

ALGORITHMES STOCHASTIQUES

Notes de cours du module *Algorithmes Stochastiques* du LPSM⁰, Sorbonne Université 2018–2019

TABLE DES MATIÈRES

1	Calcul récursif de moyenne et variance empiriques	2
1.1	Moyenne empirique	2
1.2	Variance empirique	3
2	Modèle générique d'algorithmes stochastiques	4
2.1	Heuristique	4
2.2	Descente de gradient déterministe	5
2.3	Premier retour au stochastique	8
2.4	Probabilités numériques versus datasciences	8
3	Exemples d'algorithmes stochastiques	8
3.1	Science des données	9
3.2	Apprentissage supervisé : un pas plus loin	10
3.3	Du perceptron multi-couche vers le deep learning	11
3.4	Approximation universelle	12
	Références	13

0. Laboratoire de Probabilités, Statistiques et Modélisation : <https://www.lpsm.paris>

INTRODUCTION

La descente de gradient peut être vue comme un problème d'optimisation ou de recherche de zéros. Le premier réduit souvent le second à un problème de minimisation d'une fonction en cherchant le zéro de son gradient, en l'occurrence si la fonction est convexe. L'exemple le plus courant est l'extraction de paramètres implicites : ces exemples peuvent être vus comme des moyennes, c'est-à-dire comme des fonctions définies par

$$h(y) = \mathbf{E}[H(y, Z)]$$

où Z est un q -vecteur aléatoire. Le propos de ce chapitre est de fournir un outil - l'approximation stochastique basé sur des simulations résolvants ces problèmes d'optimisation ou recherche de zéros. Cela peut être vu comme une extension des méthodes de Monte Carlo.

L'approximation stochastique peut être représentée comme une extension probabiliste de la méthode de Newton-Raphson sous sa forme récursive

$$y_{n+1} = y_n - \gamma_{n+1} h(y_n)$$

pour tout $n > 0$ et $0 < \gamma_n < \gamma_0$, où $h : \mathbb{R}^d \rightarrow \mathbb{R}^d$ est continue et a un accroissement linéaire à l'infini. Sous quelques hypothèses, on peut montrer qu'une telle méthode est bornée et converge vers un zéro y^* de h . Si on pose $\gamma_n = (Dh(y_{n-1}))^{-1}$, la méthode récursive proposée n'est autre que la méthode de Newton-Raphson. On peut aussi poser $\gamma_n = 1$ et remplacer h par $(Dh)^{-1} \circ h$.

1 CALCUL RÉCURSIF DE MOYENNE ET VARIANCE EMPIRIQUES

Un premier exemple d'approximation stochastique est méthode de Monte Carlo classique, c'est-à-dire par calcul de moyenne empirique et de variance empirique.

1.1 MOYENNE EMPIRIQUE

Soit $X \in L^1(\Omega, \mathcal{A}, \mathbf{P})$ de loi μ et $(X_n, n \geq 1)$ une suite *i.i.d.* de loi μ et la moyenne empirique

$$\bar{X}_n := \frac{X_1 + \cdots + X_n}{n}$$

D'abord, la loi forte des grands nombres nous donne presque-sûrement $m := \mathbf{E}[X] = \lim_n \bar{X}_n$.

D'autre part, on peut réécrire

$$\begin{aligned} \bar{X}_{n+1} &= \frac{n}{n+1} \bar{X}_n + \frac{X_{n+1}}{n+1} \\ &= \bar{X}_n - \frac{1}{n+1} (\bar{X}_n - X_{n+1}) \end{aligned}$$

Exercice : Faire la même chose pour la récursivité par paquet $\bar{X}_{n+p} = \bar{X}_n - \frac{p}{n+p} (\bar{X}_n - \frac{X_{n+1} + \cdots + X_{n+p}}{p})$

Soit maintenant la filtration naturelle $\mathcal{F}_n = \sigma(X_1, \dots, X_n), n \geq 1$, $\mathcal{F}_0 := \{\emptyset, \Omega\}$.

On constate que pour tout n , X_{n+1} est indépendante de la tribu \mathcal{F}_n et \bar{X}_n est \mathcal{F}_n -mesurable. (*)
Alors

$$\begin{aligned}\mathbf{E}[\bar{X}_n - X_{n+1} | \mathcal{F}_n] &= \bar{X}_n - \mathbf{E}[X_{n+1} | \mathcal{F}_n] \\ &= h_1(\bar{X}_n)\end{aligned}$$

où $h_1(\bar{x}) = \bar{x} - m$. Soit $\gamma_n = \frac{1}{n}$. On obtient

$$\bar{X}_{n+1} = \bar{X}_n - \gamma_{n+1}h_1(\bar{X}_n) - \gamma_{n+1}\Delta M_{n+1}$$

où

$$\Delta M_{n+1} = (\bar{X}_n - X_{n+1}) - h_1(\bar{X}_n), \quad n \geq 0$$

est le processus d'accroissement d'une \mathcal{F}_n -martingale (donc $\mathbf{E}[\Delta M_n] = 0$ pour tout $n \geq 0$), qu'on considère comme une *perturbation*. C'est donc comme cela qu'on simulera la moyenne empirique de façon récursive. Et donc presque-sûrement on a bien

$$\bar{X}_n \rightarrow m = \{h_1 = 0\}$$

1.2 VARIANCE EMPIRIQUE

On suppose maintenant que $X \in L^2(\mathbf{P})$ avec $\sigma := \text{Var}(X)$ et on pose $\bar{V}_n = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2$.

