Modèles Linéaires

I. Rappels sur les applications linéaires - Paramétrisation des modèles linéaires.

1) Applications Linéaire.

<u>Définition :</u> $f : \mathbb{R}^d \longrightarrow \mathbb{R}^{d'}$ est linéaire si

$$\forall x, \ f(x) = M_f \, x \quad \text{où } M_f \in \mathbb{R}^{d' \times d}$$

est la matrice associée à f.

<u>Remarque :</u> On a alors,
$$M_f = \begin{array}{c} \left(\begin{array}{c|c|c|c} f(e_1) & f(e_2) & \cdots & f(e_d) \end{array}\right) \end{array} \begin{array}{l} \leftarrow e'_1 \\ \leftarrow e'_2 \\ \vdots \\ \leftarrow e'_{d'} \end{array}$$

où $(e_1, \dots, e_d)$ est la base canonique de $\mathbb{R}^d$ et $(e'_1, \dots, e'_{d'})$ est la base canonique de $\mathbb{R}^{d'}$.

<u>Définition :</u> Le rang de f est la dimension du sous-espace vectoriel de $\mathbb{R}^{d'}$ engendré par les colones de M_f, qui est la même que la dimension du sous-espace de $\mathbb{R}^d$ engendré par les lignes de M_f.

<u>Théorème (du rang) :</u> $\dim(\mathbb{R}^d) = \dim(\ker(f)) + \dim(\mathrm{Im}(f))$.

2) Le cas des formes linéaires

<u>Définition</u> : Lorsque $d' = 1$, on parle de forme linéaire.

Comme toute application linéaire f peut être décomposée comme

$f = f_1 e_1' + f_2 e_2' + \dots + f_{d'} e_{d'}'$, on va s'intéresser

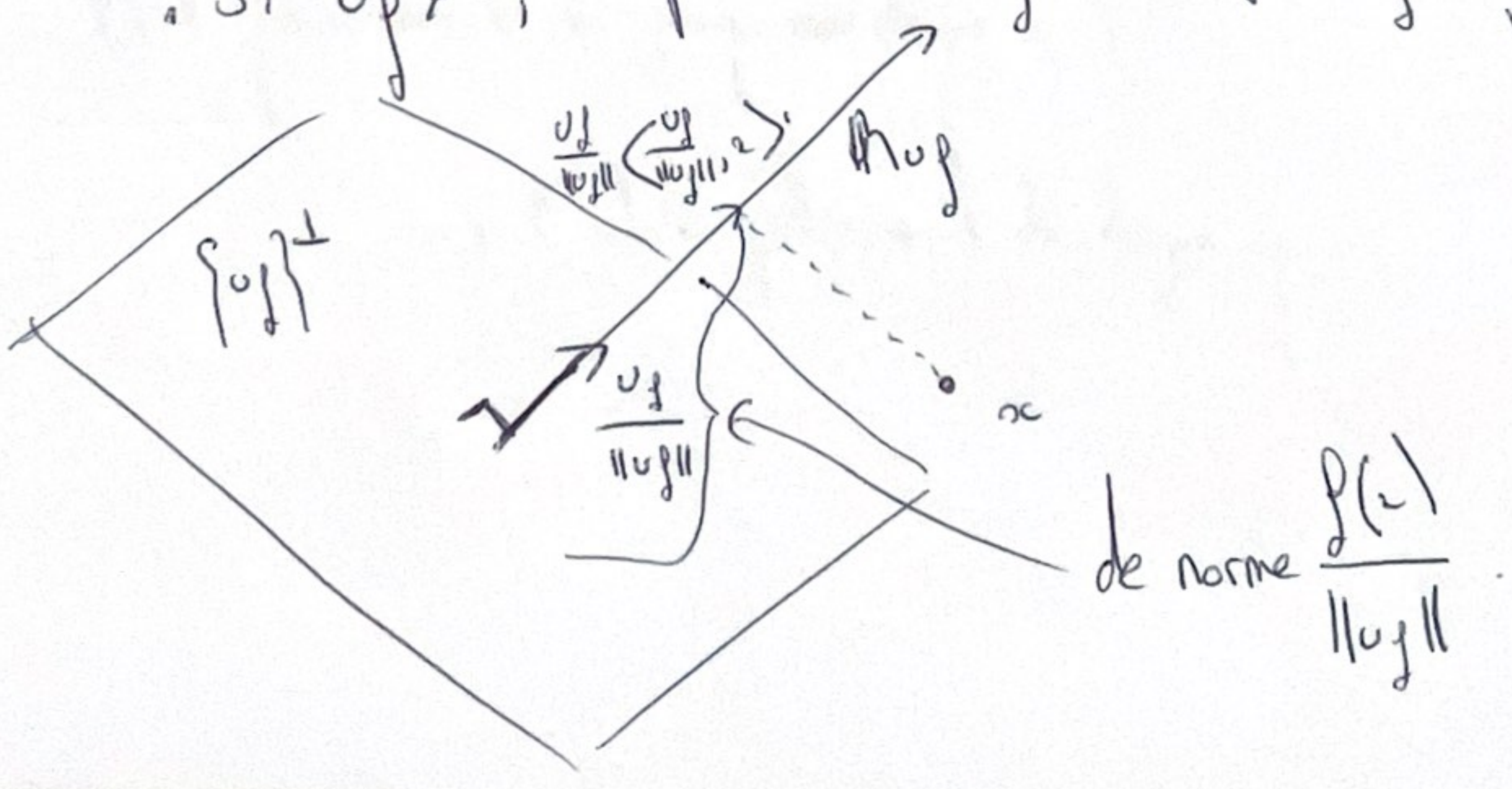
au cas des formes linéaires.

