



# IFSBM Module 11



**IFSBM**  
INSTITUT DE FORMATION  
SUPÉRIEURE BIOMÉDICALE

Méthodes de classification supervisée

Yoann Pradat

# Qui suis-je?

Education Projects/internships

## Preparatory class

CPGE Lycée-Louis-Le-Grand  
Paris

## General engineering school

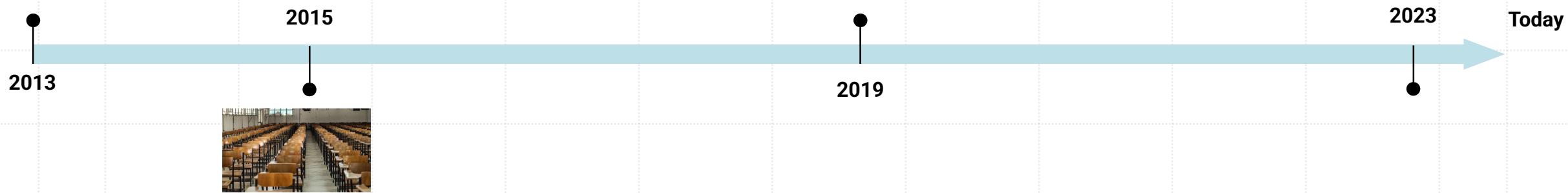
Mines ParisTech  
Paris

## MVA master's degree

ENS Paris-Saclay  
Paris area

## PhD

CentraleSupélec & Gustave Roussy  
Paris area



French national exams for  
engineering schools!

# Qui suis-je?

Education  
Projects/internships

## Preparatory class

CPGE Lycée-Louis-Le-Grand  
Paris

## General engineering school

Mines ParisTech  
Paris

## MVA master's degree

ENS Paris-Saclay  
Paris area

## PhD

CentraleSupélec & Gustave Roussy  
Paris area



## Natixis, London

Equity derivatives pricing

## MSKCC, New-York

Papaemmanuil's lab  
Analysis of a large-scale cohort (>2,500) of AML patients

## G-Research, London

Hedge fund

## EMBL-EBI, Hinxton

Ulhmann's lab  
Studied spline-based models for the 3D representation of surfaces with a sphere topology.

# Qui suis-je?

Education  
Projects/internships  
Postdoc

## Preparatory class

CPGE Lycée-Louis-Le-Grand  
Paris

## General engineering school

Mines ParisTech  
Paris

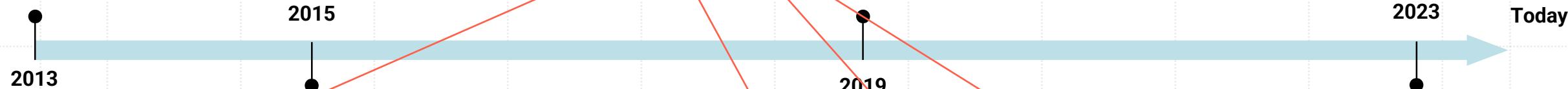
## MVA master's degree

ENS Paris-Saclay  
Paris area

## PhD

CentraleSupélec & Gustave Roussy  
Paris area

Gustave Roussy  
Elsa Bernard  
team



## Natixis, London

Equity derivatives  
pricing

## MSKCC, New-York

Papaemmanuil's lab  
Analysis of a large-scale  
cohort (>2,500) of AML  
patients

## G-Research, London

Hedge fund

## EMBL-EBI, Hinxton

Ulhmann's lab  
Studied spline-based models for  
the 3D representation of surfaces  
with a sphere topology.

# Sommaire

---

- 1. Données et modélisation
  - 1. Le jeu de données
  - 2. L'estimation (model fitting)
- 2. Modèles de régression
  - 1. Régression logistique binaire
  - 2. Régression logistique multinomiale
- 3. Réseaux de neurones
  - 1. Réseau monocouche
  - 2. Réseau multicouche
  - 3. Réseau convolutionnel

# 1.1 Le jeu de données

---

## Notations

$X : \Omega \mapsto \mathcal{X}$  : variable aléatoire (scalaire ou vectorielle)

$x \in \mathcal{X}$  : une observation de la variable aléatoire

### Exemples:

1. On lance un dé,  $\mathcal{X} = \{1, 2, \dots, 6\}$

# 1.1 Le jeu de données

---

## Notations

$X : \Omega \mapsto \mathcal{X}$  : variable aléatoire (scalaire ou vectorielle)

$x \in \mathcal{X}$  : une observation de la variable aléatoire

### Exemples:

1. On lance un dé,  $\mathcal{X} = \{1, 2, \dots, 6\}$

2. On mesure la longueur et la largeur de fleurs  $\mathcal{X} = \mathbb{R}^+ \times \mathbb{R}^+$

# 1.1 Le jeu de données

## Notations

$X : \Omega \mapsto \mathcal{X}$  : variable aléatoire (scalaire ou vectorielle)

$x \in \mathcal{X}$  : une observation de la variable aléatoire

### Exemples:

1. On lance un dé,  $\mathcal{X} = \{1, 2, \dots, 6\}$

2. On mesure la longueur et la largeur de fleurs  $\mathcal{X} = \mathbb{R}^+ \times \mathbb{R}^+$

3. On mesure le niveau d'expression de gènes relatif à 1M (K gènes)

$$\mathcal{X} = [0, 1M]^K = [0, 1M] \times [0, 1M] \times \dots \times [0, 1M]$$

# 1.1 Le jeu de données

## Notations

$X : \Omega \mapsto \mathcal{X}$  : variable aléatoire (scalaire ou vectorielle)

$x \in \mathcal{X}$  : une observation de la variable aléatoire

### Exemples:

1. On lance un dé,  $\mathcal{X} = \{1, 2, \dots, 6\}$

2. On mesure la longueur et la largeur de fleurs  $\mathcal{X} = \mathbb{R}^+ \times \mathbb{R}^+$

3. On mesure le niveau d'expression de gènes relatif à 1M (K gènes)

$$\mathcal{X} = [0, 1M]^K = [0, 1M] \times [0, 1M] \times \dots \times [0, 1M]$$

$n \in \mathbb{N}^*$  : nombre d'observations (=individus, échantillons)

$p \in \mathbb{N}^*$  : nombre de variables (=covariables, prédicteurs, features)

$x_{1:n} = (x_1, \dots, x_n)$  : ensemble d'observations (=dataset)

# 1.1 Le jeu de données

---

Exemple:

1. On réalise un essai clinique de réponse à un traitement sur 100 patients

$X = (\text{Age, Poids, Nb Tx antérieurs,}$   
 $\text{Exp gene 1, Exp gene 2, Mutation gene 1,}$   
 $\text{Temps avant rechute, Meilleure réponse})$

# 1.1 Le jeu de données

---

Exemple:

1. On réalise un essai clinique de réponse à un traitement sur 100 patients

$X = (\text{Age, Poids, Nb Tx antérieurs,}$   
 $\text{Exp gene 1, Exp gene 2, Mutation gene 1,}$   
 $\text{Temps avant rechute, Meilleure réponse})$

$x_1 = (45.2, 78.2, 3, 1032, 258, 1, 85, \text{PR})$

$x_2 = (81, 63, 6, 589, 903, 0, 390, \text{SD})$

...

# 1.1 Le jeu de données

Exemple:

1. On réalise un essai clinique de réponse à un traitement sur 100 patients

$X = (\text{Age, Poids, Nb Tx antérieurs,}$   
 $\text{Exp gene 1, Exp gene 2, Mutation gene 1,}$   
 $\text{Temps avant rechute, Meilleure réponse})$

$x_1 = (45.2, 78.2, 3, 1032, 258, 1, 85, \text{PR})$   
 $x_2 = (81, 63, 6, 589, 903, 0, 390, \text{SD})$   
...



$(x_1, x_2, \dots, x_{100})$   
= jeu de données

Objectifs

1. **Proposer un modèle mathématique pour modéliser une variable en fonction d'autres.** Exemple Meilleure réponse vs (Exp gene 1, Exp gene 2, Age)  
Modèle = Régression logistique multinomiale

# 1.1 Le jeu de données

Exemple:

1. On réalise un essai clinique de réponse à un traitement sur 100 patients

$X = (\text{Age, Poids, Nb Tx antérieurs,}$   
 $\text{Exp gene 1, Exp gene 2, Mutation gene 1,}$   
 $\text{Temps avant rechute, Meilleure réponse})$

$x_1 = (45.2, 78.2, 3, 1032, 258, 1, 85, \text{PR})$   
 $x_2 = (81, 63, 6, 589, 903, 0, 390, \text{SD})$   
...

$\left. \begin{array}{l} \\ \\ \dots \end{array} \right\} (x_1, x_2, \dots, x_{100})$   
= jeu de données

Objectifs

1. **Proposer un modèle mathématique pour modéliser une variable en fonction d'autres.** Exemple Meilleure réponse vs (Exp gene 1, Exp gene 2, Age)  
Modèle = Régression logistique multinomiale
2. **Estimer les paramètres du modèle à partir de**  $(x_1, x_2, \dots, x_{100})$

# 1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes  
->  $\mathcal{V}$  = volume  $(\mathcal{G}^1, \dots, \mathcal{G}^{5000})$  profil d'expression

$v_{1:n}, g_{1:n}^1, \dots, g_{1:n}^{5000}$  observations

- > on prend comme modèle un *modèle de regression linéaire*

$$\begin{aligned}\mathcal{V} &= f_\theta(\mathcal{G}^1, \mathcal{G}^{5000}) \\ &= \theta_1 \mathcal{G}^1 + \dots + \theta_{5000} \mathcal{G}^{5000}\end{aligned}$$

Schématique, pas mathématiquement exact



# 1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes  
->  $\mathcal{V}$  = volume  $(\mathcal{G}^1, \dots, \mathcal{G}^{5000})$  profil d'expression

$v_{1:n}, g_{1:n}^1, \dots, g_{1:n}^{5000}$  observations

- > on prend comme modèle un *modèle de regression linéaire*

$$\begin{aligned}\mathcal{V} &= f_\theta(\mathcal{G}^1, \mathcal{G}^{5000}) \\ &= \theta_1 \mathcal{G}^1 + \dots + \theta_{5000} \mathcal{G}^{5000}\end{aligned}$$

Schématique, pas mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi “coller au mieux aux données” ?

Mesure de distance entre l'observation  $v_i$  et la prediction  $\hat{v}_i = f_\theta(g_i^1, \dots, g_i^{5000})$  ?

# 1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes  
->  $\mathcal{V}$  = volume  $(\mathcal{G}^1, \dots, \mathcal{G}^{5000})$  profil d'expression

$v_{1:n}, g_{1:n}^1, \dots, g_{1:n}^{5000}$  observations

- > on prend comme modèle un *modèle de regression linéaire*

$$\begin{aligned}\mathcal{V} &= f_\theta(\mathcal{G}^1, \mathcal{G}^{5000}) \\ &= \theta_1 \mathcal{G}^1 + \dots + \theta_{5000} \mathcal{G}^{5000}\end{aligned}$$

Schématique, pas mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi “coller au mieux aux données” ?

Mesure de distance entre l'observation  $v_i$  et la prediction  $\hat{v}_i = f_\theta(g_i^1, \dots, g_i^{5000})$  ?

$$d(v_i, \hat{v}_i) = (v_i - \hat{v}_i) ?$$

# 1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes  
->  $\mathcal{V}$  = volume ( $\mathcal{G}^1, \dots, \mathcal{G}^{5000}$ ) profil d'expression

$v_{1:n}, g_{1:n}^1, \dots, g_{1:n}^{5000}$  observations

- > on prend comme modèle un *modèle de regression linéaire*

$$\begin{aligned}\mathcal{V} &= f_\theta(\mathcal{G}^1, \mathcal{G}^{5000}) \\ &= \theta_1 \mathcal{G}^1 + \dots + \theta_{5000} \mathcal{G}^{5000}\end{aligned}$$

Schématique, pas mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi “coller au mieux aux données” ?

Mesure de distance entre l'observation  $v_i$  et la prediction  $\hat{v}_i = f_\theta(g_i^1, \dots, g_i^{5000})$  ?

$$d(v_i, \hat{v}_i) = (v_i - \hat{v}_i) ? \quad (v_i - \hat{v}_i)^2 ?$$

# 1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes  
->  $\mathcal{V}$  = volume ( $\mathcal{G}^1, \dots, \mathcal{G}^{5000}$ ) profil d'expression

$v_{1:n}, g_{1:n}^1, \dots, g_{1:n}^{5000}$  observations

- > on prend comme modèle un *modèle de regression linéaire*

$$\begin{aligned}\mathcal{V} &= f_\theta(\mathcal{G}^1, \mathcal{G}^{5000}) \\ &= \theta_1 \mathcal{G}^1 + \dots + \theta_{5000} \mathcal{G}^{5000}\end{aligned}$$

Schématique, pas mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi "coller au mieux aux données" ?

Mesure de distance entre l'observation  $v_i$  et la prediction  $\hat{v}_i = f_\theta(g_i^1, \dots, g_i^{5000})$  ?

$$d(v_i, \hat{v}_i) = (v_i - \hat{v}_i) ? \quad (v_i - \hat{v}_i)^2 ? \quad (v_i - \hat{v}_i)^4 ?$$

# 1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes  
->  $\mathcal{V}$  = volume  $(\mathcal{G}^1, \dots, \mathcal{G}^{5000})$  profil d'expression

$v_{1:n}, g_{1:n}^1, \dots, g_{1:n}^{5000}$  observations

- > on prend comme modèle un *modèle de regression linéaire*

$$\begin{aligned}\mathcal{V} &= f_\theta(\mathcal{G}^1, \mathcal{G}^{5000}) \\ &= \theta_1 \mathcal{G}^1 + \dots + \theta_{5000} \mathcal{G}^{5000}\end{aligned}$$

Schématique, pas mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi “coller au mieux aux données” ?

Mesure de distance entre l'observation  $v_i$  et la prediction  $\hat{v}_i = f_\theta(g_i^1, \dots, g_i^{5000})$  ?

$$d(v_i, \hat{v}_i) = (v_i - \hat{v}_i) ? \quad (v_i - \hat{v}_i)^2 ? \quad (v_i - \hat{v}_i)^4 ? \quad e^{|v_i - \hat{v}_i|} - 1 ?$$

# 1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes  
->  $\mathcal{V}$  = volume  $(\mathcal{G}^1, \dots, \mathcal{G}^{5000})$  profil d'expression

$v_{1:n}, g_{1:n}^1, \dots, g_{1:n}^{5000}$  observations

- > on prend comme modèle un *modèle de regression linéaire*

$$\begin{aligned}\mathcal{V} &= f_\theta(\mathcal{G}^1, \mathcal{G}^{5000}) \\ &= \theta_1 \mathcal{G}^1 + \dots + \theta_{5000} \mathcal{G}^{5000}\end{aligned}$$

Schématique, pas mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi “coller au mieux aux données” ?

