On the asymptotic rate of convergence of Stochastic Newton algorithms and their Weighted Averaged versions

Claire Boyer, Antoine Godichon-Baggioni

On the asymptotic rate of convergence of Stochastic Newton algorithms and their Weighted Averaged versions

Claire Boyer and Antoine Godichon-Baggioni,
Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne-Université, 75005 Paris, France
claire.boyer@upmc.fr , antoine.godichon_baggioni@upmc.fr

January 9, 2021

Abstract

The majority of machine learning methods can be regarded as the minimization of an unavailable risk function. To optimize the latter, with samples provided in a streaming fashion at hand, we define a general stochastic Newton algorithm and its weighted average versions. In several use cases, both implementations will be shown not to require the inversion of a Hessian estimate at each iteration, but a direct update of the estimate of the inverse Hessian instead will be favored. This generalizes a trick introduced in [2] for the specific case of logistic regression, and results in a cost of $O(d^2)$ operations per iteration, for d the ambient dimension. Under mild assumptions such as local strong convexity at the optimum, we establish almost sure convergences and rates of convergence of the algorithms, as well as central limit theorems for the constructed parameter estimates. The unified framework considered in this paper covers the case of linear, logistic or softmax regressions to name a few. Numerical experiments on simulated and real data give the empirical evidence of the pertinence of the proposed methods, which outperform popular competitors particularly in case of bad initializations.

Keywords: Stochastic optimization; Newton algorithm; Averaged stochastic algorithm; Statistical learning

1 Introduction

Recently, machine learning challenges are encountered in many different scientific applications facing streaming or large amounts of data. First-order online algorithms have become hegemonic: by a cheap computational cost per iteration, they allow to perform machine learning task on large datasets, processing each observation only once, see for instance the review paper [3]. The stochastic gradient methods (SGDs) and their averaged versions are shown to be theoretically asymptotically efficient [16, 15, 11] while recent works focus on the non-asymptotic behavior of these estimates [1, 10, 9]: more precisely, it was proven that under mild assumptions, averaged estimates can converge at a rate of order $O(1/n)$ where we let n denote the size of the dataset (and the number of iterations as well, in a streaming setting). However, these first-order online algorithms can be shown in practice to be very sensitive to the Hessian structure of the risk they are supposed to minimize. For instance, when the spectrum of local Hessian matrices shows large variations among their eigenvalues, the stochastic gradient algorithm may be stuck far from the optimum, see for instance the application of [2, Section 5.2].

To address this issue, (quasi) online second-order optimization has been also considered in the literature. In view of avoiding highly costly iterations, most online (quasi) second-order algorithms rely on approximating the Hessian matrix by exclusively using gradient information or by assuming a diagonal structure of it (making its inversion much easier). These methods result to choose a different step size with respect to the components of the current gradient estimate, hence the name of adaptive stochastic gradient algorithms, such as Adagrad [7] or Adadelta [17] methods.

In this paper, we aim at minimizing the general convex function G defined for any parameter $h \in \mathbb{R}^d$ as

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

where X denotes a random variable taking value in $\mathcal{X}$ and standing for a data sample, and $g : \mathcal{X} \times \mathbb{R}^d \longrightarrow \mathbb{R}$ is a loss function. The function G may encode the risk of many supervised or unsupervised machine learning procedures, encompassing for instance linear, logistic or even softmax regressions. Having only access to the data points $X_1, \ldots, X_n$, i.i.d. copies of X, instead of the true underlying distribution of X, we propose new stochastic Newton methods in order to perform the optimization of G, relying on estimates of both the Hessian and its inverse, using second-order information of g.

1.1 Related works

A Stochastic Quasi-Newton method was introduced in [4], relying on limited-memory BFGS updates. Specifically, local curvature is captured through (subsampled) Hessian-vector products, instead of differences of gradients. The authors provide a stochastic Quasi-Newton algorithm which cost is close to the one of standard SGDs. The convergence study in [4] requires the boundedness from above and from below of the spectrum of the estimated Hessian inverses, uniformly over the space of parameters, which can be very restrictive. Furthermore, the theoretical analysis does not include the convergence of the estimates to the inverse of the Hessian, which cannot ensure an optimal asymptotic behavior of the algorithm.

A hybrid algorithm combining gradient descent and Newton-like behaviors as well as inertia is proposed in [5]. Under the Kurdyka-Łojasiewicz (KL) property, the theoretical analysis of the associated continuous dynamical model is conducted, which significantly departs from the type of convergence guarantees established in this paper.

A truncated Stochastic Newton algorithm has been specifically introduced for and dedicated to logistic regression in [2]. The recursive estimates of the inverse of the Hessian are updated through the Ricatti's formula (also called the Sherman-Morrison's formula) leading to only $O(d^2)$ operations at each iteration. Only in the particular case of logistic regression, optimal asymptotic behaviour of the algorithm is established under assumptions close to the ones allowed by the general framework considered in the present paper. Furthermore, the considered step sequence of the order of $1/n$ in [2] may freeze the estimates dynamics in practice, leading to poor results in case of bad initialization.

In [12], the authors introduced a conditioned SGD based on a preconditioning of the gradient direction. The preconditioning matrix is typically an estimate of the inverse Hessian at the optimal point, for which they obtain asymptotic optimality of the procedure under L-smoothness assumption of the objective function, and boundedness assumption of the spectrum of all the preconditioning matrices used over the iterations. They propose to use preconditioning matrices in practice as inverse of weighted recursive estimates of the Hessian. The proposed conditioned SGD thus entails a full inversion of the estimated Hessian, requiring $O(d^3)$ operations per iteration in general, which is less compatible with large-scale data. Note that the weighting procedure only concerns the estimation of the Hessian, and not the whole algorithm as we will suggest in this paper. The choice of the step sequence in the conditioned SGD remains problematic: choosing steps

of the order $1/n$, which is theoretically sound to obtain optimal asymptotic behaviour, may result in the saturation of the algorithm far from the optimum, particularly in case of bad initializations.

In order to reduce the sensitivity to the initialization, an averaged Stochastic Gauss-Newton algorithm has been proposed in the restricted setting of non-linear regression in [6]. Despite the peculiar case of non-linear regression, stochastic algorithms are shown to benefit from averaging in simulations.

1.2 Contributions

In this paper, we consider an unified and general framework that includes various applications of machine learning tasks, for which we propose a stochastic Newton algorithm: an estimate of the Hessian is constructed and *easily* updated over iterations using genuine second order information. Given a particular structure of the Hessian estimates that will be encountered in various applications, this algorithm leverages from the possibility to directly update the inverse of the Hessian matrix at each iteration in $O(d^2)$ operations, with d the ambient dimension, generalizing a trick introduced in the context of logistic regression in [2]. For the sake of simplicity, a first version of this algorithm is studied choosing the step size of the order $O(1/n)$ where we let n denote the number of iterations. Under suitable and standard assumptions, we establish the following asymptotic results: (i) the almost sure convergence, and (ii) almost sure rates of convergence of the iterates to the optimum, as well as (iii) a central limit theorem for the iterates. Nevertheless, as mentioned before, considering step sequences of order $1/n$ can lead to bad results in practice [6]. In order to alleviate this problem, we thus introduce a weighted averaged stochastic Newton algorithm (WASNA) which can benefit from better step size choices and above all from weighted averaging over the iterates. We then establish the almost sure rates of convergence of its estimates, preserving an optimal asymptotic behavior of the whole procedure. This work allows a unified framework encompassing the case of linear, logistic and softmax regressions, for which WASNAs are derived, coming with their convergence guarantees. To our knowledge, this is the first time that the softmax regression is covered by the proposed second order online method (not specifically designed for the softmax regression). To put in a nutshell, we propose new and very general weighted stochastic Newton algorithms, (i) that can be implemented efficiently in regard to their second order characteristics, (ii) that allow various choices of weighting discussed at the light of the induced convergence

optimality, and (iii) for which theoretical locks have been lifted resulting in strong convergence guarantees without requiring a global strong convexity assumption. The relevance of the proposed algorithms is illustrated in numerical experiments, challenging the favorite competitors such as SGDs with adaptive learning rate. Even without a refined tuning of the hyperparameters, the method is shown to give good performances on the real MNIST[1] dataset in the context of multi-label classification. For reproducibility purposes, the code of all the numerical experiments is available at godichon.perso.math.cnrs.fr/research.html.

1.3 Organization of the paper

The general framework is presented in Section 2 introducing all the notation. The set of mild assumptions on G and g is also discussed. Section 3 presents a new general stochastic Newton algorithm and the associated theoretical guarantees. Section 4 introduces a new weighted averaged stochastic Newton algorithm, followed by its theoretical study in which optimal asymptotic convergence is obtained. The versatility and the relevance of the proposed algorithms is illustrated in Section 5 in the case of linear, logistic and softmax regressions, both in terms of theoretical guarantees and numerical implementation on simulated data.

1.4 Notation

In the following, we will denote by $\|\cdot\|$ the Euclidean norm in dimension d, and by $\|\cdot\|_{op}$ the operator norm corresponding to the largest singular value in finite dimension. The Euclidean ball centered at c and of radius r will be noted as $\mathcal{B}(c, r)$.

2 Framework

Let X be a random variable taking values in a space $\mathcal{X}$. The aim of this work is to estimate the minimizer of the convex function $G : \mathbb{R}^d \longrightarrow \mathbb{R}$ defined for all $h \in \mathbb{R}^d$ by

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

for a loss function $g : \mathcal{X} \times \mathbb{R}^d \longrightarrow \mathbb{R}$.

[1] http://yann.lecun.com/exdb/mnist/

In this paper, under the differentiability of G, we assume that the first order derivatives also meet the following assumptions:

(A1) For almost every $x \in \mathcal{X}$, the functional $g(x, .)$ is differentiable and

- **(A1a)** there is $\theta \in \mathbb{R}^d$ such that $\nabla G(\theta) = 0$;
- **(A1b)** there are non-negative constants C and C' such that for all $h \in \mathbb{R}^d$,
$$\mathbb{E}\left[\|\nabla_h g(X, h)\|^2\right] \leq C + C'\left(G(h) - G(\theta)\right);$$
- **(A1c)** the functional
$$\Sigma : h \longmapsto \mathbb{E}\left[\nabla_h g(X, h)\, \nabla_h g(X, h)^T\right]$$
is continuous at θ.

Furthermore, second order information will be crucial for the definition of the Newton algorithms to come, for which we require the following assumptions to hold:

(A2) The functional G is twice continuously differentiable and

- **(A2a)** the Hessian of G is bounded, i.e. there is a positive constant $L_{\nabla G} > 0$ such that for all $h \in \mathbb{R}^d$,
$$\left\|\nabla^2 G(h)\right\|_{op} \leq L_{\nabla G};$$
- **(A2b)** the Hessian of G is positive at θ and we denote by $\lambda_{\min}$ its smallest eigenvalue.
- **(A2c)** the Hessian of G is Lipshitz on a neighborhood of θ: there are positive constants $A_{\nabla^2 G} > 0$ and $L_{A,\nabla^2 G} > 0$ such that for all $h \in \mathcal{B}(\theta, A_{\nabla^2 G})$,
$$\left\|\nabla^2 G(h) - \nabla^2 G(\theta)\right\|_{op} \leq L_{A,\nabla^2 G} \|\theta - h\|.$$

Remark that Assumption **(A2a)** leads the gradient of G to be Lipschitz continuous, and in particular at the optimum θ, for any $h \in \mathbb{R}^d$, one has

$$\|\nabla G(h)\| \leq L_{\nabla G} \|h - \theta\|.$$

Overall, note that all these assumptions are very close to the ones given in [14], [15], [11] or [10]. One of the main differences concerns Assumption

(A1b) in which the second order moments of $\nabla_h g(X, .)$ are not assumed to be upper-bounded by the squared errors $\|\cdot - \theta\|^2$, but by the risk error instead, i.e. by $G(\cdot) - G(\theta)$. Note that the first condition may entail the second one, when considering the functional G to be μ-strongly convex, since for any $h \in \mathbb{R}^d$, $\|h - \theta\|^2 \leq \frac{2}{\mu}(G(h) - G(\theta))$. In this respect, Assumption **(A1b)** can be seen as nearly equivalent to counterparts encountered in the literature.

3 The stochastic Newton algorithm

In order to lighten the notation, let us denote the Hessian of G at the optimum θ by $H = \nabla^2 G(\theta)$. As already mentioned in [2], usual stochastic gradient algorithms and their averaged versions are shown to be theoretically efficient, but can be very sensitive to the case where H has eigenvalues at different scales. With the aim of alleviating this problem, we first focus on the stochastic Newton algorithm (SNA).

3.1 Implementation of the SNA

3.1.1 The algorithm

Consider $X_1, \ldots, X_n, \ldots$ i.i.d. copies of X. The Stochastic Newton Algorithm (SNA) can be iteratively defined as follows: for all $n \geq 0$,

$$\theta_{n+1} = \theta_n - \frac{1}{n + 1 + c_\theta} \overline{H}_n^{-1} \nabla_h g\left(X_{n+1}, \theta_n\right) \tag{1}$$

given a finite initial point θ_0, for $c_\theta \geq 0$ and with $\overline{H}_n^{-1}$ a recursive estimate of H^{-1}, chosen symmetric and positive at each iteration. Remark that the constant c_θ plays a predominant role in the first iterations (for small n) since it balances the impact of the first (and probably bad) estimates out. This effect diminishes as the number n of iterations grows and hopefully as the iterates improve. On a more technical side, let us suppose that one can construct a filtration $(\mathcal{F}_n)$ verifying that for any $n \geq 0$, (i) $\overline{H}_n^{-1}$ and θ_n are $\mathcal{F}_n$-measurable, and (ii) X_{n+1} is independent from $\mathcal{F}_n$. Note that if the estimate $\overline{H}_n^{-1}$ only depends on $X_1, \ldots, X_n$, one can thus consider the filtration $\mathcal{F}_n$ generated by the current sample, i.e. for all $n \geq 1$, $\mathcal{F}_n = \sigma\left(X_1, \ldots, X_n\right)$.

