# OPTIMISATION POUR L'APPRENTISSAGE AUTOMATIQUE

22 janvier 2025

- 3 séances restantes / 3 sujets
- Examen ⇒ Vendredi 7 février 14H-16H
  Documents : Feuille A4 recto-verso
- Projet ⇒ Deadline mars ?

# SOUS-GRADIENTS ET DIFFÉRENTIATION AUTOMATIQUE

## Résumé des épisodes précédents

→ Problème d'optimisation pour le ML typique :

* Données (ex : $\{(x_i, y_i)\}_{i=1..n}$)
* Objectif :

$$\underset{w \in \mathbb{R}^d}{\text{minimiser}} \quad \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Erreur d'apprentissage d'un modèle paramétré par $w$ sur le $i$-ème échantillon des données
(ex : $(x_i, y_i)$, $f_i(w) = \frac{1}{2}(x_i^T w - y_i)^2$)

→ Si les $f_i$ sont des fonctions $C^1$ de $w$, on peut résoudre le problème :

→ Par descente de gradient :

$$w_{k+1} = w_k - \alpha_k \nabla f(w_k)$$

avec $f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w)$

$\alpha_k > 0$

→ Par gradient stochastique

$$w_{k+1} = w_k - \alpha_k \nabla f_{i_k}(w_k)$$

avec $\alpha_k > 0$

et $i_k$ tiré aléatoirement dans $\{1, \dots, n\}$

$\rightarrow$ Variante par "fournées" (batch)

$$w_{k+1} = w_k - \alpha_k \frac{1}{|S_k|} \sum_{i \in S_k} \nabla f_i(w_k)$$

avec $S_k$ un ensemble d'indices tirés aléatoirement (avec/sans remise) dans $\{1, \dots, n\}$

Q) Comment calcule-t-on les $\nabla f_i$ ?

<u>Exemple</u> : • Soit $\{(x_i, y_i)\}_{i=1..n}$ un jeu de données avec $x_i \in \mathbb{R}^{d_0}$ et $y_i \in \mathbb{R}^{d_3}$ $\forall i$.

• Régression : on cherche un modèle $h(x; w)$ avec $w \in \mathbb{R}^d$ tel que $\|h(x_i; w) - y_i\|^2$ soit le plus faible possible

<u>Pb d'optimisation</u> : $\underset{w \in \mathbb{R}^d}{\text{minimiser}} \ \frac{1}{n} \sum_{i=1}^{n} \underbrace{\|h(x_i; w) - y_i\|^2}_{f_i(w)}$

• On considère un réseau de neurones à 3 couches comme modèle :

$$\forall x \in \mathbb{R}^{d_0}, \quad h(x; w) = W_3 \, \mathrm{ReLU}\left(W_2 \, \mathrm{ReLU}\left(W_1 x + b_1\right) + b_2\right) + b_3$$

avec $\mathrm{ReLU}(t) = \max(t, 0)$ appliqué composante à composante

$W_1 \in \mathbb{R}^{d_1 \times d_0}$, $b_1 \in \mathbb{R}^{d_1}$, $W_2 \in \mathbb{R}^{d_2 \times d_1}$, $b_2 \in \mathbb{R}^{d_2}$, $W_3 \in \mathbb{R}^{d_3 \times d_2}$, $b_3 \in \mathbb{R}^{d_3}$

$w$ : concaténation des paramètres $(W_1, b_1, W_2, b_2, W_3, b_3)$

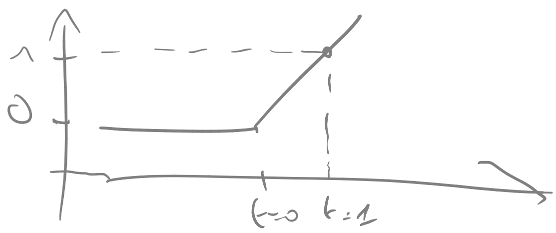
$\Rightarrow d = d_1 d_0 + d_1 + d_2 d_1 + d_2 + d_3 d_2 + d_3$

→ Si on veut appliquer la descente de gradient / le gradient stochastique, on a besoin de dériver $h(x_i; w)$ par rapport à $w$ (pour obtenir $\nabla f_i(w)$)

## 1 Sous-gradients

→ La fonction ReLU : $\mathbb{R} \to \mathbb{R}$ n'est pas dérivable en 0

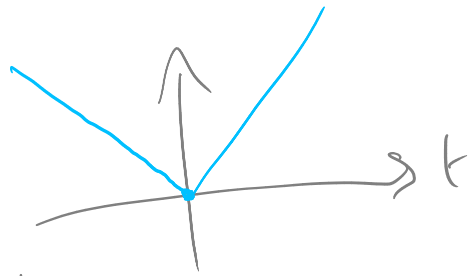
$\Rightarrow$ La notion de gradient d'une fonction basée sur ReLU n'est pas bien définie




Def : On dit que $f : \mathbb{R}^d \to \mathbb{R}$ est non lisse (nonsmooth) s'il existe des points de $\mathbb{R}^d$ en lesquels $f$ n'est pas dérivable.

