Estimation en ligne de l'inverse de la Hessienne pour l'optimisation stochastique avec application aux algorithmes de Newton stochastiques universels

Antoine Godichon-Baggioni [1] , Wei Lu [2] , Bruno Portier [2]

[1] *Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, Paris, France.*
antoine.godichon_baggioni@upmc.fr
[2] *Laboratoire de Mathématiques de l'INSA Rouen Normandie (LMI), Normandie Université, Rouen, France.*
wei.lu@insa-rouen.fr ; bruno.portier@insa-rouen.fr

Résumé. On propose un algorithme stochastique du second-ordre (de type Newton) pour estimer le minimiseur d'une fonction convexe écrite comme une espérance. Nous introduisons une technique d'estimation récursive directe pour la matrice inverse de la Hessienne en utilisant une procédure de Robbins-Monro. Cette approche permet de réduire la complexité computationnelle. Surtout, elle permet de développer des méthodes de Newton stochastiques universelles tout en assurant l'efficacité asymptotique des estimateurs obtenus.

Mots-clés. Algorithme de Newton stochastique; Optimisation Stochastique; Algorithme de Robbins-Monro; Estimation en ligne

Abstract. This work addresses second-order stochastic optimization for estimating the minimizer of a convex function written as an expectation. A direct recursive estimation technique for the inverse Hessian matrix using a Robbins-Monro procedure is introduced. This approach enables to drastically reduce computational complexity. Above all, it allows to develop universal stochastic Newton methods and investigate the asymptotic efficiency of the proposed estimates.

Keywords. Stochastic Newton algorithm; Stochastic Optimization; Robbins-Monro algorithm; Online estimation

1 Introduction

Dans ce travail, nous considérons le problème d'optimisation stochastique, consistant à estimer le paramètre $\theta \in \mathbb{R}^d$ défini par

$$\theta = \arg\min_{h \in \mathbb{R}^d} G(h).$$

La fonction G est définie pour tout $h \in \mathbb{R}^d$ par : $G(h) = \mathbb{E}\left[g(X, h)\right]$ où X est un vecteur aléatoire de $\mathbb{R}^p$ et $g : \mathbb{R}^p \times \mathbb{R}^d \longrightarrow \mathbb{R}$ est une fonction connue, supposée deux fois différentiable.

Dans ce qui suit, on notera $\nabla_h g$ et $\nabla_h^2 g$, le gradient et la matrice Hessienne de g par rapport à la seconde variable h, et ∇G et $\nabla^2 G$, le gradient et la matrice Hessienne de G. On suppose que la matrice $\nabla^2 G(\theta)$ est définie positive.

Nous nous intéressons à l'estimation récursive (ou en ligne) du paramètre θ à partir d'une suite de vecteurs aléatoires indépendants $(X_n)_{n\geq 1}$ ayant la même distribution que X. L'une des méthodes les plus connues dans ce contexte est l'algorithme du gradient stochastique, défini de manière récursive pour tout $n \geq 1$ par :

$$\theta_n^{SG} = \theta_{n-1}^{SG} - \nu_n \nabla_h g\left(X_n, \theta_{n-1}^{SG}\right)$$

où θ_0^{SG} est une valeur initiale choisie arbitrairement et $(\nu_n)_{n\geq 1}$ est une suite de nombres réels positifs décroissant vers 0.

Malgré leur efficacité reconnue, ces méthodes peuvent être très sensibles aux problèmes dit "mal conditionné", où la Hessienne a des valeurs propres à différentes échelles (voir par exemple, Bercu et al., 2020; Leluc et Portier, 2023). Pour surmonter ce problème, des algorithmes stochastiques du second ordre de la forme

$$\theta_n = \theta_{n-1} - \nu_n \, A_n \nabla_h g(X_n, \theta_{n-1})$$

ont été proposés et récemment étudiés. Ici, $(\nu_n)_{n\geq 1}$ est une suite de nombres réels positifs décroissant vers 0 et la matrice A_n est un estimateur récursif de l'inverse de la matrice Hessienne de G en θ, c'est-à-dire un estimateur récursif de H^{-1} avec $H = \nabla^2 G(\theta)$. La principale difficulté réside donc dans la construction de cette suite A_n.

