Introduction à l'optimisation stochastique

Antoine Godichon-Baggioni

Table des matières

Chapitre 1

Algorithmes de gradient stochastiques

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

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

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

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

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

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

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

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

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

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

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 $\hat{m}_n = \overline{X}_n$.

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 $\mathbf{X} = (X_1, \dots X_n)^T$ 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 $\mathbf{Y} = (Y_1, \dots, Y_n)^T$.

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.

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.

1.2.2 Un résultat de convergence

Le théorème suivant donne la normalité asymptotique des M-estimateurs.

Théorème 1.2.1. *Sous certaines hypothèses,*

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

avec

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

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 $X_1, \dots, X_n, X_{n+1}, \dots$ 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 $X_1, \dots, X_n, X_{n+1}, \dots$ 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

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

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 **[?]**, est défini de manière récursive pour tout $n \geq 0$ par

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

avec m_0 borné et (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

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

avec $\xi_{n+1} = \nabla G(m_n) - \nabla_h g(X_{n+1}, m_n)$. De plus, en considérant la filtration $(\mathcal{F}_n)$ engendrée par l'échantillon, i.e $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$, (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,

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

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 $(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, Y_{n+1}), \dots$ 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 $(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, Y_{n+1}), \dots$ 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

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

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 il existe une constante C' telle que pour tout $n \geq 1$,

$$\mathbb{E}\left[\|m_n - m\|^2 \right] \leq C' \gamma_n.$$

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

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

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 à l'aide du lemme suivant :

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

$$\delta_{n+1} \leq \left(1-a\gamma_{n+1}+b\gamma_{n+1}^2\right)\delta_n + c\gamma_{n+1}^2$$

avec $\gamma_n = c_\gamma n^{-\alpha}$, $c_\gamma, a, b, c > 0$, et $\alpha \in (1/2, 1)$. Alors il existe une constante C' telle que pour tout $n \geq 1$,

$$\delta_n \leq C'\gamma_n$$

□

1.4.3 Exemple : regression lineaire

On se place dans le cadre du modèle linéaire défini par (1.1). 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

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

De plus, on choisit $c_\gamma = 1$ et $\alpha = 0.5, 0.66, 0.75$ 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 $-\alpha$.

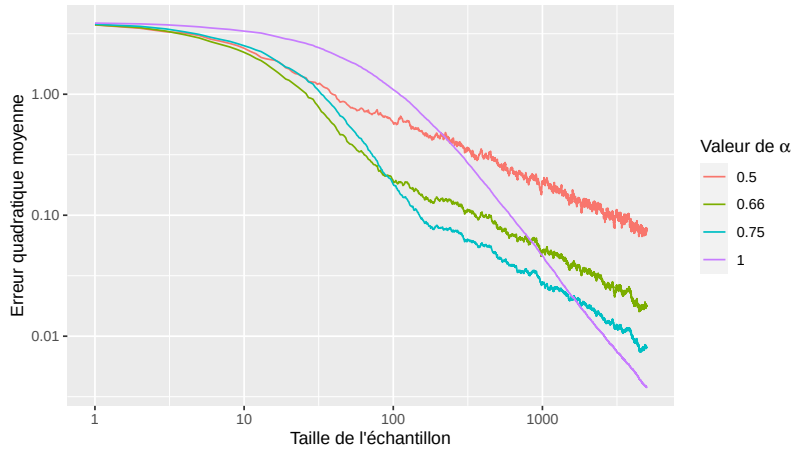

FIGURE 1.1 – Evolution de l'erreur quadratique moyenne de θ_n en fonction de la taille d'échantillon n dans le cadre de la régression linéaire.

1.5 Théorème de Robbins-Siegmund et convergence presque sûre

1.5.1 Théorème de Robbins-Siegmund

Les résultats de convergence presque sûre pour les algorithmes de gradient stochastiques reposent essentiellement sur le Théorème de Robbins-Siegmund suivant :

Théorème 1.5.1 (Robbins-Siegmund). *Soit $(V_n), (A_n), (B_n), (C_n)$ trois suites de variables aléatoires positives adaptées à une filtration $(\mathcal{F}_n)$. 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) et (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 finie et

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

On admettra ce théorème.