Ex) $d = 1$ : $t \mapsto |t|$ pas dérivable en 0

$t \mapsto \max(t, 0)$

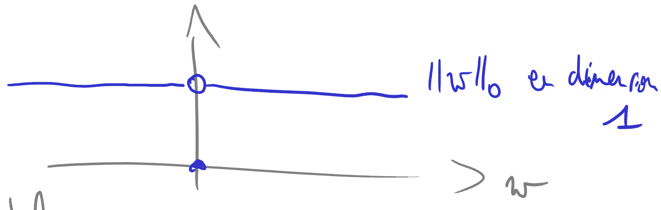



$d \geq 1$ : $w \mapsto \|w\|_1 = \sum_{j=1}^{d} |w_j|$

pas dérivable en tout point qui a au moins une coordonnée nulle.

$w \mapsto \|w\|_0 :=$ nombre de coefficients non nuls de $w$

$d = 1 \qquad \|w\|_0 = \begin{cases} 1 & \text{si } w \neq 0 \\ 0 & \text{sinon} \end{cases}$




pas dérivable en 0

→ Focus dans ce cours : fonctions non lisses mais continues et convexes (écarte notamment $\|w\|_0$)

<u>Rappel</u> : Si $f$ est $C^1$ convexe, alors

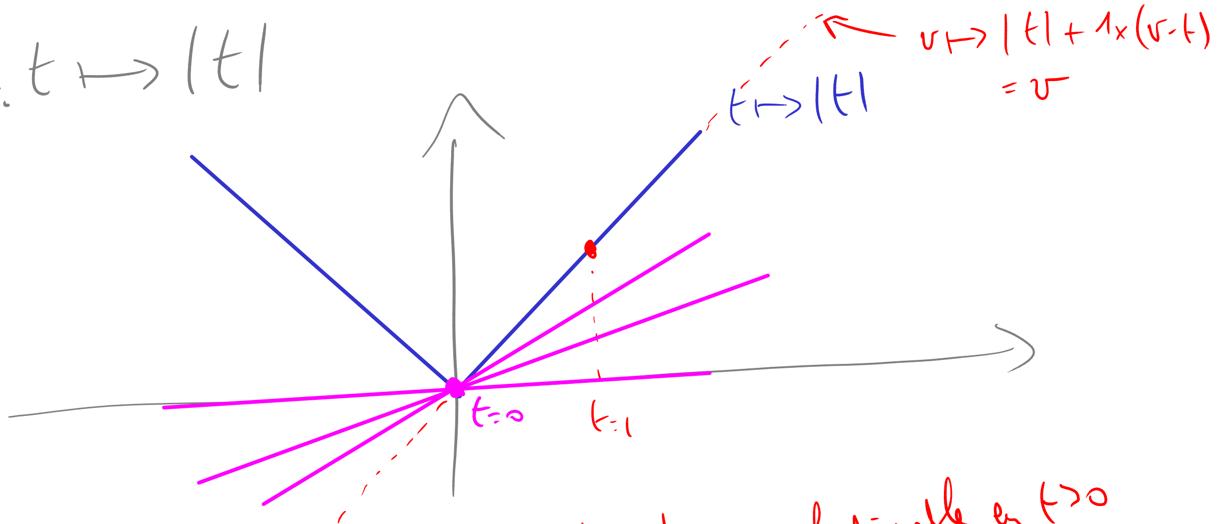
$$\forall (v,w) \in (\mathbb{R}^d)^2, \quad f(v) \geq f(w) + \nabla f(w)^T (v - w)$$

<u>Définition</u> : Soit $f$ convexe et $w \in \mathbb{R}^d$.
On dit que $g \in \mathbb{R}^d$ est un sous-gradient de $f$ en $w$

si $\quad \forall v \in \mathbb{R}^d, \quad f(v) \geq f(w) + g^T(v - w)$

L'ensemble des sous-gradients de $f$ en $w$ s'appelle le sous-différentiel de $f$ en $w$, et on le note $\partial f(w)$.