Exercice : Montrer que pour $R_{n+1} = \frac{1}{n+1}(\bar{X}_n - X_{n+1})^2$,

$$\bar{V}_{n+1} = \bar{V}_n - \gamma_{n+1}[\bar{V}_n - (\bar{X}_n - X_{n+1})^2 + R_{n+1}]$$

Grâce aux mêmes conclusions que (*),

$$\begin{aligned}\mathbf{E}[\bar{V}_n - (\bar{X}_n - X_{n+1})^2 | \mathcal{F}_n] &= \mathbf{E}[\bar{V}_n - [(\bar{X}_n - m) - (X_{n+1} - m)]^2 | \mathcal{F}_n] \\ &= \bar{V}_n - (\bar{X}_n - m)^2 - \mathbf{E}[(X - m)^2] \\ &= h_2(\bar{X}_n, \bar{V}_n)\end{aligned}$$

où $h_2(\bar{x}, \bar{v}) = \bar{v} - [(\bar{x} - m)^2 + \sigma^2]$. On a donc

$$\begin{bmatrix} \bar{X}_{n+1} \\ \bar{V}_{n+1} \end{bmatrix} = \begin{bmatrix} \bar{X}_n \\ \bar{V}_n \end{bmatrix} - \gamma_{n+1} \begin{bmatrix} h_1(\bar{X}_n) \\ h_2(\bar{X}_n, \bar{V}_n) \end{bmatrix} - \gamma_{n+1} \begin{bmatrix} \Delta M_{n+1} \\ \Delta M_{n+1}^2 \end{bmatrix}$$

où $\Delta M_{n+1}^2 = [\bar{V}_n - (\bar{X}_n - X_{n+1})^2 + R_{n+1}] - h_2(\bar{X}_n, \bar{V}_n)$ et $\mathbf{E}[\Delta M_n^2] = \mathbf{E}[R_n]$.

Que peut-on dire de R_n ? Un argument à la Borel-Cantelli nous donne une condition sur la suite $R_n, n \geq 1$.

$$\mathbf{E}[\sum_{n \geq 1} R_n^2] = \sum_{n \geq 1} \frac{\mathbf{E}[(\bar{X}_n - X_{n+1})^4]}{(n+1)^2}$$

On utilise la convexité de $x \rightarrow x^4$ et en payant le prix que $X \in L^4(\mathbf{P})$, on obtient

$$\begin{aligned}\mathbf{E}[(\bar{X}_n - X_{n+1})^4] &\leq 2^3(\mathbf{E}[\bar{X}_n^4] + \mathbf{E}[X_{n+1}^4]) \\ &\leq 2^4 \mathbf{E}[X^4]\end{aligned}$$

Exercice : Montrer que $X \in L^{2+\varepsilon}(\mathbf{P})$ suffit.

Le lemme de Borel-Cantelli nous donne alors que $\mathbf{E}[\sum_{n \geq 1} R_n^2] < \infty$ implique que presque-sûrement $\sum_{n \geq 1} R_n^2 < \infty$ et donc $R_n \rightarrow 0$.

Remarque : Pourquoi ΔM_n peut être vu comme une perturbation ?
Si on suppose que $\sum_k \gamma_k^2 < \infty$ alors $\sum_k \gamma_k^2 \mathbf{E}[\Delta M_k^2] < \infty$ si $\sup_n \mathbf{E}[\Delta M_n^2] < \infty$. On remarque que

$$\sum_{k=n+1}^{n+p} \gamma_k^2 \mathbf{E}[\Delta M_k^2] = o\left(\sum_{k=n+1}^{n+p} \gamma_k\right)$$

car $\sum_k \gamma_k^2 < \infty$ implique $\gamma_n \rightarrow 0$ et $\gamma_n^2 = o(\gamma_n)$.

Pour conclure, pour $\theta_n = \begin{bmatrix} \bar{X}_n \\ \bar{V}_n \end{bmatrix}$, on a

$$\theta_{n+1} = \theta_n - \gamma_{n+1}(h(\theta_n) + \Delta M_{n+1}) \rightarrow \{h = 0\} = (m, \sigma^2)$$

2 MODÈLE GÉNÉRIQUE D'ALGORITHMES STOCHASTIQUES

Soit $\Theta_0, \pi_1, \pi_2, \dots$ des d -vecteurs aléatoires sur $(\Omega, \mathcal{A}, \mathbf{P})$ indépendants et à valeurs dans \mathbb{R}^d . On définit la suite $\Theta_n, n \geq 1$ par

$$\Theta_{n+1} = \Theta_n - \gamma_{n+1}(h(\Theta_n) + \pi_{n+1})$$

où $h : \mathbb{R}^d \rightarrow \mathbb{R}^d$ est borélienne et $(\gamma_n, n \geq 1)$ est une suite de pas strictement positifs.

Soient les filtrations $\mathcal{F}_n = \sigma(\Theta_0, \pi_1, \dots, \pi_n), n \geq 0$ et $\mathcal{F}_n^\Theta = \sigma(\Theta_0, \dots, \Theta_n), n \geq 0$.

2.1 HEURISTIQUE

On peut voir la proposition

$$\Theta_n \rightarrow \{h = 0\}$$

comme $dist(\Theta_n, \{h = 0\}) \rightarrow 0$ ou bien $\exists \Theta^* : (\Omega, \mathcal{A}, \mathbf{P}) \rightarrow \{h = 0\}$ tel que $\Theta_n \rightarrow \Theta^*$ presque-sûrement et dans $L^p(\mathbf{P})$.

