Sorbonne Université Année 2022/2023

M2 Statistique Deuxième semestre

Algorithmes stochastiques

Antoine Godichon-Baggioni

Table des matières

1 Introduction **5**

1.1 Cadre . . . 5

1.1.1 Fonctions fortement convexes . . . 5

1.1.2 Fonctions strictement convexes . . . 6

1.2 M-estimateurs . . . 8

1.2.1 Définition et exemples . . . 8

1.2.2 Un résultat de convergence . . . 11

1.3 Estimation en ligne . . . 13

1.4 Algorithmes de gradient stochastiques . . . 14

1.4.1 Algorithmes de gradient stochastiques . . . 14

1.4.2 Approche non-asymptotique . . . 15

1.4.3 Exemple : regression lineaire . . . 18

2 Martingales **21**

2.1 Martingales réelles . . . 21

2.1.1 Définitions . . . 21

2.1.2 Théorème de Robbins-Siegmund . . . 22

2.1.3 Lois des grands nombres . . . 24

2.1.4 Théorème limite centrale . . . 28

2.2 Martingales vectorielles . . . 33

2.2.1 Définition . . . 33

2.2.2 Vitesses de convergence des martingales vectorielles . . . 34

2.2.3 Théorème Central Limite . . . 35

3 Vitesses de convergence des algorithmes de gradient stochastiques **37**

3.1 Convergence presque sûre . . . 37

3.1.1 Approche directe . . . 37

3.1.2 Approche via le développement de Taylor de la fonction G . . . 39

3.1.3 Approche Lyapunov . . . 40

3.1.4 Application au modèle linéaire . . . 41

3.1.5 Application à la régréssion logistique . . . 42

3.2 Vitesses de convergence presque sûre . . . 43
3.2.1 Cadre . . . 43
3.2.2 Vitesses de convergence . . . 44
3.2.3 Application au modèle linéaire . . . 47
3.2.4 Application à la régression logistique . . . 49
3.2.5 Remarques . . . 50

4 Accélération des méthodes de gradient stochastiques **53**
4.1 Algorithmes de gradient stochastiques moyennés . . . 53
4.1.1 Vitesse de convergence presque sûre . . . 54
4.1.2 Normalité asymptotique . . . 59
4.1.3 Application au modèle linéaire . . . 61
4.1.4 Application à la régression logistique . . . 64
4.1.5 Remarques . . . 67
4.2 Algorithme de Newton stochastique . . . 67
4.2.1 Idée de l'algorithme de Newton stochastique . . . 67
4.2.2 L'algorithme de Newton stochastique . . . 70
4.2.3 Vitesses de convergence . . . 73
4.2.4 Normalité asymptotique . . . 76
4.2.5 Application au modèle linéaire . . . 78
4.2.6 Application à la régression logistique . . . 83

Bibliographie **87**

Chapitre 1

Introduction

1.1 Cadre

Dans ce cours, on s'intéresse à l'estimation du minimiseur m d'une fonction convexe $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ définie pour tout $h \in \mathbb{R}^d$ par

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

avec $g : \mathcal{X} \times \mathbb{R}^d \longrightarrow \mathbb{R}$, et $\mathcal{X}$ un espace mesurable. On peut par exemple considérer $\mathcal{X} = \mathbb{R}$ ou $\mathcal{X} = \mathbb{R}^{d'}$. Dans ce cours, on se concentrera sur deux catégories de fonctions : fortement convexes et strictement convexes.

1.1.1 Fonctions fortement convexes

Moyenne d'une variable aléatoire : Soit X une variable aléatoire à valeurs dans $\mathbb{R}^d$. On peut voir la moyenne m de X comme le minimiseur de la fonction G définie pour tout $h \in \mathbb{R}^d$ par

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

A noter que le terme $\|X\|^2$ dans la définition de la fonction G permet de ne pas avoir à faire d'hypothèse sur l'existence du moment d'ordre 2 de X. De plus, on a

$$\nabla G(h) = -\mathbb{E}\left[X - h\right]$$

et donc $m = \mathbb{E}[X]$ est l'unique zéro du gradient de la fonction G. Enfin, on a

$$\nabla^2 G(h) = I_d$$

et la Hessienne est donc définie positive et uniformément minorée sur $\mathbb{R}^d$. Ainsi, la fonction G est fortement convexe et la moyenne m est bien son unique minimiseur.

Régression linéaire : Soit (X, Y) un couple de variables aléatoires à valeurs dans $\mathcal{X} = \mathbb{R}^d \times \mathbb{R}$ tel

que

$$Y = \theta^T X + \epsilon, \tag{1.1}$$

où $\theta \in \mathbb{R}^d$ et ϵ est une variable aléatoire indépendante de X vérifiant $\mathbb{E}[\epsilon] = 0$. On suppose que X et ϵ admettent des moments d'ordre 2 et que la matrice $\mathbb{E}\left[XX^T\right]$ est définie positive. Alors le paramètre θ est l'unique minimiseur de la fonction $G : \mathbb{R}^d \longrightarrow \mathbb{R}_+$ définie pour tout $h \in \mathbb{R}^d$ par

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

En effet, comme X et ϵ admettent un moment d'ordre 2, la fonction G est différentiable et

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

En particulier, on a

$$\nabla G(\theta) = -\mathbb{E}\left[\left(Y - \theta^T X\right) X\right] = -\mathbb{E}\left[\epsilon X\right] = -\mathbb{E}\left[\mathbb{E}\left[\epsilon | X\right] X\right] = 0.$$

De plus, comme X admet un moment d'ordre 2, la fonction G est deux fois différentiable et pour tout $h \in \mathbb{R}^d$,

$$\nabla^2 G(h) = \mathbb{E}\left[XX^T\right].$$

Cette matrice étant (supposée) définie positive, la Hessienne est uniformément minorée sur $\mathbb{R}^d$ et la fonction G est donc fortement convexe, ce qui fait de θ son unique minimiseur.

Remarque 1.1.1. *Bien que la matrice XX^T soit au plus de rang 1, la matrice $\mathbb{E}\left[XX^T\right]$ peut être définie positive. En effet, si considère un vecteur aléatoire gaussien $Z \sim \mathcal{N}(0, I_d)$, sa matrice de variance-covariance*

$$\mathbb{E}\left[ZZ^T\right] = Var\left[Z\right] = I_d$$

est définie positive.

1.1.2 Fonctions strictement convexes

Régression logistique : On considère un couple de variables aléatoires (X, Y) à valeurs dans $\mathbb{R}^d \times \{0, 1\}$ tel que

$$\mathcal{L}(Y|X) = \mathcal{B}\left(\pi\left(\theta^T X\right)\right), \tag{1.2}$$

avec $\pi(x) = \frac{\exp(x)}{1+\exp(x)}$. On peut voir le paramètre θ comme un minimiseur de la fonction $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ définie pour tout $h \in \mathbb{R}^d$ par

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

En effet, si la variable aléatoire X admet un moment d'ordre 2, la fonction G est différentiable et pour tout $h \in \mathbb{R}^d$,

$$\nabla G(h) = \mathbb{E}\left[\frac{\exp\left(h^T X\right)}{1+\exp\left(h^T X\right)} X - XY\right] = \mathbb{E}\left[\pi\left(h^T X\right) X - XY\right].$$

En particulier, comme $\mathbb{E}[Y|X] = \pi\left(\theta^T X\right)$, on a

$$\nabla G(\theta) = \mathbb{E}\left[\left(\pi\left(\theta^T X\right) - Y\right) X\right] = \mathbb{E}\left[\mathbb{E}\left[\left(\pi\left(\theta^T X\right) - Y\right) X|X\right]\right] = \mathbb{E}\left[\left(\pi\left(\theta^T X\right) - \mathbb{E}[Y|X]\right) X\right] = 0.$$

De plus, comme X admet un moment d'ordre 2, la fonction G est deux fois différentiable et

$$\nabla^2 G(h) = \mathbb{E}\left[\pi\left(h^T X\right)\left(1 - \pi\left(h^T X\right)\right) XX^T\right]$$

qui est au moins semi-définie positive. On supposera par la suite que la Hessienne en θ est définie positive, et donc que la fonction G est strictement convexe et que θ est son unique minimiseur.

Médiane d'une variable aléatoire : Soit $X \in \mathbb{R}$ une variable aléatoire et on suppose que sa fonction de répartition est strictement croissante et continue au voisinage de sa médiane notée m. Celle ci est alors le minimiseur de la fonction $G : \mathbb{R} \longrightarrow \mathbb{R}$ définie pour tout $h \in \mathbb{R}$ par

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

En effet, on rappelle que pour toute variable aléatoire X, grâce au théorème de Fubini-Tonelli,

$$\int_0^{+\infty} \mathbb{P}[Z \geq t]\, dt = \int_0^{+\infty} \mathbb{E}[\mathbf{1}_{Z \geq t}]\, dt = \mathbb{E}\left[\int_0^{+\infty} \mathbf{1}_{t \leq Z} dt\right] = \mathbb{E}\left[\int_0^{Z} 1 dt\right] = \mathbb{E}[Z].$$

Ainsi, on peut réécrire, pour tout $h \in \mathbb{R}$

$$\begin{aligned} G(h) = \int_0^{+\infty} \mathbb{P}[|X - h| \geq t]\, dt &= \int_0^{+\infty} \mathbb{P}[X \geq t + h]\, dt + \int_0^{+\infty} \mathbb{P}[X \leq h - t]\, dt \\ &= \int_h^{+\infty} (1 - F(t))\, dt + \int_{-\infty}^{h} F(t) dt \end{aligned}$$

Ainsi, pour tout h, on a $G'(h) = 2F(h) - 1$ et donc m est bien l'unique minimiseur de la fonction

G. A noter que pour ne pas avoir à faire d'hypothèse sur l'existence du moment d'ordre 1 de $|X|$, on peut réécrire la fonction G comme

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

Médiane géométrique : On considère une variable aléatoire X à valeurs dans $\mathbb{R}^d$. La médiane

géométrique de X est un minimum de la fonction $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ définie pour tout $h \in \mathbb{R}^d$ par

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

où $\|.\|$ est la norme euclidienne de $\mathbb{R}^d$. A noter que la fonction G est différentiable avec pour tout $h \in \mathbb{R}^d$,

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

Sous certaines hypothèses, la fonction G est deux fois continûment différentiable avec pour tout $h \in \mathbb{R}^d$,

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

La Hessienne de G est donc au moins semi-définie positive. On supposera par la suite qu'elle est strictement positive en m, et donc que la fonction G est strictement convexe.

1.2 M-estimateurs

1.2.1 Définition et exemples

On rappelle que l'objectif est d'estimer le minimisieur m de la fonction G définie pour tout $h \in \mathbb{R}^d$ par

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

Comme on ne sait généralement pas calculer explicitement G (ou son gradient), on ne peut généralement pas calculer (ou approcher) directement la solution. Pour pallier ce problème, on considère maintenant des variables aléatoires indépendantes et identiquement distribuées $X_1, \ldots, X_n$ de même loi que X. Une possibilité pour estimer m est alors de considérer la fonction empirique G_n définie pour tout $h \in \mathbb{R}^d$ par

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

A noter que par la loi des grands nombres, si $g(X, h)$ admet un moment d'ordre 1 (ce qui est la condition sine qua none pour que la fonction G soit bien définie en h),

$$G_n(h) \xrightarrow[n\to+\infty]{p.s} G(h).$$

Un M-estimateur de m est un minimiseur $\hat{m}_n$ de la fonction empirique G_n. A noter que $\hat{m}_n$ n'est pas toujours explicite, mais il existe tout de même quelques exemples où on sait le calculer.

Exemple 1 : estimation de la moyenne. Dans le cas de l'estimation de la moyenne, la fonction

empirique est définie pour tout $h \in \mathbb{R}^d$ par

$$G_n(h) = \frac{1}{2n} \sum_{k=1}^{n} \|X_k - h\|^2 - \|X_k\|^2 ,$$

et on obtient donc $\hat{m}_n = \overline{X}_n$. A noter que si X admet un moment d'ordre 2, le TLC nous donne

$$\sqrt{n} \left(\overline{X}_n - m \right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N} \left(0, \mathrm{Var}(X) \right).$$

Remarque 1.2.1. *En notant que* $H := \nabla^2 G(m) = I_d$ *et que*

$$\Sigma := \mathbb{E} \left[\nabla_h g \left(X, m \right) \nabla_h g \left(X, m \right)^T \right] = \mathbb{E} \left[(X - m)(X - m)^T \right] = Var(X),$$

où $\nabla_h g(X, h)$ *est le gradient de* g *par rapport à la deuxième variable, on peut réécrire le TLC précédent comme*

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

Exemple 2 : la régression linéaire. Dans le cas de l'estimation du paramètre de la régression linéaire, on a

$$G_n(h) = \frac{1}{2n} \sum_{k=1}^{n} \left(X_k^T h - Y_k \right)^2 .$$

Si la matrice $\mathbf{X} = \left(X_1, \dots X_n \right)^T$ est de rang plein, on rappelle que le minimiseur de la fonction G_n est l'estimateur des moindres carrés défini par

$$\hat{\theta}_n = \left(\mathbf{X} \mathbf{X}^T \right)^{-1} \mathbf{X}^T \mathbf{Y}$$

avec $\mathbf{Y} = \left(Y_1, \dots, Y_n \right)^T$.

Remarque 1.2.2. *Sous certaines hypothèses, on peut montrer que si la matrice Hessienne* $H := \mathbb{E} \left[X X^T \right]$ *est définie positive (et donc inversible), on a la normalité asymptotique*

$$\sqrt{n} \left(\hat{\theta}_n - \theta \right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N} \left(0, Var \left[\epsilon \right] H^{-1} \right).$$

De plus, en remarquant que

$$\begin{aligned} \Sigma := \mathbb{E} \left[\nabla_h g \left(X, \theta \right) \nabla_h g \left(X, \theta \right)^T \right] &= \mathbb{E} \left[\left(Y - X^T \theta \right) X X^T \left(Y - X^T \theta \right) \right] \\ &= \mathbb{E} \left[\epsilon^2 X X^T \right] \\ &= \mathbb{E} \left[\mathbb{E} \left[\epsilon^2 | X \right] X X^T \right] \\ &= Var \left[\epsilon \right] H, \end{aligned}$$

on peut réécrire la normalité asymptotique comme

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

Exemple 3 : médiane d'une variable aléatoire réelle. Dans le cas de l'estimation de la médiane d'une variable aléatoire X, la fonction empirique s'écrit

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

et on rappelle qu'un minimiseur de G_n est la médiane empirique $\hat{m}_n = X_{\left(\lceil \frac{n}{2} \rceil\right)}$.

Cependant, dans une grande majorité des cas tels que l'estimation de la médiane géométrique ou l'estimation des paramètres de la régression logistique, on ne sait pas calculer explicitement le minimum de la fonction G_n. On peut néanmoins utiliser les méthodes d'optimisation déterministes usuelles pour approcher $\hat{m}_n$.

Exemple 1 : la régression logistique. Dans le cadre de la régression logistique, on a

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

et il n'existe pas de solution explicite. Cependant, on peut, par exemple, utiliser un algorithme de gradient pour approcher $\hat{m}_n$. Celui-ci est défini de manière itérative pour tout $t \geq 0$ par

$$m_{n,t+1} = m_{n,t} - \eta_t \nabla G_n\left(m_{n,t}\right)$$

où η_t est une suite de pas positifs (cf cours d'optimisation pour plus de précisions).

Exemple 2 : estimation de la médiane géométrique. Dans le cas de la médiane géométrique, la fonction empirique est définie pour tout $h \in \mathbb{R}^d$ par

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

et un minimiseur de G_n est donc un zéro du gradient de G_n, i.e

$$0 = -\frac{1}{n}\sum_{k=1}^{n} \frac{X_k - \hat{m}_n}{\|X_k - \hat{m}_n\|}.$$

Là encore, il n'existe pas de solution explicite, mais on peut réécrire l'égalité précédente comme

$$\hat{m}_n = \frac{\sum_{k=1}^{n} \frac{X_k}{\|X_k - \hat{m}_n\|}}{\sum_{k=1}^{n} \frac{1}{\|X_k - \hat{m}_n\|}}$$

et on peut donc approcher $\hat{m}_n$ à l'aide d'un algorithme de point fixe, conduisant à l'algorithme

itératif suivant (algorithme de Weiszfeld [Wei37])

$$m_{n,t+1} = \frac{\sum_{k=1}^{n} \frac{X_k}{\|X_k - m_{n,t}\|}}{\sum_{k=1}^{n} \frac{1}{\|X_k - m_{n,t}\|}}$$

Remarque 1.2.3. *A noter que l'on peut réécrire l'algorithme de Weiszfeld comme*

$$m_{n,t+1} = m_{n,t} + \frac{n}{\sum_{k=1}^{n} \frac{1}{\|X_k - m_{n,t}\|}} \frac{1}{n} \sum_{k=1}^{n} \frac{X_k - m_{n,t}}{\|X_k - m_{n,t}\|} = m_{n,t} - \frac{n}{\sum_{k=1}^{n} \frac{1}{\|X_k - m_{n,t}\|}} \nabla G_n \left(m_{n,t}\right)$$

et l'on peut donc voir l'algorithme de Weiszfeld comme un algorithme de gradient avec

$$\eta_t = \frac{n}{\sum_{k=1}^{n} \frac{1}{\|X_k - m_{n,t}\|}}.$$

1.2.2 Un résultat de convergence

Le théorème suivant généralise les résultats donnés dans les remarques 1.2.1 et 1.2.2. A noter que les hypothèses présentées ne sont pas minimales mais rendent la preuve plus accessible.

Théorème 1.2.1. *On suppose que les hypothèses suivantes sont vérifiées :*

- *$\hat{m}_n$ converge en probabilité vers m.*
- *Pour presque tout x, la fonction $g(x,.)$ est deux fois continûment différentiable.*
- *Pour presque tout x, la Hessienne $\nabla_h^2 g(x,.)$ est $L(x)$-lipschitz, i.e pour tout $h, h' \in \mathbb{R}^d$,*

$$\left\|\nabla_h^2 g\left(x,h\right) - \nabla_h^2 g\left(x,h'\right)\right\|_{op} \leq L(x) \left\|h - h'\right\|$$

  où $\|.\|_{op}$ est la norme spectrale.
- *$L(X)$ et $\nabla_h^2 g(X,m)$ admettent des moments d'ordre 1.*
- *La Hessienne de G en m est inversible.*

Alors

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

avec

$$H = \nabla^2 G(m) \qquad et \qquad \Sigma = \mathbb{E}\left[\nabla_h g\left(X,m\right) \nabla_h g\left(X,m\right)^T\right].$$

Démonstration. Comme $\hat{m}_n$ est un minimiseur local de G_n, on a

$$0 = \frac{1}{n} \sum_{k=1}^{n} \nabla_h g\left(X_k, \hat{m}_n\right).$$

Comme pour presque tout x, la fonction $g(x,.)$ est deux fois continûment différentiable, à l'aide

d'un développement de Taylor, on a presque sûrement

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{k=1}^{n} \nabla_h g\left(X_k, \hat{m}_n\right) - \frac{1}{n}\sum_{k=1}^{n} \nabla_h g\left(X_k, m\right) + \frac{1}{n}\sum_{k=1}^{n} \nabla_h g\left(X_k, m\right) \\ &= \frac{1}{n}\sum_{k=1}^{n} \nabla_h g\left(X_k, m\right) + \frac{1}{n}\sum_{k=1}^{n} \int_0^1 \nabla_h^2 g\left(X_k, m + t\left(\hat{m}_n - m\right)\right)\left(\hat{m}_n - m\right) dt \end{aligned}$$

On peut donc réécrire l'égalité précédente comme

$$-\frac{1}{n}\sum_{k=1}^{n} \nabla_h g\left(X_k, m\right) = H_n\left(\hat{m}_n - m\right)$$

avec

$$\begin{aligned} H_n &= \frac{1}{n}\sum_{k=1}^{n} \int_0^1 \nabla_h^2 g\left(X_k, m + t\left(\hat{m}_n - m\right)\right) dt \\ &= \frac{1}{n}\sum_{k=1}^{n} \nabla_h^2 g\left(X_k, m\right) + \int_0^1 \left(\nabla_h^2 g\left(X_k, m + t\left(\hat{m}_n - m\right)\right) - \nabla_h^2 g\left(X_k, m\right)\right) dt \end{aligned}$$

Comme $\nabla_h g\left(X, m\right)$ admet un moment d'ordre 2 et comme $\mathbb{E}\left[\nabla_h g\left(X, m\right)\right] = \nabla G(m) = 0$, on obtient à l'aide d'un TLC

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} \nabla_h g\left(X_k, m\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \Sigma\right)$$

avec $\Sigma := \mathrm{Var}\left[\nabla_h g(X, m)\right] = \mathbb{E}\left[\nabla_h g\left(X_k, m\right) \nabla_h g\left(X_k, m\right)^T\right]$, et donc

$$\sqrt{n} H_n\left(\hat{m}_n - m\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \Sigma\right).$$

De plus, comme $\hat{m}_n$ converge en probabilité vers m, on a

$$\begin{aligned} &\left\| \frac{1}{n}\sum_{k=1}^{n} \int_0^1 \left(\nabla_h^2 g\left(X_k, m + t\left(\hat{m}_n - m\right)\right) - \nabla_h^2 g\left(X_k, m\right)\right) dt \right\|_{op} \\ &\leq \frac{1}{n}\sum_{k=1}^{n} \int_0^1 \left\| \left(\nabla_h^2 g\left(X_k, m + t\left(\hat{m}_n - m\right)\right) - \nabla_h^2 g\left(X_k, m\right)\right) \right\|_{op} dt \leq \frac{1}{2}\left\|\hat{m}_n - m\right\| \frac{1}{n}\sum_{k=1}^{n} L\left(X_k\right) \xrightarrow[n\to+\infty]{\mathbb{P}} 0. \end{aligned}$$

De plus, par la loi des grands nombres,

$$\frac{1}{n}\sum_{k=1}^{n} \nabla_h^2 g\left(X_k, m\right) \xrightarrow[n\to+\infty]{\mathbb{P}} \nabla^2 G(m),$$

et donc H_n converge en probabilité vers $H = \nabla^2 G(m)$. Comme H est positive, l'application $M \longmapsto M^{-1}$ est continue en H, et par continuité, H_n^{-1} converge en probabilité vers H^{-1} (même si H_n^{-1}

n'est pas nécessairement définie). Ainsi, par le lemme de Slutsky,

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

□

Ainsi, sous une certaine forme de régularité du modèle, on a une certaine forme d'efficacité asymptotique, i.e un résultat optimal "bornant" les résultats possibles pour les estimateurs. De plus, lorsque l'on ne sait pas calculer explicitement $\hat{m}_n$ et que l'on doit donc l'approcher via $m_{n,t}$, si $m_{n,t}$ converge vers $\hat{m}_n$, on obtient la convergence en loi

$$\lim_{n\to+\infty}\lim_{t\to+\infty}\sqrt{n}\left(m_{n,t} - m\right) = \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right)$$

Ainsi les méthodes itératives sont particulièrement efficaces, mais sous une certaine forme de régularité du modèle, on ne pourra pas avoir de meilleurs résultats que la normalité asymptotique, problème que l'on rencontrera quel que soit le type d'estimateur choisi. Cependant, pour les méthodes itératives, un gros problème peut être rencontré si l'on doit traiter des données en ligne : si on a déjà traité 1000 données (avec le temps de calcul que cela implique), que devons-nous faire si 1000 nouvelles données arrivent ? Une option serait de tout reprendre depuis le début, ce qui nécessite fatalement plus de calculs. S'intéresser aux estimateurs en ligne permet entre autre de régler ce problème.

1.3 Estimation en ligne

Dans ce qui suit, on considère des variables aléatoires $X_1, \ldots, X_n, X_{n+1}, \ldots$ arrivant de manière séquentielle. L'objectif est de mettre en place des algorithmes permettant de mettre facilement à jours les estimateurs. Un exemple simple est celui de la moyenne. Considérant que l'on a calculé $\overline{X}_n$ et qu'une nouvelle donnée arrive, comment obtenir $\overline{X}_{n+1}$ en faisant le moins de calculs possibles ? Une version brutale serait de repasser sur toutes les données afin de calculer $\overline{X}_{n+1}$, ce qui représenterait $O((n+1)d)$ nouvelles opérations. Une version plus subtile est de calculer $\overline{X}_{n+1}$ comme suit :

$$\overline{X}_{n+1} = \overline{X}_n + \frac{1}{n+1}\left(X_{n+1} - \overline{X}_n\right),$$

ce qui, quand n est grand, représente bien évidemment beaucoup moins de calculs ($O(d)$). Pour l'estimateur de la variance S_n^2, il faut se casser un peu plus la tête. Il faut d'abord remarquer que celui-ci peut s'écrire

$$S_n^2 = \frac{n}{n-1}\Sigma_n^2 - \frac{n}{n-1}\overline{X}_n\overline{X}_n^T \qquad \text{et} \qquad \Sigma_n^2 = \frac{1}{n}\sum_{k=1}^{n} X_k X_k^T$$

Ainsi, on peut construire S_n^2 de manière récursive comme suit

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

A noter qu'en plus du fait que le temps de calcul pour mettre à jours les estimateurs est réduit, l'estimation en ligne permet de ne pas avoir à garder en mémoire toutes les données, i.e on peut "jeter" les données dès qu'elles ont été traitées, ce qui peut permettre de réduire les coûts en terme de mémoire. Même si dans de nombreux cas, il n'est pas possible ou évident de construire des estimateurs en ligne, dans le cas de la minimisation de la fonction G, on peut souvent construire ce type d'estimateurs à l'aide d'algorithmes de gradient stochastiques.

1.4 Algorithmes de gradient stochastiques

On rappelle que l'on s'intéresse à l'estimation du minimiseur de la fonction $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ définie pour tout $h \in \mathbb{R}^d$ par

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

et on considère des variables aléatoires indépendantes et identiquement distribuées $X_1, \ldots, X_n, X_{n+1}, \ldots$ de même loi que X et arrivant de manière séquentielle. On rappelle qu'un algorithme de gradient pour approcher m aurait été de la forme

$$m_{t+1} = m_t - \eta_t \nabla G\left(m_t\right).$$

Cependant, ici, on n'a pas accès au gradient de G (car sous forme d'espérance). Pour régler ce problème, on a vu précédemment que l'on peut remplacer ∇G par le gradient de la fonction empirique, ce qui représente $O(nd)$ opérations à chaque itération, et en notant T ce nombres d'itérations, on arrive donc à $O(ndT)$ opérations. De plus, on a vu que ces méthodes ne permettent pas de traiter les données en ligne. Pour pallier ces problèmes, on s'intéresse à l'algorithme de gradient stochastique, qui permet de traiter les données en ligne, et ce, avec seulement $O(nd)$ opérations.

1.4.1 Algorithmes de gradient stochastiques

L'algorithme de gradient stochastique, introduit par [RM51], est défini de manière récursive pour tout $n \geq 0$ par

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

avec m_0 borné et (γ_n) est une suite de pas positifs vérifiant

$$\sum_{n\geq 1} \gamma_n = +\infty \qquad \text{et} \qquad \sum_{n\geq 1} \gamma_n^2 < +\infty. \tag{1.3}$$

A noter que l'on peut réécrire l'algorithme de gradient stochastique comme

$$m_{n+1} = m_n - \gamma_{n+1} \nabla G\left(m_n\right) + \gamma_{n+1} \xi_{n+1} \tag{1.4}$$

avec $\xi_{n+1} = \nabla G\left(m_n\right) - \nabla_h g\left(X_{n+1}, m_n\right)$. De plus, en considérant la filtration (notion que l'on définira par la suite) $(\mathcal{F}_n)$ engendrée par l'échantillon, i.e $\mathcal{F}_n = \sigma\left(X_1, \ldots, X_n\right)$, (ξ_{n+1}) est une suite de différences de martingale adaptée à la filtration $(\mathcal{F}_n)$, i.e comme m_n est $\mathcal{F}_n$-mesurable, et par indépendance,

$$\mathbb{E}\left[\xi_{n+1} | \mathcal{F}_n\right] = \nabla G\left(m_n\right) - \mathbb{E}\left[\nabla_h g\left(X_{n+1}, m_m\right) | \mathcal{F}_n\right] = 0.$$

On peut donc voir l'algorithme comme un algorithme de gradient bruité (par $\gamma_{n+1}\xi_{n+1}$).

Exemple 1 : la régression linéaire. On se place dans le cadre de la régression linéaire et on considère des couples de variables aléatoires i.i.d $\left(X_1, Y_1\right), \ldots, \left(X_n, Y_n\right), \left(X_{n+1}, Y_{n+1}\right), \ldots$ à valeurs dans $\mathbb{R}^d \times \mathbb{R}$. L'algorithme de gradient stochastique est alors défini récursivement pour tout $n \geq 0$ par

$$\theta_{n+1} = \theta_n + \gamma_{n+1} \left(Y_{n+1} - \theta_n^T X_{n+1}\right) X_{n+1}. \tag{1.5}$$

Exemple 2 : la régression logistique. On se place dans le cadre de la régression logistique et on considère des couples de variables aléatoires i.i.d $\left(X_1, Y_1\right), \ldots, \left(X_n, Y_n\right), \left(X_{n+1}, Y_{n+1}\right), \ldots$ à valeurs dans $\mathbb{R}^d \times \{0, 1\}$. L'algorithme de gradient stochastique est alors défini de manière récursive pour tout $n \geq 0$ par

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \left(\pi\left(\theta_n^T X_{n+1}\right) - Y_{n+1}\right) X_{n+1} \tag{1.6}$$

avec $\pi(x) = \frac{\exp(x)}{1+\exp(x)}$.

Exemples 3 : la médiane géométrique. On considère des variables aléatoires i.i.d $X_1, \ldots, X_n, X_{n+1}, \ldots$ à valeurs dans $\mathbb{R}^d$. L'algorithme de gradient stochastique est défini de manière récursive pour tout $n \geq 0$ par

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

1.4.2 Approche non-asymptotique

On suppose à partir de maintenant que la suite de pas γ_n est de la forme $\gamma_n = c_\gamma n^{-\alpha}$ avec $c_\gamma > 0$ et $\alpha \in (1/2, 1)$. On voit bien que cette suite de pas vérifie la condition (1.3). Le théorème suivant nous donne la vitesse de convergence en moyenne quadratique des estimateurs obtenus à l'aide de l'algorithme de gradient stochastique dans le cas où la fonction G est fortement convexe.

Théorème 1.4.1. *On suppose que les hypothèses suivantes sont vérifiées :*

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

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

De plus, on suppose que l'hypothèse suivante est vérifiée :

(PS0) *Il existe une constante positive C telle 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 pour tout $n \geq 1$,

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

avec $C' = \max\{C, 2\mu^2\}$.

En d'autres termes, les estimateurs convergent en moyenne quadratique à la vitesse $\frac{1}{n^\alpha}$.