Plusieurs algorithmes récursifs du second ordre ont été proposés et étudiés. Par exemple, Bercu et al. (2020) proposent un algorithme de Newton stochastique efficace pour estimer les paramètres d'un modèle de régression logistique. Dans un travail récent, Bercu et al. (2023) proposent un algorithme de Gauss-Newton stochastique pour estimer le coût de Transport Optimal entre deux mesures de probabilité discrètes. Godichon-Baggioni et al. (2024) proposent des algorithmes du second ordre pour résoudre le problème de régression Ridge dans le cadre linéaire et logistique, tandis que le cas de la médiane géométrique est introduit et étudié par Godichon-Baggioni et Lu (2023). Dans tous ces algorithmes, les estimateurs de l'inverse de la matrice Hessienne sont mis à jour de manière récursive en utilisant la formule d'inversion de Riccati (également appelée formule de Sherman-Morrison, voir par exemple Duflo (1996)). Ce calcul est rendu possible grâce à la forme particulière de l'estimateur de la matrice Hessienne H, présentée comme $(1/n)\sum_{k=1}^{n} a_k \phi_k \phi_k^T$, où $(a_n)_{n\geq 1}$ est une suite de variables aléatoires réelles positives et $(\phi_n)_{n\geq 1}$ est une suite de vecteurs aléatoires dans $\mathbb{R}^d$.

Cependant, il n'est pas toujours possible d'obtenir un tel estimateur de la matrice Hessienne. Dans ce travail, nous proposons de construire un estimateur récursif de H^{-1} sans tenter d'abord de construire un estimateur de H. Cette approche est basée sur le fait que nous avons $HH^{-1} = H^{-1}H = I_d$ et, par conséquent, la relation suivante :

$$\mathbb{E}\left[H^{-1}\nabla_h^2 g(X,\theta) + \nabla_h^2 g(X,\theta)H^{-1} - 2I_d\right] = 0 \tag{1}$$

où I_d désigne la matrice identité d'ordre d. En utilisant un algorithme de type Robbins-Monro, nous proposons un estimateur récursif de la matrice H^{-1} définie pour tout $n \geq 1$

par

$$A_n = A_{n-1} - \gamma_n \left(A_{n-1} \nabla_h^2 g(X_n, \theta_{n-1}) + \nabla_h^2 g(X_n, \theta_{n-1}) A_{n-1} - 2\, I_d \right)$$

où $(\gamma_n)_{n\geq 1}$ est une suite de nombres réels positifs décroissant vers 0 et θ_{n-1} est un estimateur de θ.

Nous obtenons ainsi un estimateur universel de l'inverse de la Hessienne. Après avoir apporté de légères modifications pour réduire la complexité du calcul et pour contrôler les valeurs propres de A_n, nous établissons la vitesse de convergence presque sûre pour l'estimateur proposé. Ces résultats restent vrais pour tout estimateur consistant de θ_n. Sur la base de ce concept, nous introduisons un algorithme de Newton stochastique universel. Pour améliorer davantage la vitesse de convergence, nous considérons également sa version moyennée pondérée, comme discuté par Mokkadem et Pelletier(2011); Boyer et Godichon-Baggioni (2023). Enfin, nous donnerons leurs vitesses de convergence et démontrons l'efficacité asymptotique des estimateurs moyennés pondérés.

2 Contexte

Nous considérons le problème de minimisation de la fonction convexe $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ définie pour tout $h \in \mathbb{R}^d$ par :

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

où la perte $g(X, \cdot)$ est une fonction convexe, dérivable deux fois et X est un vecteur aléatoire de $\mathbb{R}^p$. Nous supposons qu'il existe une valeur unique $\theta \in \mathbb{R}^d$ telle que

$$\nabla G(\theta) = 0.$$

Introduisons maintenant les hypothèses pour le cadre d'estimation du paramètre θ:

(A1) Il existe $C > 0$ tel que pour tout $h \in \mathbb{R}^d$,

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

(A2) La fonction G est deux fois différentiable et $\nabla^2 G(\theta)$ est définie positive. De plus, la Hessienne est uniformément bornée, c'est-à-dire qu'il existe une constante positive $L_{\nabla G}$ telle que pour tout $h \in \mathbb{R}^d$,

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

(A3) La fonction $\nabla^2 G$ est Lipschitzienne dans un voisinage de θ, c'est-à-dire qu'il existe des constantes positives $r > 0$ et L_r telles que pour tout $h \in \mathcal{B}(\theta, r)$

$$\left\| \nabla^2 G(h) - \nabla^2 G(\theta) \right\|_{op} \leq L_r \left\| \theta - h \right\|,$$

où $\mathcal{B}(\theta, r)$ désigne une boule de rayon r centrée en θ.