3.1.2 Construction of $\overline{H}_n^{-1}$

When a natural online estimate of the Hessian $\overline{H}_n = (n+1)^{-1} H_n$ of the form

$$H_n = H_0 + \sum_{k=1}^{n} u_k \left(X_k, \theta_{k-1}\right) \Phi_k \left(X_k, \theta_{k-1}\right) \Phi_k \left(X_k, \theta_{k-1}\right)^T, \tag{2}$$

with H_0 symmetric and positive, is available, a computationally-cheap estimate of its inverse can be constructed. Indeed, the inverse H_{n+1}^{-1} can be easily updated thanks to Riccati's formula [8], i.e.

$$H_{n+1}^{-1} = H_n^{-1} - u_{n+1} \left(1 + u_{n+1} \Phi_{n+1}^T H_n^{-1} \Phi_{n+1}\right)^{-1} H_n^{-1} \Phi_{n+1} \Phi_{n+1}^T H_n^{-1} \tag{3}$$

with $\Phi_{n+1} = \Phi_{n+1}\left(X_{n+1}, \theta_n\right)$ and $u_{n+1} = u_{n+1}\left(X_{n+1}, \theta_n\right)$. In such a case, one can consider the filtration generated by the sample again, i.e. $\mathcal{F}_n = \sigma\left(X_1, \ldots, X_n\right)$. In Section 5, the construction of the recursive estimates of the inverse of the Hessian will be made explicit in the cases of linear, logistic and softmax regressions.

3.2 Convergence results for the SNA

A usual key ingredient to establish the almost sure convergence of the estimates constructed by stochastic algorithms is the Robbins-Siegmund theorem [8]. To do so for the stochastic Newton algorithm proposed in (1), we need to prevent the possible divergence of the eigenvalues of $\overline{H}_n^{-1}$. In addition, to ensure the convergence to the true parameter solution θ, the smallest eigenvalue of $\overline{H}_n^{-1}$ should be bounded from below, i.e. $\lambda_{\max}\left(\overline{H}_n\right)$ should be bounded from above. To this end, we require the following assumption to be satisfied:

(H1) The largest eigenvalue of $\overline{H}_n$ and $\overline{H}_n^{-1}$ can be controlled: there is $\beta \in (0, 1/2)$ such that

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

If verifying Assumption **(H1)** in practice may seem difficult, we will see in Section 5 how to modify the recursive estimates of the inverse the Hessian to obtain such a control on their spectrum, while preserving a suitable filtration. The following theorem gives the strong consistency of the stochastic Newton estimates constructed in (1).

Theorem 3.1. *Under Assumptions* ***(A1a)***, ***(A1b)***, ***(A2a)*** ***(A2b)*** *and* ***(H1)***, *the iterates of the stochastic Newton algorithm given in* (1) *satisfy*

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

The proof is given in Supplementary Materials A.1 and consists in a particular case of the proof of the almost sure convergence for a more general algorithm than the SNA.

The convergence rate of the iterates for the SNA is a more delicate result to obtain, that requires the convergence of $\overline{H}_n$ and $\overline{H}_n^{-1}$. This is why we suppose from now on that the following assumption is fulfilled:

(H2a) If Assumptions **(A1a)**, **(A1b)**, **(A2a)**, **(A2b)**, and **(H1)** hold,

$$\overline{H}_n \xrightarrow[n\to+\infty]{a.s} H.$$

Assumption **(H2a)** formalizes how the convergence of θ_n may yield the convergence of $\overline{H}_n$. Remark that Assumption **(H2a)** would also imply the convergence of $\overline{H}_n^{-1}$. Such a hypothesis will be verified in practical use cases, mainly relying on the continuity of the Hessian at the solution θ (see for instance the proof of the coming Theorem 5.2 in the setting of the softmax regression). The following theorem gives the rate of convergence associated to the SNA (1).

Theorem 3.2. *Under Assumptions* ***(A1)***, ***(A2)***, ***(H1)*** *and* ***(H2a)***, *the iterates of the stochastic Newton algorithm given in* (1) *satisfy for all* $\delta > 0$,

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

In addition, if there exist positive constants $a > 2$ *and* C_a *such that for all* $h \in \mathbb{R}^2$,

$$\mathbb{E}\left[\|\nabla_h g(X,h)\|^a\right] \leq C_a\left(1 + \|h-\theta\|^a\right), \tag{4}$$

then,

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

The proof of Theorem 3.2 is given in Supplementary Materials A.2. This type of results are analogous to the usual ones dedicated to online estimation based on Robbins-Monro procedures [16, 15, 10]. Besides, one could

note that the requirements on the functional G to obtain rates of convergence for the SNA are very close to the ones requested to obtain rates of convergence for the averaged stochastic gradient descent [15].

Refining the theoretical analysis of SNA defined in (1), we now aim at studying the variance optimality of such a procedure. To establish strong results as the asymptotic efficiency of the parameter estimates, the estimates of the Hessian should admit a (weak) rate of convergence. In this sense, we consider the following assumption:

(H2b) Under Assumptions **(A1)**, **(A2)**, **(H1)** and **(H2a)**, there exists a positive constant p_H such that

$$\left\|\overline{H}_n - H\right\|^2 = O\left(\frac{1}{n^{p_H}}\right) \quad a.s.$$

.

Assumption **(H2b)** captures how a rate of convergence of θ_n may lead to a rate of convergence of $\overline{H}_n$. With this hypothesis at hand, we are able to establish the optimal asymptotic normality of the iterates (1).

Theorem 3.3. *Under Assumptions **(A1)**,**(A2)**, **(H1)**, **(H2a)** and **(H2b)**, the iterates of the stochastic Newton algorithm given in* (1) *satisfy*

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

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

The proof is given in Supplementary Materials A.3. In Theorem 3.3, the estimates (1) of the SNA are ensured to be asymptotically efficient provided usual assumptions on the functional G matching the ones made in [16, 15, 11]. This result highlights the benefit of stochastic Newton algorithms over standard online gradient methods, which have been shown not to be asymptotically optimal [14], unless considering an averaged version [15].

Note that this result has been achieved independently of the work of [12].

The central limit theorem established in Theorem 3.3 requires convergence rates of the Hessian estimates by Assumption **(H2b)**. This can been often ensured at the price of technical results (such as the Lipschitz continuity of the Hessian or a quite large number of bounded moments of X). In Section 4.2, we will see how to relax Assumption **(H2b)** for a modified version of the stochastic Newton algorithm.

4 The weighted averaged stochastic Newton algorithm

In this section, we propose a modified version of the SNA, by allowing non-uniform or uniform averaging over the iterates. This leads to the weighted averaged stochastic Newton algorithm (WASNA). To our knowledge, this is the first time that a general WASNA covering a large spectrum of applications is studied, allowing non-uniform weighting as well.

4.1 Implementation of the WASNA

4.1.1 The algorithm

As mentioned in [6], considering the stochastic Newton algorithm ends up taking decreasing steps at the rate $1/n$ (up to a matrix multiplication), which can clog up the dynamics of the algorithm. A bad initialization can then become a real obstacle to the high performance of the method. To circumvent this issue, we consider the Weighted Averaged Stochastic Newton Algorithm (WASNA) defined recursively for all $n \geq 0$ by

$$\tilde{\theta}_{n+1} = \tilde{\theta}_n - \gamma_{n+1} \overline{S}_n^{-1} \nabla_h g\left(X_{n+1}, \tilde{\theta}_n\right) \tag{5}$$

$$\theta_{n+1,\tau} = (1 - \tau_{n+1})\, \theta_{n,\tau} + \tau_{n+1} \tilde{\theta}_{n+1}, \tag{6}$$

given

- finite starting points $\theta_{\tau,0} = \tilde{\theta}_0$,
- $\gamma_n = \frac{c_\gamma}{\left(n + c'_\gamma\right)^\gamma}$ with $c_\gamma > 0$, $c'_\gamma \geq 0$ and $\gamma \in (1/2, 1)$,
- $\overline{S}_n^{-1}$ a recursive estimate of H^{-1}, chosen symmetric and positive at each iteration,
- the weighted averaging sequence (τ_n) that should satisfy
  - (τ_n) is $\mathcal{GS}(\nu)$ for some $\nu < 0$ (see [13]), i.e.

$$n\left(1 - \frac{\tau_{n-1}}{\tau_n}\right) \xrightarrow[n \to +\infty]{} \nu. \tag{7}$$

  - There is a constant $\tau > \max\{1/2, -\nu/2\}$ such that

$$n\tau_n \xrightarrow[n \to +\infty]{} \tau. \tag{8}$$

Note that the WASNA may output both $\tilde{\theta}_n$ and $\theta_{n+1,\tau}$, but only the last iterate $\theta_{n+1,\tau}$ benefits from averaging, and with this respect, only the last iterate $\theta_{n+1,\tau}$ should be considered.

Moreover, as for the SNA, one can construct a filtration $(\mathcal{F}_n)$ such that (i) $\overline{S}_n$ and $\tilde{\theta}_n$ are $\mathcal{F}_n$-measurable and (ii) X_{n+1} is independent from $\mathcal{F}_n$.

4.1.2 Construction of $\overline{S}_n^{-1}$

The only difference between $\overline{H}_n^{-1}$ and $\overline{S}_n^{-1}$ is that the former depends on estimates $(\theta_n)_n$ defined in (1) and the latter is constructed using the estimates $(\theta_{n,\tau})_n$ (or eventually $(\tilde{\theta}_n)_n$) given in (6). Here again, if natural Hessian estimates admit the form (2), the Riccatti's trick used in (3) can be applied to directly update $\overline{S}_n^{-1}$. Section 5 will exemplify the construction of recursive estimates $(\overline{S}_n^{-1})_n$ for the Hessian inverse in the cases of linear, logistic and softmax regressions.

4.1.3 Different weighted versions

By choosing different sequences $(\tau_n)_n$, one can play more or less on the strength of the last iterates in the optimization. For instance, choosing $\tau_n = \frac{1}{n+1}$ (which is compatible with (7) and (8)) leads to the "usual" averaging in stochastic algorithms (see [6] for instance for an averaged version of a stochastic Newton algorithm specific to the non-linear regression setting)

$$\overline{\theta}_n = \frac{1}{n+1} \sum_{k=0}^{n} \tilde{\theta}_k. \tag{9}$$

Remark that the estimate obtained with standard averaging in (9) is denoted by $\overline{\theta}_n$. When considering $\tau_n = \frac{(n+1)^\omega}{\sum_{k=0}^{n}(k+1)^\omega}$ with $\omega \geq 0$ instead leads to another version of the weighted averaged stochastic Newton algorithm, for which the iterates are denoted by $\theta_{n,\omega}$ defined as follows

$$\theta_{n,\omega} = \frac{1}{\sum_{k=0}^{n}(k+1)^\omega} \sum_{k=0}^{n} (k+1)^\omega \tilde{\theta}_k. \tag{10}$$

This particular choice enables to give more importance to last estimates $\tilde{\theta}_n$ that should improve on the first iterates. This strategy can be motivated by limiting the effect of bad initialization of the algorithms.

4.2 Convergence results

As in the case of the SNA, we need to control the possible divergence of the eigenvalues of $\overline{S}_n$ and $\overline{S}_n^{-1}$ to ensure the convergence of the estimates. To this end, suppose that the following assumption holds true:

(H1') The largest eigenvalues of $\overline{S}_n$ and $\overline{S}_n^{-1}$ can be controlled in the sense that, there exists $\beta \in (0, \gamma - 1/2)$ such that

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

Remark that the main difference between Assumptions **(H1')** and **(H1)** is the condition $\beta < \gamma - 1/2$ instead of $1/2$. This condition on β in Assumption **(H1')** allows to counterbalance the choice of the steps sequence γ_n used in (5). Section 5 will exemplify recursive estimates of the Hessian to ensure that **(H1')** is indeed verified in practice. We now dispose of all the ingredients to ensure the strong consistency of the estimates given by the WASNA.

Theorem 4.1. *Under Assumptions **(A1a)**, **(A1b)**, **(A2a)**, **(A2b)** and **(H1')**, the iterates of the weighted averaged stochastic Newton algorithm given in (5) and (6) satisfy*

$$\tilde{\theta}_n \xrightarrow[n\to+\infty]{a.s.} \theta \qquad \text{and} \qquad \theta_{n,\tau} \xrightarrow[n\to+\infty]{a.s.} \theta.$$

The proof is given in Supplementary Materials A.1. As expected, averaging does not introduce bias in the estimates. In order to go further and to derive rates of convergence for the WASNA, consider the following assumption:

(H2a') If Assumptions **(A1a)**, **(A1b)**, **(A2a)**, **(A2b)**, and **(H1')** hold,

$$\overline{S}_n \xrightarrow[n\to+\infty]{a.s.} H.$$

This hypothesis is nothing more than the counterpart of Assumption **(H2a)** for the Hessian estimates in the averaged setting. It ensures that the almost sure convergence of the estimates leads to the convergence of $\overline{S}_n$ and $\overline{S}_n^{-1}$. With this at hand, we can theoretically derive the rate of convergence of the *non-averaged* iterates of the WASNA defined in (5).

Theorem 4.2. *Under Assumptions **(A1)**, **(A2a)** **(A2b)**, **(H1')** and **(H2a')** hold, suppose also that there are positive constants $\eta > \frac{1}{\gamma} - 1$ and C_η such that*

$$\mathbb{E}\left[\left\|\nabla_h g\left(X, h\right)\right\|^{2+2\eta}\right] \leq C_\eta \left(1 + \|h - \theta\|^{2+2\eta}\right). \tag{11}$$

Then the iterates of the weighted averaged stochastic Newton algorithm given in (5) *verify that*

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

The proof is given in Supplementary Materials A.4. Theorem 4.2 involves a standard extra assumption in (11), which matches the usual ones for stochastic gradient algorithms [14, 11].

In order to derive a central limit theorem for the WASNA, suppose the following holds:

(H2b') Under Assumptions **(A1)**, **(A2a)**, **(A2b)**, **(H1')**, **(H2a')** and (11), there is a positive constant $p_S > 1/2 - \gamma/2$ such that

$$\left\|\overline{S}_n - S\right\| = O\left(\frac{1}{n^{p_S}}\right) \quad a.s.$$

Being the counterpart of Assumption **(H2b)** in the case of the Stochastic Newton algorithm, **(H2b')** allows to translate a rate of convergence of $\tilde{\theta}_n$ into a rate of convergence of $\overline{S}_n$.

Theorem 4.3. *Under Assumptions* ***(A1)***, ***(A2)***, ***(H1')***, ***(H2a')***, ***(H2b')*** *and inequality* (11)*, the iterates of the weighted averaged stochastic Newton algorithm defined in* (6) *satisfy*

$$\left\|\theta_{\tau,n} - \theta\right\|^2 = O\left(\frac{\ln n}{n}\right) \quad a.s.$$

and

$$\sqrt{n}\left(\theta_{\tau,n} - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\tau^2}{2\tau+\nu} H^{-1}\Sigma H^{-1}\right)$$

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

The proof is given in Supplementary Materials A.5. One could note that Assumption **(H2b')** is required not only to establish the asymptotic normality of the iterates, but also to get the rate of convergence of the WASNA iterates defined in (6). This was not the case for the SNA.