Mesure de distance entre l'observation  $v_i$  et la prediction  $\hat{v}_i = f_\theta(g_i^1, \dots, g_i^{5000})$  ?

$$d(v_i, \hat{v}_i) = (v_i - \hat{v}_i) ? \quad (v_i - \hat{v}_i)^2 ? \quad (v_i - \hat{v}_i)^4 ? \quad e^{|v_i - \hat{v}_i|} - 1 ?$$

Estimation

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^n d(v_i, \hat{v}_i)$$

# 1.2 L'estimation

---

Quels modèles ? Le modèle peut avoir une interpretation probabiliste ou non.

-> si interpretation probabiliste

**fonction objectif = - maximum de vraisemblance**

Exemples: régression linéaire, régression logistique, classification bayésienne naïve

# 1.2 L'estimation

---

Quels modèles ? Le modèle peut avoir une interpretation probabiliste ou non.

-> si interpretation probabiliste

fonction objectif = **- maximum de vraisemblance**

Exemples: régression linéaire, régression logistique, classification bayésienne naïve

-> si pas d'interpretation probabiliste

fonction objectif = **a la main**

Exemples: machine à vecteur de support (SVM), forêt aléatoire, réseaux de neurones

# 1.2 L'estimation

Quels modèles ? Le modèle peut avoir une interpretation probabiliste ou non.

-> si interpretation probabiliste (=modèle statistique)  
fonction objectif = **- maximum de vraisemblance**

Exemples: régression linéaire, régression logistique, classification bayésienne naïve

-> si pas d'interpretation probabiliste (=modèle statistique?)  
fonction objectif = **a la main**

Exemples: machine à vecteur de support (SVM), forêt aléatoire, réseaux de neurones

Les réseaux de neurones sont-ils des modèles statistiques ?

<https://ai.stackexchange.com/questions/10289/are-neural-networks-statistical-models/18580#18580>

Cf aussi sur la définition des modèles statistiques <https://www.stat.uchicago.edu/~pmcc/pubs/AOS023.pdf>

# 1.2 L'estimation

---

-> si interpretation probabiliste (=modèle statistique)

Modèle statistique = ensemble de lois (=mesures) de probabilité  $\mathbb{P}$  sur l'espace des observations  $\mathcal{X}$ . Si proba paramétriques, alors on parle de modèle paramétrique. Enfin si les proba ont une densité  $p$  alors le modèle s'écrit

$$\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\}$$

# 1.2 L'estimation

---

-> si interprétation probabiliste (=modèle statistique)

Modèle statistique = ensemble de lois (=mesures) de probabilité  $\mathbb{P}$  sur l'espace des observations  $\mathcal{X}$ . Si proba paramétriques, alors on parle de modèle paramétrique. Enfin si les proba ont une densité  $p$  alors le modèle s'écrit

$$\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\}$$

Quand on modélise, on fait l'hypothèse qu'il existe  $\theta^*$  tel que  $\mathbb{P}_X = \mathbb{P}_{\theta^*}$

Modéliser = calculer  $\hat{\theta}$  en espérant que  $\hat{\theta} \approx \theta^*$

# 1.2 L'estimation

Modèle statistique  $\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\}$  Il existe  $\theta^*$  tel que  $\mathbb{P}_X = \mathbb{P}_{\theta^*}$

Pour le modèle candidat  $\mathbb{P}_\theta$

$p_\theta(x_i)$  = vraisemblance échantillon i

$\prod_{i=1}^n p_\theta(x_i)$  = vraisemblance jeu de données

# 1.2 L'estimation

Modèle statistique  $\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\}$  Il existe  $\theta^*$  tel que  $\mathbb{P}_X = \mathbb{P}_{\theta^*}$

Pour le modèle candidat  $\mathbb{P}_\theta$

$p_\theta(x_i)$  = vraisemblance échantillon i

$\prod_{i=1}^n p_\theta(x_i)$  = vraisemblance jeu de données

$$L(\theta) = - \prod_{i=1}^n p_\theta(x_i)$$

Fonction de coût à minimiser



# 1.2 L'estimation

Modèle statistique  $\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\}$  Il existe  $\theta^*$  tel que  $\mathbb{P}_X = \mathbb{P}_{\theta^*}$

Pour le modèle candidat  $\mathbb{P}_\theta$

$p_\theta(x_i)$  = vraisemblance échantillon i

$\prod_{i=1}^n p_\theta(x_i)$  = vraisemblance jeu de données

$L(\theta) = - \prod_{i=1}^n p_\theta(x_i)$  équivalent à

$$\ell(\theta) = - \sum_{i=1}^n \log p_\theta(x_i)$$

Fonction de coût à minimiser = - log de la vraisemblance

# Sommaire

---

- 1. Données et modélisation
  - 1. Le jeu de données
  - 2. L'estimation (model fitting)
- 2. Modèles de régression
  - 1. Régression logistique binaire
  - 2. Régression logistique multinomiale
- 3. Réseaux de neurones
  - 1. Réseau monocouche
  - 2. Réseau multicouche
  - 3. Réseau convolutionnel

# 2.1 Régression logistique binaire

Modèle statistique de regression Prédire  $Y$  à partir de  $X$

Jeu de données  $x_{1:n}, y_{1:n}$

Un modèle de regression est une famille de lois de probabilités *conditionnelles*

$$\mathcal{M}_\Theta = \{\mathbb{P}_{Y|X=x}^\theta | \theta \in \Theta, x \in \mathcal{X}\}$$

# 2.1 Régression logistique binaire

Modèle statistique de regression Prédire  $Y$  à partir de  $X$

Jeu de données  $x_{1:n}, y_{1:n}$

Un modèle de regression est une famille de lois de probabilités *conditionnelles*

$$\mathcal{M}_\Theta = \{\mathbb{P}_{Y|X=x}^\theta | \theta \in \Theta, x \in \mathcal{X}\}$$

Regression logistique  $Y \in \{0, 1\}, \quad X \in \mathbb{R}^p$

$$\mathbb{P}_{Y|X=x}^\theta = \text{Binomial}(\sigma(x^\top \theta))$$

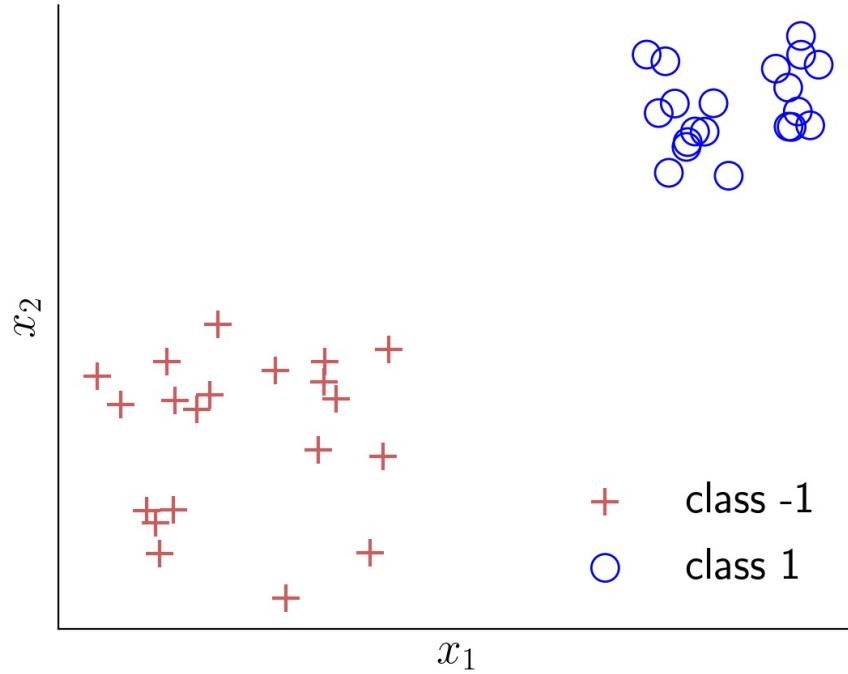
$$\sigma(u) = \frac{1}{1 + e^{-u}}$$

# 2.1 Régression logistique binaire

Regression logistique  $Y \in \{0, 1\}$ ,

Exemple:  $p=2$ ,  $X = (X^1, X^2)$

$$X \in \mathbb{R}^p \quad \mathbb{P}_{Y|X=x}^\theta = \text{Binomial}(\sigma(x^\top \theta))$$



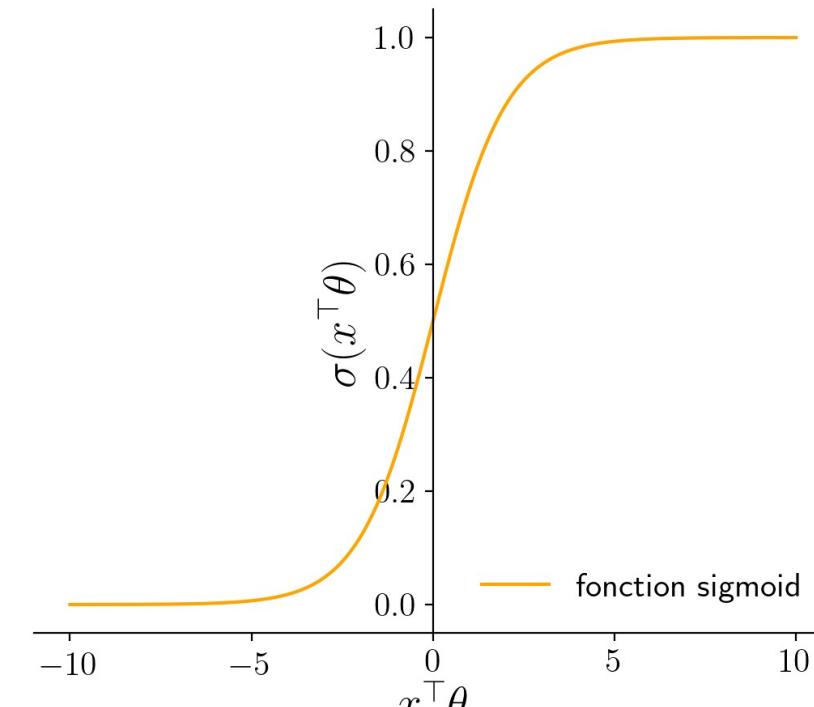
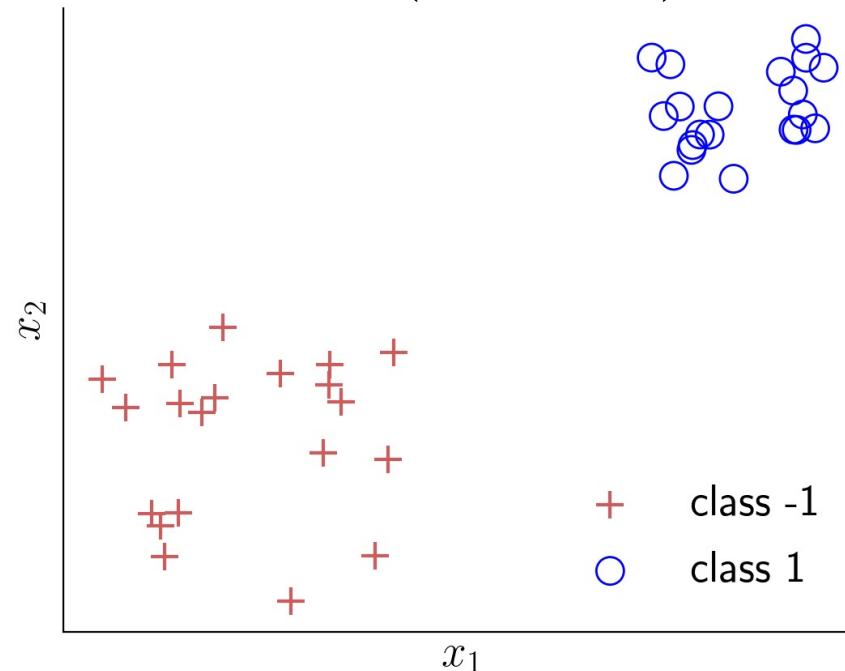
# 2.1 Régression logistique binaire

Regression logistique  $Y \in \{0, 1\}$ ,

$X \in \mathbb{R}^p$

$\mathbb{P}_Y^{\theta}|_{X=x} = \text{Binomial}(\sigma(x^\top \theta))$

Exemple:  $p=2$ ,  $X = (X^1, X^2)$



Pour  $X = x$ , prédition =  $\begin{cases} 1 & \text{si } \sigma(x^\top \theta) > 0.5, \text{i.e } x^\top \theta > 0 \\ 0 & \text{si } x^\top \theta < 0 \end{cases}$

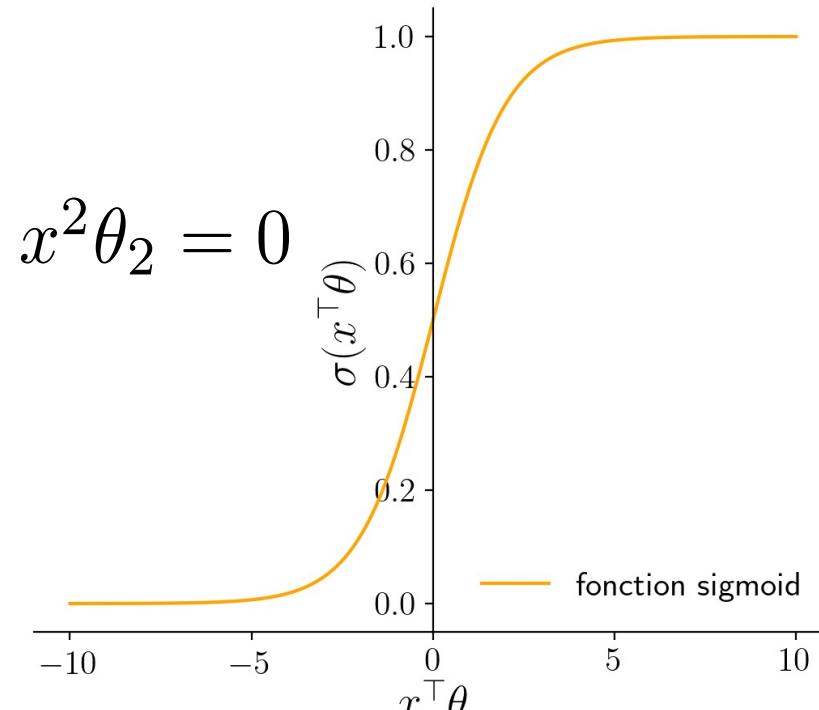
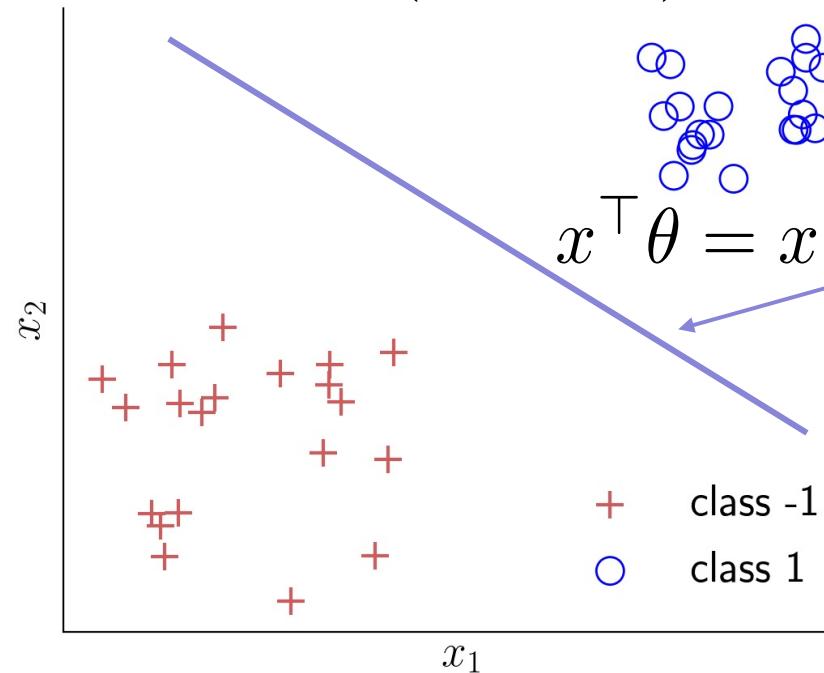
# 2.1 Régression logistique binaire