(A4) Il existe $q > 2$ et C_q tel que pour tout $h \in \mathbb{R}^d$,

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

Ces hypothèses sont très proches de celles présentées dans la littérature (Pelletier, 2000; Gadat et Panloup, 2017; Godichon-Baggioni, 2019).

3 Estimation de l'inverse de la Hessienne

On s'intéresse dans cette section à l'estimation de l'inverse de la Hessienne de la fonction G en θ, noté H^{-1} où $H = \nabla^2 G(\theta)$. Soit $(X_n)_{n\geq 1}$ une suite de vecteurs aléatoires indépendants dans $\mathbb{R}^p$ ayant la même distribution que X. Supposons d'abord que θ est connu. À partir de l'égalité (1), la matrice H^{-1} satisfait une équation de la forme $\Phi(H^{-1}) = 0$. Nous pouvons alors employer la procédure de Robbins-Monro (Robbins et Monro, 1951) pour estimer récursivement le paramètre H^{-1}. En désignant cet estimateur par $\hat{A}_n$, pour tout $n \geq 1$, nous avons :

$$\hat{A}_n = \hat{A}_{n-1} - \gamma_n \left(\hat{A}_{n-1}\nabla_h^2 g(X_n,\theta) + \nabla_h^2 g(X_n,\theta)\hat{A}_{n-1} - 2\, I_d\right),$$

où $\hat{A}_0$ est une matrice symétrique définie positive choisie arbitrairement, et $\gamma_n = c_\gamma n^{-\gamma}$ avec $\frac{1}{2} < \gamma < 1$ et $c_\gamma > 0$. Il est important de noter que par construction, la matrice $\hat{A}_n$ est symétrique pour tout $n \geq 1$. Cependant, puisque θ est inconnu, nous devons l'estimer. En supposant que nous disposons d'un estimateur récursif efficace de θ (par exemple, un estimateur du gradient stochastique), nous pouvons facilement déduire un estimateur de H^{-1} en utilisant une procédure de substitution :

$$\hat{A}_n = \hat{A}_{n-1} - \gamma_n \left(\hat{A}_{n-1}\nabla_h^2 g(X_n,\tilde{\theta}_{n-1}) + \nabla_h^2 g(X_n,\tilde{\theta}_{n-1})\hat{A}_{n-1} - 2\, I_d\right).$$

Cet estimateur est toujours symétrique mais pas nécessairement défini positif. Pour garantir cette dernière propriété, nous introduisons une troncature basée sur la norme de $\nabla_h^2 g(X_n,\tilde{\theta}_{n-1})$, conduisant à l'estimateur suivant de H^{-1} :

$$\hat{A}_n = \hat{A}_{n-1} - \gamma_n \left(\hat{A}_{n-1}\nabla_h^2 g(X_n,\tilde{\theta}_{n-1}) + \nabla_h^2 g(X_n,\tilde{\theta}_{n-1})\hat{A}_{n-1} - 2I_d\right) \mathbb{1}_{\{\|\nabla_h^2 g(X_n,\tilde{\theta}_{n-1})\|_{op}\leq \beta_n\}}, \tag{2}$$

où $\beta_n = c_\beta n^\beta$ avec $\frac{1-\gamma}{q-1} < \beta < \gamma - \frac{1}{2}$ et $0 < c_\gamma c_\beta < \frac{1}{2}$. En outre, cette troncature permet de contrôler la plus petite valeur propre de $\hat{A}_n$, ce qui est utile pour étudier un estimateur du paramètre θ impliquant la matrice $\hat{A}_n$. Cela est particulièrement important pour établir la consistance de l'algorithme de Newton stochastique présenté dans la Section 4.