Tout d'abord, regardons le cas où $\{h = 0\}$ est fini ou localement fini (fini sur ses parties compactes).

Exemple typique : $H : \mathbb{R}^q \times \mathbb{R}^q \rightarrow \mathbb{R}^q$,

$$\theta_{n+1} = \theta_n - \gamma_{n+1}H(\theta_n, Z_{n+1})$$

où $(Z_n)_{n \geq 1}$ est *i.i.d.* et telle que $\forall \theta \in \mathbb{R}^q, \mathbf{E}[H(\theta, Z_1)] < \infty$. On pose

$$h(\theta) = \mathbf{E}[H(\theta, Z_1)]$$

telle qu'on puisse ensuite définir par récurrence

$$h(\Theta_n) = \mathbf{E}[H(\Theta_n, Z_{n+1}) | \mathcal{F}_n]$$

où $\mathcal{F}_n = \sigma(\Theta_0, Z_1, \dots, Z_n)$, ce qui donne

$$\Theta_{n+1} = \Theta_n - \gamma_{n+1}(h(\Theta_n) + \Delta M_{n+1})$$

où $\Delta M_{n+1} = H(\Theta_n, Z_{n+1}) - h(\Theta_n)$ et donc on a finalement $\pi_n = \Delta M_n$.

2.2 DESCENTE DE GRADIENT DÉTERMINISTE

Dans la suite $| \cdot |$ est une norme sur \mathbb{R}^q .

Proposition 2.1 (Descente de gradient local) *Pour $h : \mathbb{R}^q \rightarrow \mathbb{R}^q$ continue telle que :*

1. $\forall \theta \neq \theta^*, \langle h(\theta) | \theta - \theta^* \rangle > 0$
2. $\exists V^* \text{ voisinage de } \theta^* \text{ tel que } \forall \theta \in V^*, \langle h(\theta) | \theta - \theta^* \rangle \rightarrow C^* |h(\theta)|^2, \text{ où } C^* > 0$
3. $h(\theta^*) = 0$

Alors, $\exists \gamma_0 > 0$ tel que $\forall \gamma \in]0, \gamma_0]$ on ait presque-sûrement

$$\theta_{n+1} = \theta_n - \gamma h(\theta_n) \rightarrow \theta^*$$

Preuve. Soit $\alpha_0 = |\theta^* - \theta_0|$.

$$\begin{aligned} |\theta_{n+1} - \theta^*|^2 &= |\theta_n - \theta^*|^2 - 2\gamma \langle h(\theta_n) | \theta_n - \theta^* \rangle + \gamma^2 |h(\theta_n)|^2 \\ &= |\theta_n - \theta^*|^2 - \gamma |h(\theta_n)|^2 \left(\frac{2\langle h(\theta_n) | \theta_n - \theta^* \rangle}{|\theta_n - \theta^*|^2} - \gamma \right) \end{aligned}$$

Par définition ((i) et (ii)), on a

$$\gamma_0 := \inf \left\{ \frac{\langle h(\theta) | \theta - \theta^* \rangle}{|h(\theta)|^2}, |\theta - \theta^*| \leq \alpha_0 \right\} > 0$$

D'où pour $\gamma \in]0, \gamma_0]$,

$$|\theta_{n+1} - \theta^*|^2 = |\theta_n - \theta^*|^2 - \gamma(2\gamma_0 - \gamma) |h(\theta_n)|^2$$

On constate d'abord que $n \rightarrow |\theta_n - \theta^*|^2$ est décroissante et $\theta_n \in B(\theta_0, \alpha_0), \forall n \geq 0$ et $|\theta_n - \theta^*|^2 = |\theta_0 - \theta^*|^2 - \gamma(2\gamma_0 - \gamma) \sum_{k=1}^n |h(\theta_k)|^2$ ce qui implique que

$$\sum_{k=1}^n |h(\theta_k)|^2 \leq \frac{|\theta_n - \theta^*|^2}{\gamma(2\gamma_0 - \gamma)}, \quad \forall n \geq 1$$

□

On obtient grâce à cette proposition une deuxième version global qu'on démontre en exercice :

Proposition 2.2 (Descente de gradient global 1) *Pour $h : \mathbb{R}^q \rightarrow \mathbb{R}^q$ continue et telle que*

- (i) $\forall \theta \neq \theta^*, \langle h(\theta) | \theta - \theta^* \rangle > 0$
- (ii) $|h(\theta)| \leq C(1 + |\theta|)$ (*Contrôle asymptotique de h*)
- (iii) $\sum_k \gamma_k = \infty$ et $\sum_k \gamma_k^2 < \infty$ (*i.e. pas décroissant*)

Alors, $\theta_{n+1} = \theta_n - \gamma_{n+1} h(\theta_n)$ converge presque-sûrement vers $\theta^* = \{h = 0\}$ pour tout θ_0 .

Preuve. Exercice.

□

Remarque Si h est monotone sur \mathbb{R}^q , on a $\langle h(\theta) - h(\theta') \mid \theta - \theta' \rangle > 0 \forall \theta \neq \theta'$, en particulier $\langle h(\theta) \mid \theta - \theta^* \rangle > 0, \forall \theta \neq \theta^*$.

