# Réseau de neurones

Quentin Behague,
Sous la direction de François Malgouyres,
IMT Toulouse

# Introduction

## Que sont les réseaux de neurones

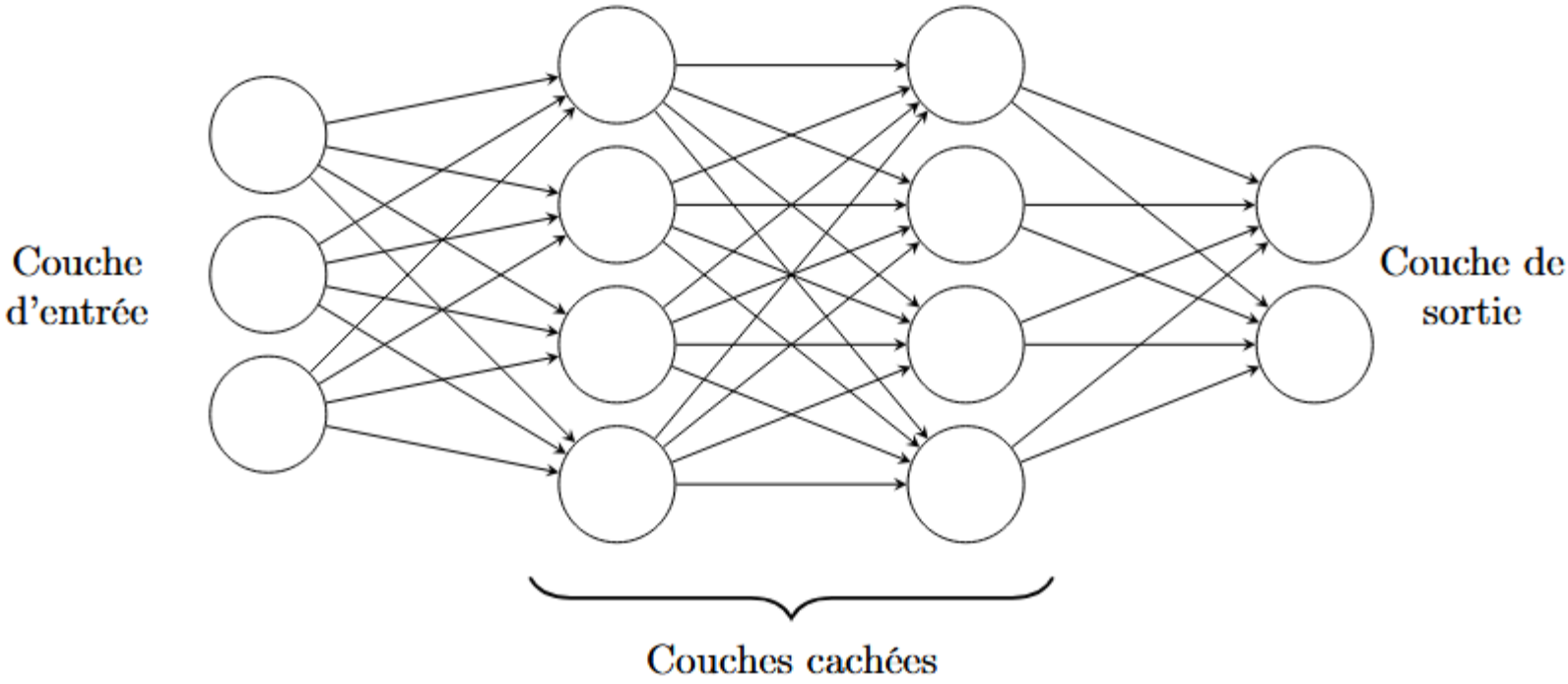


Figure – Perceptron multicouche

# Apprentissage supervisé

## Notation

- Espaces probabilisés : $(\mathcal{X}, \mathcal{P}(\mathcal{X}), \mathbb{P}_X)$ et $(\mathcal{Y}, \mathcal{P}(\mathcal{Y}), \mathbb{P}_Y)$.
- $(X, Y)$ un couple de variables aléatoires sur $\mathcal{X} \times \mathcal{Y}$
- Un échantillon $(x_i, y_i)_{i \in [\![1,n]\!]} \in (\mathcal{X} \times \mathcal{Y})^n$
- $L : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}$ mesurable et telle que $(x, y) \longmapsto L(g(x), y)$ soit $\mathbb{P}_{(X,Y)}$-intégrable sur $\mathcal{X} \times \mathcal{Y}$.

**Définition :** (Algorithme d'apprentissage)

On appelle **algorithme d'apprentissage** une fonction $\mathcal{A} : (\mathcal{X}, \mathcal{Y})^n \longrightarrow \mathcal{Y}^{\mathcal{X}}$. Le résultat d'un algorithme d'apprentissage est une **fonction de prédiction** $g : \mathcal{X} \longrightarrow \mathcal{Y}$ mesurable.