Cependant, bien que cet estimateur soit efficace, chaque mise à jour nécessite des multiplications matricielles, induisant une complexité computationnelle de l'ordre de $\mathcal{O}(d^3)$, qui est la même que pour l'inversion matricielle. Néanmoins, nous pouvons introduire un algorithme d'une complexité en $\mathcal{O}(d^2)$, basé sur l'observation suivante : soit Z un vecteur aléatoire centré dans $\mathbb{R}^d$ avec une matrice de variance-covariance I_d, indépendant du vecteur X. Alors,

$$\mathbb{E}\left[H^{-1}\, Z\, Z^T \nabla_h^2 g(X,\theta) \; + \nabla_h^2 g(X,\theta) Z\, Z^T H^{-1} \; - \; 2I_d\right] = 0.$$

Ainsi, en considérant une suite $(Z_n)_{n\geq 1}$ de vecteurs aléatoires indépendants et identiquement distribués bornés de $\mathbb{R}^d$ tels que $\mathbb{E}[Z_n] = 0$ et $\mathbb{E}\left[Z_n Z_n^T\right] = I_d$, et indépendants de $(X_n)_{n\geq 1}$, nous pouvons proposer un autre estimateur de H^{-1} défini pour tout $n \geq 1$ comme suit :

$$\begin{aligned} P_n &= A_{n-1} Z_n \\ Q_n &= \nabla_h^2 g(X_n, \tilde{\theta}_{n-1}) Z_n \\ A_n &= A_{n-1} - \gamma_n \left(P_n Q_n^T + Q_n P_n^T - 2\, I_d\right) \mathbb{1}_{\{M\|Q_n\| \leq \beta_n\}} \end{aligned} \tag{3}$$

où A_0 est une matrice symétrique et définie positive et M vérifie $\|Z_n\| \leq M$.

4 Algorithme de Newton Stochastique Universel

Dans cette section, nous introduisons l'Algorithme de Newton Stochastique Universel défini pour tout $n \geq 1$ par

$$\begin{aligned} \widehat{P}_n &= \widehat{A}_{n-1} Z_n \\ \widehat{Q}_n &= \nabla_h^2 g(X_n, \widehat{\theta}_{n-1}) Z_n \\ \widehat{A}_n &= \widehat{A}_{n-1} - \gamma_n \left(\widehat{P}_n \widehat{Q}_n^T + \widehat{Q}_n \widehat{P}_n^T - 2 I_d\right) \mathbb{1}_{M\|Q_n\| \leq \beta_n} \\ \widehat{\theta}_n &= \widehat{\theta}_{n-1} - \nu_n \widehat{A}_{n-1} \nabla_h g(X_n, \widehat{\theta}_{n-1}). \end{aligned}$$

En suivant le même schéma de preuve que pour le Théorème 4.2, on peut montrer que :

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

De plus, suivant Mokkadem et Pelletier (2011); Boyer et Godichon-Baggioni (2022), nous pouvons proposer un algorithme de Newton Stochastique Universel Moyenné Pondéré. Il est donné par

$$\begin{aligned} P_n &= A_{n-1} Z_n \\ Q_n &= \nabla_h^2 g(X_n, \theta_{n-1,\tau'}) Z_n \\ \theta_n &= \theta_{n-1} - \nu_n A_{n-1,\tau} \nabla_h g(X_n, \theta_{n-1}) && (4) \\ \theta_{n,\tau'} &= \left(1 - \frac{\ln(n+1)^{\tau'}}{\sum_{k=0}^n \ln(k+1)^{\tau'}}\right) \theta_{n-1,\tau'} + \frac{\ln(n+1)^{\tau'}}{\sum_{k=0}^n \ln(k+1)^{\tau'}} \theta_n && (5) \\ A_n &= A_{n-1} - \gamma_n \left(P_n Q_n^T + Q_n P_n^T - 2\, I_d\right) \mathbb{1}_{M\|Q_n\| \leq \beta_n} && (6) \\ A_{n,\tau} &= \left(1 - \frac{\ln(n+1)^{\tau}}{\sum_{k=0}^n \ln(k+1)^{\tau}}\right) A_{n-1,\tau} + \frac{\ln(n+1)^{\tau}}{\sum_{k=0}^n \ln(k+1)^{\tau}} A_n && (7) \end{aligned}$$

où $(\nu_n)_{n\geq 1}$ est une suite de nombres réels positifs définie pour tout $n \geq 1$ par $\nu_n = c_\nu n^\nu$ avec $c_\nu > 0$ et $\nu \in (1/2, 1-\beta)$ satisfaisant $\gamma + \nu > 3/2$. De plus, $\tau, \tau' \geq 0$. Le théorème suivant donne la consistance des estimateurs définis par (4) et (5).