Regression logistique  $Y \in \{0, 1\}$ ,

$X \in \mathbb{R}^p$

$\mathbb{P}_Y^{\theta}|_{X=x} = \text{Binomial}(\sigma(x^{\top} \theta))$

Exemple:  $p=2$ ,  $X = (X^1, X^2)$



Pour  $X = x$ , prédition =  $\begin{cases} 1 & \text{si } \sigma(x^{\top} \theta) > 0.5, \text{i.e } x^{\top} \theta > 0 \\ 0 & \text{si } x^{\top} \theta < 0 \end{cases}$

# 2.1 Régression logistique binaire

Regression logistique  $Y \in \{0, 1\}, X \in \mathbb{R}^p$

$$\mathbb{P}_{Y|X=x}^\theta = \text{Binomial}(\sigma(x^\top \theta))$$

$$\sigma(u) = \frac{1}{1 + e^{-u}}$$

Comment s'écrit la vraisemblance de  $y_i|x_i$  ?

$$p_\theta(\cdot|x_i) = \begin{cases} 1 & \text{avec probabilité } \sigma(x_i^\top \theta) \\ 0 & \text{avec probabilité } 1 - \sigma(x_i^\top \theta) \end{cases}$$

# 2.1 Régression logistique binaire

Regression logistique  $Y \in \{0, 1\}, X \in \mathbb{R}^p$

$$\mathbb{P}_{Y|X=x}^\theta = \text{Binomial}(\sigma(x^\top \theta))$$

$$\sigma(u) = \frac{1}{1 + e^{-u}}$$

Comment s'écrit la vraisemblance de  $y_i|x_i$  ?

$$p_\theta(\cdot|x_i) = \begin{cases} 1 & \text{avec probabilité } \sigma(x_i^\top \theta) \\ 0 & \text{avec probabilité } 1 - \sigma(x_i^\top \theta) \end{cases}$$

$$p_\theta(y_i|x_i) = \sigma(\theta^\top x_i)^{y_i} (1 - \sigma(\theta^\top x_i))^{1-y_i}$$

# 2.1 Régression logistique binaire

Regression logistique  $Y \in \{0, 1\}$ ,  $X \in \mathbb{R}^p$

$$\mathbb{P}_{Y|X=x}^\theta = \text{Binomial}(\sigma(x^\top \theta))$$

$$\sigma(u) = \frac{1}{1 + e^{-u}}$$

Comment s'écrit la vraisemblance de  $y_i|x_i$  ?

$$p_\theta(\cdot|x_i) = \begin{cases} 1 & \text{avec probabilité } \sigma(x_i^\top \theta) \\ 0 & \text{avec probabilité } 1 - \sigma(x_i^\top \theta) \end{cases}$$

$$p_\theta(y_i|x_i) = \sigma(\theta^\top x_i)^{y_i} (1 - \sigma(\theta^\top x_i))^{1-y_i}$$

**Fonction de coût**

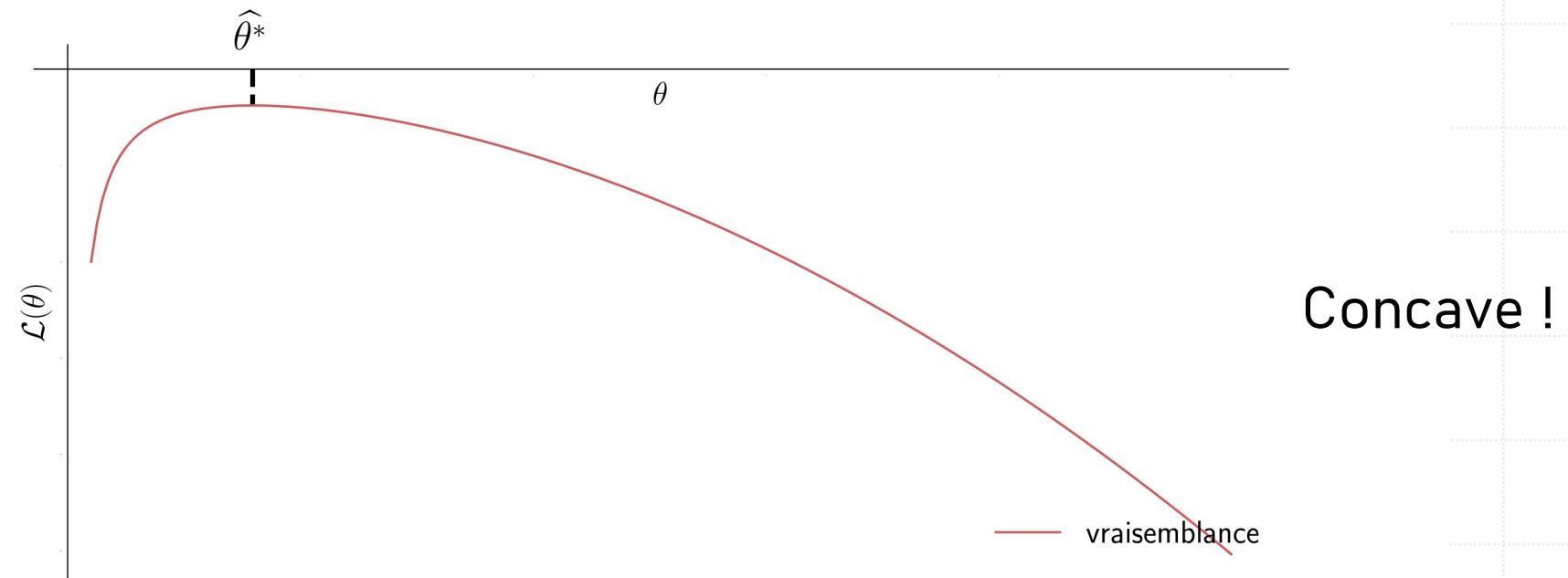
$$\ell(\theta; x_{1:n}, y_{1:n}) = - \sum_{i=1}^n y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i))$$

# 2.1 Régression logistique binaire

Regression logistique

**Fonction de coût**

$$\ell(\theta; x_{1:n}, y_{1:n}) = - \sum_{i=1}^n y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i))$$



# 2.1 Régression logistique binaire

Regression logistique

**Fonction de coût**

$$\ell(\theta; x_{1:n}, y_{1:n}) = - \sum_{i=1}^n y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i))$$

Minimisation par descente de gradient

$$\ell(\theta^{t+1}) = \ell(\theta^t) + (\theta^{t+1} - \theta^t) \nabla \ell(\theta^t) + o(\theta^{t+1} - \theta^t)$$

(dvp de Taylor ordre 1)

# 2.1 Régression logistique binaire

Regression logistique

**Fonction de coût**

$$\ell(\theta; x_{1:n}, y_{1:n}) = - \sum_{i=1}^n y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i))$$

Minimisation par descente de gradient

$$\ell(\theta^{t+1}) = \ell(\theta^t) + (\theta^{t+1} - \theta^t) \nabla \ell(\theta^t) + o(\theta^{t+1} - \theta^t)$$

(dvp de Taylor ordre 1)

Idée. Choisir  $\theta^{t+1}$  tel que  $(\theta^{t+1} - \theta^t) = -\nabla \ell(\theta^t)$

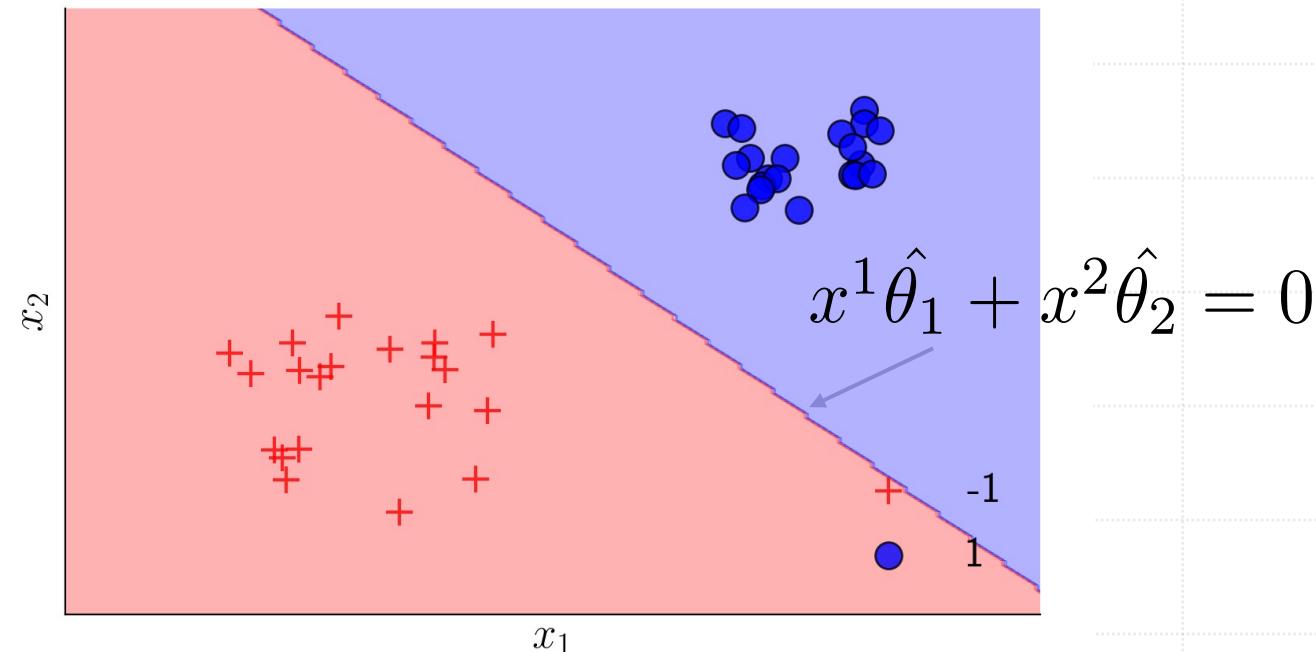
# 2.1 Régression logistique binaire

## Regression logistique

## Fonction de coût

$$\ell(\theta; x_{1:n}, y_{1:n}) = - \sum_{i=1}^n y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i))$$

## Minimisation par descente de gradient



## 2.2 Régression logistique multinomiale

Regression logistique multinomiale  $Y \in \{1, 2, \dots, K\}$ ,  $X \in \mathbb{R}^p$

$$\mathbb{P}_{Y|X=x}^\Theta = \text{Multinomial}(\sigma(\Theta^\top x))$$

$$\sigma(z) = \frac{1}{\sum_{k=1}^K e^{z_k}} \begin{bmatrix} e^{z_1} \\ \vdots \\ e^{z_K} \end{bmatrix}$$

$$\Theta = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,p} \\ \vdots & \ddots & \vdots \\ \theta_{K,1} & \cdots & \theta_{K,p} \end{bmatrix}$$

## 2.2 Régression logistique multinomiale

Regression logistique multinomiale  $Y \in \{1, 2, \dots, K\}$ ,  $X \in \mathbb{R}^p$

$$\mathbb{P}_{Y|X=x}^\Theta = \text{Multinomial}(\sigma(\Theta^\top x))$$

$$\sigma(z) = \frac{1}{\sum_{k=1}^K e^{z_k}} \begin{bmatrix} e^{z_1} \\ \vdots \\ e^{z_K} \end{bmatrix}$$

$$\Theta = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,p} \\ \vdots & \ddots & \vdots \\ \theta_{K,1} & \cdots & \theta_{K,p} \end{bmatrix}$$

$$\begin{bmatrix} \mathbb{P}_{Y|X=x}^\Theta(Y=1) \\ \vdots \\ \mathbb{P}_{Y|X=x}^\Theta(Y=K) \end{bmatrix} = \frac{1}{\sum_{k=1}^K e^{(\Theta^\top x)_k}} \begin{bmatrix} e^{(\Theta^\top x)_1} \\ \vdots \\ e^{(\Theta^\top x)_K} \end{bmatrix}$$

## 2.2 Régression logistique multinomiale

Regression logistique multinomiale  $Y \in \{1, 2, \dots, K\}$ ,  $X \in \mathbb{R}^p$

Comment s'écrit la vraisemblance de  $y_i | x_i$  ?

$$p_{\Theta}(\cdot | x_i) = k \text{ avec probabilité } \sigma(\Theta^\top x)_k$$

$$p_{\Theta}(y_i | x_i) = \prod_{k=1}^K \sigma(\Theta^\top x_i)_k^{1_{y_i=k}}$$

## 2.2 Régression logistique multinomiale

Régression logistique multinomiale  $Y \in \{1, 2, \dots, K\}$ ,  $X \in \mathbb{R}^p$

Comment s'écrit la vraisemblance de  $y_i | x_i$  ?

