Algorithmes de Newton stochastiques avec $O(Nd)$ opérations

Antoine Godichon-Baggioni [1] & Nicklas Werge [2]

[1] *LPSM, Sorbonne Université, France, antoine.godichon_baggioni@upmc.fr*
[2] *Department of Mathematics and Computer Science, University of Southern Denmark, werge@sdu.dk*

Résumé. On s'intéresse ici au traitement de données arrivant par blocs (en streaming)à l'aide d'algorithmes stochastiques dits adaptatifs. Plus précisément, on s'intéressera à des méthodes de Newton stochastiques, qui sont très utiles pour traiter des problèmes mal conditionnés. De plus, on verra que l'on peut obtenir de telles méthodes avec un temps de calculs de l'ordre de $O(Nd)$ opérations, i.e du même ordre que les algorithmes de gradient stochastiques classiques.

Mots-clés. Optimisation stochastique, méthodes adaptatives, algorithme de Newton, apprentissage en ligne.

Abstract. We focus on the processing of data arriving in blocks (streaming) using adaptive stochastic algorithm methods. Specifically, the focus is on stochastic Newton methods, which are highly useful for handling ill-conditioned problems. Furthermore, it will be shown that such methods can be achieved with a computational time of order $O(Nd)$ operations, i.e., of the same order as usual stochastic gradient algorithms.

Keywords. Stochastic optimization, adaptive methods, Newton's method, online learning.

1 Introduction

Un problème usuel consiste à estimer le minimiseur d'une fonction convexe $F : \mathbb{R}^d \to \mathbb{R}$ de la forme

$$F(\theta) := \mathbb{E}_{\xi \sim \Xi}[f(\theta; \xi)], \tag{1}$$

où f est une fonction de perte, ξ est une variable aléatoire suivant une distribution inconnue Ξ. Ce problème est fréquemment rencontré dans de nombreuses applications en apprentissage automatique [Kushner and Yin, 2003, Bottou et al., 2018].

Nous nous concentrons ici sur l'acquisition de données volumineuses arrivant en streaming. Plus précisément, les données arrivent sous forme de blocs [Godichon-Baggioni et al., 2023b, Godichon-Baggioni et al., 2023a], formant des sous-échantillons indépendants. Formellement, nous considérons une suite de copies i.i.d. : $\{\xi_{1,1}, \ldots, \xi_{1,n_1}\}, \ldots, \{\xi_{t,1}, \ldots, \xi_{t,n_t}\}, \ldots$, où $\{\xi_{t,1}, \ldots, \xi_{t,n_t}\}$ représente un bloc de n_t données arrivant au temps t.

Nous nous intéressons ici à des méthodes de gradient stochastiques adaptatives, c'est-à-dire que nous incorporons une matrice aléatoire A_t dans le pas. En particulier, nous étudierons le cas où A_t est un estimateur de l'inverse de la Hessienne de F, correspondant à un algorithme de Newton en streaming. Ces méthodes sont particulièrement utiles lorsqu'il faut traiter des fonctions mal conditionnées, c'est-à-dire lorsque les valeurs propres de la Hessienne de la fonction à minimiser sont à des échelles très différentes [Bercu et al., 2020, Boyer and Godichon-Baggioni, 2023]. Ces méthodes adaptatives peuvent s'écrire de manière récursive comme suit :

$$\theta_{t+1} = \theta_t - \gamma_{t+1} A_t \nabla_\theta f(\theta_t; \xi_{t+1}), \quad \theta_0 \in \mathbb{R}^d, \tag{2}$$

où $\nabla_\theta f(\theta_t; \xi_{t+1}) = n_{t+1}^{-1} \sum_{i=1}^{n_{t+1}} \nabla_\theta f(\theta_t, \xi_{t+1,i})$ et (γ_t) est une suite de pas positifs.