# Apprentissage supervisé

## Notion de risque

**Définition :** (Risque en population)

On appelle **risque en population** associé à la fonction de prédiction $g$ la quantité :

$$\mathcal{R}(g) := \mathbb{E}[L(g(X), Y)] = \int_{\mathcal{X}\times\mathcal{Y}} L(g(x), y) d\mathbb{P}_{(X,Y)}(x, y)$$

où $\mathbb{P}_{(X,Y)}$ désigne une loi de probabilité sur $\mathcal{X} \times \mathcal{Y}$.

**Définition :** (Risque empirique)

Soit $(x_i, y_i)_{i\in[\![1,n]\!]} \in \mathcal{X} \times \mathcal{Y}$ un jeu de données, on appelle **risque empirique** la quantité :

$$\hat{\mathcal{R}}(g) := \frac{1}{n} \sum_{i=1}^{n} L(g(x_i), y_i)$$

**Définition :** (Fonction de prédiction associée à un réseau de neurone)

On note, pour tout $h \in [\![0, H]\!]$, $g_h$ la fonction définie par récurrence par :

$$\forall x \in \mathbb{R}^{n_0} : g_h(x) = \begin{cases} x & \text{si } h = 0, \\ \sigma_h(W^{(h)} g_{h-1}(x) + b^{(h)}) & \text{si } h \in [\![1, H]\!] \end{cases}$$

La fonction $g_H : \mathbb{R}^{n_0} \longrightarrow \mathbb{R}^{n_H}$ est alors appelée la prédiction, ou l'inférence du réseau.

# Réseaux de neurones

## Entraînement d'un modèle : descente de gradient

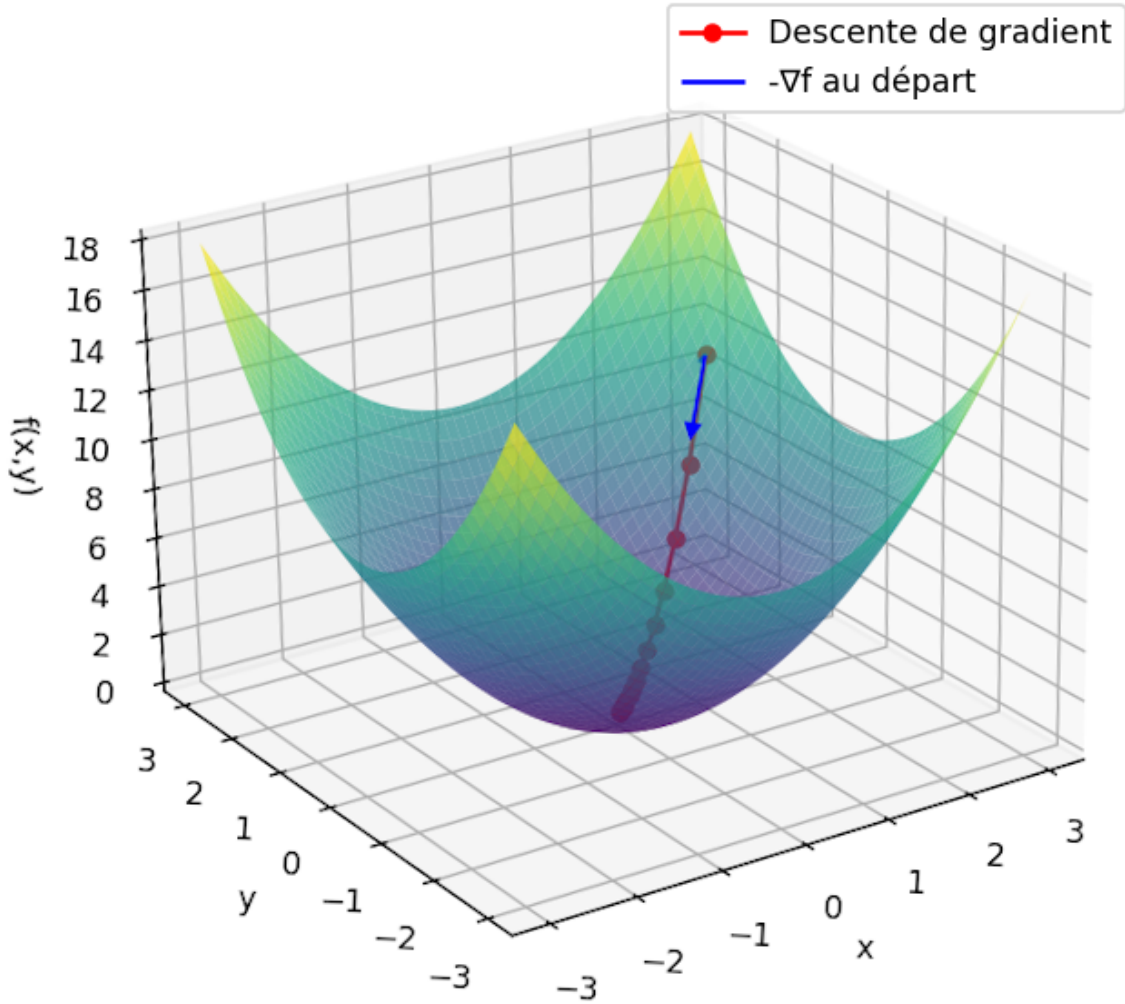


Figure – Descente de gradient

**Propriété :**
Si $f$ est convexe, à gradient $L$-lipschitzien et admet un minimiseur $x^*$, alors l'algorithme de descente de gradient, pour $(\eta_k)_{k\in\mathbb{N}} = \left(\frac{1}{L}\right)_{k\in\mathbb{N}}$ converge vers un minimum de $f$, et, pour tout $k \in \mathbb{N}$ :