Remark 4.1 (On the asymptotic optimality of weighted versions)**.** *Particularizing Theorem 4.3 to the "usual" averaged stochastic Newton algorithm defined by* (9) *gives*

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

while considering the weighted version (10) *leads to*

$$\sqrt{n}\left(\theta_{\tau,n}-\theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{(1+\omega)^2}{2\omega+1} H^{-1}\Sigma H^{-1}\right).$$

This is striking how using some weighted version of the WASNA can theoretically degrade the asymptotic variance of the constructed estimates. A possible compromise remains, however, to consider for all $n \geq 1$, $\tau_{n+1} = \frac{\ln(n+1)^\omega}{\sum_{k=0}^n \ln(k+1)^\omega}$, *such that*

$$\theta_{n,\tau} = \frac{1}{\sum_{k=0}^n \ln(k+1)^\omega} \sum_{k=0}^n \ln(k+1)^\omega \tilde{\theta}_k.$$

Indeed, with this logarithmic weights choice, one has

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

Here, the non-uniform weighting in play for the WASNA will give more importance to the last estimates $\tilde{\theta}_n$, *while keeping an optimal asymptotic behavior.*

If Assumption **(H2b')** may represent a theoretical lock for the application of Theorem 4.3, it can be by-passed by the following theorem. The idea is to exploit a particular structure of the Hessian estimates, to derive the rate of convergence and the asymptotic normality of the WASNA iterates.

Theorem 4.4. *Suppose that the Hessian estimates* $(\overline{S}_n)_n$ *in the WASNA iteration* (5) *is of the form*

$$\overline{S}_n = \frac{1}{n+1}\left(S_0 + \sum_{k=1}^n \overline{u}_k \overline{\Phi}_k \overline{\Phi}_k^T + \sum_{k=1}^n \frac{c_\beta}{k^\beta} Z_k Z_k^T\right)$$

with S_0 *a symmetric and positive matrix,* $c_\beta \geq 0$, $\beta \in (\gamma - 1/2)$ *and*

$$\overline{u}_k = u_k\left(X_k, \theta_{\tau,k-1}\right) \in \mathbb{R} \qquad \overline{\Phi}_k = \Phi_k\left(X_k, \theta_{\tau,k-1}\right) \in \mathbb{R}^d,$$

and $(Z_k)_k$ *standard Gaussian vectors in dimension* d. *Assume that*

- *for all* $\delta > 0$, *there is a positive constant* C_δ *such that for all* k,

$$\mathbb{E}\left[\left\|\overline{u}_k \overline{\Phi}_k \overline{\Phi}_k^T\right\| \mathbf{1}_{\left\{\|\theta_{\tau,k-1}-\theta\| \leq (\ln k)^{1/2+\delta}\sqrt{\gamma_k}\right\}} \middle| \mathcal{F}_{k-1}\right] \leq C_\delta, \tag{12}$$

- *there is* $\alpha \in (1/2, \tau)$ *and* $\delta > 0$ *such that*

$$\sum_{k\geq 0}(k+1)^{2\alpha}\frac{\tau_{k+1}^2}{\gamma_{k+1}}\frac{(\ln k)^{1+\delta}}{(k+1)^2}$$
$$\cdot\,\mathbb{E}\left[\left\|\overline{u}_k\overline{\Phi}_k\overline{\Phi}_k^T\right\|^2\mathbf{1}_{\left\{\|\theta_{\tau,k-1}-\theta\|\leq(\ln k)^{1/2+\delta}\sqrt{\gamma_k}\right\}}|\mathcal{F}_{k-1}\right] < +\infty \quad a.s. \tag{13}$$

Under the additional Assumptions ***(A1)***, ***(A2a)***, ***(A2b)***, ***(H1')***, ***(H2a')*** *and* (11), *the iterates of the weighted averaged stochastic Newton algorithm defined in* (6) *satisfy*

$$\|\theta_{\tau,n}-\tau\|^2 = O\left(\frac{\ln n}{n}\right) \quad a.s.$$

and

$$\sqrt{n}\left(\theta_{\tau,n}-\theta\right)\xrightarrow[n\to+\infty]{}\mathcal{N}\left(0,\frac{\tau^2}{2\tau+\nu}H^{-1}\Sigma H^{-1}\right)$$

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

The proof is given in Supplementary Materials A.6. The main interest of this theorem is that it enables to get the asymptotic normality of the estimates without having the rate of convergence of the estimates of the Hessian, at the price of a special structure of the latter. More specifically, contrary to [2], no Lipschitz assumption on the functional

$$h \longmapsto \mathbb{E}\left[u_k\left(X_k,h\right)\Phi_k\left(X_k,h\right)\Phi_k\left(X_k,h\right)^T|\mathcal{F}_{k-1}\right]$$

is needed. Note in particular that (13) is verified for all $\alpha < \frac{3-\gamma}{2}$ since

$$\mathbb{E}\left[\left\|\overline{u}_k\overline{\Phi}_k\overline{\Phi}_k^T\right\|^2\mathbf{1}_{\left\{\|\theta_{\tau,k-1}-\theta\|\leq(\ln(k))^{1/2+\delta}\sqrt{\gamma_k}\right\}}|\mathcal{F}_{k-1}\right]$$

is uniformly bounded.

Theorem 4.4 will be particularly useful for the coming practical applications, since this special structure in the Hessian estimates will be met in the case of linear, logistic and softmax regressions.

5 Applications

In this section, for different machine learning tasks, we make the weighted stochastic Newton algorithm explicit, and verify that the associated theoretical guarantees hold. Then, for each application, we perform simulations by comparing second-order online algorithms:

- the stochastic Newton algorithm (SNA) defined in (1) with a step in $1/n$, similar to the one studied in [2] specifically for the logistic regression;
- the stochastic Newton algorithm (SNA) defined in (1) with a step in $1/n^{3/4}$;
- the weighted averaged stochastic Newton algorithm (WASNA) given in (5) and (6) with standard weighting ($\tau_n = 1/(n+1)$);
- the weighted averaged stochastic Newton algorithm (WASNA) given in (5) and (6) with logarithmic weighting ($\tau_n = \frac{(n+1)^\omega}{\sum_{k=0}^{n}(k+1)^\omega}$ and $\omega = 2$);

with first-order online methods:

- the stochastic gradient algorithm (SGD) [14] with step $1/n^{3/4}$;
- the averaged Stochastic Gradient Algorithm (ASGD) [1];

and finally with first-order online algorithms mimicking second-order ones:

- the Adagrad algorithm [7], which uses adaptive step sizes using only first-order information,
- the averaged Adagrad algorithm, with standard weighting.

We illustrate their performances in the case of linear, logistic and softmax regressions, for simple and more complex structured data.

5.1 Least-square regression

5.1.1 Setting & Algorithm

Consider the least square regression model defined by

$$\forall n \geq 1, \quad Y_n = X_n^T \theta + \epsilon_n$$

where $X_n \in \mathbb{R}^d$ is a random features vector, $\theta \in \mathbb{R}^d$ is the parameters vector, and ϵ_n is a zero-mean random variable independent from X_n, and $(X_i, Y_i, \epsilon_i)_{i\geq 1}$ are independent and identically distributed. Then, θ can be seen as the minimizer of the functional $G : \mathbb{R}^d \longrightarrow \mathbb{R}_+$ defined for all parameter $h \in \mathbb{R}^p$ by

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

The Hessian of G at h is given by $\nabla^2 G(h) = \mathbb{E}\left[XX^T\right]$ and a natural estimate is

$$\overline{H}_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$$

whose inverse can be easily updated thanks to the Riccati's formula leading to

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

where $H_n = (n+1)\overline{H}_n$. Then, assuming that $H = \mathbb{E}\left[XX^T\right]$ is positive and that X and ϵ admits 4-th order moments, entails that all assumptions of Section 2 are verified, and by the law of the iterated logarithm,

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

Then the convergence results of the Stochastic Newton algorithm in Section 3.2 hold. Furthermore, if there is $\eta > 0$ verifying $\eta > \frac{1}{\gamma} - 1$ and such that X and ϵ admit moment of order $4 + 4\eta$, the convergence results of the averaged version in Section 4.2 hold.

5.1.2 Simulations

In this section, we fix $d = 10$, and we choose

$$\theta = (-4, -3, 2, 1, 0, 1, 2, 3, 4, 5)^T \in \mathbb{R}^{10},$$

$X \sim \mathcal{N}\left(0, \text{diag}\left(\frac{i^2}{d^2}\right)_{i=1,\ldots,10}\right)$ and $\epsilon \sim \mathcal{N}(0,1)$. Note that in such a case the Hessian associated to this model is equal to $\text{diag}\left(\frac{i^2}{d^2}\right)_{i=1,\ldots,10}$, meaning that the largest eigenvalue is 100 times larger than the smallest one. Therefore, considering stochastic gradient estimates leads to a step sequence which cannot be adapted to each direction. In Figure 1, we monitor the quadratic mean error of the different estimates, for three different type of initializations. In Figure 1, one can see that both averaged Newton methods and the stochastic Newton method with step size of the order $1/n$ outperform all the other algorithms, specially for far intializations (right). The faster convergence of Newton methods or of the Adagrad algorithm compared to the one of standard SGD can be explained by their ability to manage the diagonal structure of the Hessian matrix, with eigenvalues at different scales.

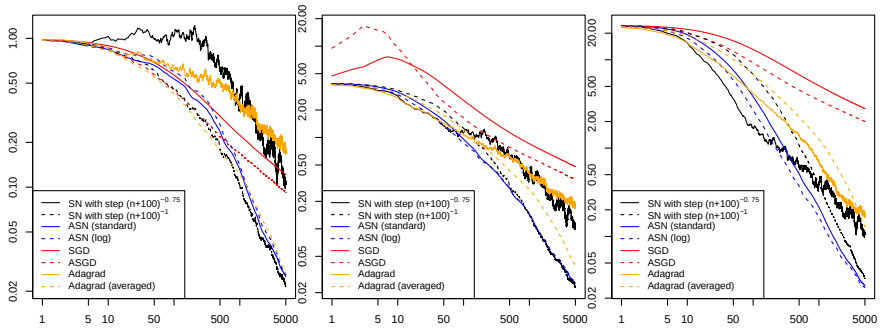

Figure 1: (Linear regression with uncorrelated variables) Mean-squared error of the distance to the optimum θ with respect to the sample size for different initializations: $\theta_0 = \theta + r_0 U$, where U is a uniform random variable on the unit sphere of $\mathbb{R}^d$ with $r_0 = 1$ (left), $r_0 = 2$ (middle) or $r_0 = 5$ (right). Each curve is obtained by an average over 50 different samples (drawing a different initial point each time).

Consider now a more complex covariance structure of the data, such as follows

$$X \sim \mathcal{N}\left(0, A \operatorname{diag}\left(\frac{i^2}{d^2}\right)_{i=1,\dots,d} A^T\right)$$

where A is a random orthogonal matrix. This particular choice of the covariates distribution, by the action of A, allows to consider strong correlations between the coordinates of X. In Figure 2, one can notice that the choice of adaptive step size used in the Adagrad algorithm is no longer sufficient to give the best convergence result in presence of highly-correlated data. In such a case, both averaged Newton algorithms remarkably perform, showing their ability to handle complex second-order structure of the data, and all the more so for bad initializations (right).

5.2 Logistic regression

5.2.1 Setting

We turn out to a binary classification problem: the logistic regression model defined by

$$\forall n \geq 1, \quad Y_n | X_n \sim \mathcal{B}\left(\pi\left(\theta^T X\right)\right)$$

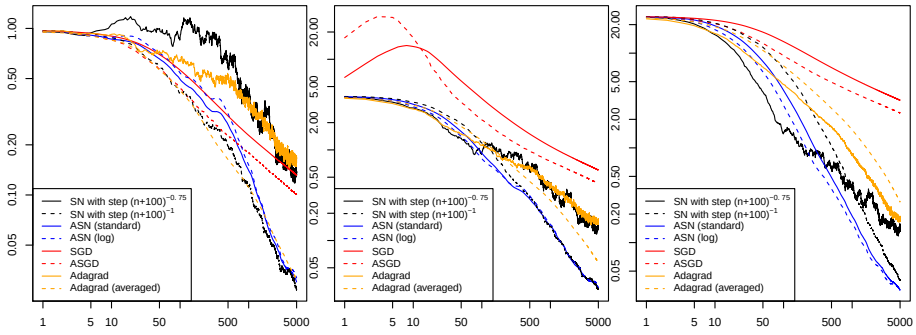

Figure 2: (Linear regression with correlated variables) Mean-squared error of the distance to the optimum θ with respect to the sample size for different initializations: $\theta_0 = \theta + r_0 U$, where U is a uniform random variable on the unit sphere of $\mathbb{R}^d$, with $r_0 = 1$ (left), $r_0 = 2$ (middle) or $r_0 = 5$ (right). Each curve is obtained by an average over 50 different samples (drawing a different initial point each time).

where $X_1, \ldots, X_n, \ldots$ are independent and identically distributed random vectors lying in $\mathbb{R}^d$ and for all $x \in \mathbb{R}$,

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

Due to the intrinsic non-linear feature of this model, this is not clear how the covariance structure of the covariates may affect the training phase, clearly departing from the linear regression setting.

5.2.2 The WASNA

In the particular case of logistic regression, the weighted averaged version the stochastic Newton algorithm in Section 4.1 can be rewritten as:

$$\begin{cases} \overline{a}_{n+1} = \pi\left(\theta_{n,\tau}^T X_{n+1}\right)\left(1 - \pi\left(\theta_{n,\tau}^T X_{n+1}\right)\right) \\ \tilde{\theta}_{n+1} = \tilde{\theta}_n + \gamma_{n+1}\overline{S}_n^{-1} X_{n+1}\left(Y_{n+1} - \pi\left(\tilde{\theta}_n^T X_{n+1}\right)\right) \\ \theta_{n+1,\tau} = (1 - \tau_{n+1})\,\theta_{n,\tau} + \tau_{n+1}\tilde{\theta}_{n+1} \\ S_{n+1}^{-1} = S_n^{-1} - \left(1 + a_{n+1} X_{n+1}^T S_n^{-1} X_{n+1}\right)^{-1} a_{n+1} S_n^{-1} X_{n+1} X_{n+1}^T S_n^{-1} \end{cases} \tag{14}$$

with $\tilde{\theta}_0 = \theta_{0,\tau}$ bounded, $\overline{S}_n^{-1} = (n+1)S_n^{-1}$ and S_0^{-1} symmetric and positive, $\gamma_n = c_\gamma n^{-\gamma}$ with $c_\gamma > 0$ and $\gamma \in (1/2, 1)$, $\tau_n = \frac{\ln(n+1)^\omega}{\sum_{k=0}^n \ln(k+1)^\omega}$, with $\omega \geq 0$,

$$\overline{a}_{n+1} = \pi\left(X_{n+1}^T \theta_{n,\tau}\right)\left(1 - \pi\left(X_{n+1}^T \theta_{n,\tau}\right)\right)$$

and

$$a_{n+1} = \max\left\{\overline{a}_{n+1}, \frac{c_\beta}{(n+1)^\beta}\right\},$$

with $c_\beta > 0$ and $\beta \in (\gamma - 1/2)$.