Theorem 4.1 *Supposons que les hypothèses* ***(A1)*** *à* ***(A4)*** *soient vérifiées. Soient* θ_n *et* $\theta_{n,\tau'}$ *définis comme dans* (4) *et* (5)*. Alors,*

$$\theta_n \xrightarrow[n\to\infty]{p.s.} \theta \quad et \quad \theta_{n,\tau'} \xrightarrow[n\to\infty]{p.s.} \theta.$$

Notez que la contrainte $\gamma + \nu > 3/2$ est de nature purement technique mais est cruciale pour l'application du Théorème de Robbins-Siegmund et donc pour obtenir la consistance des estimateurs. Cependant, nous pensons que cette condition pourrait ne pas être nécessaire en pratique. Nous pouvons maintenant donner les vitesses de convergence presque sûr des estimateurs.

Theorem 4.2 *Supposons que les hypothèses* ***(A1)*** *à* ***(A4)*** *soient vérifiées. Alors* θ_n *et* $\theta_{n,\tau'}$ *définis par* (4) *et* (5) *vérifient pour tout* $\delta > 0$

$$\|\theta_n - \theta\|^2 = o\left(\frac{\ln n^{1+\delta}}{n^\nu}\right) p.s. \quad et \quad \|\theta_{n,\tau'} - \theta\|^2 = o\left(\frac{\ln n^{1+\delta}}{n}\right) p.s.$$

De plus, A_n *et* $A_{n,\tau}$ *définis par* (6) *et* (7) *vérifient*

$$\|A_n - H^{-1}\|^2 = o\left(\frac{\ln n^{1+\delta}}{n^\gamma}\right) p.s. \quad et \quad \|A_{n,\tau} - H^{-1}\|^2 = o\left(\frac{\ln n^{1+\delta}}{n}\right) p.s.$$

De plus, les estimateurs $\theta_{n,\tau'}$ *définies par* (5) *vérifient*

$$\sqrt{n}(\theta_{n,\tau'} - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, H^{-1}\Sigma H^{-1}),$$

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

Les estimateurs de Newton Stochastique Universel Moyenné Pondéré atteignent ainsi l'efficacité asymptotique sous des hypothèses très faibles.

Bibliographie

Bercu, B., Godichon, A., & Portier, B. (2020). An efficient stochastic newton algorithm for parameter estimation in logistic regressions. SIAM Journal on Control and Optimization, 58(1), 348-367.

Bercu, B., Bigot, J., Gadat, S., & Siviero, E. (2023). A stochastic Gauss–Newton algorithm for regularized semi-discrete optimal transport. Information and Inference: A Journal of the IMA, 12(1), 390-447.

Boyer, C., & 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, M. (1996). Algorithmes stochastiques (Vol. 23, pp. xiv+-319). Berlin: Springer.

Gadat, S., & Panloup, F. (2017). Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity. arXiv preprint arXiv:1709.03342.

Godichon-Baggioni, A. (2019). Online estimation of the asymptotic variance for averaged stochastic gradient algorithms. Journal of Statistical Planning and Inference, 203, 1-19.

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

Godichon-Baggioni, A., & Lu, W. (2023). Online stochastic Newton methods for estimating the geometric median and applications. arXiv preprint arXiv:2304.00770.

Leluc, R., & Portier, F. (2023). Asymptotic Analysis of Conditioned Stochastic Gradient Descent. Transactions on Machine Learning Research.

Mokkadem, A., & 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.

Pelletier, M. (2000). Asymptotic almost sure efficiency of averaged stochastic algorithms. SIAM Journal on Control and Optimization, 39(1), 49-72.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. The annals of mathematical statistics, 400-407.