$$f(x_n) - \min f \leq \frac{2L\|x_0 - x^*\|^2}{n}$$

**Propriété :**
En notant, pour tout $h \in [\![1, H]\!]$, $z_h(x) = W^{(h)} g_{h-1}(x) + b^{(h)}$, on a, pour tout $(x, y) \in \mathcal{X} \times \mathcal{Y}$, pour tout $(i, j) \in [\![1, n_h]\!] \times [\![1, n_{h-1}]\!]$ :

$$\begin{cases} \dfrac{\partial C}{\partial W_{i,j}^{(h)}}(\theta) = (g_{h-1}(x))_j (\sigma_h'(z_h(x))_i \Delta_i^{(h)} \\ \\ \dfrac{\partial C}{\partial b_i^{(h)}}(\theta) = (\sigma_h'(z_H(x))_i \Delta_i^{(h)} \end{cases}$$

Avec, pour tout $h \in [\![1, H-1]\!], i \in [\![1, n_h]\!]$ :

$$\Delta_i^{(h)} = \begin{cases} \left( \dfrac{\partial C}{\partial x}(g_\theta(x), y) \right)_i & \text{si } h = H, \\ \displaystyle\sum_{j=1}^{n_{h+1}} \left( W_{j,i}^{(h+1)} \sigma'_{(h+1)}(z_{h+1}(x))_j \Delta_j^{(h+1)} \right) & \text{sinon.} \end{cases}$$

**Propriété :**
Pour toute architecture $(n_1, \cdots, n_H)$ et pour tout échantillon $(x_i, y_i)_{i \in [\![1,n]\!]} \in (\mathbb{R}^{n_0} \times \mathbb{R}^{n_H})^n$, $\theta \longmapsto \hat{\mathcal{R}}(\theta)$ n'est pas coercive.

# Réseaux ReLu

## Exemple non-convexe

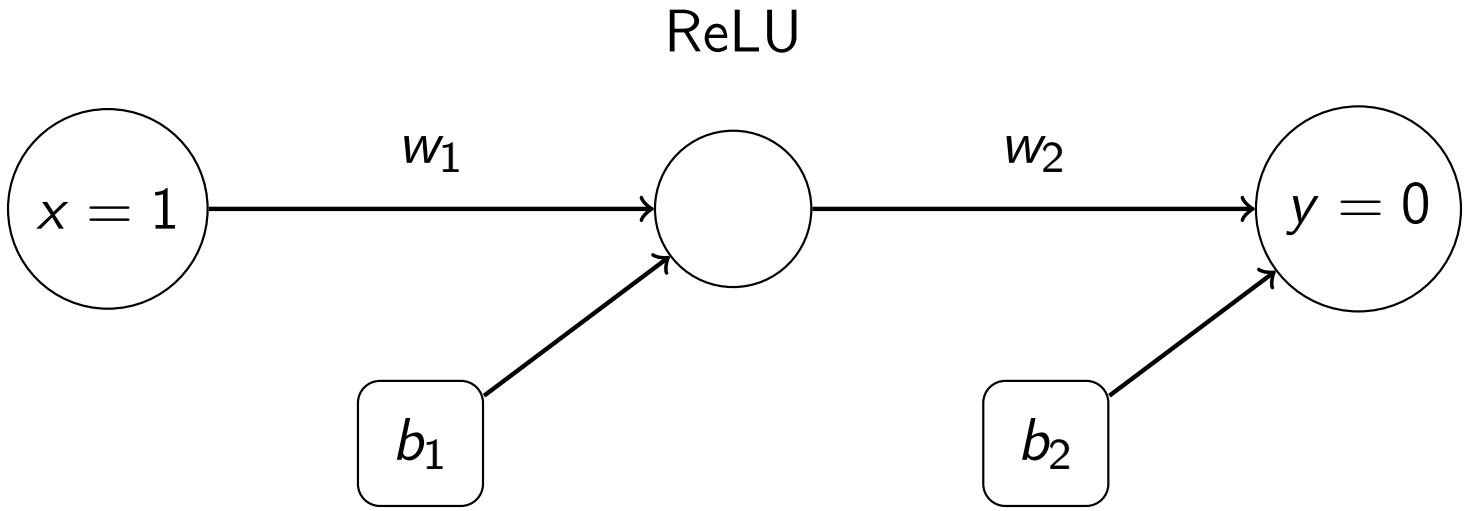


Figure – Réseau $(1, 1, 1)$

Sur cet exemple, avec $b_1 = b_2 = 0$ et $w_1 > 0$, on a :

$$\hat{\mathcal{R}}(\theta) = (w_1 w_2)^2$$

Dans la suite on note, pour tout $\theta \in \Theta$, $h \in [\![1, H-1]\!]$, $x \in \mathbb{R}^{n_0}$, $a_h(x,\theta) \in \{0,1\}^{n_h}$ le vecteur défini par :

$$\forall i \in [\![1, n_h]\!], (a_h(x,\theta))_i = \begin{cases} 1 & \text{si } (W^{(h)} g_{h-1}(x) + b^{(h)})_i > 0 \\ 0 & \text{sinon.} \end{cases}$$