Démonstration. On a

$$\|m_{n+1} - m\|^2 = \|m_n - m\|^2 - 2\gamma_{n+1}\langle \nabla_h g(X_{n+1}, m_n), m_n - m\rangle + \gamma_{n+1}^2 \|\nabla_h g(X_{n+1}, m_n)\|^2 .$$

En passant à l'espérance conditionnelle et par linéarité, on obtient donc, comme m_n est $\mathcal{F}_n$-mesurable,

$$\begin{aligned}\mathbb{E}\left[\|m_{n+1} - m\|^2 | \mathcal{F}_n\right] &= \|m_n - m\|^2 - 2\gamma_{n+1}\langle \mathbb{E}\left[\nabla_h g(X_{n+1}, m_n) | \mathcal{F}_n\right], m_n - m\rangle + \gamma_{n+1}^2 \mathbb{E}\left[\|\nabla_h g(X_{n+1}, m_n)\|^2 | \mathcal{F}_n\right] \\ &\leq \|m_n - m\|^2 - 2\gamma_{n+1}\langle \nabla G(m_n), m_n - m\rangle + \gamma_{n+1}^2 \mathbb{E}\left[\|\nabla_h g(X_{n+1}, m_n)\|^2 | \mathcal{F}_n\right]\end{aligned}$$

Grâce à l'hypothèse **(PS0)**, il vient

$$\mathbb{E}\left[\|m_{n+1} - m\|^2 | \mathcal{F}_n\right] \leq \left(1 + C\gamma_{n+1}^2\right)\|m_n - m\|^2 - 2\gamma_{n+1}\langle \nabla G(m_n), m_n - m\rangle + C\gamma_{n+1}^2.$$

Par forte convexité, on obtient donc

$$\mathbb{E}\left[\|m_{n+1} - m\|^2 | \mathcal{F}_n\right] \leq \left(1 - 2\mu\gamma_{n+1} + C\gamma_{n+1}^2\right)\|m_n - m\|^2 + C\gamma_{n+1}^2.$$

En passant à l'espérance, on a

$$\mathbb{E}\left[\|m_{n+1} - m\|^2\right] \leq \left(1 - 2\mu\gamma_{n+1} + C\gamma_{n+1}^2\right)\mathbb{E}\left[\|m_n - m\|^2\right] + C\gamma_{n+1}^2,$$

et on conclut la preuve à l'aide du lemme suivant [BM13]

Lemma 1.4.1. *Soit (δ_n) une suite positive vérifiant*

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

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 $n \geq 1$,

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

On admettra ce lemme. En prenant $C' = \max\left\{C, 2\mu^2\right\}$, il suffit de vérifier que $\sqrt{C'/2} \geq \mu$ et que l'on a également

$$\mathbb{E}\left[\|m_{n+1} - m\|^2\right] \leq \left(1 - 2\mu\gamma_{n+1} + C'\gamma_{n+1}^2\right)\mathbb{E}\left[\|m_n - m\|^2\right] + C\gamma_{n+1}^2,$$

et les hypothèses du lemme sont donc vérifiées, ce qui conclut la preuve. □

Afin de simplifier les preuves par la suite, on donne maintenant une proposition (admise) que l'on peut voir comme une généralisation du lemme précédent.

Proposition 1.4.1. *Soit δ_n, γ_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 $c_0 \in \left(0, \gamma_{n_0}^{-1}\right)$ et une suite positive v_n telle que pour tout $n \geq n_0$,*

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

Alors pour tout $n \geq 2n_0$,

$$\delta_n \leq \exp\left(-\frac{c_0c_\gamma}{4}n^{1-\alpha}\right)\left(\delta_{n_0} + \sum_{k=n_0}^{n/2-1}\gamma_{k+1}v_{k+1}\right) + \max_{n/2 \leq k+1 \leq n-1} v_{k+1}$$

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

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

On donne maintenant un lemme qui sera très important pour établir les vitesses de convergence preque sûre des algorithmes de gradient stochastiques. Ce lemme est une application directe de la proposition précédente.

Lemma 1.4.2. *Soient A_n, B_n, r_n des suites de variables aléatoires positives telles que r_n converge presque sûrement vers 0 et*

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

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

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

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

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

Proof of Lemma 1.4.2. Afin de simplifier la preuve (et quitte a considérer n suffisamment grand), on va supposer que pour tout $n \geq 0$, $c\gamma_{n+1} \leq 1$. On considère maintenant l'évènement $E_{n,c} = \{|r_n| \leq c/2\}$, et on a donc $\mathbf{1}_{E_{n,c}^C}$ qui converge presque sûrement vers 0. On peut donc réécrire A_{n+1} comme

$$\begin{aligned} A_{n+1} &\leq (1 - c\gamma_{n+1}) A_n + \frac{c}{2}\gamma_{n+1}(A_n + B_n) + \overbrace{\gamma_{n+1} r_n (A_n + B_n)}^{=:\delta_n} \mathbf{1}_{E_{n,c}^C} \\ &\leq \left(1 - \frac{c}{2}\gamma_{n+1}\right) A_n + \frac{c}{2}\gamma_{n+1} B_n + \delta_n \mathbf{1}_{E_{n,c}^C} \end{aligned}$$

Par récurrence, on peut facilement montrer que pour tout $n \geq 0$, on a

$$A_n \leq \tilde{\beta}_{n,0} A_0 + \underbrace{\frac{c}{2}\sum_{k=0}^{n-1} \tilde{\beta}_{n,k+1}\gamma_{k+1} B_k}_{=:A_{1,n}} + \underbrace{\sum_{k=0}^{n-1} \tilde{\beta}_{n,k+1}\delta_k \mathbf{1}_{E_{k,c}^C}}_{=:A_{2,n}}$$

avec $\tilde{\beta}_{n,k} = \prod_{j=k+1}^{n} \left(1 - \frac{c}{2}\gamma_j\right)$ et $\tilde{\beta}_{n,n} = 1$. Avec des calculs classiques, on peut facilement montrer que $\tilde{\beta}_{n,0}$ converge à vitesse exponentielle. De plus, on peut réécrire $A_{2,n} = \tilde{\beta}_{n,0} \sum_{k=0}^{n-1} \tilde{\beta}_{k,0}^{-1} \delta_k \mathbf{1}_{E_{k,c}^C}$ et comme $\mathbf{1}_{E_{n,c}^C}$ converge presque sûrement vers 0, la somme est presque sûrement finie et on obtient donc

$$A_{2,n} = O\left(\tilde{\beta}_{n,0}\right) \quad p.s$$

et ce terme converge donc à vitesse exponentielle. Enfin, il existe une variable aléatoire B telle que pour tout $n \geq 1$ on a $B_n \leq B v_n$ presque sûrement, et on obtient donc la relation de récurrence

$$A_{1,n+1} = \left(1 - \frac{c}{2}\gamma_{n+1}\right) A_{1,n} + \frac{c}{2}\gamma_{n+1} B_n \leq \left(1 - \frac{c}{2}\gamma_{n+1}\right) A_{1,n} + \frac{c}{2} B \gamma_{n+1} v_n$$

et en appliquant la proposition 1.4.1, on obtient

$$A_{1,n} = O(v_n) \quad p.s.$$

□

1.4.3 Exemple : regression lineaire

On se place dans le cadre du modèle linéaire défini par (1.1) et on rappelle que sous certaines hypothèses, θ est l'unique minimiseur de la fonction $G : \mathbb{R}^d \longrightarrow \mathbb{R}_+$ définie pour tout $h \in \mathbb{R}^d$ par

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

Théorème 1.4.2. *On suppose que ϵ admet un moment d'ordre 2 et que X admet un moment d'ordre 4. De plus, on suppose que la matrice $\mathbb{E}\left[XX^T\right]$ est définie positive et on note μ sa plus petite valeur propre. La suite d'estimateurs du gradient (θ_n) définie par (1.5) vérifie pour tout $n \geq 0$,*

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

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

On admettra ce théorème. Dans la Figure 1.1, on s'intéresse à l'évolution de l'erreur quadratique moyenne des estimateurs (obtenus à l'aide d'algorithmes de gradient stochastiques) du paramètre de la régression linéaire en fonction de la taille d'échantillon n. Pour cela, on considère le modèle

$$\theta = (-4, -3, -2, -1, 0, 1, 2, 3, 4, 5)^T \in \mathbb{R}^{10}, \qquad X \sim \mathcal{N}(0, I_{10}), \qquad \epsilon \sim \mathcal{N}(0, 1).$$

De plus, on choisit $c_\gamma = 1$ et $\alpha = 0.5, 0.660.75$ ou 1. Enfin, on a calculé l'erreur quadratique moyenne des estimateurs en générant 50 échantillons de taille $n = 1000$. On voit bien que l'on a une décroissance assez rapide de l'erreur quadratique moyenne, et lorsque l'on prend l'échelle logarithmique, une heuristique de pente nous donne que pour $n \geq 100$, on a bien une pente proche de $-\alpha$.

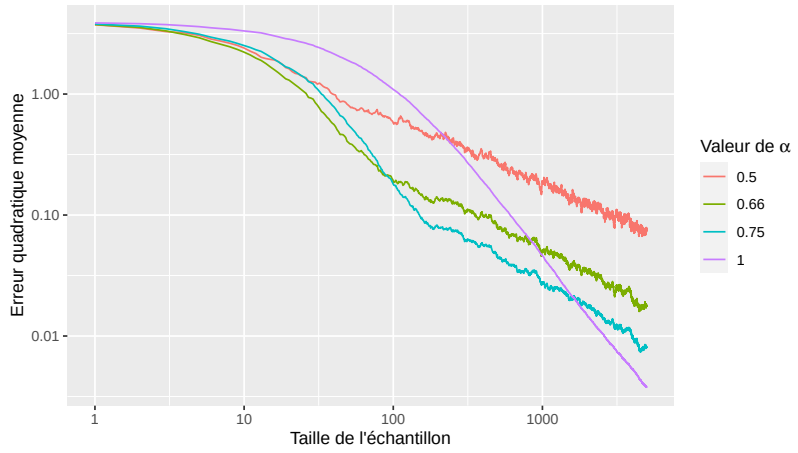

FIGURE 1.1 – Evolution de l'erreur quadratique moyenne de θ_n en fonction de la taille d'échantillon n dans le cadre de la régression linéaire.

Chapitre 2

Martingales

On a vu que l'on peut considérer l'algorithme de gradient stochastique comme un algorithme de gradient bruité par le terme $\gamma_{n+1}\xi_{n+1}$, où (ξ_n) est une suite de différences de martingale par rapport à une filtration $(\mathcal{F}_n)$. L'objectif de ce chapitre est donc, dans un premier temps, de définir les notions de filtration et de martingale. Dans un deuxième temps, on donnera des résultats de convergence type loi des grands nombres et TLC pour les martingales, qu'elles soient réelles ou vectorielles.

2.1 Martingales réelles

Cette section s'inspire très largement de [BC07] et [Duf90].

2.1.1 Définitions

Dans ce qui suit, on considère un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$.

Definition 2.1.1. *On appelle filtration $(\mathcal{F}_n)_{n\geq 0}$ de $(\Omega, \mathcal{A}, \mathbb{P})$ une suite croissante de sous-tribus de $\mathcal{A}$, i.e une suite de tribus telle que*

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset ... \subset \mathcal{A}.$$

On dit qu'une suite de variables aléatoires $(X_n)_{n\geq 0}$ est adaptée à la filtration $(\mathcal{F}_n)_{n\geq 0}$ si pour tout n, X_n est $\mathcal{F}_n$-mesurable.

En considérant une suite de variables aléatoires $(X_n)_{n\geq 1}$, un exemple classique de filtration est de considérer la suite des tribus engendrée par les X_i. Plus précisément, pour tout n, on considère la plus petite tribu rendant $(X_1, \ldots, X_n)$ mesurable et on la note $\mathcal{F}_n = \sigma(X_1 \ldots, X_n)$. Alors, $\mathcal{F} = (\mathcal{F}_n)$ est une filtration. Rappelons également au passage que par définition de l'espérance conditionnelle, $\mathbb{E}[X_{n+1}|\mathcal{F}_n]$ est $\mathcal{F}_n$-mesurable.

Definition 2.1.2. *Soit $M = (M_n)_{n\geq 0}$ une suite de variables aléatoires adaptée à une filtration $\mathcal{F} = (\mathcal{F}_n)$.*

On dit que M est une martingale par rapport à la filtration $\mathcal{F}$ si pour tout $n \geq 0$, M_n est intégrable et

$$\mathbb{E}\left[M_{n+1}|\mathcal{F}_n\right] = M_n.$$

Exemple : On considère une suite de variables aléatoires indépendantes $(X_n)_{n\geq 1}$ et on note

$$M_n = \sum_{k=1}^{n} X_k - \mathbb{E}\left[X_k\right].$$

Alors M_n est une martingale par rapport à la filtration engendrée par les X_i.

2.1.2 Théorème de Robbins-Siegmund

Le théorème de Robbins-Siegmund suivant est crucial pour démontrer la consistance des estimateurs obtenus à l'aides d'algorithmes stochastiques, et particulièrement pour les algorithmes de gradient.

Théorème 2.1.1 (Robbins-Siegmund). *Soit $(V_n), (A_n), (B_n), (C_n)$ trois suites de variables réelles positives adaptées à une filtration $\mathcal{F}$. On suppose que*

$$\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq (1 + A_n) V_n + B_n - C_n$$

et que les suites $(A_n), (B_n)$ vérifient

$$\sum_{n\geq 0} A_n < +\infty \quad p.s \qquad et \qquad \sum_{n\geq 0} B_n < +\infty \quad p.s.$$

Alors V_n converge presque sûrement vers une variable aléatoire V_∞ finie et

$$\sum_{n\geq 0} C_n < +\infty \quad p.s.$$

On admettra ce théorème mais sa preuve est disponible dans [Duf90] ou [BC07], ainsi que dans la version longue.

Exemple : estimation en ligne des quantiles. Soit $X_1, \ldots, X_{n+1}, \ldots$ des variables aléatoires indépendantes et identiquement distribuées. Soit $p \in (0,1)$, et on s'intéresse à l'estimation en ligne du quantile d'ordre p, que l'on notera m. Pour cela, on considère l'estimateur obtenu à l'aide de l'algorithme de Robbins-Monro [RM51], défini de manière récursive pour tout $n \geq 0$ par

$$m_{n+1} = m_n - \gamma_{n+1}\left(\mathbf{1}_{X_{n+1}\leq m_n} - p\right).$$

avec

$$\sum_{n\geq 0} \gamma_{n+1} = +\infty \qquad \text{et} \qquad \sum_{n\geq 0} \gamma_{n+1}^2 < +\infty.$$

On pose alors $V_n = (m_{n+1} - m)^2$, et on a

$$\begin{aligned} V_{n+1} &= V_n - 2\gamma_{n+1}(m_n - m)\left(\mathbf{1}_{X_{n+1}\leq m_n} - p\right) + \gamma_{n+1}^2\left(\mathbf{1}_{X_{n+1}\leq m_n} - p\right)^2 \\ &\leq V_n - 2\gamma_{n+1}(m_n - m)\left(\mathbf{1}_{X_{n+1}\leq m_n} - p\right) + \gamma_{n+1}^2 \end{aligned}$$

On considère la filtration $(\mathcal{F}_n)$ engendrée par l'échantillon, i.e définie pour tout $n \geq 1$ par $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. Comme l'estimateur m_n ne dépend que de $X_1, \ldots, X_n$, il est $\mathcal{F}_n$-mesurable, et on a donc, en notant F_X la fonction de répartition de X_1, et en remarquant que $p = F_X(m)$,

$$\begin{aligned} \mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] &\leq V_n - 2\gamma_{n+1}(m_n - m)\left(\mathbb{E}\left[\mathbf{1}_{X_{n+1}\leq m_n}|\mathcal{F}_n\right] - p\right) + \gamma_{n+1}^2 \\ &= V_n - \underbrace{2\gamma_{n+1}(m_n - m)\left(F_X(m_n) - F_X(m)\right)}_{=:A_n} + \gamma_{n+1}^2 \end{aligned}$$

Comme la fonction de répartition est croissante, on a $A_n \geq 0$ et comme $\sum_{n\geq 0}\gamma_{n+1}^2 < +\infty$, le théorème de Robbins-Siegmund nous donne que V_n converge presque sûrement vers une variable aléatoire finie et que

$$\sum_{n\geq 0} A_n < +\infty \quad p.s.$$

Or , comme la somme des γ_n diverge, cela implique que $\liminf_n (m_n - m)\left(F_X(m_n) - F(m)\right) = 0$. Si on suppose que la fonction F est strictement croissante sur un voisinage de F, on obtient

$$\liminf_n |m_n - m| = 0,$$

et comme $|m_n - m|$ converge vers une variable aléatoire finie, cela implique que $|m_n - m|$ converge presque sûrement vers 0, i.e que l'estimateur est fortement consistant.

Exemple : estimation des quantiles de la loi exponentielle. On considère $X \sim \mathcal{E}(1)$ et on s'intéresse à l'estimation des quantiles d'ordre 0.25 et 0.75. Figure 2.1, on voit que dans les deux cas les estimateurs convergent assez rapidement vers le quantile.

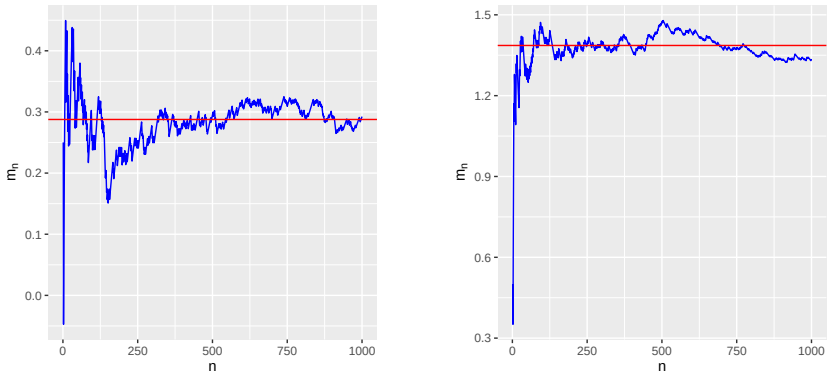

FIGURE 2.1 – Evolution de l'estimation des quantiles d'ordre 0.25 (à gauche) et 0.75 à droite pour la loi exponentielle de paramètre 1.

2.1.3 Lois des grands nombres

Dans ce qui suit, on dit qu'une martingale (M_n) est de carré intégrable si pour tout $n \geq 0$, $\mathbb{E}\left[M_n^2\right] < +\infty$.

Definition 2.1.3. *Soit (M_n) une martingale de carré intégrable. On appelle processus croissant associé à (M_n) la suite $(\langle M\rangle_n)_n$ définie par $\langle M\rangle_0 = 0$ et pour tout $n \geq 0$ par*

$$\langle M\rangle_{n+1} = \langle M\rangle_n + \mathbb{E}\left[(M_{n+1} - M_n)^2 \,|\mathcal{F}_n\right].$$

En d'autres termes, si pour tout $k \geq 0$ on note $\xi_{k+1} = M_{k+1} - M_k$ la différence de martingale, on a pour tout $n \geq 1$

$$\langle M\rangle_n = \sum_{k=1}^{n} \mathbb{E}\left[\xi_k^2 | \mathcal{F}_{k-1}\right].$$

Ainsi, on peut voir le processus croissant comme la somme des variances conditionnelles des différences de martingales. On peut maintenant introduire les lois des grands nombres pour les martingales de carré intégrable.

Théorème 2.1.2 (Première loi des grands nombres)**.** *Soit (M_n) une martingale de carré intégrable.*

1. *Si $\lim_{n\to+\infty}\langle M\rangle_n < +\infty$ presque sûrement, alors la suite (M_n) converge presque sûrement vers une variable aléatoire M_∞.*
2. *Si $\lim_{n\to+\infty}\langle M\rangle_n = +\infty$ presque sûrement, alors la suite $\left(\frac{M_n}{\langle M\rangle_n}\right)$ converge presque sûrement vers 0.*

En d'autres termes, si le processus croissant converge presque sûrement, alors $(M_n) = O(1)$ presque sûrement, et si il diverge, alors

$$|M_n| = o\left(\langle M\rangle_n\right) \quad p.s.$$

On admettra ce résultat. La preuve est disponible dans la version longue ainsi que dans [BC07] et [Duf97].

Application au bandit à deux bras : Le problème du bandit à deux bras consiste à considérer une machine à sous avec deux leviers (A et B). Pour chacun des leviers, le gain est de 1 avec probabilité θ_A ou θ_B et de 0 avec probabilité $1-\theta_A$ ou $1-\theta_B$. Dans ce qui suit, on suppose $\theta_A, \theta_B \in (0,1)$ et $\theta_A \neq \theta_B$. A chaque temps n, le joueur decide d'actionner le levier U_n (et donc $U_n = A$ ou B) par rapport à ce qu'il a pu observer avant. On note X_n son gain au temps n et on a donc $\mathbb{P}\left[X_n = 1 | \mathcal{U}_n = A\right] = \theta_A$ ou $\mathbb{P}\left[X_n = 1 | U_n = B\right] = \theta_B$, i.e $X_n | U_n \sim \mathcal{B}\left(\theta_{U_n}\right)$. L'objectif du joueur est de trouver la stratégie (U_n) qui maximise son gain moyen asymptotique, i.e en notant

$$G_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

l'objectif est de trouver la stratégie (U_n) qui permette d'obtenir

$$G_n \xrightarrow[n\to+\infty]{p.s} \max\{\theta_A, \theta_B\}.$$

En effet, si $\theta_A > \theta_B$, par exemple, il faudrait toujours actionner le levier A, et on a donc G_n qui convergerait presque sûrement vers θ_A et inversement. Dans ce qui suit, on note $N_{A,n}$ et $N_{B,n}$ le nombre de fois où chacun des leviers a été tiré au temps n, i.e

$$N_{A,n} = \sum_{k=1}^{n} \mathbf{1}_{U_k=A}, \qquad \text{et} \qquad N_{B,n} = \sum_{k=1}^{n} \mathbf{1}_{U_k=B}$$

On note maintenant

$$M_n = \sum_{k=1}^{n} X_k - \theta_A N_{A,n} - \theta_B N_{B,n}$$

On considère la filtration $(\mathcal{F}_n)$ définie pour tout $n \geq 0$ par $\mathcal{F}_n = \sigma(X_1, \dots, X_n, U_1, \dots, U_{n+1})$, et on a

$$\mathbb{E}[M_{n+1}|\mathcal{F}_n] = \mathbb{E}[X_{n+1}|\mathcal{F}_n] - \theta_A \mathbf{1}_{U_{n+1}=A} - \theta_B \mathbf{1}_{U_{n+1}=B} + M_n$$

et comme $X_{n+1}|U_{n+1} \sim \mathcal{B}\left(\theta_{U_{n+1}}\right)$, M_n est une martingale, et elle est clairement de carré intégrable. De plus, on a

$$\begin{aligned}
\langle M\rangle_n &= \sum_{k=1}^{n} \mathbb{E}\left[\left(X_k - \theta_A \mathbf{1}_{U_k=A} - \theta_B \mathbf{1}_{U_k=B}\right)^2 |\mathcal{F}_{k-1}\right] \\
&= \sum_{k=1}^{n} \mathbb{E}\left[\left(X_k - \theta_{U_k}\right)^2 |U_k\right] \\
&= \sum_{k=1}^{n} \mathbb{V}\left[X_k|U_k\right] \\
&= \sum_{k=1}^{n} \theta_A(1-\theta_A)\mathbf{1}_{U_k=A} + \theta_B(1-\theta_B)\mathbf{1}_{U_k=B} \\
&= \theta_A(1-\theta_A)N_{A,n} + \theta_B(1-\theta_B)N_{B,n}.
\end{aligned}$$

Afin d'appliquer le 2ème point de la loi des grands nombres, on va montrer que le crochet diverge. Pour cela, il suffit de remarquer que $N_{A,n} + N_{B,n} = n$, et donc soit $N_{A,n} \geq n/2$, soit $N_{B,n} \geq n/2$. Ainsi,

$$\langle M\rangle_n \geq \underbrace{\min\{\theta_A(1-\theta_A), \theta_B(1-\theta_B)\}}_{>0} \frac{n}{2} \xrightarrow[n\to+\infty]{} +\infty.$$

Par la loi des grands nombres, $\frac{M_n}{\langle M\rangle_n} \xrightarrow[n\to+\infty]{p.s} 0$. On suppose maintenant qu'il existe des constantes $l_A, l_B \in [0,1]$ telles que

$$\frac{N_{A,n}}{n} \xrightarrow[n\to+\infty]{p.s} l_A \qquad \text{et} \qquad \frac{N_{B,n}}{n} \xrightarrow[n\to+\infty]{p.s} l_B.$$

On va maintenant montrer que G_n converge presque sûrement vers $\theta_A l_A + \theta_B l_B$. On a

$$\begin{aligned} G_n - \theta_A l_A - \theta_B l_B &= \frac{1}{n} M_n + \theta_A \frac{N_{A,n}}{n} + \theta_B \frac{N_{B,n}}{n} - \theta_A l_A - \theta_B l_B \\ &= \frac{\langle M \rangle_n}{n} \underbrace{\frac{M_n}{\langle M \rangle_n}}_{\xrightarrow[n\to+\infty]{p.s} 0} + \underbrace{\theta_A \left(\frac{N_{A,n}}{n} - l_A \right) + \theta_B \left(\frac{N_{B,n}}{n} - l_B \right)}_{\xrightarrow[n\to+\infty]{p.s} 0} \end{aligned}$$

et comme $\frac{\langle M \rangle_n}{n} \in [0,1]$, on obtient

$$G_n \xrightarrow[n\to+\infty]{p.s} \theta_A l_A + \theta_B l_B.$$

Ainsi, une bonne stratégie consiste à obtenir $l_A = 1$ si $\theta_A > \theta_B$ et inversement.

On peut remarquer que la différence avec la loi des grands nombres pour des variables aléatoires i.i.d est que l'on se passe des hypothèses d'indépendance et d'identique distribution, mais le prix à payer est que l'on doit faire des hypothèses sur le comportement du crochet $\langle M \rangle_n$. A noter également que dans les exemples précédents, on aurait pu s'attendre à obtenir une meilleure vitesse de convergence, ce que nous donne la deuxième loi des grands nombres suivante.

Théorème 2.1.3 (Deuxième loi des grands nombres)**.** *Soit (M_n) une martingale de carré intégrable.*

1. *Si $\langle M \rangle_n \xrightarrow[n\to+\infty]{p.s} +\infty$, alors*

$$M_n^2 = o\left(\langle M \rangle_n \ln\left(\langle M \rangle_n \right) \right)^{1+\delta} \quad p.s.$$

2. *De plus, si il existe des constantes $a > 2$ et $b > 0$ telles que*

$$\mathbb{E}\left[|M_{n+1} - M_n|^a \,|\mathcal{F}_n \right] \leq b \left(\mathbb{E}\left[(M_{n+1} - M_n)^2 \,|\mathcal{F}_n \right] \right)^{a/2} \quad p.s$$

alors

$$M_n^2 = O\left(\langle M \rangle_n \ln\left(\langle M \rangle_n \right) \right) \quad p.s.$$

On admettra ce résultat. La preuve est disponible dans la version longue ainsi que dans [BC07] et [Duf97]. A noter que les hypothèses (notamment pour la deuxième partie du théorème) peuvent sembler indigestes, mais on peut voir dans l'exemple suivant qu'elles sont généralement "facilement" vérifiables.

Exemple : Soit $M_n = \sum_{k=1}^n \xi_k$ avec $\mathbb{E}\left[\xi_k | \mathcal{F}_{k-1}\right] = 0$. Si il existe une constante C telle que pour tout k, $\mathbb{E}\left[\xi_k^2 | \mathcal{F}_{k-1}\right] \leq C$, alors pour tout $\delta > 0$,

$$M_n^2 = o\left(n (\ln n)^{1+\delta} \right) \quad p.s$$

ce que l'on peut réécrire comme

$$\left|\frac{1}{n}\sum_{k=1}^{n}\xi_k\right|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) \quad p.s.$$

Remarque 2.1.1. *A noter que l'on aurait pu retrouver directement le premier point de cet exemple avec une application directe du théorème de Robbins-Siegmund. En effet, pour tout $\delta > 0$, on peut considérer la variabe aléatoire $V_n = \frac{1}{n(\ln n)^{1+\delta}} |M_n|^2 = \frac{n}{(\ln n)^{1+\delta}} \left\|\frac{1}{n}M_n\right\|^2$. Comme M_n est $\mathcal{F}_n$-mesurable et par linéarité, on a*

$$\begin{aligned}\mathbb{E}\left[|M_{n+1}|^2 |\mathcal{F}_n\right] &= M_n^2 + 2\mathbb{E}\left[\xi_{n+1}|\mathcal{F}_n\right] M_n + \mathbb{E}\left[|\xi_{n+1}|^2 |\mathcal{F}_n\right] \\ &= |M_n|^2 + \mathbb{E}\left[|\xi_{n+1}|^2 |\mathcal{F}_n\right]\end{aligned}$$

Ainsi, on obtient

$$\begin{aligned}\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] &= \frac{1}{(n+1)(\ln(n+1))^{1+\delta}} |M_n|^2 + \frac{1}{(n+1)(\ln(n+1))^{1+\delta}}\mathbb{E}\left[|\xi_{n+1}|^2 |\mathcal{F}_n\right] \\ &= \frac{n(\ln n)^{1+\delta}}{(n+1)(\ln(n+1))^{1+\delta}} V_n + \frac{1}{(n+1)(\ln(n+1))^{1+\delta}}\mathbb{E}\left[|\xi_{n+1}|^2 |\mathcal{F}_n\right] \\ &\leq V_n + \frac{1}{(n+1)(\ln(n+1))^{1+\delta}} C.\end{aligned}$$

En appliquant le théorème de Robbins-Siegmund, on a donc V_n qui converge presque sûrement vers une variable aléatoire finie, ce que l'on peut réécrire comme

$$\left\|\frac{1}{n}M_n\right\|^2 = O\left(\frac{(\ln n)^{1+\delta}}{n}\right) \quad a.s.$$

La proposition suivante donne une encore meilleure vitesse de convergence, mais au prix d'hypothèses un peu plus fortes.

Proposition 2.1.1. *Soit (ξ_n) une suite de différences de martingales adaptée à une filtration $(\mathcal{F}_n)_n$. Supposons qu'il existe des constantes $a > 2$ et $C_a \geq 0$ telles que $\mathbb{E}\left[|\xi_k|^a |\mathcal{F}_{k-1}\right] \leq C_a$ presque sûrement, alors,*

$$\left|\frac{1}{n}\sum_{k=1}^{n}\xi_k\right|^2 = O\left(\frac{\ln n}{n}\right) \quad p.s.$$

Application au bandit à deux bras : En reprenant les notations de l'exemple du bandit à deux bras, on a

$$|M_{n+1} - M_n| \leq 1,$$

et on peut donc appliquer la Proposition 2.1.1, on a

$$\left|\frac{1}{n}M_n\right|^2 = O\left(\frac{\ln n}{n}\right) \quad p.s$$

et si on a les vitesses de converges de $\frac{N_{A,n}}{n}$ et $\frac{N_{B,n}}{n}$ vers l_A, l_B, on obtient donc une meilleur vitesse de convergence pour le gain moyen G_n.

A noter que pour ces exemples, avec des hypothèses un peu plus restrictives, on peut trouver une meilleure vitesse grâce à la loi du log-itéré pour les martingales. Celle-ci est disponible dans la verson longue ainsi que dans [DST90] et [Duf90] (page 31).

Estimation en ligne des quantiles : A l'aide de la loi du log-itéré, on peut montrer (cf version longue) que l'estimateur en ligne des quantiles vérifie

$$|m_n - m|^2 = O\left(\frac{\ln n}{n^\alpha}\right) p.s$$

Exemple : estimation de la médiane de la loi exponentielle. On considère $X \sim \mathcal{E}(1)$ et on s'intéresse à la vitesse de convergence des estimateurs de la médiane. Figure 2.2, on se concentre sur l'erreur quadratique moyenne estimée à l'aide de 50 échantillons. On voit bien que les estimateurs convergent très rapidement vers la médiane.

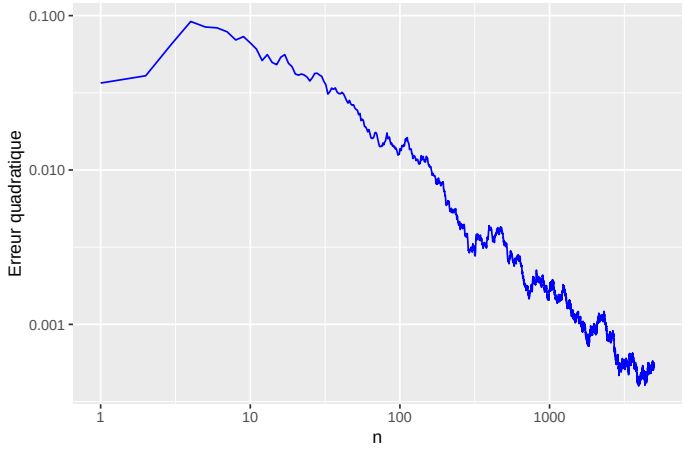

FIGURE 2.2 – Evolution de l'erreur quadratique moyenne des estimateurs en ligne de la médiane d'une loi exponentielle de paramètre 1.