Choosing $\gamma = 1$, $c_\gamma = 1$ and $\tau_n = 1$ at each iteration leads to the Newton algorithm of Section 3 for the logistic regression, which matches the specific truncated Newton algorithm developed in [2].

We study the convergence rates associated to this instance (14) of the WASNA in the following theorem.

Theorem 5.1. *Suppose that X admits a 4-th order moment and that*

$$\mathbb{E}\left[\pi\left(\theta^T X\right)\left(1 - \pi\left(\theta^T X\right)\right) XX^T\right] \succ 0$$

is a positive matrix. The iterates given by the WASNA defined in (14) verify

$$\left\|\tilde{\theta}_n - \theta\right\|^2 = O\left(\frac{\ln n}{n^\gamma}\right) \quad a.s. \qquad \text{and} \qquad \left\|\theta_{n,\tau} - \theta\right\|^2 = O\left(\frac{\ln n}{n}\right) \quad a.s.$$

Furthermore,

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

Finally,

$$\left\|\overline{S}_n - H\right\|^2 = O\left(\frac{1}{n^{2\beta}}\right) \quad a.s \qquad \text{and} \qquad \left\|\overline{S}_n^{-1} - H^{-1}\right\|^2 = O\left(\frac{1}{n^{2\beta}}\right) \quad a.s.$$

The proof is given in Supplementary Materials A.7, and consists in verifying Assumptions **(A1)**, **(A2a)**, **(A2b)**, **(H1')**, **(H2a')** to get the convergence rate of $\left\|\tilde{\theta}_n - \theta\right\|^2$, then in verifying Assumptions **(A2c)** with Inequalities (12) and (13) to get the convergence rate of $\left\|\theta_{n,\tau} - \theta\right\|^2$ and the central limit theorem. Remark that in the context of the WASNA, we get the rates of convergence as for the direct truncated stochastic Newton algorithm [2], without additional assumptions.

5.2.3 Simulations

The first logistic regression setting that we consider is given in [2] and defined by the model parameter $\theta = (9, 0, 3, 9, 4, 9, 15, 0, 7, 1, 0)^T \in \mathbb{R}^{11}$, with an intercept and standard Gaussian variables, i.e. $X = \left(1, \Phi^T\right)^T$, $\Phi \sim \mathcal{N}(0, I_{10})$. In Figure 3, we display the evolution of the quadratic mean error of the different estimates, for three different initializations. The Newton methods converge faster than online gradient descents, which can be again explained by the Hessian structure of the model: even if the features are standard Gaussian random variables, the non-linearity introduced by the logistic model leads to a covariance structure difficult to apprehend theoretically and numerically by first-order online algorithms. In case of bad initialization (right), the best performances are given by the averaged Adagrad algorithm, the stochastic Newton algorithm with steps in $O(1/n^{3/4})$, closely followed by averaged stochastic Newton algorithms. For the latter, this optimal asymptotic behaviour is enabled in particular by the use of weights giving more importance to the last estimates (being compared to the "standard" averagind Newton algorithm). One can see that in such an example the step choice for the non-averaged Newton algorithm is crucial: choosing a step sequence of the form $1/n$ as in [2] significantly slows down the optimization dynamics, whereas a step choice in $O(1/n^{3/4})$ allows to reach the optimum much more quickly.

Let us now consider a second example, consisting in choosing $\theta \in \mathbb{R}^d$ with all components equal to 1, and $X \sim \mathcal{N}\left(0, A \mathrm{diag}\left(\frac{i^2}{d^2}\right)_{i=1,\dots,d} A^T\right)$ where A is a random orthogonal matrix. The results are displayed in Figure 4. In presence of such highly-structured data, the averaged stochastic Newton algorithms are shown to perform greatly, even more for initializations far from the optimum (middle and right). One could note that in all initial configurations (left to right), the stochastic Newton algorithm with step size in $O(1/n)$ proves to be a worthy competitor, whereas the Adagrad algorithm becomes less and less relevant as the starting point moves away from the solution.

5.3 Softmax regression

In this section, we focus on the softmax regression, which consists in a multi-label classification case.

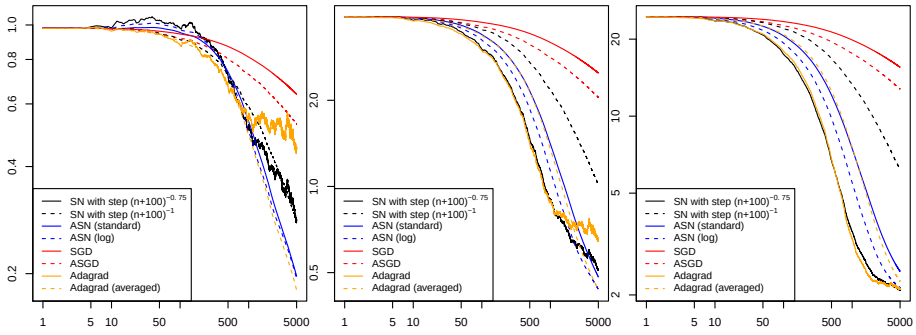

Figure 3: (Logistic regression with standard Gaussian variables) Mean-squared error of the distance to the optimum θ with respect to the sample size for different initializations: $\theta_0 = \theta + r_0 U$, where U is a uniform random variable on the unit sphere of $\mathbb{R}^d$, with $r_0 = 1$ (left), $r_0 = 2$ (middle) or $r_0 = 5$ (right). Each curve is obtained by an average over 50 different samples (drawing a different initial point each time).

5.3.1 Setting

Assume that the number of different classes is K, and that the model parameters are $\theta_1 \in \mathbb{R}^p, \ldots, \theta_K \in \mathbb{R}^p$. Consider the samples $(X_1, Y_1), \ldots, (X_n, Y_n), \ldots$ being i.i.d. random vectors in $\mathbb{R}^d \times [1, \ldots, K]$ with for all $n \geq 1$ and $k = 1, \ldots, K$,

$$\mathbb{P}\left[Y_n = k | X_n\right] = \frac{e^{\theta_k^T X_n}}{\sum_{k'=1}^K e^{\theta_{k'}^T X_n}}.$$

Then, the likelihood can be written as

$$L_n\left(\theta_1, \ldots, \theta_K\right) = \prod_{i=1}^n \frac{\sum_{k'=1}^K e^{\theta_{k'}^T X_i} \mathbf{1}_{Y_i = k}}{\sum_{k'=1}^K e^{\theta_{k'}^T X_n}} = \prod_{i=1}^n \frac{e^{\theta_{Y_i}^T X_i}}{\sum_{k'=1}^K e^{\theta_{k'}^T X_i}},$$

which leads to the following log-likelihood

$$\ell_n\left(\theta_1, \ldots, \theta_K\right) = \sum_{i=1}^n \log\left(\frac{e^{\theta_{Y_i}^T X_i}}{\sum_{k'=1}^K e^{\theta_{k'}^T X_i}}\right).$$

Then, considering the asymptotic objective function, the aim is to minimize the convex function $G : \mathbb{R}^d \times \ldots \times \mathbb{R}^d \longrightarrow \mathbb{R}$ defined for all h by

$$G(h) = -\mathbb{E}\left[\log\left(\frac{e^{h_Y^T X}}{\sum_{k'=1}^K e^{h_{k'}^T X}}\right)\right] =: \mathbb{E}\left[g(X, Y, h)\right].$$

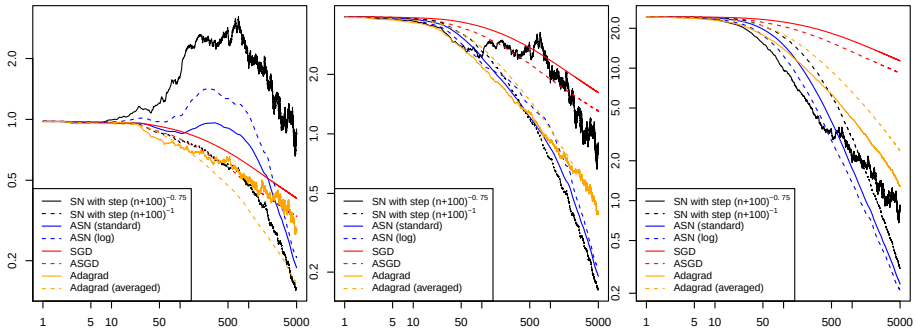

Figure 4: (Logistic regression with correlated Gaussian variables) Mean-squared error of the distance to the optimum θ with respect to the sample size for different initializations: $\theta_0 = \theta + r_0 U$, where U is a uniform random variable on the unit sphere of $\mathbb{R}^d$, with $r_0 = 1$ (left), $r_0 = 2$ (middle) or $r_0 = 5$ (right). Each curve is obtained by an average over 50 different samples (drawing a different initial point each time).

where (X, Y) is a i.i.d. copy of (X_1, Y_1). In order to establish convergence rate for the weighted averaged Newton algorithm in such a setting, we suppose that the following assumptions hold:

(HS1a) The random vector X admits a second order moment.

(HS1b) The random vector X admits a fourth order moment.

Under Assumption **(HS1a)**, the functional G is twice differentiable and for all $h = (h_1, \ldots, h_K) \in \mathbb{R}^d \times \ldots \times \mathbb{R}^d$,

$$\nabla G(h) = \mathbb{E}\left[\begin{pmatrix} X\left(\frac{e^{h_1^T X}}{\sum_{k=1}^K e^{h_k^T X}} - \mathbf{1}_{Y=1}\right) \\ \vdots \\ X\left(\frac{e^{h_K^T X}}{\sum_{k=1}^K e^{h_k^T X}} - \mathbf{1}_{Y=K}\right) \end{pmatrix}\right] = \mathbb{E}\left[\nabla_h g\left(X, Y, h\right)\right],$$

and one can check that $\nabla G(\theta) = 0$. Furthermore, computing second-order derivatives leads to the Hessian, defined for all h by

$$\nabla^2 G(h) = \mathbb{E}\left[\left(\mathrm{diag}\left(\sigma(X,h)\right) - \sigma(X,h)\sigma(X,h)^T\right) \otimes XX^T\right]$$

where $\sigma(X,h) = \left(\frac{e^{h_1^T X}}{\sum_{k=1}^K e^{h_k^T X}}, \ldots, \frac{e^{h_K^T X}}{\sum_{k=1}^K e^{h_k^T X}}\right)^T$, and $\otimes$ denotes the Kronecker product. In addition, suppose that

(HS2) The Hessian of G at θ is positive.

Finally, one can easily check that at $h = \theta$,

$$\nabla^2 G(\theta) = \mathbb{E}\left[\nabla_h g\left(X, Y, \theta\right) \nabla_h g\left(X, Y, \theta\right)^T\right].$$

5.3.2 The SNA

In the case of softmax regression, the stochastic Newton algorithm can be defined for all $n \geq 1$ by

$$\begin{cases} \Phi_{n+1} = \nabla_h g\left(X_{n+1}, Y_{n+1}, \theta_n\right) \\ \theta_{n+1} = \theta_n - \dfrac{1}{n+1} \overline{H}_n^{-1} \nabla_h g\left(X_{n+1}, Y_{n+1}, \theta_n\right) \\ H_{n+\frac{1}{2}}^{-1} = H_n^{-1} - \left(1 + \beta_{n+1} Z_{n+1}^T H_n^{-1} Z_{n+1}\right)^{-1} \beta_{n+1} H_n^{-1} Z_{n+1} Z_{n+1}^T H_n^{-1} \\ H_{n+1} = H_{n+\frac{1}{2}}^{-1} - \left(1 + \Phi_{n+1}^T H_{n+\frac{1}{2}}^{-1} \Phi_{n+1}\right)^{-1} H_{n+\frac{1}{2}}^{-1} \Phi_{n+1} \Phi_{n+1}^T H_{n+\frac{1}{2}}^{-1}, \end{cases} \tag{15}$$

where θ_0 is bounded, $\overline{H}_n = (n+1) H_n^{-1}$, H_0^{-1} is symmetric and positive, $\beta_n = c_\beta n^{-\beta}$, with $c_\beta > 0$ and $\beta \in (0, 1/2)$, and $Z_1, \ldots, Z_{n+1}$ are i.i.d. with $Z_1 \sim \mathcal{N}(0, I_d)$. To our knowledge, this is the first time that a stochastic Newton algorithm is made explicit for the softmax regression problem. The associated convergence guarantees follow.

Theorem 5.2. *Under Assumptions **(HS1a)**, **(HS1b)** and **(HS2)**, the iterates of the stochastic Newton algorithm defined by* (15) *satisfy*

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

and

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

Furthermore,

$$\left\|\overline{H}_n - H\right\|^2 = O\left(\frac{1}{n^{2\beta}}\right) \quad a.s. \qquad \text{and} \qquad \left\|\overline{H}_n^{-1} - H^{-1}\right\|^2 = O\left(\frac{1}{n^{2\beta}}\right) \quad a.s.$$

The proof is given in Supplementary Materials A.8.2.

As far as we know, this is the first theoretical result covering the softmax regression using a stochastic Newton algorithm. In such a setting, the SNA proposes efficient online estimates and the convergence guarantees can be established with weak assumptions on the first moments of the features and about local strong convexity at the optimum only.

