Online estimation of the inverse of the Hessian for stochastic optimization with application to universal stochastic Newton algorithms

Antoine Godichon-Baggioni, Wei Lu, Bruno Portier

Online estimation of the inverse of the Hessian for stochastic optimization with application to universal stochastic Newton algorithms

Antoine Godichon-Baggioni [*], Wei Lu[†] and Bruno Portier[†]

January 12, 2024

Abstract

This paper 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 reduces computational complexity. Above all, it allows to develop universal stochastic Newton methods and investigate the asymptotic efficiency of the proposed approach. This work so expands the application scope of second-order algorithms in stochastic optimization.

Keywords: Stochastic Newton algorithm; Stochastic Optimization; Robbins-Monro algorithm; online estimation

1 Introduction

In this paper, we consider the usual stochastic optimization problem, which consists of estimating the parameter $\theta \in \mathbb{R}^d$ defined by

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

where the function G is defined for all $h \in \mathbb{R}^d$ by : $G(h) = \mathbb{E}\left[g(X,h)\right]$ and X is a random vector of $\mathbb{R}^p$. The function $g : \mathbb{R}^p \times \mathbb{R}^d \longrightarrow \mathbb{R}$ is assumed to be twice continuously differentiable. This problem arises in various contexts such as estimating the parameters of logistic regressions (Bach, 2014; Cohen et al., 2017), geometric median and quantiles (Cardot et al., 2013, 2017), or superquantiles (Bercu et al., 2020a; Costa and Gadat, 2020). We denote $\nabla_h g$ and $\nabla_h^2 g$ as the gradient and Hessian matrix of g with respect to the second variable h, and ∇G and $\nabla^2 G$ as the gradient and Hessian matrix of G. It is assumed that the matrix $\nabla^2 G(\theta)$ is positive definite.

Starting from a sequence of independent random vectors $(X_n)_{n\geq 1}$ with the same distribution as X, we aim to online estimate the parameter θ. One of the most well-known methods in this context is certainly the stochastic gradient algorithm, recursively defined for all $n \geq 1$ by:

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

where θ_0^{SG} is an arbitrarily chosen initial value and $(\nu_n)_{n\geq 1}$ is a sequence of positive real numbers decreasing towards 0. These algorithms have been extensively studied, with asymptotic results found by Pelletier (1998, 2000) and non-asymptotic results, such as uniform bounds of the quadratic mean error, presented by Moulines and Bach (2011); Gadat and Panloup (2017); Godichon-Baggioni (2021) to name a few. To ensure asymptotic efficiency, an additional step consists of considering an averaged version of the estimates (Polyak and Juditsky, 1992).

[*]antoine.godichon_baggioni@upmc.fr, Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, 4 Place Jussieu, 75005 Paris, France.

[†]Laboratoire de Mathématiques de l'INSA Rouen Normandie, INSA Rouen Normandie, BP 08 - Avenue de l'Université, 76800 Saint-Etienne du Rouvray, France

Despite their known efficiency, these methods can be very sensitive to ill-conditioned problems, where the Hessian has eigenvalues at different scales (Leluc and Portier, 2020; Bercu et al., 2020b). To overcome this problem, second-order stochastic algorithms of the form

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

have been proposed and recently studied. Here, $(\nu_n)_{n\geq 1}$ is a sequence of positive real numbers decreasing towards 0 and the matrix A_n is a recursive estimate of the inverse of the Hessian matrix of G at θ, i.e a recursive estimate of H^{-1} with $H = \nabla^2 G(\theta)$. The challenge lies in constructing the recursive estimate A_n.

Several recursive second-order algorithms have been proposed and studied. For example, Bercu et al. (2020b) propose an efficient stochastic Newton algorithm for estimating the parameters of a logistic regression model. In a recent work, Bercu et al. (2023) propose a stochastic Gauss-Newton algorithm to estimate the entropically regularized Optimal Transport cost between two discrete probability measures. Cénac et al. (2020) study the asymptotic properties of a stochastic Gauss-Newton algorithm for estimating the parameters of a non-linear regression model. Godichon-Baggioni et al. (2022) propose second-order algorithms to solve the Ridge regression problem in the linear and logistic framework, while the case of the geometric median is introduced and studied by Godichon-Baggioni and Lu (2023). In all these algorithms, the estimate of the inverse of the Hessian matrix is recursively computed using the Riccati inversion formula (also called Sherman-Morrison formula, see e.g. Duflo (1990) p. 96). This calculation is made possible thanks to the particular form of the estimate of the Hessian matrix H, presented as $(1/n)\sum_{k=1}^{n} a_k \phi_k \phi_k^T$, where $(a_n)_{n\geq 1}$ is a sequence of positive real random variables and $(\phi_n)_{n\geq 1}$ is a sequence of random vectors in $\mathbb{R}^d$.

However, it is not always possible to obtain such an estimate of the Hessian matrix. In this work, we propose to construct a direct recursive estimate of H^{-1} without first attempting to construct an estimate of H. This approach is based on the fact that we have $HH^{-1} = H^{-1}H = I_d$ and, consequently, the following relation:

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

where I_d denotes the identity matrix of order d. Using a Robbins-Monro type algorithm, we propose a recursive estimate of the matrix H^{-1} defined for all $n \geq 1$ by:

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

where $(\gamma_n)_{n\geq 1}$ is a sequence of positive real numbers, decreasing towards 0 and θ_{n-1} is an estimate of θ.

However, the complexity of computing this estimate is of order $\mathcal{O}(d^3)$, which is the same as directly calculating the inverse of an estimate of matrix H. Nevertheless, we can introduce an algorithm with complexity of order $\mathcal{O}(d^2)$ based on the following observation: let Z be a centered random vector in $\mathbb{R}^d$ with variance-covariance matrix I_d, independent of the vector X. Then,

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

Therefore, considering a sequence $(Z_n)_{n\geq 1}$ of random vectors in $\mathbb{R}^d$ independent of the sequence $(X_n)_{n\geq 1}$ leads to an estimate of the form:

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

We thus obtain a universal estimate of the inverse of the Hessian, and thus with reduced calculus time. To further enhance convergence rate, we also consider its weighted averaged version, as discussed by Mokkadem and Pelletier (2011); Boyer and Godichon-Baggioni (2022). We establish the almost sure rates of convergence for the proposed estimates, after making slight modifications. These results remain true for any consistent estimates θ_n. Based on this concept, we introduce a universal recursive Newton algorithm and its weighted averaged version. Additionally, we provide their convergence rates and demonstrate the asymptotic efficiency of the weighted averaged estimates.

This paper is organized as follows: Section 2 concerns the framework and the main assumptions. Section 3 deals with the estimation of H^{-1} and the main convergence results while Section 4 concerns the Universal Weighted Averaged Stochastic Newton algorithm. A simulation study highlights the performance of the proposed methods in Section 5. The proofs of the different results are postponed in Section 6.

2 Framework

We consider the problem of minimizing the convex function $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ defined for all $h \in \mathbb{R}^d$ by:

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

where the loss $g(X, \cdot)$ is a convex, twice-differentiable function and X is a random vector of $\mathbb{R}^p$. We assume that there exists a unique value $\theta \in \mathbb{R}^d$ such that

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

This assumption, couple with strict convexity, ensures the existence of a minimizer for G and provides a well-defined optimization problem. Now, let's introduce the assumptions that underlie the parameter estimation framework for θ:

(A1) There exists $C > 0$ such that for all $h \in \mathbb{R}^d$,

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

(A2) The functional G is twice continuously differentiable and $\nabla^2 G(\theta)$ is positive. In addition, the Hessian is uniformly bounded, i.e there exists a positive constant $L_{\nabla G}$ such that for all $h \in \mathbb{R}^d$,

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

(A3) The function $\nabla^2 G$ is Lipschitz on a neighborhood of θ, i.e. there exist positive constants $r > 0$ and L_r such that for all $h \in \mathcal{B}(\theta, r)$

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

where $\mathcal{B}(\theta, r)$ denotes a ball of radius r centered at θ.
(A4) There exists $q > 2$ and C_q such that for all $h \in \mathbb{R}^d$,

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

These assumptions are very close to those presented in the literature (Pelletier, 2000; Gadat and Panloup, 2017; Godichon-Baggioni, 2019). Assumption **(A1)** controls the growth of the gradient and guarantees the stability of the estimation process. It ensures that the gradient remains bounded as the estimation progresses. Assumption **(A2)** ensures that the curvature of G at θ is well-behaved, allowing the estimation algorithm to reliably exploit the local structure of G. This assumption also guarantees that the gradient of G is Lipschitz continuous with a constant $L_{\nabla G}$. This Lipschitz continuity is crucial as convergence results are obtained using a Taylor's expansion of G up to the second order. Assumption **(A3)** indicates that the Hessian matrix does not exhibit abrupt changes within a neighborhood around θ. This assumption ensures the stability of the Hessian estimates during the estimation process. Assumption **(A4)** guarantees that the second-order derivative of G does not exhibit excessive fluctuations. It imposes a bound on the variation of the Hessian matrix, providing further stability to the estimation algorithm. It's worth noting that Hölder's inequality leads to

$$\left\|\nabla^2 G(h)\right\|_F \leq \mathbb{E}\left[\left\|\nabla_h^2 g(X, h)\right\|_F^q\right]^{1/q} \leq C_q^{1/q}.$$

Of course, this inequality intertwines with the bound in Assumption **(A2)**, but we keep the notation $L_{\nabla G}$ for the sake of clarity.

These assumptions, along with the differentiability properties of g and G, provide a solid foundation for developing efficient second order methods to solve the minimization problem and obtain reliable estimates of θ.

3 Estimation of the Hessian inverse

In this section, our focus is solely on estimating the inverse Hessian of function G in θ, denoted as H^{-1} with $H = \nabla^2 G(\theta)$. Even if our motivation is to estimate H^{-1} for proposing a second-order algorithm, the estimation of H^{-1} can also be valuable for recursively constructing confidence intervals or significance statistical tests for a component of parameter θ when the parameter θ is estimated using another asymptotically efficient algorithm like the averaged stochastic gradient algorithm. Indeed, in most cases, the asymptotic variance involving in the central limit theorem, generally depends on matrix H^{-1} and its estimation is then required.

Let $(X_n)_{n\geq 1}$ be a sequence of independent random vectors in $\mathbb{R}^p$ with the same distribution as X. Assume first that θ is known. From equality (1), the matrix H^{-1} satisfies an equation of the form $\Phi(H^{-1}) = 0$. We can then employ the Robbins-Monro procedure (Robbins and Monro, 1951) to recursively estimate the parameter H^{-1}. Denoting this estimator as $\widehat{A}_n$, for any $n \geq 1$, we have:

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

where $\widehat{A}_0$ is an arbitrary symmetric positive definite matrix, and $\gamma_n = c_\gamma n^{-\gamma}$ with $\frac{1}{2} < \gamma < 1$ and $c_\gamma > 0$. It is important to note that $\widehat{A}_n$ is symmetric for any $n \geq 1$ due to its construction. However, since θ is unknown, we need to estimate it. Assuming we have an efficient recursive estimator of θ (e.g., a stochastic gradient estimator), we can easily derive an estimator of H^{-1} using a plug-in procedure:

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

This estimator is always symmetric but not necessarily positive definite. To ensure positive definiteness, we introduce a truncation based on the norm of $\nabla_h^2 g(X_n, \widehat{\theta}_{n-1})$, leading to the following estimator of H^{-1}:

$$\widehat{A}_n = \widehat{A}_{n-1} - \gamma_n \left(\widehat{A}_{n-1} \nabla_h^2 g(X_n, \widehat{\theta}_{n-1}) + \nabla_h^2 g(X_n, \widehat{\theta}_{n-1}) \widehat{A}_{n-1} - 2 I_d \right) \mathbf{1}_{\left\{ \left\| \nabla_h^2 g(X_n, \widehat{\theta}_{n-1}) \right\|_{op} \leq \beta_n \right\}}, \tag{3}$$

where $\beta_n = c_\beta n^\beta$ with $\frac{1-\gamma}{q-1} < \beta < \gamma - \frac{1}{2}$ and $0 < c_\gamma c_\beta < \frac{1}{2}$. Additionally, this truncation enables control over the smallest eigenvalue of $\widehat{A}_n$, which is useful for studying an estimator of the parameter θ involving the matrix $\widehat{A}_n$. This is particularly important in establishing the consistency of the Stochastic Newton algorithm presented in Section 4.

However, although this estimator is efficient, each update requires matrix multiplications, resulting in a computational complexity of order $\mathcal{O}(d^3)$, which is the same as matrix inversion. Hence, it is necessary to improve the complexity of each update of $\widehat{A}_n$.

Building on equality (2), considering a sequence $(Z_n)_{n\geq 1}$ of independent and identically distributed bounded random vectors of $\mathbb{R}^d$ such that $\mathbb{E}[Z_n] = 0$ and $\mathbb{E}\left[Z_n Z_n^T\right] = I_d$, and independent of $(X_n)_{n\geq 1}$, we can propose another estimate of H^{-1} defined for any $n \geq 1$ as follows:

$$\begin{aligned} P_n &= A_{n-1} Z_n \\ Q_n &= \nabla_h^2 g(X_n, \widehat{\theta}_{n-1}) Z_n \\ A_n &= A_{n-1} - \gamma_n \left(P_n Q_n^T + Q_n P_n^T - 2\, I_d \right) \mathbf{1}_{\{\|Q_n\| \|Z_n\| \leq \beta_n\}} \end{aligned} \tag{4}$$

where A_0 is an arbitrary symmetric and positive definite matrix.

We can observe that in this algorithm the truncation is only based on $\|Q_n\|\,\|Z_n\|$, and not on $\left\|Q_n Z_n^T\right\|_{op}$ as expected, because we have $\left\|Q_n Z_n^T\right\|_{op} \leq \|Q_n\|\,\|Z_n\|$. Notably, the computational complexity of each update of A_n is now reduced to $\mathcal{O}(d^2)$. Moreover, following Mokkadem and Pelletier (2011); Boyer and Godichon-Baggioni (2022), we can propose a weighted averaged estimate $A_{n,\tau}$, which performs better in practice when the initialization is poor. It is given by

$$A_{n,\tau} = \left(1 - \frac{\ln(n+1)^\tau}{\sum_{k=0}^n \ln(k+1)^\tau} \right) A_{n-1,\tau} + \frac{\ln(n+1)^\tau}{\sum_{k=0}^n \ln(k+1)^\tau} A_n. \tag{5}$$

This estimator can be recursively computed since for any $n \geq 1$, $\sum_{k=0}^{n} \ln(k+1)^{\tau} = \ln(n+1)^{\tau} + \sum_{k=0}^{n-1} \ln(k+1)^{\tau}$. The following theorem establishes the consistency of the estimators A_n and $A_{n,\tau}$ for the parameter H^{-1} in the context of estimating the inverse of the Hessian matrix. It states that the convergence rate depends on multiple factors, including the step sequence γ_n, the truncation parameter β_n, the regularization parameter τ, and the convergence rate of the estimate $\widehat{\theta}_n$.

Theorem 3.1. *Assume that Assumptions **(A2)** to **(A4)** hold, and that there is an estimate $\hat{\theta}_n$ satisfying for all $\delta > 0$*

$$\left\|\hat{\theta}_n - \theta\right\|^2 = o\left(\frac{\ln n^{1+\delta}}{n^a}\right) a.s.,$$

with $a > 0$. Then A_n and $A_{n,\tau}$ defined by (4) and (5) satisfy

$$\left\|A_n - H^{-1}\right\|^2 = o\left(\frac{\ln n^{1+\delta}}{n^{\min\{\gamma, a, 2\beta(q-1)\}}}\right) a.s. \quad \text{and} \quad \left\|A_{n,\tau} - H^{-1}\right\|^2 = o\left(\frac{\ln n^{1+\delta}}{n^{\min\{1, a, 2\beta(q-1)\}}}\right) a.s.$$

The proof is given in Section 6. Observe that if $a = 1$, one can achieve the usual rate of convergence taking $\beta > \frac{1}{2(q-1)}$, which is only possible if $q > 1 + \frac{1}{2\gamma-1}$. For instance, taking the usual parametrization $\gamma = 3/4$ or $2/3$, q must satisfy $q > 3$ or $q > 4$.