2.1.4 Théorème limite centrale

On s'intéresse ici à l'obtention d'un TLC pour les martingales de carré intégrable, i.e on cherche à transposer le TLC usuel aux martingales.

Théorème 2.1.4. *Soient (M_n) une martingale de carré intégrable et a_n une suite positive, croissante et divergente. On suppose que les hypothèses suivantes sont vérifiées :*

1. *Il existe* $\sigma^2 \geq 0$ *telle que*
$$\frac{\langle M \rangle_n}{a_n} \xrightarrow[n\to+\infty]{\mathbb{P}} \sigma^2.$$

2. *La condition de Lindeberg est vérifiée, i.e pour tout* $\epsilon > 0$,
$$\frac{1}{a_n} \sum_{k=1}^{n} \mathbb{E}\left[(M_k - M_{k-1})^2 \mathbf{1}_{|M_k - M_{k-1}| \geq \epsilon \sqrt{a_n}} | \mathcal{F}_{k-1} \right] \xrightarrow[n\to+\infty]{\mathbb{P}} 0.$$

Alors
$$\frac{1}{\sqrt{a_n}} M_n \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right).$$

A noter que l'on s'est débarrassé des conditions d'indépendance et d'identique distribution, mais cela au prix de la condition (indigeste) de Lindeberg. Cependant, celle-ci est vérifiée dès que la condition (relativement plus digeste) de Lyapunov est vérifiée, i.e elle est vérifiée si il existe $a > 2$ tel que
$$\frac{1}{a_n^{\frac{a}{2}}} \sum_{k=1}^{n} \mathbb{E}\left[|M_k - M_{k-1}|^a | \mathcal{F}_{k-1} \right] \xrightarrow[n\to+\infty]{\mathbb{P}} 0.$$
En effet, en notant $\xi_{n+1} = M_{n+1} - M_n$, et comme $a > 2$
$$\begin{aligned} \mathbb{E}\left[\xi_k^2 \mathbf{1}_{|\xi_k| \geq \epsilon \sqrt{a_n}} | \mathcal{F}_{k-1} \right] \leq \mathbb{E}\left[|\xi_k|^a |\xi_k|^{2-a} \mathbf{1}_{|\xi_k| \geq \epsilon \sqrt{a_n}} | \mathcal{F}_{k-1} \right] &\leq \epsilon^{2-a} a_n^{\frac{2-a}{2}} \mathbb{E}\left[|\xi_k|^a \mathbf{1}_{|\xi_k| > \epsilon \sqrt{a_n}} | \mathcal{F}_{k-1} \right] \\ &\leq \epsilon^{2-a} a_n^{\frac{2-a}{2}} \mathbb{E}\left[|\xi_k|^a | \mathcal{F}_{k-1} \right]. \end{aligned}$$

On obtient donc
$$\frac{1}{a_n} \sum_{k=1}^{n} \mathbb{E}\left[(M_k - M_{k-1})^2 \mathbf{1}_{|M_k - M_{k-1}| \geq \epsilon \sqrt{a_n}} | \mathcal{F}_{k-1} \right] \leq \epsilon^{2-a} \frac{1}{a_n^{\frac{a}{2}}} \sum_{k=1}^{n} \mathbb{E}\left[|\xi_k|^a | \mathcal{F}_{k-1} \right]$$
et on a donc bien que la condition de Lindeberg est vérifiée si la condition de Lyapunov l'est. En pariculier, lorsque $a_n = n$, si il existe des constantes $a > 2$ et C_a telles que pour tout $k \geq 0$, $\mathbb{E}\left[|\xi_{k+1}|^a | \mathcal{F}_{k-1} \right] \leq C_a$, alors la condition de Lindeberg est vérifiée.

Exemple : Soit $M_n = \sum_{k=1}^{n} \xi_k$ avec $\mathbb{E}\left[\xi_k | \mathcal{F}_{k-1} \right] = 0$ et telle qu'il existe σ^2 vérifiant
$$\frac{1}{n} \sum_{k=1}^{n} \mathbb{E}\left[\xi_k^2 | \mathcal{F}_{k-1} \right] \xrightarrow[n\to+\infty]{\mathbb{P}} \sigma^2.$$
Si il existe des constantes $a > 2$ et $C_a \geq 0$ telles que $\mathbb{E}\left[|\xi_k|^a | \mathcal{F}_{k-1} \right] \leq C_a$, alors
$$\frac{1}{\sqrt{n}} M_n \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right).$$

Application au bandit à deux bras : On reprend les notations de l'exemple du bandit à deux bras,

et on cherche maintenant à estimer θ_A, θ_B. Pour cela, on considère les estimateurs naturels

$$\theta_{A,n} = \frac{1}{N_{A,n}} \sum_{k=1}^{n} \mathbf{1}_{U_k=A, X_k=1} \qquad \text{et} \qquad \theta_{B,n} = \frac{1}{N_{B,n}} \sum_{k=1}^{n} \mathbf{1}_{U_k=B, X_k=1}$$

On note $M_{A,n} = \sum_{k=1}^{n} \mathbf{1}_{U_k=A, X_k=1} - N_{A,n}\theta_A$, i.e on a $\theta_{A,n} - \theta_A = \frac{M_{A,n}}{N_{A,n}}$. On remarque que

$$\mathbb{E}\left[M_{A,n+1}|\mathcal{F}_n\right] = M_{A,n} + \mathbb{E}\left[\mathbf{1}_{X_{n+1}=1}|U_{n+1}\right]\mathbf{1}_{U_{n+1}=A} - \theta_A \mathbf{1}_{U_{n+1}=A} = M_{A,n}$$

et $M_{A,n}$ est donc une martingale (de carré intégrable). De plus,

$$\langle M_A \rangle_n = \sum_{k=1}^{n} \mathbb{E}\left[\left(\mathbf{1}_{X_k=1} - \theta_A\right)^2 |U_k\right] \mathbf{1}_{U_k=A} = \sum_{k=1}^{n} \mathbb{V}\left[X_k|U_k = A\right] \mathbf{1}_{U_k=A} = \theta_A\left(1-\theta_A\right) N_{A,n}$$

et par la première loi des grands nombres, si $\lim_{n\to+\infty} N_{A,n} < +\infty$ p.s, alors $M_{A,n}$ converge presque sûrement vers une variable aléatoire finie, mais $\theta_{A,n}$ ne converge pas nécéssairement vers θ_A. Par contre, si $N_{A,n}$ diverge presque sûrement, la loi des grands nombres nous donne que $M_{A,n} = o\left(\langle M_A \rangle_n\right) p.s$, i.e $\theta_{A,n}$ converge presque sûrement vers θ_A. On peut bien évidemment faire le même travail pour obtenir la convergence de $\theta_{B,n}$. On s'intéresse maintenant à la normalité asymptotique de ces estimateurs. On suppose maintenant que $\frac{N_{A,n}}{n} \xrightarrow[n\to+\infty]{p.s} l_A > 0$, et on a

$$\frac{1}{n}\langle M_A \rangle_n \xrightarrow[n\to+\infty]{p.s} \theta_A\left(1-\theta_A\right) l_A.$$

De plus, on remarque que $|M_{A,k} - M_{A,k-1}| \leq 1$, et la condition de Lindeberg est donc vérifiée. Le TLC pour les martingales nous donne alors

$$\frac{1}{\sqrt{n}} M_{A,n} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \theta_A\left(1-\theta_A\right) l_A\right)$$

Ainsi, à l'aide du Théorème de Slutsky, on obtient

$$\sqrt{n}\left(\theta_{A,n} - \theta\right) = \sqrt{n}\frac{M_{A,n}}{N_{A,n}} = \underbrace{\frac{n}{N_{A,n}}}_{\xrightarrow[n\to+\infty]{p.s} l_A^{-1}} \sqrt{n} M_{A,n} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\theta_A\left(1-\theta_A\right)}{l_A}\right).$$

Bandits à deux bras et efficacité asymptotique : On a vu que l'on peut avoir des estimateurs efficaces et asymptotiquement normaux de θ_A, θ_B. Se pose maintenant la question de savoir comment obtenir une stratégie (U_n) permettant d'obtenir le meilleur gain moyen asymptotique ainsi que d'obtenir son efficacité asymptotique. Une façon naturelle de choisir un levier au temps n serait de prendre

$$U_n = A\mathbf{1}_{\theta_{A,n} \geq \theta_{B,n}} + B\mathbf{1}_{\theta_{B,n} > \theta_{A,n}},$$

i.e on choisi le levier A si l'estimateur de θ_A est "meilleur" que celui de θ_B (et inversement). Bien

que cette stratégie soit naturelle, elle pose le problème quelle peut se faire piéger. Par exemple, si on choisit $U_1 = A$ et que l'on perd le premier tirage, i.e $X_1 = 0$, puis que l'on choisisse $U_2 = B$ et que l'on obtienne $X_2 = 1$, alors $\theta_{A,2} = 0 < \theta_{B,2} = 1$. Pour toute la suite, on aura $\theta_{B,n} > 0 = \theta_{A,n}$, i.e on ne choisira jamais le levier A. On a donc G_n qui converge presque sûrement vers θ_B, et si $\theta_A > \theta_B$, on ne peut donc pas converger vers la bonne solution. Une solution pour pallier ce problème est de forcer le choix du levier de temps en temps. Plus précisément, on considère (c_n) une suite croissante de $\mathbb{N}$, et note $I_c = \{c_n, n \geq 1\}$ l'ensemble des valeurs de la suite (en d'autres termes, on a pris un sous-ensemble de $\mathbb{N}$). On adopte alors la stratégie suivante :

$$U_n = \begin{cases} A & \text{si } \theta_{A,n-1} \geq \theta_{B,n-1} \text{ et } n \notin I_c \\ B & \text{si } \theta_{B,n-1} > \theta_{A,n-1} \text{ et } n \notin I_c \\ A & \text{si } \exists k \geq 1, n = c_{2k} \\ B & \text{si } \exists k \geq 0, n = c_{2k+1} \end{cases}$$

A noter qu'avec cette stratégie, on choisit obligatoirement une infinité de fois A et B, et on a donc $N_{A,n}$ et $N_{B,n}$ qui divergent presque sûrement, i.e $\theta_{A,n}$ et $\theta_{B,n}$ convergent presque sûrement vers θ_A et θ_B. De plus, pour simplifier les notations, on suppose que $\theta_A > \theta_B$ (l'autre cas étant analogue). On suppose maintenant que $n = o(c_n)$, et on remarque que

$$\frac{N_{B,n}}{n} = \frac{1}{n}\text{Card}\left\{k, c_{2k+1} \leq n\right\} + \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\theta_{B,k} > \theta_{A,k}} \leq \frac{1}{n}\underbrace{\text{Card}\left\{k, c_k \leq n\right\}}_{=:C_n} + \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\theta_{B,k} > \theta_{A,k}}.$$

Comme les deux estimateurs $\theta_{A,n}$ et $\theta_{B,n}$ sont consistants, on a $\mathbf{1}_{\theta_{B,n} > \theta_{A,n}}$ qui converge presque sûrement vers 0 et donc $\sum_{k \geq 0} \mathbf{1}_{\theta_{B,k} > \theta_{A,k}} < +\infty$ p.s, i.e on a

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\theta_{B,k} > \theta_{A,k}} = O\left(\frac{1}{n}\right) \quad p.s.$$

Il reste à montrer que $C_n = o(n)$. Pour cela, on suppose par l'absurde que ce n'est pas le cas, i.e $\liminf_n \frac{C_n}{n} > 0$, i.e il existe $\eta > 0, n_\eta \geq 0$ tels que pour tout $n \geq n_\eta$, $C_n \geq \lceil \eta n \rceil$. On a alors pour tout $n \geq n_\eta$

$$C_n = \max\left\{k, c_k \leq n\right\} \geq \lceil \eta n \rceil \Leftrightarrow c_{\lceil \eta n \rceil} \leq n \Leftrightarrow \frac{n}{c_{\lceil \eta n \rceil}} \geq 1$$

ce qui est contradictoire avec $n = o(c_n)$ car pour tout $n \geq n_\eta$, $\frac{n}{c_{\lceil \eta n \rceil}} \leq \frac{n}{\lceil \eta n \rceil}\frac{\lceil \eta n \rceil}{c_{\lceil \eta n \rceil}} \xrightarrow[n \to +\infty]{} 0$. On obtient donc que $\frac{N_{B,n}}{n}$ converge presque sûrement vers 0 et donc que $\frac{N_{A,n}}{n}$ converge presque sûrement vers 1, i.e que $l_A = 1$ et $l_B = 0$. On obtient donc

$$G_n \xrightarrow[n \to +\infty]{p.s} \max\left\{\theta_A, \theta_B\right\}.$$

On s'intéresse maintenant à l'efficacité asymptotique de G_n. Pour cela, on suppose maintenant que

$n^2 = o(c_n)$. On rappelle que l'on peut écrire

$$\sqrt{n}(G_n - l_A\theta_A - l_B\theta_B) = \frac{1}{\sqrt{n}}M_n - \sqrt{n}\theta_A\left(\frac{N_{A,n}}{n} - l_A\right) - \sqrt{n}\theta_B\left(\frac{N_{B,n}}{n} - l_B\right)$$

Ici, on a $l_A = 1$, $l_B = 0$ et on a donc

$$\frac{1}{n}\langle M\rangle_n = \frac{N_{A,n}}{n}\theta_A(1-\theta_A) + \frac{N_{B,n}}{n}\theta_B(1-\theta_B) \xrightarrow[n\to+\infty]{p.s} \theta_A(1-\theta_A).$$

De plus, pour tout $n \geq 1$, $|M_n - M_{n-1}| \leq 1$ et la condition de Lindeberg est donc vérifiée, et on obtient

$$\sqrt{n}M_n \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \theta_A(1-\theta_A)).$$

Comme $l_A = 1$ et $l_B = 0$, on a

$$\sqrt{n}\theta_A\left(\frac{N_{A,n}}{n} - l_A\right) + \sqrt{n}\theta_B\left(\frac{N_{B,n}}{n} - l_B\right) = \theta_A\frac{N_{A,n}-n}{\sqrt{n}} + \theta_B\frac{N_{B,n}}{\sqrt{n}} = \frac{N_{B,n}}{\sqrt{n}}(\theta_B - \theta_A)$$

et il suffit d'avoir la vitesse de convergence de $N_{B,n}$ pour conclure. Plus précisément, il suffit de montrer que $\frac{C_n}{\sqrt{n}} \xrightarrow[n\to+\infty]{} 0$, i.e $C_n = o\left(\sqrt{n}\right)$. Supposons par l'absurde que ce n'est pas le cas, i.e il existe $\eta' > 0$ et $n_{\eta'}$ tels que pour tout $n \geq n_{\eta'}$, $C_n \geq \lceil \eta'\sqrt{n}\rceil$, on a alors pour tout $n \geq n_{\eta'}$,

$$C_n \geq \lceil \eta'\sqrt{n}\rceil \Leftrightarrow c_{\lceil \eta'\sqrt{n}\rceil} \leq n \Leftrightarrow \frac{n}{c_{\lceil \eta'\sqrt{n}\rceil}} \geq 1,$$

ce qui est en contradiction avec $c_n = o\left(n^2\right)$ et on a donc $\frac{N_{n,B}}{\sqrt{n}} \xrightarrow[n\to+\infty]{p.s} 0$. Avec des calculs analogues pour le cas où $\theta_A < \theta_B$, on obtient donc

$$\sqrt{n}(G_n - \max\{\theta_A, \theta_B\}) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2_{A,B}\right)$$

avec $\sigma^2_{A,B} = \theta_A(1-\theta_A)\mathbf{1}_{\theta_A>\theta_B} + \theta_B(1-\theta_B)\mathbf{1}_{\theta_B>\theta_A}$.

Application à l'estimation en ligne des quantiles : On reprend le cadre de l'estimation en ligne du quantile m d'ordre p. A l'aide du TLC pour les martingales, on peut montrer que (cf version longue)

$$\frac{1}{\sqrt{\gamma_n}}M_n \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{p(1-p)}{2c_\gamma f(m)}\right)$$

et on obtient la loi asymptotique asymptotique des estimateurs, i.e

$$\frac{1}{\sqrt{\gamma_n}}(m_n - m) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{p(1-p)}{2f(m)}\right),$$

ce que l'on peut réécrire

$$Q_n := \frac{\sqrt{2f(m)}}{c_\gamma \sqrt{p(1-p)}} n^{\alpha/2} (m_n - m) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

En effet, Figure 2.3, on s'intéresse à la densité de Q_n que l'on compare à celle d'une loi normale centrée réduite. La densité de Q_n est estimée à l'aide de 500 échantillons. On voit bien que les deux densités sont de plus en plus proches lorsque la taille d'échantillon n augmente.

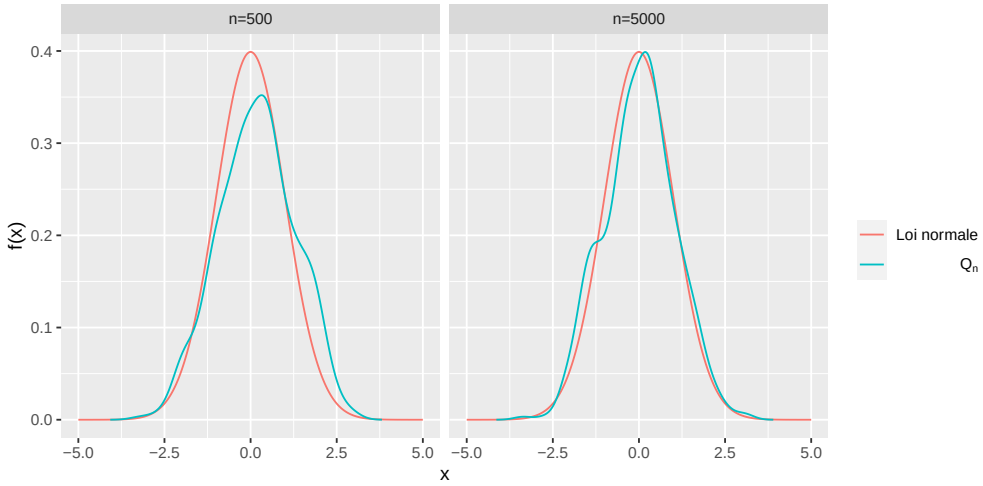

FIGURE 2.3 – Comparaison de la densité de Q_n, pour $n = 500$ (à gauche) et $n = 5000$ (à droite), et de la densité d'une loi normal centrée réduite.

2.2 Martingales vectorielles

Dans cette section, on s'intéresse aux vitesses de convergence des martingales vectorielles. Plus précisément, on établit, sous certaines conditions, les vitesses de convergence presque sûre ainsi que la normalité asymptotique des martingales.

2.2.1 Définition

Dans ce qui suit, on suppose que $M_n \in \mathbb{R}^d$ est de carré intégrable, i.e $\mathbb{E}\left[\|M_n\|^2\right] < +\infty$, où $\|.\|$ est la norme euclidienne.

Definition 2.2.1. *Soit $\mathcal{F} = (\mathcal{F}_n)$ une filtration, et (M_n) une suite adaptée à $\mathcal{F}$.*

— *(M_n) est une martingale de carré intégrable adaptée à $\mathcal{F}$ si*

$$\mathbb{E}\left[M_{n+1}|\mathcal{F}_n\right] = M_n.$$

— *La variation quadratique prévisible de (M_n), aussi appelée crochet, est le processus $\langle M \rangle = (\langle M \rangle_n)_n$*

défini par $\langle M\rangle_0 = M_0 M_0^T$ *et pour tout* $n \geq 1$, *par* $\langle M\rangle_n = \langle M\rangle_{n-1} + \Delta_n$, *avec*

$$\Delta_n = \mathbb{E}\left[(M_n - M_{n-1})(M_n - M_{n-1})^T |\mathcal{F}_{n-1}\right] = \mathbb{E}\left[M_n M_n^T - M_{n-1}M_{n-1}^T|\mathcal{F}_{n-1}\right]$$

Remarquons que l'on peut réécrire le crochet comme

$$\langle M\rangle_n = \langle M\rangle_0 + \sum_{k=1}^{n} \mathbb{E}\left[(M_n - M_{n-1})(M_n - M_{n-1})^T |\mathcal{F}_{n-1}\right].$$

En d'autres termes, en notant $\xi_n = M_n - M_{n-1}$ la différence de martingale (on peut remarquer que $\mathbb{E}[\xi_n|\mathcal{F}_{n-1}] = 0$), on a

$$\langle M\rangle_n = \langle M\rangle_0 + \sum_{k=1}^{n} \mathbb{E}\left[\xi_n \xi_n^T|\mathcal{F}_{n-1}\right]$$

et on peut donc voir le crochet, à division par n près, comme la moyenne des variances conditionnelles des différences de martingale, i.e comme la moyenne des variances conditionnelles entre les M_k et M_{k-1}.

2.2.2 Vitesses de convergence des martingales vectorielles

Le théorème suivant donne un premier résultat sur les vitesses de convergence des martingales vectorielles.

Théorème 2.2.1. *On considère* $M_n = \sum_{k=1}^n \xi_k$ *avec pour tout* $k \geq 1$, $\mathbb{E}[\xi_k|\mathcal{F}_{k-1}] = 0$. *On suppose également qu'il existe une variable aléatoire positive* C *telle que pour tout* $k \geq 1$, $\mathbb{E}\left[\|\xi_k\|^2 |\mathcal{F}_{k-1}\right] \leq C$. *On a alors pour tout* $\delta > 0$,

$$\left\|\frac{1}{n}M_n\right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) \quad p.s.$$

Démonstration. On pose $V_n = \frac{1}{(\ln n)^{1+\delta} n}\|M_n\|^2$, et comme M_n est une martingale on a

$$\begin{aligned}\mathbb{E}\left[\|V_{n+1}\|^2 |\mathcal{F}_n\right] &= \frac{1}{(n+1)\ln(n+1)^{1+\delta}}\|M_n\|^2 + \frac{1}{(n+1)\ln(n+1)^{1+\delta}}\mathbb{E}\left[\|\xi_{n+1}\|^2 |\mathcal{F}_n\right] \\ &\leq \frac{n(\ln n)^{1+\delta}}{(n+1)\ln(n+1)^{1+\delta}}V_n + \frac{1}{(n+1)\ln(n+1)^{1+\delta}}C\end{aligned}$$

et on obtient le résultat en appliquant le théorème de Robbins-Siegmund. □

On peut encore faire un peu mieux avec des hypothèses légèrement plus restrictives grâce au théorème de suivant (admis, voir la preuve du Théorème 4.3.16 dans [Duf97]).

Théorème 2.2.2. *Soit* $M_n = \sum_{k=1}^n \xi_k$, *avec* ξ_k *adapté à la filtration, de carré intégrable et* $\mathbb{E}[\xi_k|\mathcal{F}_{k-1}] = 0$. *Si il existe une matrice symétrique semi-définie positive* Γ *telle que*

$$\frac{1}{n}\langle M\rangle_n = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[\xi_k \xi_k^T|\mathcal{F}_{k-1}\right] \xrightarrow[n\to+\infty]{p.s} \Gamma,$$

et si il existe $\eta > 0$ *tel que pour tout* $\sup_k \mathbb{E}\left[\|\xi_k\|^{2+\eta} \,|\mathcal{F}_{k-1}\right] < +\infty$, *alors*

$$\left\|\frac{1}{n}M_n\right\|^2 = O\left(\frac{\ln n}{n}\right) \quad p.s.$$

On peut même faire encore mieux en terme de vitesse de convergence en utilisant la loi du log-itéré pour les martingales (cf version longue, ou bien [DST90, Duf90]).

2.2.3 Théorème Central Limite

On commence par donner une version relativement indigeste mais générale du théorème central limite pour les martingales vectorielles.

Théorème 2.2.3 (Théorème Limite Centrale)**.** *Soit* (M_n) *une martingale de carré intégrable et on suppose qu'il existe une suite croissante et divergente* a_n *et une matrice* Γ *telles que*

1. $a_n^{-1}\langle M\rangle_n \xrightarrow[n\to+\infty]{\mathbb{P}} \Gamma$,
2. *la condition de Lindeberg est satisfaite, i.e pour tout* $\epsilon > 0$,

$$a_n^{-1}\sum_{k=1}^{n}\mathbb{E}\left[\|M_k - M_{k-1}\|^2 \mathbf{1}_{\|M_k - M_{k-1}\|\geq \epsilon a_n^{1/2}}\right] \xrightarrow[n\to+\infty]{\mathbb{P}} 0.$$

Alors

$$a_n^{-1/2}M_n \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,\Gamma).$$

Là encore, l'énoncé du TLC est plutôt indigeste (notamment la condition de Lindeberg), et on s'intéressera donc plutôt à la version plus "soft" mais plus restrictive suivante :

Corollaire 2.2.1. *Soit* $M_n = \sum_{k=1}^{n}\xi_k$, *où* (ξ_n) *est une suite de différences de martingale adaptée à la filtration. On suppose que les hypothèses suivantes sont vérifiées :*

1. *Il existe une matrice* Γ *telle que*

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\left[\xi_k\xi_k^T|\mathcal{F}_{k-1}\right] \xrightarrow[n\to+\infty]{\mathbb{P}} \Gamma$$

2. *Il existe des constantes positives* $a > 2$ *et* C_a *telles que* $\mathbb{E}\left[\|\xi_k\|^a \,|\mathcal{F}_{k-1}\right] \leq C_a$.

Alors

$$\frac{1}{\sqrt{n}}M_n \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,\Gamma).$$

La preuve est disponible dans la version longue.

Chapitre 3

Vitesses de convergence des algorithmes de gradient stochastiques

3.1 Convergence presque sûre

On rappelle que l'on cherche à estimer le minimiseur m de la fonction convexe $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ définie pour tout $h \in \mathbb{R}^d$ par

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

avec $g : \mathcal{X} \times \mathbb{R}^d \longrightarrow \mathbb{R}$, et X une variable aléatoire à valeurs dans un espace mesurable $\mathcal{X}$. On suppose également que pour presque tout x, la fonction $g(x, .)$ est différentiable et considérant des variables aléatoires i.i.d $X_1, \ldots, X_n, X_{n+1}, \ldots$ de même loi que X et arrivant de manière séquentielle, l'algorithme de gradient stochastique est défini de manière récursive pour tout $n \geq 0$ par

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

avec γ_n une suite de pas positifs vérifiant

$$\sum_{n \geq 0} \gamma_{n+1} = +\infty \qquad \text{et} \qquad \sum_{n \geq 0} \gamma_{n+1}^2 < +\infty.$$

3.1.1 Approche directe

L'approche directe repose sur l'écriture récursive de $\|m_n - m\|^2$. Celle-ci nous permet, à l'aide du théorème de Robbins-Siegmund, d'obtenir la forte consistance des estimateurs.

Théorème 3.1.1 (Approche directe)**.** *On suppose que la fonction G est strictement convexe, i.e que pour tout $h \in \mathbb{R}^d$ tel que $h \neq m$*

$$\left\langle \nabla G\left(h\right), h - m \right\rangle > 0 \tag{3.1}$$

et qu'il existe une constante C telle que pour tout $h \in \mathbb{R}^d$,

$$\mathbb{E}\left[\|\nabla_h g(X,h)\|^2\right] \leq C\left(1+\|h-m\|^2\right), \tag{3.2}$$

alors

$$m_n \xrightarrow[n\to+\infty]{p.s} m.$$

Démonstration. On a

$$\|m_{n+1}-m\|^2 \leq \|m_n-m\|^2 - 2\gamma_{n+1}\langle \nabla_h g(X_{n+1},m_n), m_n-m\rangle + \gamma_{n+1}^2 \|\nabla_h g(X_{n+1},m_n)\|^2.$$

On considère la filtration $(\mathcal{F}_n)$ engendrée par l'échantillon, i.e $\mathcal{F}_n = \sigma(X_1,\dots,X_n)$. En passant à l'espérance conditionnelle, on obtient, comme m_n est $\mathcal{F}_n$-mesurable et par linéarité de l'espérance,

$$\begin{aligned}\mathbb{E}\left[\|m_{n+1}-m\|^2 |\mathcal{F}_n\right] &= \|m_n-m\|^2 - 2\gamma_{n+1}\langle \mathbb{E}\left[\nabla_h g(X_{n+1},m_n)|\mathcal{F}_n\right], m_n-m\rangle + \gamma_{n+1}^2 \mathbb{E}\left[\|\nabla_h g(X_{n+1},m_n)\|^2|\mathcal{F}_n\right] \\ &= \|m_n-m\|^2 - 2\gamma_n\langle \nabla G(m_n), m_n-m\rangle + \gamma_{n+1}^2 \mathbb{E}\left[\|\nabla_h g(X_{n+1},m_n)\|^2|\mathcal{F}_n\right].\end{aligned}$$

Grâce à l'inégalité (3.2), on obtient

$$\mathbb{E}\left[\|m_{n+1}-m\|^2|\mathcal{F}_n\right] \leq \left(1+C\gamma_{n+1}^2\right)\|m_n-m\|^2 - 2\gamma_{n+1}\langle \nabla G(m_n), m_n-m\rangle + C\gamma_{n+1}^2.$$

Comme $\sum_{n\geq 0}\gamma_{n+1}^2 C < +\infty$, en appliquant le théorème de Robbins-Siegmund, $\|m_n-m\|^2$ converge presque sûrement vers une variable aléatoire finie et

$$\sum_{n\geq 0}\gamma_{n+1}\langle \nabla G(m_n), m_n-m\rangle < +\infty \quad p.s.$$

Comme $\sum_{n\geq 1}\gamma_n = +\infty$, on a $\liminf_n \langle \nabla G(m_n), m_n-m\rangle = 0$ presque sûrement, et comme la fonction G est strictement convexe, cela implique que $\liminf_n \|m_n-m\| = 0$ presque sûrement. Ainsi, $\|m_n-m\|$ converge presque sûrement vers une variable aléatoire finie et sa limite inférieure est 0, donc cette suite converge presque sûrement vers 0. □

A noter que l'on a choisi une suite de pas déterministe. Cependant, la preuve précédente reste vraie si on prend une suite de pas aléatoires. Plus précisément, le théorème précédent reste vrai si on prend une suite de variables aléatoires Γ_{n+1} vérifiant

$$\sum_{n\geq 0}\Gamma_{n+1} = +\infty \quad p.s \qquad \text{et} \qquad \sum_{n\geq 0}\Gamma_{n+1}^2 < +\infty \quad p.s,$$

et telle que pour tout $n \geq 0$, Γ_{n+1} est $\mathcal{F}_n$-mesurable.

3.1.2 Approche via le développement de Taylor de la fonction G

On peut également utiliser une approche basée sur le développement de Taylor de la fonction G. Bien que cette approche nécessite des hypothèses beaucoup plus restrictives, on verra par la suite qu'elle est cruciale pour obtenir la convergence des algorithmes de Newton stochastiques.

Théorème 3.1.2. *On suppose que m est l'unique minimiseur de G et l'unique zéro du gradient. On suppose également qu'il existe des constantes C, C' telles que pour tout $h \in \mathbb{R}^d$,*

$$\left\|\nabla^2 G(h)\right\|_{op} \leq C \qquad et \qquad \mathbb{E}\left[\left\|\nabla_h g\left(X, h\right)\right\|^2\right] \leq C'\left(1 + G(h) - G(m)\right).$$

Alors m_n converge presque sûrement vers m.