5.3.3 The WASNA

Let us now consider the weighted averaged stochastic Newton algorithm defined recursively for all $n \geq 1$ by

$$\begin{cases} \overline{\Phi}_{n+1} = \nabla_h g\left(X_{n+1}, Y_{n+1}, \theta_{n,\tau}\right) \\ \tilde{\theta}_{n+1} = \tilde{\theta}_n - \gamma_{n+1} \overline{S}_n^{-1} \nabla_h g\left(X_{n+1}, Y_{n+1}, \theta_n\right) \\ \theta_{n+1,\tau} = \left(1 - \tau_{n+1}\right) \theta_{n,\tau} + \tau_{n+1} \tilde{\theta}_{n+1} \\ S_{n+\frac{1}{2}}^{-1} = S_n^{-1} - \left(1 + \beta_{n+1} Z_{n+1}^T S_n^{-1} Z_{n+1}\right)^{-1} \beta_{n+1} S_n^{-1} Z_{n+1} Z_{n+1}^T S_n^{-1} \\ S_{n+1}^{-1} = S_{n+\frac{1}{n}}^{-1} - \left(1 + \overline{\Phi}_{n+1}^T S_{n+\frac{1}{n}}^{-1} \overline{\Phi}_{n+1}\right)^{-1} S_{n+\frac{1}{n}}^{-1} \overline{\Phi}_{n+1} \overline{\Phi}_{n+1}^T S_{n+\frac{1}{n}}^{-1}, \end{cases} \tag{16}$$

with $\tilde{\theta}_0 = \theta_{0,\tau}$ bounded, $\overline{S}_n^{-1} = (n+1) S_n^{-1}$ and S_0^{-1} symmetric and positive, $\gamma_n = c_\gamma n^{-\gamma}$ with $c_\gamma > 0$ and $\gamma \in (1/2, 1)$, $\tau_n = \frac{\ln(n+1)^\omega}{\sum_{k=1}^n \ln(k+1)^\omega}$, with $\omega \geq 0$, and $\beta_n = c_\beta n^{-\beta}$ with $c_\beta > 0$ and $\beta \in (\gamma - 1/2)$. Finally, Z_n are i.i.d random vectors with $Z_1 \sim \mathcal{N}(0, I_d)$. The following theorem gives rates of convergence of the weighted averaged stochastic Newton algorithm for the softmax regression.

Theorem 5.3. *Under Assumptions **(HS1a)**, **(HS1b)** and **(HS2)**, the iterates of the weighted averaged stochastic Newton algorithm defined by (16) satisfy*

$$\left\|\tilde{\theta}_n - \theta\right\|^2 = O\left(\frac{\ln n}{n^\gamma}\right) \quad a.s. \qquad \text{and} \qquad \left\|\theta_{n,\tau} - \theta\right\|^2 = O\left(\frac{\ln n}{n}\right) \quad a.s,$$

Furthermore,

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

Finally,

$$\left\|\overline{S}_n - S\right\|^2 = O\left(\frac{1}{n^{2\beta}}\right) \quad a.s. \qquad \text{and} \qquad \left\|\overline{S}_n^{-1} - S^{-1}\right\|^2 = O\left(\frac{1}{n^{2\beta}}\right) \quad a.s.$$

The proof is given in Supplementary Materials A.8.2. This is the first time to our knowledge that a stochastic Newton algorithm is considered in the softmax regression setting, for which convergence results hold under weak assumptions. Note as well that the Riccatti's trick to update the inverse of the Hessian estimates is particularly appropriate, as the dimensionality of θ gets larger.

5.3.4 Numerical experiments

We have performed numerical experiments by running the SNA and the WASNA with different tuning, as well as the previous considered benchmark on simulated data. As the results are similar to the ones obtained in the case of the logistic regression, we discuss them in Supplementary Materials B.

We focus instead on the MNIST[2] real dataset, in order to illustrate the performance of the WASNA in a context of multi-label classification. It consists in 70000 pictures of 28×28 pixels representing handwritten digits recast into vectors of dimension 784. The goal is to predict the digit $Y \in \{0, \dots, 9\}$ represented on each vectorized image $X \in \mathbb{R}^{784}$, where each coordinate gives the contrast (between 0 and 255) of each pixel. This is then a multi-label classification setting with 10 different classes. In a preprocessing step, we normalize the features between 0 and 1 before applying the softmax regression. More formally, the model can be defined for any $k \in \{0, \dots 9\}$ by

$$\mathbb{P}\left[Y = k|X\right] = \frac{e^{\theta_k^T X}}{\sum_{k=0}^{9} e^{\theta_k^T X}}$$

with the parameters $\theta = \left(\theta_0^T, \dots, \theta_9^T\right)^T$ and the normalized features $X \in [0,1]^{784}$. Despite the simplicity of this model which is not really adapted to imaging data, we will see that it will lead to good performances even applied directly on the raw pixels data. The dataset is randomly split into a training set of size 60000 and a test set of size 10000 and the WASNA is run with *default* parameters, i.e. $\gamma = 0.75$, $c_\gamma = 1$, $c'_\gamma = 0$ and $\omega = 2$ on the training set. The constructed estimates of the parameter θ are then used to "read" the digit displayed in pictures of the test set, leading to an overall performance of 88% accurate predictions. For completeness, and to understand which digits are mistaken, we provide the resulting confusion matrix in Figure 5. Remark that Averaged Stochastic gradient algorithms and the

[2] http://yann.lecun.com/exdb/mnist/

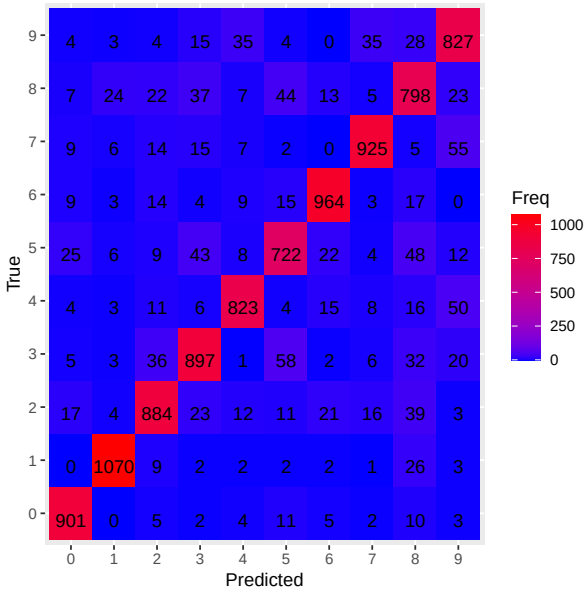

Figure 5: (Softmax regression on the MNIST dataset) Confusion matrix for the predictions given by the default WASNA on a test set of size 10000.

Adagrad one leads to analogous (or slightly better) results. The comparison in terms of accuracy may not be totally fair, as the hyperparameters of the WASNA has not been optimized at all but chosen as default values. This numerical experiment on the MNIST real dataset proves the proposed WASNA to be a second order online method able to tackle large-scale data. And if the number of hyperparameters can be a legitimate delicate point raised by some readers, it should be noted that a default choice however already leads to very good results on a difficult problem such as the classification of images into 10 classes.

6 Conclusion

In this paper, we have given a unified framework to derive stochastic Newton algorithms and their new averaged versions. Under mild assumptions, we have established convergence results, such as almost sure convergence rates and asymptotic efficiency of the constructed estimates. The different proposed methods require the calibration of several hyperparameters, which can be seen as a limitation for the implementation. Nevertheless,

we believe this work paves the way of genuine second-order online algorithms for machine learning. Indeed, a few different and arbitrary choices for these hyperparameters, as highlighted in the numerical experiments, prove the averaged stochastic Newton algorithm to give the best practical results in most of cases, providing more stability and less sensitivity to bad initializations. The Riccatti's trick to directly update the Hessian inverse estimates is also very beneficial in practice to reduce the iteration cost usually attributed to second-order methods. The next step to explore is to optimize the storage needed by such algorithms, that could become a lock for very high-dimensional data.

References

[1] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in neural information processing systems*, pages 773–781, 2013.

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

[3] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[4] Richard H Byrd, Samantha L Hansen, Jorge Nocedal, and Yoram Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

[5] Camille Castera, Jérôme Bolte, Cédric Févotte, and Edouard Pauwels. An inertial newton algorithm for deep learning. *arXiv preprint arXiv:1905.12278*, 2019.

[6] Peggy Cénac, Antoine Godichon-Baggioni, and Bruno Portier. An efficient averaged stochastic Gauss-Newton algorithm for estimating parameters of non linear regressions models. *arXiv preprint arXiv:2006.12920*, 2020.

[7] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[8] Marie Duflo. *Random iterative models*, volume 34 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997. Translated from the 1990 French original by Stephen S. Wilson and revised by the author.

[9] Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic bound of the ruppert-polyak averaging without strong convexity. *arXiv preprint arXiv:1709.03342*, 2017.

[10] Antoine Godichon-Baggioni. Lp and almost sure rates of convergence of averaged stochastic gradient algorithms: locally strongly convex objective. *ESAIM: Probability and Statistics*, 23:841–873, 2019.

[11] Antoine Godichon-Baggioni. Online estimation of the asymptotic variance for averaged stochastic gradient algorithms. *Journal of Statistical Planning and Inference*, 203:1–19, 2019.

[12] Rémi Leluc and François Portier. Towards asymptotic optimality with conditioned stochastic gradient descent. *arXiv preprint arXiv:2006.02745*, 2020.

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

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

[15] Mariane Pelletier. Asymptotic almost sure efficiency of averaged stochastic algorithms. *SIAM J. Control Optim.*, 39(1):49–72, 2000.

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

[17] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

A Proofs

Remark that for the sake of simplicity, in all the proofs we consider that $c_\theta = c'_\gamma = 0$. Indeed, for the cases where $c_\theta \neq 0$ or $c'_\gamma \neq 0$, one can consider

$\overline{H'}_n^{-1} = \frac{n+1}{n+1+c_\theta}\overline{H}_n^{-1}$ and $\overline{S'}_n^{-1} = \frac{(n+1)^\gamma}{((n+1+c_\gamma)^\gamma}$. Then, these estimates have the same asymptotic behaviors as $\overline{H}_n^{-1}$ and $\overline{S}_n^{-1}$.

A.1 Proof of Theorems 3.1 and 4.1

We only give the proof of Theorem 4.1 since taking $c_\gamma = 1$ and $\gamma = 1$ and exchanging $\overline{S}_n$ with $\overline{H}_n$ lead to the proof of Theorem 3.1.

With the help of a Taylor's decomposition of G, and thanks to Assumption **(A2a)**,

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

Then, since $\tilde{\theta}_{n+1} - \tilde{\theta}_n = -\gamma_{n+1}\overline{S}_n^{-1}\nabla_h g\left(X_{n+1}, \tilde{\theta}_n\right)$,

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

Let us define $V_n = G\left(\tilde{\theta}_n\right) - G(\theta)$. Then, we can rewrite the previous inequality as

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

and considering the conditional expectation w.r.t. $\mathcal{F}_n$, since $\tilde{\theta}_n$ and $\overline{S}_n^{-1}$ are $\mathcal{F}_n$-measurable,

$$\begin{aligned} \mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] &\leq V_n - \gamma_{n+1}\nabla G\left(\tilde{\theta}_n\right)^T \overline{S}_n^{-1}\nabla G\left(\tilde{\theta}_n\right) \\ &\quad + \frac{L_{\nabla G}}{2}\gamma_{n+1}^2 \left\|\overline{S}_n^{-1}\right\|_{op}^2 \mathbb{E}\left[\left\|\nabla_h g\left(X_{n+1}, \tilde{\theta}_n\right)\right\|^2 |\mathcal{F}_n\right]. \end{aligned}$$

Then, thanks to Assumption **(A1b)**, it comes

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

Remark that thanks to Assumption **(H1')**,

$$\sum_{n\geq 0} \gamma_{n+1}^2 \left\|\overline{S}_n^{-1}\right\|_{op}^2 < +\infty \quad a.s.$$

Then, since $\overline{S}_n^{-1}$ is positive and applying Robbins-Siegmund Theorem, V_n almost surely converges to a finite random variable and

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

and since, thanks to Assumption **(H1')**,

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

this necessarily implies that $\liminf_n \left\|\nabla G\left(\tilde{\theta}_n\right)\right\| = 0$ almost surely. Since G is strictly convex, this also implies that

$$\liminf_n \left\|\tilde{\theta}_n - \theta\right\| = 0 \quad a.s \qquad \text{and} \qquad \liminf_n V_n = G\left(\tilde{\theta}_n\right) - G(\theta) = 0 \quad a.s,$$

and since V_n converges almost surely to a random variable, this implies that $G\left(\tilde{\theta}_n\right)$ converges almost surely to $G(\theta)$. Finally, by strict convexity, $\tilde{\theta}_n$ converges almost surely to θ.