<u>Proposition</u> : Si f est une forme linéaire, il existe $u_f \in \mathbb{R}^d$

tel que $\forall x,\ f(x) = u_f^T x = \langle u_f, x \rangle$

<u>Interprétation Géométrique</u> :

- Si $u_f = 0$, la fonction est nulle.
- Si $u_f \neq 0$, on peut écrire $f(x) = \|u_f\| \langle \frac{u_f}{\|u_f\|}, x \rangle$ $\forall x$.



3) Reparamétrisation et feature maps


Il est souvent possible, lorsqu'une fonction $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ n'est pas linéaire d'écrire

$$f = g \circ \varphi \quad \text{où} \quad \varphi : \mathbb{R}^d \longrightarrow \mathbb{R}^{d''}$$

est appelée une feature map, et où

$g : \mathbb{R}^{d''} \to \mathbb{R}^{d'}$ est linéaire.

Exemples:

- $\varphi(x) = x$ → Modèles linéaires classiques.
- $\varphi(x) = \begin{pmatrix} x \\ 1 \end{pmatrix}$ → Modèle affine classique (souvent simplement appelé modèle linéaire).
- $\varphi(x) = \big($monomes en x jusqu'au degré $q\big)$

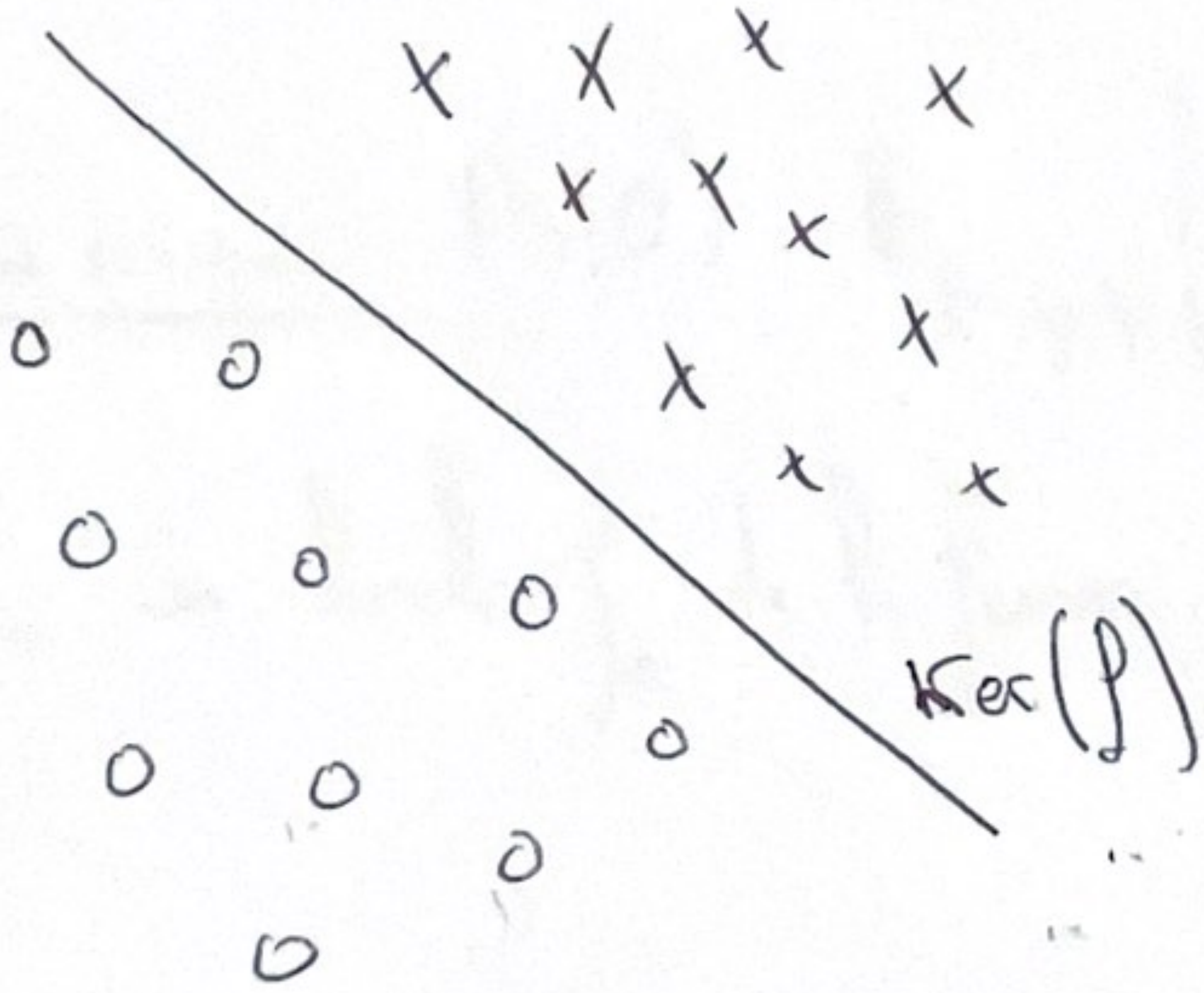
  → Modèle polynômial de degré q.

4) Séparabilité Linéaire.

Soient $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \{\underbrace{-1}_{\substack{\text{Identifié} \\ \text{à "O"}}}, \underbrace{1}_{\substack{\text{Identifié} \\ \text{à "X"}}}\}$