Proposition 2.3 (Descente de gradient global 2) Soit $V : \mathbb{R}^q \rightarrow \mathbb{R}_+$ continue et telle que ∇V soit lipschitzienne. On notera $[\nabla V]_{Lip}$ sa norme lipschitz. On suppose que $\lim_{|\theta| \rightarrow \infty} V(\theta) = \infty$ et $\{\nabla V = 0\} = \theta^*$. Soit $h : \mathbb{R}^q \rightarrow \mathbb{R}^q$ continue et telle que

$$(i) \quad \langle \nabla V(\theta) \mid h(\theta) \rangle > 0, \forall \theta \neq \theta^*$$

$$(ii) \quad |h(\theta)| \leq C(1 + V(\theta))^{1/2}$$

et $(\gamma_n, n \geq 1)$ suite de pas strictement positifs telle que

$$\sum_k \gamma_k = \infty \text{ et } \sum_k \gamma_k^2 < \infty$$

Alors,

$$1. \quad \arg \min V = \theta^*$$

$$2. \quad \theta_{n+1} = \theta_n - \gamma_{n+1} h(\theta_n) \rightarrow \theta^* \text{ presque-sûrement.}$$

Preuve. Tout d'abord, V atteint son minimum de façon locale, donc $\arg \min V \subset \{\nabla V = 0\} = \{\theta^*\}$. Ensuite,

$$\begin{aligned} V(\theta_{n+1}) &= V(\theta_n) + \langle \nabla V(\xi_n) \mid \Delta \theta_{n+1} \rangle \\ &= V(\theta_n) + \langle \nabla V(\theta_n) \mid \Delta \theta_{n+1} \rangle - \langle \nabla V(\theta_n) - \nabla V(\xi_n) \mid \Delta \theta_{n+1} \rangle \\ &\leq V(\theta_n) - \gamma_{n+1} \langle \nabla V \mid h \rangle(\theta_n) + [\nabla V]_{Lip} |\theta_{n+1} - \theta_n|^2 \\ &\leq V(\theta_n) - \gamma_{n+1} \langle \nabla V \mid h \rangle(\theta_n) + [\nabla V]_{Lip} \gamma_{n+1}^2 |h(\theta_n)|^2 \end{aligned}$$

et grâce à l'hypothèse (ii), on obtient

$$V(\theta_{n+1}) \leq V(\theta_n)(1 + [\nabla V]_{Lip} C^2 \gamma_{n+1}^2) - \gamma_{n+1} \langle \nabla V \mid h \rangle(\theta_n) + [\nabla V]_{Lip} C^2 \gamma_{n+1}^2$$

On ajoute une somme qui *a priori* sort de nulle part, qui se télescope pour retrouver l'inégalité précédente. On pose $\tilde{C} = [\nabla V]_{Lip} C^2$:

$$\begin{aligned} V(\theta_{n+1}) + \sum_{k=1}^{n+1} \gamma_k \langle \nabla V \mid h \rangle(\theta_{k-1}) + \tilde{C} \sum_{k \geq n+2} \gamma_k^2 &\leq V(\theta_n)(1 + \tilde{C} \gamma_{n+1}^2) + \tilde{C} \sum_{k \geq n+1} \gamma_k^2 + \sum_{k=1}^n \gamma_k \langle \nabla V \mid h \rangle(\theta_{k-1}) \\ &\leq (1 + \tilde{C} \gamma_{n+1}^2)(V(\theta_n) + \tilde{C} \sum_{k \geq n+1} \gamma_k^2 + \sum_{k=1}^n \gamma_k \langle \nabla V \mid h \rangle(\theta_{k-1})) \end{aligned}$$

Si on définit pour tout n , A_{n+1} comme étant le terme de gauche, on obtient

$$A_{n+1} \leq (1 + \tilde{C} \gamma_{n+1}^2) \times A_n$$

et donc

$$\tilde{A}_{n+1} := A_{n+1} \left[\prod_{k=1}^{n+1} (1 + \tilde{C} \gamma_k^2) \right]^{-1} \leq A_n \left[\prod_{k=1}^n (1 + \tilde{C} \gamma_k^2) \right]^{-1}$$

Comme pour tout $n \geq 1$, $\tilde{A}_n > 0$, on peut conclure que presque-sûrement

$$\tilde{A}_n \rightarrow \tilde{A}_\infty < \infty$$

or, $\Lambda_n := \prod_{k=1}^n (1 + \tilde{C}\gamma_k^2) = \exp(\sum_{k=1}^n \log(1 + \tilde{C}\gamma_k^2)) \rightarrow \Lambda_\infty < \infty$. Ce qui nous donne

$$A_n = \Lambda_n \tilde{A}_n \rightarrow \Lambda_\infty \tilde{A}_\infty = A_\infty$$

d'où

$$V(\theta_n) + \sum_{k=1}^n \gamma_k \langle \nabla V \mid h \rangle(\theta_n) + \tilde{C} \sum_{k \geq n+2} \gamma_k^2 \rightarrow A_\infty$$

donc $\sum_{n \geq 1} \gamma_n \langle \nabla V \mid h \rangle(\theta_{n-1}) < \infty$ presque-sûrement. On obtient enfin

$$V(\theta_n) \rightarrow V_\infty \text{ presque-sûrement}$$

Si $(V(\theta_n))_n$ est bornée dans $L^1(\mathbf{P})$ et $\lim_{|\theta| \rightarrow \infty} V(\theta) \rightarrow \infty$ alors $(|\theta_n|)_n$ bornée dans $L^1(\mathbf{P})$. Donc $\sum_n \gamma_n \langle \nabla V \mid h \rangle(\theta_n) < \infty$ et $\sum_n \gamma_n = \infty$ impliquent