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 [**?**], 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 $\sum_{n\geq 0} \gamma_{n+1}^2 < +\infty$ p.s, 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 $\liminf_n (m_n - m)(F_X(m_n) - F(m)) = 0$ p.s. Si on suppose que la fonction F est strictement croissante sur un voisinage de F, on obtient

$$\liminf_n |m_n - m| = 0 \quad p.s,$$

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 1.2, on voit que dans les deux cas les estimateurs convergent assez rapidement vers le quantile.

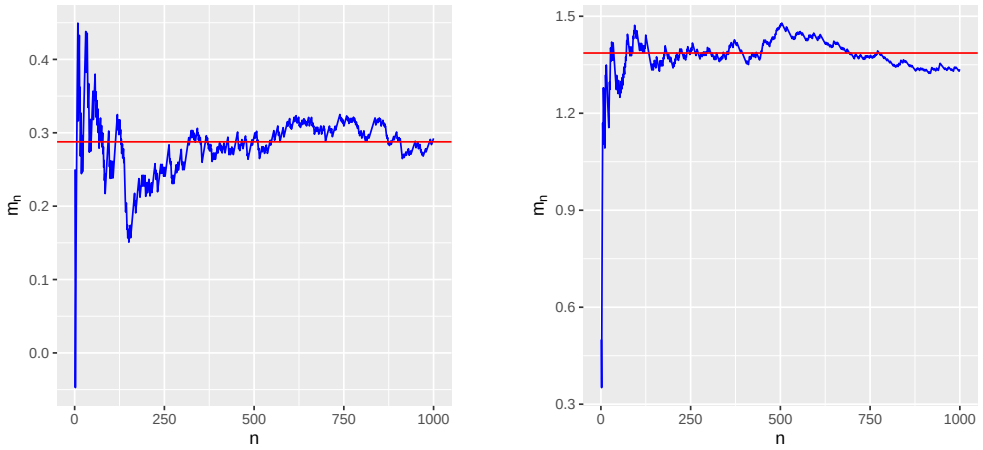

FIGURE 1.2 – Evolution de l'estimation des quantiles d'ordre 0.25 (à gauche) et 0.75 à droite pour la loi exponentielle de paramètre 1.

1.5.2 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

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

Une approche directe pour obtenir la convergence des estimateurs 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 1.5.2 (Approche directe). *On suppose que la fonction G est strictement convexe, i.e que pour tout $h \in \mathbb{R}^d$ tel que $h \neq m$*

$$\langle \nabla G\left(h\right), h - m \rangle > 0 \tag{1.7}$$

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

$$\mathbb{E}\left[\|\nabla_h g\left(X,h\right)\|^2\right] \leq C\left(1 + \|h - m\|^2\right), \tag{1.8}$$

alors

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

Démonstration. On a

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

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

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

Grâce à l'inégalité (1.8), on obtient

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

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

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

Comme $\sum_{n\geq 1}\gamma_n = +\infty$, on a $\liminf_n \left\langle \nabla G\left(m_n\right), m_n-m\right\rangle = 0$ 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. 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 1.5.3. *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

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

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

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

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

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

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

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

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

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

Comme $\sum_{n \geq 1} \gamma_{n+1} = +\infty$, on a $\liminf_n \| \nabla G(m_n) \| = 0$ presque sûrement, et comme m est l'unique zéro du gradient, $\liminf_n \| m_n - m \| = 0$ 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. □

1.5.3 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 $\lambda_{\min} = \lambda_{\min}\left(\nabla^2 G(m)\right)$. 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,

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

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

Chapitre 2

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.

2.1 Algorithmes de gradient stochastiques moyennés

Une méthode usuelle pour accélérer la converge, introduite par [**?**] et [**?**], 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 $\alpha \in (1/2, 1)$.

