-1, 1\}$   $\underbrace{f_x(z)}_{P(X=x)}$   
 $P(X=x, Y=c) = P(X=x) P(Y=c | X=x)$   
de plus  $P(X=x, Y=c) = \underbrace{P(Y=c)}_{\pi_c} \underbrace{P(X=x | Y=c)}_{f_{X|Y=c}(z)}$

donc  $\forall x, \forall c, \underbrace{f_x(z)}_{>0} P(Y=c | X=x) = \pi_c f_{X|Y=c}(z)$

d'où:  $P(Y=c | X=x) = \frac{\pi_c f_{X|Y=c}(z)}{f_x(z)}$

or, d'après la formule des probabilités totales,

$$f_x(z) = \pi_1 f_{x|y=1}(z) + \pi_{-1} f_{x|y=-1}(z)$$

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

$$\text{4) } P_{\hat{\pi}_1}(y_1, \dots, y_n) = \prod_{i=1}^n \pi_{y_i=1}^{y_i=1} (1 - \pi_i)^{1 - y_i=1}$$

donc on prend le log:

$$\log P_{\hat{\pi}_1}(y_1, \dots, y_n) = \sum_{i=1}^n \left( \frac{y_i=1}{\pi_i} \log(\pi_i) + (1 - y_i=1) \log(1 - \pi_i) \right)$$

$$\frac{\partial}{\partial \hat{\pi}_i} \log P_{\hat{\pi}_1}(y_1, \dots, y_n) = \sum_{i=1}^n \left( \frac{\frac{y_i=1}{\pi_i}}{1 - \hat{\pi}_i} - \frac{1 - y_i=1}{1 - \hat{\pi}_i} \right)$$

$$\frac{\partial}{\partial \hat{\pi}_i} \log P_{\hat{\pi}_1}(y_1, \dots, y_n) = 0 \text{ iff}$$

$$\sum_{i=1}^n \left( \frac{\frac{y_i=1}{\pi_i}}{1 - \hat{\pi}_i} - \frac{1 - y_i=1}{1 - \hat{\pi}_i} \right) = 0$$

$$\text{iff } \frac{1 - \hat{\pi}_i}{\hat{\pi}_i} \left( \sum_{i=1}^n \frac{y_i=1}{\pi_i} \right) = n - \sum_{i=1}^n \frac{y_i=1}{\pi_i}$$

iff

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

iff

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

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

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

5) Remarquons que  $V_i, \frac{1}{\mathbb{1}_{Y_i=1}} \sim \underbrace{B(\pi_i)}$ , et  
Lai de Bernoulli

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

Plus •  $E\left(\frac{1}{n} \sum_{i=1}^n \frac{1}{\mathbb{1}_{Y_i=1}}\right) = \frac{1}{n} \sum_{i=1}^n E\left(\frac{1}{\mathbb{1}_{Y_i=1}}\right) = \pi_1 \Rightarrow$  sans biais

•  $\text{Var}\left(\frac{1}{n} \sum_{i=1}^n \frac{1}{\mathbb{1}_{Y_i=1}}\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}\left(\frac{1}{\mathbb{1}_{Y_i=1}}\right) = \frac{\pi_1(1-\pi_1)}{n}$

6) Utiliser la propriété Poincaré de Helly

$$V_i, Z_i \in L_{p_i, \mu_i}, b_i, 1/d_i > 0$$

et t, s en  $\mathbb{R}^d$

$$\exists n \in \mathbb{N} \quad P\left(\|Z_i - b_i\|_p > t\right) \leq \exp(-nd_i)$$

Alors, on extrait une suite d'échantillons de  
S<sub>n</sub> de  $\mathbb{R}^d$

2)  $V_n, P_k(\cdot), \frac{1}{n^k} \sum_{i=1}^n k\left(\frac{\|\cdot - b_i\|_p}{t}\right)$

soit  $P_k(\cdot) \geq 0$

donc  $P_k(\cdot) \geq 0$

$$\begin{aligned} \int_{B_d} P_k(\cdot) d\mu &= \int_{B_d} \frac{1}{n^k} \sum_{i=1}^n k\left(\frac{\|\cdot - b_i\|_p}{t}\right) d\mu \\ &= \frac{1}{n^k} \sum_{i=1}^n \int_{B_d} k\left(\frac{\|\cdot - b_i\|_p}{t}\right) d\mu \end{aligned}$$