$$\liminf \langle \nabla V \mid h \rangle(\theta_n) \rightarrow 0$$

Soit une suite extraite $(\theta'_n)_n$ telle que $\langle \nabla V \mid h \rangle(\theta'_n) \rightarrow 0$. On peut en extraire une sous-suite $(\theta''_n)_n$ telle que $\langle \nabla V \mid h \rangle(\theta''_n) \rightarrow 0$ et

$$\theta''_n \rightarrow \theta_\infty$$

Comme $\langle \nabla V \mid h \rangle$ continue, $\langle \nabla V \mid h \rangle(\theta_\infty) = 0$ et (i), on a forcément

$$\theta_\infty = \theta^*$$

et par continuité de V on obtient

$$\theta_n \rightarrow \theta^*$$

□

Définition 2.1 V est appelée **fonction de Lyapunov** pour h . De manière générale, une fonction de Lyapunov pour h $L : \mathbb{R}^d \rightarrow \mathbb{R}_+$ est telle que, pour toute solution $t \rightarrow x(t)$ de l'ODE définit par

$$\nabla y = -h(y)$$

on a $t \rightarrow L(x(t))$ décroissante. Si L est différentiable, c'est équivalent de dire que $\langle \nabla L \mid h \rangle \geq 0$ puisque

$$\frac{d}{dt} L(y(t)) = \langle \nabla L(y(t)) \mid \nabla y(t) \rangle = -\langle \nabla L \mid h \rangle(y(t))$$

Si un telle fonction de Lyapunov existe, le système est dit dissipatif.

Remarques Une descente de gradient c'est prendre $h = \nabla V$, où V vérifie les hypothèses de la proposition précédente.

Que se passe-t-il lorsque ∇V n'est pas lipschitz ? Il arrive qu'elle ne le soit pas, mais avec des conditions plus restrictives on peut s'en sortir : Si V est convexe, on peut trouver $h(\theta) = \rho(\theta)\nabla V(\theta)$ où $\rho(\theta)$ est proportionnel à la quantité $(1 + |\theta|)/(1 + \nabla V)$.

On peut aussi se passer de l'hypothèse de continuité pour h tant que $\langle \nabla V \mid h \rangle$ est semi-continue inférieurement.

2.3 PREMIER RETOUR AU STOCHASTIQUE

On veut $V : \mathbb{R}^d \rightarrow \mathbb{R}_+$ de la forme $V(\theta) = \mathbf{E}[v(\theta, Z)]$, où $Z : (\Omega, \mathcal{A}, \mathbf{P}) \rightarrow \mathbb{R}^q$ et $v : \mathbb{R}^q \times \mathbb{R}^q \rightarrow \mathbb{R}_+$ pour finalement trouver la quantité $\min_\theta V(\theta)$. On suppose que V est différentiable et

$$\nabla V(\theta) = \mathbf{E}[\partial_\theta v(\theta, Z)]$$

∂_θ est une pure notation qui peut être représentée de manières différentes, l'important étant de vérifier l'identité du paragraphe précédent.

Supposons que nous savons simuler Z à bas coût. On a par la loi forte des grands nombres, presque-sûrement

$$\nabla V(\Theta_n) := \mathbf{E}[\partial_\Theta v(\Theta_n, Z) | \mathcal{F}_n^\Theta] = \lim_m \frac{1}{m} \sum_{k=1}^m \partial_\theta v(\Theta_n, Z_k)$$

où $(Z_k)_k$ i.i.d. de même loi que Z et indépendants de $\Theta_1, \Theta_2, \dots$.

Théorème 2.1 (Robbins-Monroe) *La descente de gradient stochastique (SGD) correspond au cas "m = 1", c'est-à-dire*

$$\begin{aligned}\Theta_{n+1} &= \Theta_n - \gamma_{n+1} \partial_\theta v(\Theta_n, Z_{n+1}) \\ &= \Theta_n - \gamma_{n+1} (\nabla V(\Theta_n) + \Delta M_{n+1})\end{aligned}$$

avec $\Delta M_{n+1} = \partial_\Theta v(\Theta_n, Z_{n+1}) - \nabla V(\Theta_n)$.

2.4 PROBABILITÉS NUMÉRIQUES VERSUS DATASCIENCES

Les probabilités numériques permettent de simuler des variables aléatoires de lois à densité et régularisantes. Soit Z une variable gaussienne centrée réduite. Pour tout fonction bornée f ,

$$\begin{aligned}f_\varepsilon(x) &= \mathbf{E}[f(x + \varepsilon Z)] \\ &= (g_\varepsilon * f)(x) \rightarrow f(x)\end{aligned}$$

où $(g_\varepsilon)_{\varepsilon>0}$ est une approximation de l'unité gaussienne. On a donc pour nos fonctions précédentes

$$h(\theta) = \mathbf{E}[H(\theta, Z)] \text{ et } V(\theta) = \mathbf{E}[v(\theta, Z)]$$

Z ayant un effet régularisant, h et V sont plus régulières que

$$z \rightarrow H(\cdot, z) \text{ et } z \rightarrow v(\cdot, z)$$

Les algorithmes stochastique consistent à tirer les données au hasard. Soit $N > 0$ la taille des données et $(k_n)_n$ i.i.d. de loi uniforme sur $\{1, \dots, N\}$,

$$\Theta_{n+1} = \Theta_n - \gamma_{n+1} \partial_\theta v(\Theta_n, Z_{k_{n+1}})$$

correspond au SGD.

3 EXEMPLES D'ALGORITHMES STOCHASTIQUES

On désignera la transposée d'un vecteur a par a^* , on écrira le produit scalaire $\langle a | b \rangle = a^* b$.