Nous introduirons également une version moyennée pondérée de ces algorithmes [Polyak and Juditsky, 19 Mokkadem and Pelletier, 2011, Boyer and Godichon-Baggioni, 2023]. Cela permet, via la moyennisation, d'obtenir des estimateurs asymptotiquement efficaces, tandis que les pondérations permettent de donner plus d'importance aux derniers estimateurs de gradient et ainsi réduire les éventuels problèmes d'initialisation. Enfin, nous montrons que l'approche en streaming permet d'obtenir (dans certains cas tels que les régressions linéaires, logistiques et softmax, ainsi que l'estimation de la médiane) des estimateurs de Newton stochastiques ne nécessitant que $O(Nd)$ opérations (où N est la taille totale de l'échantillon), c'est-à-dire aussi peu coûteux que les algorithmes de premier ordre tels que les algorithmes de gradient stochastiques ou Adagrad. Ces estimateurs sont donc "optimaux" en termes de temps de calculs, tout en restant asymptotiquement efficaces.

2 Cadre

Dans cette section, on introduit les hypothèses nécessaires pour l'obtention de nos résultats théoriques. Ces hypothèses sont usuelles en optimisation stochastique, et en particulier pour les méthodes adaptatives [Leluc and Portier, 2023, Boyer and Godichon-Baggioni, 2023, Kushner and Yin, Duflo, 2013, Godichon-Baggioni and Tarrago, 2023].

Hypothèse 1 *Pour presque tout ξ, la fonction $f(\cdot; \xi)$ est différentiable et il existe des constantes positives C et C' telles que pour tout $\theta \in \mathbb{R}^d$*

$$\mathbb{E}[\|\nabla_\theta f(\theta; \xi)\|^2] \leq C + C'(F(\theta) - F(\theta^*)). \tag{3}$$

De plus, il existe $\theta^* \in \mathbb{R}^d$ tel que $\nabla_\theta F(\theta^*) = 0$, et la fonction $\Sigma : \theta \to \mathbb{E}[\nabla_\theta f(\theta; \xi) \nabla_\theta f(\theta; \xi)^\top]$ est continue en θ^*.

A noter que dans l'Hypothèse 1, $\mathbb{E}[\|\nabla_\theta f(\theta; \xi)\|^2]$ n'est pas majoré par une constante auquel on ajoute l'erreur quadratique $\|\theta - \theta^*\|^2$. Au lieu de cela, nous utilisons l'erreur fonctionnelle $F(\theta) - F(\theta^*)$ [Gower et al., 2019, Gazagnadou et al., 2019]. Cependant, si la fonctionnelle F est μ fortement convexe et $L_{\nabla F}$-lisse, on a $\frac{2}{L_{\nabla F}}(F(\theta) - F(\theta^*)) \leq \|\theta - \theta^*\|^2 \leq \frac{2}{\mu}(F(\theta) - F(\theta^*))$ pour tout $\theta \in \mathbb{R}^d$.

Afin d'assurer la forte consistance des estimateurs, nous introduisons une deuxième hypothèse. Cette dernière permet l'utilisation d'un développement de Taylor à l'ordre 2 de la fonctionnelle F.

Hypothèse 2 *La fonctionnelle F est deux fois continûment différentiable avec une hessienne uniformément bornée, i.e il existe $L_{\nabla F}$ tel que $\|\nabla^2_\theta F(\theta)\|_{\rm op} \leq L_{\nabla F}$.*

A noter que cela implique, entre autre, que le gradient de F est $L_{\nabla F}$-Lipschitz. La troisième hypothèse permet d'assurer l'unicité du minimiseur θ^* de la fonctionnelle F.

Hypothèse 3 *La fonctionnelle F est localement fortement convexe: $\lambda_{\min} := \lambda_{\min}(\nabla^2_\theta F(\theta^*)) > 0$.*

3 Méthodes adaptatives en streaming

Pour simplifier les résultats et notations, on considère maintenant que la taille des sous échantillons est constante, i.e $n_t = n$ pour tout $t \geq 0$. Néanmoins, les résultats pour des tailles de sous-échantillons croissantes sont disponibles dans [Godichon-Baggioni and Werge, 2023]. Dans tout ce qui suit, on note N_t le nombre total de données traitées au temps t, i.e $N_t = nt$. On suppose également que A_t est symétrique et définie positive pour tout $t \geq 0$. De plus, on suppose à partir de maintenant que la suite de pas (γ_t) et la suite de matrices aléatoires (A_t) vérifient les conditions suivantes:

$$\sum_{t\geq 1} \gamma_t \lambda_{\min}(A_{t-1}) = +\infty \text{ p.s.}, \quad \text{et} \quad \sum_{t\geq 1} \gamma_t^2 \lambda_{\max}(A_{t-1})^2 < +\infty \text{ p.s.} \tag{4}$$