<u>Exemple</u> : $f : t \mapsto |t|$




Pour $t > 0$, $|t| = t$ → $f$ dérivable en $t > 0$ et $f'(t) = 1$

$g \in \mathbb{R}$ tq $\quad \forall v \in \mathbb{R}, \quad f(v) \geq f(t) + 1 \times (v - t)$ → 1 est la seule valeur qui vérifie cette propriété

$f(v) \geq f(1) + g(v-1) \quad \forall v \in \mathbb{R}$ $\quad\quad |v| \geq |t| + v - t$

$\iff g = 1$ $\quad\quad t=1 \quad |v| \geq v \quad \forall v \in \mathbb{R}$

Pour $t = 0$, l'on a $\quad f(v) \geq f(0) + g(v-0)$

$\forall g \in [-1, 1]$

$|v| \geq 0 + gv \quad \forall g \in [-1,1]$

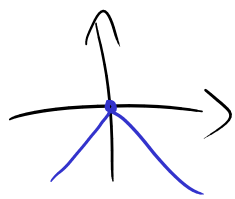
$\Rightarrow \partial f(0) = [-1, 1]$

↳ Si $f$ est dérivable en $w$, alors $\underbrace{\partial f(w)}_{\subseteq \mathbb{R}^d} = \{ \underbrace{\nabla f(w)}_{\in \mathbb{R}^d} \}$

(le sous-différentiel est un singleton)

<u>NB</u> : Si $f$ est non convexe, $\partial f(w)$ peut être vide

Ex) $f(t) = -|t|$



## Sous-gradients et optimisation

<u>Pb</u> : $\quad \underset{w \in \mathbb{R}^d}{\text{minimiser}} \; f(w) \quad$ avec $f$ continue, <u>convexe</u>

<u>Cas $f$ $C^1$</u> : $\quad$ . <u>Algorithme</u> : $\quad w_{k+1} = w_k - \alpha_k \nabla f(w_k)$

. <u>Condition d'optimalité</u> : $\quad w^* \in \underset{w}{\operatorname{argmin}} f(w) \iff \nabla f(w^*) = 0_{\mathbb{R}^d}$

Cas $f$ non lisse : . Condition d'optimalité :

$$w^* \in \underset{w}{\operatorname{argmin}}\, f(w) \iff 0_{\mathbb{R}^d} \in \underbrace{\partial f(w^*)}_{\subseteq \mathbb{R}^d}$$

(+) Généralise le cas $C^1$ ($f$ $C^1$ en $w^*$ $\Rightarrow$ $0_{\mathbb{R}^d} \in \{\nabla f(w^*)\}$)

(−) Vérifier cette condition requiert de calculer tout le sous-différentiel (*), ce qu'on ne fait pas en pratique

(*) ou de résoudre $\underset{g \in \mathbb{R}^d}{\text{minimiser}} \ \|g\|^2$ s.c. $g \in \partial f(w^*)$

. Algorithme :

$$w_{k+1} = w_k - \alpha_k g_k \quad \text{avec } \alpha_k > 0 \text{ et } g_k \in \partial f(w_k)$$

(+) Généralise la descente de gradient

$f$ dérivable en $w_k$ $\Rightarrow$ $w_{k+1} = w_k - \alpha_k \nabla f(w_k)$

(−) Le choix du sous-gradient influence fortement la performance de l'algorithme

Ex) $d=1$ $\quad f : w \longmapsto |w|$ $\qquad w_k = 0$, $\quad g_k \neq 0$, $g_k \in [-1,1] = \partial f(w_k)$

$$0 - \overset{>0}{\alpha_k}\ \overset{\neq 0}{g_k}$$

$$\forall \alpha_k > 0, \quad |w_k - \alpha_k g_k| > \underset{0}{\underset{\parallel}{|w_k|}}$$

$$e_j = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow j$$

Ex) $d \geq 2$, $\quad f : w \longmapsto \|w\|_1 = \sum_{j=1}^{d} |w_j|$

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \updownarrow 1 \\ \updownarrow d-1 \end{matrix} \qquad \partial f(e_1) = \left\{ g \in \mathbb{R}^d \ \middle|\ g = e_1 + \sum_{j=2}^{d} t_j e_j, \ t_j \in [-1,1] \ \forall j \right\}$$

$\forall g \in \partial f(e_1)$ tel que $g = e_1 + \sum_{j=2}^{d} t_j e_j$, $\sum_{j=2}^{d} |t_j| > 1$

$\forall \alpha > 0$, $\|e_1 - \alpha g\|_1 > \|e_1\|_1$

<u>TP</u>:

$$f(w) = \frac{1}{m} \| Xw - y \|_1 = \frac{1}{m} \sum_{i=1}^{m} | x_i^T w - y_i |$$

$$\partial f(w) = \left\{ g = \sum_{i=1}^{m} \varepsilon_i x_i \right., \text{ avec } \varepsilon_i \begin{cases} = 1 & \text{si } x_i^T w - y_i > 0 \\ = -1 & \text{si } x_i^T w - y_i < 0 \\ \in [-1, 1] & \text{si } x_i^T w - y_i = 0 \end{cases}$$

$\Rightarrow$ on choisit $g \in \partial f(w)$ tel que $\varepsilon_i = 0$ si $x_i^T w - y_i = 0$

$$g = X^T \varepsilon \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad X = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix}$$