Dès lors que le R. Sane, nous effectuons le changement de  
variables  $v_i = \frac{\|\cdot - b_i\|_p}{t}$

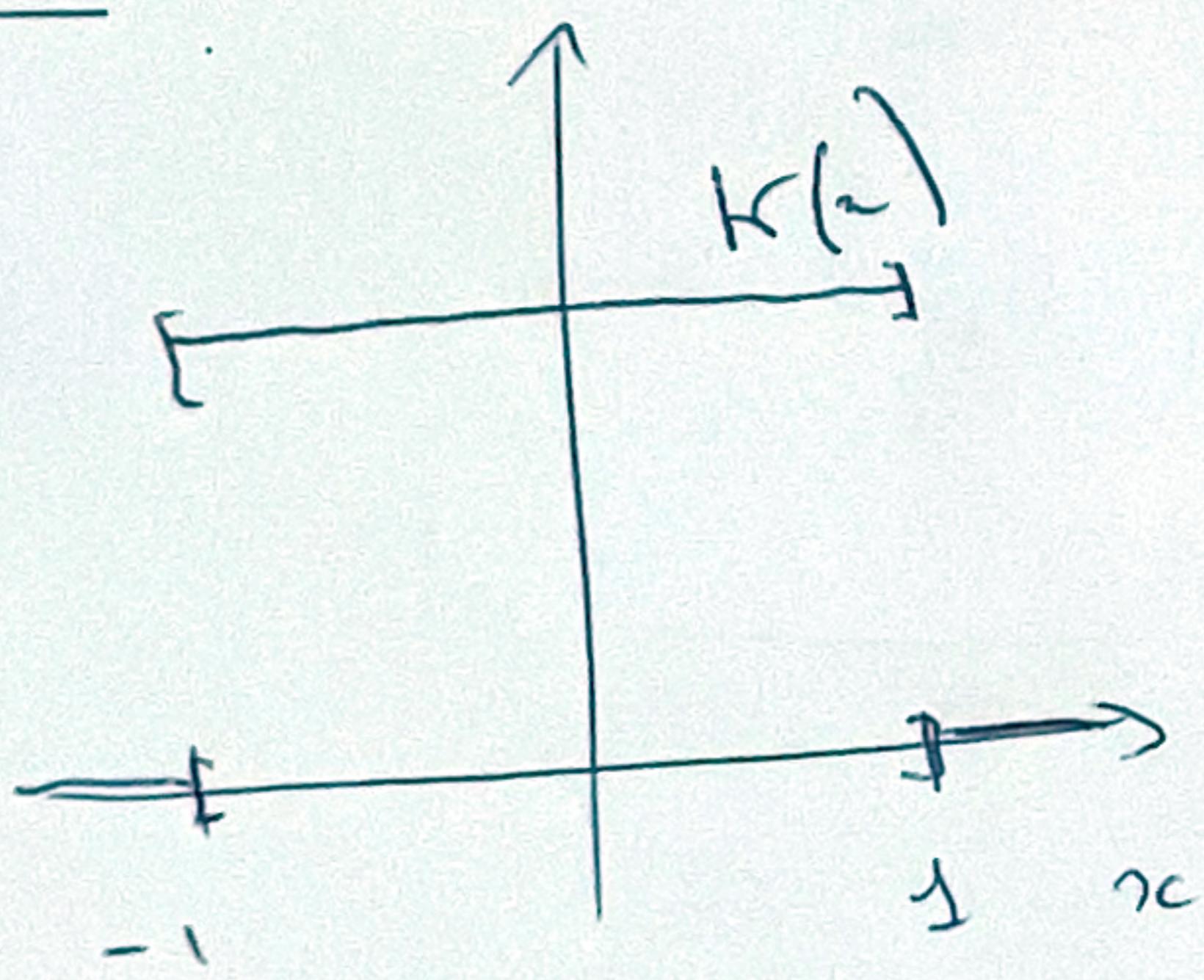
$$\frac{1}{n^k} \left( \sum_{i=1}^n \int_{B_d} k(v_i) t^d d\mu \right) = \frac{1}{n^k} \sum_{i=1}^n \int_{B_d} k(v_i) d\mu$$

$$\int_{\mathbb{R}^d} f_h(z) du = \frac{1}{n h^d} \sum_{i=1}^n \int_{\mathbb{R}^d} k(u_i) h^d du; \quad \text{Jacobien du changement de variable.}$$