$$p_{\Theta}(\cdot | x_i) = k \text{ avec probabilité } \sigma(\Theta^\top x)_k$$

$$p_{\Theta}(y_i | x_i) = \prod_{k=1}^K \sigma(\Theta^\top x_i)_k^{1_{y_i=k}}$$

**Fonction de coût**

$$\ell(\Theta; x_{1:n}, y_{1:n}) = - \sum_{i=1}^n \sum_{k=1}^K \mathbf{1}_{y_i=k} \log \sigma(\Theta^\top x_i)_k$$

## 2.2 Régression logistique multinomiale

Estimation par descente de gradient  $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

## 2.2 Régression logistique multinomiale

Estimation par descente de gradient  $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

NOTE 1: modèle surparamétré, car

$$\text{si } \Theta \leftarrow \Theta - \psi = \begin{bmatrix} \Theta_{1,:} - \psi \\ \vdots \\ \Theta_{K,:} - \psi \end{bmatrix}$$

## 2.2 Régression logistique multinomiale

Estimation par descente de gradient  $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

NOTE 1: modèle surparamétré, car

si  $\Theta \leftarrow \Theta - \psi = \begin{bmatrix} \Theta_{1,:} - \psi \\ \vdots \\ \Theta_{K,:} - \psi \end{bmatrix}$  alors  $\begin{bmatrix} \mathbb{P}_{Y|X=x}^\Theta(Y=1) \\ \vdots \\ \mathbb{P}_{Y|X=x}^\Theta(Y=K) \end{bmatrix} = \begin{bmatrix} \mathbb{P}_{Y|X=x}^{\Theta-\psi}(Y=1) \\ \vdots \\ \mathbb{P}_{Y|X=x}^{\Theta-\psi}(Y=K) \end{bmatrix}$

## 2.2 Régression logistique multinomiale

Estimation par descente de gradient  $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

NOTE 1: modèle surparamétré, car

si  $\Theta \leftarrow \Theta - \psi = \begin{bmatrix} \Theta_{1,:} - \psi \\ \vdots \\ \Theta_{K,:} - \psi \end{bmatrix}$  alors  $\begin{bmatrix} \mathbb{P}_{Y|X=x}^\Theta(Y=1) \\ \vdots \\ \mathbb{P}_{Y|X=x}^\Theta(Y=K) \end{bmatrix} = \begin{bmatrix} \mathbb{P}_{Y|X=x}^{\Theta-\psi}(Y=1) \\ \vdots \\ \mathbb{P}_{Y|X=x}^{\Theta-\psi}(Y=K) \end{bmatrix}$

On fixe donc  $\Theta_{K,:} = 1$

## 2.2 Régression logistique multinomiale

Estimation par descente de gradient  $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

NOTE 1: modèle surparamétré, car

NOTE 2: généralisation de la régression logistique binaire

## 2.2 Régression logistique multinomiale

Estimation par descente de gradient  $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

NOTE 1: modèle surparamétré, car

NOTE 2: généralisation de la régression logistique binaire

NOTE 3: donne le même modèle de classification que K modèles de regression logistique binaire combinés en stratégie multiclasse one-vs-rest

$f_{\theta^1}^1 = \text{Reg. log. binaire } Y = 1 \text{ vs } Y \neq 1$

...

$f_{\theta^K}^K = \text{Reg. log. binaire } Y = K \text{ vs } Y \neq K$

# Sommaire

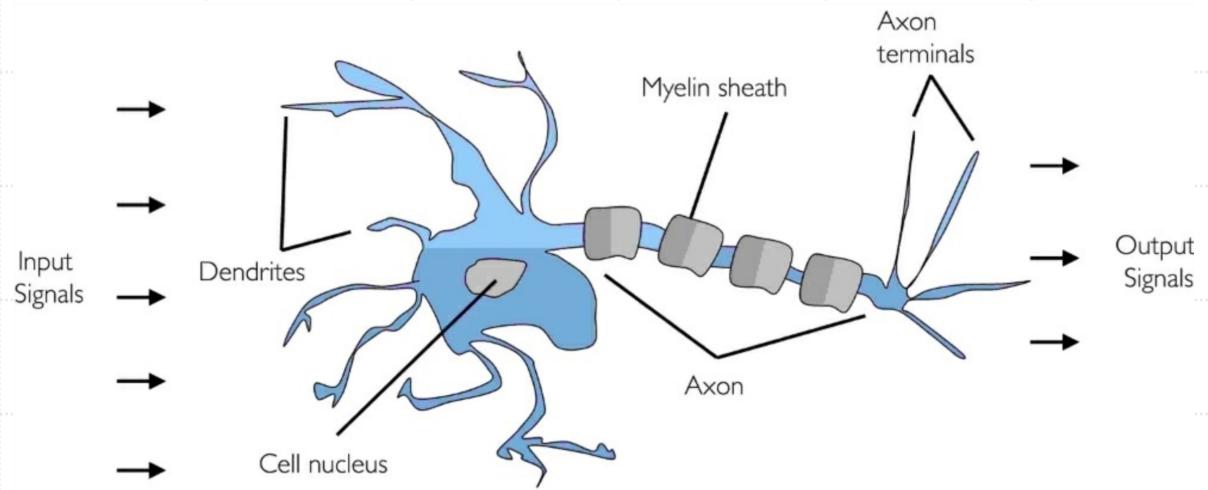
---

1. Données et modélisation
  1. Le jeu de données
  2. L'estimation (model fitting)
2. Modèles de régression
  1. Régression logistique binaire
  2. Régression logistique multinomiale
3. Réseaux de neurones
  1. Réseau monocouche
  2. Réseau multicouche
  3. Réseau convolutionnel

# 3.1 Réseau monocouche

Un neurone biologique:

si somme des signaux entrées > seuil  
→ potential d'action généré  
sinon inactif.



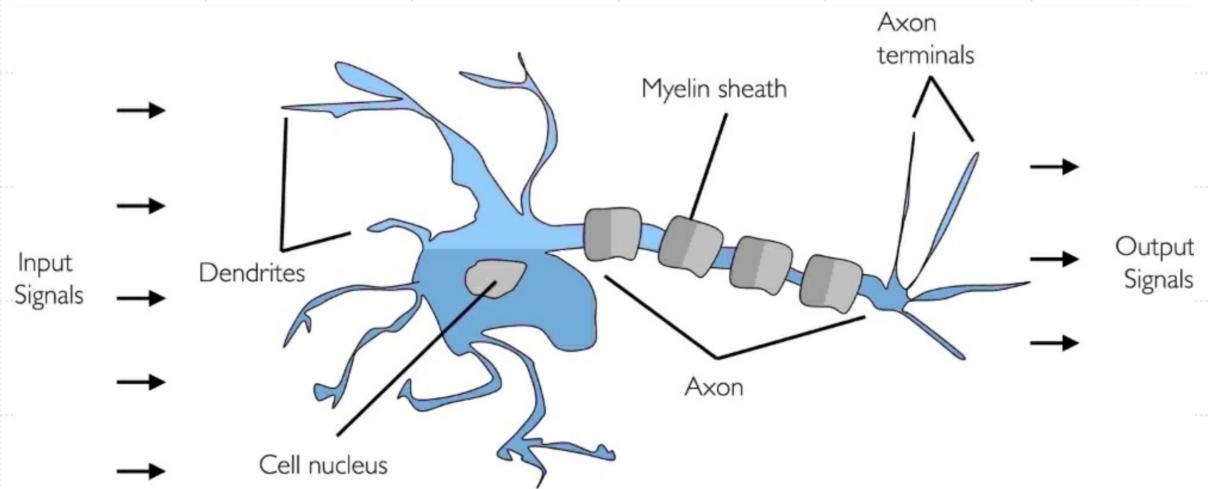
# 3.1 Réseau monocouche

## Un neurone biologique:

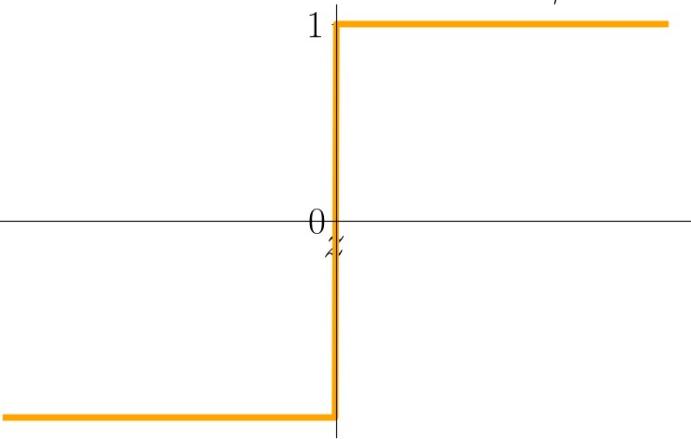
si somme des signaux entrées > seuil  
→ potential d'action généré  
sinon inactif.

## Un neurone artificiel Rosenblatt 1957

- Entrées  $x = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$ , poids  $w = \begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix}$
- Entrées agrégées  $z = x^\top w$
- Activation  $\phi(z)$ .



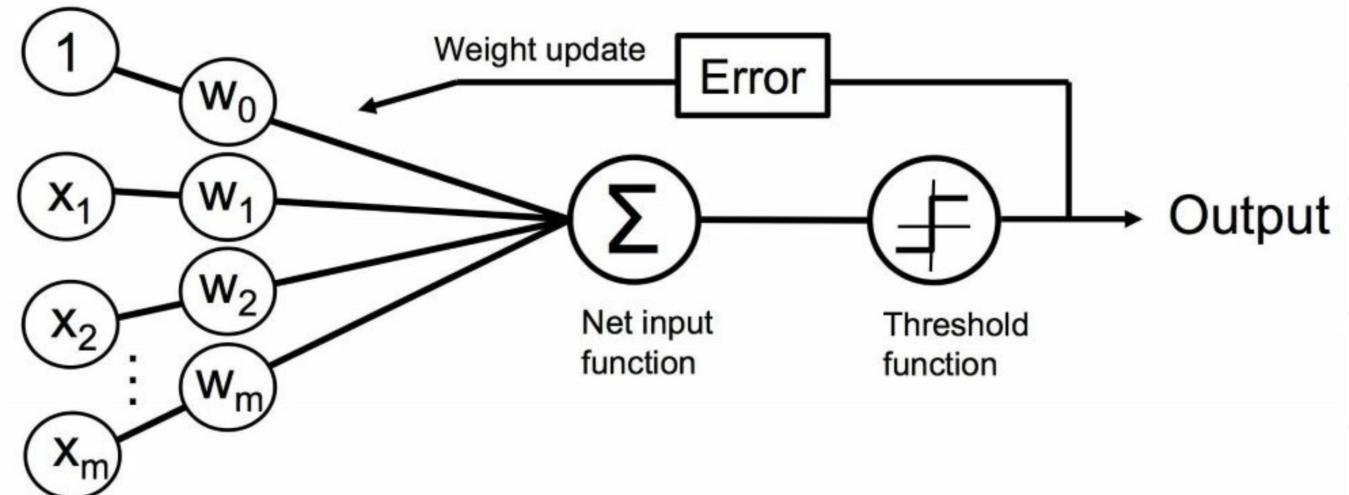
Fonction activation  $\phi$



# 3.1 Réseau monocouche

Règle du Perceptron: jeu de données  $x_{1:n}, y_{1:n}$

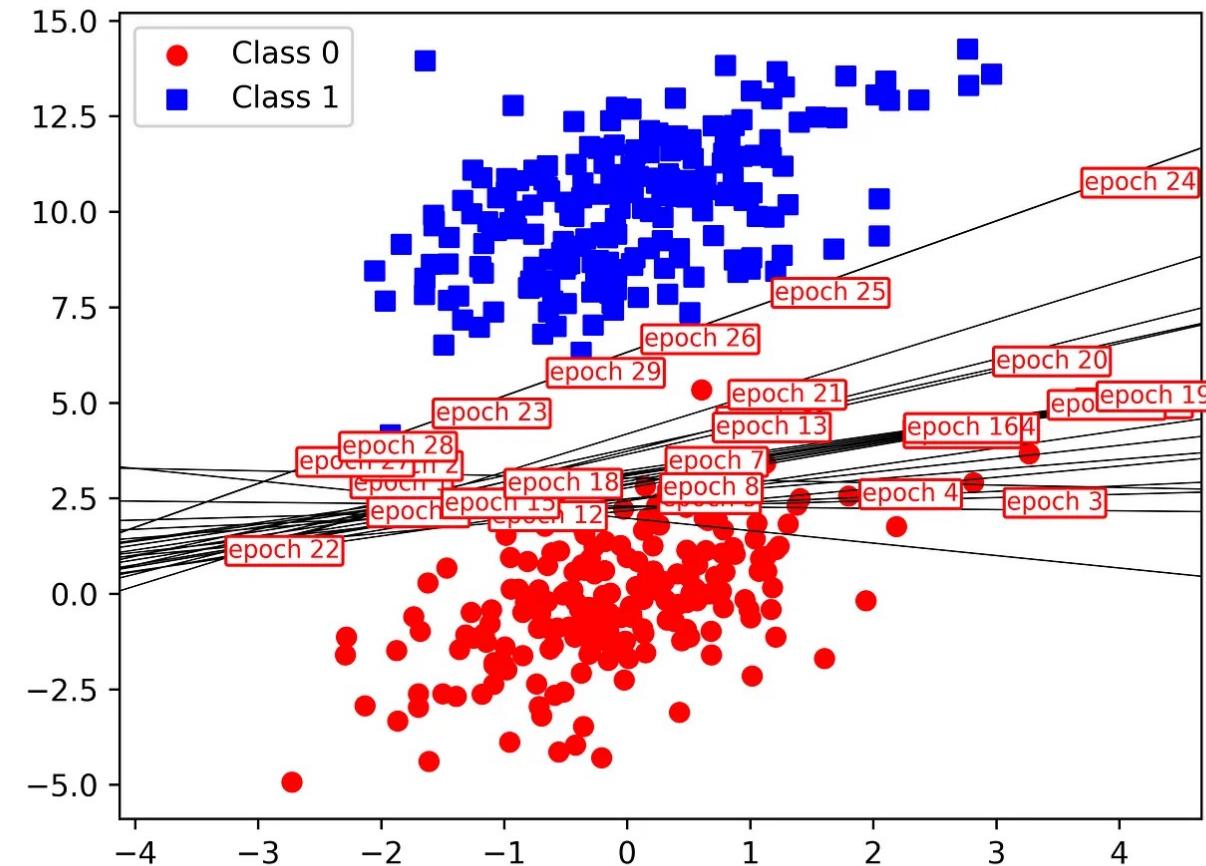
1. initialisation  $w$  aléatoire;
2. pour chaque  $i = 1, \dots, n$ ,
  - (a)  $\hat{y}_i = \phi(x_i^\top w)$
  - (b)  $\delta = \eta(y_i - \hat{y}_i)$ .
  - (c)  $w = w + \delta x_i$