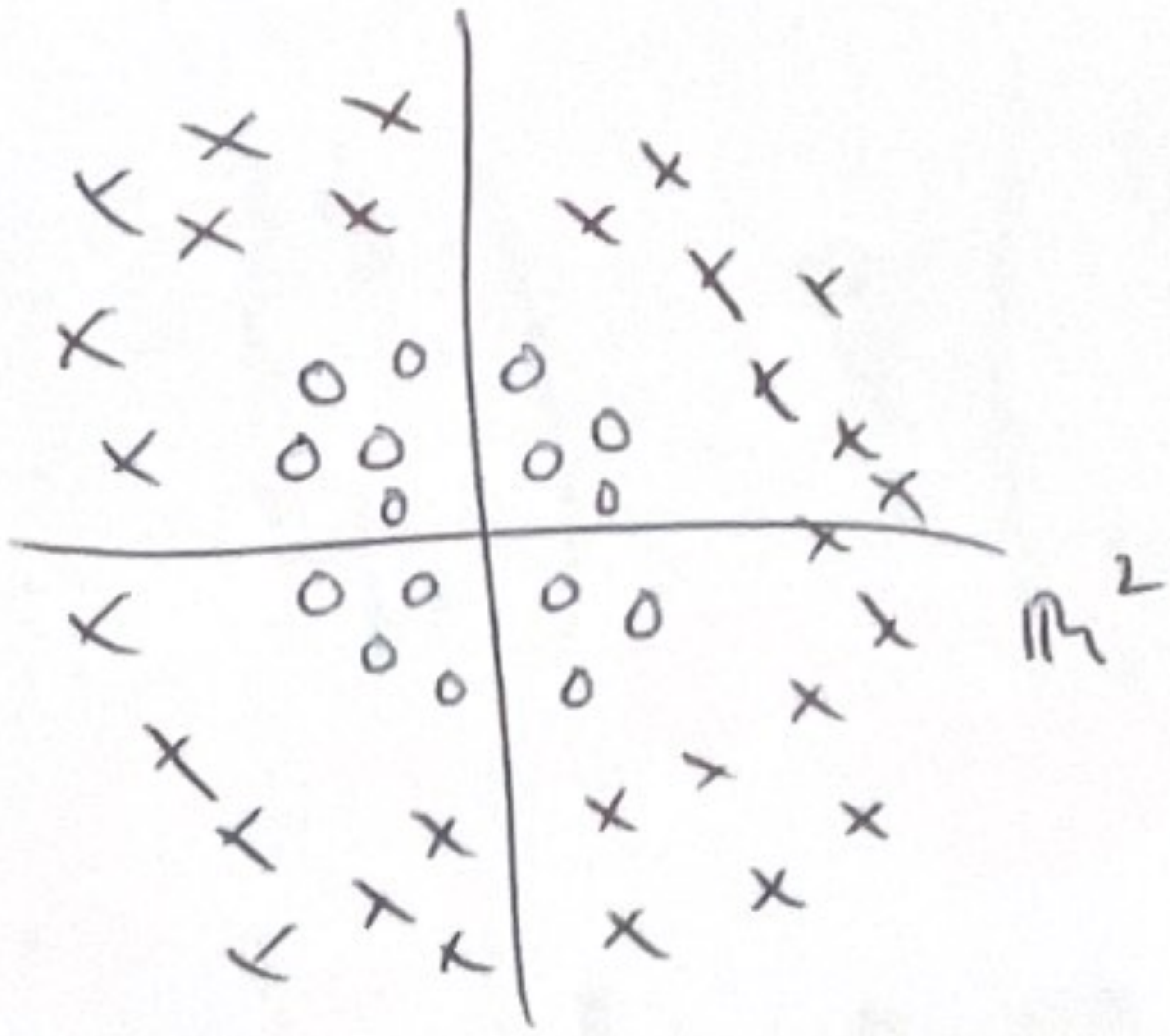
On dit que ces points sont séparable linéairement si il existe une application linéaire $f : \mathbb{R}^d \to \mathbb{R}$ telle que

$$\forall i, \quad y_i f(x_i) > 0.$$



Exercice: Trouver un ensemble de points qui ne sont pas séparables linéairement mais qui le deviennent en introduisant une "feature map"

Solution :



$$\varphi(x) = \begin{pmatrix} \|x\| \\ 1 \end{pmatrix} \in \mathbb{R}^2$$

rend séparable les pts linéairement.

II. Régression Linéaire.

Contexte : $(x_1, y_1), \dots, (x_n, y_n) \in \underbrace{\mathcal{X}}_{\subset \mathbb{R}^d} \times \underbrace{\mathcal{Y}}_{\subset \mathbb{R}}$.

Minimisation du risque $\mathcal{R}(\theta) = \mathbb{E}_{(x,y)}\left((y - f_\theta(x))^2 \right), \theta \in (H)$

Modèles linéaires : On cherche f_θ de la forme $f_\theta(x) = \theta^T \varphi(x)$

1) Moindres Carrés - Minimisation du Risque Empirique.

Empirical Risk Minimization (ERM)

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^{d'}}{\operatorname{argmin}} \hat{\mathcal{R}}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \theta^T \varphi(x_i) \right)^2$$

Notation Matricielle

$$y := \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \Phi = \begin{pmatrix} \varphi(x_1) \\ \vdots \\ \varphi(x_n) \end{pmatrix}$$

$$\hat{R}(\theta) = \frac{1}{n} \|y - \Phi\theta\|_2^2$$

Question: Comment trouver $\hat{\theta} \in \underset{\theta}{\operatorname{argmin}}\, \hat{R}(\theta)$?