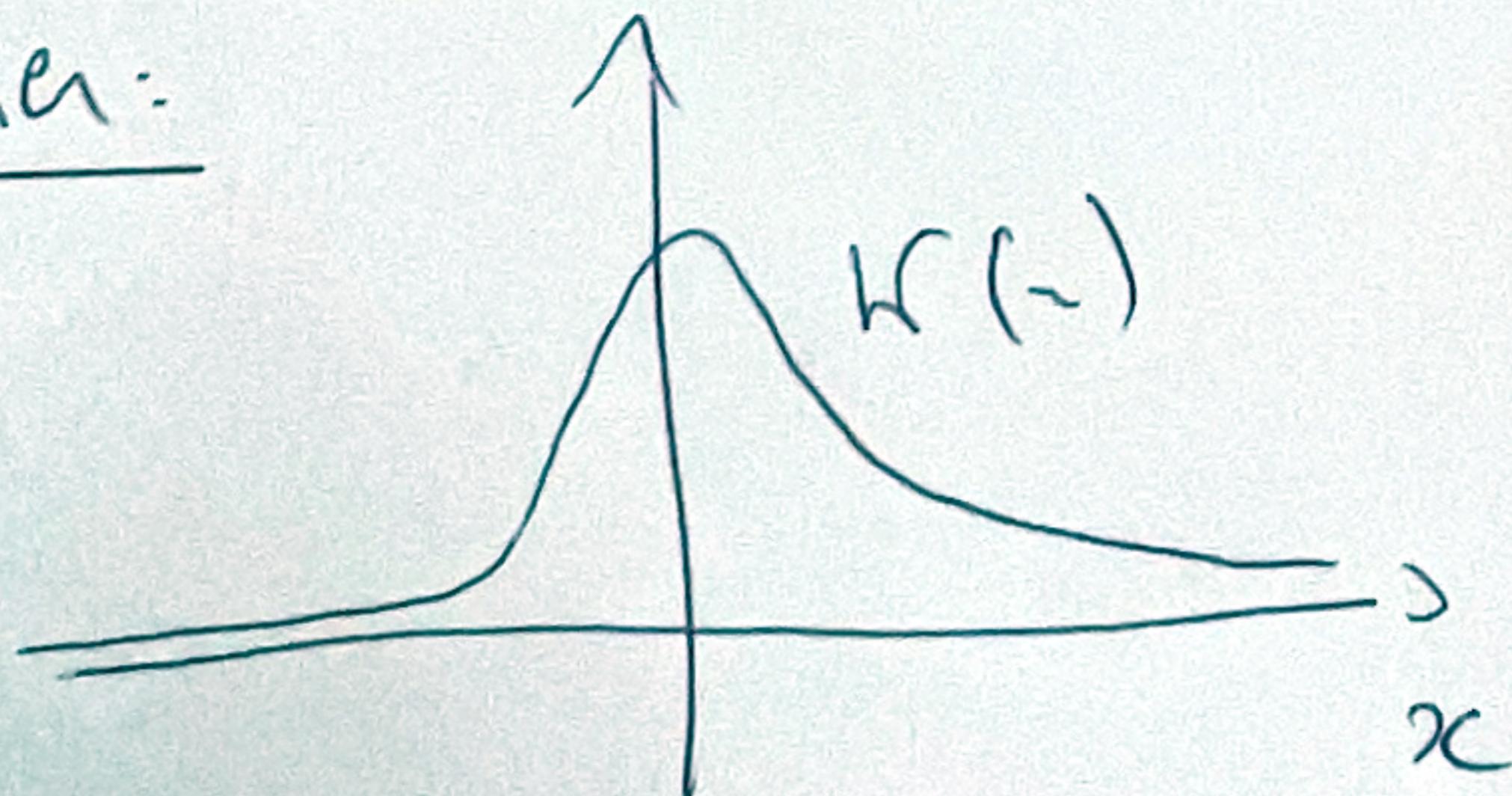
$$= \frac{1}{n} \sum_{i=1}^n \underbrace{\int_{\mathbb{R}^d} k}_{=1}$$