4 Universal Weighted Averaged Stochastic Newton Algorithm

In this section, we introduce the Universal Weighted Averaged Stochastic Newton algorithm and discuss its main properties. As mentioned in Theorem 3.1, using a Weighted Averaged version of the inverse Hessian estimate can yield improved theoretical results. Therefore, we incorporate this choice into the Stochastic Newton algorithm. Furthermore, we have observed that the convergence rate of the estimate for θ significantly influences the theoretical behavior of the estimates of H^{-1}. Consequently, we incorporate the best possible estimate for parameter θ, namely the Weighted Averaged Stochastic Newton estimates, into the latter. This reasoning leads to the following Weighted Averaged Stochastic Newton algorithm defined for all $n \geq 1$ by

$$\begin{aligned}
P_n &= A_{n-1} Z_n \\
Q_n &= \nabla_h^2 g(X_n, \theta_{n-1,\tau'}) Z_n \\
\theta_n &= \theta_{n-1} - \nu_n A_{n-1,\tau} \nabla_h g(X_n, \theta_{n-1}) && (6) \\
\theta_{n,\tau'} &= \left(1 - \frac{\ln(n+1)^{\tau'}}{\sum_{k=0}^{n} \ln(k+1)^{\tau'}}\right) \theta_{n-1,\tau'} + \frac{\ln(n+1)^{\tau'}}{\sum_{k=0}^{n} \ln(k+1)^{\tau'}} \theta_n && (7) \\
A_n &= A_{n-1} - \gamma_n \left(P_n Q_n^T + Q_n P_n^T - 2\, I_d\right) \mathbf{1}_{\{\|Q_n\| \|Z_n\| \leq \beta_n\}} && (8) \\
A_{n,\tau} &= (1 - \frac{\ln(n+1)^{\tau}}{\sum_{k=0}^{n} \ln(k+1)^{\tau}}) A_{n-1,\tau} + \frac{\ln(n+1)^{\tau}}{\sum_{k=0}^{n} \ln(k+1)^{\tau}} A_n && (9)
\end{aligned}$$

where $(\nu_n)_{n \geq 1}$ is a sequence of positive real numbers defined for any $n \geq 1$ by $\nu_n = c_\nu n^\nu$ with $c_\nu > 0$ and $\nu \in (1/2, 1-\beta)$ satisfying $\gamma + \nu > 3/2$. In addition, $\tau, \tau' \geq 0$. The following theorem gives the consistency of the estimates defined by (6) and (7).

Theorem 4.1. *Assume that Assumptions **(A1)** to **(A4)** hold. Let θ_n and $\theta_{n,\tau'}$ be defined as in (6) and (7). Then,*

$$\theta_n \xrightarrow[n\to\infty]{a.s.} \theta \quad \text{and} \quad \theta_{n,\tau'} \xrightarrow[n\to\infty]{a.s.} \theta.$$

The proof is given in Section 6. Note that the constraint $\gamma + \nu > 3/2$ is of a purely technical nature and is crucial for the application of the Robbins-Siegmund Theorem and so that to get the consistency of the estimates. However, we believe this condition might not be necessary in practical applications. We can now give the almost sure rate of convergence of the estimates.

Theorem 4.2. *Assume that Assumptions* ***(A1)*** *to* ***(A4)*** *hold. Then* θ_n *and* $\theta_{n,\tau'}$ *defined by* (6) *and* (7) *satisfy for all* $\delta > 0$

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

In addition, A_n *and* $A_{n,\tau}$ *defined by* (8) *and* (9) *satisfy*

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

Moreover, the estimates $\theta_{n,\tau'}$ *defined by* (7) *satisfy*

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

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

The proof is given in Section 6. The Universal Weighted Averaged Stochastic Newton estimates so achieve the asymptotic efficiency, and so, under very weak assumptions.

Remark 4.1. *Mention that for estimating parameter* θ*, it is also possible to consider the following simpler algorithm, which we refer to as the Universal Stochastic Newton Algorithm, and only relies on* A_n*:*

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

By following the same scheme of proof as for Theorem 4.2, one could check that:

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

However, it can be observed that the convergence rate of $\widehat{\theta}_n$ *is not optimal. Nevertheless, this algorithm has the merit of being much simpler. In addition, mention that following the reasoning presented by Bercu et al. (2020b), one could take a step sequence of the form* $\nu_n = \frac{1}{n}$ *leading to the Stochastic Newton algorithm. However, we are unfortunately not able to obtain the consistency of the estimates in this context.*

5 Applications

The simulation section of this paper focuses on evaluating the performance of our novel methods, called Universal Stochastic Newton Algorithm (USNA) and Universal Weighted Averaged Stochastic Newton Algorithms (UWASNA). We begin by analyzing the performance of USNA and UWASNA in the context of logistic regression and the geometric median estimation. In both contexts, it was already feasible to employ second-order algorithms such as the Stochastic Newton Algorithms (SNA) and its Weighted Averaged version (WASNA). These two algorithms use the Riccati formula to recursively compute the inverse of the Hessian estimator. By demonstrating comparable results to SNA and WASNA, we establish USNA and UWASNA as viable alternatives with efficient performance. Additionally, we investigate the applicability of our method in challenging scenarios where using the Riccati formula in SNA is not feasible, particularly in estimating p-means and parameters of a spherical distribution. In these two cases, we compare the performances of USNA and UWASNA with the one of the Averaged Stochastic Gradient Descent (ASGD) introduced by Polyak and Juditsky (1992). Through comprehensive simulations, our goal is to verify that USNA and UWASNA consistently exhibit favourable performances even in these contexts. To close this section, we extend our evaluation to real-world datasets to showcase the practicality of deploying USNA and UWASNA algorithm in applications.

5.1 Choice of the hyperparameters

In our experiments, the choice of hyperparameters involved in the different algorithms, plays a significant role in achieving desired outcomes. The decision to set these values is based on both theoretical justifications and empirical observations.

1. **Setting of ν_n for USNA**: Despite the lack of a theoretical proof demonstrating the convergence rate of USNA when $\nu_n = 1/n$, this setting is adopted in our experiments for a direct comparison with SNA. Empirical results, as presented later, validate that this choice is effective in practice.

2. **Conditions on β and γ**: Condition $\beta < \gamma - 1/2$ is only used to apply the Robbins-Siegmund theorem. It serves a theoretical purpose, and we advocate for its removal. On the contrary, condition $\beta_{n+1}\gamma_{n+1} \leq 1/2$ is essential. It ensures the positiveness of the estimate of the inverse of the Hessian. For this reason, in all simulations, we set $\beta_{n+1} = \frac{1}{2}\gamma_n^{-1}$.

3. **Initialization of Estimators**: For initializing the estimators of the Hessian inverse, we consistently use $S_0^{-1} = I_d$ for both SNA and WASNA. Similarly, $A_0 = I_d$ is chosen for USNA and UWASNA.

5.2 Comparison with Riccati Newton

The objective of this section is to demonstrate the comparable performance of USNA and UWASNA when contrasted with SNA and WASNA, particularly in scenarios where the use of the Riccati formula is applicable to recursively compute the inverse of an estimator of the Hessian. Let us begin by revisiting this context. If we can estimate the Hessian matrix $H = \nabla^2 G(\theta)$ by an estimator of the form S_n/n with S_n defined by $S_n = \sum_{j=1}^n \varphi_j \varphi_j^T$ where $(\varphi_n)_{n\geq 1}$ is a sequence of random vectors of $\mathbb{R}^d$, then, thanks to the Riccati formula, we can recursively calculated matrix S_n^{-1} for any $n \geq 1$:

$$S_n^{-1} = S_{n-1}^{-1} - \frac{1}{1 + \varphi_n^T S_{n-1}^{-1} \varphi_n} S_{n-1}^{-1} \varphi_n \varphi_n^T S_{n-1}^{-1}, \tag{10}$$

with $S_0 = I_d$ to avoid the invertibility problem. This formula finds application in various scenarios, as previously demonstrated by Bercu et al. (2020b), for instance, to obtain efficient stochastic Newton algorithms. In light of this, we can define the stochastic Newton algorithm (SN) for estimating parameter θ as followed:

$$U_n = S_{n-1}^{-1} \varphi_n \tag{11}$$

$$S_n^{-1} = S_{n-1}^{-1} - (1 + \varphi_n^T U_n)^{-1} U_n U_n^T \tag{12}$$

$$\theta_n^{SN} = \theta_{n-1}^{SN} - S_n^{-1} \nabla_h g(X_n, \theta_{n-1}^{SN}) \tag{13}$$

where $S_0^{-1} = I_d$ and θ_0^{SN} is arbitrarily chosen. Note that the random vector φ_n is dependent on the current observation X_n and the previous estimation θ_{n-1}^{SN}. The Weighted Averaged Stochastic Newton Algorithm is defined by:

$$\begin{aligned}
\overline{U}_n &= \overline{S}_{n-1}^{-1} \overline{\varphi}_n \\
\overline{S}_n^{-1} &= \overline{S}_{n-1}^{-1} - (1 + \overline{\varphi}_n^T \overline{U}_n)^{-1} \overline{U}_n \overline{U}_n^T \\
\overline{\theta}_n &= \overline{\theta}_{n-1} - \gamma_{n+1} \overline{S}_n^{-1} \nabla_h g(X_n, \overline{\theta}_{n-1}) \\
\theta_n^{WASN} &= \left(1 - \frac{\ln(n+1)^{\tau'}}{\sum_{k=0}^n \ln(k+1)^{\tau'}}\right) \theta_{n-1}^{WASN} + \frac{\ln(n+1)^{\tau'}}{\sum_{k=0}^n \ln(k+1)^{\tau'}} \overline{\theta}_n
\end{aligned}$$

where $S_0^{-1} = I_d$, $\overline{\theta}_0$ and θ_0^{WASN} are arbitrarily chosen, the random vector $\overline{\varphi}_n$ is dependent on the current observation X_n and the previous estimation θ_{n-1}^{WASN}.

5.2.1 Logistic regression

Let (X, Y) be a random vector taking values in $\mathbb{R}^p \times \{0, 1\}$ and let us set $\phi = (1, X^T)^T$. In the binary logistic regression framework, function G to minimize is defined for any $h \in \mathbb{R}^{p+1}$ by:

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

where the conditional distribution of the binary response Y knowing ϕ is a Bernoulli distribution of parameter $\pi(\theta^T\phi)$ with for any $x \in \mathbb{R}$, $\pi(x) = \exp(x)/(1 + \exp(x))$ and $\theta \in \mathbb{R}^{p+1}$ is the unknown parameter to be estimated. It is easy to show that

$$\theta = \arg\min_{h \in \mathbb{R}^{p+1}} G(h)$$

and θ is the unique solution of equation $\nabla G(h) = 0$. In addition, we have

$$H = \mathbb{E}\left[a(\theta^T\phi)\phi\phi^T\right] \quad \text{with} \quad a(z) = \pi(z)(1 - \pi(z)).$$

Let $(\phi_n, Y_n)_{n\geq 1}$ be a sequence of independent random vectors in $\mathbb{R}^{p+1} \times \{0, 1\}$ with the same distribution as (ϕ, Y). In this context, Bercu et al. (2020b) uses algorithm (11)-(13) to estimate θ with $\varphi_n = \sqrt{a(\phi_n^T\theta_{n-1}^{SN})}\phi_n$ and $\nabla_h g(X_n, \theta_{n-1}^{SN}) = -\Phi_n\left(Y_n - \pi(\phi_n^T\theta_{n-1}^{SN})\right)$.

We conduct extensive simulations to evaluate the performance of our novel methods, USNA and UWASNA, in comparison to SNA and WASNA. For this purpose, we consider the logistic regression model introduced by Bercu et al. (2020b) where $p = 10$ and the true coefficients are set as follows:

$$\theta = (0, 3, -9, 4, -9, 15, 0, -7, 1, 0)^T.$$

We compare USNA and UWASNA against SNA and WASNA in terms of their ability to accurately estimate the true coefficients. The evaluation of the algorithms' performance is carried out using the mean squared error (MSE) metric. We simulate $N = 100$ independent sample of size $n = 10\,000$. The results are averaged to mitigate the effects of sampling fluctuations. For all algorithms, we initialize the estimate of the parameter with $\theta_{init} = \theta + e\epsilon$, where $\epsilon \sim \mathcal{N}(0, I_{p+1})$ and $e = 1$ or 2.

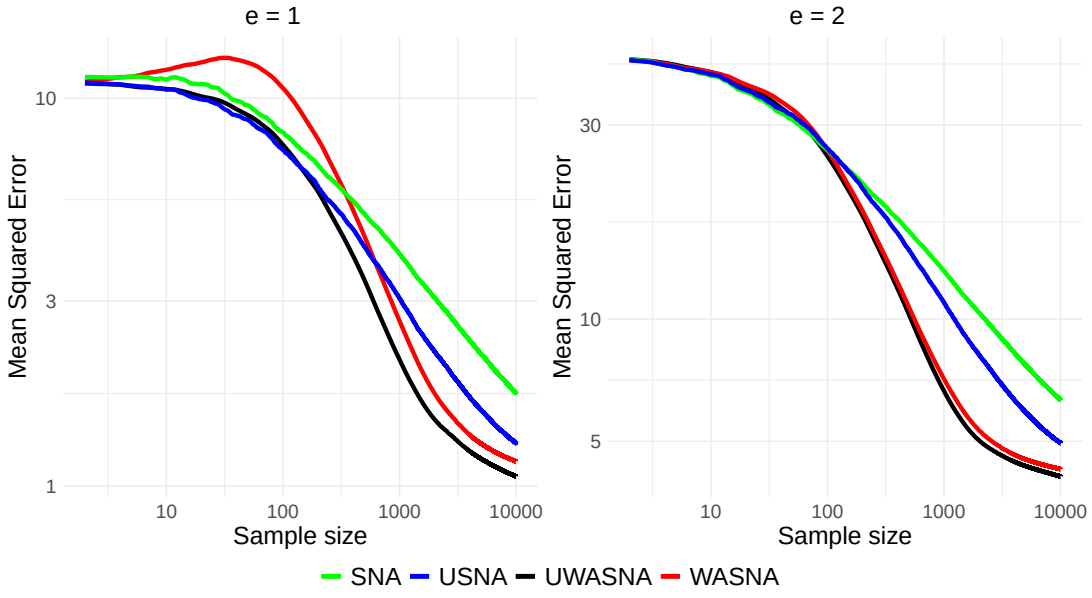

Figure 1: Evolution of the mean squared error with respect to the sample size for logistic regression.

As shown in the figure 1, the weighted averaged estimators converge more rapidly than the other two methods. Both USNA and UWASNA perform comparably to SNA and WASNA in accurately estimating the true coefficients of the logistic regression model. Remarkably, USNA and UWASNA achieve this without using the Riccati formula.

5.2.2 Geometric Median

Next, we conduct simulations in the context of geometric median estimation for a multivariate distribution. We focus on the model introduced by Godichon-Baggioni and Lu (2023). We generate $n = 10\,000$ copies of the random vector X of $\mathbb{R}^p$ with $p = 10$, where $X \sim \mathcal{N}(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma_{i,j}} = |\mathbf{i} - \mathbf{j}|^{\mathbf{0.5}}$. Recall that the geometric median is defined by:

$$m = \arg\min_{h \in \mathbb{R}^p} \mathbb{E}\left[\|X - h\| - \|X\|\right].$$

In this model, the result leads to $m = (0, \ldots, 0)^T$ in this model. In addition, the Hessian matrix H is defined by:

$$H = \mathbb{E}\left[\frac{1}{\|X - m\|}\left(I_p - \frac{(X-m)(X-m)^T}{\|X-m\|^2}\right)\right].$$

In this context, algorithm (11)-(13) can be used to estimate m taking $\nabla_h g(X_n, m_{n-1}^{SN}) = -\frac{X_n - m_{n-1}}{\|X_n - m_{n-1}\|}$ and $\varphi_n = \nabla_h g(X_n, m_{n-1}^{SN} + \alpha_n Z_n) - \nabla_h(X_n, m_{n-1}^{SN})$, where $\alpha_n = \frac{1}{n \ln(n+1)}$ and $(Z_n)_{n \geq 1}$ is a sequence of independent standard Gaussian vectors.

For a comprehensive evaluation of our methods' effectiveness in geometric median estimation, we also compare them against two baselines : SNA and WASNA. For all four algorithms, we initialize the estimate of the geometric median with $m_{init} = e\epsilon$, where $\epsilon \sim \mathcal{N}(0, I_{p+1})$ and $e = 1$ or 2. Throughout the simulations, we recorded the MSE of the estimated medians for the four algorithms. The simulation results were averaged over multiple iterations ($N = 100$).