Hypothèse 1: Les colonnes de Φ sont indépendantes

$\iff$ Il n'y a pas de dépendance linéaire entre les features.

En pratique, ce n'est souvent pas un problème si n est assez grand.

Définition - Proposition:

Sous l'hypothèse 1, $\hat{R}(\cdot)$ admet un unique minimiseur sur $\mathbb{R}^d$ appelé estimateur des moindres carrés ordinaires ($\hat{\theta}$).

Il satisfait $\quad \Phi^T \Phi \hat{\theta} = \Phi^T y \quad$ (Équation Normale)

Donc $\quad \hat{\theta} = (\Phi^T \Phi)^{-1} \Phi^T y$.

Preuve Algébrique :

La fonction $\theta \mapsto \hat{R}(\theta)$ est différentiable sur $\mathbb{R}^d$ qui est ouvert.

Alors $\nabla \hat{R}(\hat{\theta}) = 0$ est une condition nécessaire à l'optimalité de $\hat{\theta}$.

Or, $\nabla \hat{R}(\theta) = \frac{2}{n} \Phi^T (\Phi \theta - y) \; \forall \theta$

ce qui donne $\Phi^T \Phi \hat{\theta} = \Phi^T y$ (équation normale).

Cette équation admet une unique solution, donc il existe un unique θ qui satisfait les conditions nécessaires d'optimalité.

Pour prouver que ces conditions sont suffisantes, il y a trois possibilités (au moins).

- $\nabla^2 \hat{R}(\theta) = \frac{2}{n} \Phi^T \Phi \succcurlyeq 0$

  $\Rightarrow \hat{\theta}$ est un minimum local donc global car c'est le seul à vérifier la condition nécessaire.

- $\hat{R}(.)$ est convexe donc les conditions nécessaires deviennent suffisantes

- Coercivité $\Rightarrow$ on peut se ramener à l'étude sur un compact.

2) Le cas où l'hypothèse 1 n'est pas satisfaite

Si l'hypothèse 1 n'est pas satisfaite, $\ker \Phi \neq 0$ et l'équation normale peut ne pas avoir de solution.

Par exemple, lorsqu'on peut écrire $\hat{\theta} = (\Phi^T \Phi)^{-1} \Phi^T y$ et que $\lambda_{min}(\Phi^T \Phi) \to 0$, $\hat{\theta}$ "part à l'∞".

Pour résoudre ce problème, on "force" $\hat{\theta}$ à rester proche de 0 en pénalisant par sa norme.

$$\hat{\theta}_\lambda \in \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \|y - \Phi\theta\|_2^2 + \lambda \|\theta\|_2^2$$

$$\Rightarrow \hat{\theta}_\lambda = \frac{1}{n} \left(\frac{1}{n} \Phi^T \Phi + \lambda I \right)^{-1} \Phi^T y$$

<u>Estimateur "Ridge"</u>.