$$= 1.$$

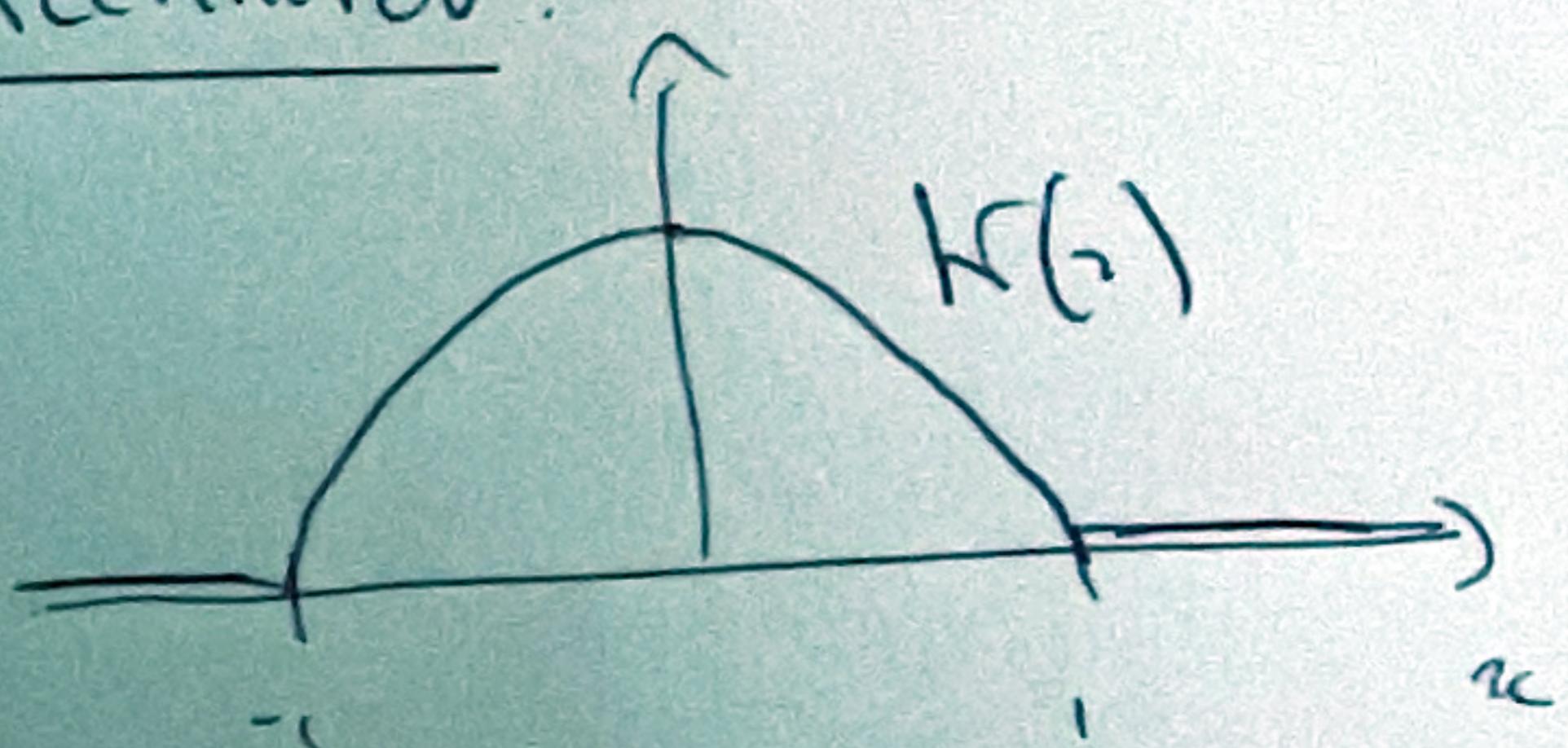
8) Uniforme:



Gaussian:



Épanechnikov:



## • L'angle $\theta$ est grand:

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

## • L'angle $\theta$ est petit :

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

g)



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

~~régression~~  
classification

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

D

variables .. - - -

8

Exercice 2:

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

$$\begin{aligned} \sum_{i,j} \alpha_i \alpha_j P(x_i, x_j) &= \sum_{i,j} \alpha_i \alpha_j \langle \phi(x_i), \phi(x_j) \rangle \quad (\text{par définition}) \\ &= \left\langle \sum_i \alpha_i \phi(x_i), \sum_j \alpha_j \phi(x_j) \right\rangle \quad (\text{bilinearité}) \\ &= \left\| \sum_i \alpha_i \phi(x_i) \right\|^2 \quad (\text{définition de } P \text{ norme}) \\ &\geq 0 \quad (\text{car carré toujours positif}). \end{aligned}$$

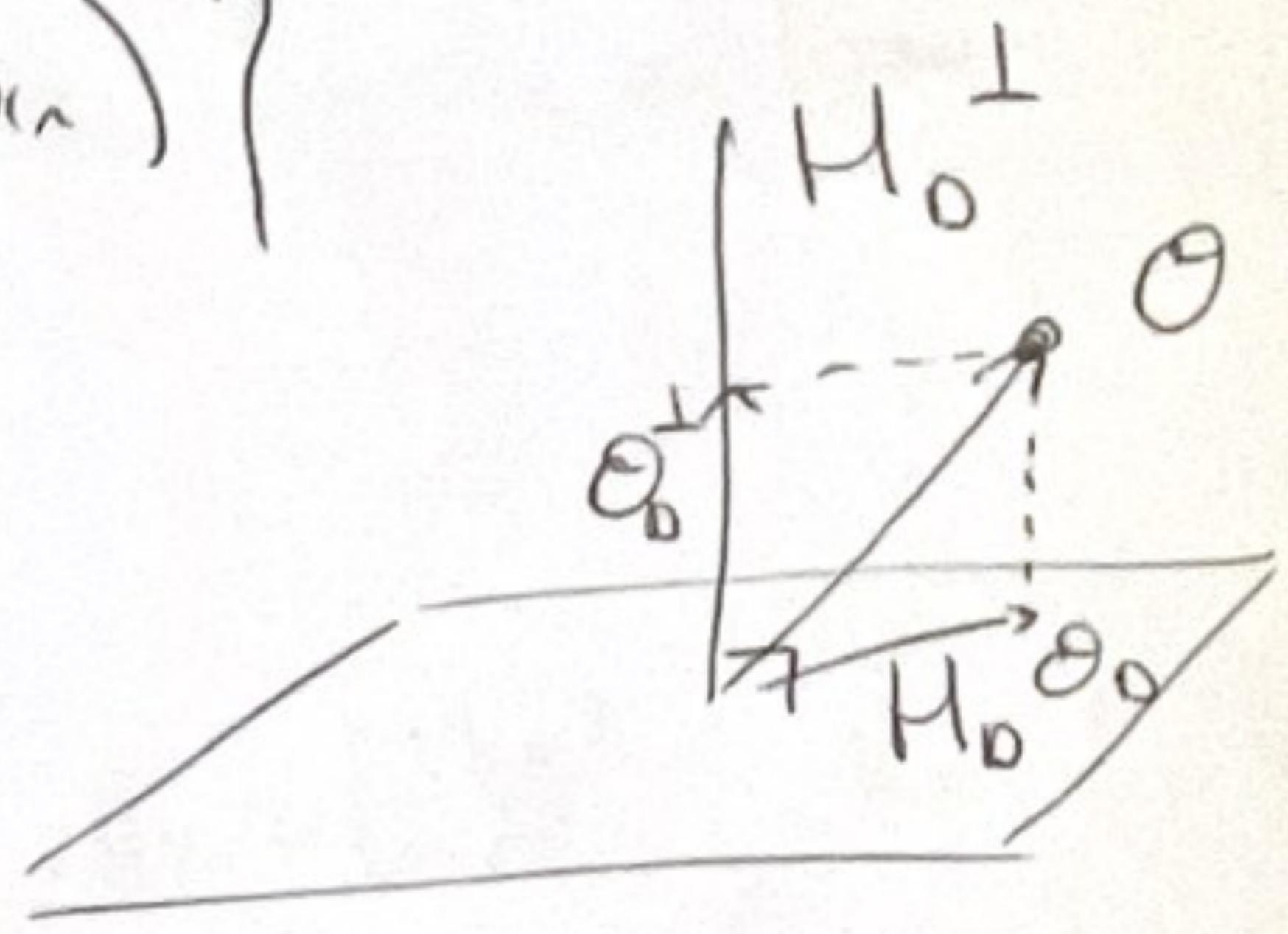
2) D'après le théorème d'Aronszajn, si  $P$  est un noyau symétrique positif, alors il existe un espace de Hilbert et  $\phi: X \rightarrow H$  tel que