Soient $\mathbf{X} = \{x_k \in \mathbb{R}^d, k = 1, \dots, N\}$ nos *input* et $\mathbf{Y} = \{y_k \in \mathbb{R}^q, k = 1, \dots, N\}$ nos *output*. - On pose

$$D = \{(x_k, y_k)\}_{k=1}^N \subset \mathbb{R}^d \times \mathbb{R}^q = S$$

notre ensemble de données.

Remarque Les x_k et les y_k sont des vecteurs colonnes.

3.1 SCIENCE DES DONNÉES

Apprentissage supervisé Ce qu'on appelle *apprentissage supervisé* est en fait un simple problème de minimisation. En effet on cherche le θ^* optimal pour le problème d'optimisation

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{N} \sum_{(x,y) \in D} |y - f(\theta, x)|^2$$

On peut y ajouter des poids, changer la fonction de coût, mais concrètement on reste avec le même problème d'optimisation. C'est en fait une simple régression linéaire (OLS, MCO). Dans la suite on prendra $f(\theta, x) = \theta^* x - w_0$ où w_0 représente le terme qui nous permet de contrôler le *biais* du modèle.

Apprentissage non supervisé On parle alors de *classification automatique*. En effet, dans ce cas nous sommes privé des *output* \mathbf{Y} . Soit $\theta = (\theta^1, \dots, \theta^r) \in (\mathbb{R}^q)^r$ notre *prototype*. Le problème d'optimisation est le suivant :

$$\min_{\theta \in (\mathbb{R}^q)^r} \frac{1}{N} \sum_{x \in \mathbf{X}} \min_{1 \leq i \leq r} |\theta^i - x|^2$$

Lemme 3.1 Si $\theta = (\theta^1, \dots, \theta^r)$ vérifie

$$\forall x \in \mathbf{X}, \# \{i : |x - \theta^i| = \min_{1 \leq j \leq r} |x - \theta^j|\} = 1$$

alors $V : (\mathbb{R}^q)^r \rightarrow \mathbb{R}_+$, $\theta \mapsto \frac{1}{N} \sum_{x \in \mathbf{X}} \min_{1 \leq i \leq r} |\theta^i - x|^2$ est différentiable, et

$$(\nabla V(\theta))_i = \frac{1}{N} \sum_{x \in \mathbf{X}} |\theta^i - x| \mathbf{1}_{|\theta^i - x| < \min_{j \neq i} |\theta^j - x|}, \quad i = 1, \dots, r$$

Preuve. Exercice.

□

Théorème 3.1 $\emptyset \neq \arg \min_{\theta} V(\theta) \subset \{\nabla V = 0\}$

Preuve. Exercice.

□

3.2 APPRENTISSAGE SUPERVISÉ : UN PAS PLUS LOIN

Dans la suite, on appliquera les résultats de la partie 1.2 pour

$$Z_n(\omega) = (x_n, y_n)$$

a la différence qu'on ne connaît pas la loi de Z_n : c'est bien ça que l'on cherche à estimer ! Soit donc Z un vecteur aléatoire suivant la loi dont nos données sont

la réalisation. On utilise donc des méthodes bayésiennes et empiriques pour traduire en termes probabilistes un problème d'optimisation. On cherche la meilleure approximation affine des sorties \mathbf{y} par les données initiales \mathbf{x} . Ce qui revient à chercher pour $\tilde{x} = \begin{bmatrix} 1 \\ x \end{bmatrix}$ et $f(w, x) = w^* \tilde{x} = \sum_{i=1}^d w_i x_i + w_0$,

$$\arg \min_{w \in \mathbb{R}^{d+1}} \sum_{(x,y) \in D} |y - f(w, x)|^2$$

La *fonction de coût local* sera définie pour $z = (x, y)$ par

$$v(w, z) = \frac{1}{2} |y - f(w, x)|^2$$

et la *fonction de coût global* par

$$\begin{aligned} V(w) &= \frac{1}{2N} \sum_{(x,y) \in D} |y - f(w, x)|^2 \\ &= \int_S v(w, z) \mu_N(dz) \\ &= \mathbf{E}_{Z \sim \mu_N}[v(w, Z)] \end{aligned}$$

où $\mu_N = \frac{1}{N} \sum_{z \in D} \delta_z$ est la mesure empirique issue de nos données. On a directement le vecteur

$$\begin{aligned} (\nabla V(w))_i &= \mathbf{E}_{Z \sim \mu_N}[\partial_{w_i} v(w, Z)] \\ &= \int_S (y - w^* \tilde{x}) \tilde{x}_i \mu_N(d\tilde{x}, dy) \end{aligned}$$

En remarquant que $(w^* \tilde{x}) \tilde{x} = (\tilde{x} \tilde{x}^*) w$ et $w \rightarrow V(w)$ est convexe,

$$\nabla V = 0 \iff \left[\int_S (\tilde{x} \tilde{x}^*) \mu_N(d\tilde{x}, dy) \right] w = \int_S (\tilde{x} y) \mu_N(d\tilde{x}, dy)$$

et donc le vecteur de poids optimal est

$$w_* = \left[\int_S (\tilde{x} \tilde{x}^*) \mu_N(d\tilde{x}, dy) \right]^{-1} \int_S (\tilde{x} y) \mu_N(d\tilde{x}, dy) \quad \nabla^2 V(w_*) = \int_S (\tilde{x} \tilde{x}^*) \mu_N(d\tilde{x}, dy) \geq \beta \text{Id}$$

Problème $\int_S (\tilde{x} \tilde{x}^*) \mu_N$ (qui est une matrice !) est-elle inversible ? Le calcul $[\int_S (\tilde{x} \tilde{x}^*) \mu_N]^{-1}$ étant en $O((d+1)^3)$ il est difficile de trouver une application : d'où l'idée de la récursivité !

