

# Algorithmes stochastiques

## Introduction

A. Godichon-Baggioni

# I. Cadre

# CADRE

**Objectif :** minimiser la fonction  $G : \mathbb{R}^d \longrightarrow \mathbb{R}$  définie par

$$G(h) := \mathbb{E} [g(X, h)]$$

où  $X$  est une variable aléatoire à valeurs dans un espace  $\mathcal{X}$ .

**Cadre :** On supposera que  $G$  est convexe et on considèrera deux cas

- ▶  $G$  est fortement convexe
- ▶  $G$  est strictement convexe

# FONCTIONS FORTEMENT CONVEXES

**Moyenne d'une variable aléatoire :** Soit  $X$  un vecteur aléatoire de  $\mathbb{R}^d$ . Sa moyenne minimise la fonction  $G$  définie par

$$G(h) = \frac{1}{2} \mathbb{E} \left[ \|X - h\|^2 - \|X\|^2 \right].$$

**Modèle linéaire :** On considère le modèle

$$Y = \theta^T X + \epsilon$$

avec  $X, \epsilon$  des vecteurs aléatoires de  $\mathbb{R}^d$  indépendants et  $\mathbb{E}[\epsilon] = 0$ . Le paramètre  $\theta$  est un minimiseur de la fonction

$$G(h) = \frac{1}{2} \mathbb{E} \left[ (Y - h^T X)^2 \right].$$

# FONCTIONS STRICTEMENT CONVEXES

**Régression logistique :** Soit  $(X, Y)$  vérifiant

$$Y|X \sim \mathbb{B}(\pi(\theta^T X))$$

avec  $\pi(x) = \frac{\exp(x)}{1+\exp(x)}$ . Alors  $\theta$  est un minimiseur de la fonction

$$G(h) = \mathbb{E} [\log(1 + \exp(h^T X)) - h^T XY]$$

**Médiane d'une variable aléatoire :** Soit  $X$  à valeurs dans  $\mathbb{R}$ . Sa médiane minimise la fonction

$$G(h) = \mathbb{E} [|X - h|]$$

**Médiane géométrique :** Soit  $X$  à valeurs dans  $\mathbb{R}^d$ . Sa médiane géométrique est un minimum de la fonction

$$G(h) = \mathbb{E} [\|X - h\| - \|X\|]$$

## II. M-estimateurs

# DÉFINITION

Soient  $X_1, \dots, X_n$  des variables aléatoires i.i.d de même loi que  $X$ . On considère la fonction empirique

$$G_n(h) = \frac{1}{n} \sum_{k=1}^n g(X_k, h)$$

Un M-estimateur est un minimiseur  $\hat{m}_n$  de  $G_n$ .

## EXEMPLES

**Moyenne d'un vecteur aléatoire :** On a

$$G_n(h) = \frac{1}{2n} \sum_{i=1}^n \|X_i - h\|^2 - \|X_i\|^2$$

et on retrouve  $\hat{m}_n = \bar{X}_n$ .

**Modèle linéaire :** On a

$$G_n(h) = \frac{1}{2n} \sum_{i=1}^n (Y_i - X_i^T h)^2$$

on obtient l'estimateur des moindres carrés  $\hat{\theta}_n = (\mathbf{X}\mathbf{X}^T)^{-1} \mathbf{X}^T \mathbf{Y}$ .

**Médiane d'une variable aléatoire :** On a

$$G_n(h) = \frac{1}{n} \sum_{i=1}^n |X_i - h|$$

on retrouve la médiane empirique  $\hat{m}_n = X_{(\lceil \frac{n}{2} \rceil)}$ .



# "CONTRE-EXEMPLES"

**Régression logistique :** On a

$$G_n(h) = \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(h^T X_i)) - h^T X_i Y_i.$$

- Algorithme de gradient

**Médiane géométrique :** On a

$$G_n(h) = \frac{1}{n} \sum_{i=1}^n \|X_i - h\|$$

- Algorithme de Weiszfeld

# UN RÉSULTAT DE CONVERGENCE

## Théorème

*On suppose que les hypothèses suivantes sont vérifiées :*

- $\hat{m}_n$  converge en probabilité vers  $m$ .
- Pour presque tout  $x$ , la fonction  $g(x, \cdot)$  est deux fois continument différentiable.
- Pour presque tout  $x$ , la fonction  $\nabla_h^2 g(x, \cdot)$  est  $L(x)$ -lipschitz, i.e

$$\forall h, h', \quad \|\nabla^2 g(x, h) - \nabla^2 g(x, h')\|_{op} \leq L(x) \|h - h'\|.$$

- $L(X)$  et  $\nabla^2 g(X, m)$  admettent des moments d'ordre 2.
- $H := \nabla^2 G(m)$  est inversible.