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

Donc  $P$  est un noyau pas un produit scalaire et une fonction  $\phi$  n'est pas restreinte, il y a équivalence entre les deux.

3) Notons  $H_0 = \text{Vect}\{\phi(x_1), \dots, \phi(x_n)\}$

Sat  $\theta \in H$ ,  $\theta = \theta_0 + \theta_0^\perp$   
 $\uparrow \quad \uparrow$   
 $H_0 \quad H_0^\perp$



$$\text{Notons } M_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log \left( \frac{1 - \gamma_n(\theta, \phi(x_i))}{1 + e^{-\gamma_n(\theta, \phi(x_i))}} \right) + \|\theta\|^2$$

$$\begin{aligned} \text{Alors } M_n(\theta) &= \frac{1}{n} \sum_{i=1}^n \log \left( \frac{1 - \gamma_n(\theta_0 + \theta_0^\perp, \phi(x_i))}{1 + e^{-\gamma_n(\theta_0 + \theta_0^\perp, \phi(x_i))}} \right) + \|\theta_0 + \theta_0^\perp\|^2 \\ &= \frac{1}{n} \sum_{i=1}^n \underbrace{\log \left( \frac{1 - \gamma_n(\theta_0, \phi(x_i)) - \gamma_n(\theta_0^\perp, \phi(x_i))}{1 + e^{-(\gamma_n(\theta_0, \phi(x_i)) + \gamma_n(\theta_0^\perp, \phi(x_i)))}} \right)}_{\text{par Pythagore}} + \underbrace{\|\theta_0\|^2 + \|\theta_0^\perp\|^2}_{\text{par Pythagore}}. \end{aligned}$$