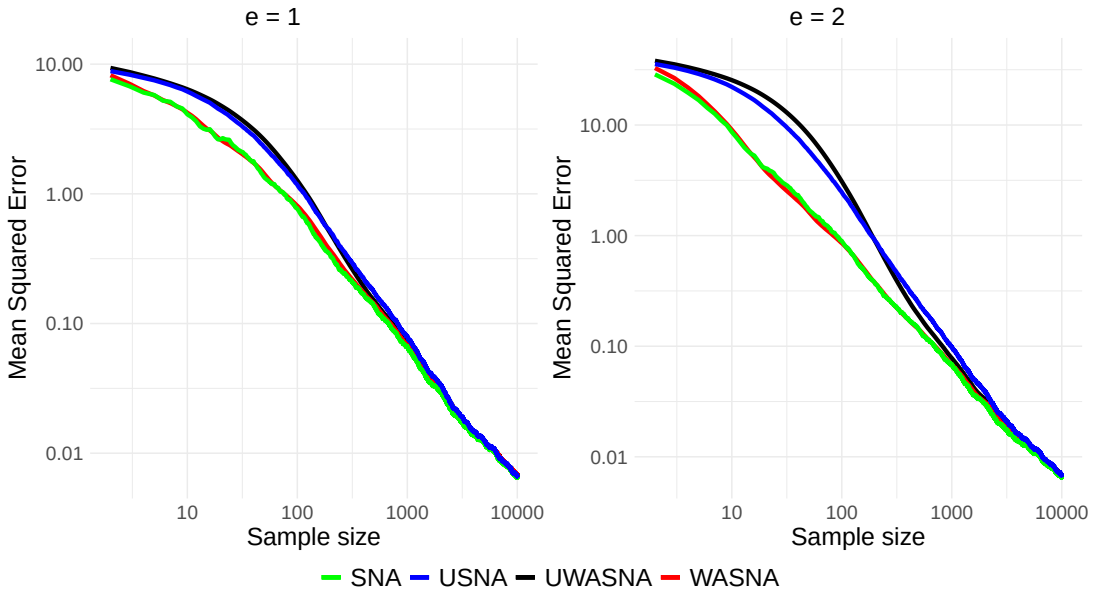

Figure 2: Evolution of the mean squared error with respect to the sample size for geometric median estimation.

In Figure 2, the results consistently indicate that when the sample size is relatively small (around 100), USNA and UWASNA converge slightly slower. However, beyond that point, they achieve performance on par with both WASNA and SNA in estimating the geometric median.

5.3 Cases where the Riccati formula cannot be used

In this section, we focus on the scenarios where using the Riccati formula in SNA or WASNA is not feasible. We explore the performance of USNA and UWASNA in contrast to an alternative averaged stochastic gradient-based method ("ASGD") proposed by Polyak and Juditsky (1992), which is defined as:

$$\theta_n^{SGD} = \theta_{n-1}^{SGD} - \eta_n \nabla_h g(X_n, \theta_{n-1}^{SGD}) \tag{14}$$

$$\theta_n^{ASGD} = \theta_{n-1}^{ASGD} + 1/n(\theta_n^{SGD} - \theta_{n-1}^{ASGD}) \tag{15}$$

where $(\eta_n)_{n \geq 1}$ is a sequence of learning rates and $\theta_0^{ASGD} = \theta_0^{SGD}$.

5.3.1 Spherical Distribution

In this paragraph, we focus on the estimation of the parameters of a spherical distribution (Godichon-Baggioni and Portier, 2017). The aim of the task is to fit a sphere onto a 3D point cloud with noise. In this context, we assume that the observations represent independent realizations of a random vector X, which is defined as

$$X = \mu + rWU,$$

where U is uniformly distributed on the unit sphere of $\mathbb{R}^3$, $W \sim \mathcal{U}([1-\delta, 1+\delta])$ with $\delta > 0$, W and U are independent. The radius $r > 0$ and the center $\mu \in \mathbb{R}^3$ are parameters to be estimated. The unknown parameter $\theta = (\mu, r)^T$ is a local minimizer of the function $G : \mathbb{R}^3 \times \mathbb{R}_+^* \longrightarrow \mathbb{R}$ defined for all $h = (a, b) \in \mathbb{R}^3 \times \mathbb{R}_+^*$ by:

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

In this scenario, one can use algorithm (14)-(15) to estimate θ with

$$\nabla_h g(X, h) = \begin{pmatrix} a - X + \frac{b(X-a)}{\|X-a\|} \\ b - \|X - a\| \end{pmatrix}.$$

Second-order methods USNA and UWASNA can also be used with

$$\nabla_h^2 g(X, h) = \begin{pmatrix} \left(1 - \frac{b}{\|X-a\|}\right) I_3 + \frac{b(X-a)(X-a)^T}{\|X-a\|^3} & \frac{X-a}{\|X-a\|} \\ \frac{(X-a)^T}{\|X-a\|} & 1 \end{pmatrix}.$$

We emphasize that in this specific case, neither the conventional Stochastic Newton Algorithms (SNA) nor the Weighted Averaged version (WASNA) using the Riccati formula are applicable due to the nature of the problem. Thus, we conduct simulations to compare the performance of the USNA, UWASNA, and ASGD. The synthetic datasets were generated with a sample size of 10 000 and the true parameters of the spherical distribution were set as follows: $\mu = (0, 0, 0)$ and $r = 2$. In addition, we set $\delta = 0.2$, which results in a Hessian matrix with eigenvalues of different order sizes. For all three algorithms, we initialize the estimate of the parameter by

$$\theta_{init} = (\mu_{init}, r_{init})^T = (0, 0, 0, 2)^T + e\epsilon$$

where $\epsilon \sim \mathcal{N}(0, I_4)$ and $e = 0.5$ or 1. Multiple iterations ($N = 100$) were performed to reduce the impact of sampling variations.

Figure 3 illustrates the comparison between the performances of USNA, UWASNA and ASGD in terms of the mean squared error (MSE) of the estimated parameters. Throughout the simulations, UWASNA amd USNA demonstrate superior performance in accurately estimating the parameters of the spherical distribution when compared to the gradient-based method. Additionally, the Hessian matrix H of the model can be explicitly calculated (Godichon-Baggioni and Portier, 2017), one has

$$H = \begin{pmatrix} I_3 - \frac{2}{3} I_3 \mathbb{E}\left[W\right]\left[W^{-1}\right] & 0 \\ 0 & 1 \end{pmatrix}.$$

Note that this matrix is diagonal, making its inverse computation straightforward. Therefore, we investigate the Frobenius norm of the difference between the estimated matrix A_n and the true matrix H^{-1}. From Figure 4, it's evident that our methods provide a good estimation of the inverse of the Hessian matrix. Moreover, UWASNA offers a better estimation compared to USNA.

5.3.2 p-means

Now we focus on the estimation of p-means of a multivariate distribution (Fréchet, 1948). We consider a random vector X of $\mathbb{R}^d$ with $d = 40$, where $X \sim \mathcal{N}(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma_{i,j}} = |\mathbf{i} - \mathbf{j}|^{\mathbf{0.5}}$. The p-mean m of X is defined as the minimizer of the functional $G_p : \mathbb{R}^d \to \mathbb{R}$ given for all $h \in \mathbb{R}^d$ by:

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

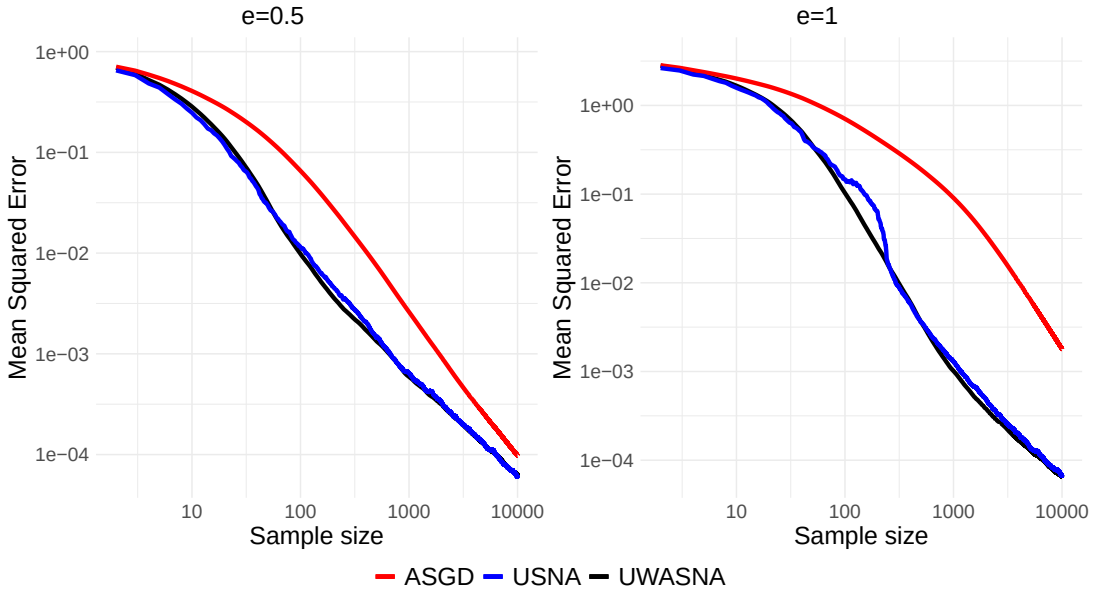

Figure 3: Evolution of the mean squared error with respect to the sample size for parameters estimation in a spherical Gaussian distribution.

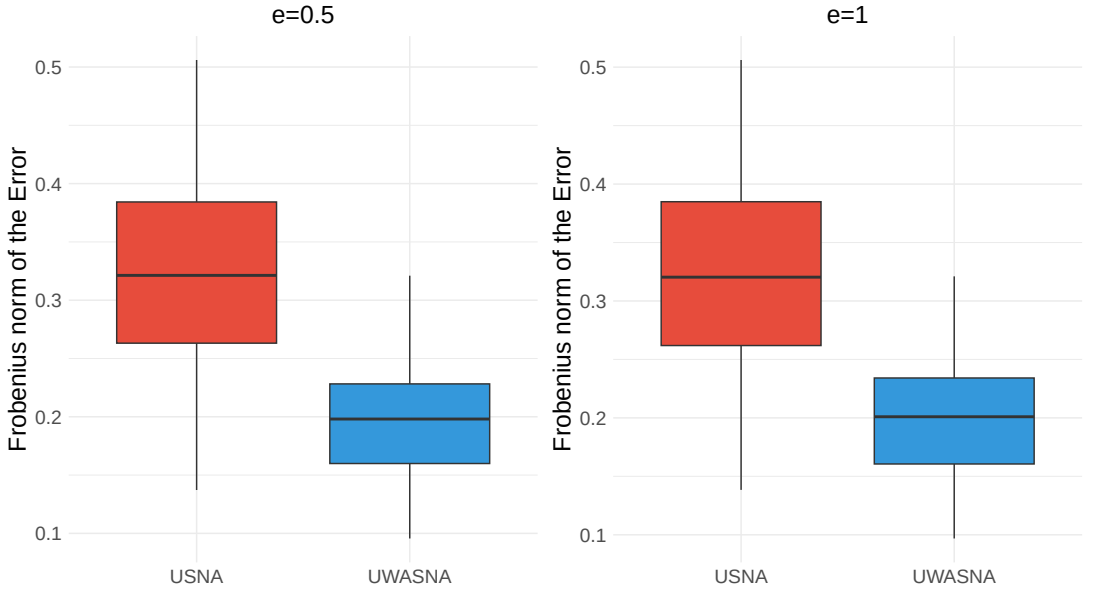

Figure 4: Frobenius norm of the difference between the estimated matrix A_n and the true matrix H^{-1}.

where $1 \leq p < +\infty$. Note that in our model $m = (0, \dots, 0)^T$. We can easily verify that the gradient and the Hessian of G_p are given by:

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

We aim to compare the performance of USNA, UWASNA, and ASGD for estimating the p-mean of X through simulations. We consider the case $p = 1.5$ and we simulate $N = 100$ independent samples of size $n = 10\,000$. For all three algorithms, we initialize the estimate with $m_{init} = e\epsilon$ where $\epsilon \sim \mathcal{N}(0, I_d)$ and $e = 1$ or 2.

As depicted in the Figure 5, we plotted the MSE versus sample size. It is evident that USNA and UWASNA consistently outperforms ASGD in estimating the p-means. Their superior performance can be attributed to the incorporation of information from the Hessian matrix. These results further emphasize

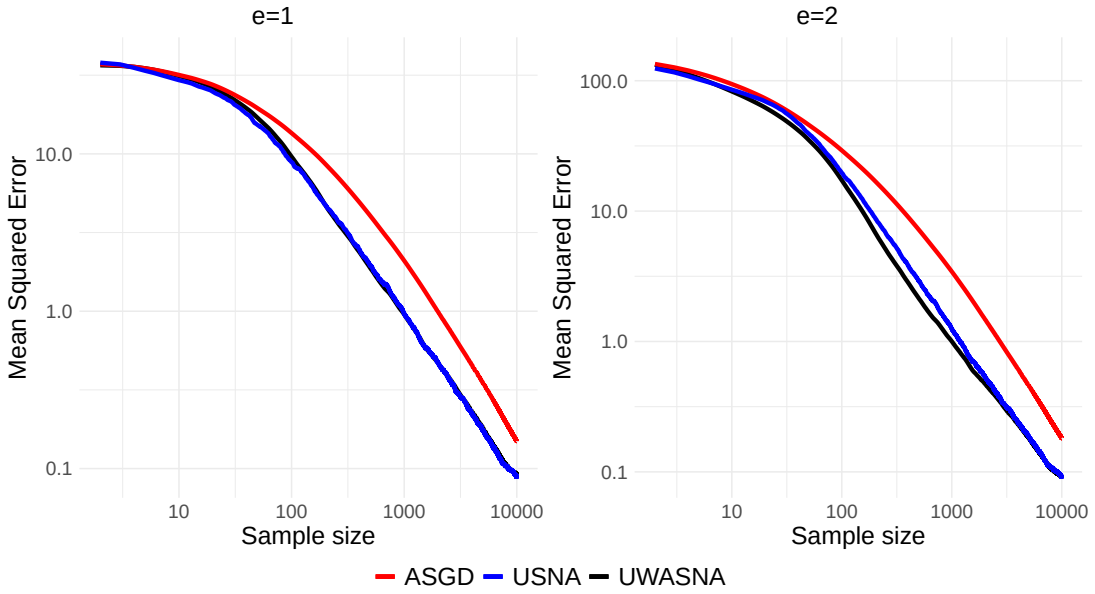

Figure 5: Evolution of the mean squared error with respect to the sample size for p-means estimation.

the advantage of using USNA and UWASNA as alternative methods when it is impossible to use the Riccati formula.

5.4 Application to real data

We apply the algorithms to "COVTYPE" dataset, a well-known dataset used for classification tasks (Blackard and Dean, 1999; Schmidt et al., 2017; Toulis et al., 2016). The original study was based on a dataset comprising 581,011 observations and 54 predictors. The primary objective was to predict the cover type of forests within Roosevelt National Park. In our current investigation, we will narrow our focus to the most prevalent category of the target variable, which is "Spruce/Fir," accounting for 49% of the observations. As a result, we convert the "covertype" variable into a binary form, with the "fir" category assigned a value of 1, while all other categories are assigned a value of 0. The objective is to use logistic regression to predict the variable "covertype". We split the dataset into training (80%) and test (20%) sets.

We implement the UWASNA, USNA, WASNA and SNA on the training set. As a baseline, we also apply a first-order algorithm ASGD on it. We evaluate the performance of each algorithm by calculating the accuracy of each one on both training set and test set. The results are summarized in Table 1.

	UWASNA	USNA	WASNA	SNA	ASGD
Training Accuracy(%)	**75.63**	75.31	75.52	75.38	74.54
Test Accuracy(%)	**75.61**	75.34	75.50	75.33	74.64

Table 1: Accuracy of UWASNA, USNA, WASNA, SNA and ASGD algorithms on "COVTYPE" dataset.

We observe that UWASNA, USNA, WASNA and SNA demonstrate similar performances, and they achieve higher accuracies on both training set and test set compared to ASGD. By successfully applying USNA and UWASNA on this dataset, we illustrate their practicality in real-world applications.

Conclusion

In this study, we thoroughly examined the stochastic optimization problem, primarily focusing on accurately estimating the unknown parameter. Our significant contribution is the introduction of a direct method to estimate the inverse of the Hessian matrix. Instead of the traditional approach, which estimates the Hessian matrix first, we directly addressed its inverse, using the Robbins-Monro procedure.