# 3.1 Réseau monocouche

Règle du Perceptron: jeu de données  $x_{1:n}, y_{1:n}$

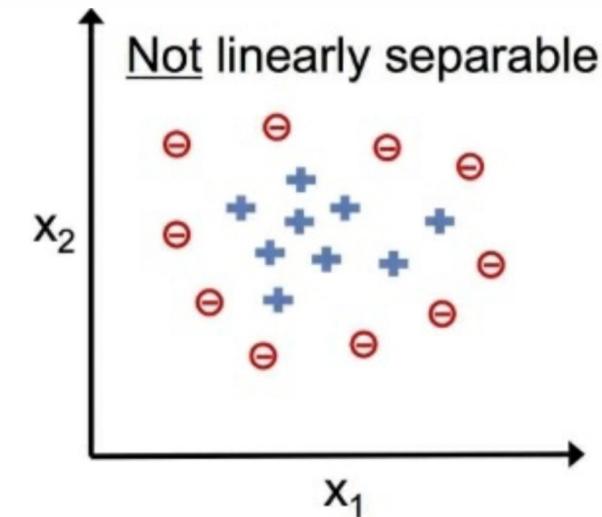
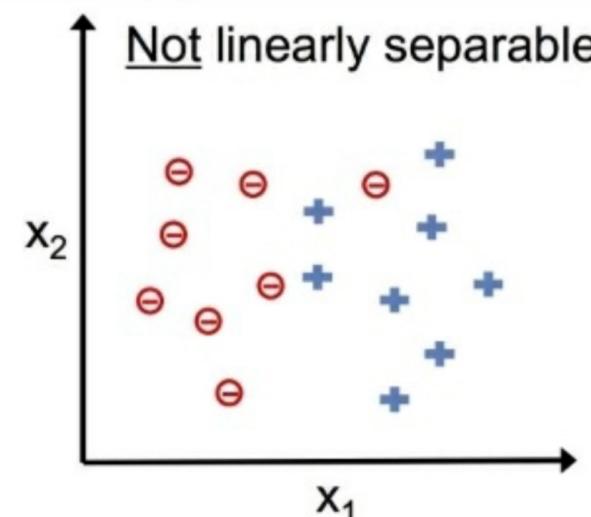
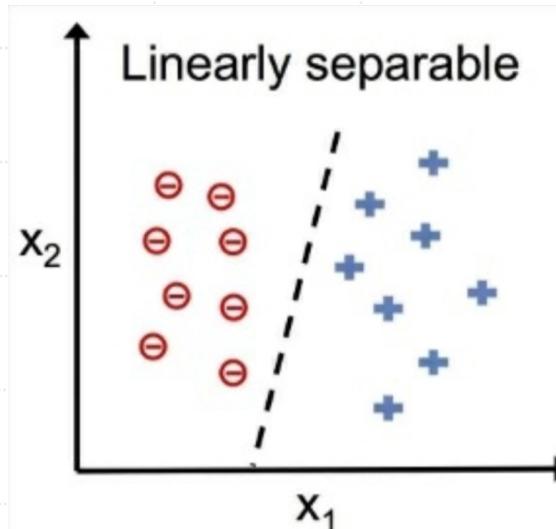
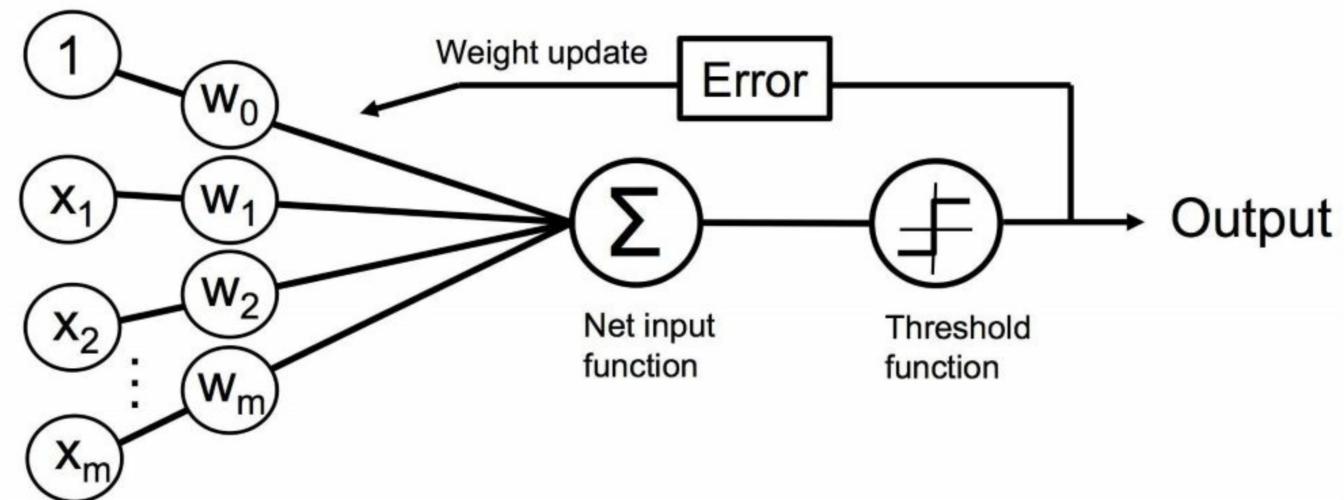
1. initialisation  $w$  aléatoire;
2. pour chaque  $i = 1, \dots, n$ ,
  - (a)  $\hat{y}_i = \phi(x_i^\top w)$
  - (b)  $\delta = \eta(y_i - \hat{y}_i)$ .
  - (c)  $w = w + \delta x_i$



# 3.1 Réseau monocouche

Règle du Perceptron: jeu de données  $x_{1:n}, y_{1:n}$

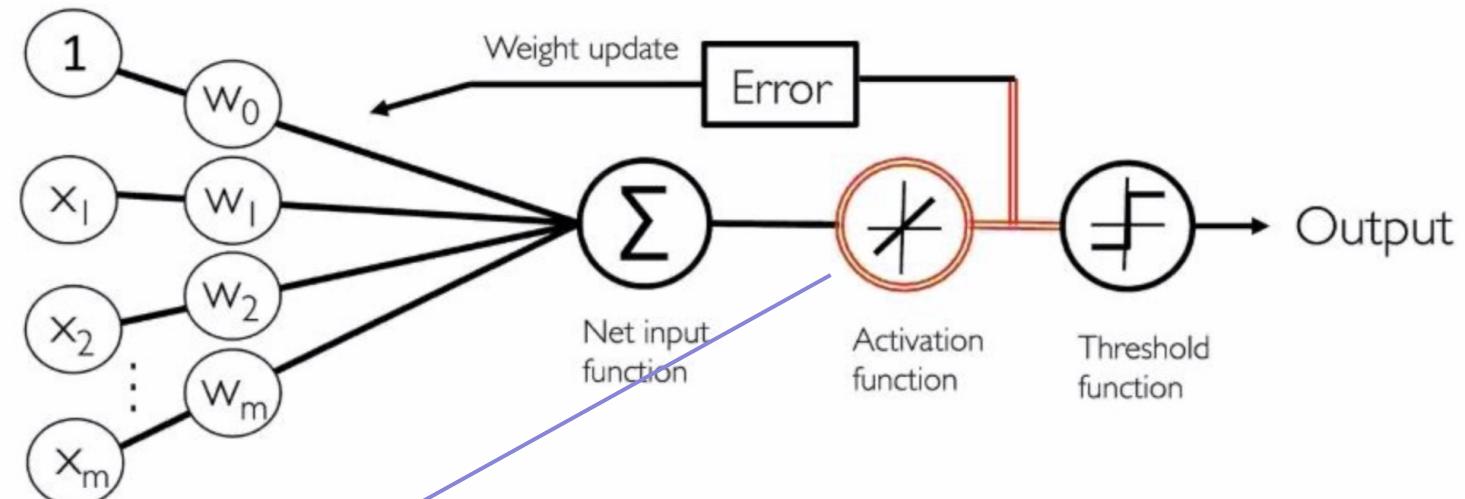
1. initialisation  $w$  aléatoire;
2. pour chaque  $i = 1, \dots, n$ ,
  - (a)  $\hat{y}_i = \phi(x_i^\top w)$
  - (b)  $\delta = \eta(y_i - \hat{y}_i)$ .
  - (c)  $w = w + \delta x_i$



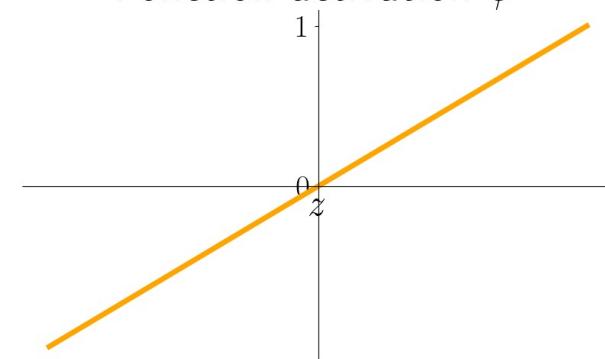
# 3.1 Réseau monocouche

Réseau Adaline: jeu de données  $x_{1:n}, y_{1:n}$

1. initialisation  $w$  aléatoire;
2. pour chaque  $i = 1, \dots, n$ ,
  - (a)  $\hat{y}_i = \phi(x_i^\top w)$
  - (b)  $\delta = \eta(y_i - \hat{y}_i)$ .
  - (c)  $w = w + \delta x_i$



Fonction activation  $\phi$

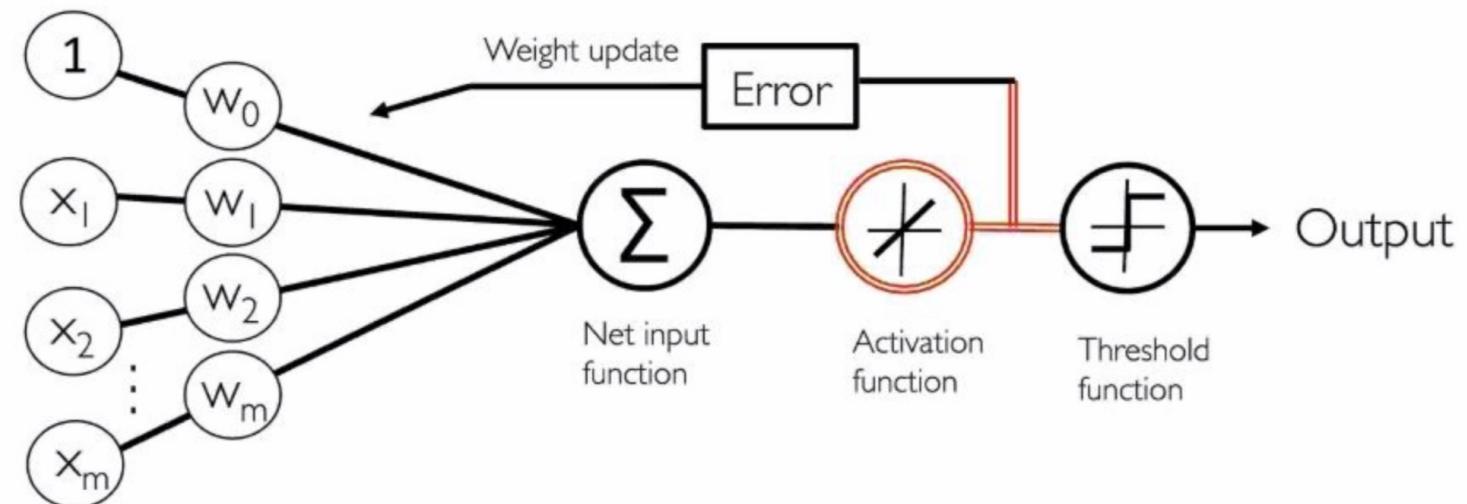


Adaptive Linear Neuron (Adaline)

# 3.1 Réseau monocouche

Réseau Adaline: jeu de données  $x_{1:n}, y_{1:n}$

1. initialisation  $w$  aléatoire;
2. pour chaque  $i = 1, \dots, n$ ,
  - (a)  $\hat{y}_i = \phi(x_i^\top w)$
  - (b)  $\delta = \eta(y_i - \hat{y}_i)$ .
  - (c)  $w = w + \delta x_i$



Pourquoi:  $y_i - x_i^\top w$

$$\hat{w}(x_{1:n}, y_{1:n}) \in \underset{w \in \mathbb{R}^p}{\operatorname{argmin}} J(w) = \frac{1}{2} \sum_{i=1}^n (y_i - x_i^\top w)^2$$

Adaptive Linear Neuron (Adaline)

descente de gradient sur un échantillon    pas =  $-\nabla J_i(w) = x_i(y_i - x_i^\top w)$

# 3.1 Réseau monocouche

Réseau Adaline: jeu de données  $x_{1:n}, y_{1:n}$

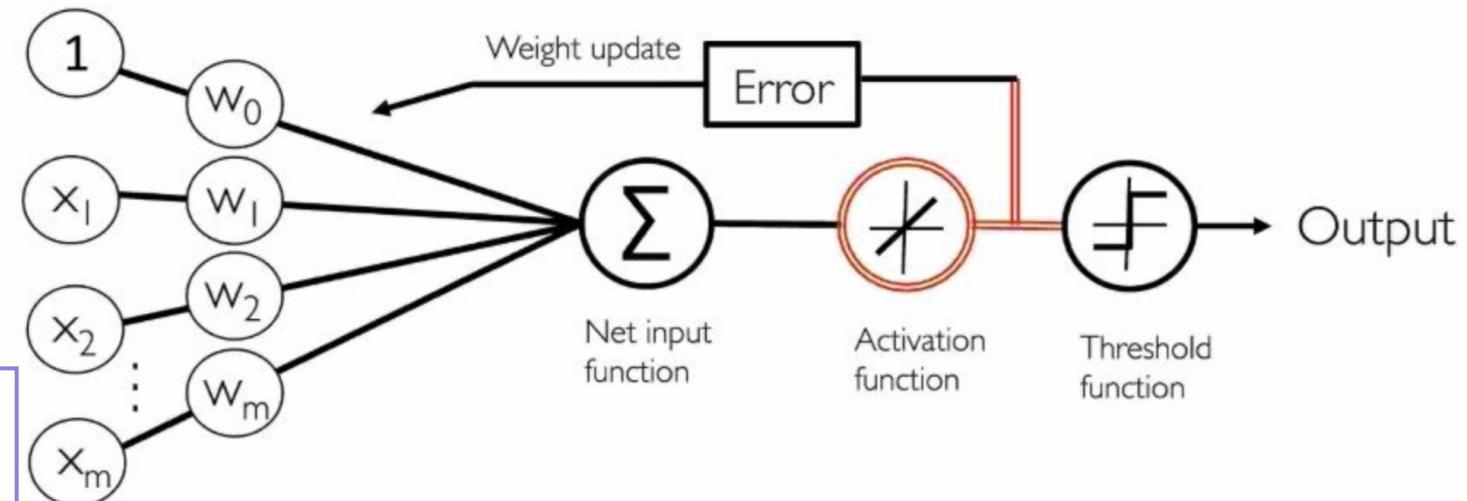
1. initialisation  $w$  aléatoire;