Enfin, il existe une filtration $\mathcal{F}_t$ telle que A_t soit $\mathcal{F}_t$ mesurable et ξ_{t+1} soit indépendant de $\mathcal{F}_t$. Ces hypothèses sont assez usuelles et sont même vitales pour prouver la forte consistance des estimateurs à l'aide du théorème de Robbins-Siegmund [Boyer and Godichon-Baggioni, 2023]. Dans ce qui suit, on prendra $\gamma_t = C_\gamma t^{-\gamma}$ avec $C_\gamma > 0$ et $\gamma \in (1/2, 1)$.

Le théorème suivant établit la forte consistance de nos estimateurs de gradient stochastiques adaptatifs (θ_t) définis par (2).

Théorème 1 *Supposons que les Hypothèses 1 à 3 sont vérifiées, ainsi que les conditions* (4)*. Alors, θ_t converge presque sûrement vers θ^*.*

L'hypothèse suivante permet d'obtenir les vitesses de convergence des estimateurs (θ_t).

Hypothèse 4 *La matrice aléatoire A_t converge presque sûrement vers une matrice définie positive A.*

Par exemple, dans les méthodes de Newton, la matrice A correspond à l'inverse de la Hessienne, et dans le cas d'Adagrad, elle correspond à l'inverse de la racine carrée de la diagonale de la variance du gradient. A noter également que la forte consistance de θ_t (cf Théorème 1) implique souvent la forte consistance de A_t.

Théorème 2 *On suppose que les Hypothèses 1 à 4 sont vérifiées, ainsi que les conditions* (4). *De plus, on suppose qu'il existe des constantes positives* C_η *et* $\eta > \frac{1}{\gamma} - 1$ *telles que pour tout* $\theta \in \mathbb{R}^d$,

$$\mathbb{E}\left[\|\nabla_\theta f(\theta;\xi)\|^{2+2\eta}\right] \leq C_\eta \left(1 + F(\theta) - F(\theta^*)\right)^{1+\eta}. \tag{5}$$

Alors,

$$\|\theta_t - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t^\gamma}\right) \ p.s.$$

4 La version moyennée pondérée

La version moyennée pondérée des méthodes de gradient stochastiques adaptatives pour les données en streaming est définie comme suit:

$$\theta_{t,w} = \frac{1}{\sum_{i=0}^{t-1} \ln(i+1)^w} \sum_{i=0}^{t-1} \ln(i+1)^w \theta_i, \tag{6}$$

ce qui peut être écrit récursivement comme

$$\theta_{t+1,w} = \theta_{t,w} + \frac{\ln(t+1)^w}{\sum_{i=0}^{t} \ln(i+1)^w}(\theta_t - \theta_{t,w}).$$

Cette moyenne pondérée dans (6) améliore le comportement des estimateurs en attribuant plus de poids aux dernières estimations de (θ_t). La pondération logarithmique met donc l'accent sur les estimations récentes, présumées plus précises, tout en assurant l'efficacité des estimateurs [Mokkadem and Pelletier, 2011, Boyer and Godichon-Baggioni, 2023]. Pour établir la vitesse de convergence des estimateurs moyennés, on introduit une nouvelle hypothèse.

Hypothèse 5 *Il existe des constantes positives* L_r *et* r *telles que pour tout* $\theta \in \mathcal{B}(\theta^*, r)$

$$\|\nabla_\theta F(\theta) - \nabla^2_\theta F(\theta^*)(\theta - \theta^*)\| \leq L_r \|\theta - \theta^*\|^2.$$

Cette hypothèse est satisfaite dès que la Hessienne de F est localement Lipschitzienne sur un voisinage de θ^*.