Alors, en notant  $\Sigma = \mathbb{E} \left[ \nabla_h g(X, m) \nabla_h g(X, m)^T \right]$ , on a

$$\sqrt{n} (\hat{m}_n - m) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, H^{-1} \Sigma H^{-1}).$$

## ESTIMATION EN LIGNE

On considère des variables aléatoires  $X_1, \dots, X_n, X_{n+1}, \dots$  arrivant de manière séquentielle.

**Objectifs :** Mettre à jours les estimateurs avec le moins de calculs possibles.

**Estimation de la moyenne :** On a

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

**Estimateur sans biais de la variance :** On a

$$\begin{aligned}\Sigma_{n+1}^2 &= \Sigma_n^2 + \frac{1}{n+1} (X_{n+1} X_{n+1}^T - \Sigma_n^2) \\ S_{n+1}^2 &= \frac{n+1}{n} \Sigma_{n+1}^2 - \frac{n+1}{n} \bar{X}_{n+1} \bar{X}_{n+1}^T.\end{aligned}$$

# III. Algorithmes de gradient stochastiques

# ALGORITHME DE GRADIENT

On cherche à minimiser la fonction convexe  $G$  définie par

$$G(h) = \mathbb{E} [g(X, h)] .$$

L'algorithme de gradient est défini de manière itérative par

$$m_{t+1} = m_t - \gamma_t \nabla G(m_t)$$

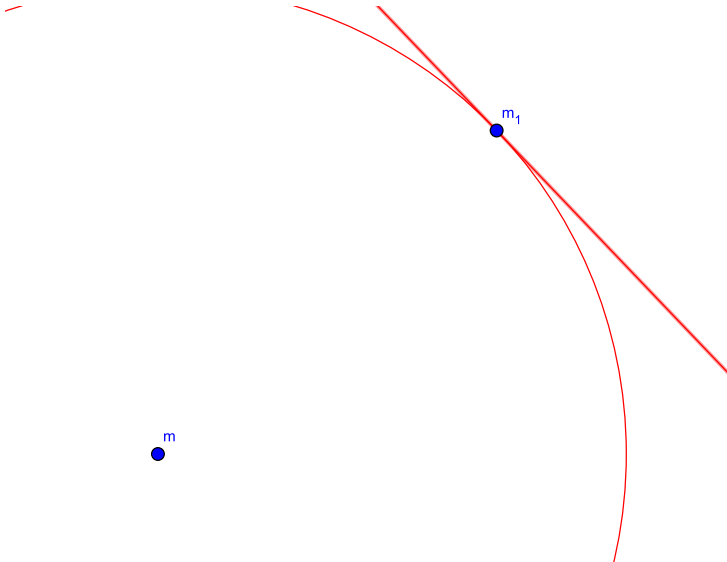
où  $\eta_t$  est une suite de pas positifs.