**Définition :** On appelle motif d'activation du réseau la fonction :

$$a(x,\cdot) : \begin{array}{ccc} \Theta & \longrightarrow & \{0,1\}^{n_1} \times \cdots \times \{0,1\}^{n_{H-1}} \\ \theta & \longmapsto & (a_h(x,\theta))_{1 \leq h \leq H}. \end{array}$$

# Réseaux ReLu

## Motif d'activation

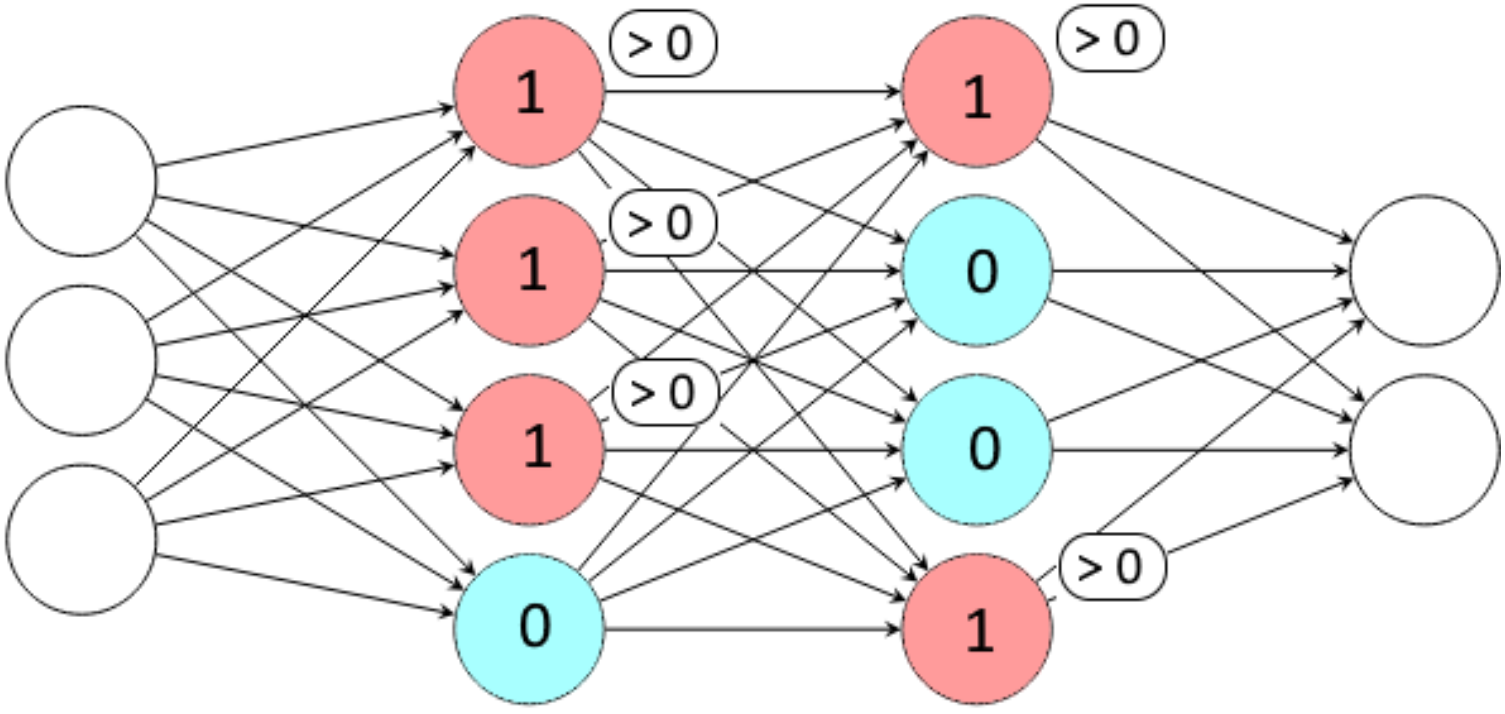


Figure – Motif d'activation

# Réseau ReLu

## Prédiction affine par morceaux

**Définition :** Soit $\theta \in \Theta$ fixé, et soit $\delta \in \{0,1\}^{n_1} \times \cdots \times \{0,1\}^{n_H - 1}$ un motif d'activation, on note :

$$D_\delta(\theta) := \{x \in \mathbb{R}^{n_0} \mid a(x,\theta) = \delta\}$$

**Propriété :**
Pour un réseau $\mathrm{ReLU}$, avec $\theta \in \Theta$ fixé, et $\delta \in \{0,1\}^{n_1} \times \cdots \times \{0,1\}^{n_{H-1}}$ un motif d'activation, les fonctions $x \longmapsto g_h(x)$ sont continues et affines par morceaux.