Théorème 3 *On suppose que les Hypothèses 1 à 5 sont vérifiées ainsi que l'inégalité* (5). *De plus, on suppose qu'il existe* $v' > 1/2$ *tel que*

$$\frac{1}{\sum_{i=0}^{t-1} \ln(i+1)^w} \sum_{i=0}^{t-1} \ln(i+1)^{w+1/2+\delta} \|A_{i+1}^{-1} - A_i^{-1}\|_{\text{op}} (i+1)^{\frac{\gamma}{2}} = \mathcal{O}\left(\frac{1}{t^{v'}}\right) \ p.s, \tag{7}$$

pour n'importe quel $\delta > 0$. *Alors*

$$\|\theta_{t,w} - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right) \ p.s \quad et \quad \sqrt{N_t}(\theta_{t,w} - \theta^*) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}(0, \nabla^2_\theta F(\theta^*)^{-1} \Sigma \nabla^2_\theta F(\theta^*)^{-1}).$$

La variance $\nabla^2_\theta F(\theta^*)^{-1} \Sigma \nabla^2_\theta F(\theta^*)^{-1}$ correspond à l'inverse de l'Information de Fisher, et les estimateurs sont donc asymptotiquement efficaces.

5 Applications aux algorithmes de Newton

5.1 Algorithmes de Newton en streaming

Dans le cas particulier des méthodes de Newton stochastiques, on peut obtenir l'efficacité asymptotique sans moyennisation en choisissant une suite de pas de la forme $\gamma_t = \frac{1}{t}$. L'algorithme de Newton stochastique est alors défini de manière récursive pour tout $t \geq 0$ par

$$\theta_{t+1} = \theta_t - \frac{1}{t+1}\overline{H}_t^{-1}\nabla_\theta f(\theta_t; \xi_{t+1}), \quad \theta_0 \in \mathbb{R}^d, \tag{8}$$

où $\nabla_\theta f(\theta_t; \xi_{t+1}) = n^{-1}\sum_{i=1}^n \nabla_\theta f(\theta_t; \xi_{t+1,i})$ et $\overline{H}_t$ est un estimateur de la hessienne de F. De plus, on supposera que $\overline{H}_t$ est de la forme $\overline{H}_t = N_t^{-1}H_t$ avec

$$H_t = H_0 + \sum_{i=1}^{t}\sum_{j=1}^{n} \alpha_{i,j}\Phi_{i,j}\Phi_{i,j}^\top,$$

avec H_0 symétrique et positive, $\alpha_{i,j} = \alpha(\theta_{i-1}; \xi_{i,j})$, et $\Phi_{i,j} = \Phi(\theta_{i-1}; \xi_{i,j})$. On peut alors mettre à jour H_t^{-1} de manière récursive et avec un temps de calculs réduit à l'aide de la formule de Riccati/Sherman-Morrisson [Duflo, 2013, Sherman and Morrison, 1950] utilisée n fois, i.e pour tout $j = 1, \ldots, n$,

$$H_{t-1,j}^{-1} = H_{t-1,j-1}^{-1} - \alpha_{t,j}\left(1 + \alpha_{t,j}\Phi_{t,j}^\top H_{t-1,j-1}^{-1}\Phi_{t,j}\right)^{-1} H_{t-1,j-1}^{-1}\Phi_{t,j}\Phi_{t,j}^\top H_{t-1,j-1}^{-1}.$$

avec la convention $H_{t-1,0}^{-1} = H_{t-1}$. La construction explicite des estimations récursives de l'inverse de la hessienne est détaillée dans diverses applications, notamment les régressions linéaires, logistiques, softmax et ridge [Bercu et al., 2020, Boyer and Godichon-Baggioni, 2023, Godichon-Baggioni et al., 2024].

Théorème 4 *On suppose que les Hypothèses 1, 2, 3 et 5 sont vérifiées, ainsi que l'inégalité* (5). *Alors, θ_t converge presque sûrement vers θ^*. De plus, supposons que $\overline{H}_t^{-1}$ converge presque sûrement vers $\nabla_\theta^2 F(\theta^*)^{-1}$, alors*

$$\|\theta_t - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right) \ p.s.$$

Suppose de plus qu'il existe une constante positive ν telle que $\|\overline{H}_t^{-1} - \nabla_\theta^2 F(\theta^*)^{-1}\|_{op} = \mathcal{O}(\frac{1}{t^\nu})$ p.s. Alors

$$\sqrt{N_t}(\theta_t - \theta^*) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}(0, \nabla_\theta^2 F(\theta^*)^{-1}\Sigma\nabla_\theta^2 F(\theta^*)^{-1}).$$

5.2 Algorithmes de Newton stochastiques avec $\mathcal{O}(dN_t)$ opérations

La méthode de Newton stochastique nécessite $\mathcal{O}(d^2 N_t)$ opérations, ce qui peut être coûteux en terme de temps de calcul, surtout dans des contextes de (relativement) grande dimension.