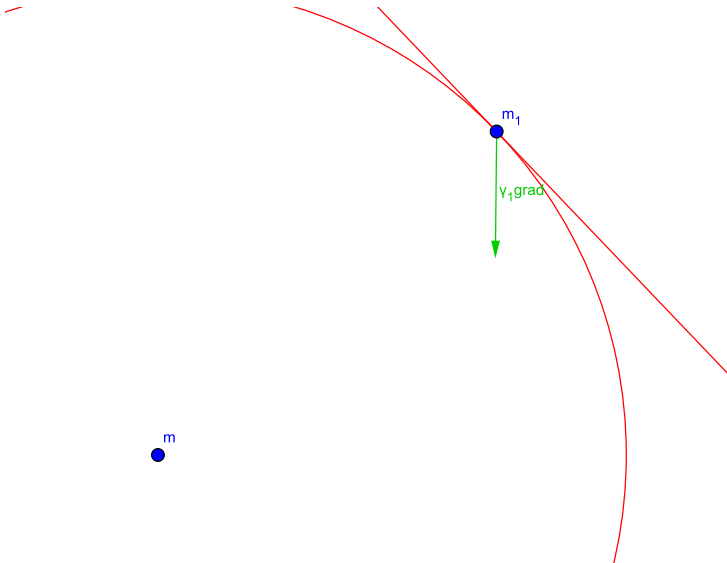
$m$



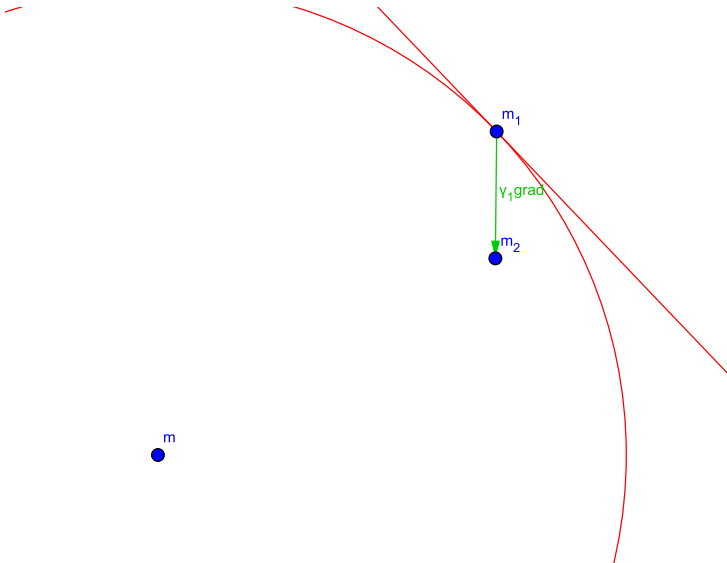
$m_1$

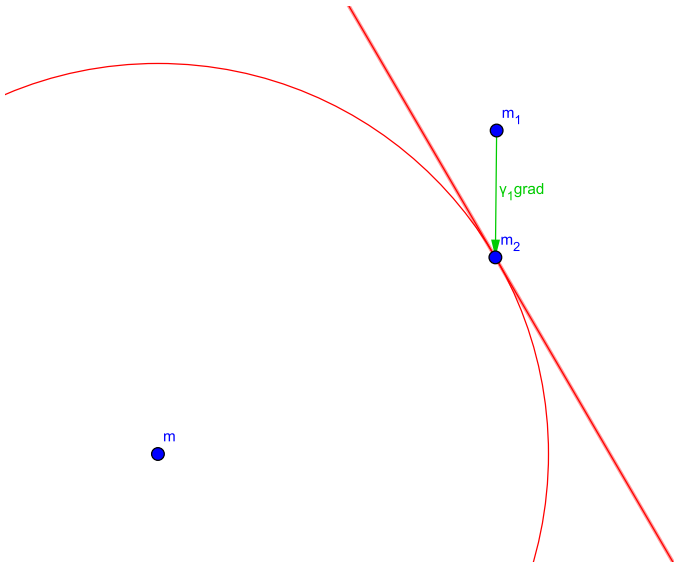


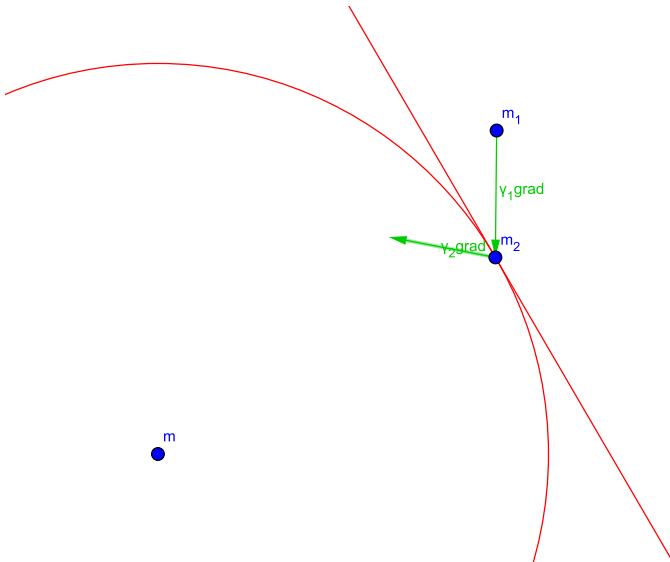


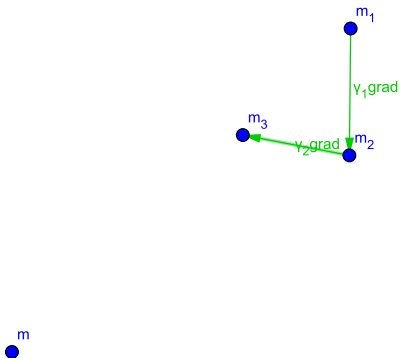


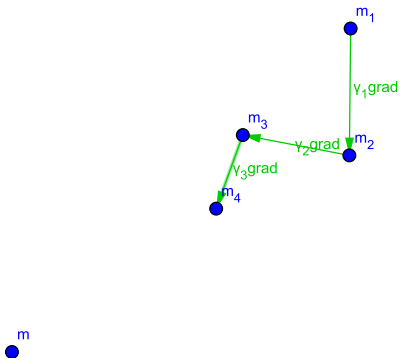


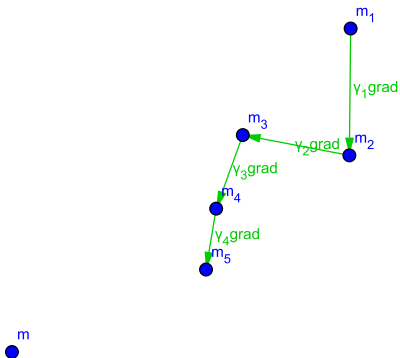


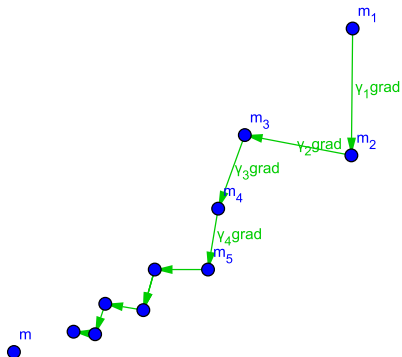












# ALGORITHMES DE GRADIENT STOCHASTIQUES

L'algorithme de gradient stochastique est défini de manière récursive pour tout  $n \geq 0$  par

$$m_{n+1} = m_n - \gamma_{n+1} \nabla_h g(X_{n+1}, m_n)$$

avec  $m_0$  borné et  $(\gamma_n)$  une suite de pas positifs vérifiant

$$\sum_{n \geq 1} \gamma_n = +\infty \quad \text{et} \quad \sum_{n \geq 1} \gamma_n^2 < +\infty.$$



# EXEMPLES

**Régression linéaire :** On a

$$\theta_{n+1} = \theta_n + \gamma_{n+1} (Y_{n+1} - \theta_n^T X_{n+1}) .$$

**Régression logistique :** On a

$$\theta_{n+1} = \theta_n - \gamma_{n+1} (\pi(\theta_n^T X_{n+1}) - Y_{n+1}) X_{n+1} .$$

**Médiane géométrique :** On a

$$m_{n+1} = m_n + \gamma_{n+1} \frac{X_{n+1} - m_n}{\|X_{n+1} - m_n\|} .$$

## APPROCHE NON ASYMPTOTIQUE

On prend  $\gamma_n = c_\gamma n^{-\alpha}$  avec  $c_\gamma > 0$  et  $\alpha \in (1/2, 1)$  et on suppose que les hypothèses suivantes sont vérifiées :

1. Il existe un minimiseur  $m$  de la fonction  $G$ .