2. pour chaque  $i = 1, \dots, n$ ,

(a)  $\hat{y}_i = \phi(x_i^\top w)$

(b)  $\delta = \eta(y_i - x_i^\top w)$

(c)  $w = w + \delta x_i$



Pourquoi:  $y_i - x_i^\top w$

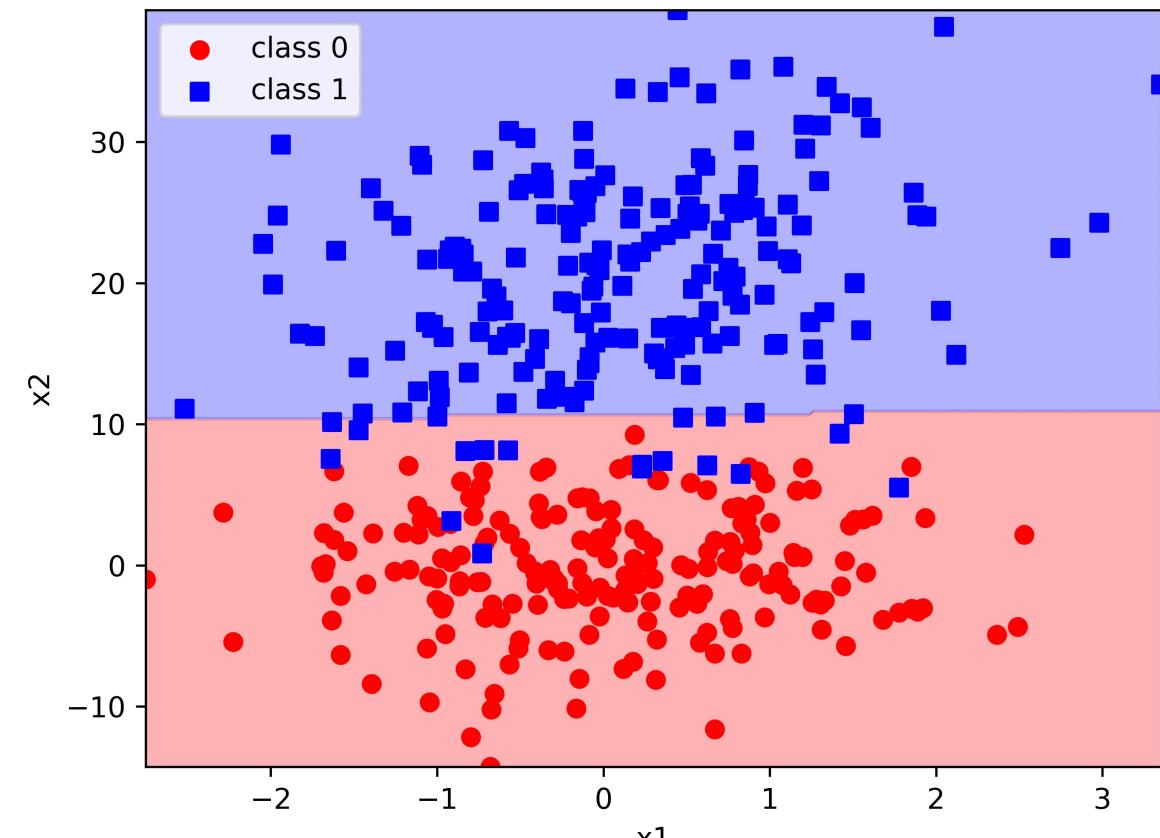
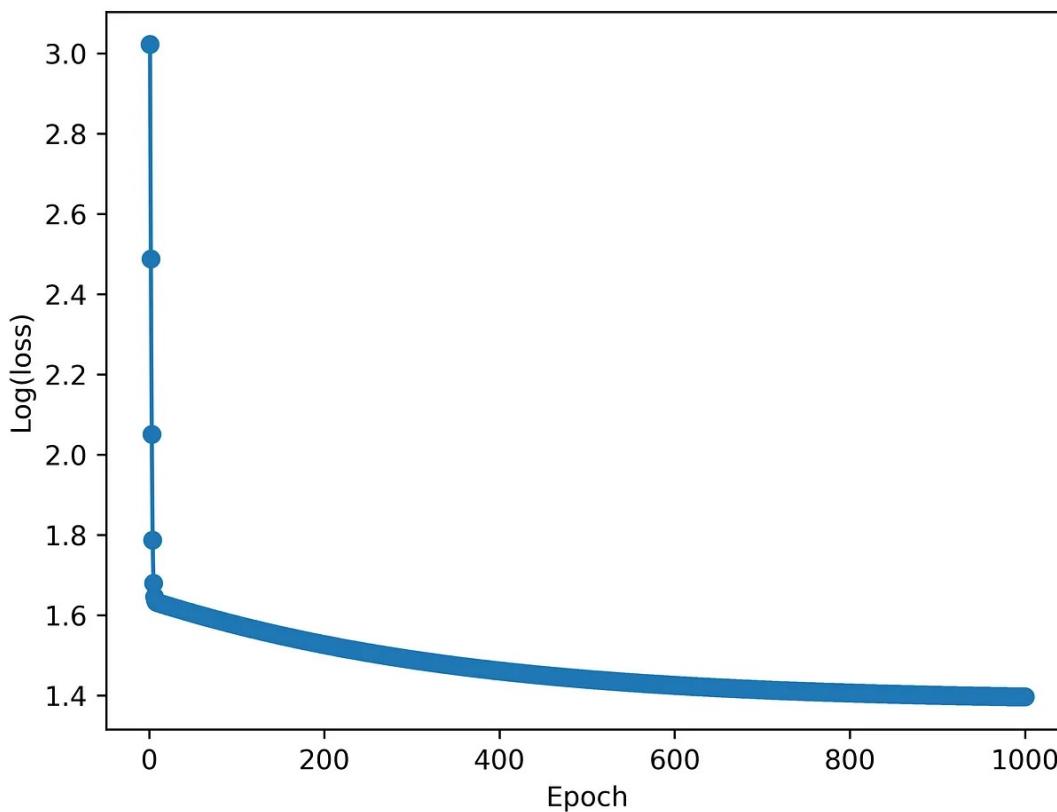
$$\hat{w}(x_{1:n}, y_{1:n}) \in \underset{w \in \mathbb{R}^p}{\operatorname{argmin}} J(w) = \frac{1}{2} \sum_{i=1}^n (y_i - x_i^\top w)^2$$

descente de gradient sur un échantillon

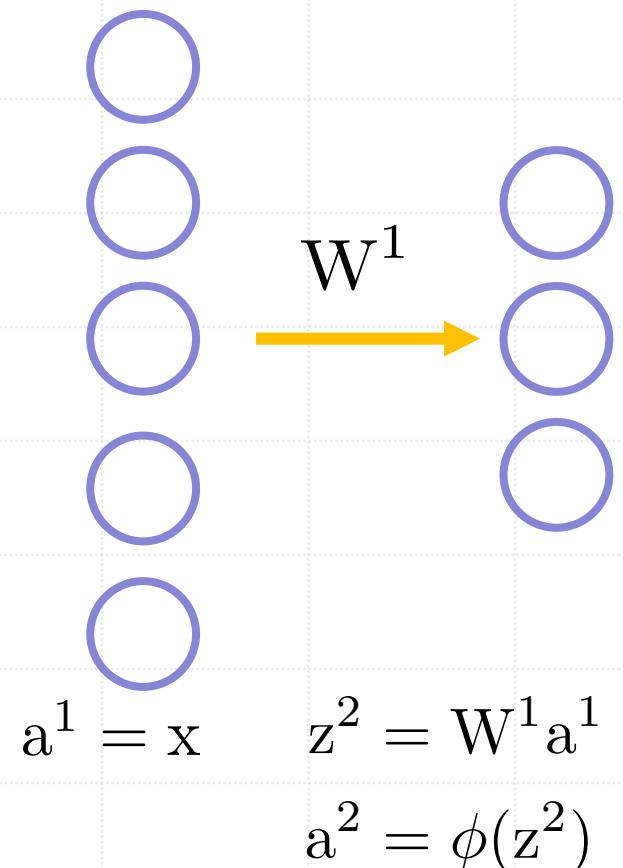
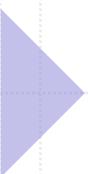
$$\text{pas} = -\nabla J_i(w) = x_i(y_i - x_i^\top w)$$

# 3.1 Réseau monocouche

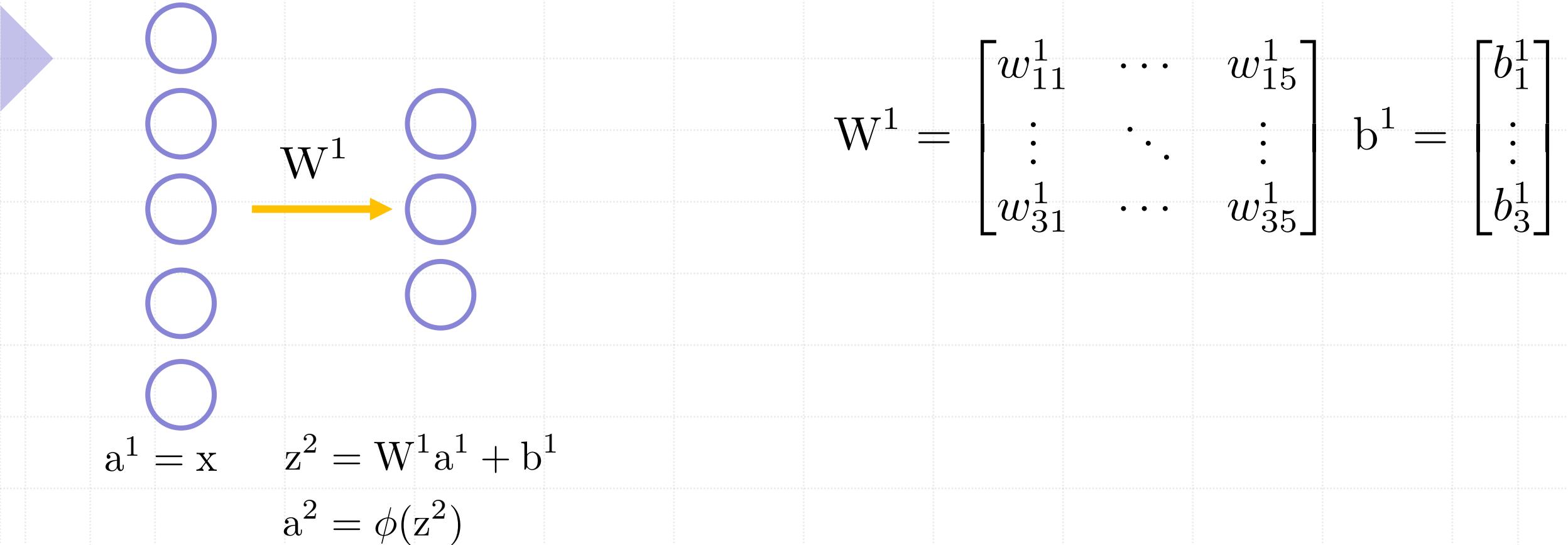
Réseau Adaline: jeu de données  $x_{1:n}, y_{1:n}$