This approach led us to develop the Universal Weighted Averaged Stochastic Newton Algorithm. Through our extensive testing, we found that our newly proposed methods are both efficient and robust. When compared with standard first and second-order algorithms, our method consistently performed well. In certain scenarios, it even outperformed these conventional algorithms.

In summary, our findings emphasize the potential of second-order methods in optimization. Our approach to directly estimating the Hessian matrix's inverse represents a significant advancement, combining simplicity with efficiency.

6 Proofs

For the sake of simplicity, in the following we denote $T_n := \nabla_h^2 g(X_n, \hat{\theta}_{n-1}) Z_n Z_n^T$. Recall that $\|Z_n\|$ is bounded, i.e there is M such that $\|Z\| \leq M$. In addition, let us recall that if $\|Q_{n+1}\| \, \|Z_{n+1}\| \leq \beta_{n+1}$, it then follows that $\|T_{n+1}\|_{op} \leq \beta_{n+1}$.

6.1 Study on the largest eigenvalue of A_n

The following proposition provides an initial asymptotic bound of the largest eigenvalue of A_n without requiring knowledge on the behavior of the estimate $\hat{\theta}_n$. This result is crucial to prove Theorems 3.1 and 4.1.

Proposition 6.1. *Under Assumptions **(A3)** and **(A5)**, the largest eigenvalue of A_n denoted by $\lambda_{\max}(A_n)$ satisfies for all $\delta > 0$*

$$\lambda_{\max}(A_n) = o\left(n^{1-\gamma} \ln n^{1+\delta}\right) a.s. \quad \textit{and} \quad \lambda_{\max}(A_n, \tau) = o\left(n^{1-\gamma} \ln n^{1+\delta}\right) a.s.$$

Proof of Proposition 6.1. Define $W_n := A_{n-1} T_n^T + T_n A_{n-1}$. By definition of A_n,

$$\begin{aligned} \|A_{n+1}\|_F^2 = \|A_n\|_F^2 &- 2\gamma_{n+1} \langle A_n, W_{n+1} - 2I_d \rangle_F \mathbf{1}_{\|Q_{n+1}\| \|Z_{n+1}\| \leq \beta_{n+1}} \\ &+ \gamma_{n+1}^2 \|W_{n+1} - 2I_d\|_F^2 \mathbf{1}_{\|Q_{n+1}\| \|Z_{n+1}\| \leq \beta_{n+1}}. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[\|A_{n+1}\|_F^2 |\mathcal{F}_n\right] = \|A_n\|_F^2 &- 2\gamma_{n+1} \langle A_n, \nabla^2 G(\hat{\theta}_n) A_n + A_n \nabla^2 G(\hat{\theta}_n) - 2I_d \rangle_F \\ &+ \gamma_{n+1}^2 \mathbb{E}\left[\|W_{n+1} - 2I_d\|^2 \mathbf{1}_{\|Q_{n+1}\| \|Z_{n+1}\| \leq \beta_{n+1}} |\mathcal{F}_n\right] \\ &+ 2\gamma_{n+1} \mathbb{E}\left[\langle A_n, W_{n+1} - 2I_d \rangle_F \mathbf{1}_{\|Q_{n+1}\| \|Z_{n+1}\| > \beta_{n+1}} |\mathcal{F}_n\right]. \end{aligned}$$

Assumption **(A5)** ensures that

$$\begin{aligned} &\mathbb{E}\left[\|W_{n+1} - 2I_d\|_F^2 \mathbf{1}_{\|Q_{n+1}\| \|Z_{n+1}\| \leq \beta_{n+1}} |\mathcal{F}_n\right] \\ &\leq 8\left(\mathbb{E}\left[\left\|\nabla_h^2 g(X_{n+1}, \hat{\theta}_n)\right\|_F^2\right] \mathbb{E}\left[\|Z_{n+1}\|^4\right] \|A_n\|_F^2 + d\right) \\ &\leq 8\left(C_q^{2/q} \|A_n\|_F^2 + d\right). \end{aligned}$$

Since $\|Q_{n+1}\| = \|Q_{n+1}\|^q \|Q_{n+1}\|^{1-q}$ with $q > 1$, for any $\zeta_n > 0$,

$$\begin{aligned} &\gamma_{n+1} \mathbb{E}\left[\langle A_n, W_{n+1} - 2I_d \rangle_F \mathbf{1}_{\|Q_{n+1}\| \|Z_{n+1}\| > \beta_{n+1}} |\mathcal{F}_n\right] \\ &\leq 2\gamma_{n+1} \|A_n\|_F^2 \mathbb{E}\left[\|Q_{n+1}\| \, \|Z_{n+1}\| \mathbf{1}_{\|Q_{n+1}\| \|Z_{n+1}\| \geq \beta_{n+1}} |\mathcal{F}_n\right] \\ &+ \frac{\gamma_{n+1}}{\zeta_{n+1}} \|A_n\|_F^2 + \zeta_{n+1} d \\ &\leq \left(2C_q M^{2q} \gamma_{n+1} \beta_{n+1}^{1-q} + \frac{\gamma_{n+1}}{\zeta_{n+1}}\right) \|A_n\|_F^2 + \gamma_{n+1} \zeta_{n+1} d. \end{aligned}$$

Since $\nabla^2 G(\hat{\theta}_n)$ and A_n are positive, one has for any $\zeta_n > 0$,

$$-\gamma_{n+1} \langle A_n, \nabla^2 G(\hat{\theta}_n) A_n + A_n \nabla^2 G(\hat{\theta}_n) - 2I_d \rangle_F \leq \gamma_{n+1} \langle A_n, 2I_d \rangle \leq \frac{\gamma_{n+1}}{\zeta_{n+1}} \|A_n\|_F^2 + \gamma_{n+1}\zeta_{n+1} d.$$

Finally,

$$\mathbb{E}\left[\|A_{n+1}\|_F^2 \,|\mathcal{F}_n\right] \leq \left(1 + \frac{4\gamma_{n+1}}{\zeta_{n+1}} + 4C_q\gamma_{n+1}\beta_{n+1}^{1-q} + 8\gamma_{n+1}^2 C_q^{2/q}\right) \|A_n\|_F^2 + 4\gamma_{n+1}\zeta_{n+1} d + 8\gamma_{n+1}^2 d.$$

Setting $\zeta_n = n^{1-\gamma} \ln n^{1+\delta}$ with $\delta > 0$, we can apply Lemma 6.2 with $V_n = \|A_n\|_F^2$, and $a_n = \zeta_n^2$. One then obtains $\|A_n\|_F^2 = o\left(\zeta_n^2\right) = o\left(n^{2-2\gamma} \ln n^{2+2\delta}\right)$, so that $\lambda_{\max}(A_n) = o\left(n^{1-\gamma} \ln n^{1+\delta}\right)$ $a.s.$ □

6.2 Proof of Theorem 3.1

The aim is to provide an initial rate of convergence for A_n. A more refined or faster rate will be established later. First, note that

$$\begin{aligned}
A_{n+1} - H^{-1} &= A_n - H^{-1} - \gamma_{n+1}(W_{n+1} - 2I_d)\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\| \leq \beta_{n+1}} \\
&= A_n - H^{-1} - \gamma_{n+1}(W_{n+1} - 2I_d) + \gamma_{n+1}(W_{n+1} - 2I_d)\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\| > \beta_{n+1}} \\
&= A_n - H^{-1} - \gamma_{n+1}\left(\mathbb{E}\left[W_{n+1}|\mathcal{F}_n\right] - 2I_d\right) + \gamma_{n+1}\xi_{n+1} \\
&\quad + \gamma_{n+1}(W_{n+1} - 2I_d)\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\| > \beta_{n+1}} \\
&= A_n - H^{-1} - \gamma_{n+1}\left(\nabla^2 G(\hat{\theta}_n) A_n + A_n \nabla^2 G(\hat{\theta}_n) - 2I_d\right) + \gamma_{n+1}\xi_{n+1} + \\
&\quad \gamma_{n+1}(W_{n+1} - 2I_d)\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\| > \beta_{n+1}}
\end{aligned}$$

where $\xi_{n+1} := -W_{n+1} + \nabla^2 G(\hat{\theta}_n) A_n + A_n \nabla^2 G(\hat{\theta}_n)$. Let α_k^* be the function defined for all $h \in \mathcal{M}_p(\mathbb{R})$ by

$$\alpha_k^*(h) = (I_d - \gamma_{k+1} H)\, h\, (I_d - \gamma_{k+1} H),$$

we then have

$$\begin{aligned}
A_{n+1} - A &= \alpha_n^*\left(A_n - H^{-1}\right) - \gamma_{n+1}^2 H(A_n - H^{-1})H \\
&\quad + \gamma_{n+1} r_{1,n} + \gamma_{n+1} r_{2,n} + \gamma_{n+1} s_n + \gamma_{n+1} s_n' + \gamma_{n+1}\xi_{n+1}
\end{aligned} \tag{16}$$

where

$$\begin{aligned}
r_{1,n} &= (\nabla^2 G(\hat{\theta}_n) - H)H^{-1} + H^{-1}(\nabla^2 G(\hat{\theta}_n) - H), \\
r_{2,n} &= (W_{n+1} - 2I_d)\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\| > \beta_{n+1}} - \mathbb{E}\left[(W_{n+1} - 2I_d)\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\| > \beta_{n+1}} |\mathcal{F}_n\right], \\
s_n &= \mathbb{E}\left[\left(T_{n+1}\left(A_n - H^{-1}\right) + \left(A_n - H^{-1}\right) T_{n+1}\right)\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\| > \beta_{n+1}} |\mathcal{F}_n\right] \\
&\quad + \left(\nabla^2 G(\hat{\theta}_n) - H\right)\left(A_n - H^{-1}\right) + \left(A_n - H^{-1}\right)\left(\nabla^2 G(\hat{\theta}_n) - H\right), \\
s_n' &= \mathbb{E}\left[\left(T_{n+1}H^{-1} + H^{-1}T_{n+1} - 2I_d\right)\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\| > \beta_{n+1}} |\mathcal{F}_n\right]
\end{aligned}$$

By induction,

$$\begin{aligned}
A_n - H^{-1} &= \Psi_{n,0}^*\left(A_0 - H^{-1}\right) + \sum_{k=0}^{n-1} \Psi_{n,k+1}^* \gamma_{k+1}\xi_{k+1} + \sum_{k=0}^{n-1} \Psi_{n,k+1}^* \gamma_{k+1} r_{1,k} \\
&\quad + \sum_{k=0}^{n-1} \Psi_{n,k+1}^* \gamma_{k+1} r_{2,k} + \sum_{k=0}^{n-1} \Psi_{n,k+1}^* \gamma_{k+1} s_k + \sum_{k=0}^{n-1} \Psi_{n,k+1}^* \gamma_{k+1} s_k',
\end{aligned} \tag{17}$$

where $\Psi^*_{n,k}$ is the function defined for all $h \in \mathcal{M}_p(\mathbb{R})$ by :

$$\Psi^*_{n,k}(h) = \left(\prod_{j=k+1}^{n} \alpha_j^*\right)(h) = \left(\prod_{j=k+1}^{n} (I_d - \gamma_j H)\right) h \left(\prod_{j=k+1}^{n} (I_d - \gamma_j H)\right).$$

Next, we will determine the rate of convergence for each term on the right-hand side of equation (17).

Rate of convergence of $M_n := \sum_{k=0}^{n-1} \Psi^*_{n,k+1}\gamma_{k+1}\xi_{k+1}$: Recall that for all $\delta > 0$, $\lambda_{\max}(A_n) = o\left(n^{1-\gamma} \ln n^{1+\delta}\right)$ $a.s.$ Thus, there exists a constant $c' > 0$ such that $\mathbb{E}\left[\|\xi_{n+1}\|_F^2 |\mathcal{F}_n\right] \leq c'\left(1 + \|A_n\|_F^2\right) = o\left(n^{2-2\gamma} \ln n^{2+2\delta}\right)$ a.s., and according to Lemma 6.3, it follows that

$$\left\|\sum_{k=0}^{n-1} \Psi^*_{n,k+1}\gamma_{k+1}\xi_{k+1}\right\|_F^2 = \mathcal{O}\left(n^{2-3\gamma} \ln n^{2+2\delta}\right) \quad a.s. \tag{18}$$

Rate of convergence of $R_{1,n} := \sum_{k=0}^{n-1} \Psi^*_{n,k+1}\gamma_{k+1}r_{1,k}$: we have

$$R_{1,n+1} = (I_d - \gamma_{n+1}H) R_{1,n} (I_d - \gamma_{n+1}H) + \gamma_{n+1} r_{1,n+1}.$$

Therefore, for n large enough

$$\begin{aligned}\|R_{1,n+1}\|_{op} &\leq \|I_d - \gamma_{n+1}H\|_{op}^2 \|R_{1,n}\|_{op} + \gamma_{n+1} \|r_{1,n+1}\|_{op} \\ &\leq (1 - \lambda_{\min}(H)\gamma_{n+1})^2 \|R_{1,n}\|_{op} + \gamma_{n+1} \|r_{1,n+1}\|_{op} \, a.s.\end{aligned}$$

By Assumption **(A4)**, $\|r_{1,n+1}\|_{op} = o\left(n^{-a/2} \ln n^{(1+\delta)/2}\right)$ $a.s.$ According to Lemma 6.4,

$$\|R_{1,n+1}\|_{op} = o\left(n^{-a/2} \ln n^{(1+\delta)/2}\right) \quad a.s. \tag{19}$$

Rate of convergence of $R_{2,n} := \sum_{k=0}^{n-1} \Psi^*_{n,k+1}\gamma_{k+1}r_{2,k}$:

$$\begin{aligned}&\mathbb{E}\left[\left\|(W_{n+1} - 2I_d)\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\| \geq \beta_{n+1}}\right\|_{op}^2 |\mathcal{F}_n\right] \\ \leq &12\mathbb{E}\left[\left(\|Q_{n+1}\|^2 \|Z_{n+1}\|^2 \|A_n\|^2 + d\right) \mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\| \geq \beta_{n+1}} |\mathcal{F}_n\right].\end{aligned}$$

Applying Markov's inequality, and since $\|Z_n\|$ is supposed to be bounded by M,

$$\begin{aligned}\mathbb{E}\left[\left\|(W_{n+1} - 2I_d)\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\| \geq \beta_{n+1}}\right\|^2 |\mathcal{F}_n\right] &\leq 12d\mathbb{E}\left[\left\|\nabla^2 g(X_{n+1}, \hat{\theta}_n)\right\|_{op}^q \left\|Z_n Z_n^T\right\|^q |\mathcal{F}_n\right] \beta_n^{-q} \\ &+ 12 \|A_n\|^2 \mathbb{E}\left[\|Q_n\|^2 M^2 \mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\| \geq \beta_{n+1}}\right] \\ &\leq 12dM^{2q}\mathbb{E}\left[\left\|\nabla^2 g(X_{n+1}, \hat{\theta}_n)\right\|_{op}^q |\mathcal{F}_n\right] \beta_n^{-q} \\ &+ 12C_q \|A_n\|^2 \beta_{n+1}^{2-q}\end{aligned}$$

Thus, by Lemma 6.4,

$$\|R_{2,n}\|_{op}^2 = \mathcal{O}\left(n^{2-3\gamma} \ln n^{2+2\delta} n^{2\beta-q\beta}\right) \quad a.s. \tag{20}$$

Note that the rate of convergence for $R_{2,n}$ is faster than that of M_n.