La méthode de descente de gradient pour l'apprentissage supervisé s'écrit pour $i = 1, \dots, q$,

$$\begin{aligned}(w_{n+1})_i &= (w_n)_i - \gamma_{n+1}(\nabla V(w_n))_i \\ &= (w_n)_i - \gamma_{n+1} \mathbf{E}_{(x,y) \sim \mu_N}[(w_n^* x - y)x_i]\end{aligned}$$

et sa version stochastique pour $(k_n)_n$ i.i.d. de loi uniforme sur $\{1, \dots, N\}$,

$$\begin{aligned}(w_{n+1})_i &= (w_n)_i - \gamma_{n+1} \partial_{(w)_i} v(w_n, Z_{k_{n+1}}) \\ &= (w_n)_i - \gamma_{n+1} (w_n^* x^{(k_{n+1})} - y^{(k_{n+1})}) x^{(k_{n+1})}_i\end{aligned}$$

Exercices Faire la même chose avec une fonction d'activation Φ :

$$\begin{aligned}V(w) &= \frac{1}{2} \mathbf{E}_{(x,y) \sim \mu_N}[(\Phi(w^* \tilde{x}) - y)^2] ; \quad \nabla V(w) = \mathbf{E}_{(x,y) \sim \mu_N}[(\Phi(w^* \tilde{x}) - y)\Phi'(w^* \tilde{x})\tilde{x}] \\ \nabla^2 V(w) &= \mathbf{E}_{(x,y) \sim \mu_N}[(\Phi'(w^* \tilde{x})^2 + (\Phi(w^* \tilde{x}) - y))\Phi''(w^* \tilde{x})\tilde{x}\tilde{x}^*]\end{aligned}$$

Remarquer qu'on a besoin de $\Phi \in C(\mathbb{R}, [0, 1])$. On se retrouve cependant avec une fonction qui n'est plus convexe, on a donc plusieurs extrema.

3.3 DU PERCEPTRON MULTI-COUCHE VERS LE DEEP LEARNING

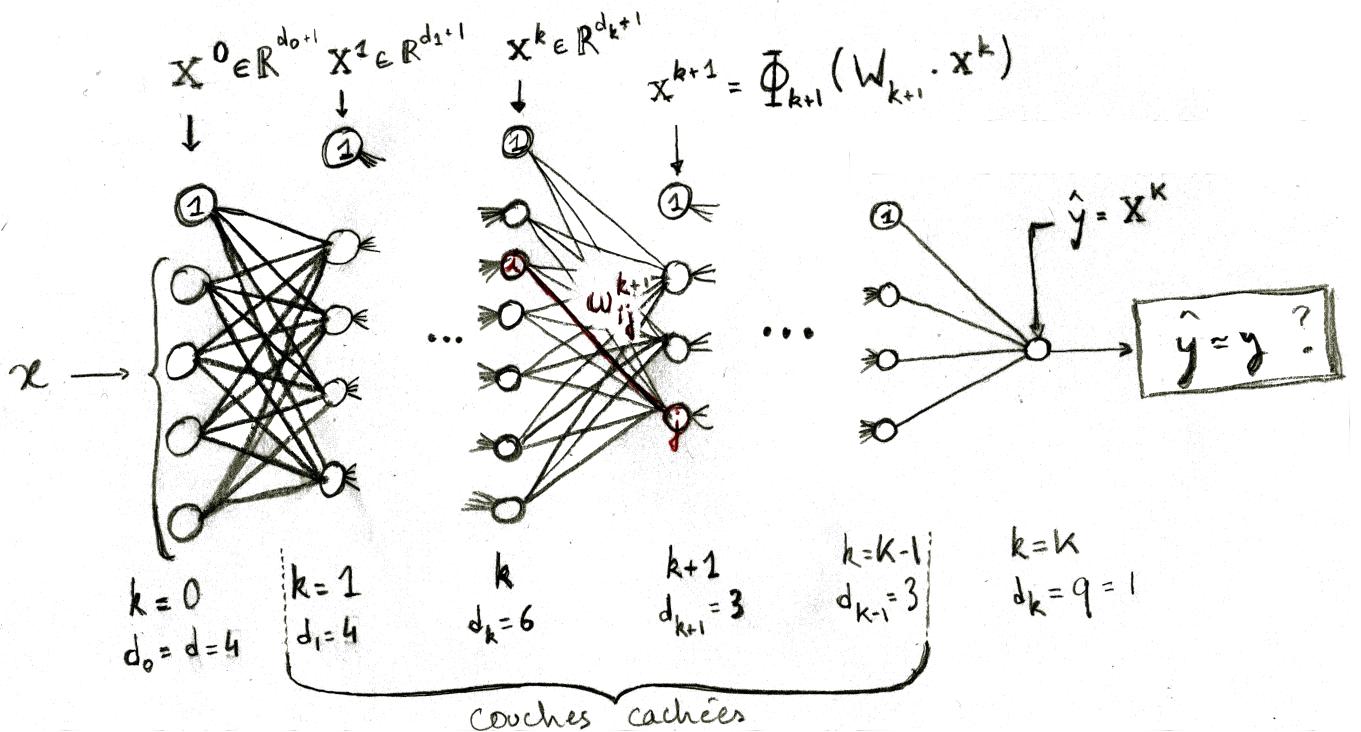


FIGURE 1 – Exemple de réseau de neurones à $K - 1$ couches cachées

Rétropropagation du gradient Dans la suite, on considère $K + 1$ couches de perceptron, d_0, \dots, d_K une suite d'entiers tels que $d_0 = d$ et $d_K = q$, K fonctions d'activation $\phi_k : \mathbb{R} \rightarrow [0, 1]$, $k = 1, \dots, K$. On considère