Pour pallier ce problème, on remplace H_t dans (8) par $H_{t,w'}$ défini par

$$H_{t,w'} = H_{0,w'} + \sum_{i=1}^{t} \ln(i+1)^{w'} \sum_{j=1}^{n} Z_{i,j} \left(\iota_{i,j} \tilde{e}_{i,j} \tilde{e}_{i,j}^\top + \alpha_{i,j} \Phi_{i,j} \Phi_{i,j}^\top \right), \tag{9}$$

avec $N_{t,Z} = 1 + \sum_{i=1}^{t} \ln(i+1)^{w'} \sum_{j=1}^{n} Z_{i,j}$, $H_{0,w'}$ symétrique et positive, $w' \geq 0$, et $Z_{i,j}$ i.i.d avec $Z_{i,j} \sim \mathcal{B}(p)$ pour un certain $p \in (0,1]$. De plus, soit $N_{t,k,Z} = (1 + \sum_{i=1}^{t-1} \sum_{j=1}^{n} Z_{i,j} + \sum_{j=1}^{k} Z_{t,j})$, $\iota_{i,j} = c_\iota N_{i,j,Z}^{-\iota}$ pour $\iota \in (0, 1/2)$, et $e_{i,j}$ soit le composant ($N_{i,j,Z}$ modulo $d+1$)-ème de la base canonique. A noter

$$\tilde{H}_{t,j,w'}^{-1} = H_{t,j-1,w'}^{-1} - \frac{Z_{t+1,j} \iota_{t+1,j}}{1 + \iota_{t+1,j} e_{t+1,j} \tilde{H}_{t,j-1,w'}^{-1} e_{t+1,j}^T} H_{t,j-1,w'}^{-1} e_{t+1,j} e_{t+1,j}^T H_{t,j,w'}^{-1}$$

$$H_{t,j,w'}^{-1} = H_{t,j-1,w'}^{-1} - \frac{Z_{t+1,j} \ln(t+1)^{w'} \alpha_{t+1,j}}{1 + \ln(t+1)^{w'} \alpha_{t+1,j} \Phi_{t+1,j}^T H_{t,j-1,w'}^{-1} \Phi_{t+1,j}} H_{t,j-1,w'}^{-1} \Phi_{t+1,j} \Phi_{t+1,j}^T H_{t,j-1,w'}^{-1}$$

avec $\tilde{H}_{t,0,w'}^{-1} = H_{t-1,w'}$ et $H_{t,0,w'}^{-1} = \tilde{H}_{t,n,w'}^{-1}$. La mise à jour de H_{t+1}^{-1} ne coûte en moyenne que $\mathcal{O}(pd^2n)$ opérations, conduisant à un nombre total d'opérations d'ordre (en moyenne)

$$\underbrace{pd^2N_t}_{\text{estimation de l'inverse de la Hessienne}} + \underbrace{dN_t}_{\text{estimation du gradient}} + \underbrace{\frac{d^2N_t}{n}}_{\text{multiplication Hessienne*gradient}} .$$

Ainsi, on peut jouer avec la valeur de p pour réduire le coût de la mise à jour de l'inverse de la Hessienne. En effet, on peut obtenir un coût computationnel moyen au temps t de l'ordre de $\mathcal{O}(dN_t)$ opérations en prenant $p = d^{-1}$ et $n = d$. En d'autres termes, il est possible d'obtenir une méthode de Newton stochastique avec seulement $\mathcal{O}(dN_t)$ opérations.