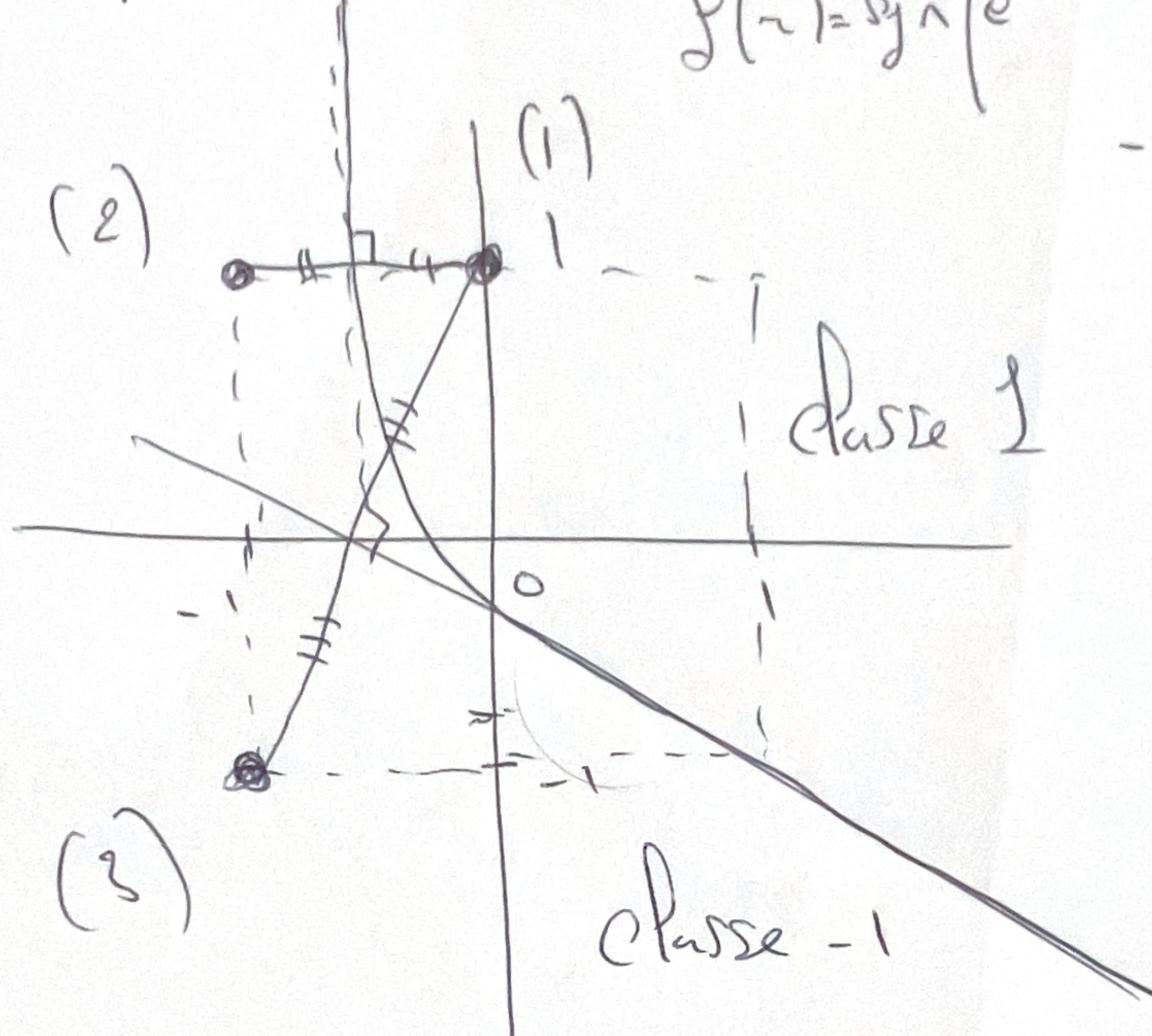
$$\text{or, } \forall i, \langle \theta_0^\perp, \phi(x_i) \rangle = 0 \text{ car } \theta_0^\perp \in \text{Vect}(\phi(x_1), \dots, \phi(x_n))^\perp.$$