# Réseau ReLu

## Prédiction affine par morceaux

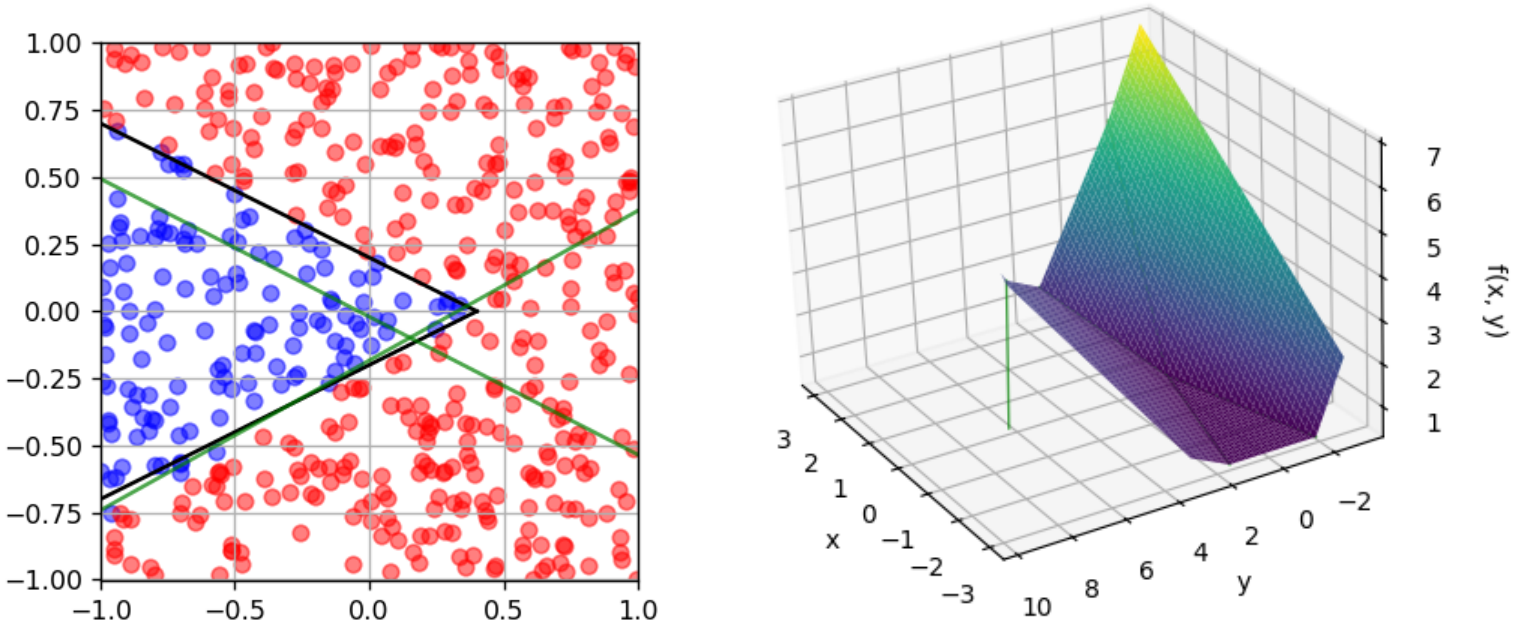


Figure – Représentation des zones affines $D_\delta(\theta)$ pour un réseau entraîné à la classification binaire

**Propriété :**
Les ensembles $D_\delta(\theta)$ sont des polyèdres, et, pour tout $h$ la restriction de $g_h$ à $D_\delta(\theta)$ est affine pour tout motif d'activation $\delta$.