2.1.1 Vitesse de convergence presque sûre

Théorème 2.1.1. *Sous certaines hypothèses,*

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

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.

2.1.2 Application au modèle linéaire

Dans la Figure 2.1, 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

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

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.

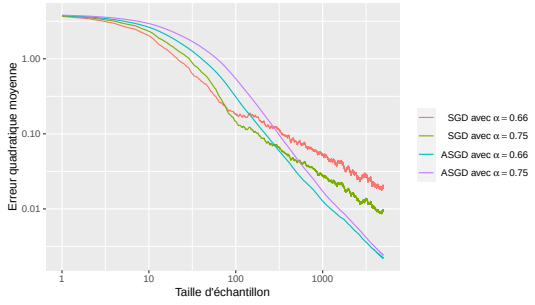

FIGURE 2.1 – Evolution de l'erreur quadratique moyenne de l'estimateur de gradient θ_n (SGD) et de sa version moyennée $\overline{\theta}_n$ (ASGD) en fonction de la taille d'échantillon n dans le cadre de la régression linéaire.

A noter que l'on peut réécrire la normalité asymptotique comme

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

Dans la Figure 2.2, 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.

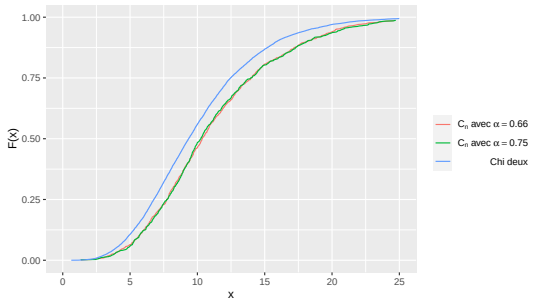

FIGURE 2.2 – Comparaison de la fonction de répartition de C_n avec $n = 5000$, pour $\alpha = 0.66$ et $\alpha = 0.75$, et de celle d'une Chi 2 à 10 degrés de liberté dans le cadre du modèle linéaire.

2.2 Algorithme de Newton stochastique

2.2.1 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 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 2.3. A noter que pour la Figure 2.3, 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.

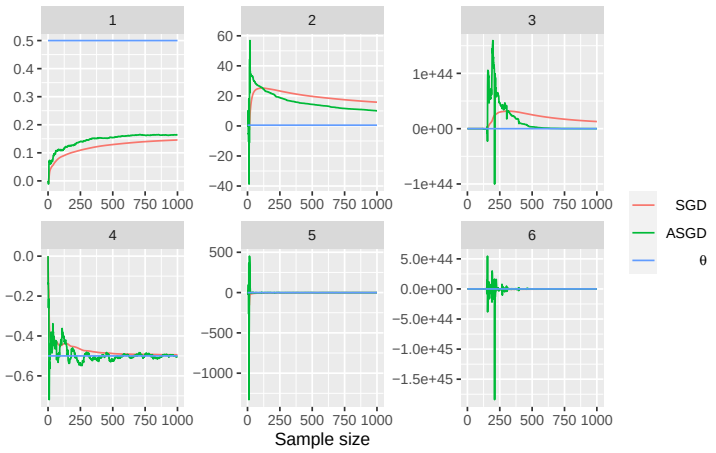

FIGURE 2.3 – Evolution des estimateurs de la première coordonnée (en haut) et de la deuxième (en bas) pour, de gauche à droite, $c_\gamma = 0.1, c_\gamma = 1$ et $c_\gamma = 10$.

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 2.4, on voit bien que les estimateurs des deux coordonnées arrivent rapidement à convergence.