Dans ce qui suit, on suppose que pour tout $\theta \in \mathbb{R}^d$,

$$\nabla_\theta^2 F(\theta) = \mathbb{E}\left[\alpha(\theta;\xi)\Phi(\theta;\xi)\Phi(\theta;\xi)^T\right]. \tag{10}$$

Théorème 5 *On suppose que les Hypothèses 1, 2, 3 et 5 sont vérifiées, ainsi que les inégalités* (5) *et* (10)*. De plus, supposons que pour tout θ, il existe des constantes positives $C_{\eta'}$ et $\eta' > 1$ telles que pour tout $\theta \in \mathbb{R}^d$,*

$$\mathbb{E}[\|\alpha(\theta;\xi)\Phi(\theta;\xi)\Phi(\theta;\xi)^\top\|^{\eta'}] \leq C_{\eta'}^{\eta'}.$$

Alors,

$$\|\theta_t - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right) \ p.s.$$

De plus, on suppose que la Hessienne de F est localement $L_{\nabla^2 F}$-Lipschitz sur un voisinage autour de θ^* et que $\eta' \geq 2$. Alors

$$\sqrt{N_t}(\theta_t - \theta^*) \xrightarrow[t \to +\infty]{\mathcal{L}} \mathcal{N}(0, \nabla_\theta^2 F(\theta^*)^{-1} \Sigma \nabla_\theta^2 F(\theta^*)^{-1}).$$

5.3 Version moyennée pondérée de la méthode de Newton stochastique avec $\mathcal{O}(dN_t)$ opérations

Bien que la méthode de Newton stochastique "directe" soit très performante, elle peut être assez sensible à une mauvaise initialisation [Boyer and Godichon-Baggioni, 2023]. On considère donc une version moyennée pondérée définie récursivement par

$$\theta_{t+1} = \theta_t - \gamma_{t+1}\overline{S}_{t,w'}^{-1}\nabla_\theta f(\theta_t;\xi_{t+1}),\ \theta_0 \in \mathbb{R}^d, \tag{11}$$

$$\theta_{t+1,w} = \theta_{t,w} + \frac{\ln(t+1)^w}{\sum_{i=0}^t \ln(i+1)^w}(\theta_t - \theta_{t,w}), \tag{12}$$

$\overline{S}_{t,w'} = N_{t,Z}^{-1}S_{t,w'}$ avec $N_{t,Z}\overline{S}_{t,w'} =: S_{t,w'} = S_{0,w'} + \sum_{i=1}^t \ln(i+1)^{w'}\sum_{j=1}^n Z_{i,j}\left(\iota_i e_{i,j}e_{i,j}^\top + \alpha_{i,j}\Phi_{i,j}\Phi_{i,j}^\top\right)$ et $S_{0w'}$ symétrique et positive. On peut suivre le même schéma récursif que pour $H_{t,w'}^{-1}$ pour mettre à jour l'inverse de $S_{t,w'}^{-1}$. En effet, la seule différence entre $S_{t,w'}^{-1}$ et $H_{t,w'}^{-1}$ est le choix de l'estimateur de θ^* choisi pour la méthode du plug-in.

Théorème 6 *On suppose que les hypothèses 1, 2, 3 et 5 sont vérifiées, ainsi que les inégalités* (5) *and* (10)*. De plus, supposons que pour tout θ, il existe des constantes positives $C_{\eta'}$ et $\eta' > 1$ telles que pour tout $\theta \in \mathbb{R}^d$,*

$$\mathbb{E}[\|\alpha(\theta;\xi)\Phi(\theta;\xi)\Phi(\theta;\xi)^\top\|^{\eta'}] \leq C_{\eta'}^{\eta'}.$$

Alors,

$$\|\theta_t - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t^\gamma}\right)\ p.s., \qquad \|\theta_{t,w} - \theta^*\|^2 = \mathcal{O}\left(\frac{\ln(N_t)}{N_t}\right)\ p.s.,$$

et

$$\sqrt{N_t}(\theta_{t,w} - \theta^*) \xrightarrow[t\to+\infty]{\mathcal{L}} \mathcal{N}(0, \nabla_\theta^2 F(\theta^*)^{-1}\Sigma\nabla_\theta^2 F(\theta^*)^{-1}).$$