Rate of convergence of $S'_n := \sum_{k=0}^{n-1} \Psi^*_{n,k+1}\gamma_{k+1}s'_k$: We have similarly

$$\begin{aligned}\left\|s'_n\right\| &\leq \mathbb{E}\left[\left\|\left(T_{n+1}H^{-1} + H^{-1}T_{n+1} - 2I_d\right)\right\|_F \mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\| \geq \beta_{n+1}} |\mathcal{F}_n\right] \\ &\leq 2\left\|H^{-1}\right\|_F \mathbb{E}\left[\|Z_{n+1}\|^2 \left\|\nabla_h^2 g\left(X_{n+1}, \hat{\theta}_n\right)\right\|_F \mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\| \geq \beta_{n+1}} |\mathcal{F}_n\right] \\ &+ 2q\mathbb{E}\left[\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\| \geq \beta_{n+1}} |\mathcal{F}_n\right] = \mathcal{O}\left(\beta_n^{1-q}\right).\end{aligned}$$

Therefore, thanks to Lemma 6.4,

$$S'_n = \mathcal{O}\left(\beta_n^{1-q}\right). \tag{21}$$

A first result for A_n: Thanks to equalities (18) to (21), one can rewrite $A_n - H^{-1}$ as

$$A_n - H^{-1} = \sum_{k=0}^{n-1} \Psi^*_{n,k+1}\gamma_{k+1}s_k + \widetilde{S}_n$$

with $\left\|\widetilde{S}_n\right\|_F = o\left(n^{\max\left(1-\frac{3}{2}\gamma,-a/2,\beta(1-q)\right)}\ln n^{1+\delta}\right)$. In addition, we have

$$\begin{aligned}\|s_n\|_{op} &\leq \left\|A_n - H^{-1}\right\|_{op}\mathbb{E}\left[\|Q_{n+1}\|\,\|Z_{n+1}\|\,\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\geq\beta_n}\|\mathcal{F}_n\right] \\ &\quad + \left\|\nabla^2 G(\hat{\theta}_n) - H\right\|\left\|A_n - H^{-1}\right\|_{op} \\ &= o\left(\left\|A_n - H^{-1}\right\|_{op}\right) \quad a.s.\end{aligned}$$

Define $S_n := \sum_{k=0}^{n-1}\Psi^*_{n,k+1}\gamma_{k+1}s_k$. There exists a positive sequence $(\widetilde{r}_n)_{n\geq 0}$ with $\widetilde{r}_n \xrightarrow[n\to\infty]{a.s.} 0$ such that

$$\begin{aligned}\|S_{n+1}\|_{op} &\leq (1-\lambda_{\min}(H)\gamma_{n+1})^2\|S_n\|_{op} + \gamma_{n+1}\widetilde{r}_n\left\|A_n - H^{-1}\right\|_{op} \\ &\leq (1-\lambda_{\min}(H)\gamma_{n+1})^2\|S_n\|_{op} + \gamma_{n+1}\widetilde{r}_n\left(\|S_n\|_{op} + \left\|\widetilde{S}_n\right\|_{op}\right).\end{aligned}$$

Applying Lemma 6.4,

$$\|S_n\|_{op} = o\left(n^{\max\left(1-\frac{3}{2}\gamma,-a/2,\beta(1-q)\right)}\ln n^{1+\delta}\right),$$

which implies that

$$\left\|A_n - H^{-1}\right\|_F = o\left(n^{\max\left(1-\frac{3}{2}\gamma,-a/2,\beta(1-q)\right)}\ln n^{1+\delta}\right) a.s.$$

Final rate of convergence of A_n: The aim here is, with the help of this first result on A_n, to give better rates of convergence for M_n and $R_{2,n}$. First, note that if $\gamma > 2/3$, then $1-\frac{3}{2}\gamma < 0$ and we have directly $A_n \xrightarrow[n\to\infty]{a.s.} H^{-1}$. If $\gamma \leq 2/3$, we have $1-\frac{3}{2}\gamma > 0$, so that $\max\{1-\frac{3}{2}\gamma, -a/2, \beta(1-q)\} = \max\{1-\frac{3}{2}\gamma, \beta(1-q)\}$ and $\|A_n\| = o\left(n^{\max\{1-\frac{3}{2}\gamma,\beta(1-q)\}}\ln n^{1+\delta}\right) a.s.$. Thus, applying Lemma 6.3 in the worse case (i.e when $\gamma > 2/3$), one now has

$$\|M_n\|_F^2 = o\left(n^{\max\{2-4\gamma,\beta(2-2q)-\gamma\}}\ln n^{2+2\delta}\right) \quad a.s.$$

and

$$\|R_{2,n}\|_{op}^2 = \mathcal{O}\left(n^{2-4\gamma}\ln n^{2+2\delta}n^{2\beta-q\beta}\right) \quad a.s.$$

Following the same process as before, we now have that for any $\gamma \in (1/2, 1)$, A_n converges almost surely to H^{-1}, and in particular, one has $\|A_n\| = \mathcal{O}(1)$. Applying another time Lemma 6.3, we have

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

and

$$\|R_{2,n}\|_{op}^2 = \mathcal{O}\left(n^{-\gamma}\ln n^{2+2\delta}n^{2\beta-q\beta}\right) \quad a.s.$$

Finally, we have

$$\left\|A_n - H^{-1}\right\|_F^2 = o\left(\frac{\ln n^{1+\delta}}{n^{\min\{\gamma,a,2\beta(q-1)\}}}\right) a.s. \tag{22}$$

Rate of convergence of $A_{n,\tau}$. Decomposition (16) can be written as

$$H(A_n - H^{-1}) + (A_n - H^{-1})H = \frac{A_n - H^{-1} - (A_{n+1} - H^{-1})}{\gamma_{n+1}} + r_{1,n} + r_{2,n} + s_n + s'_n + \xi_{n+1},$$

so that

$$H(A_{n,\tau}-H^{-1})+(A_{n,\tau}-H^{-1})H = t_n\sum_{k=0}^{n}u_k\left((A_k-H^{-1})-(A_{k+1}-H^{-1})\right) + t_n\sum_{k=0}^{n}u_k\gamma_{k+1}(r_{1,n}+r_{2,n}+s_n+s'_n+\xi_{n+1}) \tag{23}$$

where $t_n=\frac{1}{\sum_{k=0}^{n}\log(k+1)^\tau}$ and $u_{k+1}=\frac{\ln(k+1)^\tau}{\gamma_{k+1}}$. Our objective is to determine the rate of convergence for each term on the right-hand side of the decomposition given in (23).

Rate of convergence of $t_n\sum_{k=0}^{n}u_{k+1}\left((A_k-H^{-1})-(A_{k+1}-H^{-1})\right)$: with the help of an Abel's transform,

$$t_n\sum_{k=0}^{n}u_{k+1}\left((A_k-H^{-1})-(A_{k+1}-H^{-1})\right) = -t_n\left(A_{n+1}-H^{-1}\right)u_{n+1}+t_n\left(A_0-H^{-1}\right)u_1 + t_n\sum_{k=1}^{n}(A_k-H^{-1})(u_{k+1}-u_k).$$

It is obvious that $t_n\left(A_0-H^{-1}\right)u_1$ is negligible while thanks to equality (22),

$$t_n\left\|A_{n+1}-H^{-1}\right\|u_{n+1}=o\left(\frac{\ln n^{1+\delta}}{n^{1-\gamma+\frac{1}{2}\min\{\gamma,a,2\beta(q-1)\}}}\right)\quad a.s$$

In addition, by the mean value theorem, $|u_{k+1}-u_k|\leq 2k^{\gamma-1}\max\{1,\tau\}\ln(k+1)^\tau$, which implies that

$$\left\|\sum_{k=1}^{n}(A_k-H^{-1})(u_{k+1}-u_k)\right\|_F\leq\sum_{k=1}^{n}\left\|A_k-H^{-1}\right\|_F 2k^{\gamma-1}\max\{1,\tau\}\ln(k+1)^\tau$$

and with the help of equation (22)

$$\left\|t_n\sum_{k=0}^{n}u_k\left((A_k-H^{-1})-(A_{k+1}-H^{-1})\right)\right\|_F=o\left(\frac{\ln n^{1+\delta}}{n^{1-\gamma+\frac{1}{2}\min\{\gamma,a,2\beta(q-1)\}}}\right)\quad a.s$$

Rate of convergence of $t_n\sum_{k=0}^{n}u_{k+1}\gamma_{k+1}\xi_{k+1}$. Remark that there exists $c'>0$ such that for all $\delta>0$

$$\mathbb{E}\left[\|\xi_{n+1}\|_F^2\,\|\mathcal{F}_n\right]\leq c'\left(1+\|A_n\|_F^2\right)\leq c'\left(1+2\left\|A_n-H^{-1}\right\|_F^2+2\left\|H^{-1}\right\|_F^2\right).$$

By applying a law of large numbers for martingales, one obtains for all $\delta>0$

$$\left\|t_n\sum_{k=0}^{n}u_{k+1}\gamma_{k+1}\xi_{k+1}\right\|_F^2=o\left(\frac{\ln n^{1+\delta}}{n}\right)a.s.$$

Rate of convergence of $t_n\sum_{k=0}^{n}u_{k+1}\gamma_{k+1}r_{1,k}$. First, let us give a useful equality: consider $\sum_{k=0}^{n}k^{-\tilde{a}/2}\ln k^{(1+\delta)/2}$ with $\tilde{a}>0$, since $\int_1^n x^{-\tilde{a}/2}\ln x^{(1+\delta)/2}dx\leq c(\ln n)^{(1+\delta)/2}x^{1-\tilde{a}}$ if $\tilde{a}\neq 2$ and $\int_1^n x^{-\tilde{a}/2}\ln x^{(1+\delta)/2}dx\leq c'(\ln n)^{(3+\delta)/2}$ if $\tilde{a}=2$ then one can verify that

$$\left\|t_n\sum_{k=0}^{n}u_k\gamma_{k+1}k^{-\tilde{a}/2}\ln k^{(1+\delta)/2}\right\|_F^2=o\left(\frac{\ln n^{1+\delta+2\mathbf{1}_{\tilde{a}=2}}}{n^{\min\{2,\tilde{a}\}}}\right)a.s. \tag{24}$$

Then, recalling that $\|r_{1,k+1}\|_{op}=o\left(n^{-a/2}\ln k^{(1+\delta)/2}\right)$

$$\left\|t_n\sum_{k=0}^{n}u_k\gamma_{k+1}r_{1,k}\right\|_F^2=o\left(\frac{\ln n^{1+\delta+2\mathbf{1}_{a=2}}}{n^{\min\{2,a\}}}\right)a.s.$$

Rate of convergence of $t_n \sum_{k=0}^n u_{k+1}\gamma_{k+1}s'_k$. Recalling that $s'_k = O\left(\beta_k^{1-q}\right)$, it comes

$$\left\| t_n \sum_{k=0}^{n} u_k \gamma_{k+1} s'_k \right\|_F^2 = o\left(\frac{\ln n^{2\mathbf{1}_{2\beta(q-1)=2}}}{n^{\min\{2,2\beta(q-1)\}}} \right) \quad a.s.$$

Rate of convergence of $t_n \sum_{k=0}^n u_{k+1}\gamma_{k+1}s_k$. Recall that

$$\begin{aligned}\|s_n\|_{op} \leq \left\|A_n - H^{-1}\right\|_{op} \mathbb{E}\left[\|Q_{n+1}\|\,\|Z_{n+1}\|\,\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\geq\beta_n}|\mathcal{F}_n\right] \\ + \left\|\nabla^2 G(\hat{\theta}_n) - H\right\| \left\|A_n - H^{-1}\right\|_{op},\end{aligned}$$

Recalling that $\mathbb{E}\left[\|T_{n+1}\|_{op}\,\mathbf{1}_{\|T_{n+1}\|\geq\beta_n}|\mathcal{F}_n\right] = O\left(\beta_n^{1-q}\right)$, and with the help of equality (22)

$$\begin{aligned}&\left\|A_n - H^{-1}\right\|_{op} \mathbb{E}\left[\|Q_{n+1}\|\,\|Z_{n+1}\|\,\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\geq\beta_n}|\mathcal{F}_n\right] \\ &= o\left(\frac{\ln n^{(1+\delta)/2}}{n^{\beta(q-1)+\frac{1}{2}\min\{a,\gamma,2\beta(q-1)\}}} \right) \quad a.s\end{aligned}$$

In addition, one has

$$\left\|\nabla^2 G(\hat{\theta}_n) - H\right\| \left\|A_n - H^{-1}\right\|_{op} = o\left(\frac{\ln n^{1+\delta}}{n^{a/2+\frac{1}{2}\min\{a,\gamma,2\beta(q-1)\}}} \right) \quad a.s.$$

leading to

$$\|s_n\| = o\left(\frac{\ln n^{1+\delta}}{n^{\min\{a/2,\beta(q-1)\}+\frac{1}{2}\min\{a,\gamma,2\beta(q-1)\}}} \right) \quad a.s.$$

Then, since $\min\{a/2, \beta(q-1)\} + \frac{1}{2}\min\{a,\gamma,2\beta(q-1)\} \geq \min\{a,\gamma,2\beta(q-1)\}$,

$$\left\| t_n \sum_{k=0}^{n} u_k \gamma_{k+1} s_k \right\|_F^2 = o\left(\frac{\ln n^{1+\delta}}{n^{\min\{2,2\min\{a,\gamma,2\beta(q-1)\}\}}} \right) a.s.$$

Rate of convergence of $t_n \sum_{k=0}^n u_{k+1}\gamma_{k+1}r_{2,k}$. Let us recall that $r_{2,n}$ is a martingale difference, and with the help of a law of large numbers for martingales it comes that

$$\left\| t_n \sum_{k=0}^{n} u_{k+1} \gamma_{k+1} r_{2,k} \right\|_F^2 = o\left(\frac{\ln n^{1+\delta}}{n} \right) \quad a.s.$$

Rate of convergence of $A_{n,\tau}$. We so have

$$\left\|H(A_{n,\tau} - H^{-1}) + (A_{n,\tau} - H^{-1})H\right\|^2 = o\left(\frac{\ln n^{1+\delta}}{n^{\min\{1,a,2\beta(q-1)\}}} \right) a.s.$$

Note that H is diagonalisable, and its eigenvalues are denoted by $\lambda_1, ..., \lambda_d$. Thus, for any matrix $B \in \mathcal{M}_d$, we have

$$\begin{aligned}\|HB + BH\|_F^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} (HB + BH)_{i,j}^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} (\lambda_i + \lambda_j)^2 B_{i,j}^2 \\ &\geq 4\lambda_{\min}(H)^2 \sum_{i=1}^{n}\sum_{j=1}^{n} B_{i,j}^2 = 4\lambda_{\min}(H)^2 \|B\|_F^2 .\end{aligned}$$

$$\begin{aligned}\left\|A_{n,\tau} - H^{-1}\right\|^2 &\leq \frac{1}{4\lambda_{\min}(H)^2} \left\|H(A_{n,\tau} - H^{-1}) + (A_{n,\tau} - H^{-1})H\right\|^2 \\ &= o\left(\frac{\ln n^{1+\delta}}{n^{\min\{1,a,2\beta(q-1)\}}} \right) a.s.\end{aligned}$$

6.3 Proof of Theorem 4.1

In order to prove this theorem, we first show that the eigenvalues of A_n and $A_{n,\tau}$ are well controlled, which implies the consistency of θ_n. Then, we give an initial rate of convergence of θ_n, which allows us to obtain the rate of convergence of A_n and $A_{n,\tau}$. The last step is to deduce the rate of convergence of θ_n and $\theta_{n,\tau}$. We first study the smallest eigenvalue of A_n.