# Réseaux ReLu

## Régularisation implicite

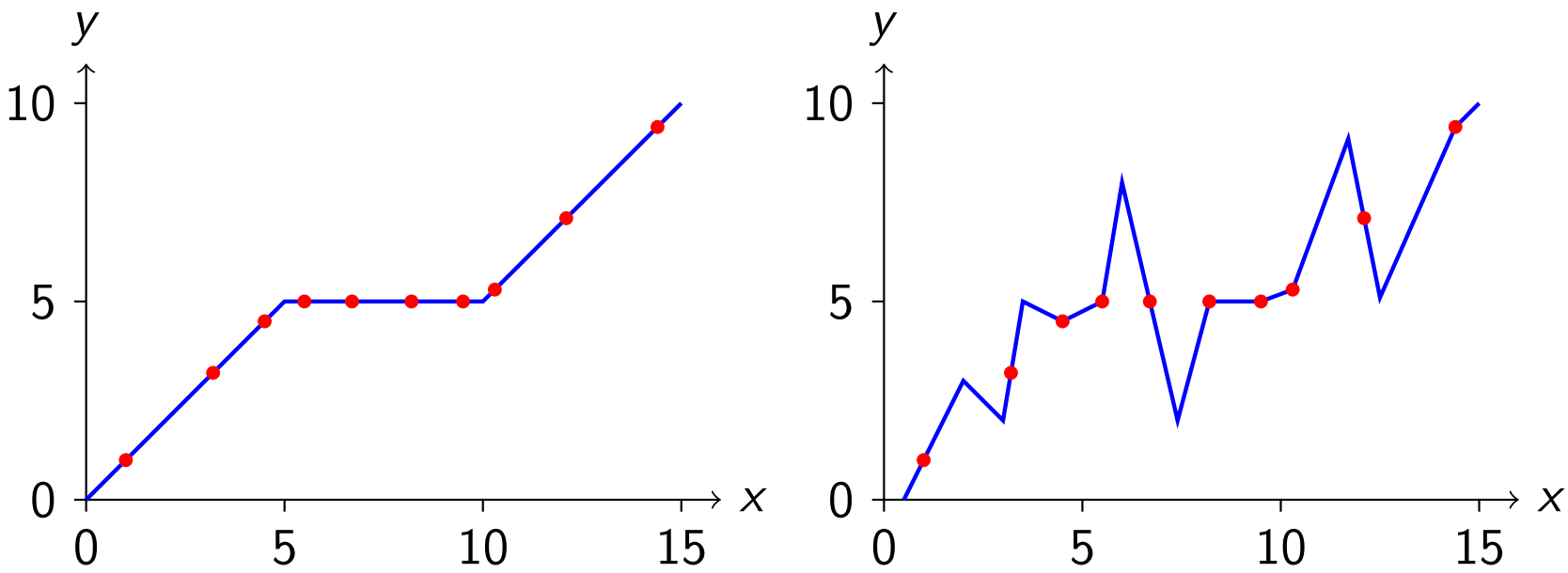


Figure – Deux prédictions (en bleu) minisant le risque empirique ($\hat{\mathcal{R}}(g_\theta) = 0$) sur un même échantillon d'apprentissage (en rouge)