Démonstration. A l'aide d'un développement de Taylor de la fonction G, il existe $h \in [m_n, m_{n+1}]$ tel que

$$\begin{aligned} G\left(\theta_{n+1}\right) &= G\left(\theta_n\right) + \left\langle \nabla G\left(m_n\right), m_{n+1} - m_n \right\rangle + \frac{1}{2}\left\langle m_{n+1} - m_n, \nabla^2 G(h)\left(m_{n+1} - m_n\right)\right\rangle \\ &= G\left(\theta_n\right) - \gamma_{n+1}\left\langle \nabla G\left(m_n\right), \nabla_h g\left(X_{n+1}, m_n\right)\right\rangle + \frac{1}{2}\gamma_{n+1}^2\left\langle \nabla_h g\left(X_{n+1}, m_n\right), \nabla^2 G(h)\nabla_h g\left(X_{n+1}, m_n\right)\right\rangle \end{aligned}$$

De plus, comme $\left\|\nabla^2 G(h)\right\|_{op} \leq C$, on obtient

$$\begin{aligned} G\left(m_{n+1}\right) &\leq G\left(m_n\right) - \gamma_{n+1}\left\langle \nabla G\left(m_n\right), \nabla_h g\left(X_{n+1}, m_n\right)\right\rangle + \frac{1}{2}\gamma_{n+1}^2\left\|\nabla^2 G(h)\right\|_{op}\left\|\nabla_h g\left(X_{n+1}, m_n\right)\right\|^2 \\ &\leq G\left(m_n\right) - \gamma_{n+1}\left\langle \nabla G\left(m_n\right), \nabla_h g\left(X_{n+1}, m_n\right)\right\rangle + \frac{1}{2}\gamma_{n+1}^2 C\left\|\nabla_h g\left(X_{n+1}, m_n\right)\right\|^2 \end{aligned}$$

Attention ! On n'a pas supposé que la fonction G est positive, et on ne peut pas espérer appliquer directement le théorème de Robbins-Siegmund. Cependant, on peut réécrire l'inégalité précédente comme

$$G\left(m_{n+1}\right) - G(m) \leq G\left(m_n\right) - G(m) - \gamma_{n+1}\left\langle \nabla G\left(m_n\right), \nabla_h g\left(X_{n+1}, m_n\right)\right\rangle + \frac{1}{2}C\gamma_{n+1}^2\left\|\nabla_h g\left(X_{n+1}, m_n\right)\right\|^2$$

Pour tout $n \geq 0$, on note $V_n = G\left(m_n\right) - G(m)$ et $\mathcal{F}_n = \sigma\left(X_1, \ldots, X_n\right)$. En passant à l'espérance conditionnelle, on obtient, comme m_n est $\mathcal{F}_n$-mesurable,

$$\begin{aligned} \mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] &\leq V_n - \left\langle \nabla G\left(m_n\right), \mathbb{E}\left[\nabla_h g\left(X_{n+1}, m_n\right)|\mathcal{F}_n\right]\right\rangle + \frac{1}{2}C\gamma_{n+1}^2\mathbb{E}\left[\left\|\nabla_h g\left(X_{n+1}, m_n\right)\right\|^2|\mathcal{F}_n\right] \\ &\leq \left(1 + \frac{1}{2}CC'\gamma_{n+1}^2\right)V_n - \gamma_{n+1}\left\|\nabla G\left(m_n\right)\right\|^2 + \frac{1}{2}CC'\gamma_{n+1}^2 \end{aligned}$$

Comme par définition de m, V_n est positif et grâce au théorème de Robbins-Siegmund, V_n converge

presque sûrement vers une variable aléatoire finie et

$$\sum_{n\geq 1} \gamma_{n+1} \left\|\nabla G\left(m_n\right)\right\|^2 < +\infty \quad p.s.$$

Comme $\sum_{n\geq 1} \gamma_{n+1} = +\infty$, on a $\liminf_n \left\|\nabla G\left(m_n\right)\right\| = 0$ presque sûrement, et comme m est l'unique zéro du gradient, $\liminf_n \left\|m_n - m\right\| = 0$ presque sûrement. Ainsi, $\liminf_n V_n = 0$ presque sûrement et donc V_n converge presque sûrement vers 0, ce qui implique, comme m est l'unique minimiseur de la fonction G, que m_n converge presque sûrement vers m. □

3.1.3 Approche Lyapunov

En réalité, les approches précédentes peuvent être "généralisées" aux fonctions Lyapunov $V : \mathbb{R}^d \longrightarrow \mathbb{R}_+$.

Théorème 3.1.3. *On suppose qu'il existe une fonction $V : \mathbb{R}^d \longrightarrow \mathbb{R}$ vérifiant*

- *$V(m) = 0$ et pour tout $h \neq m$, $V(m) > 0$.*
- *V est continument différentiable et à gradient L-lipschitz, i.e pour tout $h, h' \in \mathbb{R}^d$,*

$$\left\|\nabla V(h) - \nabla V(h')\right\| \leq L \left\|h - h'\right\|.$$

- *Il existe une constante positive C telle que pour tout $h \in \mathbb{R}^d$,*

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

- *Il existe $\alpha > 0$ tel que pour tout $h \in \mathbb{R}^d$,*

$$\langle \nabla G(h), \nabla V(h)\rangle \geq \alpha V(h),$$

alors

$$m_n \xrightarrow[n\to+\infty]{p.s} m.$$

Démonstration. En regardant le développement de Taylor de la fonction V à l'ordre 1, on obtient

$$\begin{aligned} V\left(m_{n+1}\right) &= V\left(m_n\right) + \left\langle \int_0^1 \nabla V\left(m_{n+1} + t\left(m_n - m_{n+1}\right)\right) dt, m_{n+1} - m_n \right\rangle \\ &= V\left(m_n\right) + \left\langle \nabla V\left(m_n\right), m_{n+1} - m_n \right\rangle + \left\langle \int_0^1 \nabla V\left(m_{n+1} + t\left(m_n - m_{n+1}\right)\right) - \nabla V\left(m_n\right) dt, m_{n+1} - m_n \right\rangle \end{aligned}$$

En utilisant l'inégalité de Cauchy-Schwarz et comme le gradient de V est L-lipschitz, on obtient

$$\begin{aligned} V(m_{n+1}) &\leq V(m_n) + \langle \nabla V(m_n), m_{n+1} - m_n \rangle + \int_0^1 \|\nabla V(m_{n+1} + t(m_n - m_{n+1})) - \nabla V(m_n)\| \, dt \, \|m_{n+1} - m_n\| \\ &\leq V(m_n) + \langle \nabla V(m_n), m_{n+1} - m_n \rangle + L \int_0^1 (1-t) dt \, \|m_{n+1} - m_n\|^2 \\ &= V(m_n) + \langle \nabla V(m_n), m_{n+1} - m_n \rangle + \frac{L}{2} \|m_{n+1} - m_n\|^2 \end{aligned}$$

Ainsi, en remplaçant m_{n+1} et en passant à l'espérance conditionnelle, on obtient

$$\begin{aligned} \mathbb{E}[V(m_{n+1}) | \mathcal{F}_n] &\leq V(m_n) - \gamma_{n+1} \langle \nabla V(m_n), \nabla G(m_n) \rangle + \frac{L}{2} \gamma_{n+1}^2 \mathbb{E}\left[\|\nabla_h g(X_{n+1}, m_n)\|^2 | \mathcal{F}_n \right] \\ &\leq \left(1 + \frac{LC}{2} \gamma_{n+1}^2\right) V(m_n) - \gamma_{n+1} \alpha V(m_n) + \frac{LC}{2} \gamma_n^2. \end{aligned}$$

En appliquant le théorème de Robbins-Siegmund on obtient que $V(m_n)$ converge presque sûrement vers une variable aléatoire finie et que

$$\sum_{n \geq 0} \gamma_{n+1} V(m_n) < +\infty \quad p.s$$

et on peut alors conclure de la même façon que pour les théorèmes précédents. □

3.1.4 Application au modèle linéaire

On se place dans le cadre du modèle linéaire défini par (1.1), et on rappelle que l'on cherche à minimiser la fonction $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ définie pour tout $h \in \mathbb{R}^d$,

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

Le théorème suivant nous donne la forte consistance des estimateurs de gradient dans le cas du modèle linéaire.

Théorème 3.1.4. *On suppose que X admet un moment d'ordre 4, que ϵ admet un moment d'ordre 2 et que la matrice $\mathbb{E}[XX^T]$ est définie positive. Alors, les estimateurs de gradient stochastiques définis par (1.5) vérifient*

$$\theta_n \xrightarrow[n \to +\infty]{p.s} \theta.$$

Démonstration. On rappelle que pour tout $h \in \mathbb{R}^d$, on a $\nabla^2 G(h) = \mathbb{E}[XX^T]$ qui est supposée définie positive. De plus, pour tout $h \in \mathbb{R}^d$ tel que $h \neq \theta$, on a, à l'aide d'un développement de Taylor du gradient,

$$\nabla G(h) = \int_0^1 \nabla^2 G(\theta + t(h - \theta) dt \, (h - \theta) = \mathbb{E}\left[XX^T \right] (h - \theta)$$

et on a donc, comme $\mathbb{E}\left[XX^T\right]$ est définie positive,

$$\langle \nabla G(h), h-\theta \rangle = (h-\theta)^T \mathbb{E}\left[XX^T\right](h-\theta) > 0.$$

De plus, on a pour tout $h \in \mathbb{R}^d$,

$$\begin{aligned}\|\nabla_h g(X,Y,h)\|^2 = \left\|\left(Y - X^T h\right) X\right\|^2 &= \left(\underbrace{\left(Y - X^T\theta\right)}_{=\epsilon} + X^T(\theta - h)\right)^2 \|X\|^2 \\ &\leq 2\epsilon^2 \|X\|^2 + 2\|X\|^4 \|\theta - h\|^2\end{aligned}$$

Et donc comme X admet un moment d'ordre 4, que ϵ admet un moment d'ordre 2, et comme X, ϵ sont indépendants, on a

$$\mathbb{E}\left[\|\nabla_h g(X,Y,h)\|^2\right] \leq 2\mathbb{E}\left[\epsilon^2\right]\mathbb{E}\left[\|X\|^2\right] + 2\mathbb{E}\left[\|X\|^4\right]\|h-\theta\|^2$$

et les hypothèses du Théorème 3.1.1 sont donc vérifiées, i.e θ_n converge presque sûrement vers θ. □

3.1.5 Application à la régréssion logistique

On se place dans le cadre de la régression logistique (1.2), et on rappelle que l'on cherche à minimiser

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

Le théorème suivant donne la forte consistance des estimateurs de gradient stochastique dans le cadre de la régression logistique.

Théorème 3.1.5. *On suppose que la variable aléatoire X admet un moment d'ordre 2 et que la Hessienne de G en θ est positive. Alors, les estimateurs (θ_n) de gradient stochastiques définis par (1.6) vérifient*

$$\theta_n \xrightarrow[n\to+\infty]{p.s} \theta.$$

Démonstration. Pour tout $h \in \mathbb{R}^d$, on a

$$\|\nabla_h g(X,Y,h)\| = \left\|\left(\pi\left(h^T X\right) - Y\right) X\right\| \leq \left|\pi\left(h^T X\right) - Y\right| \|X\| \leq \|X\|.$$

On a donc

$$\mathbb{E}\left[\|\nabla_h g(X,Y,h)\|^2\right] \leq \mathbb{E}\left[\|X\|^2\right].$$

De plus, à l'aide d'un développement de Taylor, on a pour tout h

$$\nabla G(h) = \int_0^1 \nabla^2 G\left(\theta + t(h-\theta)\right) dt\,(h-\theta).$$

De plus, la fonction π est continue et comme X admet un moment d'ordre 2, l'application

$$h \longmapsto \nabla^2 G(h) = \mathbb{E}\left[\pi\left(h^T X\right)\left(1-\pi\left(h^T X\right)\right) XX^T\right]$$

est continue. De plus, comme la Hessienne de G est positive en θ, en notant $\lambda_{\min}$ sa plus petite valeurs propre, il existe $r_\theta > 0$ tel que pour tout $h \in \mathcal{B}(\theta, r_\theta)$,

$$\lambda_{\min}\left(\nabla^2 G(h)\right) \geq \frac{\lambda_{\min}}{2}.$$

Comme la Hessienne de G est au moins semi-définie positive, pour tout $h \neq m$,

$$\begin{aligned}
\langle \nabla G(h), h-\theta\rangle &= \left\langle \int_0^1 \nabla^2 G(\theta + t(\theta-h))dt(h-\theta), h-\theta \right\rangle \\
&= \int_0^1 \left\langle \nabla^2 G(\theta + t(h-\theta))(h-\theta), h-\theta\right\rangle dt \\
&\geq \int_0^1 \lambda_{\min}\left(\nabla^2 G(\theta + t(h-\theta))\right) \|h-\theta\|^2 dt
\end{aligned}$$

Si $h \in \mathcal{B}(\theta, r_\theta)$, on a

$$\langle \nabla G(h), h-\theta\rangle \geq \int_0^1 \frac{\lambda_{\min}}{2}\|h-\theta\|^2 dt = \frac{\lambda_{\min}}{2}\|h-\theta\|^2.$$

Si $\|h-\theta\| \geq r_\theta$, on a

$$\langle \nabla G(h), h-\theta\rangle \geq \int_0^{\frac{r_\theta}{\|h-\theta\|}} \lambda_{\min}\left(\nabla^2 G(\theta + t(h-\theta))\right)\|h-\theta\|^2 dt \geq \int_0^{\frac{r_\theta}{\|h-\theta\|}} \frac{\lambda_{\min}}{2}\|h-\theta\|^2 dt = \frac{r_\theta \lambda_{\min}}{2}\|h-\theta\|.$$

Ainsi, si $h \neq \theta$,

$$\langle \nabla G(h), h-\theta\rangle > 0$$

et les hypothèses du théorème sont donc vérifiées.

□

3.2 Vitesses de convergence presque sûre

On s'intéresse maintenant à la vitesse de convergence des estimateurs obtenus à l'aide de l'algorithme de gradient stochastique. Pour cela, on suppose maintenant que la suite de pas (γ_n) vérifie $\gamma_n = c_\gamma n^{-\alpha}$ avec $c_\gamma > 0$ et $\alpha \in (1/2, 1)$.

3.2.1 Cadre

Afin d'obtenir les vitesses de convergence presque sûre des estimateurs, on suppose maintenant que la fonction G que l'on cherche à minimiser est différentiable, convexe, et qu'il existe m qui soit

un zéro du gradient. De plus, on suppose que les hypothèses suivantes sont vérifiées :

(PS1) Il existes des constantes positives $\nu > \frac{1}{\alpha} - 1$ et C_ν telles que pour tout $h \in \mathbb{R}^d$,

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

(PS2) La fonction G est deux fois continûment différentiable sur un voisinage de m et

$$\lambda_{\min} := \lambda_{\min}\left(\nabla^2 G(m)\right) > 0.$$

A noter que l'hypothèse **(PS1)** est vérifiée, par exemple, dès que $\nabla_h g(X,.)$ admet un moment d'ordre 4 tandis que l'hypothèse **(PS2)** implique la stricte convexité (et même forte convexité locale) de la fonction G. A noter que les hypothèses pour obtenir les vitesses de convergence presque sûre sont beaucoup moins restrictives (de manière générale) que celles pour obtenir la convergence en moyenne quadratique.

3.2.2 Vitesses de convergence

Le théorème suivant nous donne alors la vitesse de convergence presque sûre des estimateurs.

Théorème 3.2.1. *On suppose que les hypothèses **(PS1)** et **(PS2)** sont vérifiées. Alors*

$$\|m_n - m\|^2 = O\left(\frac{\ln n}{n^\alpha}\right) \quad p.s.$$

Démonstration. A noter que les hypothèses du Théorème 3.1.1 sont vérifiées et que l'on a donc

$$m_n \xrightarrow[n\to+\infty]{p.s} m.$$

N'ayant pas forte convexité de la fonction G, on ne peut pas utiliser l'approche brutale de la preuve du Théorème 1.4.1, mais on va s'en approcher en linéarisant le gradient. Plus précisément, rappelons que l'on peut réécrire l'algorithme comme

$$m_{n+1} - m = m_n - m - \gamma_{n+1}\nabla G(m_n) + \gamma_{n+1}\xi_{n+1}$$

avec $\xi_{n+1} = \nabla G(m_n) - \nabla_h g(X_{n+1}, m_n)$. Rappelons également que considérant la filtration $\mathcal{F} = (\mathcal{F}_n)$, (ξ_n) est une suite de différences de martingale. En linéarisant le gradient, on obtient

$$\begin{aligned} m_{n+1} - m &= m_n - m - \gamma_{n+1}\nabla^2 G(m)(m_n - m) + \gamma_{n+1}\xi_{n+1} + \gamma_{n+1}\nabla^2 G(m)(m_n - m) - \gamma_{n+1}\nabla G(m_n) \\ &= \left(I_d - \gamma_{n+1}\nabla^2 G(m)\right)(m_n - m) + \gamma_{n+1}\xi_{n+1} - \gamma_{n+1}\delta_n \end{aligned} \tag{3.3}$$

où $\delta_n = \nabla G(m_n) - \nabla^2 G(m)(m_n - m)$ est le terme de reste dans la décomposition de Taylor du gradient. De plus, grâce à l'hypothèse **(PS2)**, il existe un voisinage V_m de m telle que G soit deux

fois continûment différentiable sur V_m et donc, pour tout $h \in V_m$

$$\begin{aligned}\left\|\nabla G\left(h\right) - \nabla^2 G\left(m\right)\left(h-m\right)\right\| &= \left\|\int_0^1 \nabla^2 G\left(m + t(h-m)\right) dt \left(h-m\right) + \nabla^2 G(m)(h-m)\right\| \\ &\leq \int_0^1 \left\|\nabla^2 G(m + t(h-m)) - \nabla^2 G(m)\right\|_{op} dt \left\|h-m\right\|\end{aligned}$$

Par continuité et comme m_n converge presque sûrement vers m, on a donc

$$\left\|\delta_n\right\| = o\left(\left\|m_n - m\right\|\right) \quad p.s.$$

De plus, grâce à la décomposition (3.3) on peut montrer par récurrence que l'on peut réécrire $m_n - m$ comme

$$m_n - m = \beta_{n,0}\left(m_0 - m\right) + \sum_{k=0}^{n-1} \beta_{n,k+1}\gamma_{k+1}\xi_{k+1} - \sum_{k=0}^{n-1} \beta_{n,k+1}\gamma_{k+1}\delta_k \tag{3.4}$$

avec pour tout $k, n \geq 0$ tel que $k \leq n$,

$$\beta_{n,k} = \prod_{j=k+1}^{n} \left(I_d - \gamma_j \nabla^2 G(m)\right) \qquad \text{et} \qquad \beta_{n,n} = I_d.$$

Cette décomposition est clairement vérifiée pour $n = 0$. De plus, pour $n+1$, en utilisant la décomposition (3.3) et par hypothèse de récurrence, on a

$$\begin{aligned}m_{n+1} - m &= \left(I_d - \gamma_{n+1}\nabla^2 G(m)\right)\left(m_n - m\right) + \gamma_{n+1}\xi_{n+1} - \gamma_{n+1}\delta_n \\ &= \left(I_d - \gamma_{n+1}\nabla^2 G(m)\right)\left(\beta_{n,0}\left(m_0 - m\right) + \sum_{k=0}^{n-1} \beta_{n,k+1}\gamma_{k+1}\xi_{k+1} - \sum_{k=0}^{n-1} \beta_{n,k+1}\gamma_{k+1}\delta_k\right) + \gamma_{n+1}\xi_{n+1} - \gamma_{n+1}\delta_n.\end{aligned}$$

En remarquant que que pour tout $k \leq n-1$, $\left(I_d - \gamma_{n+1}\nabla^2 G\left(m\right)\right)\beta_{n,k+1} = \beta_{n+1,k+1}$, on obtient

$$m_{n+1} - m = \beta_{n+1,0}\left(m_0 - m\right) + \sum_{k=0}^{n-1} \beta_{n+1,k+1}\gamma_{k+1}\xi_{k+1} - \sum_{k=0}^{n-1} \beta_{n+1,k+1}\gamma_{k+1}\delta_k + \gamma_{n+1}\xi_{n+1} - \gamma_{n+1}\delta_n$$

Comme $\gamma_{n+1}\xi_{n+1} = \beta_{n+1,n+1}\gamma_{n+1}\xi_{n+1}$ et que l'on peut faire de même pour δ_n, on obtient

$$\begin{aligned}m_{n+1} - m &= \beta_{n+1,0}\left(m_0 - m\right) + \sum_{k=0}^{n-1} \beta_{n+1,k+1}\gamma_{k+1}\xi_{k+1} - \sum_{k=0}^{n-1} \beta_{n+1,k+1}\gamma_{k+1}\delta_k + \gamma_{n+1}\beta_{n+1,n+1}\xi_{n+1} \\ &\quad - \gamma_{n+1}\beta_{n+1,n+1}\delta_n \\ &= \beta_{n+1,0}\left(m_0 - m\right) + \sum_{k=0}^{n} \beta_{n+1,k+1}\gamma_{k+1}\xi_{k+1} - \sum_{k=0}^{n} \beta_{n+1,k+1}\gamma_{k+1}\delta_k\end{aligned}$$

et la décomposition (3.4) est donc exacte. De plus, en remarquant que l'on peut réécrire $\beta_{n,k} =$

$\beta_{n,0}\beta_{k,0}^{-1}$, on peut réécrire

$$\sum_{k=0}^{n-1} \beta_{n,k+1}\gamma_{k+1}\xi_{k+1} = \beta_n \sum_{k=0}^{n-1} \beta_{k+1}^{-1}\xi_{k+1} =: \beta_n M_n$$

et M_n est alors un terme de martingale. Cependant, il est compliqué d'utiliser directement la loi des grands nombres ou la loi du log itéré pour les martingales vectorielles directement pour ce terme. On peut par contre réécrire la martingale dans la base orthonormée de la Hessienne et d'utiliser la loi des grands nombres pour les martingales réelles à chacune de ces coordonnées (approche développée dans [Pel98] par exemple)ce qui conduit au lemme suivant :

Lemma 3.2.1. *On suppose que les hypothèses **(PS1)** et **(PS2)** sont vérifiées. Alors*

$$\left\| \sum_{k=0}^{n-1} \beta_{n,k+1}\gamma_{k+1}\xi_{k+1} \right\|^2 = O\left(\frac{\ln n}{n^\alpha}\right) \quad p.s.$$

On admettra ce lemme. Cependant, la preuve basée sur la loi du log itérée développée dans [Pel98], ainsi qu'une autre preuve basée sur l'obtention d'inégalités exponentielles sont disponible dans la version longue. Revenons maintenant à la preuve du théorème. Maintenant que l'on a donné la vitesse de convergence du terme de martingale, on peut s'intéresser à la vitesse de convergence du terme dû à l'erreur d'initialisation, i.e du terme $\beta_{n,0}(m_0 - m)$. Rappelons qu'il existe n_0 tel que pour tout $n \geq n_0$, $\|I_d - \gamma_{n+1}H\|_{op} \leq 1 - \lambda_{\min}\gamma_{n+1}$. On a donc, comme $1 + x \leq \exp(x)$,

$$\begin{aligned}\|\beta_{n,0}\|_{op} \leq \prod_{k=1}^{n} \|I_d - \gamma_k H\|_{op} &\leq \prod_{k=1}^{n_0} \|I_d - \gamma_k H\|_{op} \prod_{k=n_0+1}^{n} (1 - \gamma_k\lambda_{\min}) \\ &\leq \prod_{k=1}^{n_0} \|I_d - \gamma_k H\|_{op} \exp\left(-\lambda_{\min} \sum_{k=n_0+1}^{n} \gamma_k\right)\end{aligned}$$

et ce terme converge donc à vitesse exponentielle, et en particulier, couplé avec le Lemme 3.2.1, il vient

$$\left\| \beta_{n,0}(m_0 - m) + \sum_{k=0}^{n-1} \beta_{n,k+1}\gamma_{k+1}\xi_{k+1} \right\|^2 = O(\gamma_{n+1}\ln n) \quad p.s$$

et il existe donc une variable aléatoire A telle que pour tout $n \geq 0$,

$$\left\| \beta_{n,0}(m_0 - m) + \sum_{k=0}^{n-1} \beta_{n,k+1}\gamma_{k+1}\xi_{k+1} \right\|^2 \leq A\gamma_{n+1}\ln(n+1) \quad p.s.$$

On va maintenant se concentrer sur le dernier terme, i.e sur

$$\Delta_n = \sum_{k=0}^{n-1} \beta_{n,k+1}\gamma_{k+1}\delta_k.$$

Rappelons que l'on a vu que pour tout $n \geq 0$, $\|\delta_n\| = o\left(\|m_n - m\|\right)$ $p.s$. On a donc

$$\begin{aligned}\|\Delta_{n+1}\| &= \|(I_d - \gamma_{n+1}H)\,\Delta_n + \gamma_{n+1}\delta_n\| \leq (1 - \lambda_{\min}\gamma_{n+1})\,\|\Delta_n\| + \gamma_{n+1}\,\|\delta_n\| \\ &\leq (1 - \lambda_{\min}\gamma_{n+1})\,\|\Delta_n\| + \gamma_{n+1}r_n\,\|m_n - m\| \end{aligned} \tag{3.5}$$

avec $r_n := \frac{\|\delta_n\|}{\|m_n - m\|}\mathbf{1}_{\|m_n - m\| \neq 0}$ qui converge presque sûrement vers 0. De plus, par inégalité triangulaire, $\|m_n - m\| \leq \left\|\beta_{n,0}\,(m_0 - m) + \sum_{k=0}^{n-1}\beta_{n,k+1}\gamma_{k+1}\xi_{k+1}\right\| + \|\Delta_n\|$. On peut donc réécrire l'inégalité (3.5), pour tout $n \geq n_0$, comme

$$\|\Delta_{n+1}\| \leq (1 - \lambda_{\min}\gamma_{n+1})\,\|\Delta_n\| + r_n\gamma_{n+1}\left(\left\|\beta_{n,0}\,(m_0 - m) + \sum_{k=0}^{n-1}\beta_{n,k+1}\gamma_{k+1}\xi_{k+1}\right\| + \|\Delta_n\|\right)$$

Comme $\left\|\beta_{n,0}\,(m_0 - m) + \sum_{k=0}^{n-1}\beta_{n,k+1}\gamma_{k+1}\xi_{k+1}\right\| \leq \sqrt{A}\frac{\sqrt{\ln n}}{n^{\alpha/2}}$ presque sûrement et en appliquant le Lemme 1.4.2, on obtient donc

$$\|\Delta_n\| = O\left(\frac{\sqrt{\ln n}}{n^{\alpha/2}}\right) \quad p.s,$$

ce qui conclut la preuve. □

3.2.3 Application au modèle linéaire

On se place dans le cadre du modèle linéaire défini par (1.1). Le théorème suivant donne la vitesse de convergence presque sûre des estimateurs obtenus grâce à l'algorithme de gradient stochastique.

Théorème 3.2.2. *Soit $\eta > \frac{1}{\alpha} - 1$ tel que que X admette un moment d'ordre $4 + 4\eta$ et tel que ϵ admette un moment d'ordre $2 + 2\eta$. On suppose également que la matrice $\mathbb{E}\left[XX^T\right]$ est définie positive. Alors les estimateurs de gradient définis par (1.5) vérifient*

$$\|\theta_n - \theta\|^2 = O\left(\frac{\ln n}{n^{\alpha}}\right) \quad p.s.$$

Démonstration. On a vu dans la preuve du Théorème 3.1.4 que l'hypothèse **(PS2)** est vérifiée. De plus, comme (voir la preuve du Théorème 3.1.4 pour plus de détails)

$$\|\nabla_h g\,(X, Y, h)\| \leq |\epsilon|\,\|X\| + \|X\|^2\,\|h - \theta\|$$

on a donc

$$\mathbb{E}\left[\|\nabla_h g\,(X, Y, h)\|^{2+2\eta}\right] = 2^{1+2\eta}\mathbb{E}\left[|\epsilon|^{2+2\eta}\right]\mathbb{E}\left[\|X\|^{2+2\eta}\right] + \mathbb{E}\left[\|X\|^{4+4\eta}\right]\|h - \theta\|^{2+2\eta}$$

et l'hypothèse **(PS1)** est donc vérifée, ce qui conclut la preuve. □

Dans la Figure 3.1, on s'intéresse à l'évolution de l'erreur quadratique des estimateurs de gradient

du paramètre de la régression linéaire en fonction de la taille d'échantillon n. Pour cela, on considère $\theta = (-4,-3,-2,-1,0,1,2,3,4,5)^T \in \mathbb{R}^{10}$, et on prend $X \sim \mathcal{N}(0, I_{10})$ et $\epsilon \sim \mathcal{N}(0,1)$. De plus, on a choisi $c_\gamma = 1$ et $\alpha = 0.5, 0.66, 0.75$ ou 1. On voit bien que l'on a une décroissance assez rapide de l'erreur quadratique, et lorsque l'on prend l'échelle logarithmique, une heuristique de pente nous donne que pour n suffisamment grand, on a bien une pente proche de $-\alpha$. Enfin, on peut remarquer que plus α est grand, plus l'erreur semble stable.

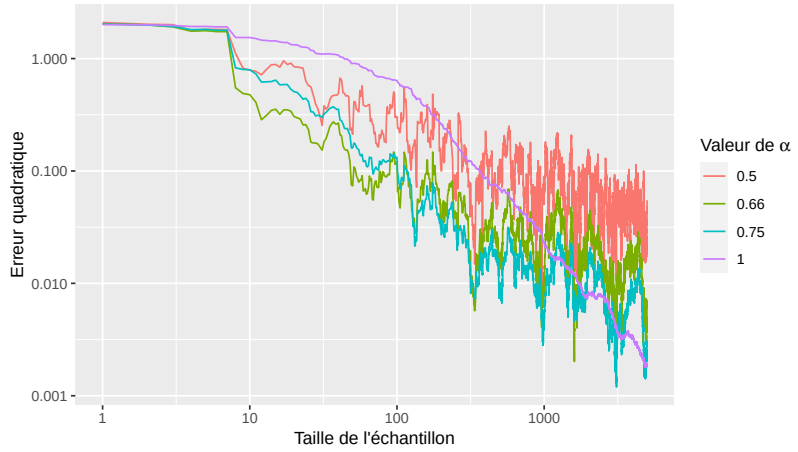

FIGURE 3.1 – Evolution de l'erreur quadratique de θ_n en fonction de la taille d'échantillon n dans le cadre de la régression linéaire.

Dans la figure 3.2, afin de mieux visualiser les pentes, on considère l'erreur quadratique moyenne des estimateurs en générant 50 échantillons de taille $n = 5000$. A noter que bien que l'erreur quadratique moyenne est meilleure pour $\alpha = 1$, les estimateurs semblent dans ce cas beaucoup plus sensibles à une possible mauvaise initialisation.

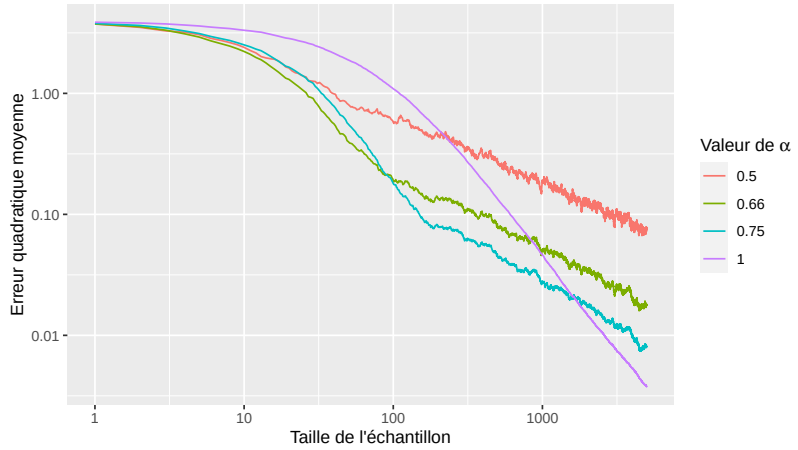

FIGURE 3.2 – Evolution de l'erreur quadratique moyenne de θ_n en fonction de la taille d'échantillon n et du choix de α dans le cadre de la régression linéaire.

3.2.4 Application à la régression logistique

On se replace dans le cadre de la régression logistique défini par (1.2). On rappelle que le paramètre θ est un minimiseur de la fonction $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ définie pour tout $h \in \mathbb{R}^d$ par

$$G(h) = \mathbb{E}\left[\log\left(1 + \exp\left(h^T X\right)\right) - h^T XY\right] =: \mathbb{E}\left[g\left(X, Y, h\right)\right].$$

Le théorème suivant donne la vitesse de convergence des estimateurs obtenus via l'algorithme de gradient stochastique dans le cadre de la régression logistique.