Proposition 6.2. *Under Assumptions **(A3)** and **(A5)**, the smallest eigenvalue of A_n denoted by $\lambda_{\min}(A_n)$ satisfies*

$$\lambda_{\min}(A_n)^{-1} = \mathcal{O}\left(n^{\beta}\right) a.s. \qquad \text{and} \qquad \lambda_{\min}(A_{n,\tau})^{-1} = \mathcal{O}\left(n^{\beta}\right) a.s.$$

Proof of Proposition 6.2. First of all, note that for all $n \geq 0$, $\lambda_{\min}(A_n) > 0$. According to the definition of A_n, one has

$$\begin{aligned}
\lambda_{\min}(A_{n+1}) &= \lambda_{\min}\left(A_n - \gamma_{n+1}\left(T_{n+1}A_n + A_nT_{n+1}^T - 2I_d\right)\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\leq\beta_{n+1}}\right) \\
&\geq \lambda_{\min}\left(I_d - \gamma_{n+1}T_{n+1}\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\leq\beta_{n+1}}\right)A_n \\
&\quad \times \left(I_d - \gamma_{n+1}T_{n+1}^T\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\leq\beta_{n+1}}\right) \\
&\quad + \left(\gamma_{n+1}I_d - \gamma_{n+1}^2T_{n+1}A_nT_{n+1}^T\right)\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\leq\beta_{n+1}}\Big) \\
&\geq \lambda_{\min}\left(I_d - \gamma_{n+1}T_{n+1}\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\leq\beta_{n+1}}\right)A_n \\
&\quad \times \left(I_d - \gamma_{n+1}T_{n+1}^T\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\leq\beta_{n+1}}\right) \\
&\quad + \gamma_{n+1} - \gamma_{n+1}^2\beta_{n+1}^2\lambda_{\max}(A_n) - \gamma_{n+1}\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\geq\beta_{n+1}}
\end{aligned}$$

For all $h \in \mathbb{R}^p$, one has

$$\begin{aligned}
&h^t\left(I_d - \gamma_{n+1}T_{n+1}\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\leq\beta_{n+1}}\right)A_n\left(I_d - \gamma_{n+1}T_{n+1}^T\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\leq\beta_{n+1}}\right)h \\
&\geq \left\|A_n^{1/2}\left(I_d - \gamma_{n+1}T_{n+1}\mathbf{1}_{\|\|Q_{n+1}\|\|Z_{n+1}\|\leq\beta_{n+1}\|}\right)h\right\|^2 \\
&\geq \lambda_{\min}(A_n)\left\|\left(I_d - \gamma_{n+1}T_{n+1}\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\leq\beta_{n+1}}\right)h\right\|^2 \\
&\geq \lambda_{\min}\left(A_n\right)\left(1 - \gamma_{n+1}\beta_{n+1}\right)^2\|h\|^2
\end{aligned}$$

Thus,

$$\begin{aligned}
&\lambda_{\min}\left(\left(I_d - \gamma_{n+1}T_{n+1}\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\leq\beta_{n+1}}\right)A_n\left(I_d - \gamma_{n+1}T_{n+1}^T\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\leq\beta_{n+1}}\right)\right) \\
&\quad \geq \lambda_{\min}\left(A_n\right)\left(1 - \gamma_{n+1}\beta_{n+1}\right)^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lambda_{\min}(A_{n+1}) \geq \left(1 - \gamma_{n+1}\beta_{n+1}\right)^2\lambda_{\min}\left(A_n\right) + \gamma_{n+1} - \gamma_{n+1}^2\beta_{n+1}^2\lambda_{\max}(A_n) \\
- \gamma_{n+1}\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|>\beta_{n+1}}.
\end{aligned}$$

Note that $(\beta_n\gamma_n)_n$ is a decreasing sequence and $\beta_n\gamma_n \leq 1/2$. Let U_n and V_n be sequences defined as

$$U_n := \sum_{k=1}^{n}\prod_{j=k+1}^{n}\left(1 - \gamma_k\beta_k\right)^2\gamma_{k+1}^2\beta_{k+1}^2\lambda_{\max}(A_n),$$

and

$$V_n := \sum_{k=0+1}^{n}\prod_{j=k+1}^{n}\left(1 - \gamma_k\beta_k\right)^2\gamma_{k+1}\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\geq\beta_{k+1}}.$$

One can prove by induction that

$$\lambda_{\min}(A_n) \geq \lambda_{\min}\left(A_0\right)\prod_{k=1}^{n}\left(1 - \gamma_k\beta_k\right)^2 + \sum_{k=1}^{n}\prod_{j=k+1}^{n}\left(1 - \gamma_j\beta_j\right)^2\gamma_k - U_n - V_n,$$

with the convention $\prod_{j=n+1}^{n}\left(1-\gamma_j\beta_j\right)^2=1$. One has

$$U_{n+1}=\left(1-\gamma_{n+1}\beta_{n+1}\right)^2U_n+\gamma_{n+1}^2\beta_{n+1}^2\lambda_{\max}(A_n),$$

since $\lambda_{\max}(A_n)=o\left(\ln n^{1+\delta}n^{1-\gamma}\right)\,a.s.$,

$$U_n=o\left(\ln n^{1+\delta}\beta_{n+1}n^{1-\gamma}\gamma_{n+1}\right)=o\left(\frac{1}{\beta_n}\right)$$

since $\beta<\gamma-1/2$. One also has

$$V_{n+1}=\left(1-\gamma_{n+1}\beta_{n+1}\right)^2V_n+\gamma_{n+1}\mathbf{1}_{\|Q_{n+1}\|\|Z_{n+1}\|\geq\beta_{n+1}}.$$

We define $V_n':=n^v\ln(n+1)^{-(1+\delta)}V_n$ for some $v>0$, then, for n large enough,

$$\begin{aligned}\mathbb{E}\left[V_{n+1}'|\mathcal{F}_n\right]&\leq\frac{\ln n^{1+\delta}}{\ln(n+1)^{1+\delta}}\frac{(n+1)^v}{n^v}\left(1-\gamma_{n+1}\beta_{n+1}\right)^2V_n'+\frac{\gamma_{n+1}(n+1)^vC_q}{\ln(n+1)^{1+\delta}\beta_n^q}M^{2q}\\&\leq V_n'+\frac{\gamma_{n+1}(n+1)^vC_q}{\ln(n+1)^{1+\delta}\beta_n^q}M^{2q}.\end{aligned}$$

In order to apply the Robbins-Siegmund theorem, let us take $v=\gamma+q\beta-1$, then V_n' converges almost surely to a finite random variable for all δ, which can be translated by

$$V_n=o\left(\frac{\ln(n+1)^{1+\delta}}{n^{\gamma+q\beta-1}}\right)\quad a.s$$

which is negligible as soon as $\beta>\frac{1-\gamma}{q-1}$. It is obvious that $\lambda_{\min}\left(A_0\right)\prod_{k=1}^{n}\left(1-\gamma_k\beta_k\right)^2\geq0$. Finally

$$\begin{aligned}\sum_{k=1}^{n}\prod_{j=k+1}^{n}\left(1-\gamma_j\beta_j\right)^2\gamma_k&\geq\sum_{k=1}^{n}\frac{1}{2\beta_k}\prod_{j=k+1}^{n}\left(1-\gamma_j\beta_j\right)^2 2\gamma_k\beta_k\\&\geq\sum_{k=1}^{n}\frac{1}{2\beta_k}\prod_{j=k+1}^{n}\left(1-\gamma_j\beta_j\right)^2\left(2\gamma_k\beta_k-\beta_k^2\gamma_k^2\right)\\&=\sum_{k=1}^{n}\frac{1}{2\beta_k}\left(\prod_{j=k+1}^{n}\left(1-\gamma_j\beta_j\right)^2-\prod_{j=k}^{n}\left(1-\gamma_j\beta_j\right)^2\right)\end{aligned}$$

Since $\left(\frac{1}{\beta_n}\right)_n$ is a decreasing sequence, one has

$$\begin{aligned}\sum_{k=1}^{n}\prod_{j=k+1}^{n}\left(1-\gamma_j\beta_j\right)^2\gamma_k&\geq\frac{1}{2\beta_n}\sum_{k=1}^{n}\left(\prod_{j=k+1}^{n}\left(1-\gamma_j\beta_j\right)^2-\prod_{j=k}^{n}\left(1-\gamma_j\beta_j\right)^2\right)\\&\geq\frac{1}{2\beta_n}\left(1-\prod_{j=1}^{n}\left(1-\gamma_j\beta_j\right)^2\right)\\&\geq\frac{1-\left(1-\gamma_1\beta_1\right)^2}{2\beta_n}.\end{aligned}$$

Thus, $\lambda_{\min}(A_n)\geq\frac{c_1}{\beta_n}+o\left(\frac{1}{\beta_n}\right)$ a.s with $c_1=1-\left(1-\gamma_1\beta_1\right)^2/2$, which implies that

$$\frac{1}{\lambda_{\min}(A_n)}=O\left(\beta_n\right)\quad a.s.$$

□

We now prove the convergence of θ_n. According to Taylor's theorem, and using Assumption **(A2)**, we obtain

$$\begin{aligned} G(\theta_{n+1}) &= G(\theta_n) + \nabla G(\theta_n)^T(\theta_{n+1}-\theta_n) \\ &\quad + \frac{1}{2}(\theta_{n+1}-\theta_n)^T \int_0^1 \nabla^2 G(\theta_{n+1} + t(\theta_n - \theta_{n+1}))dt(\theta_{n+1}-\theta_n) \\ &\leq G(\theta_n) + \nabla G(\theta_n)^T(\theta_{n+1}-\theta_n) + \frac{L_{\nabla G}}{2}\left\|\theta_{n+1}-\theta_n\right\|^2. \end{aligned}$$

From the definition of θ_n (6),

$$G(\theta_{n+1}) \leq G(\theta_n) - \nu_{n+1}\nabla G(\theta_n)^T A_{n,\tau}\nabla g(X_{n+1},\theta_n) + \frac{L_{\nabla G}}{2}\nu_{n+1}^2\left\|A_{n,\tau}\right\|_{op}^2\left\|\nabla g(X_{n+1},\theta_n)\right\|^2.$$

Letting $K_n := G(\theta_n) - G(\theta)$, we can rewrite the inequality as

$$K_{n+1} \leq K_n - \nu_{n+1}\nabla G(\theta_n)^T A_{n,\tau}\nabla^2 g(X_{n+1},\theta_n) + \frac{L_{\nabla G}}{2}\nu_{n+1}^2\left\|A_{n,\tau}\right\|_{op}^2\left\|\nabla g(X_{n+1},\theta_n)\right\|^2,$$

which implies

$$\begin{aligned} \mathbb{E}\left[K_{n+1}|\mathcal{F}_n\right] &\leq K_n - \nu_{n+1}\nabla G(\theta_n)^T A_{n,\tau}\nabla G\left(\theta_n\right) \\ &\quad + \frac{L_{\nabla G}}{2}\nu_{n+1}^2\left\|A_{n,\tau}\right\|_{op}^2\mathbb{E}\left[\left\|\nabla g(X_{n+1},\theta_n)\right\|^2|\mathcal{F}_n\right]. \end{aligned}$$

According to Assumption **(A1)**, we have

$$\begin{aligned} \mathbb{E}\left[K_{n+1}|\mathcal{F}_n\right] &\leq \left(1 + \frac{CL_{\nabla G}}{2}\nu_{n+1}^2\left\|A_{n,\tau}\right\|_{op}^2\right)K_n - \nu_{n+1}\lambda_{\min}(A_{n,\tau})\left\|\nabla G(\theta_n)\right\|^2 \\ &\quad + \frac{CL_{\nabla G}}{2}\nu_{n+1}^2\left\|A_{n,\tau}\right\|_{op}^2. \end{aligned}$$

With the help of Proposition 6.1,

$$\sum_{n=0}^{\infty}\nu_{n+1}^2\left\|A_{n,\tau}\right\|_{op}^2 < +\infty \quad a.s.$$

Subsequently, according to Robbins-Siegmund Theorem, K_n is guaranteed to converge almost surely to a finite random variable, and

$$\sum_{n=0}^{\infty}\nu_{n+1}\lambda_{\min}(A_{n,\tau})\left\|\nabla G(\theta_n)\right\|^2 < +\infty \quad a.s.$$

With the help of Proposition 6.2,

$$\sum_{n=0}^{\infty}\nu_{n+1}\lambda_{\min}(A_{n,\tau}) = +\infty \quad a.s.$$

It suggests that $\liminf_n \left\|\nabla G(\theta_n)\right\| = 0$ almost surely and $\liminf_n K_n = 0$ almost surely. As K_n converges almost surely to a random variable, $G(\theta_n)$ converges almost surely to $G(\theta)$. By local strict convexity,

$$\theta_n \xrightarrow[n\to\infty]{a.s.} \theta \quad \text{and} \quad \theta_{n,\tau'} \xrightarrow[n\to\infty]{a.s.} \theta.$$

6.4 Proof of Theorem 4.2

Rate of convergence of θ_n. Recall that

$$\begin{aligned} \mathbb{E}\left[G(\theta_{n+1}) - G(\theta)|\mathcal{F}_n\right] &\leq \left(1 + \frac{L_{\nabla G}C}{2}\nu_{n+1}^2\lambda_{\max}(A_{n,\tau})^2\right)(G(\theta_n) - G(\theta)) \\ &\quad - \nu_{n+1}\nabla G(\theta_n)^T A_{n,\tau}\nabla G(\theta_n) + \frac{L_{\nabla G}C}{2}\nu_{n+1}^2\lambda_{\max}(A_{n,\tau})^2. \end{aligned}$$

According to equation (6) in Godichon-Baggioni (2021), there is a positive constant c_0 such that

$$\nabla G(\theta_n)^T A_{n,\tau} \nabla G(\theta_n) \geq \lambda_{\min}(A_{n,\tau}) \|\nabla G(\theta_n)\|^2 \geq c_0 \lambda_{\min}(A_{n,\tau}) \left(G(\theta_n) - G(\theta)\right).$$

Recall that $2\gamma + 2\nu - 2 > 1$, so that there exists $\eta > 0$ such that $\eta < 2\gamma + 2\nu - 3$. We define $\widetilde{V}_n := n^\eta \left(G(\theta_n) - G(\theta)\right)$, so that

$$\begin{aligned}\mathbb{E}\left[\widetilde{V}_{n+1}|\mathcal{F}_n\right] = \left(\frac{n+1}{n}\right)^\eta \left(\left(1 + \frac{L_{\nabla GC}}{2}\nu_{n+1}^2 \lambda_{\max}(A_{n,\tau})^2\right) - c_0 \nu_{n+1} \lambda_{\min}(A_{n,\tau})\right) \widetilde{V}_n \\ + \frac{L_{\nabla GC}}{2}\nu_{n+1}^2 \lambda_{\max}(A_{n,\tau})^2 (n+1)^\eta.\end{aligned}$$

Let $\widetilde{U}_n := \left(\frac{n+1}{n}\right)^\eta \left(\left(1 + \frac{L_{\nabla GC}}{2}\nu_{n+1}^2 \lambda_{\max}(A_{n,\tau})^2\right) - c_0 \nu_{n+1}^2 \lambda_{\min}(A_{n,\tau})\right)$, then

$$\mathbb{E}\left[\widetilde{V}_{n+1}|\mathcal{F}_n\right] \leq \widetilde{V}_n + \frac{L_{\nabla GC}}{2}\nu_{n+1}^2 \lambda_{\max}(A_{n,\tau})^2 (n+1)^\eta + \widetilde{V}_n \mathbf{1}_{\widetilde{U}_n > 1}.$$

As $\nu + \beta < 1$, one can easily verify that $\mathbf{1}_{\widetilde{U}_n > 1} \xrightarrow{a.s.} 0$. We can now apply the Robbins-Siegmund theorem. We therefore have $\|G(\theta_{n+1}) - G(\theta)\| = o\left(n^{-\eta}\right)$ for all $\eta < 2\gamma + 2\nu - 3$. Thanks to the local strong convexity of G, this leads to $\|\theta_n - \theta\|^2 = o\left(n^{-\eta}\right)$ a.s. We are now able to apply Theorem 3.1 and one obtains

$$\|A_n - A\|^2 = o\left(\frac{\ln n^{1+\delta}}{n^{\min\{\gamma,\eta,2\beta(q-1)\}}}\right) a.s. \quad \text{and} \quad \|A_{n,\tau} - A\|^2 = o\left(\frac{\ln n^{1+\delta}}{n^{\min\{\eta,2\beta(q-1)\}}}\right) a.s.$$

According to Theorem 4.2 and Theorem 4.3 in Boyer and Godichon-Baggioni (2022), one so has

$$\|\theta_n - \theta\|^2 = o\left(\frac{\ln n^{1+\delta}}{n^\nu}\right) a.s. \tag{25}$$

Rate of convergence of $\theta_{n,\tau'}$. For all non-negative integers n,

$$\theta_{n+1} - \theta_n = -\nu_{n+1} A_{n,\tau} \nabla_h g(X_{n+1}, \theta_n),$$

so that

$$\frac{\theta_n - \theta - (\theta_{n+1} - \theta)}{\nu_{n+1}} = \frac{\theta_n - \theta_{n+1}}{\nu_{n+1}} = A_{n,\tau} \nabla_h g(X_{n+1}, \theta_n).$$

In addition,

$$\begin{aligned}\theta_n - \theta = H^{-1} \nabla_h g(X_{n+1}, \theta_n) + \left(H^{-1} \nabla G(\theta_n) - H^{-1} \nabla_h g(X_{n+1}, \theta_n)\right) \\ - \left(H^{-1} \nabla G(\theta_n) - (\theta_n - \theta)\right).\end{aligned}$$

Finally,

$$\begin{aligned}\theta_n - \theta = H^{-1} A_{n,\tau}^{-1} \frac{\theta_n - \theta - (\theta_{n+1} - \theta)}{\nu_{n+1}} + H^{-1} \left(\nabla G(\theta_n) - \nabla_h g(X_{n+1}, \theta_n)\right) \\ - H^{-1} \left(\nabla G(\theta_n) - H(\theta_n - \theta)\right),\end{aligned}$$

which implies that

$$\begin{aligned}\theta_{n,\tau} - \theta = t'_n \sum_{k=0}^{n} \ln(k+1)^{\tau'} H^{-1} A_{k,\tau}^{-1} \frac{\theta_k - \theta - (\theta_{k+1} - \theta)}{\nu_{k+1}} \\ + t'_n \sum_{k=0}^{n} \ln(k+1)^{\tau'} H^{-1} \left(\nabla G(\theta_k) - \nabla_h g(X_{k+1}, \theta_k)\right) \\ - t'_n \sum_{k=0}^{n} \ln(k+1)^{\tau'} H^{-1} \left(\nabla G(\theta_k) - H(\theta_k - \theta)\right)\end{aligned}$$

where $t'_n = \frac{1}{\sum_{k=0}^{n} \log(k+1)^{\tau'}}$.