6 Simulations

Dans cette section, on se concentre sur deux exemples classiques: la régression linéaire et la régression logistique. Pour le modèle linéaire, on a $\xi = (x, y) \in \mathbb{R}^d \times \mathbb{R}$, et on cherche à minimiser la fonction $F(\theta) = \frac{1}{2}\mathbb{E}[(y - x^\top\theta)^2]$. Dans le cas de la régression logistique, on a $\xi = (x, y) \in \mathbb{R}^d \times \{-1, 1\}$, et la fonction correspondante est $F(\theta) = \mathbb{E}[\ln(1+\exp(x^\top\theta)) - yx^\top\theta]$. Dans les deux cas, on va considérer une structure de covariance "complexe", i.e on va prendre

$$x \sim \mathcal{N}\left(0, M\,\mathrm{diag}\left(\frac{i^2}{d^2}\right)_{i=1,\dots,d} M^\top\right).$$

où M est une matrice orthogonale. Ce choix de distribution nous permet d'introduire des corrélations fortes entre les coordonnées de x. Dans ce qui suit, on fixe $d = 100$ et on note donc que la Hessienne a des valeurs propres à des échelles très différentes.

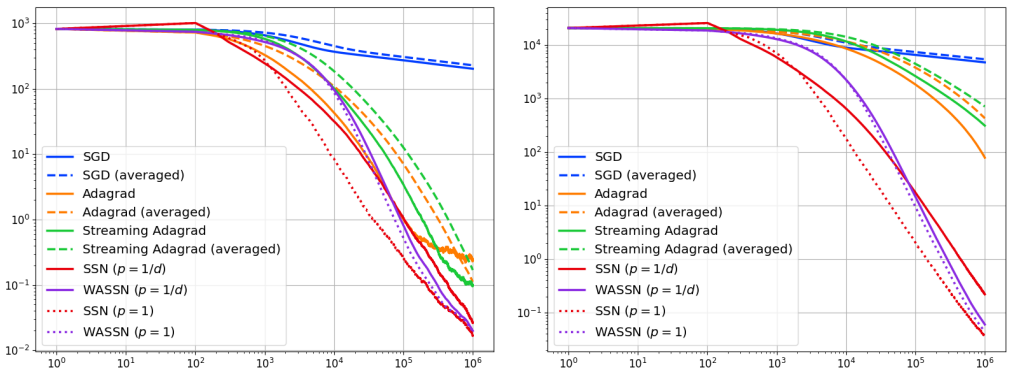

Figure 1: Modèle linéaire : Evolution de l'erreur quadratique moyenne des estimateurs en fonction de la taille d'échantillon. Les points initiaux θ_0 sont définis par $\theta_0 = \theta^*(1 + rU)$, où U suit une loi uniforme sur la sphère unitaire de $\mathbb{R}^d$, et $r = 1$ (à gauche) ou $r = 5$ (à droite).

6.1 Modèle linéaire

Dans la Figure 1, on considère le modèle linéaire avec deux types d'initialisations (plus ou moins précises). On peut voir que Adagrad et les algorithmes de Newton présentent des taux de convergence plus rapides par rapport au SGD standard (sans surprise). A noter que bien que l'algorithme Adagrad adapte ses pas, il peut être moins efficace lorsqu'il est confronté à des données fortement corrélées. C'est particulièrement criant lorsque les algorithmes sont mal initialisés, tandis les deux méthodes de Newton restent très performantes.

6.2 Régression logistique

Dans la Figure 2, on considère le problème de régression logistique avec là encore deux initialisations différentes. Pour toutes les configurations initiales, les méthodes de Newton stochastiques sont particulièrement efficace tandis qu'Adagrad semble moins adaptés.

Bibliographie

References

[Bercu et al., 2020] Bercu, B., Godichon, A., and Portier, B. (2020). An efficient stochastic newton algorithm for parameter estimation in logistic regressions. *SIAM Journal on Control and Optimization*, 58(1):348–367.

[Bottou et al., 2018] Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311.