A.2 Proof of Theorem 3.2

Following the strategy used in [2, Theorem 4.2], the main ideas of the proof are the following. Remark that

$$\begin{aligned}\theta_{n+1}-\theta &= \theta_n-\theta-\frac{1}{n+1}\overline{H}_n^{-1}\nabla G\left(\theta_n\right)+\frac{1}{n+1}\overline{H}_n^{-1}\xi_{n+1} \qquad (17)\\ &= \theta_n-\theta-\frac{1}{n+1}H^{-1}\nabla G\left(\theta_n\right)-\frac{1}{n+1}\left(\overline{H}_n^{-1}-H^{-1}\right)\nabla G\left(\theta_n\right)\\ &+\frac{1}{n+1}\overline{H}_n^{-1}\xi_{n+1}\\ &= \left(1-\frac{1}{n+1}\right)\left(\theta_n-\theta\right)-\frac{1}{n+1}H^{-1}\delta_n \qquad (18)\\ &-\frac{1}{n+1}\left(\overline{H}_n^{-1}-H^{-1}\right)\nabla G\left(\theta_n\right)+\frac{1}{n+1}\overline{H}_n^{-1}\xi_{n+1}\end{aligned}$$

with $\xi_{n+1}=\nabla G\left(\theta_n\right)-\nabla_h g\left(X_{n+1},\theta_n\right)$ and $\delta_n=\nabla G\left(\theta_n\right)-H\left(\theta_n-\theta\right)$ is the remainder term in the Taylor's decomposition of the gradient. Since θ_n and $\overline{H}_n$ are $\mathcal{F}_n$-measurable, and since X_{n+1} is independent from $\mathcal{F}_n$, (ξ_{n+1}) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$. Moreover, inductively, one can check that

$$\theta_{n+1}-\theta=\frac{1}{n+1}\underbrace{\sum_{k=0}^{n}\overline{H}_k^{-1}\xi_{k+1}}_{=:M_{n+1}}\underbrace{-\frac{1}{n+1}\sum_{k=0}^{n}H^{-1}\delta_k-\frac{1}{n}\sum_{k=0}^{n}\left(\overline{H}_k^{-1}-H^{-1}\right)\nabla G\left(\theta_k\right)}_{=:\Delta_n}. \qquad (19)$$

A.2.1 Convergence rate for M_{n+1}

Note that (M_n) is a martingale adapted to the filtration $(\mathcal{F}_n)$ and since $\overline{H}_n^{-1}$ is $\mathcal{F}_n$-measurable,

$$\begin{aligned}\langle M\rangle_{n+1} &= \sum_{k=0}^{n}\overline{H}_k^{-1}\mathbb{E}\left[\xi_{k+1}\xi_{k+1}^T|\mathcal{F}_k\right]\overline{H}_k^{-1}\\ &= \sum_{k=0}^{n}\overline{H}_k^{-1}\mathbb{E}\left[\nabla_h g\left(X_{k+1},\theta\right)\nabla_h g\left(X_{k+1},\theta_k\right)^T|\mathcal{F}_k\right]\overline{H}_k^{-1}\\ &- \sum_{k=0}^{n}\overline{H}_k^{-1}\nabla G\left(\theta_k\right)\nabla G\left(\theta_k\right)^T\overline{H}_k^{-1}\end{aligned}$$

Since $\overline{H}_k^{-1}$ converges almost surely to H^{-1}, since θ_k converges almost surely to θ, and since ∇G is $L_{\nabla G}$ lipschitz,

$$\frac{1}{n+1}\sum_{k=0}^{n}\overline{H}_k^{-1}\nabla G\left(\theta_k\right)\nabla G\left(\theta_k\right)^T\overline{H}_k^{-1} \xrightarrow[n\to+\infty]{a.s} 0.$$

Moreover, Assumption **(A1c)** entails that

$$\frac{1}{n+1}\sum_{k=0}^{n}\overline{H}_k^{-1}\mathbb{E}\left[\nabla_h g\left(X_{k+1},\theta\right)\nabla_h g\left(X_{k+1},\theta_k\right)^T|\mathcal{F}_k\right]\overline{H}_k^{-1} \xrightarrow[n\to+\infty]{a.s} H^{-1}\mathbb{E}\left[\nabla_h g\left(X,\theta\right)\nabla_h g(X,\theta)^T\right]H^{-1}.$$

Setting $\Sigma := \mathbb{E}\left[\nabla_h g\left(X,\theta\right)\nabla_h g(X,\theta)^T\right]$, one has

$$\frac{1}{n+1}\left\langle M\right\rangle_{n+1} \xrightarrow[n\to+\infty]{a.s} H^{-1}\Sigma H^{-1}. \tag{20}$$

Then, applying a law of large numbers for martingales (see [8]), for all $\delta > 0$,

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

and if inequality (4) is verified,

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

A.2.2 Convergence rate for Δ_n

Let us recall that

$$\Delta_n := -\frac{1}{n+1}\sum_{k=1}^{n}H^{-1}\delta_k - \frac{1}{n+1}\sum_{k=1}^{n}\left(\overline{H}_k^{-1}-H^{-1}\right)\nabla G\left(\theta_k\right).$$

Given that θ_k converges almost surely to θ, and since the Hessian is continuous at θ,

$$\left\|\delta_n\right\| = \left\|\int_0^1\left(\nabla^2 G\left(\theta+t\left(\theta_n-\theta\right)-\nabla^2 G(\theta)\right)\right)dt\left(\theta_n-\theta\right)\right\| = o\left(\left\|\theta_n-\theta\right\|\right) \quad a.s.$$

Similarly, since the gradient is $L_{\nabla G}$-lipschitz, one can check that

$$\left\|\left(\overline{H}_k^{-1}-H^{-1}\right)\nabla G\left(\theta_n\right)\right\| = o\left(\left\|\theta_n-\theta\right\|\right) \quad a.s$$

Then, on has

$$\begin{aligned}\|\Delta_{n+1}\| \leq \frac{n}{n+1}\|\Delta_n\| + \frac{1}{n+1}\|\delta_n\| &\leq \frac{n}{n+1}\|\Delta_n\| + \frac{1}{n+1}o\left(\|\theta_n-\theta\|\right) \quad a.s \\ &= \frac{n}{n+1}\|\Delta_n\| + \frac{1}{n+1}\left(\|\Delta_n\| + \|M_n\|\right) \quad a.s \\ &= \left(1 - \frac{1}{n+1} + o\left(\frac{1}{n+1}\right)\right)\|\Delta_n\| + \frac{1}{n+1}o\left(\|M_n\|\right) \quad a.s\end{aligned}$$

and following computation such as Equations (6.27),(6.28),(6.34) and (6.35) in [2], one has

$$\|\Delta_n\|^2 = O\left(\frac{1}{n^2}\|M_{n+1}\|^2\right) \quad a.s,$$

which concludes the proof.

A.3 Proof of Theorem 3.3

In order to derive asymptotic normality of the estimates, start again from Equation (19): the first term will dictate the speed of convergence, while the other terms will collapse. First, since (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$, given (20) and (4), the Central Limit Theorem for martingales (see [8]) reads as follows,

$$\frac{1}{\sqrt{n}}\sum_{k=0}^{n}\overline{H}_k^{-1}\xi_{k+1} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right). \tag{21}$$

It remains to show that the other terms on the right-hand side of (19) are negligible. Under Assumption **(A2c)**, and since θ_n converges almost surely to θ,

$$\|\delta_n\| = \left\|\int_0^1 \left(\nabla^2 G\left(\theta + t\left(\theta_n-\theta\right) - \nabla^2 G(\theta)\right)\right) dt\left(\theta_n-\theta\right)\right\| = O\left(\|\theta_n-\theta\|^2\right) \quad a.s.$$

Theorem 3.2 coupled with the Toeplitz lemma imply in turn

$$\frac{1}{n+1}\left\|\sum_{k=0}^{n}\delta_k\right\| = o\left(\frac{(\ln n)^{2+\delta}}{n}\right) \quad a.s \tag{22}$$

In the same way, the gradient being $L_{\nabla G}$-lipschitz and under Assumption **(H2b)**, one has

$$\frac{1}{n+1}\left\|\sum_{k=0}^{n}\left(\overline{H}_k^{-1} - H^{-1}\right)\nabla G\left(\theta_k\right)\right\| = o\left(\frac{(\ln n)^{2+\delta}}{n^{\min\left\{\frac{1}{2}+p_H,1\right\}}}\right) \quad a.s. \tag{23}$$

The rates obtained in (22) and (23) are negligible compared to the one in (21), which leads to the desired conclusion.

A.4 Proof of Theorem 4.2

Considering the Weighted Averaged Stochastic Newton Algorithm defined by (6), Inequality (18) can be adapted such as

$$\begin{aligned}\tilde{\theta}_{n+1} - \theta = (1 - \gamma_{n+1})\left(\tilde{\theta}_n - \theta\right) - \gamma_{n+1} H^{-1} \tilde{\delta}_n \\ - \gamma_{n+1}\left(\overline{S}_n^{-1} - H^{-1}\right) \nabla G\left(\tilde{\theta}_n\right) + \gamma_{n+1} \overline{S}_n^{-1} \tilde{\xi}_{n+1}\end{aligned} \tag{24}$$

with $\tilde{\xi}_{n+1} = \nabla G\left(\tilde{\theta}_n\right) - \nabla_h g\left(X_{n+1}, \tilde{\theta}_n\right)$ and $\tilde{\delta}_n = \nabla G\left(\tilde{\theta}_n\right) - H\left(\tilde{\theta}_n - \theta\right)$ is the remainder term of the Taylor's expansion of the gradient. Since $\tilde{\theta}_n, S_n$ are $\mathcal{F}_n$-measurable and X_{n+1} is independent from $\mathcal{F}_n$, $\left(\tilde{\xi}_{n+1}\right)$ is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$. Moreover, inductively, one can check that

$$\begin{aligned}\tilde{\theta}_n - \theta =& \beta_{n,0}\left(\tilde{\theta}_0 - \theta\right) \\ & \underbrace{- \sum_{k=0}^{n-1} \beta_{n,k+1}\gamma_{k+1} H^{-1}\tilde{\delta}_k - \sum_{k=0}^{n-1} \beta_{n,k+1}\gamma_{k+1}\left(\overline{S}_k^{-1} - H^{-1}\right)\nabla G\left(\tilde{\theta}_k\right)}_{:=\tilde{\Delta}_n} \\ & + \underbrace{\sum_{k=0}^{n-1} \beta_{n,k+1}\gamma_{k+1}\overline{S}_k^{-1}\tilde{\xi}_{k+1}}_{:=\tilde{M}_n}\end{aligned} \tag{25}$$

with for all $k, n \geq 0$ such that $k \leq n$, $\beta_{n,k} = \prod_{j=k+1}^{n}\left(1 - \gamma_j\right)$ and $\beta_{n,n} = 1$. Focusing on the last term of (25), applying Theorem 6.1 in [6] and thanks to Inequality (11)

$$\left\| \sum_{k=0}^{n-1} \beta_{n,k+1}\gamma_{k+1}\overline{S}_k^{-1}\tilde{\xi}_{k+1} \right\|^2 = O\left(\frac{\ln n}{n^{\gamma}}\right) \quad a.s. \tag{26}$$

Furthermore, one can check that

$$|\beta_{n,0}| = O\left(\exp\left(-\frac{c_\gamma}{1-\gamma} n^{1-\gamma}\right)\right),$$

and the term $\beta_{n,0}\left(\tilde{\theta}_0 - \theta\right)$ is negligible compared to (26). Considering now $\tilde{\Delta}_n$ and following the proof of Theorem 3.2, one can check that

$$\left\|\tilde{\delta}_n\right\| = o\left(\left\|\tilde{\theta}_n - \theta\right\|\right) \quad a.s \qquad \text{and} \qquad \left\|\left(\overline{S}_n^{-1} - H^{-1}\right)\nabla G\left(\tilde{\theta}_n\right)\right\| = o\left(\left\|\tilde{\theta}_n - \theta\right\|\right) \quad a.s$$

Let n_0 such that for all $n \geq n_0$, $\gamma_n \leq 1$. Then, for all $n \geq n_0$,

$$\begin{aligned}
\left\|\tilde{\Delta}_{n+1}\right\| &\leq (1-\gamma_{n+1})\left\|\tilde{\Delta}_n\right\| + \gamma_{n+1}\left(\left\|H^{-1}\tilde{\delta}_n\right\| + \left\|\left(\overline{S}_n^{-1} - H^{-1}\right)\nabla G\left(\tilde{\theta}_n\right)\right\|\right) \\
&= (1-\gamma_{n+1})\left\|\tilde{\Delta}_n\right\| + o\left(\gamma_{n+1}\left\|\tilde{\theta}_n - \theta\right\|\right) \quad a.s \\
&= (1-\gamma_{n+1})\left\|\tilde{\Delta}_n\right\| + o\left(\gamma_{n+1}\left\|\tilde{M}_n + \beta_{n,0}\left(\tilde{\theta}_0 - \theta\right)\right\| + \gamma_{n+1}\left\|\tilde{\Delta}_n\right\|\right) \quad a.s \\
&= (1-\gamma_{n+1}+o(\gamma_{n+1}))\left\|\tilde{\Delta}_n\right\| \\
&\quad + o\left(\gamma_{n+1}\left(\sqrt{\frac{\ln n}{n^\gamma}} + \exp\left(-\frac{c_\gamma}{1-\gamma}n^{1-\gamma}\right)\right)\right) \quad a.s
\end{aligned}$$

and applying a lemma of stabilization [8] or with analogous calculus to the ones of the proof of Lemma 3 in [14], it comes

$$\left\|\tilde{\Delta}_n\right\|^2 = O\left(\frac{\ln n}{n^\gamma}\right) \quad a.s.$$

A.5 Proof of Theorem 4.3

Remark that for all $n \geq 0$, $\theta_{n,\tau}$ can be written as

$$\theta_{n,\tau} - \theta = \underbrace{\prod_{k=j}^{n}(1-\tau_j)}_{=:\kappa_{n,0}}\left(\theta_{0,\tau} - \theta\right) + \sum_{k=0}^{n-1}\underbrace{\prod_{j=k+1}^{n}(1-\tau_j)}_{=:\kappa_{n,k}}\tau_{k+1}\left(\tilde{\theta}_k - \theta\right). \tag{27}$$

with $\kappa_{n,n} = 1$. Remark also that (24) can be written as

$$\tilde{\theta}_n - \theta = \frac{\tilde{\theta}_n - \theta - \left(\tilde{\theta}_{n+1} - \theta\right)}{\gamma_{n+1}} - H^{-1}\tilde{\delta}_n - \left(\overline{S}_n^{-1} - H^{-1}\right)\nabla G\left(\tilde{\theta}_n\right) + \overline{S}_n^{-1}\tilde{\xi}_{n+1} \tag{28}$$

Then, (27) can be written as

$$\begin{aligned}
\theta_{n,\tau} - \theta = \kappa_{n,0}\left(\theta_{0,\tau} - \theta\right) &+ \sum_{k=1}^{n}\kappa_{n,k}\tau_{k+1}\frac{\tilde{\theta}_k - \theta - \left(\tilde{\theta}_{k+1} - \theta\right)}{\gamma_{k+1}} - H^{-1}\sum_{k=0}^{n-1}\kappa_{n,k}\tau_{k+1}\tilde{\delta}_k \\
&- \sum_{k=0}^{n-1}\kappa_{n,k}\tau_{k+1}\left(\overline{S}_k^{-1} - H^{-1}\right)\nabla G\left(\tilde{\theta}_k\right) + \sum_{k=0}^{n-1}\kappa_{n,k}\tau_{k+1}\overline{S}_k^{-1}\tilde{\xi}_{k+1}
\end{aligned} \tag{29}$$

The rest of the proof consists in establishing the rate of convergence of each term on the right-hand side (29).

Bounding $\left\|\prod_{k=1}^{n}(1-\tau_k)(\theta_{0,\tau}-\theta)\right\|$: Since $n\tau_n$ converges to $\tau > 1/2$, there is a rank n_τ such that for all $n \geq n_\tau$, $0 \leq \tau_n \leq 1$, so that for all $n \geq n_\tau$,

$$\begin{aligned}\prod_{k=1}^{n}|1-\tau_k| &\leq \prod_{k=1}^{n_\tau-1}|1-\tau_k|\exp\left(\sum_{k=n_\tau}^{n}1-\tau_k\right)\\ &\leq \prod_{k=1}^{n_\tau-1}|1-\tau_k|\exp\left(-\sum_{k=n_\tau}^{n}\tau_k\right) = O\left(\frac{1}{n^\tau}\right). \qquad (30)\end{aligned}$$

Then

$$\left\|\prod_{k=1}^{n}(1-\tau_k)(\theta_{0,\tau}-\theta)\right\| = O\left(\frac{1}{n^\tau}\right) \quad a.s.$$

and this term is negligible since $\tau > 1/2$.