Rate of convergence of $t'_n \sum_{k=0}^{n} \log(k+1)^{\tau'} H^{-1} (\nabla G(\theta_k) - \nabla_h g(X_{k+1}, \theta_k))$. Analogous to the proof of Theorem 3.1, one can check with the help of a law of large numbers for martingales that

$$\left\| t'_n \sum_{k=0}^{n} \log(k+1)^{\tau'} H^{-1} (\nabla G(\theta_k) - \nabla_h g(X_{k+1}, \theta_k)) \right\|_F^2 = o\left(\frac{\ln n^{1+\delta}}{n} \right).$$

Rate of convergence of $t'_n \sum_{k=0}^{n} \log(k+1)^{\tau'} H^{-1} (\nabla G(\theta_k) - H(\theta_k - \theta))$. With the help of Assumption **(A3)** and since θ_n converges almost surely to θ, $\|G(\theta_n) - H(\theta_n - \theta)\| = O\left(\|\theta_n - \theta\|^2\right)$ a.s. Then, with the help of Thanks to equality (25) and since $\nu < 1$, we can prove with the help of the equation (24) that

$$\left\| t'_n \sum_{k=0}^{n} \log(k+1)^{\tau'} H^{-1} (\nabla G(\theta_k) - H(\theta_k - \theta)) \right\|_F = o\left(\frac{\ln n^{1+\delta}}{n^{\nu}} \right),$$

and this term is negligible as soon as $\nu > 1/2$.

Rate of convergence of $t'_n \sum_{k=0}^{n} \ln(k+1)^{\tau'} H^{-1} A_{k+1,\tau}^{-1} \frac{\theta_k - \theta - (\theta_{k+1} - \theta)}{\nu_{k+1}}$. One can easily check that

$$\begin{aligned}
&t'_n \sum_{k=0}^{n} \ln(k+1)^{\tau'} H^{-1} A_{k,\tau}^{-1} \frac{\theta_k - \theta - (\theta_{k+1} - \theta)}{\nu_{k+1}} \\
&= t'_n \sum_{k=0}^{n} \ln(k+1)^{\tau'} H^{-1} \frac{A_{k,\tau}^{-1}(\theta_k - \theta) - A_{k+1,\tau}^{-1}(\theta_{k+1} - \theta)}{\nu_{k+1}} \\
&\quad + t'_n \sum_{k=0}^{n} \ln(k+1)^{\tau'} H^{-1} \frac{(A_{k+1,\tau}^{-1} - A_{k,\tau}^{-1})(\theta_{k+1} - \theta)}{\nu_{k+1}}.
\end{aligned}$$

With the help of an Abel's transform and since $A_{n,\tau}$ converges almost surely to H^{-1}, one has

$$\left\| t'_n \sum_{k=0}^{n} \ln(k+1)^{\tau'} H^{-1} \frac{A_{k,\tau}^{-1}(\theta_k - \theta) - A_{k+1,\tau}^{-1}(\theta_{k+1} - \theta)}{\nu_{k+1}} \right\|_F^2 = o\left(\frac{\ln n^{1+\delta}}{n^{2-\nu}} \right).$$

Observe that

$$A_{k+1,\tau}^{-1} - A_{k,\tau}^{-1} = A_{k+1,\tau}^{-1}(I_d - A_{k+1,\tau} A_{k,\tau}^{-1}) = A_{k+1,\tau}^{-1}(A_{k,\tau} - A_{k+1,\tau}) A_{k,\tau}^{-1}.$$

Therefore,

$$\begin{aligned}
&t'_n \sum_{k=0}^{n} \ln(k+1)^{\tau'} H^{-1} \frac{(A_{k+1,\tau}^{-1} - A_{k,\tau}^{-1})(\theta_{k+1} - \theta)}{\nu_{k+1}} \\
&= -\sum_{k=0}^{n} \ln(k+1)^{\tau'} H^{-1} \frac{A_{k+1,\tau}^{-1}(A_{k+1} - A_{k,\tau}) A_{k,\tau}^{-1}(\theta_{k+1} - \theta)}{\nu_{k+1}} \ln(k+1)^{\tau} t_k.
\end{aligned}$$

As A_{k+1} and $A_{k,\tau}$ converge to H^{-1},

$$\begin{aligned}
&\left\| t'_n \sum_{k=0}^{n} \ln(k+1)^{\tau'} H^{-1} \frac{(A_{k+1,\tau}^{-1} - A_{k,\tau}^{-1})(\theta_{k+1} - \theta)}{\nu_{k+1}} \right\|_F \\
&\leq t'_n \ln(n+1)^{\tau'} \sum_{k=0}^{n} \frac{\left\| H^{-1} A_{k+1,\tau}^{-1}(A_{k+1} - A_{k,\tau}) A_{k,\tau}^{-1} \right\|_{op} \|\theta_{k+1} - \theta\|}{\nu_{k+1}} \ln(k+1)^{\tau} t_k.
\end{aligned}$$

Thus, as $\|\theta_n - \theta\|^2 = o\left(\frac{\ln n^{1+\delta}}{n^{\nu}}\right)$ $a.s.$ we have for all $\delta > 0$

$$\left\| t'_n \sum_{k=0}^{n} \ln(k+1)^{\tau'} H^{-1} \frac{(A_{k+1,\tau}^{-1} - A_{k,\tau}^{-1})(\theta_{k+1} - \theta)}{\nu_{k+1}} \right\|_F^2 = o\left(\frac{\ln(n+1)^{1+\delta+2\tau}}{n^{2-\nu}} \right),$$

which is negligible as soon as $2 - \nu > 1$. Finally, we can conclude that

$$\|\theta_{n,\tau} - \theta\|^2 = o\left(\frac{\ln n^{1+\delta}}{n} \right) a.s.$$

6.5 Useful lemmas

The following lemma is a corollary of the Robbins-Siegmund theorem Robbins and Siegmund (1971).

Lemma 6.1. *Let (V_n), (B_n), (D_n) and (a_n) be positive sequences adapted to $\mathbb{F} = (\mathcal{F}_n)$. Assume that V_0 is integrable and, for all $n \geq 0$,*

$$\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq V_n + B_n - D_n \quad a.s.$$

Assume also that $\sum_{n=0}^{\infty} \frac{B_n}{a_n} < +\infty$ a.s. If $a_n \to \infty$, then $V_n = o\left(a_n\right)$ a.s.

The proof of this lemma is given in Chapter 1.III in Duflo (1990), and here we give a generalized version of it.

Lemma 6.2. *Let (V_n), (B_n), (D_n), E_n and (a_n) be positive sequences adapted to $\mathbb{F} = (\mathcal{F}_n)$. Assume that V_0 is integrable and, for all $n \geq 0$,*

$$\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq \left(1 + E_n\right) V_n + B_n - D_n \quad a.s.$$

Assume also that $\sum_{n=0}^{\infty} E_n < +\infty$ a.s. and $\sum_{n=0}^{\infty} \frac{B_n}{a_n} < +\infty$ a.s. If $a_n \to \infty$, then $V_n = o\left(a_n\right)$ a.s.

Proof of Lemma 6.2. Note that the case where $E_n = 0$ is exactly the case of Lemma 6.1. Therefore, we are going to study the case where $E_n \neq 0$. We define $\alpha_n := \prod_{k=0}^{n}(1 + E_k)$. Note that since $\sum_{n=0}^{\infty} E_n$ converges almost surely, α_n converges almost surely to a finite random variable α_∞. Moreover, noting

$$V_n' = \frac{V_n}{\alpha_{n-1}}, \quad B_n' = \frac{B_n}{\alpha_n}, \quad D_n' = \frac{D_n}{\alpha_n},$$

we observe that

$$\mathbb{E}\left[V'n|\mathcal{F}_n\right] \leq V_n' + B_n' - D_n'.$$

In addition, since $\alpha_n \geq 1$, we have

$$\sum_{n=0}^{\infty} B_n' \leq \sum_{n=0}^{\infty} B_n < +\infty \quad a.s.$$

According to Lemma 6.1, we have $V_n' = o(a_n)$ a.s., and therefore $V_n = o(a_n)$ a.s. Furthermore,

$$\sum_{n=0}^{\infty} D_n \leq \alpha_\infty \sum_{n=0}^{\infty} D_n' < +\infty \quad a.s.$$

□

We now give two lemmas which will be tools for the study of the rate of convergence associated to the estimates A_n.

Lemma 6.3. *Let us denote by $\mathcal{H} = \mathcal{M}_q(\mathbb{R})$ the set of squared matrices of size $q \times q$. Let us consider*

$$M_{n+1} = \sum_{k=1}^{n} \beta_{n,k}\gamma_k R_k \xi_{k+1},$$

where

- *(ξ_n) is a $\mathcal{H}$-valued martingale differences sequence adapted to a filtration $(\mathcal{F}_n)$ such that*

$$\mathbb{E}\left[\left\|\xi_{n+1}\right\|_F^2 |\mathcal{F}_n\right] \leq C + R_{2,n} \quad a.s,$$
$$\sum_{n\geq 1} \gamma_n \mathbb{E}\left[\left\|\xi_{n+1}\right\|_F^2 \mathbf{1}_{\left\|\xi_{n+1}\right\|_F^2 \geq \gamma_n^{-1}(\ln n)^{-1}} |\mathcal{F}_n\right] < +\infty \quad a.s, \tag{26}$$

  where C is a non-negative random variable and $(R_{2,n})_n$ converges almost surely to 0;

- *$\gamma_n = cn^{-\gamma}$ with $c > 0$ and $\gamma \in (1/2, 1)$;*

- (R_n) *is a sequence of matrices lying in* $\mathcal{H}$ *such that*

$$\|R_n\|_F = o\left(v_n\right) \quad a.s \qquad \textit{where} \qquad v_n = \frac{(\ln n)^a}{n^b},$$

with $a, b \in \mathbb{R}$*;*

- *For all* $n \geq 1$ *and* $1 \leq k \leq n$*, and for all* $A \in \mathcal{H}$*,*

$$\beta_{n,k} A = \prod_{j=k+1}^{n} \left(I_q - \gamma_j \Gamma\right) A \prod_{j=k+1}^{n} \left(I_q - \gamma_j \Gamma\right) \quad \textit{and} \quad \beta_{n,n} A = A,$$

where $\Gamma \in \mathcal{H}$ *is symmetric and satisfies* $0 < \lambda_{\min}(\Gamma) \leq \lambda_{\max}(\Gamma) < +\infty$*.*

Then,

$$\|M_{n+1}\|_F^2 = O\left(\gamma_n v_n^2 \ln n\right) \quad a.s.$$

Proof of Lemma 6.3. Let us now consider the events

$$\begin{aligned}
A_n &= \{\|R_n\|_F > v_n \quad \text{or} \quad R_{2,n} > C\} \\
B_{n+1} &= \{\|R_n\|_F \leq v_n, R_{2,n} \leq C, \|\xi_{n+1}\|_F \leq \delta_n\} \\
C_{n+1} &= \{\|R_n\|_F \leq v_n, R_{2,n} \leq C, \|\xi_{n+1}\|_F > \delta_n\}
\end{aligned}$$

with $\delta_n = \gamma_n^{-1/2} (\ln n)^{-1/2}$. One can check that $A_n^c = B_{n+1} \sqcup C_{n+1}$. Then, one can write M_{n+1} as

$$\begin{aligned}
M_{n+1} &= \sum_{k=1}^{n} \beta_{n,k} \gamma_k R_k \xi_{k+1} \mathbf{1}_{A_k} + \sum_{k=1}^{n} \beta_{n,k} \gamma_k R_k \xi_{k+1} \mathbf{1}_{A_k^c} \\
&= \sum_{k=1}^{n} \beta_{n,k} \gamma_k R_k \xi_{k+1} \mathbf{1}_{A_k} + \sum_{k=1}^{n} \beta_{n,k} \gamma_k R_k \left(\xi_{k+1} \mathbf{1}_{B_{k+1}} - \mathbb{E}\left[\xi_{k+1} \mathbf{1}_{B_{k+1}} | \mathcal{F}_k\right]\right) \\
&+ \sum_{k=1}^{n} \beta_{n,k} \gamma_k R_k \left(\xi_{k+1} \mathbf{1}_{C_{k+1}} - \mathbb{E}\left[\xi_{k+1} \mathbf{1}_{C_{k+1}} | \mathcal{F}_k\right]\right).
\end{aligned}$$

Let us now give the rates of convergence of these three terms.

Bounding $M_{1,n+1} := \sum_{k=1}^{n} \beta_{n,k} \gamma_k R_k \xi_{k+1} \mathbf{1}_{A_k}$**.** There exists a rank n_0 such that for all $n \geq n_0$, $\|I_q - \gamma_n \Gamma\|_{op} \leq (1 - \lambda_{\min} \gamma_n)$. Furthermore, $M_{1,n+1} = (I_q - \gamma_n \Gamma) M_{1,n} (I_q - \gamma_n \Gamma) + \gamma_n R_n \xi_{n+1} \mathbf{1}_{A_n}$. Then, for all $n \geq n_0$,

$$\mathbb{E}\left[\|M_{1,n+1}\|_F^2 \, | \mathcal{F}_n\right] \leq (1 - \lambda_{\min} \gamma_n)^4 \|M_{1,n}\|_F^2 + \gamma_n^2 \|R_n\|_F^2 \left(C + R_{2,n}\right) \mathbf{1}_{A_n}.$$

Considering $V_{n+1} = \prod_{k=1}^{n} (1 + \lambda_{\min} \gamma_k)^4 \|M_{1,n+1}\|_F^2$, it follows that

$$\mathbb{E}\left[V_{n+1} | \mathcal{F}_n\right] \leq \left(1 - \lambda_{\min}^2 \gamma_n^2\right)^4 V_n + \prod_{k=1}^{n} \left(1 + \lambda_{\min} \gamma_k\right)^4 \gamma_n^2 \|R_n\|_F^2 \left(C + R_{2,n}\right) \mathbf{1}_{A_n}$$

Moreover, $\mathbf{1}_{A_n}$ converges almost surely to 0 implying that

$$\sum_{n \geq 1} \prod_{k=1}^{n} \left(1 + \lambda_{\min} \gamma_k\right)^4 \gamma_n^2 \|R_n\|_F^2 \left(C + R_{2,n}\right) \mathbf{1}_{A_n} < +\infty \quad a.s$$

and applying Robbins-Siegmund Theorem, V_n converges almost surely to a finite random variable, i.e

$$\|M_{1,n+1}\|_F^2 = \mathcal{O}\left(\prod_{k=1}^{n} \left(1 + \lambda_{\min} \gamma_k\right)^{-4}\right) \quad a.s$$

and converges exponentially fast.