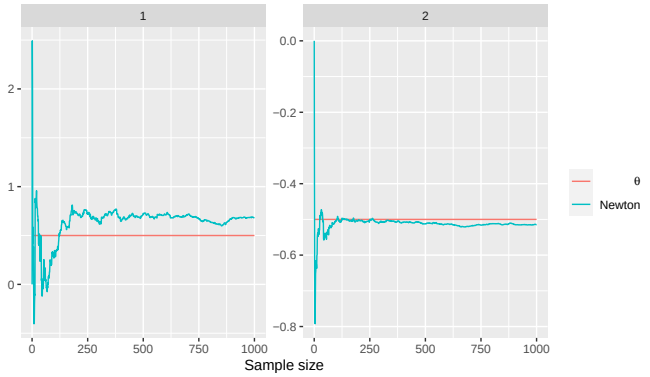

FIGURE 2.4 – Evolution des estimateurs de la première coordonnée (à gauche) et de la deuxième coordonnée (à droite) pour l'algorithme de Newton stochastique.

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.

2.2.2 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 [**?**]

$$\tilde{m}_{n+1} = \tilde{m}_n - \frac{1}{n+1}\overline{H}_n^{-1}\nabla_h g\left(X_{n+1}, \tilde{m}_n\right) \tag{2.1}$$

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 par la suite comment construire de tels estimateurs en ligne de l'inverse de la Hessienne pour la régression linéaire. 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 $h \in \mathbb{R}^d$,

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

On considère maintenant l'hypothèse suivante sur les valeurs propres de l'estimateur de la Hessienne :

(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 considère également 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 2.2.1. *On suppose que les hypothèse **(PS0")**, **(PS2)**, **(PS5)** et **(H1)** 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, il existe $h \in [m_n, m_{n+1}]$ tel que

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

De plus, comme $\left\|\nabla^2 G(h)\right\|_{op} \leq L_{\nabla G}$, on obtient grâce à l'inégalité de Cauchy Schwarz

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

En passant à l'espérance conditionnelle et comme $\overline{H}_n^{-1}$ est $\mathcal{F}_n$ mesurable,

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

avec pour tout $n \geq 0$, $V_n = G\left(\tilde{m}_n\right) - G(m)$. Alors, grâce à l'hypothèse **(PS0")**, il vient

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

Remarquons que grâce à l'hypothèse **(H1)**,

$$\sum_{n\geq 0}\frac{1}{(n+1)^2}\left\|\overline{H}_n^{-1}\right\|_{op}^2 < +\infty \quad p.s.$$

On a donc, en appliquant le théorème de Robbins-Siegmund, que V_n converge presque sûrement vers une variable aléatoire finie et

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

et comme, grâce à l'hypothèse **(H1)** on a

$$\sum_{n\geq 0}\frac{1}{n+1}\lambda_{\min}\left(\overline{H}_n^{-1}\right) = \sum_{n\geq 0}\frac{1}{n+1}\frac{1}{\lambda_{\max}\left(\overline{H}_n\right)} = +\infty \quad p.s,$$

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. □

2.2.3 Normalité asymptotique

Théorème 2.2.2. *Sous certaines hypothèses,*

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

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

On retrouve donc l'efficacité asymptotique, i.e. la même normalité asymptotique que les M-estimateurs.

2.2.4 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 $\nabla^2 G(h) = \mathbb{E}\left[XX^T\right]$, 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) qui permet que $\overline{H}_n$ soit définie positive pour tout n. 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 2.2.3 (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*

$$\left(A + uv^T\right)^{-1} = A^{-1} - \left(1 + v^T A^{-1} u\right)^{-1} A^{-1} uv^T A^{-1}.$$

Démonstration. On a

$$\begin{aligned}
&\left(A^{-1} - \left(1 + v^T A^{-1} u\right)^{-1} A^{-1} uv^T A^{-1}\right)\left(A + uv^T\right) \\
&\qquad = I + A^{-1}uv^T - \left(1 + v^T A^{-1} u\right)^{-1} A^{-1}\left(uv^T + uv^T A^{-1} uv^T\right) \\
&\qquad = I + A^{-1}uv^T - \left(1 + v^T A^{-1} u\right)^{-1} A^{-1}\left(u\left(1 + v^T A^{-1} u\right)v^T\right) \\
&\qquad = I + A^{-1}uv^T - A^{-1}uv^T \\
&\qquad = I.
\end{aligned}$$