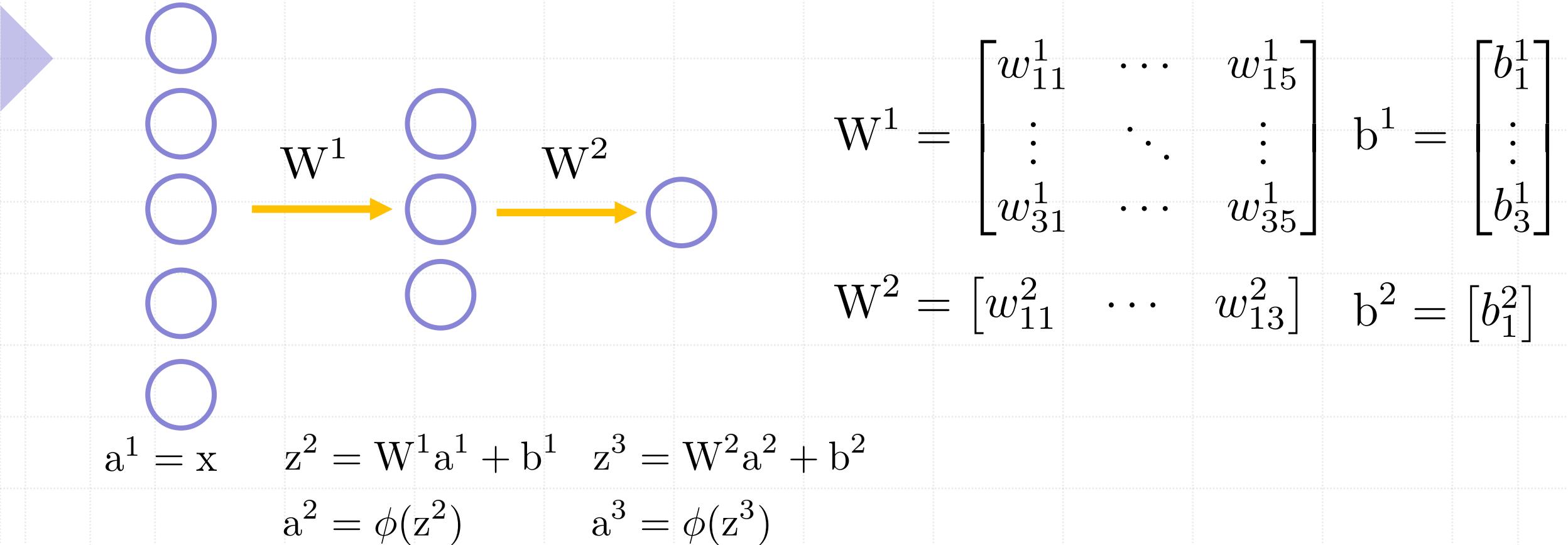
## 3.2 Réseau multicouches



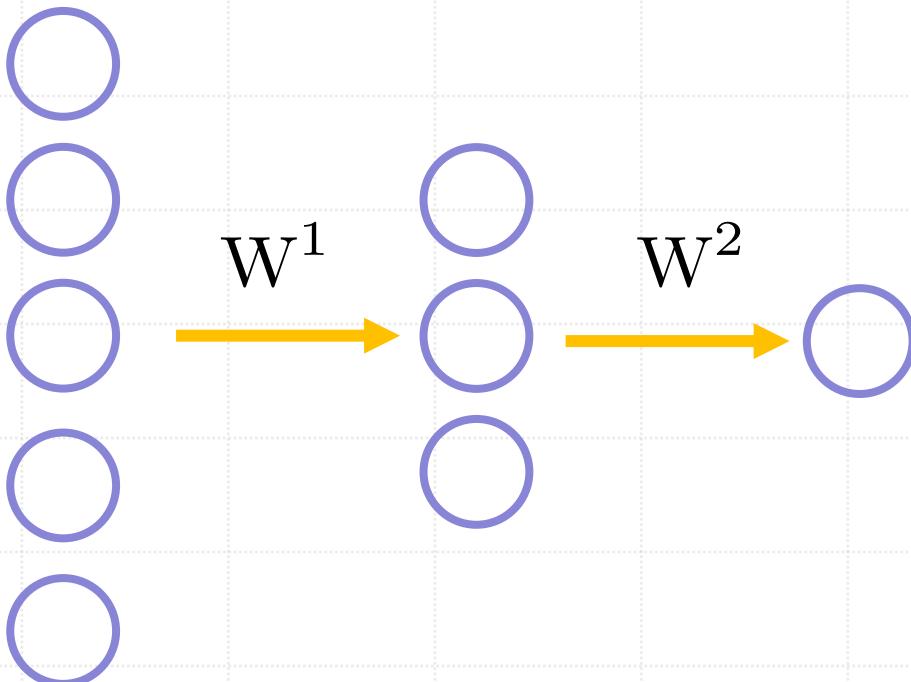
## 3.2 Réseau multicouches



## 3.2 Réseau multicouches



## 3.2 Réseau multicouches



$$a^1 = x \quad z^2 = W^1 a^1 + b^1 \quad z^3 = W^2 a^2 + b^2$$

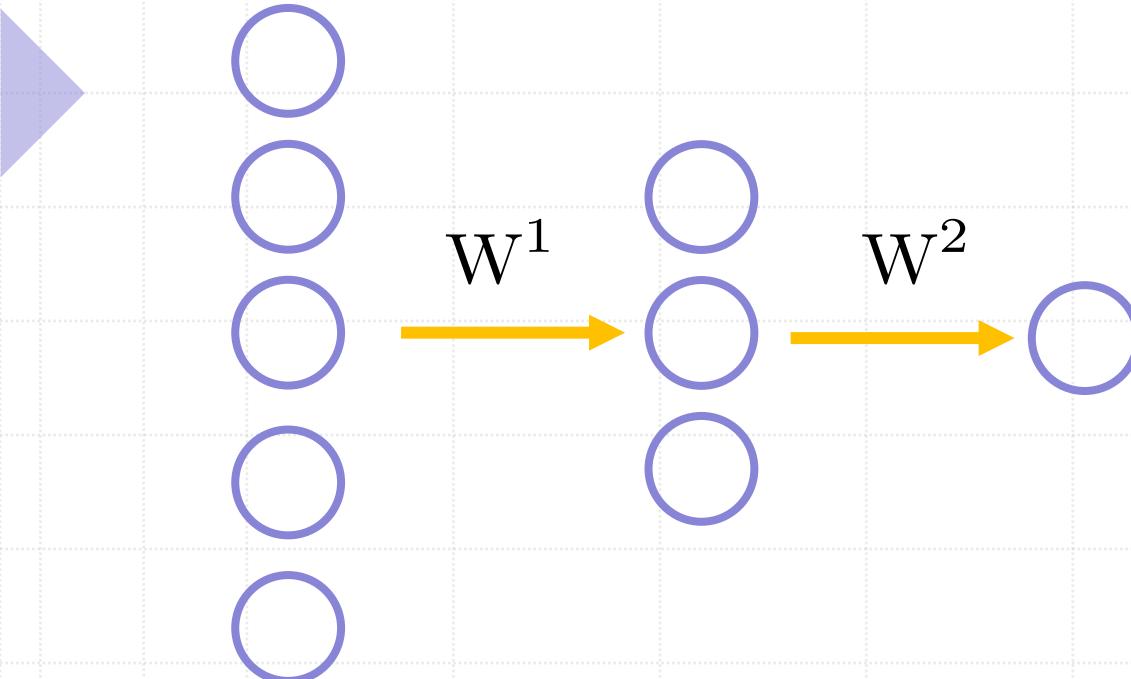
$$a^2 = \phi(z^2) \quad a^3 = \phi(z^3)$$

Q1: combien de paramètres ?

Q2: que se passe-t-il si  $\phi = \text{Id}$  ?

$$W^1 = \begin{bmatrix} w_{11}^1 & \cdots & w_{15}^1 \\ \vdots & \ddots & \vdots \\ w_{31}^1 & \cdots & w_{35}^1 \end{bmatrix} \quad b^1 = \begin{bmatrix} b_1^1 \\ \vdots \\ b_3^1 \end{bmatrix}$$
$$W^2 = [w_{11}^2 \quad \cdots \quad w_{13}^2] \quad b^2 = [b_1^2]$$

## 3.2 Réseau multicouches



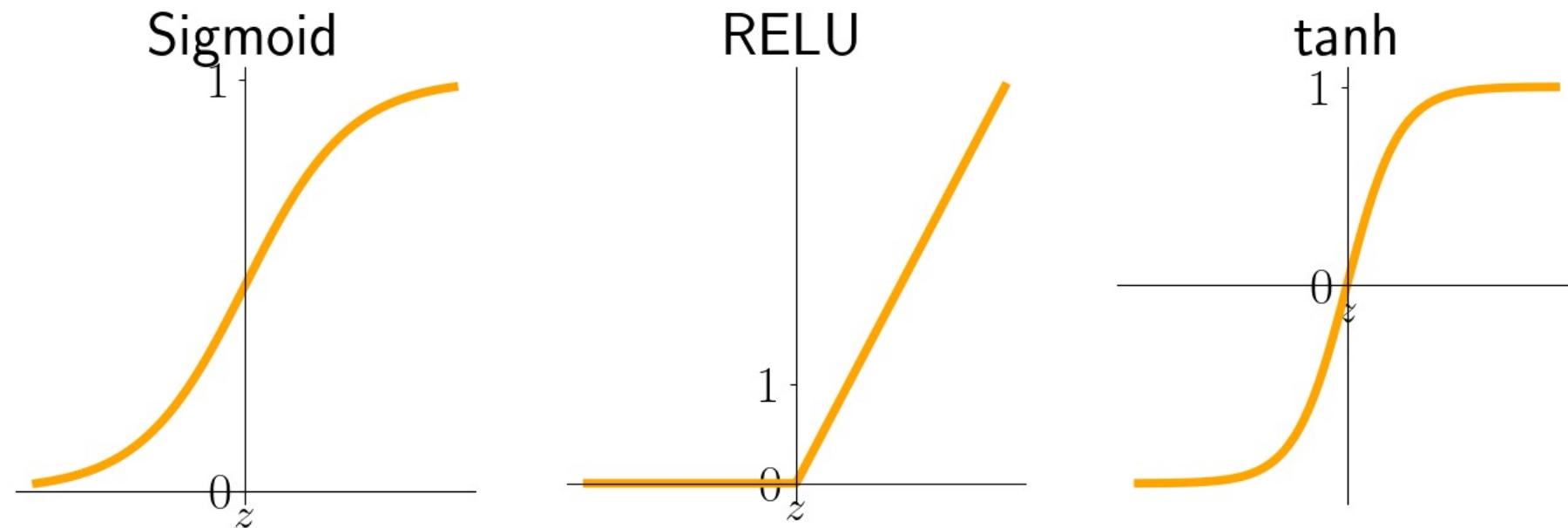
$$a^1 = x \quad z^2 = W^1 a^1 + b^1 \quad z^3 = W^2 a^2 + b^2$$

$$a^2 = \phi(z^2) \quad a^3 = \phi(z^3)$$

Q1: combien de paramètres ?  $3 \times 5 + 3 + 1 \times 3 + 1 = 22$

Q2: que se passe-t-il si  $\phi = \text{Id}$  ?  $a^3 = W^2 W^1 x + W^2 b^1 + b^2 = mx + p$

## 3.2 Réseau multicouches



Exemples de fonctions d'activation non linéaires

# 3.2 Réseau multicouches

Comment apprendre les paramètres  $\Theta = [W^1, b^1, W^2, b^2]$  ?

Jeu de données  $x_{1:n}, y_{1:n}$

Modèle  $f_\Theta(x) = a^3(x)$

# 3.2 Réseau multicouches

Comment apprendre les paramètres  $\Theta = [W^1, b^1, W^2, b^2]$  ?

Jeu de données  $x_{1:n}, y_{1:n}$

Modèle  $f_\Theta(x) = a^3(x)$

-> Choix d'une fonction de coût

Exemples:  $y_i \in \mathbb{R}$      $J(\Theta) = \frac{1}{2n} \sum_{i=1}^n \|y_i - f_\Theta(x)\|^2$     `tf.keras.losses.MeanSquaredError`

# 3.2 Réseau multicouches

Comment apprendre les paramètres  $\Theta = [W^1, b^1, W^2, b^2]$  ?

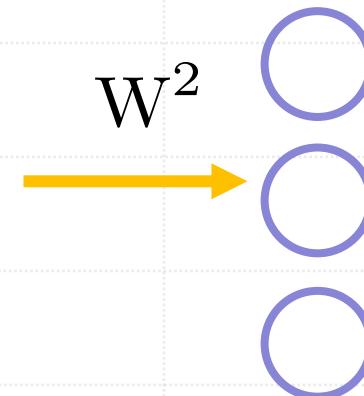
Jeu de données  $x_{1:n}, y_{1:n}$

Modèle  $f_\Theta(x) = a^3(x)$

-> Choix d'une fonction de coût

Exemples:  $y_i \in \mathbb{R}$   $J(\Theta) = \frac{1}{2n} \sum_{i=1}^n \|y_i - f_\Theta(x)\|^2$  `tf.keras.losses.MeanSquaredError`

$y_i \in \{1, \dots, K\}$   $J(\Theta) = \frac{-1}{n} \sum_{i=1}^n \sum_{k=1}^K \mathbb{1}_{y_i=k} \log ((f_\Theta(x_i))_k)$  `tf.keras.losses.CategoricalCrossEntropy`



$$\begin{aligned} z^3 &= W^2 a^2 + b^2 \\ a^3 &= \phi(z^3) \end{aligned}$$

$$\phi(z) = \begin{bmatrix} \frac{e^{z_1}}{\sum_{k=1}^K e^{z_k}} \\ \vdots \\ \frac{e^{z_K}}{\sum_{k=1}^K e^{z_k}} \end{bmatrix}$$

"probabilities" (!)

Softmax activation

# 3.2 Réseau multicouches

Comment apprendre les paramètres  $\Theta = [W^1, b^1, W^2, b^2]$  ?

Jeu de données  $x_{1:n}, y_{1:n}$

Modèle  $f_\Theta(x) = a^3(x)$

- > Choix d'une fonction de coût
- > Apprentissage par descente de gradient

$$\Theta^{t+1} = \Theta^t - \eta_t \nabla J(\Theta^t)$$

# 3.2 Réseau multicouches

Comment apprendre les paramètres  $\Theta = [W^1, b^1, W^2, b^2]$  ?

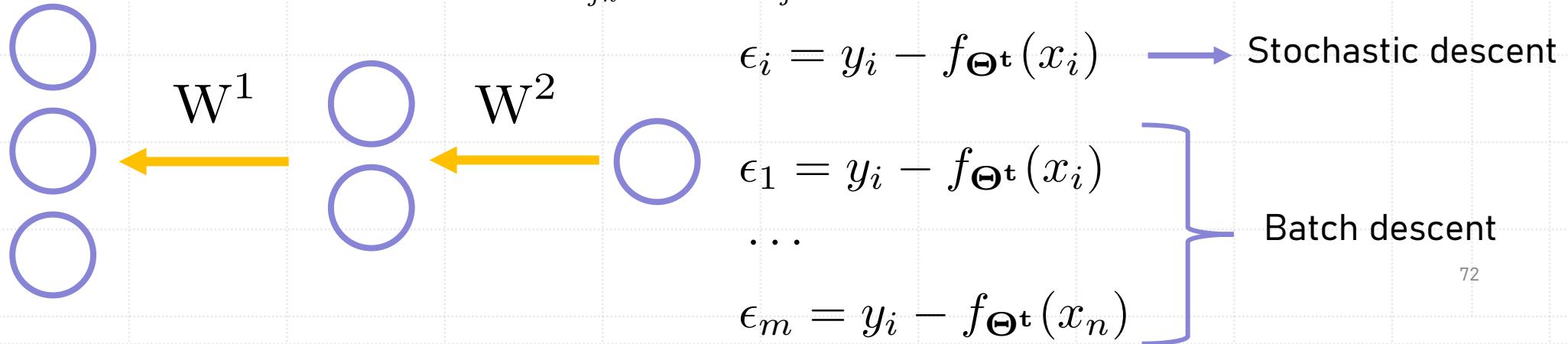
Jeu de données  $x_{1:n}, y_{1:n}$

Modèle  $f_\Theta(x) = a^3(x)$

- > Choix d'une fonction de coût
- > Apprentissage par descente de gradient

$$\Theta^{t+1} = \Theta^t - \eta_t \nabla J(\Theta^t)$$

Difficulté 1: comment calculer  $\frac{\partial J}{\partial w_{jk}^l}(\Theta^t), \frac{\partial J}{\partial b_j^l}(\Theta^t)$  pour tout  $j, k, l$  ?



## 3.2 Réseau multicouches

Comment apprendre les paramètres  $\Theta = [W^1, b^1, W^2, b^2]$  ?