Bounding $M_{2,n+1} := \sum_{k=1}^{n} \beta_{n,k}\gamma_k R_k(\xi_{k+1}\mathbf{1}_{B_{k+1}} - \mathbb{E}[\xi_{k+1}\mathbf{1}_{B_{k+1}}|\mathcal{F}_k])$**.** Let us denote $\Xi_{k+1} = R_k(\xi_{k+1}\mathbf{1}_{B_{k+1}} - \mathbb{E}[\xi_{k+1}\mathbf{1}_{B_{k+1}}|\mathcal{F}_k])$. Remark that (Ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$. As in Pinelis (1994) (proofs of Theorems 3.1 and 3.2), let $\lambda > 0$ and consider for all $t \in [0,1]$ and $j \leq n$,

$$\varphi(t) = \mathbb{E}\left[\cosh\left(\lambda \left\|\sum_{k=1}^{j-1} \beta_{n,k}\gamma_k \Xi_{k+1} + t\beta_{n,j}\gamma_j \Xi_{j+1}\right\|_F\right)|\mathcal{F}_j\right].$$

One can check that $\varphi'(0) = 0$ and (see Pinelis for more details)

$$\varphi''(t) \leq \lambda^2 \mathbb{E}\left[\|\beta_{n,j}\gamma_j \Xi_{j+1}\|_F^2 \, e^{\lambda t \|\beta_{n,j}\gamma_j \Xi_{j+1}\|_F} \cosh\left(\lambda \left\|\sum_{k=1}^{j-1} \beta_{n,k}\gamma_k \Xi_{k+1}\right\|_F\right)|\mathcal{F}_j\right]$$

Then,

$$\begin{aligned}\mathbb{E}\left[\cosh\left(\lambda \left\|\sum_{k=1}^{j} \beta_{n,k}\gamma_k \Xi_{k+1}\right\|_F\right)|\mathcal{F}_j\right] = \varphi(1) &= \varphi(0) + \int_0^1 (1-t)\varphi''(t)dt \\ &\leq (1+e_{j,n}) \cosh\left(\lambda \left\|\sum_{k=1}^{j-1} \beta_{n,k}\gamma_k \Xi_{k+1}\right\|_F\right)\end{aligned}$$

with $e_{j,n} = \mathbb{E}[e^{\lambda\|\beta_{n,j}\gamma_j\Xi_{j+1}\|_F} - 1 - \lambda\|\beta_{n,j}\gamma_j\Xi_{j+1}\|_F|\mathcal{F}_j]$, which is well defined since Ξ_{j+1} is a.s. finite. Additionally, considering

$$G_{n+1} = \frac{\cosh\left(\lambda \left\|\sum_{k=1}^{n} \beta_{n,k}\gamma_k \Xi_{k+1}\right\|_F\right)}{\prod_{j=1}^{n}(1+e_{j,n})} \qquad \text{and} \qquad G_0 = 1$$

and since $\mathbb{E}[G_{n+1}|\mathcal{F}_n] = G_n$, it comes $\mathbb{E}[G_{n+1}] = 1$. For all $r > 0$,

$$\mathbb{P}\left[\|M_{2,n+1}\|_F \geq r\right] = \mathbb{P}\left[G_{n+1} \geq \frac{\cosh(\lambda r)}{\prod_{j=1}^{n}(1+e_{j,n})}\right] \leq \mathbb{P}\left[2G_{n+1} \geq \frac{e^{\lambda r}}{\prod_{j=1}^{n}(1+e_{jn})}\right].$$

Furthermore, let $\epsilon_{j+1} = \xi_{j+1}\mathbf{1}_{B_j} - \mathbb{E}[\xi_{j+1}\mathbf{1}_{B_j}|\mathcal{F}_j]$ and note that $\mathbb{E}[\|\epsilon_{j+1}\|_F^2|\mathcal{F}_j] \leq 2C$. Then, recalling that $\delta_n = \gamma_n^{-1/2}(\ln n)^{-1/2}$, and since for all $k \geq 2$,

$$\mathbb{E}\left[\|\epsilon_{j+1}\|_F^k |\mathcal{F}_j\right] \leq 2^{k-2}\delta_j^{k-2}\mathbb{E}\left[\|\xi_{j+1}\|_F^2 \mathbf{1}_{B_j}|\mathcal{F}_j\right] \leq 2^{k-1}C\delta_j^{k-2},$$

ans since for any $A \in \mathcal{H}$ one has $\|\beta_{n,k}A\|_F \leq \|\beta_{n,j}\|_{op}\|A\|_F$,

$$\begin{aligned}e_{j,n} &\leq \sum_{k=2}^{\infty} \lambda^k \|\beta_{n,j}\|_{op}^k \gamma_j^k \mathbb{E}\left[\|\Xi_{j+1}\|_F^k |\mathcal{F}_j\right] \leq \sum_{k=2}^{\infty} \lambda^k \|\beta_{n,j}\|_{op}^k \gamma_j^k v_j^k \mathbb{E}\left[\|\epsilon_{j+1}\|_F^k |\mathcal{F}_k\right] \\ &\leq \sum_{k=2}^{\infty} \lambda^k \|\beta_{n,j}\|_{op}^k \gamma_j^k v_j^k 2^{k-1} C\delta_j^{k-2} \\ &\leq 2C\lambda^2 \|\beta_{n,j}\|_{op}^2 \gamma_j^2 v_j^2 \sum_{k=2}^{\infty} (2\lambda)^{k-2} \|\beta_{n,j}\|_{op}^{k-2} \gamma_j^{\frac{k-2}{2}} v_j^{k-2} \ln j^{-\frac{k-2}{2}} \\ &= 2C\lambda^2 \|\beta_{n,j}\|_{op}^2 \gamma_j^2 v_j^2 \exp\left(2\lambda \|\beta_{n,j}\|_{op} \sqrt{\gamma_j} v_j\right)\end{aligned}$$

Then,

$$\begin{aligned}&\mathbb{P}\left[\|M_{2,n+1}\|_F \geq r\right] \\ \leq &\mathbb{P}\left[2G_{n+1} \geq \frac{e^{\lambda r}}{\prod_{j=1}^{n}\left(1 + 2C\lambda^2 \|\beta_{n,j}\|_{op}^2 \gamma_j^2 v_j^2 \exp\left(2\lambda \|\beta_{n,j}\|_{op} v_j \sqrt{\gamma_j \ln j}\right)\right)}\right]\end{aligned}$$

Applying Markov's inequality,

$$\mathbb{P}\left[\|M_{2,n+1}\| \geq r\right] \leq 2\exp\left(-\lambda r + 2C\lambda^2 \sum_{j=1}^{n} \|\beta_{n,j}\|_{op}^2 \gamma_j^2 v_j^2 \exp\left(2\lambda \|\beta_{n,j}\|_{op} v_j \sqrt{\gamma_j \ln j}\right)\right).$$

Take $\lambda = \gamma_n^{-1/2} v_n^{-1} \sqrt{\ln n}$. Let $C_0 = \|\beta_{n_0,0}\|_{op}$ and remark that for $n \geq 2n_0$ (i.e such that $\gamma_{n/2}\lambda_{\max}(\Gamma) \leq 1$), and for all $j \leq n/2$,

$$\|\beta_{n,j}\|_{op} \leq C_0 \exp\left(-c\lambda_{\min}(n/2)^{1-\alpha}\right),$$

so that for all $j \leq n/2$,

$$\lambda \|\beta_{n,j}\|_{op} \gamma_j v_j \xrightarrow[n\to+\infty]{a.s} 0.$$

Furthermore, for all $n \geq 2n_0$, and for all $j \geq n/2$,

$$\lambda \|\beta_{n,j}\|_{op} \frac{\sqrt{\gamma_j} v_j}{\sqrt{\ln j}} \leq C_0 2^{2b+2a+\alpha+1}.$$

Then, there is a positive constant C'' such that for all $n \geq 1$ and $j \leq n$,

$$\exp\left(\lambda \|\beta_{n,j}\|_{op} \sqrt{\gamma_j} v_j\right) \leq C''$$

Finally, one can easily check that (see Lemma E.2 in Cardot and Godichon-Baggioni (2017))

$$\sum_{j=1}^{n} \|\beta_{n,j}\|_{op}^2 \gamma_j^2 \frac{(\ln j)^{2a}}{j^{2b}} = \mathcal{O}\left(\frac{(\ln n)^{2a}}{n^{2b+\alpha}}\right).$$

There is a positive constant C''' such that

$$\mathbb{P}\left[\|M_{2,n+1}\|_F \geq r\right] \leq \exp\left(-r v_n^{-1} \gamma_n^{-1/2} \sqrt{\ln n} + C''' \ln n\right)$$

Then , taking $r = (2 + C''')v_n \sqrt{\gamma_n \ln n}$, it comes

$$\mathbb{P}\left[\|M_{2,n+1}\|_F \geq \left(2 + C'''\right) v_n \sqrt{\gamma_n \ln n}\right] \leq \exp\left(-2\ln n\right) = \frac{1}{n^2}$$

and applying Borell Cantelli's lemma,

$$\|M_{2,n+1}\|_F = \mathcal{O}\left(v_n \sqrt{\gamma_n \ln n}\right) \quad a.s.$$

Bounding $M_{3,n+1} := \sum_{k=1}^{n} \beta_{n,k}\gamma_k R_n(\xi_{k+1}\mathbf{1}_{C_{k+1}} - \mathbb{E}[\xi_{k+1}\mathbf{1}_{C_{k+1}}|\mathcal{F}_k])$**.** Let us denote

$$\epsilon_{k+1} = \xi_{k+1}\mathbf{1}_{C_{k+1}} - \mathbb{E}[\xi_{k+1}\mathbf{1}_{C_{k+1}}|\mathcal{F}_k]$$

and remark that for $n \geq n_0$,

$$\begin{aligned}
\mathbb{E}\left[\|M_{3,n+1}\|_F^2 |\mathcal{F}_n\right] &\leq (1 - \lambda_{\min}\gamma_n)^4 \|M_{3,n}\|_F^2 + \gamma_n^2 v_n^2 \mathbb{E}\left[\|\epsilon_{n+1}\|_F^2 |\mathcal{F}_n\right] \\
&\leq (1 - \lambda_{\min}\gamma_n)^4 \|M_{3,n}\|_F^2 + \gamma_n^2 v_n^2 \mathbb{E}\left[\|\xi_{n+1}\|_F^2 \mathbf{1}_{\|\xi_{n+1}\|_F^2 \geq \gamma_n^{-1}} |\mathcal{F}_n\right]
\end{aligned}$$

Let $V_n' = \gamma_n^{-1} v_n^{-2} \|M_{3,n}\|_F^2$. There are a rank n_1 and a positive constant c such that for all $n \geq n_1$

$$\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq (1 - c\gamma_n) V_n + \mathcal{O}\left(\gamma_n \mathbb{E}\left[\|\xi_{n+1}\|_F^2 \mathbf{1}_{\|\xi_{n+1}\|_F^2 \geq \gamma_n^{-1}} |\mathcal{F}_n\right]\right) \quad a.s.$$

Applying Robbins-Siegmund Theorem as well as equation (26), it comes

$$\|M_{3,n+1}\|_F^2 = \mathcal{O}\left(\gamma_n v_n^2\right) \quad a.s.$$

□

Lemma 6.4. *Let J_n, K_n, r_n be sequences of positive random variables such that r_n converges almost surely to 0 and*

$$J_{n+1} = (1 - c\widetilde{\gamma}_{n+1})J_n + \widetilde{\gamma}_{n+1} r_n (J_n + K_n)$$

where $\widetilde{\gamma}_n = c_{\widetilde{\gamma}} n^{\widetilde{\gamma}}$ with $1/2 < \widetilde{\gamma} < 1$ and $c_{\widetilde{\gamma}} > 0$. In addition, it is assumed that

$$K_n = \mathcal{O}(v_n) \quad a.s.$$

where $v_n = c_v n^v (\ln n)^b$ with $v \in \mathbb{R}$ and $b \geq 0$. Then

$$J_n = \mathcal{O}(v_n) \quad a.s.$$

Proof of Lemma 6.4. For the sake of simplicity, let us assume that for every $n \geq 0$, $c\widetilde{\gamma}_{n+1} \leq 1$ (up to take n large enough). Now, consider the event $E_{n,c} = \{|r_n| \leq c/2\}$, and therefore $\mathbf{1}_{E_{n,c}^C}$ converges almost surely to 0. Hence, J_{n+1} can be rewritten as:

$$\begin{aligned} J_{n+1} &\leq (1 - c\widetilde{\gamma}_{n+1})\, J_n + \frac{c}{2}\widetilde{\gamma}_{n+1}\,(J_n + K_n) + \overbrace{\widetilde{\gamma}_{n+1} r_n\,(J_n + K_n)}^{=:\delta_n} \mathbf{1}_{E_{n,c}^C} \\ &\leq \left(1 - \frac{c}{2}\widetilde{\gamma}_{n+1}\right) J_n + \frac{c}{2}\widetilde{\gamma}_{n+1} K_n + \delta_n \mathbf{1}_{E_{n,c}^C} \end{aligned}$$

By induction, one can check that for all $n \geq 0$:

$$J_n \leq \tilde{\beta}_{n,0} J_0 + \underbrace{\frac{c}{2}\sum_{k=0}^{n-1} \tilde{\beta}_{n,k+1}\widetilde{\gamma}_{k+1} K_k}_{=:J_{1,n}} + \underbrace{\sum_{k=0}^{n-1} \tilde{\beta}_{n,k+1}\delta_k \mathbf{1}_{E_{k,c}^C}}_{=:J_{2,n}}$$

with $\tilde{\beta}_{n,k} := \prod_{j=k+1}^{n} \left(1 - \frac{c}{2}\widetilde{\gamma}_j\right)$ and $\tilde{\beta}_{n,n} := 1$. Using standard calculations, we can easily show that $\tilde{\beta}_{n,0}$ converges at an exponential rate. Furthermore, $J_{2,n}$ can be written as $\tilde{\beta}_{n,0}\sum_{k=0}^{n-1} \tilde{\beta}_{k,0}^{-1}\delta_k \mathbf{1}_{E_{k,c}^C}$ and since $\mathbf{1}_{E_{n,c}^C}$ converges almost surely to 0, the sum is almost surely finite, leading to

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

and this term thus converges at an exponential rate. Finally, there exists a random variable K such that for every $n \geq 1$, $K_n \leq K v_n$ almost surely, leading to the induction relation:

$$J_{1,n+1} = \left(1 - \frac{c}{2}\widetilde{\gamma}_{n+1}\right) J_{1,n} + \frac{c}{2}\widetilde{\gamma}_{n+1} K_n \leq \left(1 - \frac{c}{2}\widetilde{\gamma}_{n+1}\right) J_{1,n} + \frac{c}{2}\widetilde{\gamma}_{n+1} K v_n$$

By applying Proposition A.4 in Godichon-Baggioni et al. (2023), we obtain:

$$J_{1,n} = O\left(v_n\right) \quad \text{a.s.}$$

□

References

Bach, F. (2014). Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research*, 15(1):595–627.

Bercu, B., Bigot, J., Gadat, S., and 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.

Bercu, B., Costa, M., and Gadat, S. (2020a). Stochastic approximation algorithms for superquantiles estimation. *arXiv preprint arXiv:2007.14659*.

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

Blackard, J. A. and Dean, D. J. (1999). Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and electronics in agriculture*, 24(3):131–151.

Boyer, C. and Godichon-Baggioni, A. (2022). On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *Computational Optimization and Applications*, pages 1–52.

Cardot, H., Cénac, P., and Godichon-Baggioni, A. (2017). Online estimation of the geometric median in hilbert spaces: Nonasymptotic confidence balls. *The Annals of Statistics*, 45(2):591–614.

Cardot, H., Cénac, P., and Zitt, P.-A. (2013). Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19(1):18–43.

Cardot, H. and Godichon-Baggioni, A. (2017). Fast estimation of the median covariation matrix with application to online robust principal components analysis. *Test*, 26(3):461–480.

Cénac, P., Godichon-Baggioni, A., and Portier, B. (2020). An efficient averaged stochastic gauss-newton algorithm for estimating parameters of non linear regressions models. *arXiv preprint arXiv:2006.12920*.

Cohen, K., Nedić, A., and Srikant, R. (2017). On projected stochastic gradient descent algorithm with weighted averaging for least squares regression. *IEEE Transactions on Automatic Control*, 62(11):5974–5981.

Costa, M. and Gadat, S. (2020). Non asymptotic controls on a recursive superquantile approximation.

Duflo, M. (1990). Méthodes récursives aléatoires. *(No Title)*.

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pages 215–310.

Gadat, S. and 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. (2021). Convergence in quadratic mean of averaged stochastic gradient algorithms without strong convexity nor bounded gradient. *arXiv preprint arXiv:2107.12058*.

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

Godichon-Baggioni, A. and Portier, B. (2017). An averaged projected robbins-monro algorithm for estimating the parameters of a truncated spherical distribution.

Godichon-Baggioni, A., Portier, B., and Lu, W. (2022). Recursive ridge regression using second-order stochastic algorithms.

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

Leluc, R. and Portier, F. (2020). Asymptotic optimality of conditioned stochastic gradient descent. *arXiv preprint arXiv:2006.02745*.

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.

Moulines, E. and Bach, F. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24.

Pelletier, M. (1998). On the almost sure asymptotic behaviour of stochastic algorithms. *Stochastic processes and their applications*, 78(2):217–244.

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

Pinelis, I. (1994). Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Robbins, H. and Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier.

Schmidt, M., Le Roux, N., and Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112.

Toulis, P., Tran, D., and Airoldi, E. (2016). Towards stability and optimality in stochastic gradient descent. In *Artificial Intelligence and Statistics*, pages 1290–1298. PMLR.