Bounding $\sum_{k=0}^{n-1}\kappa_{n,k}\tau_{k+1}\tilde{\delta}_k$: Since

$$\left\|\tilde{\delta}_n\right\| = O\left(\left\|\tilde{\theta}_n-\theta\right\|^2\right) \quad a.s,$$

and with the help of Theorem 4.2, for all $\delta > 0$, there is a positive random variable B_δ such that

$$\left\|\sum_{k=0}^{n-1}\kappa_{n,k}\tau_{k+1}\tilde{\delta}_k\right\| \leq B_\delta\sum_{k=1}^{n}|\kappa_{n,k}|\,\tau_{k+1}\frac{(\ln(+1))^{1+\delta}}{(k+1)^\gamma} \quad a.s.$$

Then, since the sequence $\left(\frac{(\ln n)^{1+\delta}}{n^\gamma}\right)$ is in $\mathcal{GS}(-\gamma)$, applying Lemma A.1,

$$\left\|\sum_{k=0}^{n-1}\kappa_{n,k}\tau_{k+1}\tilde{\delta}_k\right\| = o\left(\frac{(\ln n)^{1+\delta}}{n^\gamma}\right) \quad a.s.$$

Bounding $\sum_{k=0}^{n-1}\kappa_{n,k}\tau_{k+1}\left(\overline{S}_k^{-1}-H^{-1}\right)\nabla G\left(\tilde{\theta}_k\right)$: Thanks to Assumption **(H2b')** and since the gradient of G is $L_{\nabla G}$ lipshitz, for all $\delta > 0$, there is a positive random variable B'_δ such that

$$\left\|\sum_{k=0}^{n-1}\kappa_{n,k}\tau_k\left(\overline{S}_k^{-1}-H^{-1}\right)\nabla G\left(\tilde{\theta}_k\right)\right\| \leq B'_\delta\sum_{k=0}^{n-1}|\kappa_{n,k}|\,\tau_{k+1}\frac{(\ln(k+1))^{1/2+\delta}}{(k+1)^{p_S+\gamma/2}} \quad a.s.$$

Then, since the sequence $\left(\frac{(\ln n)^{1/2+\delta}}{n^{p_S+\gamma/2}}\right)$ is in $\mathcal{GS}(-p_S-\gamma/2)$, applying Lemma A.1,

$$\left\|\sum_{k=0}^{n-1}\kappa_{n,k}\tau_{k+1}\left(\overline{S}_k^{-1}-H^{-1}\right)\nabla G\left(\tilde{\theta}_k\right)\right\| = o\left(\frac{(\ln n)^{1/2+\delta}}{n^{p_S+\gamma/2}}\right) \quad a.s.$$

Bounding $\sum_{k=0}^{n-1} \kappa_{n,k}\tau_{k+1}\frac{\tilde{\theta}_k-\theta-\left(\tilde{\theta}_{k+1}-\theta\right)}{\gamma_{k+1}}$: Applying an Abel's transform, one can check that

$$\begin{aligned}\sum_{k=0}^{n-1} \kappa_{n,k}\tau_{k+1}&\frac{\tilde{\theta}_k-\theta-\left(\tilde{\theta}_{k+1}-\theta\right)}{\gamma_{k+1}}\\ &=\frac{\kappa_{n,0}\tau_1}{\gamma_1}\left(\tilde{\theta}_0-\theta\right)-\frac{\tau_n}{\gamma_n}\left(\tilde{\theta}_n-\theta\right)\\ &\quad+\sum_{k=1}^{n-1}\kappa_{n,k}\tau_{k+1}\gamma_{k+1}^{-1}\left(1-\left(1-\tau_{k+1}\right)\frac{\tau_k\gamma_k^{-1}}{\tau_{k+1}\gamma_{k+1}^{-1}}\right)\left(\tilde{\theta}_k-\theta\right)\end{aligned}$$

Thanks to (30),

$$\frac{\kappa_{n,1}\tau_1}{\gamma_1}\left\|\tilde{\theta}_0-\theta\right\| = O\left(\frac{1}{n^{\tau}}\right) \quad a.s.$$

Furthermore, using Theorem 4.2,

$$\frac{\tau_{n+1}}{\gamma_{n+1}}\left\|\tilde{\theta}_{n+1}-\theta\right\| = O\left(\frac{\sqrt{\ln n}}{n^{1-\gamma/2}}\right) \quad a.s.$$

Finally, since τ_n is in $\mathcal{GS}(\nu)$,

$$\begin{aligned}1-\left(1-\tau_{n+1}\right)\frac{\tau_n\gamma_n^{-1}}{\tau_n\gamma_{n+1}^{-1}} =&1+\underbrace{\left(1-\tau_{n+1}\right)}_{=1-\frac{\tau}{n}+o\left(\frac{1}{n}\right)}\underbrace{\left(1-\frac{\tau_n}{\tau_{n+1}}\right)}_{=\frac{\nu}{n}+o\left(\frac{1}{n}\right)}\underbrace{\frac{\gamma_n^{-1}}{\gamma_{n+1}^{-1}}}_{=1+\frac{\gamma}{n}+o\left(\frac{1}{n}\right)}\\ &-\underbrace{\left(1-\tau_{n+1}\right)}_{=1-\frac{\tau}{n}+o\left(\frac{1}{n}\right)}\underbrace{\frac{\gamma_n^{-1}}{\gamma_{n+1}^{-1}}}_{=1+\frac{\gamma}{n}+o\left(\frac{1}{n}\right)}\\ =&\frac{2\nu-\gamma}{n}+o\left(\frac{1}{n}\right)\end{aligned}$$

and applying Theorem 4.2, for all $\delta > 0$,

$$\begin{aligned}\sum_{k=1}^{n-1}\kappa_{n,k}\tau_{k+1}\gamma_{k+1}^{-1}&\left(1-\left(1-\tau_{k+1}\right)\frac{\tau_k\gamma_k^{-1}}{\tau_{k+1}\gamma_{k+1}^{-1}}\right)\left\|\tilde{\theta}_k-\theta\right\|\\ &=O\left(\sum_{k=1}^{n}\kappa_{n,k}\tau_{k+1}\frac{(\ln k)^{1+\delta}}{k^{1-\gamma/2}}\right) \quad a.s.\end{aligned}$$

Then, since $\left(\frac{(\ln n)^{1/2+\delta}}{n^{1-\gamma/2}}\right)$ is in $\mathcal{GS}\left(-1+\gamma/2\right)$ and applying Lemma A.1

$$\sum_{k=1}^{n-1}\kappa_{n,k}\tau_{k+1}\gamma_{k+1}^{-1}\left(1-\left(1-\tau_{k+1}\right)\frac{\tau_k\gamma_k^{-1}}{\tau_{k+1}\gamma_{k+1}^{-1}}\right)\left\|\tilde{\theta}_k-\theta\right\|=o\left(\frac{(\ln n)^{1/2+\delta}}{n^{1-\gamma/2}}\right)\quad a.s$$

so that

$$\sum_{k=1}^{n-1}\kappa_{n,k}\tau_{k+1}\frac{\tilde{\theta}_k-\theta-\left(\tilde{\theta}_{k+1}-\theta\right)}{\gamma_{k+1}}=o\left(\frac{(\ln n)^{1/2+\delta}}{n^{1-\gamma/2}}\right)\quad a.s.$$

Bounding $\sum_{k=0}^{n-1}\kappa_{n,k}\tau_{k+1}\overline{S}_k^{-1}\tilde{\xi}_{k+1}$: First, remark that this term can be written as

$$\sum_{k=0}^{n-1}\kappa_{n,k}\tau_{k+1}\overline{S}_k^{-1}\tilde{\xi}_{k+1}=\kappa_{n,0}\underbrace{\sum_{k=0}^{n-1}\prod_{j=1}^{k}\left(1-\tau_j\right)^{-1}\tau_{k+1}\overline{S}_k^{-1}\tilde{\xi}_{k+1}}_{\overline{M}_n}$$

where $\left(\overline{M}_n\right)$ is a martingale term with respect to the filtration $\left(\mathcal{F}_n\right)$, and

$$\langle M\rangle_n=\sum_{k=0}^{n-1}\left(\prod_{j=1}^{k}\left(1-\tau_j\right)^{-2}\right)\tau_{k+1}^2\overline{S}_k^{-1}\mathbb{E}\left[\tilde{\xi}_{k+1}\tilde{\xi}_{k+1}^T|\mathcal{F}_k\right]\overline{S}_k^{-1},$$

which can be written as

$$\begin{aligned}\langle M\rangle_n=&\sum_{k=0}^{n-1}\left(\prod_{j=1}^{k}\left(1-\tau_j\right)^{-2}\right)\tau_{k+1}^2\overline{S}_k^{-1}\mathbb{E}\left[\nabla_h g\left(X_{k+1},\tilde{\theta}_k\right)\nabla_h g\left(X_{k+1},\tilde{\theta}_k\right)^T|\mathcal{F}_k\right]\overline{S}_k^{-1}\\&-\sum_{k=0}^{n-1}\left(\prod_{j=1}^{k}\left(1-\tau_j\right)^{-2}\right)\tau_{k+1}^2\overline{S}_k^{-1}\nabla G\left(\tilde{\theta}_k\right)\nabla G\left(\tilde{\theta}_k\right)^T\overline{S}_k^{-1}\end{aligned}$$

Since $\nabla G\left(\tilde{\theta}_n\right)$ and $\overline{S}_n^{-1}$ converge almost surely respectively to 0 and H^{-1} and applying Lemma A.1 (third equality),

$$\kappa_{n,0}^2\tau_n^{-1}\left\|\sum_{k=0}^{n-1}\left(\prod_{j=1}^{k}\left(1-\tau_j\right)^{-2}\right)\tau_{k+1}^2\overline{S}_k^{-1}\nabla G\left(\tilde{\theta}_k\right)\nabla G\left(\tilde{\theta}_k\right)^T\overline{S}_k^{-1}\right\|\xrightarrow[n\to+\infty]{a.s}0.$$

Furthermore, since $\tilde{\theta}_k$ converges almost surely to θ, since $\overline{S}_k^{-1}$ converges almost surely to H^{-1} and thanks to assumption **(A1c)**, there is a sequence

of random matrices R_n converging to 0 such that

$$\begin{aligned}\overline{S}_k^{-1}\mathbb{E}\left[\nabla_h g\left(X_{k+1},\tilde{\theta}_k\right)\nabla_h g\left(X_{k+1},\tilde{\theta}_k\right)^T|\mathcal{F}_k\right]\overline{S}_k^{-1}\\ = H^{-1}\underbrace{\mathbb{E}\left[\nabla_h g\left(X,\theta\right)\nabla_h g\left(X,\theta\right)^T\right]}_{:=\Sigma}H^{-1}+R_k.\end{aligned}$$

Applying Lemma A.1 (third equality),

$$\kappa_{n,0}^2\tau_n^{-1}\left\|\sum_{k=0}^{n-1}\left(\prod_{j=1}^{k}\left(1-\tau_j\right)^{-2}\right)\tau_{k+1}^2R_k\right\|\xrightarrow[n\to+\infty]{a.s}0.$$

Finally, applying Lemma A.1 (second equality),

$$\kappa_{n,0}^2\tau_n^{-1}\sum_{k=0}^{n-1}\prod_{j=1}^{k}\left(1-\tau_j\right)^{-2}\tau_{k+1}^2H^{-1}\Sigma H^{-1}\xrightarrow[n\to+\infty]{}\frac{\tau}{2\tau+\nu}H^{-1}\Sigma H^{-1},$$

and then,

$$\kappa_{n,0}^2\tau_n^{-1}\left\langle M\right\rangle_n\xrightarrow[n\to+\infty]{p.s}H^{-1}\Sigma H^{-1}.$$

Then, thanks to (11) and with the help of the law of large numbers for martingales [8], it comes

$$\|M_n\|^2=O\left(\ln(n)\frac{\tau_n}{\kappa_{n,0}^2}\right)\quad a.s$$

which can be written, since $n\tau_n$ converges to τ, as

$$\|\kappa_{n,0}M_n\|^2=O\left(\frac{\ln n}{n}\right)\quad a.s.$$

Furthermore, thanks to (11), the Lindeberg condition for the Central Limit Theorem for martingales is verified. Indeed, by Hölder's inequality, for all $\epsilon>0$,

$$\begin{aligned}L_n&:=\kappa_{n,0}^2\tau_n^{-1}\sum_{k=1}^{n-1}\prod_{j=1}^{k}\kappa_{k,0}^{-2}\tau_{k+1}^2\mathbb{E}\left[\left\|\overline{H}_k^{-1}\tilde{\xi}_{k+1}\right\|^2\mathbf{1}_{\kappa_{k,0}^{-1}\tau_{k+1}\left\|\overline{H}_k^{-1}\tilde{\xi}_{k+1}\right\|\geq\epsilon\kappa_{n,0}^{-1}\tau_n^{1/2}}\middle|\mathcal{F}_k\right]\\ &\leq\frac{\kappa_{n,0}^2}{\tau_n}\sum_{k=1}^{n-1}\frac{\tau_{k+1}^2}{\kappa_{k,0}^2}\left\|\overline{H}_k^{-1}\right\|_{op}^2\left(\mathbb{E}\left[\left\|\tilde{\xi}_{k+1}\right\|^{2+2\eta}|\mathcal{F}_k\right]\right)^{\frac{1}{1+\eta}}\left(\mathbb{P}\left[\frac{\tau_{k+1}}{\kappa_{k,0}}\left\|\overline{H}_k^{-1}\tilde{\xi}_{k+1}\right\|\geq\frac{\epsilon\sqrt{\tau_n}}{\kappa_{n,0}}|\mathcal{F}_k\right]\right)^{\frac{\eta}{1+\eta}}.\end{aligned}$$

Then, applying Markov inequality,

$$L_n \leq \frac{\kappa_{n,0}^2}{\tau_n} \sum_{k=0}^{n-1} \frac{\tau_{k+1}^2}{\kappa_{k,0}^2} \left\| \overline{H}_k^{-1} \right\|_{op}^2 \left(\mathbb{E}\left[\left\| \tilde{\xi}_{k+1} \right\|^{2+2\eta} | \mathcal{F}_k \right] \right)^{\frac{1}{1+\eta}} \left\| \overline{H}_k^{-1} \right\|_{op}^{2\eta} \left(\mathbb{E}\left[\left\| \tilde{\xi}_{k+1} \right\|^{2+2\eta} | \mathcal{F}_k \right] \right)^{\frac{\eta}{1+\eta}} \frac{\kappa_{n,0}^{2\eta} \tau_{k+1}^{2\eta}}{\epsilon^2 \tau_n^{\eta} \kappa_{k,0}^{2\eta}}$$

$$= \frac{1}{\epsilon^2} \frac{\kappa_{n,0}^{2+2\eta}}{\tau_n^{1+\eta}} \sum_{k=0}^{n-1} \frac{\tau_{k+1}^{2+2\eta}}{\kappa_{k,0}^{2+2\eta}} \left\| \overline{H}_k^{-1} \right\|_{op}^{2+2\eta} \mathbb{E}\left[\left\| \tilde{\xi}_{k+1} \right\|^{2+2\eta} | \mathcal{F}_k \right].$$

Furthermore, thanks to Theorem 4.2, inequality (11) and Assumption **(H2a')**, there is a positive random variable B such that $\left\| \overline{H}_k^{-1} \right\|_{op}^{2+2\eta} \mathbb{E}\left[\left\| \tilde{\xi}_{k+1} \right\|^{2+2\eta} | \mathcal{F}_k \right] \leq B$ so that

$$L_n \leq \frac{B}{\epsilon^2} \frac{\kappa_{n,0}^{2+2\eta}}{\tau_n^{1+\eta}} \sum_{k=0}^{n-1} \frac{\tau_{k+1}^{2+2\eta}}{\kappa_{k,0}^{2+2\eta}}$$

and applying Lemma A.1, this term converges to 0. Then the Lindeberg condition is verified and it comes

$$\kappa_{n,0} \tau_n^{-1/2} M_n \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\tau}{2\tau + \nu} H^{-1} \Sigma H^{-1} \right),$$

since $n\tau_n$ converges to τ, this can be written as

$$\sqrt{n} \kappa_{n,0} M_n \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\tau^2}{2\tau + \nu} H^{-1} \Sigma H^{-1} \right),$$

which concludes the proof.