2. La fonction  $G$  est  $\mu$ -fortement convexe : pour tout  $h \in \mathbb{R}^d$ ,

$$\langle \nabla G(h), h - m \rangle \geq \mu \|h - m\|^2.$$

**(PS0)** Il existe  $C \geq 0$  tel que pour tout  $h \in \mathbb{R}^d$ ,

$$\mathbb{E} \left[ \|\nabla_h g(X, h)\|^2 \right] \leq C \left( 1 + \|h - m\|^2 \right).$$

Alors, en notant  $C' = \max \{C, 2\mu^2\}$ , pour tout  $n \geq 1$ ,

$$\begin{aligned} \mathbb{E} \left[ \|m_n - m\|^2 \right] &\leq 2 \exp \left( C' c_\gamma^2 \frac{2\alpha}{2\alpha - 1} \right) \exp \left( -\frac{\mu c_\gamma}{4} n^{1-\alpha} \right) \left( \mathbb{E} \left[ \|m_0 - m\|^2 \right] + 1 \right) \\ &\quad + \frac{2c_\gamma C}{\mu n^\alpha} \end{aligned}$$

# PREUVE

La preuve repose sur le lemme suivant (admis) :

## Lemme

*Soit  $(\delta_n)$  une suite positive vérifiant*

$$\delta_{n+1} \leq (1 - 2\mu\gamma_{n+1} + 2L^2\gamma_{n+1}^2) \delta_n + 2\sigma^2\gamma_{n+1}^2$$

*avec  $\gamma_n = c_\gamma n^{-\alpha}$ ,  $c_\gamma, L \geq \mu > 0$ ,  $\sigma^2 \geq 0$  et  $\alpha \in (1/2, 1)$ . Alors pour tout  $n \geq 1$ ,*

$$\delta_n \leq 2 \exp\left(-\frac{\mu}{4}n^{1-\alpha}\right) \exp\left(2L^2c_\gamma^2 \frac{1-n^{1-2\alpha}}{2\alpha-1}\right) \left(\delta_0 + \frac{\sigma^2}{L^2}\right) + \frac{4c_\gamma\sigma^2}{\mu n^\alpha}$$