# Réseaux ReLu

## Régularisation implicite

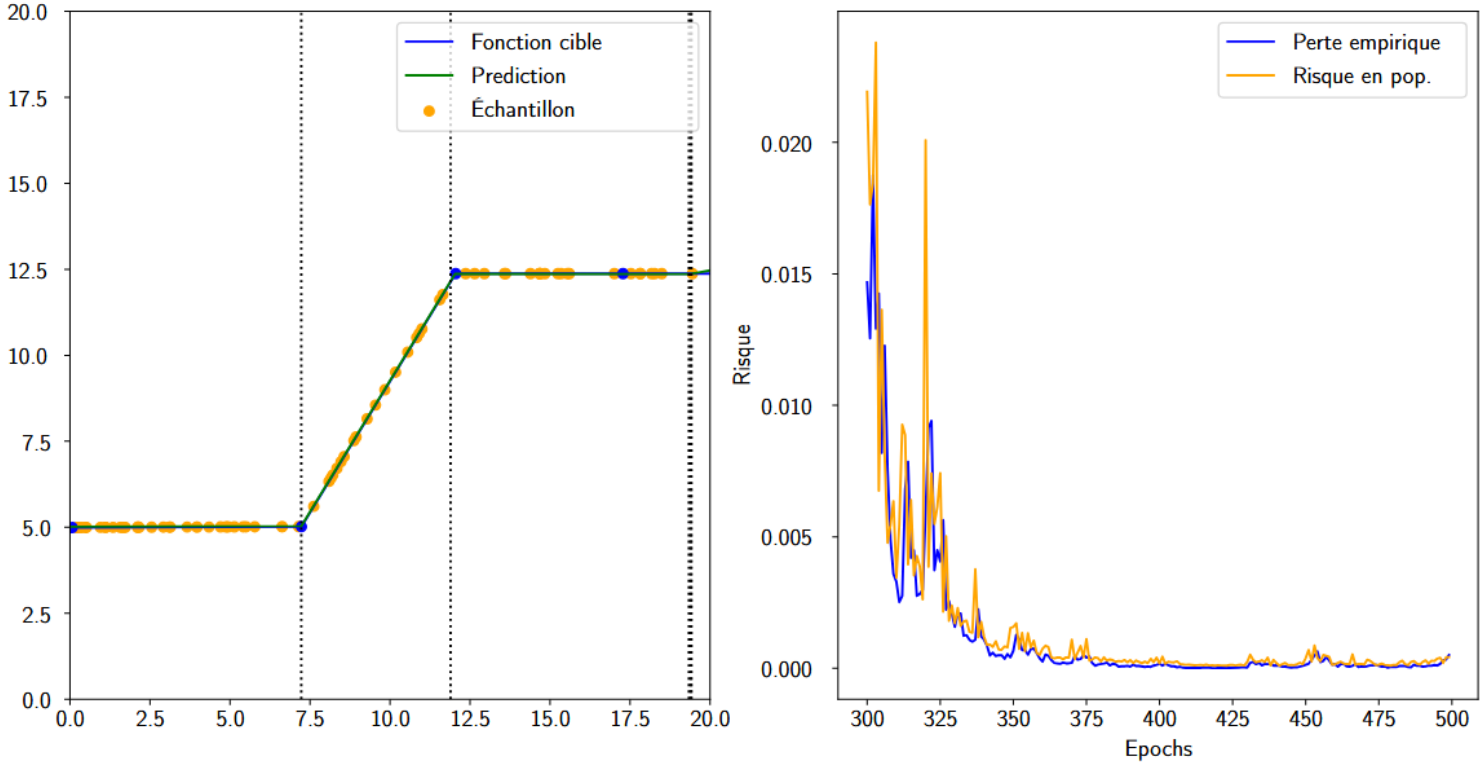


Figure – Entraînement avec taux d'apprentissage 0.05 sur 500 epochs

# Réseaux ReLu

## Régularisation implicite

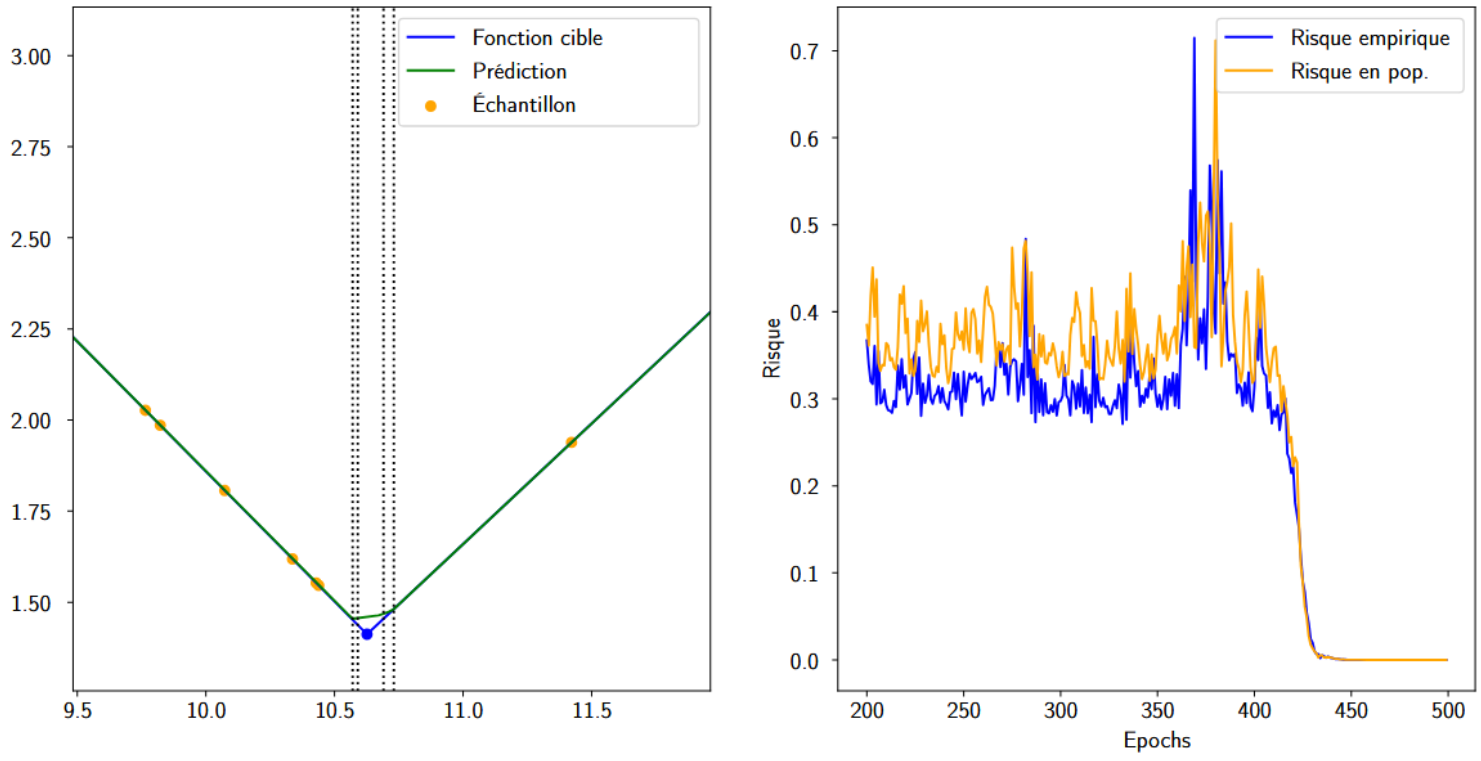


Figure – Comportement imprévisible sur un changement de zone affine peu couvert par les données

Annexes

# Annexe 1

## méthode de gradient continue

**Propriété :**
Si $f$ est continue et différentiable sur $E$, convexe, coercive, que son gradient est $L$-lipschitz et si $\theta$ est solution de :

$$\begin{cases} \theta(0) &= \theta_0 \in E \\ \dot{\theta}(t) &= -\nabla f(\theta(t)) \end{cases}$$

Alors, $\lim\limits_{t \to +\infty} f(\theta(t))$ existe et :

$$\lim_{t \to +\infty} f(\theta(t)) \in \operatorname*{argmin}_{\theta \in E} f(\theta)$$

# Annexe 1

## méthode de gradient continue

Tout d'abord, le minimum de $f$ existe, soit $\theta^* \in \underset{\theta \in E}{\operatorname{argmin}} f(\theta)$ car $f$ est coercive et convexe (voir 3.2). De plus $f(\theta(t))$ est décroissante :

$$\frac{\mathrm{d}f(\theta(t))}{\mathrm{d}t} = \nabla f(\theta(t))^\top \dot{\theta}(t) = \nabla f(\theta(t))^\top (-\nabla f(\theta(t))) = -\|\nabla f(\theta(t))\|^2 \leq 0$$