Théorème 3.2.3. *On suppose qu'il existe $\eta > 0$ tel que X admette un moment d'ordre $2 + 2\eta$. On suppose également que $\nabla^2 G(\theta)$ est inversible. Alors les estimateurs de gradient définis par (1.6) vérifient*

$$\|\theta_n - \theta\|^2 = O\left(\frac{\ln n}{n^\alpha}\right) \quad p.s.$$

Démonstration. Il faut montrer que les hypothèses **(PS1)** et **(PS2)** sont vérifiées.

Vérification de (PS2). A noter que θ est bien un zéro du gradient. De plus comme la fonction $x \longmapsto \pi(x)(1 - \pi(x))$ est continue et bornée, et comme X admet un moment d'ordre 2, la fonction G est bien deux fois continument différentiable sur $\mathbb{R}^d$ (et donc au voisinage de θ). Enfin, comme $\nabla^2 G(\theta)$ est inversible, on a bien $\lambda_{\min} > 0$, et l'hypothèse **(PS2)** est vérifiée.

Vérification de (PS1). Comme $\|\nabla_h g\left(X, Y, h\right)\| \leq \|X\|$ et comme X admet un moment d'ordre $2 + 2\eta$, l'hypothèse **(PS1)** est vérifiée.

□

Dans la Figure 3.3, on considère $\theta = (1, 1, 1, 1, 1)^T \in \mathbb{R}^5$, et on prend $X \sim \mathcal{N}(0, I_5)$. De plus, on a pris $c_\gamma = 1$ et $\alpha = 0.5, 0.66, 0.75$ ou 1. On voit bien que l'on a une décroissance assez rapide de l'erreur quadratique, sauf pour $\alpha = 1$.

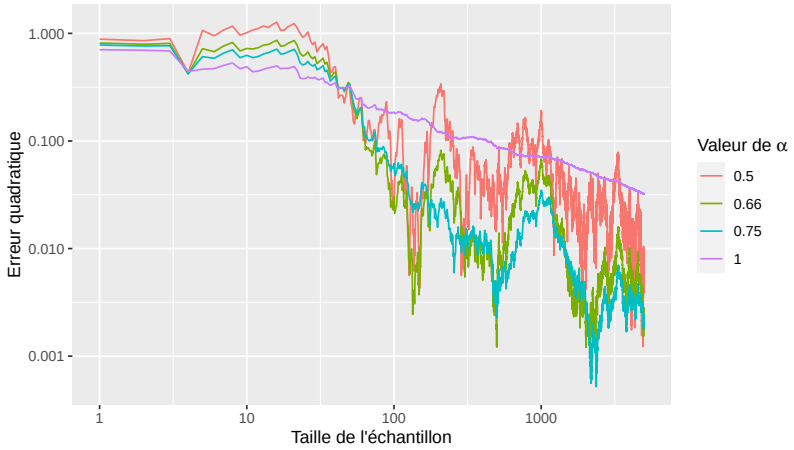

FIGURE 3.3 – Evolution de l'erreur quadratique de θ_n en fonction de la taille de l'échantillon n dans le cadre de la régression logistique.

Dans la figure 3.4, on s'intéresse à l'évolution de l'erreur quadratique moyenne afin de mieux visualiser les pentes. Cela semble confirmer que choisir $\alpha = 1$ n'est pas une très bonne option. Les estimateurs ne semblent pas du tout converger à la vitesse $1/n$. Pour les autres choix, on voit bien que plus α est grand, plus la convergence est rapide.

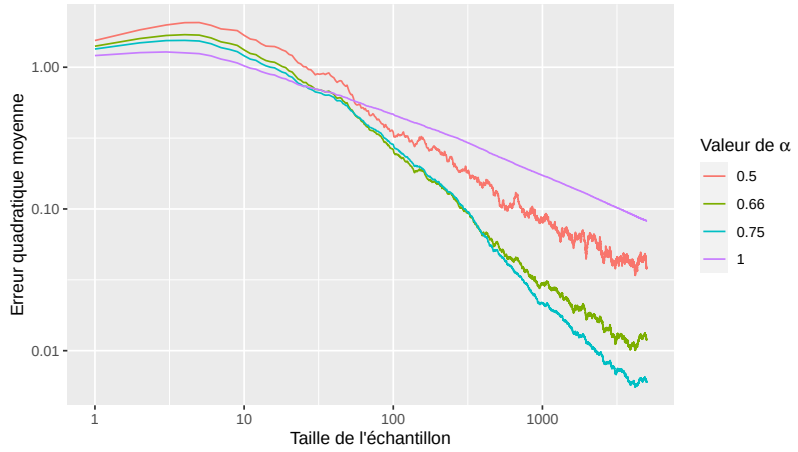

FIGURE 3.4 – Evolution de l'erreur quadratique moyenne de θ_n en fonction de la taille de l'échantillon n et du choix du paramètre α dans le cadre de la régression logistique.

3.2.5 Remarques

On a vu que sous des hypothèses faibles et en prenant un pas de la forme $c_\gamma n^{-\alpha}$, avec $\alpha \in (1/2, 1)$, on obtient une vitesse de convergence de l'ordre $\frac{1}{n^\alpha}$ (à un terme logarithmique près). Cependant, comme on a prit $\alpha < 1$, on ne peut pas avoir une vitesse "optimale", c'est à dire une vitesse de l'ordre de $1/n$. On peut alors se dire naïvement que l'on peut prendre $\alpha = 1$. Cependant, pour assurer une vitesse en $1/n$, cela implique de prendre $c_\gamma > \frac{1}{2\lambda_{\min}}$, avec $\lambda_{\min} = \lambda_{\min}\left(\nabla^2 G(m)\right)$. Dans la Figure 3.2, cette hypothèse était vérifiée et on voit bien que les estimateurs convergent à la bonne vitesse. Cependant, dans la Figure 3.4, cette hypothèse n'était pas vérifiée et on a pu remarquer que l'on ne convergeait pas du tout à la vitesse $1/n$. Cette approche a donc deux principaux inconvénients. Le premier, c'est qu'il faut calibrer le pas par rapport à la plus petite valeur propre de la Hessienne alors que l'on ne la connait pas. Dans certains cas, on est capable de la minorer par une certaine valeur $\lambda_{\inf}$ et peut alors choisir $c_\gamma > \frac{1}{2\lambda_{\inf}}$ et ainsi obtenir une vitesse en $1/n$ (à un terme logarithmique près). On peut même montrer, sous certaines hypothèses,

$$\sqrt{n}\,(m_n - m) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\,(0, \Sigma_{RM})\,,$$

où

$$\Sigma_{RM} = c_\gamma \int_0^{+\infty} e^{-s\left(H - \frac{1}{2c_\gamma} I_d\right)} \Sigma e^{-s\left(H - \frac{1}{2c_\gamma} I_d\right)} ds$$

avec $\Sigma = \mathbb{E}\left[\nabla_h g\,(X, m)\,\nabla_h g(X, m)^T\right]$. A noter que $\left(H - \frac{1}{2c_\gamma} I_d\right)$ est définie positive car $\frac{1}{2c_\gamma} < \lambda_{\min}(H)$, et donc Σ_{RM} est bien définie. Cependant, si on s'intéresse aux M-estimateur $\hat{m}_n$, on a

vu que sous certaines hypothèses de régularité

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

et si $\Sigma \neq 0$, on peut montrer que l'on a ainsi une meilleur variance, i.e on peut montrer que la matrice $H^{-1}\Sigma H^{-1} - \Sigma_{RM}$ est au moins semi-définie négative. En d'autres termes, on ne peut pas avoir, généralement, un comportement asymptotique optimal pour les estimateurs obtenus à l'aide d'algorithmes de gradient stochastiques. Pour s'en convaincre, on peut considérer l'exemple de la régression linéaire. On rappelle que dans ce cas, on a $\Sigma = \sigma^2 H$, et en écrivant Σ_{RM} dans la base orthonormée de H, et en notant $\lambda_1, \ldots, \lambda_d$ ses valeurs propres, on a

$$\begin{aligned}
\Sigma_{RM} &= \sigma^2 c_\gamma \int_0^{+\infty} e^{-s\mathrm{diag}\left(\lambda_1 - \frac{1}{2c_\gamma}, \ldots, \lambda_d - \frac{1}{2c_\gamma}\right)} \mathrm{diag}\left(\lambda_1, \ldots, \lambda_d\right) e^{-s\mathrm{diag}\left(\lambda_1 - \frac{1}{2c_\gamma}, \ldots, \lambda_d - \frac{1}{2c_\gamma}\right)} ds \\
&= \frac{\sigma^2}{2} \mathrm{diag}\left(\frac{c_\gamma \lambda_1}{\lambda_1 - \frac{1}{2c_\gamma}}, \ldots, \frac{c_\gamma \lambda_d}{\lambda_d - \frac{1}{2c_\gamma}}\right) \\
&= \frac{\sigma^2}{2} \mathrm{diag}\left(\lambda_1^{-1} \frac{c_\gamma \lambda_1}{1 - \frac{1}{2c_\gamma \lambda_1}}, \ldots, \lambda_d^{-1} \frac{c_\gamma \lambda_d}{1 - \frac{1}{2c_\gamma \lambda_d}}\right) \\
&> \sigma^2 \mathrm{diag}\left(\lambda_1^{-1}, \ldots, \lambda_d^{-1}\right).
\end{aligned}$$

car pour tout $x \in (0,1)$, on a $\frac{x}{1-\frac{1}{2x}} > 2$. On a donc $H^{-1}\Sigma H^{-1} - \Sigma_{RM}$ qui est définie négative.

Chapitre 4

Accélération des méthodes de gradient stochastiques

On a vu dans le chapitre précédent qu'il n'est pas possible, généralement, d'obtenir un comportement asymptotique optimal pour les estimateurs obtenus à l'aide d'algorithmes de gradient stochastiques. On propose dans ce chapitre des transformations de ces derniers afin d'accélérer la convergence et obtenir des estimateurs asmptotiquement efficaces.

4.1 Algorithmes de gradient stochastiques moyennés

Une méthode usuelle pour accélérer la converge, introduite par [Rup88] et [PJ92], est de considérer un algorithme de gradient stochastique moyenné. Celui-ci consiste à considérer la moyenne de tous les estimateurs de gradient obtenus au temps n, i.e l'estimateur moyenné $\overline{m}_n$ est défini pour tout $n \geq 0$ par

$$\overline{m}_n = \frac{1}{n+1} \sum_{k=0}^{n} m_k.$$

On reste ici sur des estimateurs en ligne dans le sens où on peut écrire la procédure de manière récursive pour tout $n \geq 0$ comme

$$\begin{aligned} m_{n+1} &= m_n - \gamma_{n+1} \nabla_h g\left(X_{n+1}, m_n\right) \\ \overline{m}_{n+1} &= \overline{m}_n + \frac{1}{n+2}\left(m_{n+1} - \overline{m}_n\right), \end{aligned}$$

avec $m_0 = \overline{m}_0$ borné. A noter que la mise à jour de l'estimateur reste très peu couteuse en terme de temps de calcul. En effet, l'étape de moyennisation ne représente, à chaque itération, que $O(d)$ opérations supplémentaires. On s'intéresse maintenant aux vitesses de convergence des estimateurs moyennés. Pour cela, dans ce qui suit, on considère une suite de pas de la forme $\gamma_n = c_\gamma n^{-\alpha}$ avec $c_\gamma > 0$ et $\alpha \in (1/2, 1)$.

4.1.1 Vitesse de convergence presque sûre

Avant de donner les vitesses de convergence, on introduit le Lemme de Toeplitz, qui sera très utile dans les preuves.

Lemma 4.1.1 (Toeplitz). *Soit a_n une suite positive telle que $\sum_{n\geq 0} a_n = +\infty$ et X_n une suite de variables convergeant presque sûrement vers une variable aléatoire X. Alors*

$$\frac{1}{\sum_{k=0}^{n} a_k} \sum_{k=0}^{n} a_k X_k \xrightarrow[n\to+\infty]{p.s} X.$$

En particulier, le lemme de Toeplitz nous dit que si m_n converge presque sûrement vers m, alors $\overline{m}_n$ converge aussi.

Afin de donner la vitesse de convergence des estimateurs moyennés, on introduit maintenant une nouvelle hypothèse, qui permet notamment de donner la vitesse de convergence du terme induit par le terme de reste dans la décomposition de Taylor du gradient.

(PS3) Il existe des constantes $\eta > 0$ et C_η telles que pour tout $h \in B_\eta := \mathcal{B}(m, \eta)$,

$$\left\|\nabla G(h) - \nabla^2 G(m)(h-m)\right\| \leq C_\eta \left\|h-m\right\|^2.$$

L'hypothèse **(PS3)** est vérifiée, par exemple, dès que la Hessienne de G est C_η Lipschitz sur le voisinage de m, i.e si pour tout $h \in \mathcal{B}_\eta$,

$$\left\|\nabla^2 G(m) - \nabla^2 G(h)\right\|_{op} \leq C_\eta \left\|h-m\right\|.$$

En effet, comme $\nabla G(m) = 0$, la décomposition de Taylor du gradient nous donne que pour tout $h \in \mathcal{B}_\eta$,

$$\nabla G(h) = \int_0^1 \nabla^2 G\left(m + t(h-m)\right) dt (h-m)$$

et on obtient donc

$$\begin{aligned}
\left\|\nabla G(h) - \nabla^2 G(m)(h-m)\right\| &= \left\| \int_0^1 \nabla^2 G(m + t(h-m)) dt (h-m) - \nabla^2 G(m)(h-m) \right\| \\
&= \left\| \int_0^1 \left(\nabla^2 G(m + t(h-m)) - \nabla^2 G(m)\right) dt (h-m) \right\| \\
&\leq \int_0^1 \left\|\nabla^2 G(m + t(h-m)) - \nabla^2 G(m)\right\|_{op} dt \left\|h-m\right\|
\end{aligned}$$

Comme la Hessienne est C_η-Lipschitz sur $\mathcal{B}_\eta$ et comme pour tout $h \in \mathcal{B}_\eta$ et $t \in [0,1]$, on a $m + t(h-m) \in B_\eta$, on obtient que l'hypothèse **(PS3)** est bien vérifiée. On peut maintenant revenir à la vitesse de convergence presque sûre des estimateurs moyennés, ce que nous donne le théorème suivant.

Théorème 4.1.1. *On suppose que les hypothèses **(PS1)** à **(PS3)** sont vérifiées. Alors pour tout $\delta > 0$,*

$$\|\overline{m}_n - m\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) \quad p.s.$$

On obtient donc une vitesse (à un terme log près) une vitesse de convergence en $1/n$.

Démonstration. Au vu du terme log, on se doute qu'il va falloir utiliser une loi des grands nombres pour les martingales vectorielles, et il faut donc faire apparaitre ce terme de martingale. Rappelons que l'on a la décomposition suivante pour m_{n+1} :

$$m_{n+1} - m = (I_d - \gamma_{n+1}H)(m_n - m) + \gamma_{n+1}\xi_{n+1} - \gamma_{n+1}\delta_n$$

avec $H = \nabla^2 G(m)$, $\delta_n = \nabla G(m_n) - H(m_n - m)$, et $\xi_{n+1} = \nabla G(m_n) - \nabla_h g(X_{n+1}, m_n)$. Rappelons également que (ξ_n) est une suite de différences de martingales par rapport à la filtration générée par l'échantillon. Cette décomposition de m_{n+1} peut se réécrire,

$$\gamma_{n+1}H(m_n - m) = (m_n - m) - (m_{n+1} - m) + \gamma_{n+1}\xi_{n+1} - \gamma_{n+1}\delta_n$$

et en divisant par γ_{n+1}, on obtient donc

$$H(m_n - m) = \frac{(m_n - m) - (m_{n+1} - m)}{\gamma_{n+1}} + \xi_{n+1} - \delta_n.$$

En sommant ces inégalités et par linéarité, on obtient

$$H\sum_{k=0}^{n}(m_k - m) = \sum_{k=0}^{n}\frac{(m_k - m) - (m_{k+1} - m)}{\gamma_{k+1}} + \underbrace{\sum_{k=0}^{n}\xi_{k+1}}_{:=M_n} - \sum_{k=0}^{n}\delta_k$$

A noter que M_n est une martingale par rapport à la filtration. Enfin, en divisant par $n+1$, on obtient (par définition de $\overline{m}_n$),

$$H(\overline{m}_n - m) = \underbrace{\frac{1}{n+1}\sum_{k=0}^{n}\frac{(m_k - m) - (m_{k+1} - m)}{\gamma_{k+1}}}_{:=R_{1,n}} + \frac{1}{n+1}\sum_{k=0}^{n}\xi_{k+1} - \underbrace{\frac{1}{n+1}\sum_{k=0}^{n}\delta_k}_{:=R_{2,n}} \tag{4.1}$$

et il ne reste plus qu'à donner les vitesses de convergence de chacun des termes à droite de l'égalité (4.1)

Vitesse de convergence de $\frac{1}{n+1}\mathbf{M_{n+1}}$. Rappelons que pour appliquer le Théorème 2.2.2, il faut vérifier qu'il existe $\nu > 0$ tel que les moments conditionnels d'ordre $2 + 2\nu$ soit uniformément

bornés, ce qui n'est pas exactement le cas pour nous. En effet, on a à l'aide de l'hypothèse **(PS1)**,

$$\begin{aligned}\mathbb{E}\left[\|\xi_{k+1}\|^{2+2\nu}|\mathcal{F}_k\right] &\leq \mathbb{E}\left[\|\nabla_h g(X_{n+1},m_n)+\mathbb{E}\left[\nabla_h g(X_{n+1},m_n)|\mathcal{F}_n\right]\|^{2+2\nu}|\mathcal{F}_n\right]\\ &\leq 2^{1+2\nu}\mathbb{E}\left[\|\nabla_h g(X_{n+1},m_n)\|^{2+2\nu}|\mathcal{F}_n\right]\\ &\leq 2^{1+2\nu}C_\nu\left(1+\|m_k-m\|^{2+2\nu}\right).\end{aligned}$$

Cependant, comme m_n converge presque sûrement vers m, on va utiliser un argument de troncature. Plus précisément, on introduit l'évènement $A_n = \{\|m_n - m\| \leq 1\}$ et on peut réécrire la martingale comme

$$M_n = \sum_{k=0}^{n} \xi_{k+1}\mathbf{1}_{A_k} + \sum_{k=0}^{n}\xi_{k+1}\mathbf{1}_{A_k^C}$$

Comme $\mathbf{1}_{A_n}$ est $\mathcal{F}_n$-mesurable, $(\xi_{n+1}\mathbf{1}_{A_n})$ est toujours une suite de différences de martingales, et elle vérifie

$$\mathbb{E}\left[\|\xi_{n+1}\|^{2+2\nu}\mathbf{1}_{A_n}|\mathcal{F}_n\right] \leq 2^{2+2\nu}C_\nu$$

et en appliquant le Théorème 2.2.1, on obtient pour tout $\delta > 0$,

$$\frac{1}{(n+1)^2}\left\|\sum_{k=0}^{n}\xi_{k+1}\mathbf{1}_{A_k}\right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) \quad p.s.$$

De plus comme $\mathbf{1}_{A_n^C}$ converge presque sûrement vers 0, on a

$$\sum_{n\geq 0}\|\xi_{n+1}\|\mathbf{1}_{A_n^C} < +\infty \quad p.s$$

et en particulier,

$$\frac{1}{(n+1)^2}\left\|\sum_{k=0}^{n}\xi_{k+1}\mathbf{1}_{A_k^C}\right\|^2 = O\left(\frac{1}{n^2}\right) \quad p.s.$$

et donc pour tout $\delta > 0$,

$$\frac{1}{(n+1)^2}\|M_n\|^2 \leq \frac{2}{(n+1)^2}\left(\left\|\sum_{k=0}^{n}\xi_{k+1}\mathbf{1}_{A_k}\right\|^2 + \left\|\sum_{k=0}^{n}\xi_{k+1}\mathbf{1}_{A_k^C}\right\|^2\right) = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) \quad p.s.$$

Vitesse de convergence de $\mathbf{R_{1,n}}$. Afin de simplifier les notations, notons $u_k = m_k - m$. On peut

réécrire $(n+1)R_{1,n}$ comme

$$\begin{aligned}(n+1)R_{1,n} &= \sum_{k=0}^{n} (u_k - u_{k+1})\,\gamma_{k+1}^{-1} \\ &= \sum_{k=0}^{n} u_k\gamma_{k+1}^{-1} - \sum_{k=0}^{n} u_{k+1}\gamma_{k+1}^{-1} \\ &= \sum_{k=1}^{n} u_k\gamma_{k+1}^{-1} + u_0\gamma_1^{-1} - \sum_{k=0}^{n} u_{k+1}\gamma_{k+1}^{-1} \\ &= \sum_{k=1}^{n} u_k\left(\gamma_{k+1}^{-1} - \gamma_k^{-1}\right) + u_0\gamma_1^{-1} + \sum_{k=1}^{n} u_k\gamma_k^{-1} - \sum_{k=0}^{n} u_{k+1}\gamma_{k+1}^{-1}.\end{aligned}$$

En faisant le changement d'indice $k' = k+1$ dans la dernière somme, on obtient

$$\begin{aligned}(n+1)R_{1,n} &= \sum_{k=1}^{n} u_k\left(\gamma_{k+1}^{-1} - \gamma_k^{-1}\right) + u_0\gamma_1^{-1} + \sum_{k=1}^{n} u_k\gamma_k^{-1} - \sum_{k'=1}^{n+1} u_{k'}\gamma_{k'}^{-1} \\ &= \sum_{k=1}^{n} u_k\left(\gamma_{k+1}^{-1} - \gamma_k^{-1}\right) + u_0\gamma_1^{-1} - u_{n+1}\gamma_{n+1}^{-1}\end{aligned}$$

On a clairement

$$\frac{1}{(n+1)}\left\|m_0 - m\right\|\gamma_1^{-1} = O\left(\frac{1}{n+1}\right) \quad p.s.$$

et ce terme est donc négligeable. De plus, grâce au Théorème 3.2.1, on a

$$\frac{1}{n+1}\left\|m_{n+1} - m\right\|\gamma_{n+1}^{-1} = O\left(\frac{\sqrt{\ln(n+1)}}{(n+1)^{1-\alpha/2}}\right) \quad p.s$$

et comme $\alpha < 1$, ce terme est négligeable. Il reste donc à donner la vitesse de convergence du dernier terme. A noter que la fonction $f : t \longmapsto c_\gamma^{-1}t^\alpha$ est dérivable, que $\gamma_{k+1}^{-1} - \gamma_k^{-1} = f(k+1) - f(k)$, et que

$$f'(t) = c_\gamma^{-1}\alpha t^{\alpha-1}.$$

On obtient donc $\left|\gamma_{k+1}^{-1} - \gamma_k^{-1}\right| \leq c_\gamma^{-1}\alpha k^{\alpha-1}$ et

$$\left\|\sum_{k=1}^{n}(m_k - m)\left(\gamma_{k+1}^{-1} - \gamma_k^{-1}\right)\right\| \leq \sum_{k=1}^{n}\left\|m_k - m\right\|c_\gamma^{-1}\alpha k^{\alpha-1}$$

De plus, d'après le Théorème 3.2.1, on a pour tout $\delta > 0$,

$$k^{1-\alpha}\frac{k^{\alpha/2}}{\ln(k+1)^{1/2+\delta}}\left\|m_k - m\right\|k^{\alpha-1} \xrightarrow[k\to+\infty]{p.s} 0,$$

et on a donc, d'après le Lemme de Toeplitz,

$$\sum_{k=1}^{n} \|m_k - m\| k^{\alpha-1} = \sum_{k=1}^{n} k^{\alpha/2-1} \ln(k+1)^{1/2+\delta} \left(\frac{k^{1-\alpha/2}}{\ln(k+1)^{1/2+\delta}} \|m_k - m\| k^{\alpha-1} \right)$$
$$= o\left(\sum_{k=1}^{n} k^{\alpha/2-1} \ln(k+1)^{1/2+\delta} \right) \quad p.s.$$

et comme $\sum_{k=1}^{n} k^{1-\alpha/2} \ln(k+1)^{1/2+\delta} = O\left(n^{\alpha/2} \ln(n+1)^{1/2+\delta}\right)$, il vient

$$\frac{1}{n+1} \left\| \sum_{k=1}^{n} (m_k - m) \left(\gamma_{k+1}^{-1} - \gamma_k^{-1} \right) \right\| = o\left(\frac{(\ln n)^{1+\delta}}{n^{1-\alpha/2}} \right) \quad p.s$$

et comme $\alpha < 1$, ce terme est négligeable, et donc $R_{1,n}$ est négligeable par rapport au terme de martingale.

Vitesse de convergence de $\mathbf{R_{2,n}}$. Comme m_n converge presque sûrement vers m et d'après l'hypothèse **(PS3)**, on a $\delta_n = O\left(\|m_n - m\|^2\right)$ presque sûrement, et donc, pour tout $\delta > 0$,

$$\|\delta_n\| \frac{(n+1)^{\alpha}}{\ln(n+1)^{1+\delta}} \xrightarrow[n\to+\infty]{p.s} 0.$$

Ainsi, en appliquant le lemme de Toeplitz, on obtient

$$(n+1) \|R_{2,n}\| \leq \sum_{k=0}^{n} \|\delta_k\| = \sum_{k=0}^{n} \frac{\ln(k+1)^{1+\delta}}{(k+1)^{\alpha}} \left(\|\delta_k\| \frac{(k+1)^{\alpha}}{\ln(k+1)^{1+\delta}} \right) = o\left(\sum_{k=0}^{n} \frac{\ln(k+1)^{1+\delta}}{(k+1)^{\alpha}} \right) \quad p.s$$

et comme $\sum_{k=0}^{n} \frac{\ln(k+1)^{1+\delta}}{(k+1)^{\alpha}} = O\left(\ln(n+1)^{1+\delta} n^{1-\alpha}\right)$, on obtient

$$R_{2,n} = o\left(\ln(n+1)^{1+\delta} (n+1)^{-\alpha} \right) \quad p.s$$

ce qui est négligeable car $\alpha > 1/2$, et on a donc

$$\|H(\overline{m}_n - m)\|^2 = O\left(\frac{\ln n}{n} \right) \quad p.s.$$

En particulier

$$\|(\overline{m}_n - m)\|^2 \leq \lambda_{\min}^{-2} \|H(\overline{m}_n - m)\|^2 = O\left(\frac{\ln n}{\lambda_{\min}^2 n} \right) \quad p.s.$$

□

Remarquons que l'on pourrait alléger l'hypothèse **(PS3)**. En effet, on peut la remplacer par

(PS3') Il existe des constantes positives $A > 0$, $a > 1$ et $C_{A,a}$ telles que pour tout $h \in \mathcal{B}(m, A)$,

$$\left\| \nabla G(h) - \nabla^2 G(m)(h - m) \right\| \leq C_{A,a} \|h - m\|^a .$$

Ainsi, en prenant $\alpha > \frac{1}{a}$, le terme $R_{2,n}$ resterait un terme négligeable, et on conserverait la même vitesse de convergence pour l'estimateur moyenné.

4.1.2 Normalité asymptotique

Afin d'obtenir la normalité asymptotique de l'estimateur moyenné, on a besoin d'une nouvelle hypothèse sur le gradient de g. Plus précisément, on supposera que l'hypothèse suivante est vérifiée :

(PS4) La fonction $\Sigma : \mathbb{R}^d \longrightarrow \mathcal{M}_d(\mathbb{R})$ définie pour tout $h \in \mathbb{R}^d$ par

$$\Sigma(h) = \mathbb{E}\left[\nabla_h g\left(X, h\right) \nabla_h g\left(X, h\right)^T\right]$$

est continue en m.

On peut maintenant obtenir la normalité asymptotique.

Théorème 4.1.2. *On suppose que les hypothèses **(PS1)** à **(PS4)** sont vérifiées. Alors*

$$\sqrt{n}\left(\overline{m}_n - m\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right)$$

avec $H := \nabla^2 G(m)$ *et* $\Sigma := \mathbb{E}\left[\nabla_h g\left(X, m\right) \nabla_h g\left(X, m\right)^T\right]$.

On obtient donc un comportement asymptotique des estimateurs moyennés identique à celui des M-estimateurs.