□

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

$$\left(A + uu^T\right) = A^{-1} - \left(1 + u^T A^{-1} u\right)^{-1} A^{-1} uu^T A^{-1}.$$

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 :

$$H_{n+1}^{-1} = H_n^{-1} - \left(1 + X_{n+1}^T H_n^{-1} X_{n+1}\right)^{-1} H_n^{-1} X_{n+1} X_{n+1}^T H_n^{-1}$$

avec $\overline{H}_{n+1}^{-1} = (n+2) H_{n+1}^{-1}$. 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 :

$$\begin{aligned}
\tilde{\theta}_{n+1} &= \tilde{\theta}_n + \frac{1}{n+1} \overline{H}_n^{-1} \left(Y_{n+1} - \tilde{\theta}_n^T X_{n+1}\right) X_{n+1} \\
H_{n+1}^{-1} &= H_n^{-1} - \left(1 + X_{n+1}^T H_n^{-1} X_{n+1}\right)^{-1} H_n^{-1} X_{n+1} X_{n+1}^T H_n^{-1}
\end{aligned}$$

et $\overline{H}_n^{-1} = (n+1) H_n^{-1}$. A noter que l'on pourrait réécrire l'algorithme comme

$$\begin{aligned}
\tilde{\theta}_{n+1} &= \tilde{\theta}_n + H_n^{-1} \left(Y_{n+1} - \tilde{\theta}_n^T X_{n+1}\right) X_{n+1} \\
H_{n+1}^{-1} &= H_n^{-1} - \left(1 + X_{n+1}^T H_n^{-1} X_{n+1}\right)^{-1} H_n^{-1} X_{n+1} X_{n+1}^T H_n^{-1}.
\end{aligned}$$

Dans la Figure 2.5, on considère le modèle

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

avec pour tout $i = 1, \ldots, d$, $\sigma_i^2 = \frac{i^2}{d^2}$. 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 2.5 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.

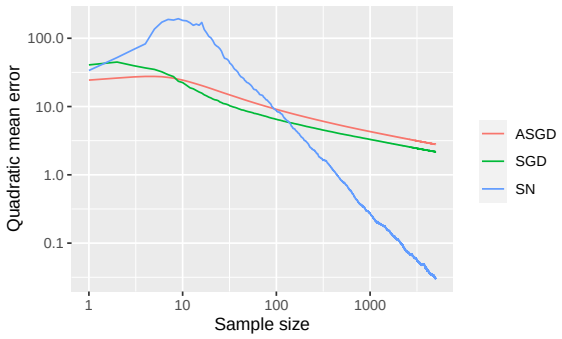

FIGURE 2.5 – Evolution de l'erreur quadratique moyenne des estimateurs de gradient θ_n (SGD), de leur version moyennée $\overline{\theta}_n$ (ASGD) et des estimateurs de Newton stochastique $\tilde{\theta}_n$ (SN) en fonction de la taille de l'échantilon dans le cadre du modèle linéaire.

De plus, on peut facilement vérifier 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 2.6, 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.

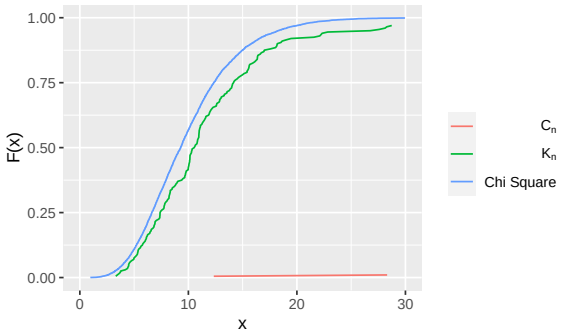

FIGURE 2.6 – Comparaison des fonctions de répartition de C_n et K_n, pour $n = 5000$, et de la fonction de répartition d'une Chi 2 à 10 degrés de liberté dans le cadre du modèle linéaire.

Bibliographie

[] 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.

[] Boris Polyak and Anatoli Juditsky. Acceleration of stochastic approximation. *SIAM J. Control and Optimization*, 30 :838–855, 1992.

[] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[] David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.