Comme $f$ admet un minimum, $t \longmapsto f(\theta(t))$ est bornée inférieurement et décroissante, donc tend vers une limite $f_\infty$. Ainsi, en posant $\theta_0 = \theta(0) < +\infty$, on en déduis que le gradient de $t \longmapsto f(\theta(t))$ est carré intégrable :

$$\frac{\mathrm{d}f(\theta(t))}{\mathrm{d}t} = -\|\nabla f(\theta(t))\|^2 \Longrightarrow \int_0^{+\infty} \|\nabla f(\theta(t))\|^2 \mathrm{d}t = f(\theta_0) - f_\infty < +\infty$$

Or, $t \longmapsto |\nabla f(\theta(t))|$ est positive, carré intégrable et lipschitzienne, donc elle converge vers $0 = |\nabla f(\theta^*)|$. Ainsi, par continuité de $\nabla f$, $t \longmapsto \theta(t)$ converge vers l'unique minimum global $\theta^*$.

# Annexe 2

## Inégalité de Taylor -Lagrange avec gradient Lipschitz

**Théorème :**
**Inégalité de Taylor-Lagrange avec gradient lipschitzien**
Soit $f : E \longrightarrow \mathbb{R}$ une fonction dérivable sur $E$, avec gradient $L$-lipschitzien. alors :

$$f(y) \leq f(x) + \langle \nabla f(x) | y - x \rangle + \frac{L}{2} \|x - y\|^2$$

Si de plus $f$ est convexe, alors, on a :

$$f(x) + \langle \nabla f(x) | y - x \rangle \leq f(y) \leq f(x) + \langle \nabla f(x) | y - x \rangle + \frac{L}{2} \|x - y\|^2$$

# Annexe 2

## Inégalité de Taylor -Lagrange avec gradient Lipschitz

Notons $g : t \in [0,1] \mapsto f((1-t)x + ty)$. C'est une fonction dérivable, et on a :

$$g'(t) = \langle \nabla f((1-t)x + ty), y - x \rangle.$$

Ainsi,

$$f(y) = g(1) = g(0) + \int_0^1 g'(t)\,\mathrm{d}t = f(x) + \int_0^1 \langle \nabla f((1-t)x + ty), y - x \rangle\,\mathrm{d}t.$$

Pour tout $t \in [0,1]$, par hypothèse sur la $L$-lipschitzianité du gradient :

$$\|\nabla f((1-t)x + ty) - \nabla f(x)\| \leq Lt\|x - y\|,$$

ce qui implique :

$$\langle \nabla f((1-t)x + ty), y - x \rangle \leq \langle \nabla f(x), y - x \rangle + Lt\|x - y\|^2.$$

En intégrant, on obtient :

$$f(y) \leq f(x) + \int_0^1 \left( \langle \nabla f(x), y - x \rangle + Lt\|x - y\|^2 \right) dt = f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2.$$

# Annexe 3

## Gradient du risque empirique

Ayant posé, pour tout $h \in [\![1, H]\!]$, $z_h(x) = W^{(h)} g_{h-1}(x) + b^{(h)}$, on applique la règle de la chaine, qui nous donne :

$$\frac{\partial C}{\partial W_{i,j}^{(h)}}(\theta) = \frac{\partial (z_h(x))_i}{\partial W_{i,j}^{(h)}} \frac{\partial (g_h(x))_i}{\partial (z_h(x))_i} \Delta_i^{(h)}$$

Or :

- $\forall h \in [\![1, H]\!], z_h(x) = W^{(h)} g_{h-1}(x) + b^{(h)} \Longrightarrow \dfrac{\partial (z_h(x))_i}{\partial W_{i,j}^{(h)}} = (g_{h-1}(x))_j$
- $\forall h \in [\![1, H]\!], g_h(x) = \sigma(z_h(x)) \Longrightarrow \dfrac{\partial (g_h(x))_i}{\partial (z_h(x))_i} = \sigma_h'(z_h(x))_j$

De même, par la règle de la chaine, on obtient, pour tout $h \in [\![1, H-1]\!], i \in [\![1, n_h]\!]$ :

$$\Delta_i^{(h)} = \sum_{j=1}^{n_{h+1}} \left( \frac{\partial (z_{h+1}(x))_j}{\partial (g_h(x))_i} \frac{\partial (g_{h+1}(x))_j}{\partial (z_{h+1}(x))_j} \Delta_j^{(h+1)} \right)$$

Or :

- $\forall h \in [\![1, H-1]\!], z_{h+1}(x) = W^{(h+1)} g_h(x) + b^{(h+1)} \Longrightarrow \dfrac{\partial (z_{h+1}(x))_j}{\partial (g_h(x))_i} = W_{j,i}^{(h+1)}$
- $\forall h \in [\![1, H-1]\!], g_{h+1}(x) = \sigma(z_{h+1}(x)) \Longrightarrow \dfrac{\partial (g_{h+1}(x))_i}{\partial (z_{h+1}(x))_i} = \sigma_h'(z_{h+1}(x))_j$

D'où le résultat.