# UNE PROPOSITION UTILE

## Proposition

Soit  $\delta_n, \gamma_n$  deux suites positives telles que

- $\gamma_n = c_\gamma n^{-\alpha}$  avec  $c_\gamma > 0$  et  $\alpha \in (1/2, 1)$ .
- Il existe un rang  $n_0$ , une constante  $c_0 \in (0, \gamma_{n_0}^{-1})$  et une suite positive  $v_n$  telle que pour tout  $n \geq n_0$ ,

$$\delta_{n+1} \leq (1 - c_0 \gamma_{n+1}) \delta_n + \gamma_{n+1} v_{n+1}.$$

Alors pour tout  $n \geq 2n_0$ ,

$$\delta_n \leq \exp\left(-\frac{c_0 c_\gamma}{4} n^{1-\alpha}\right) \left( \delta_{n_0} + \sum_{k=n_0}^{n/2-1} \gamma_{k+1} v_{k+1} \right) + \max_{n/2 \leq k+1 \leq n-1} v_{k+1}$$

# POUR SIMPLIFIER

## Corollaire

*Soit  $\delta_n, \gamma_n$  deux suites positives telles que*

- *$\gamma_n = c_\gamma n^{-\alpha}$  avec  $c_\gamma > 0$  et  $\alpha \in (1/2, 1)$ .*
- *Il existe un rang  $n_0$ , une constante  $c_0 \in (0, \gamma_{n_0}^{-1})$  et une suite positive  $v_n$  telle que pour tout  $n \geq n_0$ ,*

$$\delta_{n+1} \leq (1 - c_0 \gamma_{n+1}) \delta_n + \gamma_{n+1} v_{n+1}.$$

*Si  $v_n = C_v (\ln n)^\beta n^v$  avec  $\beta \geq 0$  et  $v \in \mathbb{R}$ , alors*

$$\delta_n = O(v_n).$$

# UNE APPLICATION IMPORTANTE

## Lemme

*Soient  $A_n, B_n, r_n$  des suites de variables aléatoires positives telles que  $r_n$  converge presque sûrement vers 0 et*

$$A_{n+1} = (1 - c\gamma_{n+1}) A_n + \gamma_{n+1} r_n (A_n + B_n)$$

*avec  $\gamma_n = c_\gamma n^{-\alpha}$ . De plus, on suppose*

$$B_n = O(v_n) \quad p.s$$

*avec  $v_n = C_v n^v (\ln n)^\beta$  avec  $v \in \mathbb{R}$  et  $\beta \geq 0$ . Alors*

$$A_n = O(v_n) \quad p.s.$$

# APPLICATION À LA RÉGRESSION LOGISTIQUE

On suppose

- ▶  $\mathbb{E} [\epsilon^2] < +\infty$ .
- ▶  $\mathbb{E} [X^4] < +\infty$
- ▶ La matrice  $\mathbb{E} [XX^T]$  est définie positive et on note  $\mu$  sa plus petite valeurs propre.

On a alors

$$\mathbb{E} [\|\theta_n - \theta\|^2] \leq 2 \exp \left( C c_\gamma^2 \frac{1 - n^{2\alpha-1}}{2\alpha - 1} \right) \exp \left( -\frac{\mu c_\gamma}{4} n^{1-\alpha} \right) \left( \mathbb{E} [\|m_0 - m\|^2] + 1 \right) + \frac{2c_\gamma C}{\mu n^\alpha}$$

$$\text{avec } C = \max \left\{ 2\mathbb{E} [\|\epsilon\|^2] \mathbb{E} [\|X\|^2], 2\mathbb{E} [\|X\|^4] \right\}.$$

# APPLICATION À LA RÉGRESSION LINÉAIRE

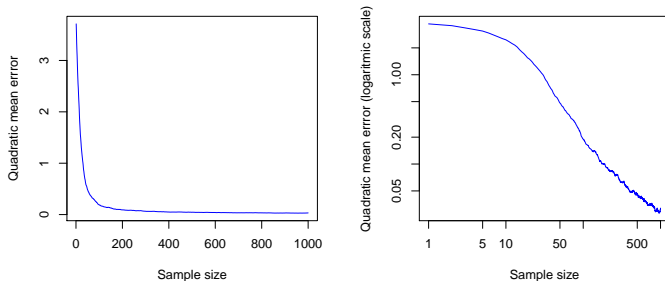


FIGURE – Evolution de l'erreur quadratique moyenne de  $\theta_n$  en fonction de la taille d'échantillon  $n$  dans le cadre de la régression linéaire.