Démonstration. On a vu dans la preuve du Théorème 4.1.1 que l'on pouvait écrire $H\left(\overline{m}_n - m\right)$ comme

$$H\left(\overline{m}_n - m\right) = R_{1,n} + R_{2,n} + \frac{1}{n+1} M_n$$

avec $\|R_{1,n}\| = o\left(\frac{1}{\sqrt{n}}\right)$ et $\|R_{1,n}\| = o\left(\frac{1}{\sqrt{n}}\right)$ presque sûrement et $M_n = \sum_{k=0}^{n} \xi_{k+1}$ est un terme de martingale. En particulier, il vient

$$\sqrt{n}\,\|R_{1,n}\| \xrightarrow[n\to+\infty]{p.s} 0 \qquad \text{et} \qquad \sqrt{n}\,\|R_{2,n}\| \xrightarrow[n\to+\infty]{p.s} 0$$

Il ne reste donc plus qu'à appliquer le TLC au terme de martingale. Attention, pour pouvoir vérifier les conditions du TLC, il faut que la martingale soit a minima de carré intégrable, ce qui n'a pas été vérifié. On va donc réécrire le terme de martingale comme

$$M_n = \sum_{k=0}^{n} \xi_{k+1} \mathbf{1}_{\|m_k - m\| \leq 1} + \sum_{k=1}^{n} \xi_{k+1} \mathbf{1}_{\|m_k - m\| > 1} =: M_{1,n} + M_{2,n}.$$

Comme m_n converge presque sûrement vers m, $\mathbf{1}_{\|m_n - m\|>1}$ converge presque sûrement vers 0, et donc

$$\sum_{n\geq 0} \|\xi_{n+1}\|\, \mathbf{1}_{\|m_n - m\|>1} < +\infty \quad p.s$$

et en particulier

$$\frac{1}{\sqrt{n}} \|M_{2,n}\| \xrightarrow[n\to+\infty]{p.s} 0 \quad p.s$$

i.e ce terme est négligeable. Pour appliquer le TLC au terme de martingale, calculons le crochet : on a

$$\begin{aligned}\langle M_1\rangle_n &= \sum_{k=0}^{n} \mathbb{E}\left[\xi_{k+1}\xi_{k+1}^T \mathbf{1}_{\|m_k-m\|\leq 1}|\mathcal{F}_k\right] \\ &= \sum_{k=0}^{n} \mathbb{E}\left[\left(\nabla G\left(m_k\right) - \nabla_h g\left(X_{k+1}, m_k\right)\right)\left(\nabla G\left(m_k\right) - \nabla_h g\left(X_{k+1}, m_k\right)\right)^T \mathbf{1}_{\|m_k-m\|\leq 1}|\mathcal{F}_k\right]\end{aligned}$$

Comme m_k est $\mathcal{F}_k$-mesurable, il vient

$$\begin{aligned}\langle M_1\rangle_n &= \sum_{k=0}^{n} \nabla G\left(m_k\right)\nabla G\left(m_k\right)^T \mathbf{1}_{\|m_k-m\|\leq 1} - \nabla G\left(m_k\right)\left(\mathbb{E}\left[\nabla_h g\left(X_{k+1}, m_k\right)|\mathcal{F}_k\right]\right)^T \mathbf{1}_{\|m_k-m\|\leq 1} \\ &- \sum_{k=0}^{n} \mathbb{E}\left[\nabla_h g\left(X_{k+1}, m_k\right)|\mathcal{F}_k\right]\nabla G\left(m_k\right)^T \mathbf{1}_{\|m_k-m\|\leq 1} + \mathbb{E}\left[\nabla_h g\left(X_{k+1}, m_k\right)\nabla_h g\left(X_{k+1}, m_k\right)^T|\mathcal{F}_k\right]\mathbf{1}_{\|m_k-m\|\leq 1} \\ &= \sum_{k=0}^{n} \mathbb{E}\left[\nabla_h g\left(X_{k+1}, m_k\right)\nabla_h g\left(X_{k+1}, m_k\right)^T|\mathcal{F}_k\right]\mathbf{1}_{\|m_k-m\|\leq 1} - \sum_{k=0}^{n} \nabla G\left(m_k\right)\nabla G\left(m_k\right)^T \mathbf{1}_{\|m_k-m\|\leq 1}\end{aligned}$$

L'hypothèse **(PS2)** nous donne que G est deux fois continument différentiable sur un voisinage V de m. Il existe donc un constante positive C_V telle que pour tout $h \in V$, $\left\|\nabla^2 G(h)\right\|_{op} \leq C_V$ et donc pour tout $h \in V$,

$$\|\nabla G(h)\| = \left\|\int_0^1 \nabla^2 G(m + t(h-m))dt(h-m)\right\| \leq C_V \|h-m\|.$$

De plus, pour tout $h \in V$, on a

$$\left\|\nabla G(h)\nabla G(h)^T\right\|_{op} \leq \|\nabla G(h)\|^2 \leq C_V^2 \|h-m\|^2.$$

Ainsi, comme m_n converge presque sûrement vers m et d'après le Théorème 3.2.1, on a pour tout $\delta > 0$,

$$\frac{(n+1)^\alpha}{\ln(n+1)^{1+\delta}} \left\|\nabla G(h)\nabla G(h)^T\right\|_{op} \xrightarrow[n\to+\infty]{p.s} 0$$

et en appliquant le lemme de Toeplitz, on obtient

$$\begin{aligned}\left\|\sum_{k=0}^{n} \nabla G\left(m_k\right)\nabla G\left(m_k\right)^T \mathbf{1}_{\|m_k-m\|\leq 1}\right\|_{op} &\leq \sum_{k=0}^{n} \frac{\ln(k+1)^{1+\delta}}{(k+1)^\alpha}\left(\frac{(k+1)^\alpha}{\ln(k+1)^{1+\delta}}\left\|\nabla G\left(m_k\right)\right\|^2\right) \\ &= o\left(\sum_{k=0}^{n} \frac{\ln(k+1)^{1+\delta}}{(k+1)^\alpha}\right) \quad p.s\end{aligned}$$

et donc

$$\frac{1}{n+1}\left\|\sum_{k=0}^{n}\nabla G\left(m_k\right)\nabla G\left(m_k\right)^T\mathbf{1}_{\|m_k-m\|\leq 1}\right\|_{op}=o\left(\frac{\ln(n+1)^{1+\delta}}{(n+1)^{\alpha}}\right)\quad p.s.$$

Enfin, comme m_n converge presque sûrement vers m et d'après l'hypothèse **(PS4)**, on a

$$\mathbb{E}\left[\nabla_h g\left(X_{n+1},m_n\right)\nabla_h g\left(X_{n+1},m_n\right)^T\mathbf{1}_{\|m_n-m\|\leq 1}|\mathcal{F}_n\right]\xrightarrow[n\to+\infty]{p.s}\Sigma$$

et en appliquant le lemme de Toeplitz, on obtient

$$\frac{1}{n+1}\sum_{k=0}^{n}\mathbb{E}\left[\nabla_h g\left(X_{k+1},m_k\right)\nabla_h g\left(X_{k+1},m_k\right)^T\mathbf{1}_{\|m_k-m\|\leq 1}|\mathcal{F}_k\right]\xrightarrow[n\to+\infty]{p.s}\Sigma.$$

Ainsi, $\frac{1}{n+1}\left\langle M_1\right\rangle_n\xrightarrow[n\to+\infty]{p.s}\Sigma$. De plus, comme $\mathbb{E}\left[\left\|\xi_{n+1}\right\|^{2+2\eta}\mathbf{1}_{\|m_n-m\|\leq 1}|\mathcal{F}_n\right]\leq 2^{2+2\nu}C_\nu$, la condition de Lindeberg est vérifiée. On peut donc appliquer le TLC pour les martingales et on obtient

$$\frac{1}{\sqrt{n}}M_{1,n}\xrightarrow[n\to+\infty]{\mathcal{L}}\mathcal{N}\left(0,\Sigma\right).$$

ce qui conclut la preuve. □

4.1.3 Application au modèle linéaire

On se place dans le cadre de la régression linéaire (1.1). On peut donc écrire l'algorithme de gradient stochastique moyenné comme

$$\begin{aligned}\theta_{n+1}&=\theta_n+\gamma_{n+1}\left(Y_{n+1}-X_{n+1}^T\theta_n\right)X_{n+1}\\ \overline{\theta}_{n+1}&=\overline{\theta}_n+\frac{1}{n+2}\left(\theta_{n+1}-\overline{\theta}_n\right),\end{aligned}$$

avec $\overline{\theta}_0=\theta_0$ borné. Le théorème suivant donne la vitesse de convergence presque sûre ainsi que la normalité asymptotique des estimateurs moyennés.

Théorème 4.1.3. *On suppose qu'il existe $\eta>\frac{1}{\alpha}-1$ tel que X et ϵ admettent respectivement des moments d'ordre $4+4\eta$ et $2+2\eta$. Alors pour tout $\delta>0$,*

$$\left\|\overline{\theta}_n-\theta\right\|^2=o\left(\frac{(\ln n)^{1+\delta}}{n}\right)\quad p.s\qquad et\qquad\sqrt{n}\left(\overline{\theta}_n-\theta\right)\xrightarrow[n\to+\infty]{\mathcal{L}}\mathcal{N}\left(0,\sigma^2H^{-1}\right).$$

Démonstration. On a vu dans la preuve du Théorème 3.2.2 que les hypothèses **(PS1)** et **(PS2)** sont vérifiées. De plus, comme pour tout $h\in\mathbb{R}^d$

$$\nabla G\left(h\right)=\int_0^1\nabla^2G\left(\theta+t(h-\theta)\right)dt(h-\theta)=\mathbb{E}\left[XX^T\right](h-\theta)$$

l'hypothèse **(PS3)** est vérifiée et on a même $\nabla G(h)-\nabla^2G(\theta)(h-\theta)=0$. Il reste donc à vérifier

que **(PS4)** l'est. A noter que pour tout $h \in \mathbb{R}^d$, comme $Y - X^T\theta = \epsilon$, on peut écrire $\Sigma(h)$ comme

$$\begin{aligned}\Sigma(h) = \mathbb{E}\left[\left(Y - X^T h\right)^2 XX^T\right] &= \mathbb{E}\left[\epsilon^2 XX^T\right] - 2\mathbb{E}\left[\epsilon X^T (h-\theta) XX^T\right] + \mathbb{E}\left[\left(X^T(h-\theta)\right)^2 XX^T\right] \\ &= \mathbb{E}\left[\epsilon^2 XX^T\right] + \mathbb{E}\left[\left(X^T(h-\theta)\right)^2 XX^T\right]\end{aligned}$$

et comme X admet un moment d'ordre 4, la fonction Σ est bien continue en θ. De plus, on a $\Sigma(\theta) = \sigma^2\mathbb{E}\left[XX^T\right] = \sigma^2 H$, ce qui conclut la preuve. □

Dans la Figure 4.1, on s'intéresse à l'évolution de l'erreur quadratique moyenne des estimateurs de gradient et de leur version moyennée, dans le cadre de la régression linéaire, en fonction de la taille d'échantillon n. Pour cela, on considère le modèle

$$\theta = (-4,-3,-2,-1,0,1,2,3,4,5)^T \in \mathbb{R}^{10}, \qquad X \sim \mathcal{N}(0, I_{10}), \qquad \text{et} \qquad \epsilon \sim \mathcal{N}(0,1)$$

De plus, on a choisi $c_\gamma = 1$ et $\alpha = 0.66$ ou 0.75. Enfin, on a calculé l'erreur quadratique moyenne des estimateurs en générant 50 échantillons de taille $n = 5000$. On peut voir que dès que l'algorithme de gradient arrive à convergence ($n \simeq 200$), la moyennisation permet une réelle accélération, et à échelle logarithmique, la pente avoisine -1. De plus, on peut noter que pour $\alpha = 0.66$, l'algorithme de gradient arrive plus vite à convergence, ce qui permet à l'étape de moyennisation d'accéler la convergence plus rapidement.

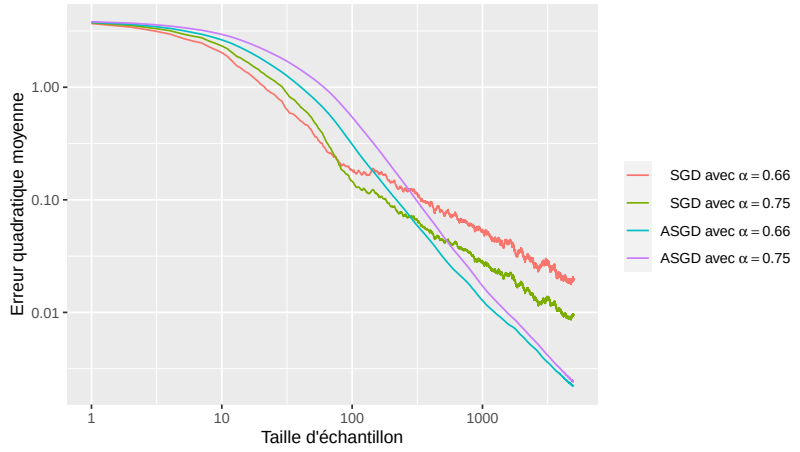

FIGURE 4.1 – Evolution de l'erreur quadratique moyenne de l'estimateur de gradient θ_n (SGD) et de sa version moyennée $\overline{\theta}_n$ (ASGD) en fonction de la taille d'échantillon n dans le cadre de la régression linéaire.

De plus, pour tout $x_0 \in \mathbb{R}^d$, on a

$$\sqrt{n}\left(x_0^T\overline{\theta}_n - x_0^T\theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2 x_0^T H^{-1} x_0\right)$$

ce que l'on peut réécrire comme

$$\sqrt{n}\frac{x_0^T\overline{\theta}_n - x_0^T\theta}{\sqrt{\sigma^2 x_0^T H^{-1} x_0}} \xrightarrow[n\to+\infty]{\mathcal{N}} (0,1)\,.$$

Ainsi, en connaissant un estimateur en ligne de H^{-1}, on pourrait construire un intervalle de confiance et un test en ligne pour $x_0^T\theta$. Un estimateur récursif de H serait défini pour tout $n \geq 0$ par

$$\overline{H}_{n+1} = \overline{H}_n + \frac{1}{n+2}\left(X_{n+1}X_{n+1}^T - \overline{H}_n\right)$$

avec $\overline{H}_0$ symétrique et définie positive. En effet, on peut réécrire $\overline{H}_n$ comme

$$\overline{H}_n = \frac{1}{n+1}\left(\overline{H}_0 + \sum_{k=1}^{n} X_k X_k^T\right)$$

et par la loi des grands nombres, c'est un estimateur consistant. Cependant, inverser $\overline{H}_n$ à chaque itération pourrait s'avérer coûteux en terme de temps de calculs, et on ne peut pas, pour le moment construire ce type d'intervalles en ligne. On verra cependant comment inverser cette matrice à chaque itération et ce, à moindre cout. Remarquons maintenant que l'on peut écrire le TLC comme

$$\sqrt{n}\frac{H^{1/2}\left(\overline{\theta}_n - \theta\right)}{\sigma} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

et on obtient, par le théorème de continuité,

$$\left\|\sqrt{n}\frac{H^{1/2}\left(\overline{\theta}_n - \theta\right)}{\sigma}\right\|^2 = \frac{n\left(\overline{\theta}_n - \theta\right)^T H\left(\overline{\theta}_n - \theta\right)}{\sigma^2} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi_d^2$$

et grâce au théorème de Slutsky, on obtient

$$\frac{n\left(\overline{\theta}_n - \theta\right)^T \overline{H}_n\left(\overline{\theta}_n - \theta\right)}{\sigma^2} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi_d^2.$$

Il ne reste plus qu'à construire un estimateur récursif de σ^2 pour pouvoir tester en ligne $\theta = \theta_0$. Un estimateur en ligne "naturel" de σ^2 est de considérer la moyenne des erreurs quadratiques des prévisions, i.e de considérer l'estimateur $\hat{\sigma}_n^2$ défini récursivement pour tout $n \geq 1$ par

$$\hat{\sigma}_{n+1}^2 = \hat{\sigma}_n^2 + \frac{1}{n+2}\left(\left(Y_{n+1} - X_{n+1}^T\overline{\theta}_n\right)^2 - \hat{\sigma}_n^2\right)$$

avec $\sigma_0^2 = 0$. On peut réécrire l'estimateur comme

$$\hat{\sigma}_n^2 = \frac{1}{n+1}\sum_{k=1}^{n}\left(Y_k - X_k^T\overline{\theta}_{k-1}\right)^2.$$

On admettra pour le moment la consistance de cet estimateur, mais on l'établira par la suite pour les estimateurs obtenus à l'aide d'algorithmes de Newton stochastiques. On a donc, par le théorème de Slutsky,

$$C_n := \frac{n\left(\overline{\theta}_n - \theta\right)^T \overline{H}_n \left(\overline{\theta}_n - \theta\right)}{\hat{\sigma}_n^2} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi_d^2.$$

Dans la Figure 4.2, on prend une taille d'échantillon $n = 5000$ et on compare la fonction de répartition d'une Chi-deux à 10 degrés de liberté, et celle de C_n avec $\alpha = 0.66$ ou $\alpha = 0.75$. On voit que dans les deux cas, la fonction de répartition de C_n s'approche de celle de la Chi-deux.Cela fait de la variable aléatoire C_n un bon candidat pour construire des tests asymptotiques en ligne.

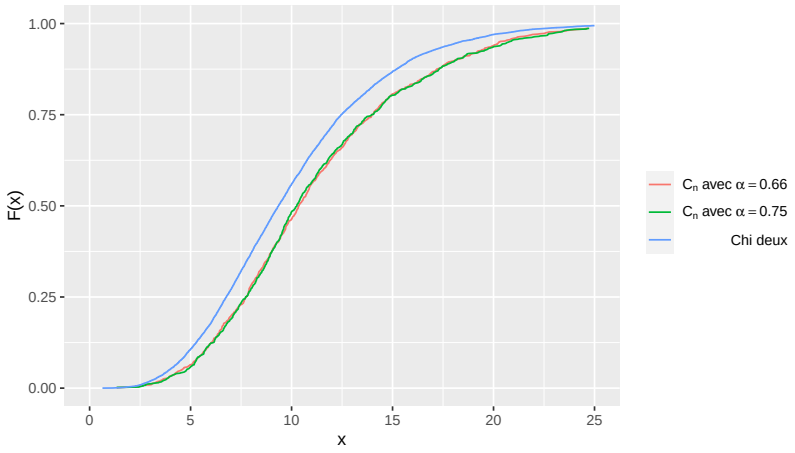

FIGURE 4.2 – Comparaison de la fonction de répartition de C_n avec $n = 5000$, pour $\alpha = 0.66$ et $\alpha = 0.75$, et de celle d'une Chi 2 à 10 degrés de liberté dans le cadre du modèle linéaire.

4.1.4 Application à la régression logistique

On se place dans le cadre de la régression logistique (1.2). On peut écrire l'algorithme de gradient stochastique moyenné comme

$$\theta_{n+1} = \theta_n - \gamma_{n+1}\left(\pi\left(X_{n+1}^T\theta_n\right) - Y_{n+1}\right)X_{n+1}$$
$$\overline{\theta}_{n+1} = \overline{\theta}_n + \frac{1}{n+2}\left(\theta_{n+1} - \overline{\theta}_n\right),$$

avec $\pi(x) = \frac{\exp(x)}{1+\exp(x)}$ et $\theta_0 = \overline{\theta}_0$ borné. Le théorème suivant donne la vitesse de convergence presque sûre ainsi que la normalité asymptotique des estimateurs moyennés.

Théorème 4.1.4. *On suppose que X admet un moment d'ordre 3 et que $\nabla^2 G(\theta)$ est inversible. Alors pour tout $\delta > 0$,*

$$\left\|\overline{\theta}_n - \theta\right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) \quad p.s \quad et \quad \sqrt{n}\left(\overline{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\right)$$

avec

$$H = \mathbb{E}\left[\pi\left(\theta^T X\right)\left(1-\pi\left(\theta^T X\right)\right) XX^T\right].$$

On admettra ce théorème mais la preuve est disponible dans la version longue. Dans la Figure 4.3, on considère le modèle $\theta = (1,1,1,1,1)^T \in \mathbb{R}^5$ et $X \sim \mathcal{N}(0, I_5)$. De plus, on a choisi $c_\gamma = 5$ et $\alpha = 0.66$ ou 0.75. Enfin, on a calculé l'erreur quadratique moyenne des estimateurs en générant 50 échantillons de taille $n = 20000$. On voit bien qu'au bout d'un certain temps ($n = 5000$), les estimateurs obtenus à l'aide de l'algorithme de gradient stochastique arrivent à convergence. Arrivé à ce moment, l'étape de moyennisation permet bel et bien d'accélérer la convergence. Cependant, on voit également que les estimateurs moyennés souffrent beaucoup en cas de mauvaise initialisation.

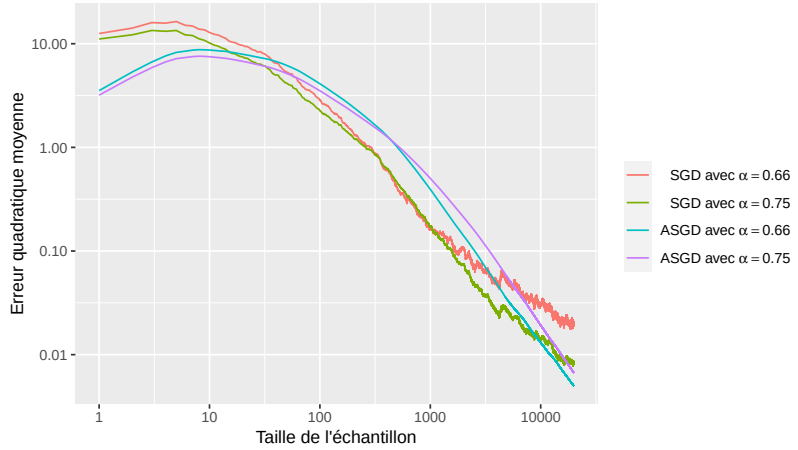

FIGURE 4.3 – Evolution de l'erreur quadratique moyenne par rapport à la taille de l'échantillon des estimateurs de gradients θ_n (SGD) et de leurs versions moyennées $\overline{\theta}_n$ (ASGD) dans le cadre de la régression logistique.

De la même façon que pour le modèle linéaire, pour tout $x_0 \in \mathbb{R}^d$, on a

$$\sqrt{n}\left(x_0^T\overline{\theta}_n - x_0^T\theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, x_0^T H^{-1} x_0\right)$$

ce que l'on peut réécrire comme

$$\sqrt{n}\frac{x_0^T\overline{\theta}_n - x_0^T\theta}{\sqrt{x_0^T H^{-1} x_0}} \xrightarrow[n\to+\infty]{} \mathcal{N}(0,1)$$

et ainsi, si on connaissait un estimateur en ligne de H^{-1}, on pourrait construire un intervalle confiance et un test en ligne pour $x_0^T\theta$. Un estimateur récursif en ligne de H serait pour tout $n \geq 0$

$$\overline{H}_{n+1} = \overline{H}_n + \frac{1}{n+2}\left(\pi\left(X_{n+1}^T\overline{\theta}_n\right)\left(1-\pi\left(X_{n+1}^T\overline{\theta}_n\right)\right) X_{n+1}X_{n+1}^T - \overline{H}_n\right)$$

ce que l'on peut réécrire comme

$$\overline{H}_n = \frac{1}{n+1}\sum_{k=0}^{n} \pi\left(X_{k+1}^T\overline{\theta}_k\right)\left(1-\pi\left(X_{k+1}^T\overline{\theta}_k\right)\right)X_{k+1}X_{k+1}^T$$

Cependant, inverser cette matrice à chaque itération peut être très coûteux en terme de temps de calculs si on s'y prend mal (on verra par la suite qu'en s'y prenant bien, cela ne représente "que" $O(d^2)$ opérations). Cependant, on peut d'ores et déjà remarquer que l'on peut réécrire le TLC comme

$$\sqrt{n}H^{1/2}\left(\overline{\theta}_n-\theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, I_d\right)$$

et en appliquant le Théorème de continuité, on obtient

$$\left\|\sqrt{n}H^{1/2}\left(\overline{\theta}_n-\theta\right)\right\|^2 = n\left(\overline{\theta}_n-\theta\right)^T H\left(\overline{\theta}_n-\theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi_d^2$$

et on obtient, via le Théorème de Slutsky,

$$C_n := \left\|\sqrt{n\overline{H}_n^{1/2}}\left(\overline{\theta}_n-\theta\right)\right\|^2 = n\left(\overline{\theta}_n-\theta\right)^T \overline{H}_n\left(\overline{\theta}_n-\theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi_d^2$$

et on peut ainsi construire un test asymptotique pour tester $\theta = \theta_0$. En effet, Figure 4.4, on voit que lorsque la taille d'échantillon augmente, on voit que la fonction de répartition de C_n s'approche de celle d'une Chi 2 à 5 degrés de liberté. A noter cependant que si α est trop grand, il semble que l'agorithme de gradient met trop de temps avant d'arriver à convergence, ce qui implique une moins bonne performance de l'algorithme moyenné dans ce cas, et donc également de l'estimateur de la Hessienne. Ceci peut expliquer en partie le fait que les résultats de la Figure 4.4 soient légèrement moins bon qu'attendu.

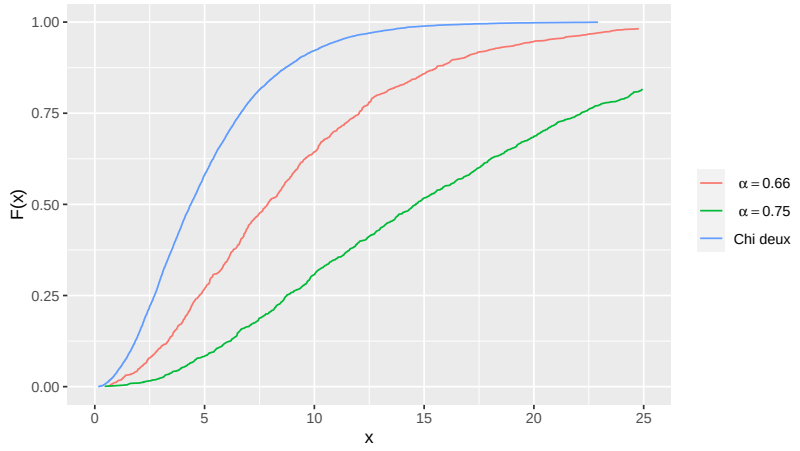

FIGURE 4.4 – Comparaison de la fonction de répartition de C_n, avec $n = 20000$ et $\alpha = 0.66$ ou $\alpha = 0.75$, et de la fonction de répartition d'une Chi deux à 5 degrés de liberté dans le cadre de la régression logistique.

4.1.5 Remarques

On a vu que l'algorithme moyenné permet d'accélérer les méthodes de gradient stochastiques, notamment lorsque celles ci arrivent à convergence. Cependant, on a également vu qu'elles sont assez sensibles à une mauvaise initialisation. Pour pallier ce problème, une solution peut être de considérer une version pondérée de la moyennisation (voir [MP11]). Cette pondération permet de donner plus de poids aux derniers estimateurs obtenus à l'aide de l'algorithme de gradient stochastique (qui sont censés être les meilleurs). On peut par exemple considérer un algorithme pondéré de la forme [BGB20]

$$\overline{m}_n = \frac{1}{\sum_{k=1}^{n} \log(k+1)^w} \sum_{k=1}^{n} \log(k+1)^w m_k$$

avec $w > 0$. Le terme $\log(k+1)$ permet donc de mettre plus de poids aux derniers estimateurs de gradient. Dans la Figure 4.5 on peut voir que dans le cas de la régression logistique où la mauvaise initialisation pouvait conduire à des résultats moyens en pratique, la pondération permet de pallier partiellement ce problème.

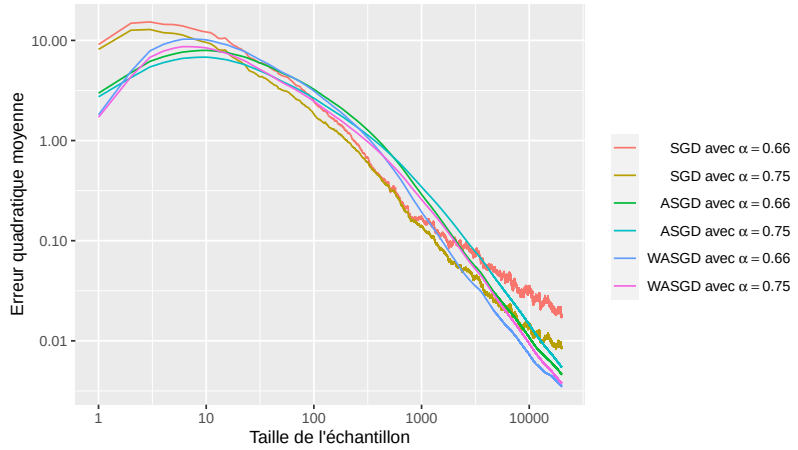

FIGURE 4.5 – Evolution de l'erreur quadratique moyenne par rapport à la taille de l'échantillon des estimateurs de gradients θ_n (SGD) et de leurs versions moyennées $\overline{\theta}_n$ (ASGD) dans le cadre de la régression logistique.

4.2 Algorithme de Newton stochastique

4.2.1 Idée de l'algorithme de Newton stochastique

Contrairement à ce que l'on peut faire en optimisation déterministe, on ne peut pas penser utiliser l'algorithme de Newton stochastique pour améliorer la vitesse de convergence par rapport aux estimateurs de gradient stochastiques moyennés. En effet, on a vu que sous certains critères de régularité, ceux-ci ont un comportement asymptotique optimal. L'idée est plutôt de créer une suite

de pas adaptée à toutes les directions du gradient, permettant de mieux traiter certains cas en pratique. En effet, rappelons que les estimateurs de gradient stochastique $(m_n)_n$ vérifient

$$\mathbb{E}\left[m_{n+1}|\mathcal{F}_n\right] = m_n - \gamma_{n+1}\nabla G\left(m_n\right).$$

Ainsi, lorsque $m_n \simeq m$, on a $\nabla G\left(m_n\right) \simeq \nabla^2 G(m)\left(m_n - m\right)$, et on a alors

$$\mathbb{E}\left[m_{n+1} - m|\mathcal{F}_n\right] \simeq m_n - m - \gamma_{n+1}\nabla^2 G(m)\left(m_n - m\right) = \left(I_d - \gamma_{n+1}\nabla^2 G\left(m\right)\right)\left(m_n - m\right).$$

Dans le cas où les valeurs propres de $\nabla^2 G(m)$ sont à des échelles très différentes, il n'est pas possible de régler le paramètre c_γ pour que le pas soit adapté à toutes les directions. Prenons l'exemple simple de la régression linéaire. Posons le modèle

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

avec $\epsilon \sim \mathcal{N}(0,1)$, $\theta \in \mathbb{R}^2$ et

$$X \sim \mathcal{N}\left(0, \begin{pmatrix} 10^{-2} & 0 \\ 0 & 10^2 \end{pmatrix}\right)$$

Il vient immédiatement que pour tout h,

$$\nabla^2 G(h) = \mathbb{E}\left[XX^T\right] = \begin{pmatrix} 10^{-2} & 0 \\ 0 & 10^2 \end{pmatrix}.$$

Ainsi, comme dans ce cadre on a exactement $\nabla G\left(m_n\right) = \nabla^2 G(m)\left(\theta_n - \theta\right)$, il vient, en notant $\theta^{(1)}$ et $\theta^{(2)}$ les premières coordonnées de θ, et en prenant les même notations pour θ_n,

$$\mathbb{E}\left[\theta_{n+1}^{(1)} - \theta^{(1)}|\mathcal{F}_n\right] = \left(1 - \frac{c_\gamma 10^{-2}}{(n+1)^\alpha}\right)\left(\theta_n^{(1)} - \theta^{(1)}\right)$$
$$\mathbb{E}\left[\theta_{n+1}^{(2)} - \theta^{(2)}|\mathcal{F}_n\right] = \left(1 - \frac{c_\gamma 10^{2}}{(n+1)^\alpha}\right)\left(\theta_n^{(2)} - \theta^{(2)}\right)$$

Ainsi, choisir c_γ proche de 10^2 permettrait d'avoir un pas adapté pour la première coordonnée mais ferait "exploser" la deuxième cordonnée, dans le sens où pour les premiers pas, on aurait des pas de l'ordre de 10^4. Faire l'inverse, i.e prendre $c_\gamma = 10^{-2}$, permettrait cette fois-ci d'avoir un pas adapté à la seconde coordonnée, mais on aurait un pas très petit pour la première coordonnée, et les estimateurs risqueraient de "ne pas bouger". Prendre un entre deux, i.e c_γ proche de 1, apporterait les deux problèmes, ce que semble indiquer la figure 4.6. A noter que pour la Figure 4.6, on a pris un cas normalement plus simple que le précédent, i.e on a pris des valeurs propres égales à 0.1 et 10.

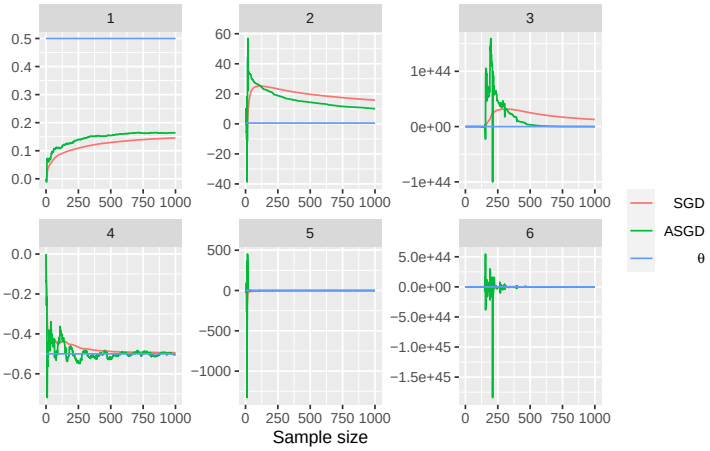

FIGURE 4.6 – Evolution des estimateurs de la première coordonnée (en haut) et de la deuxième (en bas) pour, de gauche à droite, $c_\gamma = 0.1, c_\gamma = 1$ et $c_\gamma = 10$.

Ainsi, une solution pour régler ce problème serait de supposer que $\nabla^2 G(m)$ est inversible et de considérer un algorithme de Newton stochastique, i.e un algorithme de la forme

$$m_{n+1} = m_n - \frac{1}{n+1} \nabla^2 G(m)^{-1} \nabla_h g\left(X_{n+1}, m_n\right).$$

Dans le cas de la régression linéaire, on aurait alors

$$\mathbb{E}\left[\theta_{n+1} - \theta | \mathcal{F}_n\right] = \theta_n - \theta - \frac{1}{n+1} \nabla^2 G(\theta)^{-1} \nabla^2 G(\theta) \left(\theta_n - \theta\right) = \left(1 - \frac{1}{n+1}\right) \left(\theta_n - \theta\right).$$

On aurait alors un biais $\mathbb{E}\left[\theta_n - \theta\right] = \frac{1}{n+1} \left(\theta_0 - \theta\right)$, et ce, quelles que soient les différences d'échelles entre les valeurs propres (tant qu'elles sont strictement positives). En effet, dans la figure 4.7, on voit bien que les estimateurs des deux coordonnées arrivent rapidement à convergence.

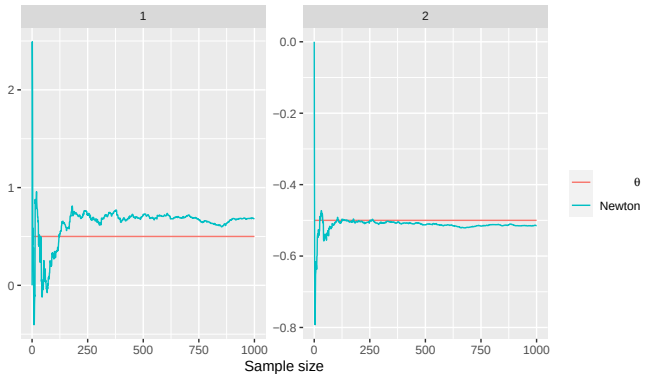

FIGURE 4.7 – Evolution des estimateurs de la première coordonnée (à gauche) et de la deuxième coordonnée (à droite) pour l'algorithme de Newton stochastique.