$$\forall k \in \{1, \dots, K\}, \Phi_k : z \in \mathbb{R}^{d_k} \rightarrow \begin{bmatrix} \phi_k(z_1) \\ \vdots \\ \phi_k(z_{d_k}) \end{bmatrix} \in \mathbb{R}^{d_k} \text{ et } x \in \mathbb{R}^{d_k} \rightarrow \tilde{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} \in \mathbb{R}^{d_{k+1}}$$

$x \rightarrow \tilde{x}$ provient du fait qu'on veut contrôler un éventuel biais w_0 dans nos paramètres estimés W_k , $k = 1, \dots, K$. On notera les vecteurs de l'algorithme en gras et les matrices de l'algorithme en majuscule pour plus de lisibilité. Soient pour n'importe quel j , une itération de l'apprentissage revient à l'algorithme

$$\begin{aligned}\mathbf{x}_0 &:= x_j \in \mathbb{R}^d \\ \mathbf{z}_0 &:= \tilde{\mathbf{x}}_0 \in \mathbb{R}^{d+1} \\ \mathbf{z}_k &:= W_{k+1} \tilde{\mathbf{x}}_k \in \mathbb{R}^{d_{k+1}} \\ \mathbf{x}_{k+1} &:= \Phi_{k+1}(\mathbf{z}_k) \in \mathbb{R}^{d_{k+1}}, \quad k = 1, \dots, K-1\end{aligned}$$

où $W_k \in (\mathbb{R}^{d_{k+1}})^{d_{k+1}}$ est la matrice définie dans la figure 1. Nos paramètres à optimiser sont les W_k , alors on s'intéresse à une fonction de coût qui est fonction de nos matrices W_k ! On cherche donc à minimiser

$$\mathcal{E}(W_1, \dots, W_K) = \frac{1}{2N} \sum_{j=1}^N |y^{(j)} - \mathbf{x}_K|_2^2$$

Remarque $\mathbf{x}_K(x^{(j)})$ est notre estimateur de y_j ! D'où le terme de *couches cachées* : on ne voit pas directement apparaître les matrices W_k dans le terme de droite. Avec un peu d'effort :

$$\begin{aligned}\mathcal{E}(W_1, \dots, W_K) &= \frac{1}{2} \mathbf{E}_{(x,y) \sim \mu_N} [|y - \mathbf{x}_K|_2^2] \\ &= \frac{1}{2} \int_S \mu_N(dx, dy) \sum_{l=1}^q (\phi_K(z_l^{K-1}) - y_l)^2\end{aligned}$$

d'où

$$\frac{\partial \mathcal{E}}{\partial w_{ji}^k} = \int_S \mu_N(dx, dy) \sum_{l=1}^q (\phi_K(z_l^{K-1}) - y_l) \phi'_K(z_l^{K-1}) \frac{\partial z_l^{K-1}}{\partial w_{ji}^k}$$

or

$$\frac{\partial z_l^{K-1}}{\partial w_{ji}^k} = \sum_{m=1}^{d_{K-1}} w_{lm}^K \frac{\partial x_m^{K-1}}{\partial w_{ji}^k}$$

Alors si on pose $e_l^K := (\phi_K(z_l^{K-1}) - y_l) \phi'_K(z_l^{K-1})$ l'erreur locale de la couche K , on obtient

$$\begin{aligned}\frac{\partial \mathcal{E}}{\partial w_{ji}^k} &= \int_S \mu_N(dx, dy) \sum_{m=1}^{d_{K-1}} \frac{\partial x_m^{K-1}}{\partial w_{ji}^k} \sum_{l=1}^q e_l^K w_{lm}^K \\ &= \int_S \mu_N(dx, dy) \sum_{m=1}^{d_{K-1}} \frac{\partial x_m^{K-1}}{\partial w_{ji}^k} \langle e^K | w_{\cdot m}^K \rangle\end{aligned}$$

3.4 APPROXIMATION UNIVERSELLE

Théorème 3.2 (K. Hornik [Hor91]) Si $\Phi : \mathbb{R} \rightarrow \mathbb{R}$ continue bornée et non constante, alors

$$\{x \rightarrow \sum_{i=1}^N \lambda_i \Phi(w^* x - w_0), N \geq 0, w \in \mathbb{R}^d, w_0 \in \mathbb{R}, \lambda_i \in \mathbb{R}\}$$

est dense dans $C([0, 1]^d, \mathbb{R})$.

Preuve. La démonstration est de Civenko, elle repose sur la théorie des distributions. Soit F l'ensemble en question. Si $F \subsetneq C([0, 1]^d, \mathbb{R})$, le théorème de Banach-Hahn nous donne une forme linéaire L bornée sur $C([0, 1]^d, \mathbb{R})$ tel que L est non nul et $L|_F = 0$.

RÉFÉRENCES

- [Hor91] HORNIK, Kurt : Approximation capabilities of multilayer feedforward networks. In : *Neural Networks* Volume 4, Issue 2 (1991), S. 251–257