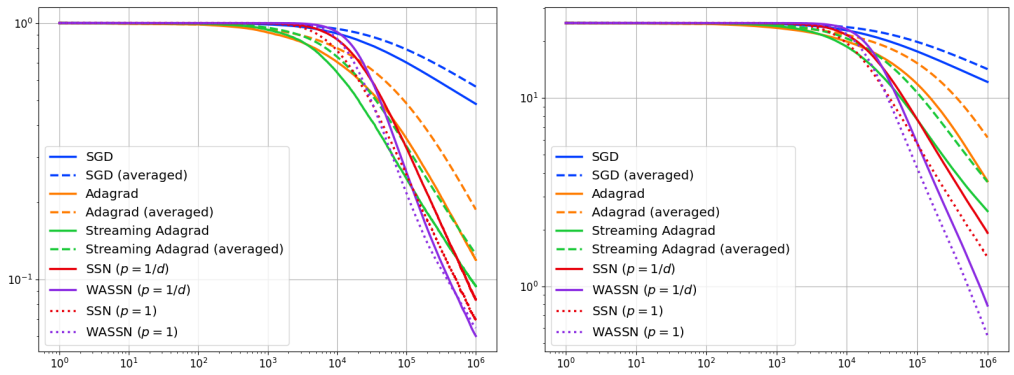

Figure 2: Régression logistique : Evolution de l'erreur quadratique moyenne des estimateurs en fonction de la taille d'échantillon. Les points initiaux θ_0 sont définis par $\theta_0 = \theta^*(1 + rU)$, où U suit une loi uniforme sur la sphère unitaire de $\mathbb{R}^d$, et $r = 1$ (à gauche) ou $r = 5$ (à droite).

[Boyer and Godichon-Baggioni, 2023] Boyer, C. and Godichon-Baggioni, A. (2023). On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *Computational Optimization and Applications*, 84(3):921–972.

[Duflo, 2013] Duflo, M. (2013). *Random iterative models*, volume 34. Springer Science & Business Media.

[Gazagnadou et al., 2019] Gazagnadou, N., Gower, R., and Salmon, J. (2019). Optimal mini-batch and step sizes for saga. In *International conference on machine learning*, pages 2142–2150. PMLR.

[Godichon-Baggioni et al., 2024] Godichon-Baggioni, A., Lu, W., and Portier, B. (2024). Recursive ridge regression using second-order stochastic algorithms. *Computational Statistics & Data Analysis*, 190:107854.

[Godichon-Baggioni and Tarrago, 2023] Godichon-Baggioni, A. and Tarrago, P. (2023). Non asymptotic analysis of adaptive stochastic gradient algorithms and applications. *arXiv preprint arXiv:2303.01370*.

[Godichon-Baggioni and Werge, 2023] Godichon-Baggioni, A. and Werge, N. (2023). On adaptive stochastic optimization for streaming data: A newton's method with o (dn) operations. *arXiv preprint arXiv:2311.17753*.

[Godichon-Baggioni et al., 2023a] Godichon-Baggioni, A., Werge, N., and Wintenberger, O. (2023a). Learning from time-dependent streaming data with online stochastic algorithms. *Transactions on Machine Learning Research*.

[Godichon-Baggioni et al., 2023b] Godichon-Baggioni, A., Werge, N., and Wintenberger, O. (2023b). Non-asymptotic analysis of stochastic approximation algorithms for streaming data. *ESAIM: Probability and Statistics*, 27:482–514.

[Gower et al., 2019] Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR.

[Gower et al., 2021] Gower, R. M., Richtárik, P., and Bach, F. (2021). Stochastic quasi-gradient methods: Variance reduction via jacobian sketching. *Mathematical Programming*, 188:135–192.

[Kushner and Yin, 2003] Kushner, H. and Yin, G. (2003). *Stochastic approximation and recursive algorithms*. Springer-Verlag NY.

[Leluc and Portier, 2023] Leluc, R. and Portier, F. (2023). Asymptotic analysis of conditioned stochastic gradient descent. *Transactions on Machine Learning Research*.

[Mokkadem and Pelletier, 2011] Mokkadem, A. and Pelletier, M. (2011). A generalization of the averaging procedure: The use of two-time-scale algorithms. *SIAM Journal on Control and Optimization*, 49(4):1523–1543.

[Polyak and Juditsky, 1992] Polyak, B. and Juditsky, A. (1992). Acceleration of stochastic approximation. *SIAM J. Control and Optimization*, 30:838–855.

[Sherman and Morrison, 1950] Sherman, J. and Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127.