Cependant, on n'a généralement pas accès à la matrice Hessienne de G en m, et encore moins à son

inverse. On va donc remplacer $\nabla^2 G(m)$ par un estimateur.

4.2.2 L'algorithme de Newton stochastique

Dans tout ce qui suit, on note $H := \nabla^2 G(m)$ et on suppose que l'hypothèse **(PS2)** est vérifiée, i.e que H est inversible. L'algorithme de Newton stochastique est défini de manière récursive pour tout $n \geq 0$ par [BGB20]

$$\tilde{m}_{n+1} = \tilde{m}_n - \frac{1}{n+1} \overline{H}_n^{-1} \nabla_h g\left(X_{n+1}, \tilde{m}_n\right) \tag{4.2}$$

avec $\tilde{m}_0$ borné. De plus, $\overline{H}_n^{-1}$ est un estimateur récursif de H^{-1}, symétrique et défini positif, et il existe une filtration $(\mathcal{F}_n)$ telle que

- $\overline{H}_n^{-1}$ et $\tilde{m}_n$ soient $\mathcal{F}_n$-mesurable.
- X_{n+1} est indépendant de $\mathcal{F}_n$.

A noter que si on considère la filtration générée par l'échantillon et si $\overline{H}_n^{-1}$ ne dépend que de $X_1, \ldots, X_n$ et $\tilde{m}_0, \ldots, \tilde{m}_n$, alors les hypothèses sur la filtration sont vérifiées. On verra dans les Sections 4.2.5 et 4.2.6 comment construire de tels estimateurs en ligne de l'inverse de la Hessienne pour les régressions linéaires et logistiques. Afin d'assurer la convergence des estimateurs obtenus à l'aide de l'algorithme de Newton stochastique, on suppose maintenant que l'hypothèse suivante est vérifiée :

(PS5) La Hessienne de G est uniformément bornée, i.e il existe une constante $L_{\nabla G}$ telle que pour tout $h \in \mathbb{R}^d$,

$$\left\| \nabla^2 G(h) \right\|_{op} \leq L_{\nabla G}.$$

Cas avec gradient et estimateurs de la Hessienne bornés

Afin d'obtenir "facilement" la convergence des estimateurs obtenus à l'aide de l'algorithme de Newton stochastique, on fait d'abord l'hypothèse que l'on a un gradient "borné", i.e :

(PS0') On suppose qu'il existe une constante positive C telle que pour tout $h \in \mathbb{R}^d$:

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

De la même façon, pour simplifier les preuves et afin d'assurer la consistance des estimateurs, on supposera que les valeurs propres de l'estimateur de la Hessienne sont uniformément bornées, i.e :

(H0) Il existe des constantes strictement positives $\lambda_{\text{inf}}, \lambda_{\text{sup}}$ telles que pour tout $n \geq 0$,

$$\lambda_{\min}\left(\overline{H}_n^{-1}\right) \geq \lambda_{\text{inf}} \qquad \text{et} \qquad \lambda_{\max}\left(\overline{H}_n^{-1}\right) \leq \lambda_{\text{sup}}.$$

A noter que cette hypothèse est extrêmement restrictive. Cependant, on donne ce cadre de travail afin d'obtenir une première preuve assez simple de la consistance de $\tilde{m}_n$.

Théorème 4.2.1. *On suppose que les hypothèse **(PS0')**, **(PS2)**, **(PS5)** et **(H0)** sont vérifiées. Alors*

$$\tilde{m}_n \xrightarrow[n\to+\infty]{p.s} m.$$

Démonstration. A l'aide d'un développement de Taylor de la fonction G et grâce à l'hypothèse **(PS5)**,

$$\begin{aligned}
&G\left(\tilde{m}_{n+1}\right)\\
&\quad= G\left(\tilde{m}_n\right) + \nabla G\left(\tilde{m}_n\right)^T\left(\tilde{m}_{n+1}-\tilde{m}_n\right) + \left(\tilde{m}_{n+1}-\tilde{m}_n\right)^T \int_0^1 (1-t)\nabla^2 G\left(\tilde{m}_{n+1}+t\left(\tilde{m}_n-\tilde{m}_{n+1}\right)\right) dt \left(\tilde{m}_{n+1}-\tilde{m}_n\right)\\
&\quad\leq G\left(\tilde{m}_n\right) + \nabla G\left(\tilde{m}_n\right)^T\left(\tilde{m}_{n+1}-\tilde{m}_n\right) + \int_0^1 (1-t)\left\|\nabla^2 G\left(\tilde{m}_{n+1}+t\left(\tilde{m}_n-\tilde{m}_{n+1}\right)\right)\right\|_{op} dt \left\|\tilde{m}_{n+1}-\tilde{m}_n\right\|^2\\
&\quad\leq G\left(\tilde{m}_n\right) + \nabla G\left(\tilde{m}_n\right)^T\left(\tilde{m}_{n+1}-\tilde{m}_n\right) + \frac{L_{\nabla G}}{2}\left\|\tilde{m}_{n+1}-\tilde{m}_n\right\|^2
\end{aligned}$$

Alors, comme $\tilde{m}_{n+1}-\tilde{m}_n = -\frac{1}{n+1}\overline{H}_n^{-1}\nabla_h g\left(X_{n+1},\tilde{m}_n\right)$,

$$\begin{aligned}
G\left(\tilde{m}_{n+1}\right) &= G\left(\tilde{m}_n\right) - \frac{1}{n+1}\nabla G\left(\tilde{m}_n\right)^T \overline{H}_n^{-1}\nabla_h g\left(X_{n+1},\tilde{m}_n\right) + \frac{L_{\nabla G}}{2}\frac{1}{(n+1)^2}\left\|\overline{H}_n^{-1}\nabla_h g\left(X_{n+1},\tilde{m}_n\right)\right\|^2\\
&\leq G\left(\tilde{m}_n\right) - \frac{1}{n+1}\nabla G\left(\tilde{m}_n\right)^T \overline{H}_n^{-1}\nabla_h g\left(X_{n+1},\tilde{m}_n\right) + \frac{L_{\nabla G}}{2}\frac{1}{(n+1)^2}\left\|\overline{H}_n^{-1}\right\|_{op}^2\left\|\nabla_h g\left(X_{n+1},\tilde{m}_n\right)\right\|^2
\end{aligned}$$

On note maintenant $V_n = G\left(\tilde{m}_n\right) - G(m)$. On peut alors réécrire l'inégalité précédente comme

$$V_{n+1} \leq V_n - \frac{1}{n+1}\nabla G\left(\tilde{m}_n\right)^T \overline{H}_n^{-1}\nabla_h g\left(X_{n+1},\tilde{m}_n\right) + \frac{L_{\nabla G}}{2}\frac{1}{(n+1)^2}\left\|\overline{H}_n^{-1}\right\|_{op}^2\left\|\nabla_h g\left(X_{n+1},\tilde{m}_n\right)\right\|^2$$

et en prenant l'espérance conditionnelle, comme $\tilde{m}_n$ et $\overline{H}_n^{-1}$ sont $\mathcal{F}_n$-mesurables,

$$\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq V_n - \frac{1}{n+1}\nabla G\left(\tilde{m}_n\right)^T \overline{H}_n^{-1}\nabla G\left(\tilde{m}_n\right) + \frac{L_{\nabla G}}{2}\frac{1}{(n+1)^2}\left\|\overline{H}_n^{-1}\right\|_{op}^2 \mathbb{E}\left[\left\|\nabla_h g\left(X_{n+1},\tilde{m}_n\right)\right\|^2|\mathcal{F}_n\right]. \tag{4.3}$$

Alors, grâce à l'hypothèse **(PS0')**, il vient

$$\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq V_n - \frac{1}{n+1}\lambda_{\min}\left(\overline{H}_n^{-1}\right)\left\|\nabla G\left(\tilde{m}_n\right)\right\|^2 + \frac{CL_{\nabla G}}{2}\frac{1}{(n+1)^2}\left\|\overline{H}_n^{-1}\right\|_{op}^2$$

Remarquons que grâce à l'hypothèse **(H0)**, on peut réécrire l'inégalité précédente comme

$$\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq V_n - \frac{1}{n+1}\lambda_{\inf}\left\|\nabla G\left(\tilde{m}_n\right)\right\|^2 + \frac{CL_{\nabla G}}{2}\frac{1}{(n+1)^2}\lambda_{\sup}^2.$$

De plus, on a

$$\sum_{n\geq 0}\frac{1}{(n+1)^2}\frac{1}{2}CL_{\nabla G}\lambda_{\sup}^2 < +\infty \quad p.s.$$

et grâce au théorème de Robbins-Siegmund, V_n converge presque sûrement vers une variable aléatoire finie et

$$\sum_{n\geq 0} \frac{1}{n+1} \lambda_{\inf} \left\|\nabla G\left(\tilde{m}_n\right)\right\|^2 < +\infty \quad p.s.$$

Cela implique nécessairement que $\liminf_n \left\|\nabla G\left(\tilde{m}_n\right)\right\| = 0$ presque sûrement. Comme G est strictement convexe, cela implique aussi que

$$\liminf_n \left\|\tilde{m}_n - m\right\| = 0 \quad p.s \qquad \text{et} \qquad \liminf_n V_n = \liminf_n G\left(\tilde{m}_n\right) - G(m) = 0 \quad p.s,$$

et comme V_n converge presque sûrement vers une variable aléatoire, cela implique que $G\left(\tilde{m}_n\right)$ converge presque sûrement vers $G(m)$ et par stricte convexité, que $\tilde{m}_n$ converge presque sûrement vers m. □

Cas général

On peut obtenir la convergence des estimateurs en faisant des hypothèses moins fortes sur les valeurs propres de l'estimateur de H^{-1}. Plus précisément, on peut remplacer l'hypothèse **(H0)** par :

(H1) On peut contrôler les plus grandes valeurs propres de $\overline{H}_n$ et $\overline{H}_n^{-1}$: il existe $\beta \in (0, 1/2)$ tel que

$$\lambda_{\max}\left(\overline{H}_n\right) = O(1) \quad p.s. \qquad \text{et} \qquad \lambda_{\max}\left(\overline{H}_n^{-1}\right) = O\left(n^{\beta}\right) \quad p.s.$$

Cette hypothèse implique que, sans même savoir si $\tilde{m}_n$ converge, on doit pouvoir contrôler le comportement des valeurs propres (plus petite et plus grande) de l'estimateur de la Hessienne. On verra Section 4.2.6 comment modifier les estimateurs récursifs naturels de la Hessienne afin que cette hypothèse soit vérifiée. De plus, on peut remplacer l'hypothèse **(PS0')** par l'hypothèse suivante :

(PS0") Il existe des constantes positives C, C' telles que pour tout $h \in \mathbb{R}^d$, on ait

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

A noter que l'hypothèse **(PS0")** n'est pas beaucoup plus restrictive que **(PS0)**. En effet, dans le cas de la régression logistique, on peut voir qu'elles sont toutes les deux vérifiées avec un minimum d'hypothèses sur la variable explicative X. De plus, si la fonction est μ-fortement convexe, on a pour tout $h \in \mathbb{R}^d$, $\|h - m\|^2 \leq \frac{2}{\mu}(G(h) - G(m))$, et l'hypothèse **(PS0)** implique l'hypothèse **(PS0")**. De la même façon, si les hypothèses **(PS0")** et **(PS5)** sont vérifiées, on a $G(h) - G(m) \leq \frac{L_{\nabla G}}{2}\|h - m\|^2$, et l'hypothèse **(PS0)** est alors vérifiée. On peut maintenant montrer la forte consistance des estimateurs.

Théorème 4.2.2. *On suppose que les hypothèse **(PS0")**, **(PS2)**, **(PS5)** et **(H1)** sont vérifiées. Alors*

$$\tilde{m}_n \xrightarrow[n\to+\infty]{p.s} m.$$

On admettra ce théorème mais la preuve est disponible dans [BGB20] ainsi que dans la version longue.

4.2.3 Vitesses de convergence

Afin de d'obtenir les vitesses de convergence, on est "obligé" d'avoir la forte consistance de l'estimateur de la Hessienne et de son inverse. Dans ce but, on introduit l'hypothèse suivante :

(H2) Si les hypothèses **(PS0")**, **(PS2)**,**(PS5)** et **(H1)** sont vérifiées, alors

$$\overline{H}_n \xrightarrow[n\to+\infty]{p.s} H \qquad \text{et} \qquad \overline{H}_n^{-1} \xrightarrow[n\to+\infty]{p.s} H^{-1}.$$

Cette hypothèse veut tout simplement dire que si l'on a la convergence de $\tilde{m}_n$, on a également la convergence de l'estimateur de la Hessienne. On peut maintenant donner la vitesse de convergence de l'algorithme de Newton stochastique.

Théorème 4.2.3. *On suppose que les hypothèses **(PS0")**, **(PS2)**, **(PS4)**, **(PS5)**, **(H1)** et **(H2)** sont vérifiées. Alors pour tout $\delta > 0$,*

$$\|\tilde{m}_n - m\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) \quad p.s..$$

*De plus, si l'hypothèse **(PS1)** est vérifiée, alors*

$$\|\tilde{m}_n - m\|^2 = O\left(\frac{\ln n}{n}\right) \quad p.s.$$

Démonstration. Remarquons d'abord que l'on peut écrire l'algorithme de Newton comme

$$\tilde{m}_{n+1} - m = \tilde{m}_n - m - \frac{1}{n+1}\overline{H}_n^{-1}\nabla G\left(\tilde{m}_n\right) + \frac{1}{n+1}\overline{H}_n^{-1}\tilde{\xi}_{n+1}, \tag{4.4}$$

avec $\tilde{\xi}_{n+1} := \nabla_h g\left(X_{n+1}, \tilde{m}_n\right) - \nabla G\left(\tilde{m}_n\right)$. On a donc que $\left(\tilde{\xi}_n\right)$ est une suite de différences de martingale par rapport à la filtration $(\mathcal{F}_n)$. En linéarisant le gradient, on a alors

$$\tilde{m}_{n+1} - m = \tilde{m}_n - m - \frac{1}{n+1}\overline{H}_n^{-1}H\left(\tilde{m}_n - m\right) - \frac{1}{n+1}\overline{H}_n^{-1}\tilde{\delta}_n + \frac{1}{n+1}\overline{H}_n^{-1}\tilde{\xi}_{n+1}$$

où $\tilde{\delta}_n := \nabla G\left(\tilde{m}_n\right) - H\left(\tilde{m}_n - m\right)$ est le terme de reste dans la décomposition de Taylor du gradient. De plus, on peut réécrire l'égalité précédente comme

$$\begin{aligned}\tilde{m}_{n+1} - m &= \tilde{m}_n - m - \frac{1}{n+1}H^{-1}H\left(\tilde{m}_n - m\right) - \frac{1}{n+1}\left(\overline{H}_n^{-1} - H^{-1}\right)H\left(\tilde{m}_n - m\right) - \frac{1}{n+1}\overline{H}_n^{-1}\tilde{\delta}_n + \frac{1}{n+1}\overline{H}_n^{-1}\tilde{\xi}_{n+1} \\ &= \left(1 - \frac{1}{n+1}\right)\left(\tilde{m}_n - m\right) - \frac{1}{n+1}\left(\overline{H}_n^{-1} - H^{-1}\right)H\left(\tilde{m}_n - m\right) - \frac{1}{n+1}\overline{H}_n^{-1}\tilde{\delta}_n + \frac{1}{n+1}\overline{H}_n^{-1}\tilde{\xi}_{n+1}.\end{aligned} \tag{4.5}$$

On peut donc montrer (récurrence immédiate) que pour tout $n \geq 1$,

$$\tilde{m}_n - m = \underbrace{-\frac{1}{n}\sum_{k=0}^{n-1}\left(\overline{H}_k^{-1} - H^{-1}\right) H\left(\tilde{m}_k - m\right) - \frac{1}{n}\sum_{k=0}^{n-1}\overline{H}_k^{-1}\tilde{\delta}_k}_{=:\tilde{\Delta}_n} + \underbrace{\frac{1}{n}\sum_{k=0}^{n-1}\overline{H}_k^{-1}\tilde{\xi}_{k+1}}_{=:\tilde{M}_n}. \tag{4.6}$$

A noter que $\tilde{M}_n$ est une martingale par rapport à la filtration $(\mathcal{F}_n)$. A noter que grâce aux hypothèses **(PS0")** et **(PS5)**, on a

$$\mathbb{E}\left[\left\|\tilde{\xi}_{n+1}\right\|^2 |\mathcal{F}_n\right] \leq \mathbb{E}\left[\left\|\nabla_h g\left(X_{n+1}, \tilde{m}_n\right)\right\|^2 |\mathcal{F}_n\right] \leq C + C'\left(G\left(\tilde{m}_n\right) - G(m)\right) \leq C + \frac{L_{\nabla G} C'}{2}\left\|\tilde{m}_n - m\right\|^2.$$

Vitesse de convergence de $\tilde{M}_n$. A noter que $\tilde{M}_n$ est une martingale mais pas nécéssairement de carré intégrable. Cependant, on peut considérer la suite d'évènements $(A_n)_{n\geq 0}$ définie pour tout n par $A_n := \left\{\left\|\overline{H}_n^{-1}\right\|_{op} \leq \frac{2}{\lambda_{\min}}, \|m_n - m\| \leq 1\right\}$. Comme m_n converge presque sûrement vers m et comme $\overline{H}_n^{-1}$ converge presque sûrement vers H^{-1} (et que $\left\|H^{-1}\right\| = \lambda_{\min}^{-1}$), $\mathbf{1}_{A_n}$ converge presque sûrement vers 1. On peut maintenant réécrire $\tilde{M}_n$ comme

$$\tilde{M}_n = \underbrace{\sum_{k=0}^{n-1}\overline{H}_k\tilde{\xi}_{k+1}\mathbf{1}_{A_k}}_{=:\tilde{M}_{1,n}} + \underbrace{\sum_{k=0}^{n-1}\overline{H}_k\tilde{\xi}_{k+1}\mathbf{1}_{A_k^C}}_{=:\tilde{M}_{2,n}}$$

et comme $\mathbf{1}_{A_n^C}$ converge presque sûrement vers 0, on a

$$\frac{1}{n}\left\|\tilde{M}_{2,n}\right\| = O\left(\frac{1}{n}\right) \quad p.s.$$

et ce terme est donc négligeable. Afin d'obtenir la vitesse de convergence du terme de martingale $\tilde{M}_{1,n}$, on va calculer son crochet. Par linéarité de l'espérance, on a pour tout $n \geq 1$,

$$\begin{aligned}
\left\langle \tilde{M}_1 \right\rangle_n &= \sum_{k=0}^{n-1}\mathbb{E}\left[\overline{H}_k^{-1}\tilde{\xi}_{k+1}\left(\overline{H}_k^{-1}\tilde{\xi}_{k+1}\right)^T |\mathcal{F}_k\right]\mathbf{1}_{A_k} \\
&= \sum_{k=0}^{n-1}\overline{H}_k^{-1}\mathbb{E}\left[\tilde{\xi}_{k+1}\tilde{\xi}_{k+1}^T|\mathcal{F}_k\right]\overline{H}_k^{-1}\mathbf{1}_{A_k} \\
&= \sum_{k=0}^{n-1}\overline{H}_k^{-1}\underbrace{\mathbb{E}\left[\nabla_h g\left(X_{k+1}, \tilde{m}_k\right)\nabla_h g\left(X_{k+1}, \tilde{m}_k\right)^T |\mathcal{F}_k\right]}_{=\Sigma(\tilde{m}_k)}\overline{H}_k^{-1}\mathbf{1}_{A_k} - \sum_{k=0}^{n-1}\overline{H}_k^{-1}\nabla G\left(\tilde{m}_k\right)\nabla G\left(\tilde{m}_k\right)^T\overline{H}_k^{-1}\mathbf{1}_{A_k}
\end{aligned}$$

Comme $\overline{H}_k^{-1}, \tilde{m}_k, \mathbf{1}_{A_k}$ convergent presque sûrement vers $H^{-1}, m, 1$, et par continuité de Σ (hypothèse **(PS4)**), on a

$$\overline{H}_k^{-1}\mathbb{E}\left[\nabla_h g\left(X_{k+1}, \tilde{m}_k\right)\nabla_h g\left(X_{k+1}, \tilde{m}_k\right)^T |\mathcal{F}_k\right]\overline{H}_k^{-1}\mathbf{1}_{A_k} \xrightarrow[n\to+\infty]{p.s} H^{-1}\Sigma(m)H^{-1}.$$

Ainsi, en appliquant le lemme de Toeplitz,

$$\frac{1}{n}\sum_{k=1}^{n-1}\overline{H}_k^{-1}\mathbb{E}\left[\nabla_h g\left(X_{k+1},\tilde{m}_k\right)\nabla_h g\left(X_{k+1},\tilde{m}_k\right)^T|\mathcal{F}_k\right]\overline{H}_k^{-1}\mathbf{1}_{A_k}\xrightarrow[n\to+\infty]{p.s}H^{-1}\Sigma(m)H^{-1}.$$

De plus, comme ∇G est $L_{\nabla G}$-lipschitz, et comme $\tilde{m}_k$ converge presque sûrement vers m et comme $\overline{H}_n^{-1}$ converge vers H^{-1}, on a

$$\left\|\overline{H}_k^{-1}\nabla G\left(\tilde{m}_k\right)\nabla G\left(\tilde{m}_k\right)^T\overline{H}_k^{-1}\right\|_{op}\leq\left\|\overline{H}_k^{-1}\right\|_{op}^2\left\|\nabla G\left(\tilde{m}_k\right)\right\|^2\leq L_{\nabla G}^2\left\|\overline{H}_k^{-1}\right\|_{op}^2\left\|\tilde{m}_k-m\right\|^2\xrightarrow[n\to+\infty]{p.s}0.$$

En particulier, en appliquant le lemme de Toeplitz, on obtient

$$\text{`}\frac{1}{n}\sum_{k=0}^{n-1}\overline{H}_k^{-1}\nabla G\left(\tilde{m}_k\right)\nabla G\left(\tilde{m}_k\right)^T\overline{H}_k^{-1}\mathbf{1}_{A_k}\xrightarrow[n\to+\infty]{p.s}0$$

et donc

$$\frac{1}{n}\left\langle\tilde{M}_1\right\rangle_n\xrightarrow[n\to+\infty]{p.s}H^{-1}\Sigma(m)H^{-1},$$

et en appliquant une loi des grands nombres pour les martingales vectorielles, on a pour tout $\delta>0$

$$\left\|\frac{1}{n+1}\tilde{M}_n\right\|^2=o\left(\frac{(\ln n)^{1+\delta}}{n}\right)\quad p.s$$

De plus, si **(PS1)** est vérifiée, on a

$$\left\|\frac{1}{n+1}\tilde{M}_n\right\|^2=O\left(\frac{\ln n}{n}\right)\quad p.s$$

Vitesse de convergence de $\tilde{\Delta}_n$. Remarquons d'abord que grâce aux hypothèses **(PS3)** et **(PS5)**, et comme $\tilde{m}_k$ et $\overline{H}_n^{-1}$ convergent presque sûrement vers m et H^{-1}, on a

$$\left\|\left(\overline{H}_k^{-1}-H^{-1}\right)H\left(\tilde{m}_k-m\right)\right\|=o\left(\left\|\tilde{m}_k-m\right\|\right)\quad p.s\qquad\text{et}\qquad\left\|\overline{H}_k^{-1}\tilde{\delta}_k\right\|=o\left(\left\|\tilde{m}_k-m\right\|\right)\quad p.s.$$

Soit $c\in(0,1)$. On introduit maintenant la suite d'évènement $(A_{k,c})$ définie pour tout $k\geq1$ par

$$A_{k,c}=\left\{\left\|\left(\overline{H}_k^{-1}-H^{-1}\right)H\nabla G\left(\tilde{m}_k\right)+\overline{H}_k^{-1}\tilde{\delta}_k\right\|\leq c\left\|\tilde{m}_k-m\right\|\right\}.$$

D'après ce qui précède, $\mathbf{1}_{A_{k,c}^C}$ converge presque sûrement vers 0. On peut alors réécrire $\left\|\tilde{\Delta}_{n+1}\right\|$ comme

$$\begin{aligned}\left\|\tilde{\Delta}_{n+1}\right\|&=\left\|\left(1-\frac{1}{n+1}\right)\tilde{\Delta}_n+\frac{1}{n+1}\left(\overline{H}_n^{-1}-H^{-1}\right)H\nabla G\left(\tilde{m}_n\right)+\frac{1}{n+1}\overline{H}_n^{-1}\tilde{\delta}_n\right\|\\&\leq\left(1-\frac{1}{n+1}\right)\left\|\tilde{\Delta}_n\right\|+\frac{c}{n+1}\left\|\tilde{m}_n-m\right\|\mathbf{1}_{A_n}+\frac{1}{n+1}\left\|\left(\overline{H}_n^{-1}-H^{-1}\right)H\nabla G\left(\tilde{m}_n\right)+\overline{H}_n^{-1}\tilde{\delta}_n\right\|\mathbf{1}_{A_{n,c}^C}\end{aligned}$$

Comme $\|\tilde{m}_n - m\| \leq \left\|\tilde{\Delta}_n\right\| + \frac{1}{n}\tilde{M}_n$, on obtient

$$\left\|\tilde{\Delta}_{n+1}\right\| \leq \left(1 - \frac{(1-c)}{n+1}\right)\left\|\tilde{\Delta}_n\right\| + \frac{c}{(n+1)n}\left\|\tilde{M}_n\right\| + \frac{1}{n+1}\left\|\left(\overline{H}_n^{-1} - H^{-1}\right) H\nabla G\left(\tilde{m}_n\right) + \overline{H}_n^{-1}\tilde{\delta}_n\right\| \mathbf{1}_{A_{n,c}^C}$$

En notant $\tilde{c} = 1 - c$, comme $1 + x \leq e^x$ et à l'aide d'une comparaison série intégrale,

$$\prod_{j=k+1}^{n}\left(1 - \frac{\tilde{c}}{j}\right) \leq \exp\left(-\tilde{c}\sum_{j=k+1}^{n}\frac{1}{j}\right) \leq \exp\left(-\tilde{c}\left(\ln(n+1) - \ln(k+1)\right)\right) \leq \left(\frac{k+1}{n+1}\right)^{\tilde{c}}$$

Ainsi, on peut montrer à l'aide d'une récurrence que pour tout $n \geq 0$,

$$\left\|\tilde{\Delta}_n\right\| \leq \underbrace{\frac{1}{(n+1)^{\tilde{c}}}\sum_{k=1}^{n-1}(k+1)^{\tilde{c}}\frac{1}{(k+1)k}\left\|\tilde{M}_k\right\|}_{:=R_{3,n}} + \underbrace{\frac{1}{(n+1)^{\tilde{c}}}\sum_{k=0}^{n-1}\frac{(k+1)^{\tilde{c}}}{k+1}\left\|\left(\overline{H}_k^{-1} - H^{-1}\right) H\nabla G\left(\tilde{m}_k\right) + \overline{H}_n^{-1}\tilde{\delta}_k\right\| \mathbf{1}_{A_{k,c}^C}}_{:=R_{4,n}} \tag{4.7}$$

Vitesse de convergence de $R_{4,n}$. Comme $\mathbf{1}_{A_{k,c}^C}$ converge presque sûrement vers 0, on a

$$\sum_{k=0}^{n-1}\frac{(k+1)^{\tilde{c}}}{k+1}\left\|\left(\overline{H}_k^{-1} - H^{-1}\right) H\nabla G\left(\tilde{m}_k\right) + \overline{H}_n^{-1}\tilde{\delta}_k\right\| \mathbf{1}_{A_{k,c}^C} < +\infty \quad p.s.$$

et on obtient donc

$$R_{4,n} = O\left(\frac{1}{(n+1)^{\tilde{c}}}\right) \quad p.s$$

qui est négligeable dès que $\tilde{c} > 1/2$, i.e dès que $c < 1/2$.

Vitesse de convergence de $R_{3,n}$. Comme on a la vitesse de convergence de $\tilde{M}_n$, pour tout $\delta \geq 0$ (le cas $\delta = 0$ correspondant au cas où le gradient admet un moment conditionnel d'ordre strictement plus grand que 2), il existe une variable aléatoire positive B_δ telle que pour tout $k \geq 1$,

$$\left\|\tilde{M}_k\right\| \leq B_\delta \ln(k+1)^{1/2+\delta}\sqrt{k+1} \quad p.s.$$

Ainsi, on a

$$R_{3,n} \leq \frac{B_\delta}{(n+1)^{\tilde{c}}}\sum_{k=1}^{n-1}(k+1)^{\tilde{c}}\frac{1}{(k+1)k}\ln(k+1)^{1/2+\delta}\sqrt{k+1} = \begin{cases} O\left(\frac{\ln(n+1)^{1/2+\delta}}{(n+1)^{\tilde{c}}}\right) p.s & \text{si } \tilde{c} < 1/2 \\ O\left(\frac{\ln(n+1)^{1/2+\delta}}{\sqrt{n+1}}\right) p.s & \text{si } \tilde{c} > 1/2 \end{cases}$$

et on conclut la preuve en prenant $\tilde{c} > 1/2$, i.e en prenant $c < 1/2$. □

4.2.4 Normalité asymptotique

Afin d'obtenir la normalité asymptotique des estimateurs, on est obligé d'avoir une vitesse de convergence de l'estimateur de la Hessienne. Pour ce faire, on introduit l'hypothèse suivante :

(H3) Si les hypothèses **(PS0")**, **(PS2)**, **(PS3)**, **(PS4)**, **(PS5)**, **(H1)** et **(H2)** sont vérifiées, alors il existe $p_H > 0$ tel que

$$\left\|\overline{H}_n - H\right\|_{op}^2 = O\left(\frac{1}{n^{p_H}}\right) \quad p.s.$$

Cette hypothèse signifie juste que obtenir une vitesse de convergence presque sûre de l'estimateur $\tilde{m}_n$ permet d'obtenir une vitesse de convergence presque sûre de $\overline{H}_n$. Le théorème suivant nous donne la normalité asymptotique des estimateurs obtenus à l'aide d'algorithmes de Newton stochastiques. En particulier, il nous indique que ces estimateurs ont un comportement asymptotique optimal, dans le sens où ils ont le même comportement asymptotique que les M-estimateurs.

Théorème 4.2.4. *On suppose que les hypothèses **(PS0")**, **(PS1)**, **(PS2)**, **(PS3)**, **(PS4)**, **(PS5)** **(H1)**, **(H2)** et **(H3)** sont vérifiées. Alors*

$$\sqrt{n}\left(\tilde{m}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right)$$

avec $\Sigma = \Sigma(m) = \mathbb{E}\left[\nabla_h g\left(X, m\right) \nabla_h g\left(X, m\right)^T\right]$.

Démonstration. On rappelle que l'on a

$$\tilde{m}_n - m = \underbrace{-\frac{1}{n}\sum_{k=0}^{n-1}\left(\overline{H}_k^{-1} - H^{-1}\right) H\left(\tilde{m}_k - m\right) - \frac{1}{n}\sum_{k=0}^{n-1}\overline{H}_k^{-1}\tilde{\delta}_k}_{=:\tilde{\Delta}_n} + \underbrace{\frac{1}{n}\sum_{k=0}^{n-1}\overline{H}_k^{-1}\tilde{\xi}_{k+1}}_{=:\tilde{M}_n}.$$

et on va donc donner les vitesses de convergence de ces termes pour montrer que c'est bien le terme de martingale qui "porte" la convergence.