$$\text{donc } \boxed{M_n(\theta) = M_n(\theta_0) + \|\theta_0^\perp\|^2}$$

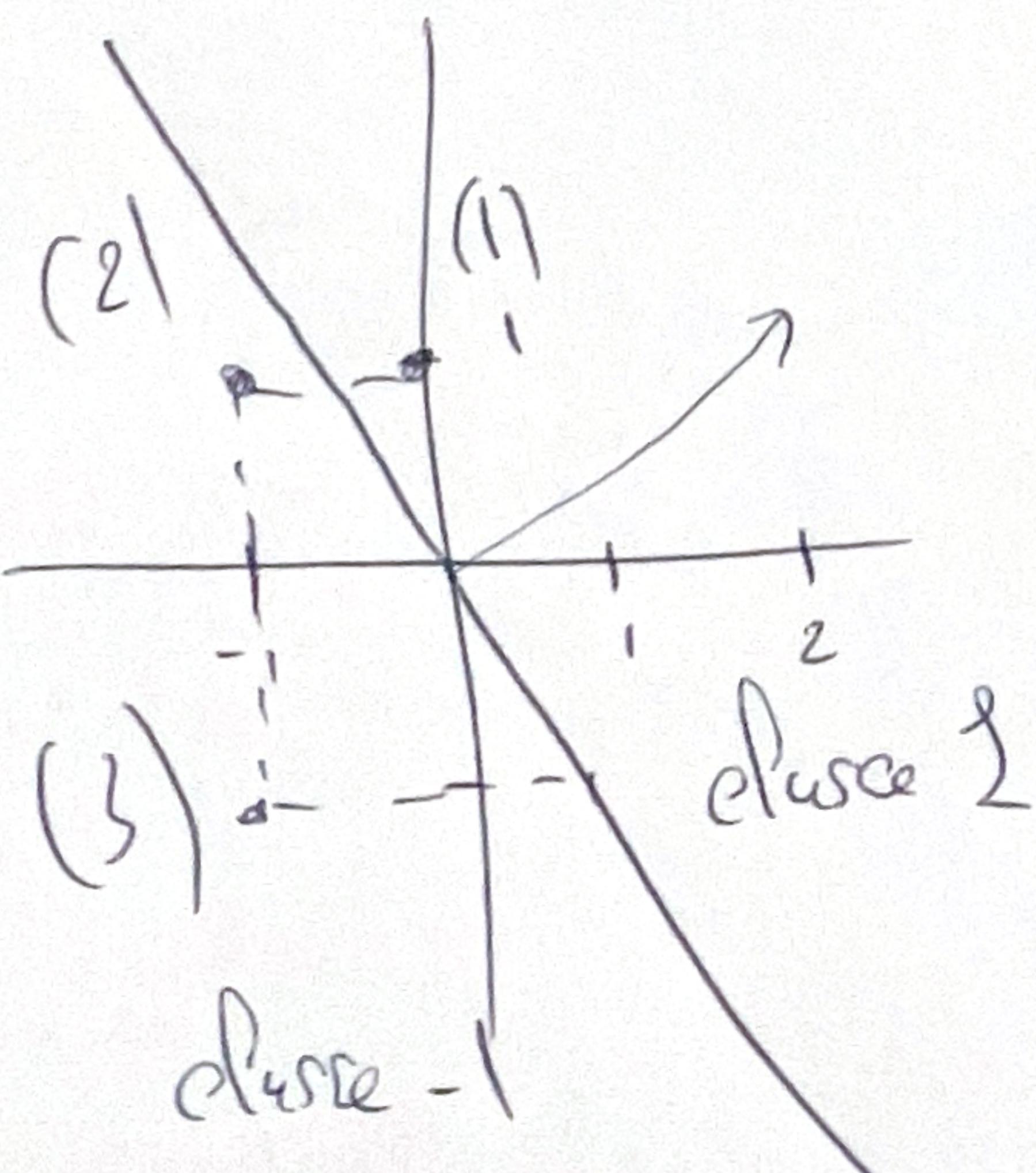
Une solution optimale (si elle existe) a donc une composante dans l'orthogonal de  $H_0$  nulle. On peut se restreindre à chercher dans  $H_0$ .

Sur  $H_2$ , le problème est portement convexe (grâce au terme  $\lambda \|\theta\|^2$ ). Il admet donc une unique solution.

a) (noyau exponentiel)



(noyau linéaire)



$$= \text{sgn}\left(\langle x, (0) \rangle - \langle x, (1) \rangle - \langle x, (2) \rangle\right)$$

5) Changement de variables  $\mathcal{O} = \alpha_1 \phi(x_1) + \dots + \alpha_n \phi(x_n)$

$$\begin{aligned}
 M_n(\alpha_1, \dots, \alpha_n) &= -\frac{1}{n} \sum_{i=1}^n \text{Pay} \left( 1 + \frac{-y_i (\sum_j \alpha_j \phi(x_j), \phi(x_i))}{1 + e^{-y_i (\sum_j \alpha_j \phi(x_j), \phi(x_i))}} \right) + \lambda \left\| \sum_j \alpha_j \phi(x_j) \right\|^2 \\
 &= \frac{1}{n} \sum_{i=1}^n \text{Pay} \left( 1 + \frac{-y_i (\sum_j \alpha_j \phi(x_j), \phi(x_i))}{1 + e^{-y_i (\sum_j \alpha_j \phi(x_j), \phi(x_i))}} \right) + \lambda \left( \sum_i \alpha_i \phi(x_i), \sum_j \alpha_j \phi(x_j) \right) \\
 &= \frac{1}{n} \sum_{i=1}^n \text{Pay} \left( 1 + \frac{-y_i (K\alpha)_i}{1 + e^{-y_i (K\alpha)_i}} \right) + \lambda \alpha^T K \alpha \\
 &\quad \text{où } K = \left( b(x_i, x_j) \right)_{i,j}.
 \end{aligned}$$

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

IP peut être capable de calculer

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

ce qui donne

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