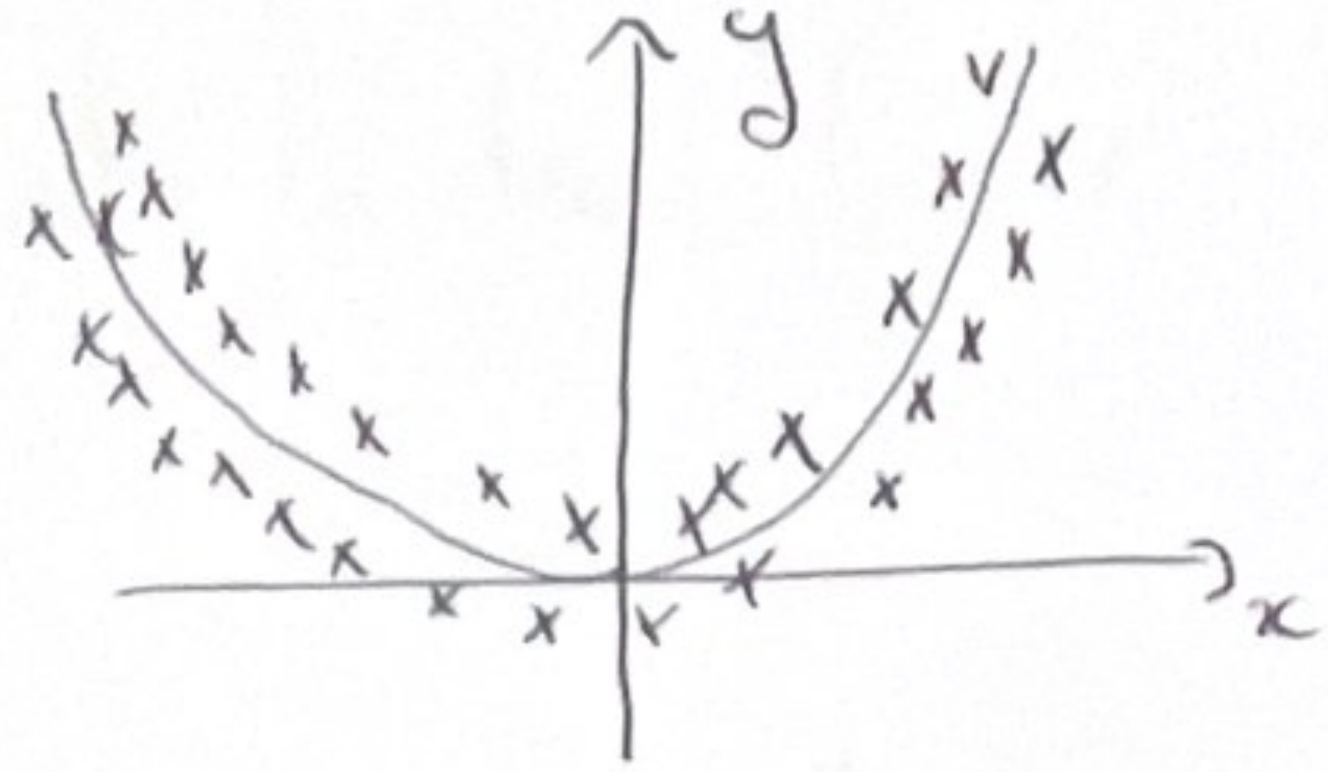
<u>Remarque</u> : En pratique c'est très utile lorsque :

- d est "grand" par rapport à n
- les features sont proches d'une situation de dépendance linéaire

3) Compromis sur/sous apprentissage.

Exemple: $y = x^2 + \text{bruit}$

centré.



Considérons : $\varphi_q(x) = \begin{pmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^q \end{pmatrix}$.

Cas extrême 1 : $q = 0$

Le modèle ne peut apprendre que des droites horizontales

=> Il n'est pas assez expressif pour comprendre le modèle

=> On parle de sous-apprentissage.

Cas extrême 2 : $q = n-1$

Il existe un polynôme interpolateur (de Lagrange) qui passe par tous les points

=> Le modèle est trop expressif et apprend tout le bruit des données

=> On parle de sur-apprentissage.

En pratique, il faut prendre un q entre les deux (ici 2 marchera).

III Modèles Linéaires pour la Classification - Régression Logistique.

3) Maximum de vraissemblance - Régression logistique

Fonction sigmoïde : $\sigma(z) := \frac{1}{1+e^{-x}}$

Modèle logistique : $p(y=1|x) = \frac{1}{1+e^{-\theta^T x}}$ $\quad y \in \{-1, 1\}$

$= \sigma(\theta^T x) = \sigma(y\theta^T x)$

remarque $p(y=-1) = 1 - \sigma(\theta^T x)$ (10)

$$= 1 - \frac{1}{1+e^{-\theta^T x}} = \frac{e^{-\theta^T x}}{1+e^{-\theta^T x}} = \frac{1}{1+e^{\theta^T x}}$$

$$= \sigma(y\theta^T x)$$

donc $p(y|x) = \sigma(y\theta^T x)$.

Maximum de vraissemblance :

$$\hat{\theta} \in \underset{\theta}{\arg\max} \sum_{i=1}^{n} \ln\left(\frac{1}{1+e^{-y\theta^T x}}\right)$$

$$= \underset{\theta}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \ln\left(1+e^{-y\theta^T x}\right)$$

Remarque : Il s'agit de l'estimateur du minimum du risque empirique pour la perte

$$\ell_{logistique}(y, z) = \ln\left(1+e^{-yz}\right).$$