Vitesse de convergence de $\tilde{\Delta}_n$. Comme $\tilde{m}_n$ converge presque sûrement vers m et grâce à l'hypothèse **(PS3)**, on a

$$\left\|\tilde{\delta}_n\right\| = O\left(\left\|\tilde{m}_n - m\right\|^2\right) \quad p.s.$$

et donc, comme $\overline{H}_n^{-1}$ converge presque sûrement vers H^{-1}, en appliquant le lemme de Toeplitz et grâce au Théorème 4.2.3, on obtient que pour tout $\delta > 0$,

$$\left\|\sum_{k=0}^{n-1}\overline{H}_k^{-1}\tilde{\delta}_k\right\| \leq \sum_{k=0}^{n-1}\frac{(\ln(k+1))^{1+\delta}}{k+1}\left\|\overline{H}_k^{-1}\right\|_{op}\left(\frac{k+1}{\ln(k+1)^{1+\delta}}\left\|\tilde{\delta}_k\right\|\right) = o\left(\sum_{k=0}^{n-1}\frac{\ln(k+1)^{2+\delta}}{k+1}\right) \quad p.s$$

et donc,

$$\left\|\frac{1}{n}\sum_{k=0}^{n-1}\overline{H}_k^{-1}\tilde{\delta}_k\right\| = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) \quad p.s.$$

Ce terme est donc négligeable. De la même façon, en appliquant le lemme de Toeplitz, grâce au

Théorème 4.2.3 et à l'hypothèse **(H3)**, on a

$$\begin{aligned}
&\left\| \sum_{k=0}^{n-1} \left(\overline{H}_k^{-1} - H^{-1} \right) H \left(\tilde{m}_k - m \right) \right\| \\
&\leq \sum_{k=0}^{n-1} \frac{\ln(k+1)^{\frac{1}{2}+\delta}}{(k+1)^{\frac{1+p_H}{2}}} \left\| \overline{H} \right\|_{op} \left((k+1)^{\frac{p_H}{2}} \left\| \overline{H}_k^{-1} - H^{-1} \right\|_{op} \right) \left(\frac{\sqrt{k+1}}{\ln(k+1)^{\frac{1}{2}+\delta}} \left\| \tilde{m}_k - m \right\| \right) \\
&= o \left(\sum_{k=0}^{n-1} \frac{\ln(k+1)^{1/2+\delta}}{(k+1)^{1/2+p_H/2}} \right) \quad p.s
\end{aligned}$$

et on obtient donc

$$\left\| \frac{1}{n} \sum_{k=0}^{n-1} \left(\overline{H}_k^{-1} - H^{-1} \right) H \left(\tilde{m}_k - m \right) \right\| = o \left(\frac{(\ln n)^{1/2+\delta}}{n^{1/2+p_H}} \right) \quad p.s$$

et ce comme $p_H > 0$, ce terme est négligeable.

Vitesse de convergence de $\tilde{M}_n$. On a déjà montré que $\tilde{M}_n = \tilde{M}_{1,n} + \tilde{M}_{2,n}$ où $\tilde{M}_{2,n}$ est négligeable et $\tilde{M}_{1,n}$ est une martingale dont le crochet vérifie

$$\frac{1}{n} \left\langle \tilde{M}_1 \right\rangle_n \xrightarrow[n \to +\infty]{p.s} H^{-1} \Sigma H^{-1}.$$

Il faut donc montrer que la condition de Lindeberg est vérifiée. Pour cela, il suffit de voir que grâce à l'hypothèse **(PS1)**, et comme $\left\| \tilde{\xi}_{k+1} \right\| \leq \left\| \nabla_h g \left(X_{k+1}, \tilde{m}_k \right) \right\| + \left\| \mathbb{E} \left[\nabla_h g \left(X_{k+1}, \tilde{m}_k \right) | \mathcal{F}_k \right] \right\|$, on a

$$\mathbb{E} \left[\left\| \overline{H}_k^{-1} \tilde{\xi}_{k+1} \right\|^{2+2\eta} \mathbf{1}_{A_k} | \mathcal{F}_k \right] \leq 2^{1+2\eta} C_\eta \left(1 + \left\| \tilde{m}_k - m \right\|^{1+2\eta} \right) \left\| \overline{H}_k^{-1} \right\|_{op}^2 \mathbf{1}_{A_k} \leq 2^{4+2\eta} \lambda_{\min}^{-2}.$$

Le TLC pour les martingales vectorielles nous donne donc

$$\sqrt{n} \tilde{M}_{1,n} \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N} \left(0, H^{-1} \Sigma H^{-1} \right)$$

ce qui conclut la preuve. □

4.2.5 Application au modèle linéaire

On se place maintenant dans le cadre de la régression linéaire défini par (1.1). On rappelle que la Hessienne de la fonction à minimiser G est définie pour tout $h \in \mathbb{R}^d$ par $\nabla^2 G(h) = \mathbb{E} \left[X X^T \right]$, et on supposera que cette matrice est définie positive. Un estimateur naturel de $H = \nabla^2 G(\theta)$ est donc

$$\overline{H}_n = \frac{1}{n+1} \left(\sum_{k=1}^{n} X_k X_k^T + H_0 \right)$$

où H_0 est une matrice symétrique définie positive (on peut prendre $H_0 = I_d$ par exemple). A noter que l'on peut écrire $\overline{H}_n$ de manière récursive comme

$$\overline{H}_{n+1} = \overline{H}_n + \frac{1}{n+2}\left(X_{n+1}X_{n+1}^T - \overline{H}_n\right).$$

On va maintenant s'intéresser à l'inversion de la matrice $\overline{H}_n$. Pour cela, on va introduire la matrice $H_n = (n+1)\overline{H}_n$ que l'on peut écrire comme

$$H_{n+1} = H_n + X_{n+1}X_{n+1}^T.$$

Afin de réduire le temps de calcul, on ne va pas inverser directement cette matrice à chaque itération, mais plutôt utiliser la formule d'inversion de Ricatti (aussi appelée formule de Sherman-Morrison) suivante :

Théorème 4.2.5 (Formule de Riccati)**.** *Soit $A \in \mathcal{M}_d(\mathbb{R})$ une matrice inversible et $u, v \in \mathbb{R}^d$. Si $1 + v^T A^{-1} u \neq 0$, alors $A + uv^T$ est inversible et*

$$\left(A + uv^T\right)^{-1} = A^{-1} - \left(1 + v^T A^{-1} u\right)^{-1} A^{-1} u v^T A^{-1}.$$

La preuve est évidente et est disponible dans la version longue. En particulier, si A est une matrice définie positive, pour tout $u \in \mathbb{R}^d$ on a $1 + u^T A^{-1} u \geq 1$ et donc

$$\left(A + uu^T\right) = A^{-1} - \left(1 + u^T A^{-1} u\right)^{-1} A^{-1} u u^T A^{-1}.$$

A noter que cette opération ne représente "que" $O\left(d^2\right)$ opérations. On peut donc mettre à jour l'estimateur de l'inverse de la Hessienne comme suit :

$$H_{n+1}^{-1} = H_n^{-1} - \left(1 + X_{n+1}^T H_n^{-1} X_{n+1}\right)^{-1} H_n^{-1} X_{n+1} X_{n+1}^T H_n^{-1}$$

et $\overline{H}_{n+1}^{-1} = (n+2)H_{n+1}^{-1}$. Ainsi, une fois que l'on a H_0^{-1}, on peut facilement mettre à jours nos estimateurs, ce qui conduit à l'algorithme de Newton stochastique suivant :

$$\begin{aligned}
\tilde{\theta}_{n+1} &= \tilde{\theta}_n + \frac{1}{n+1}\overline{H}_n^{-1}\left(Y_{n+1} - \tilde{\theta}_n^T X_{n+1}\right) X_{n+1} \\
H_{n+1}^{-1} &= H_n^{-1} - \left(1 + X_{n+1}^T H_n^{-1} X_{n+1}\right)^{-1} H_n^{-1} X_{n+1} X_{n+1}^T H_n^{-1}
\end{aligned}$$

avec $\overline{H}_n^{-1} = (n+1)H_n^{-1}$. A noter que l'on pourrait réécrire l'algorithme comme

$$\begin{aligned}
\tilde{\theta}_{n+1} &= \tilde{\theta}_n + H_n^{-1}\left(Y_{n+1} - \tilde{\theta}_n^T X_{n+1}\right) X_{n+1} \\
H_{n+1}^{-1} &= H_n^{-1} - \left(1 + X_{n+1}^T H_n^{-1} X_{n+1}\right)^{-1} H_n^{-1} X_{n+1} X_{n+1}^T H_n^{-1}.
\end{aligned}$$

On peut alors obtenir les vitesses de convergence des estimateurs, et ce, avec des hypothèses assez faibles.

Théorème 4.2.6. *Si il existe $\eta > 0$ tel que X et ϵ admettent des moments d'ordre $4+4\eta$ et $2+2\eta$, et si $\mathbb{E}\left[XX^T\right]$ est définie positive, alors*

$$\left\|\tilde{\theta}_n - \theta\right\|^2 = O\left(\frac{\ln n}{n}\right) \quad p.s. \qquad et \qquad \sqrt{n}\left(\tilde{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2 H^{-1}\right).$$

Démonstration. On a déjà vu que sous ces hypothèses, les hypothèses **(PS1)** à **(PS4)** sont vérifiées et l'hypothèse **(PS0")** est clairement vérifiée. De plus pour tout $h \in \mathbb{R}^d$, on a

$$\left\|\nabla^2 G(h)\right\|_{op} = \left\|\mathbb{E}\left[XX^T\right]\right\|_{op} \leq \mathbb{E}\left[\|X\|^2\right]$$

et l'hypothèse **(PS5)** est donc vérifiée. De plus, comme X admet un moment d'ordre 4, on a par la loi du log-itéré

$$\left\|\overline{H}_n - H\right\|^2 = O\left(\frac{\ln \ln n}{n}\right) \quad p.s$$

et les hypothèses **(H1)** à **(H3)** sont vérifiées, ce qui conclut la preuve. □

Dans la Figure 4.8, on considère le modèle

$$\theta = (-4,-3,-2,-1,0,1,2,3,4,5)^T \in \mathbb{R}^{10}, \qquad X \sim \mathcal{N}\left(0, \mathrm{diag}\left(\sigma_i^2\right)\right), \qquad \text{et} \qquad \epsilon \sim \mathcal{N}(0,1)$$

avec pour tout $i = 1,\dots,d$, $\sigma_i^2 = \frac{i^2}{d^2}$. On a donc la plus grande valeur propre de la Hessienne qui est 100 fois plus grande que la plus petite. On voit bien Figure 4.8 que les estimateurs de gradient peinent à arriver à convergence, et donc, que les estimateurs moyennés ne permettent pas d'accélérer la convergence. A contrario, on voit bien que malgré ces différences d'échelles entre les valeurs propres de la Hessienne, l'algorithme de Newton stochastique converge très rapidement.

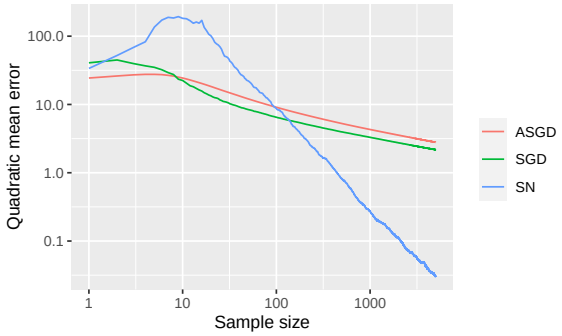

FIGURE 4.8 – Evolution de l'erreur quadratique moyenne des estimateurs de gradient θ_n (SGD), de leur version moyennée $\overline{\theta}_n$ (ASGD) et des estimateurs de Newton stochastique $\tilde{\theta}_n$ (SN) en fonction de la taille de l'échantilon dans le cadre du modèle linéaire.

De plus, on vu que pour l'algorithme de gradient stochastique moyenné, on a

$$C_n := \frac{1}{\sigma^2} n \left(\overline{\theta}_n - \theta\right)^T \overline{H}_n \left(\overline{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_d,$$

et de la même façon, on a

$$K_n := \frac{1}{\sigma^2} n \left(\tilde{\theta}_n - \theta\right)^T \overline{H}_n \left(\tilde{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_d,$$

et on peut ainsi construire un test asymptotique pour tester $\theta = \theta_0$. En effet, Figure 4.9, on s'intéresse aux fonctions de répartitions de C_n et K_n estimées à l'aide de 5000 échantillons. On voit que la fonction de répartition de K_n s'approche de celle d'une Chi 2 à 10 degrés de liberté, tandis que celle de C_n en est très loin. En effet, l'algorithme de gradient n'étant pas arrivé à convergence, la moyennisation n'accelère pas du tout la convergence, voire la dessert.

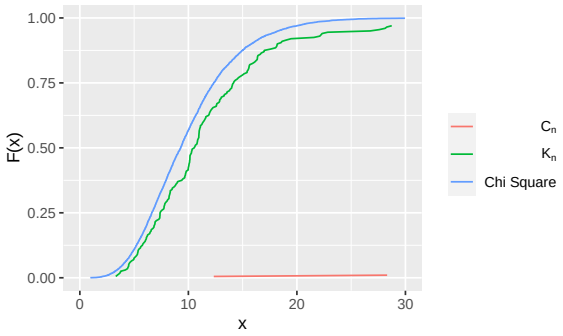

FIGURE 4.9 – Comparaison des fonctions de répartition de C_n et K_n, pour $n = 5000$), et de la fonction de répartition d'une Chi 2 à 10 degrés de liberté dans le cadre du modèle linéaire.

De plus, on peut remarquer que pour tout $x_0 \in \mathbb{R}^d$ on peut réécrire le TLC comme

$$\sqrt{n}\left(x_0^T\tilde{\theta}_n - x_0^T\theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2 x_0^T H^{-1} x_0\right)$$

ce que l'on peut réécrire comme

$$\sqrt{n}\frac{x_0^T\tilde{\theta}_n - x_0^T\theta}{\sqrt{\sigma^2 x_0^T H^{-1} x_0}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Ainsi, comme on dispose d'un estimateur en ligne de H^{-1}, il ne reste qu'à avoir un estimateur récursif de σ^2 pour obtenir des intervalles de confiance en ligne de $x_0^T\theta$. Une façon pour estimer σ^2 est de considérer l'erreur quadratique moyenne des prévisions,

$$\sigma_n^2 = \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - X_k^T\tilde{\theta}_{k-1}\right)^2$$

ce que l'on peut réécrire de manière récursive comme

$$\sigma_{n+1}^2 = \sigma_n^2 + \frac{1}{n+1}\left(\left(Y_{n+1} - X_{n+1}\tilde{\theta}_n\right)^2 - \sigma_n^2\right).$$

En effet, le théorème suivant nous confirme que cet estimateur converge rapidement vers σ^2.

Théorème 4.2.7. *Si il existe $\eta > 0$ tel que X et ϵ admettent des moments d'ordre $4+4\epsilon$ et $2+2\eta$ et que $\mathbb{E}\left[XX^T\right]$ est inversible, alors pour tout $\delta > 0$,*

$$\left|\sigma_n^2 - \sigma^2\right|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) \quad p.s.$$

On admettra ce théorème. Grâce au théorème de Slutsky,

$$C_{x_0} = \sqrt{n}\frac{x_0^T\tilde{\theta}_n - x_0\theta}{\sqrt{\sigma_n^2 x_0^T \overline{H}_n^{-1} x_0^T}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

En effet, Figure 4.10, on s'intéresse à la densité de C_{e_1}, i.e on prend $x_0 = e_1 = (1,0,\ldots,0)^T$ et on compare sa densité à celle d'une loi normale centrée réduite. La densité de C_{e_1} est estimée à l'aide de 1000 échantillons.

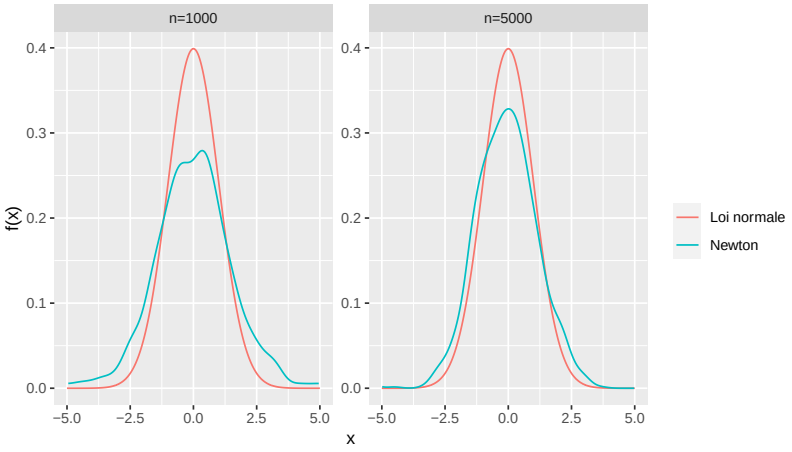

FIGURE 4.10 – Comparaison de la densité de C_{e_1}, pour $n = 1000$ (à gauche) et $n = 5000$ (à droite), et de la densité d'une loi normal centrée réduite dans le cadre de la régression linéaire.

On voit que l'estimation est plutôt bonne et que lorsque la taille d'échantillon augmente, la densité de C_{e_1} s'approche de celle de la loi normale centrée réduite, ce qui légitime l'usage de C_{x_0} pour l'obtention d'intervalles de confiance. En effet, on peut ainsi construire les intervalles de confiance asymptotiques en ligne

$$IC_{n,1-\alpha}(\theta) = \left[x_0^T\tilde{\theta}_n \pm \Phi^{-1}(1-\alpha/2)\frac{\sqrt{\sigma_n^2 x_0^T \overline{H}_n^{-1} x_0}}{\sqrt{n}}\right]$$

où $\Phi^{-1}(1-\alpha/2)$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite. On parle d'intervalles de confiance en ligne dans le sens où on peut mettre à jours ces intervalles avec un cout assez réduit en terme de temps de calculs. On peut également construire un rectangle de confiance de niveau asymptotique au moins $1-\alpha$. En effet, en considérant $B = \{e_1, \ldots, e_d\}$ la base canonique de $\mathbb{R}^d$, on a

$$R_{1-\alpha}(\theta) = \prod_{i=1}^{d} \left[e_i^T \tilde{\theta}_n \pm \Phi^{-1}\left(1 - \frac{\alpha}{2d}\right) \frac{\sqrt{\sigma_n^2 e_i^T \overline{H}_n e_i}}{\sqrt{n}} \right]$$

où $\Phi^{-1}(1-\alpha/2d)$ est le quantile d'ordre $1-\alpha/2d$ de la loi normale centrée réduite.

4.2.6 Application à la régression logistique

On se place maintenant dans le cadre de la régression logistique défini par (1.2). On rappelle également que la Hessienne de la fonction que l'on cherche à minimiser G est définie pour tout $h \in \mathbb{R}^d$ par

$$\nabla^2 G(h) = \mathbb{E}\left[\pi\left(h^T X\right)\left(1 - \pi\left(h^T X\right)\right) XX^T \right],$$

avec $\pi(x) = \frac{e^x}{1+e^x}$. Un estimateur naturel de la Hessienne serait donc

$$\overline{S}_n = \frac{1}{n+1}\left(S_0 + \sum_{k=1}^{n} \pi\left(\theta^T X_k\right)\left(1 - \pi\left(\theta^T X_k\right)\right) X_k X_k^T \right).$$

Cependant, on ne connait généralement pas θ, et il va donc falloir le remplacer par un estimateur en ligne de θ, conduisant à un premier algorithme de Newton stochastique

$$\begin{aligned}
\hat{\theta}_{n+1} &= \hat{\theta}_n + \frac{1}{n+1}\overline{S}_n^{-1}\left(Y_{n+1} - \pi\left(\hat{\theta}_n^T X_{n+1}\right)\right) X_{n+1} \\
\overline{S}_{n+1} &= \overline{S}_n + \frac{1}{n+2}\left(\pi\left(\hat{\theta}_n^T X_{n+1}\right)\left(1 - \pi\left(\hat{\theta}_n^T X_{n+1}\right)\right) X_{n+1} X_{n+1}^T - \overline{S}_n \right)
\end{aligned}$$

i.e cela reviendrait à considérer $\overline{S}_n = \frac{1}{n+1}\left(\overline{S}_0 + \sum_{k=1}^{n} \pi\left(X_k^T \hat{\theta}_{k-1}\right)\left(1 - \pi\left(X_k^T \hat{\theta}_{k-1}\right)\right) X_k X_k^T\right)$. Se pose alors deux questions. La première est de savoir comment mettre à jour $\overline{S}_{n+1}^{-1}$. Cela peut se faire en utilisant encore une fois la formule de Riccati. De plus, est-ce que notre estimateur de la Hessienne vérifie **(H1)** ? On est malheureusement incapable de suffisamment contrôler la plus petite valeur propre de $\overline{S}_n$ et on est donc obligé de proposer une version tronquée de l'algorithme de Newton stochastique, i.e on va considérer [BGBP19]

$$\begin{aligned}
\alpha_{n+1} &= \pi\left(\tilde{\theta}_n^T X_{n+1}\right)\left(1 - \pi\left(\tilde{\theta}_n^T X_{n+1}\right)\right) \\
\tilde{\theta}_{n+1} &= \tilde{\theta}_n + \frac{1}{n+1}\overline{H}_n^{-1}\left(Y_{n+1} - \pi\left(\tilde{\theta}_n^T X_{n+1}\right)\right) X_{n+1} \\
H_{n+1}^{-1} &= H_n^{-1} - a_{n+1}\left(1 + a_{n+1} X_{n+1}^T H_n^{-1} X_{n+1}\right)^{-1} H_n^{-1} X_{n+1} X_{n+1}^T H_n^{-1}
\end{aligned}$$

avec H_0^{-1} symétrique et définie positive, $\tilde{\theta}_0$ borné, $\overline{H}_n^{-1} = (n+1)H_n$, et $a_{n+1} = \max\left\{\alpha_{n+1}, \frac{c_\beta}{(n+1)^\beta}\right\}$ avec $c_\beta > 0$ et $\beta \in (0, 1/2)$. A noter que grâce à la formule de Riccati, on peut montrer que

$$H_n = H_0 + \sum_{k=1}^{n} a_k X_k X_k^T.$$

Le terme de troncature a_n permet de contrôler le comportement asymptotique de la plus petite valeurs propre de H_n. En effet, si X admet un moment d'ordre 2, on a

$$\frac{1}{\sum_{k=1}^{n} \frac{c_\beta}{k^\beta}} \sum_{k=1}^{n} \frac{c_\beta}{k^\beta} X_k X_k^T \xrightarrow[n\to+\infty]{p.s} \mathbb{E}\left[XX^T\right]$$

et en d'autres termes, si $\mathbb{E}\left[XX^T\right]$ est inversible, on a

$$\lambda_{\max}\left(\left(H_0 + \sum_{k=1}^{n} \frac{c_\beta}{k^\beta} X_k X_k^T\right)^{-1}\right) = O\left(n^{\beta-1}\right) \quad p.s,$$

ce qui nous permet d'obtenir la convergence des estimateurs [BGBP19, BGB20].

Théorème 4.2.8. *On suppose que X admet un moment d'ordre 2 et que $H := \nabla^2 G(\theta)$ est inversible. Alors $\tilde{\theta}_n$ converge presque sûrement vers θ. De plus, si X admet un moment d'ordre 4, alors*

$$\left\|\tilde{\theta}_n - \theta\right\|^2 = O\left(\frac{\ln n}{n}\right) \quad p.s. \qquad et \qquad \sqrt{n}\left(\tilde{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\right).$$

On admettra ce théorème mais la preuve est disponible dans la version longue. Figure 4.11, on considère le modèle

$$\theta = (1, \ldots, 1)^T \in \mathbb{R}^5 \qquad \text{et} \qquad X \sim \mathcal{N}\left(0, \text{diag}\left(\sigma_i^2\right)\right)$$

avec pour tout $i = 1, \ldots, d$, $\sigma_i^2 = \frac{i^2}{d^2}$. On s'attend donc à ce que la plus grande valeur propre de la Hessienne soit à peu près 25 fois plus grande que la plus petite. On voit bien Figure 4.11 que les estimateurs de gradient peinent à arriver à convergence, et donc, que les estimateurs moyennés ne permettent pas d'accélérer la convergence. A contrario, on voit bien que malgré ces différences d'échelles entre les valeurs propres de la Hessienne, l'algorithme de Newton stochastique converge très rapidement.

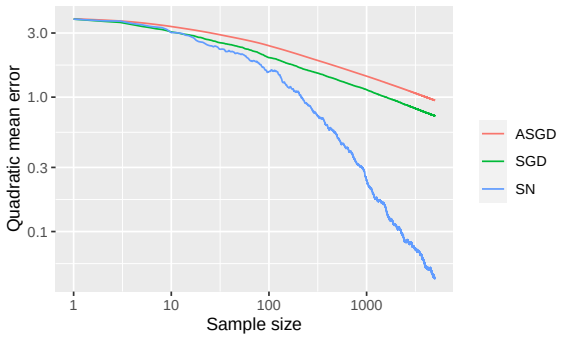

FIGURE 4.11 – Evolution de l'erreur quadratique moyenne des estimateurs de gradient θ_n (SGD), de leur version moyennée $\overline{\theta}_n$ (ASGD) et des estimateurs de Newton stochastique (SN) en fonction de la taille de l'échantillon dans le cadre de la régression logistique.

A noter que l'on a vu que l'on a, grâce au théorème de continuité,

$$n\left(\tilde{\theta}_n - \theta\right)^T H \left(\tilde{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_d.$$

Ainsi, comme $\overline{H}_n$ converge presque sûrement vers H, en appliquant le théorème de Slutsky, on obtient

$$K_n := n\left(\tilde{\theta}_n - \theta\right)^T \overline{H}_n \left(\tilde{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_d.$$

En effet, Figure 4.12, on s'intéresse à la fonction de répartition de K_n estimée à l'aide de 1000 échantillons. On voit que même pour une relativement petite taille d'échantillon ($n = 5000$), la fonction de répartition de K_n s'approche de celle d'une Chi 2 à 10 degrés de liberté.

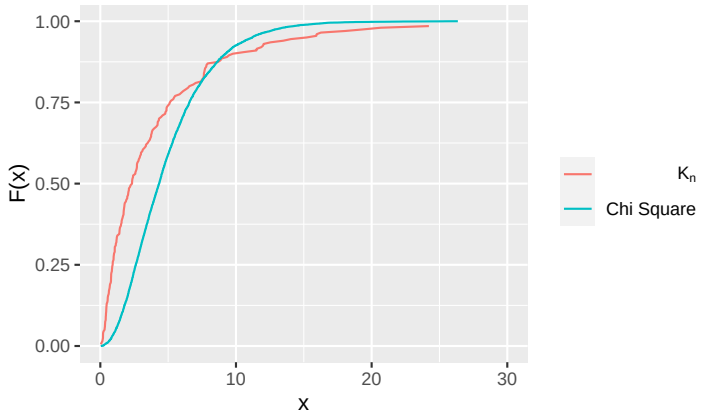

FIGURE 4.12 – Comparaison de la fonction de répartition de K_n, pour $n = 5000$, et de la fonction de répartition d'une Chi 2 à 10 degrés de liberté.

A noter également que l'on peut réécrire la normalité asymptotique comme

$$\sqrt{n}\frac{x_0^T\tilde{\theta}_n - x_0^T\theta}{\sqrt{x_0^T H^{-1} x_0}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

et grâce au théorème de Slutsky, on obtient

$$C_{x_0}\sqrt{n}\frac{x_0^T\tilde{\theta}_n - x_0^T\theta}{\sqrt{x_0^T \overline{H}_n^{-1} x_0}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

En effet, Figure 4.10, on s'intéresse à la densité de C_{e_1}, i.e on prend $x_0 = e_1 = (1,0,\dots,0)^T$ et on compare sa densité à celle d'une loi normale centrée réduite.

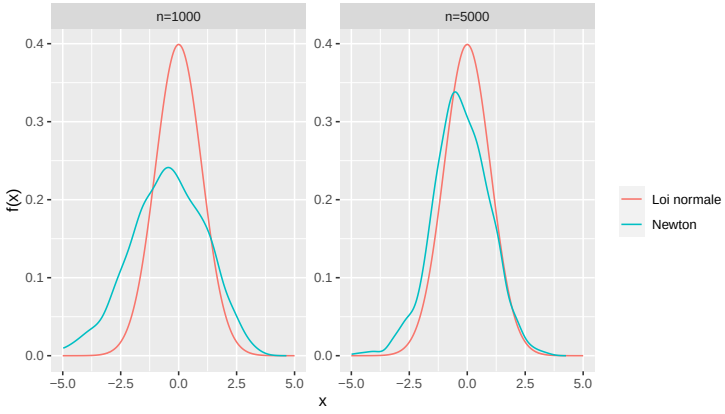

FIGURE 4.13 – Comparaison de la densité de C_{e_1}, pour $n = 1000$ (à gauche) et $n = 5000$ (à droite), et de la densité d'une loi normal centrée réduite.

On voit que l'estimation est plutôt bonne et ce, même pour une taille d'échantillon raisonnable ($n = 1000$). Ceci légitime donc l'usage de C_{x_0} pour l'obtention d'intervalles de confiance, i.e on considère les intervalles de confiance asymptotiques en ligne suivant

$$\mathbb{P}\left[x_0^T\theta \in \left[x_0^T\tilde{\theta}_n \pm \Phi^{-1}(1-\alpha/2)\frac{\sqrt{x_0^T\overline{H}_n^{-1}x_0}}{\sqrt{n}}\right]\right] \xrightarrow[n\to+\infty]{} 1-\alpha$$

où $\Phi^{-1}(1-\alpha/2)$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite. On peut également construire un rectangle de confiance de niveau asymptotique au moins $1-\alpha$. En effet, en considérant $B = \{e_1,\dots,e_d\}$ la base canonique de $\mathbb{R}^d$, on a

$$R_{1-\alpha}(\theta) = \prod_{i=1}^{d}\left[e_i^T\tilde{\theta}_n \pm \Phi^{-1}\left(1-\frac{\alpha}{2d}\right)\frac{\sqrt{e_i^T\overline{H}_n e_i}}{\sqrt{n}}\right]$$

où $\Phi^{-1}(1-\alpha/2d)$ est le quantile d'ordre $1-\alpha/2d$ de la loi normale centrée réduite.

Bibliographie

[BC07] Bernard Bercu and Djalil Chafaï. *Modélisation stochastique et simulation-Cours et applications*. 2007.

[BGB20] Claire Boyer and Antoine Godichon-Baggioni. On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *arXiv preprint arXiv :2011.09706*, 2020.

[BGBP19] Bernard Bercu, Antoine Godichon-Baggioni, and Bruno Portier. An efficient stochastic newton algorithm for parameter estimation in logistic regressions. *arXiv preprint arXiv :1904.07908*, 2019.

[BM13] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O (1/n). In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

[CGBP20] Peggy Cénac, Antoine Godichon-Baggioni, and Bruno Portier. An efficient averaged stochastic gauss-newtwon algorithm for estimating parameters of non linear regressions models. *arXiv preprint arXiv :2006.12920*, 2020.

[DST90] M Duflo, R Senoussi, and A Touati. Sur la loi des grands nombres pour les martingales vectorielles et l'estimateur des moindres carrés d'un modèle de régression. In *Annales de l'IHP Probabilités et statistiques*, volume 26, pages 549–566, 1990.

[Duf90] Marie Duflo. *Méthodes récursives aléatoires*. Masson, 1990.

[Duf97] Marie Duflo. *Random iterative models*, volume 34 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997. Translated from the 1990 French original by Stephen S. Wilson and revised by the author.

[MP11] Abdelkader Mokkadem and Mariane Pelletier. A generalization of the averaging procedure : The use of two-time-scale algorithms. *SIAM Journal on Control and Optimization*, 49(4) :1523–1543, 2011.

[Pel98] Mariane Pelletier. On the almost sure asymptotic behaviour of stochastic algorithms. *Stochastic processes and their applications*, 78(2) :217–244, 1998.

[PJ92] Boris Polyak and Anatoli Juditsky. Acceleration of stochastic approximation. *SIAM J. Control and Optimization*, 30 :838–855, 1992.

[RM51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[Rup88] David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

[Wei37] Endre Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Math. J*, 43(355-386) :2, 1937.