Jeu de données  $x_{1:n}, y_{1:n}$

Modèle  $f_\Theta(x) = a^3(x)$

- > Choix d'une fonction de coût
- > Apprentissage par descente de gradient

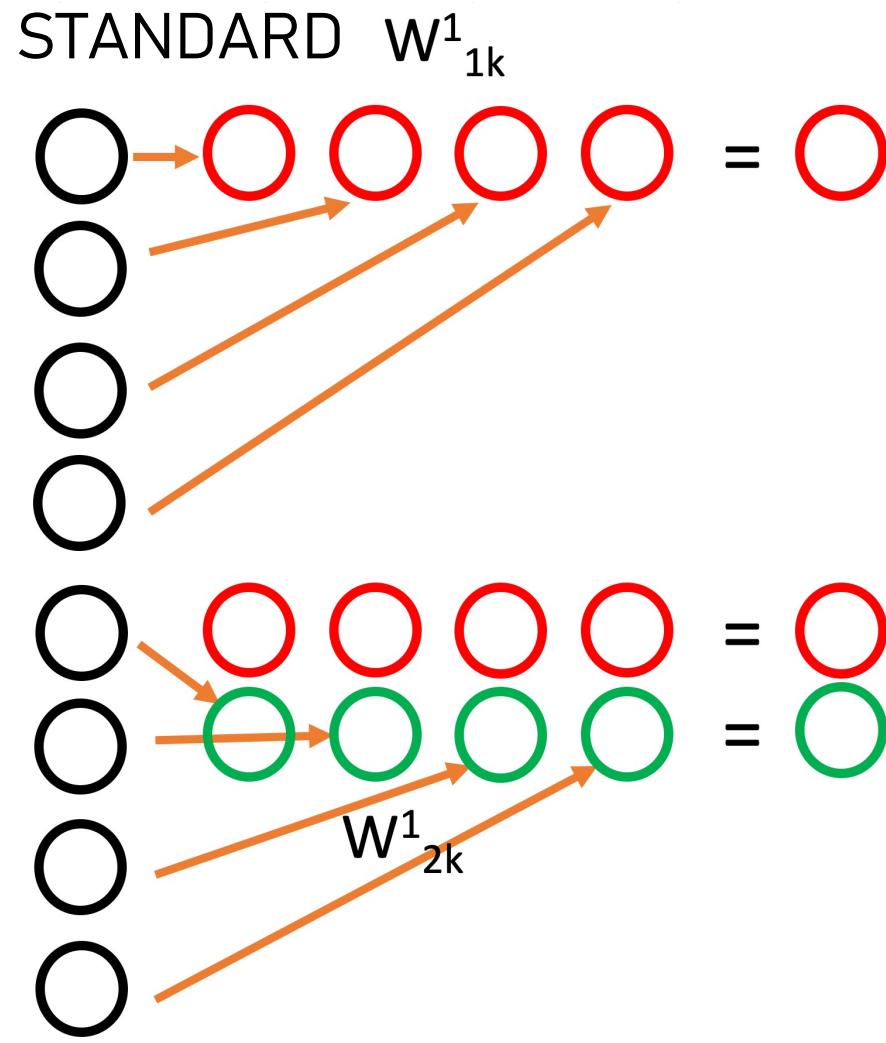
$$\Theta^{t+1} = \Theta^t - \eta_t \nabla J(\Theta^t)$$

Difficulté 1: comment calculer  $\frac{\partial J}{\partial w_{jk}^l}(\Theta^t), \frac{\partial J}{\partial b_j^l}(\Theta^t)$  pour tout  $j, k, l$  ?

Difficulté 2: comment choisir  $\eta_t$  ?

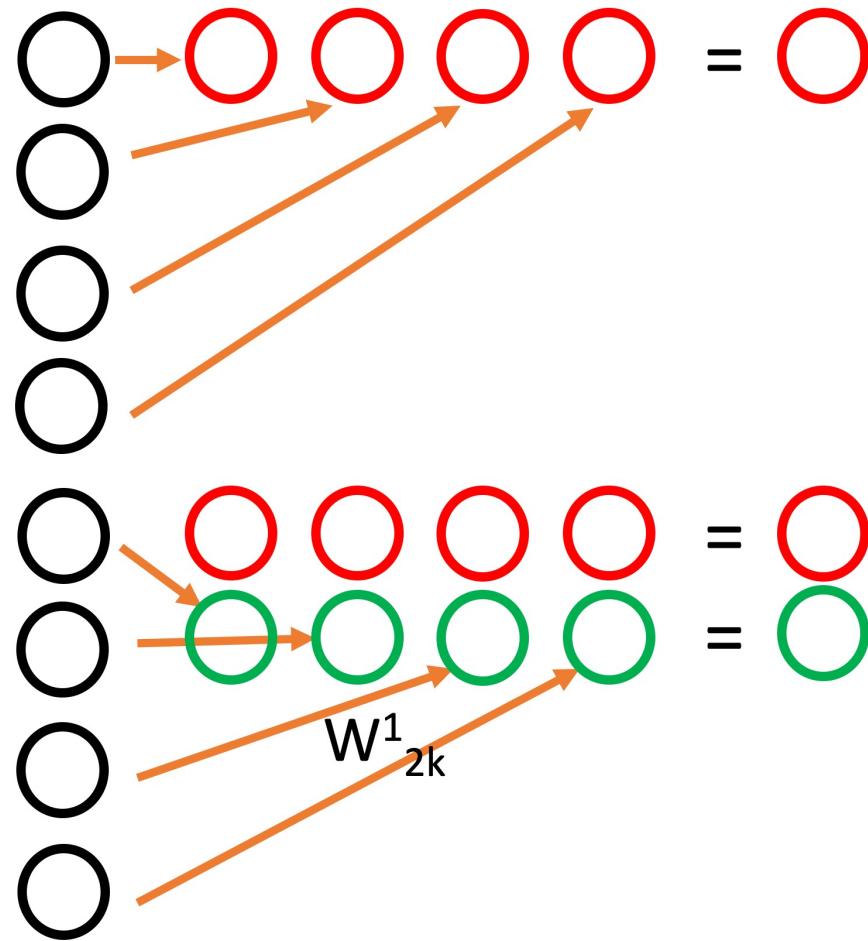
Constant, RMSProp, Adam, Adamax, Adadelta, Nadam, Adafactor, ...

## 3.2 Réseau convolutionnel

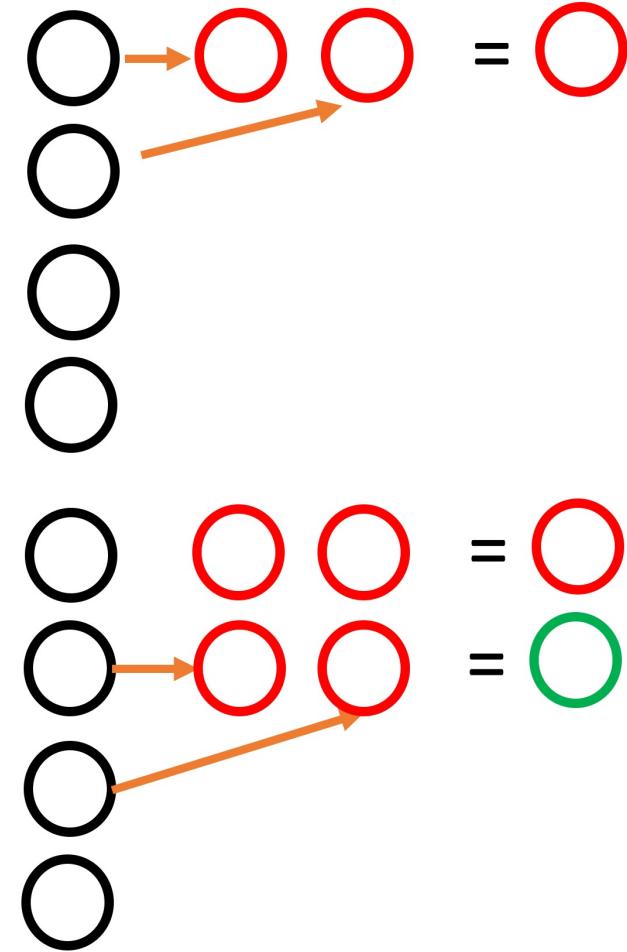


## 3.2 Réseau convolutionnel

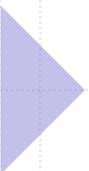
STANDARD  $W^1_{1k}$



CONVOLUTIONNEL



# 3.2 Réseau convolutionnel



$$W^1 = \begin{bmatrix} w_{11}^{1,(1)} & \dots & w_{13}^{1,(1)} \\ \vdots & \ddots & \vdots \\ w_{31}^{1,(1)} & \dots & w_{33}^{1,(1)} \end{bmatrix}$$

0	0	0	0	0	0	0	...
0	156	155	156	158	158	158	...
0	153	154	157	159	159	159	...
0	149	151	155	158	159	159	...
0	146	146	149	153	158	158	...
0	145	143	143	148	158	158	...
...	...	...	...	...	...	...	...

Input Channel #1 (Red)

-1	-1	1
0	1	-1
0	1	1

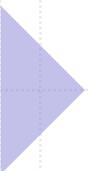
Kernel Channel #1



308

biais  
+ = ○

# 3.2 Réseau convolutionnel



$$W^1 = \begin{bmatrix} w_{11}^{1,(1)} & \dots & w_{13}^{1,(1)} \\ \vdots & \ddots & \vdots \\ w_{31}^{1,(1)} & \dots & w_{33}^{1,(1)} \\ \hline w_{11}^{1,(2)} & \dots & w_{13}^{1,(2)} \\ \vdots & \ddots & \vdots \\ w_{31}^{1,(2)} & \dots & w_{33}^{1,(2)} \end{bmatrix}$$

0	0	0	0	0	0	0	...
0	156	155	156	158	158	158	...
0	153	154	157	159	159	159	...
0	149	151	155	158	159	159	...
0	146	146	149	153	158	158	...
0	145	143	143	148	158	158	...
...	...	...	...	...	...	...	...

Input Channel #1 (Red)

-1	-1	1
0	1	-1
0	1	1

Kernel Channel #1

308

biais + = ○

0	0	0	0	0	0	0	...
0	156	155	156	158	158	158	...
0	153	154	157	159	159	159	...
0	149	151	155	158	159	159	...
0	146	146	149	153	158	158	...
0	145	143	143	148	158	158	...
...	...	...	...	...	...	...	...

Input Channel #1 (Red)

-1	-1	1
0	1	-1
0	1	1

Kernel Channel #1

310

biais + = ○

# 3.2 Réseau convolutionnel

0	0	0	0	0	0	0	...
0	156	155	156	158	158	158	...
0	153	154	157	159	159	159	...
0	149	151	155	158	158	159	...
0	146	146	149	153	158	158	...
0	145	143	143	148	158	158	...
...	...	...	...	...	...	...	...

Input Channel #1 (Red)

0	0	0	0	0	0	0	...
0	167	166	167	169	169	169	...
0	164	165	168	170	170	170	...
0	160	162	166	169	170	170	...
0	156	156	159	163	168	168	...
0	155	153	153	158	168	168	...
...	...	...	...	...	...	...	...

Input Channel #2 (Green)

0	0	0	0	0	0	0	...
0	163	162	163	165	165	165	...
0	160	161	164	166	166	166	...
0	156	158	162	165	165	166	...
0	155	155	158	162	167	167	...
0	154	152	152	157	167	167	...
...	...	...	...	...	...	...	...

Input Channel #3 (Blue)

-1	-1	1
0	1	-1
0	1	1

Kernel Channel #1

1	0	0
1	-1	-1
1	0	-1

Kernel Channel #2

0	1	1
0	1	0
1	-1	1

Kernel Channel #3

↓

161

+

↓

-9

+

↓

659

+ 1 = 812

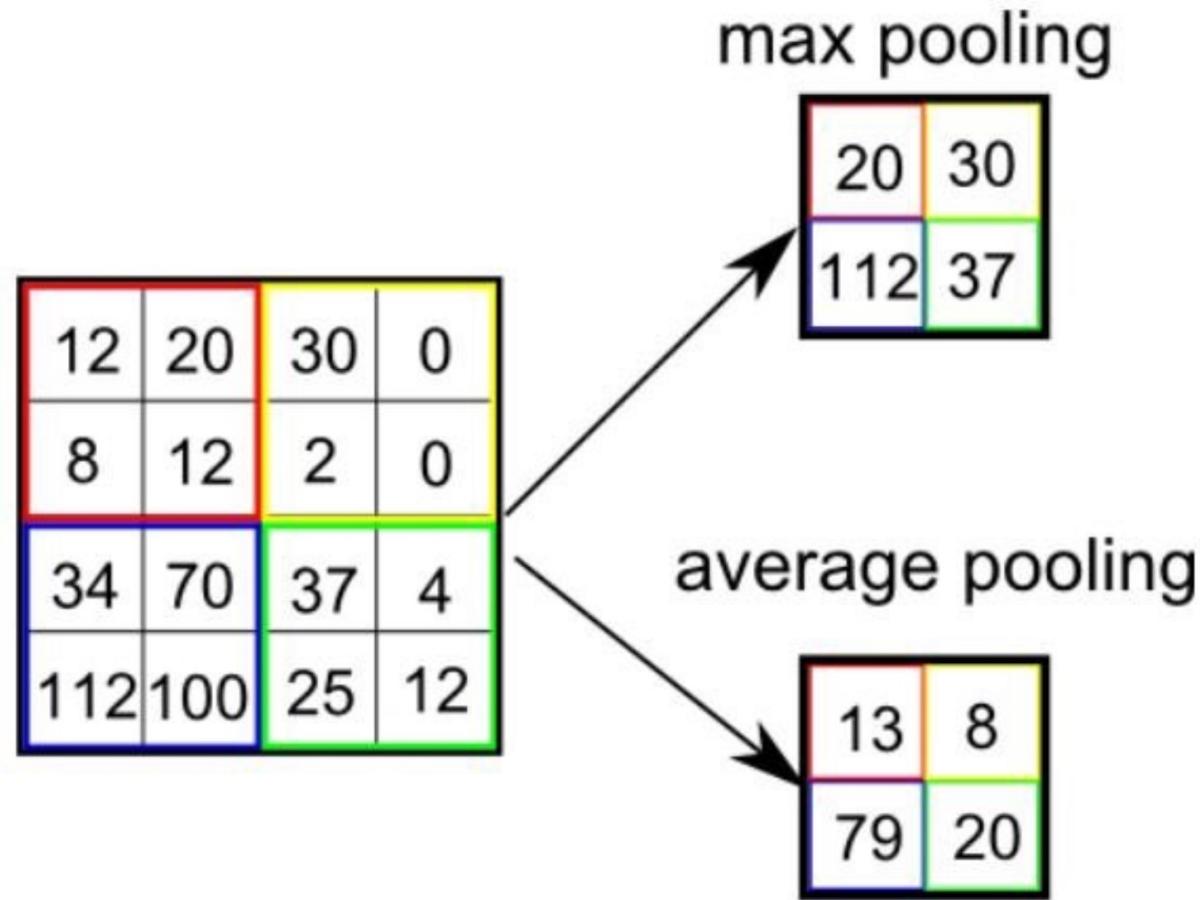
↑  
Bias = 1

Output

-25	466	466	475	...
295	787	798	812	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

Convolution operation on a MxNx3 image matrix with a 3x3x3 Kernel

# 3.2 Réseau convolutionnel



Types of Pooling

# 3.2 Réseau convolutionnel