A.6 Proof of Theorem 4.4

Let us recall that $\tilde{\theta}_{n+1}$ can be written as

$$\tilde{\theta}_{n+1} - \theta = \tilde{\theta}_n - \theta - \gamma_{n+1} \overline{S}_n^{-1} \nabla G\left(\tilde{\theta}_n \right) + \gamma_{n+1} \overline{S}_n^{-1} \tilde{\xi}_{n+1}.$$

Then, linearizing the gradient,

$$\tilde{\theta}_{n+1} - \theta = \left(I_d - \gamma_{n+1} \overline{S}_n^{-1} H \right) \left(\tilde{\theta}_n - \theta \right) - \gamma_{n+1} \overline{S}_n^{-1} \tilde{\delta}_n + \gamma_{n+1} \overline{S}_n^{-1} \tilde{\xi}_{n+1},$$

which can be written as

$$\tilde{\theta}_n - \theta = H^{-1} \overline{S}_n \frac{\left(\tilde{\theta}_n - \theta \right) - \left(\tilde{\theta}_{n+1} - \theta \right)}{\gamma_{n+1}} + H^{-1} \tilde{\delta}_n + H^{-1} \tilde{\xi}_{n+1} \qquad (31)$$

Then, thanks to decomposition (27),

$$\begin{aligned}\theta_{n,\tau} - \theta =& \kappa_{n,0}\left(\theta_{0,\tau} - \theta\right) + H^{-1}\sum_{k=0}^{n-1}\kappa_{n,k}\tau_{k+1}\overline{S}_k\frac{\left(\tilde{\theta}_k - \theta\right) - \left(\tilde{\theta}_{k+1} - \theta\right)}{\gamma_{k+1}} \\ & - H^{-1}\sum_{k=0}^{n-1}\kappa_{n,k}\tau_{k+1}\tilde{\delta}_k + H^{-1}\sum_{k=0}^{n-1}\kappa_{n,k}\tau_{k+1}\tilde{\xi}_{k+1}. \end{aligned} \tag{32}$$

Note that the rate of convergence of the first and third terms on the right-hand side of previous equality are given in the proof of Theorem 4.3. For the martingale term, with analogous calculus as the ones in the proof of Theorem 4.3, one can check that

$$\left\| H^{-1}\sum_{k=0}^{n-1}\kappa_{n,k}\tau_{k+1}\tilde{\xi}_{k+1}\right\|^2 = O\left(\frac{\ln n}{n}\right) \quad a.s$$

and

$$\sqrt{n}\left(H^{-1}\sum_{k=0}^{n-1}\kappa_{n,k}\tau_{k+1}\tilde{\xi}_{k+1}\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathbb{N}\left(0, \frac{\tau^2}{2\tau+\nu}H^{-1}\Sigma H^{-1}\right).$$

In order to conclude the proof, a rate of convergence for the second term on the right-hand side of equality (32) remains to be given. Applying an Abel's transform,

$$\begin{aligned}\sum_{k=0}^{n-1}\kappa_{n,k}\tau_{k+1}\overline{S}_k\frac{\left(\tilde{\theta}_k - \theta\right) - \left(\tilde{\theta}_{k+1} - \theta\right)}{\gamma_{k+1}} =& \frac{\kappa_{n,0}\tau_1}{\gamma_1}\overline{S}_0^{-1}\left(\tilde{\theta}_0 - \theta\right) - \frac{\tau_n}{\gamma_n}\overline{S}_{n-1}^{-1}\left(\tilde{\theta}_n - \theta\right) \\ & \underbrace{- \sum_{k=1}^{n-1}\left(\kappa_{n,k-1}\frac{\tau_k}{\gamma_k}\overline{S}_{k-1} - \kappa_{n,k}\frac{\tau_{k+1}}{\gamma_{k+1}}\overline{S}_k\right)\left(\tilde{\theta}_k - \theta\right)}_{:=R_n}\end{aligned}$$

The rate of convergence of the two first term on the right hand side of previous equality are given (since $\overline{S}_n$ converges almost surely to H^{-1}) in the proof of Theorem 4.3. Remark that since $\overline{S}_k = \overline{S}_{k-1} + \frac{1}{k+1}\left(\overline{u}_k\overline{\Phi}_k\overline{\Phi}_k^T - \overline{S}_{k-1}\right)$,

R_n can be rewritten as

$$R_n = \underbrace{\sum_{k=1}^{n-1} \left(\kappa_{n,k-1} \frac{\tau_k}{\gamma_k} - \kappa_{n,k} \frac{\tau_{k+1}}{\gamma_{k+1}} \right) \overline{S}_{k-1} \left(\tilde{\theta}_k - \theta \right)}_{:=R_{1,n}}$$
$$+ \underbrace{\sum_{k=1}^{n-1} \kappa_{n,k} \frac{\tau_{k+1}}{\gamma_{k+1}} \frac{1}{k+1} \left(\overline{S}_{k-1} - \overline{u}_k \overline{\Phi}_k \overline{\Phi}_k^T + \frac{c_\beta}{k^\beta} Z_k Z_k^T \right) \left(\tilde{\theta}_k - \theta \right)}_{:=R_{2,n}}$$

Rate of convergence of $R_{1,n}$: First, since $\kappa_{n,k-1} = (1 - \tau_k)\, \kappa_{n,k}$,

$$\begin{aligned} R_{1,n} &= \sum_{k=1}^{n-1} \kappa_{n,k} \left((1 - \tau_k) \frac{\tau_k}{\gamma_k} - \frac{\tau_{k+1}}{\gamma_{k+1}} \right) \overline{S}_{k-1} \left(\tilde{\theta}_k - \theta \right) \\ &= \sum_{k=1}^{n-1} \kappa_{n,k} \frac{\tau_{k+1}}{\gamma_{k+1}} \left((1 - \tau_k) \frac{\tau_k \gamma_{k+1}}{\tau_{k+1} \gamma_k} - 1 \right) \overline{S}_{k-1}^{-1} \left(\tilde{\theta}_k - \theta \right) \end{aligned}$$

Given that

$$1 - (1 - \tau_{n-1}) \frac{\tau_{n-1} \gamma_{n+1}}{\tau_n \gamma_n} = \frac{-2\nu + \gamma}{n} + o\left(\frac{1}{n} \right),$$

and applying Lemma A.1 coupled with Theorem 4.2, it comes that for all $\delta > 0$,

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

Rate of convergence of $R_{2,n}$: Thanks to Theorem 4.2 coupled with lemma A.1, one can check that for all $\delta > 0$,

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

Then, let us consider the sequence of events (Ω_n) defined for all $n \geq 0$ by

$$\Omega_n = \left\{ \left\| \tilde{\theta}_n - \theta \right\| < (\ln(n)^{1/2+\delta} \gamma_{n+1}^{1/2}, \| \theta_{\tau,n-1} - \theta \| < (\ln(n))^{1/2+\delta} \gamma_n^{1/2} \right\}.$$

Remark that $\mathbf{1}_{\Omega_n^C}$ converges almost surely to 0. Then, one can write $R_{2,n}$ as

$$\begin{aligned} R_{2,n} = & \sum_{k=1}^{n-1} \kappa_{n,k} \frac{\tau_{k+1}}{\gamma_{k+1}} \frac{1}{k+1} \overline{S}_{k-1} \left(\tilde{\theta}_k - \theta\right) \\ & - \sum_{k=1}^{n} \kappa_{n,k} \frac{\tau_{k+1}}{\gamma_{k+1}} \frac{1}{k+1} \left(\overline{u}_k \overline{\Phi}_k \overline{\Phi}_k^T + \frac{c_\beta}{k^\beta} Z_k Z_k^T \right) \left(\tilde{\theta}_k - \theta\right) \mathbf{1}_{\Omega_k} \\ & - \sum_{k=1}^{n-1} \kappa_{n,k} \frac{\tau_{k+1}}{\gamma_{k+1}} \frac{1}{k+1} \left(\overline{u}_k \overline{\Phi}_k \overline{\Phi}_k^T + \frac{c_\beta}{k^\beta} Z_k Z_k^T \right) \left(\tilde{\theta}_k - \theta\right) \mathbf{1}_{\Omega_k^C}. \end{aligned}$$

Applying Lemma A.1, for all $\delta > 0$,

$$\left\| \sum_{k=1}^{n-1} \kappa_{n,k} \frac{\tau_{k+1}}{\gamma_{k+1}} \frac{1}{k+1} \overline{S}_k \left(\tilde{\theta}_k - \theta\right) \right\| = o\left(\frac{(\ln n)^{1/2+\delta}}{n^{1-\gamma/2}} \right) \quad a.s.$$

Furthermore, remark that

$$\begin{aligned} & \sum_{k=1}^{n-1} \kappa_{n,k} \frac{\tau_{k+1}}{\gamma_{k+1}} \frac{1}{k+1} \left(\overline{u}_k \overline{\Phi}_k \overline{\Phi}_k^T + \frac{c_\beta}{k^\beta} Z_k Z_k^T \right) \left(\tilde{\theta}_k - \theta\right) \\ & \qquad = \kappa_{n,0} \sum_{k=1}^{n-1} \kappa_{k-1,0}^{-1} \frac{\tau_{k+1}}{\gamma_{k+1}} \frac{1}{k+1} \left(\overline{u}_k \overline{\Phi}_k \overline{\Phi}_k^T + \frac{c_\beta}{k^\beta} Z_k Z_k^T \right) \left(\tilde{\theta}_k - \theta\right) \end{aligned}$$

Since $\mathbf{1}_{\Omega_n^C}$ converges almost surely to 0,

$$\sum_{k \geq 1} \kappa_{k-1,0}^{-1} \frac{\tau_{k+1}}{\gamma_{k+1}} \frac{1}{k+1} \left\| \overline{u}_k \overline{\Phi}_k \overline{\Phi}_k^T + \frac{c_\beta}{k^\beta} Z_k Z_k^T \right\| \left\| \tilde{\theta}_k - \theta \right\| \mathbf{1}_{\Omega_k^C} < +\infty \quad a.s$$

and

$$\begin{aligned} \left\| \sum_{k=1}^{n-1} \kappa_{n,k} \frac{\tau_{k+1}}{\gamma_{k+1}} \frac{1}{k+1} \left(\overline{u}_k \overline{\Phi}_k \overline{\Phi}_k^T + \frac{c_\beta}{k^\beta} Z_k Z_k^T \right) \left(\tilde{\theta}_k - \theta\right) \mathbf{1}_{\Omega_k^C} \right\| &= O\left(\kappa_{n,0}\right) \quad a.s \\ &= O\left(\frac{1}{n^\tau} \right) \quad a.s, \end{aligned}$$

which is negligible since $\tau > 1/2$. Finally, let

$$\begin{aligned}
R_{3,n} &= \sum_{k=1}^{n-1} \kappa_{n,k} \frac{\tau_{k+1}}{\gamma_{k+1}} \frac{1}{k+1} \left\| \overline{u}_k \overline{\Phi}_k \overline{\Phi}_k^T + \frac{c_\beta}{k^\beta} Z_k Z_k^T \right\| \left\| \tilde{\theta}_k - \theta \right\| \mathbf{1}_{\Omega_k} \\
&\leq \sum_{k=1}^{n-1} \kappa_{n,k} \frac{\tau_{k+1}}{\sqrt{\gamma_{k+1}}} \frac{1}{k+1} (\ln k)^{1/2+\delta} \left\| \overline{u}_k \overline{\Phi}_k \overline{\Phi}_k^T + \frac{c_\beta}{k^\beta} Z_k Z_k^T \right\| \mathbf{1}_{\Omega_k} \\
&\leq \sum_{k=1}^{n-1} \kappa_{n,k} \frac{\tau_{k+1}}{\sqrt{\gamma_{k+1}}} \frac{1}{k+1} (\ln k)^{1/2+\delta} \underbrace{\left\| \overline{u}_k \overline{\Phi}_k \overline{\Phi}_k^T + \frac{c_\beta}{k^\beta} Z_k Z_k^T \right\|}_{:=\Delta_k} \mathbf{1}_{\Omega'_{k-1}}
\end{aligned}$$

with $\Omega'_{k-1} = \left\{ \|\theta_{\tau,k-1} - \theta\| \leq (\ln(k))^{1/2+\delta} \gamma_k \right\}$. Then, $R_{3,n}$ can be written as

$$R_{3,n} = \sum_{k=1}^{n-1} \kappa_{n,k} \frac{\tau_{k+1}}{\sqrt{\gamma_{k+1}}} \frac{1}{k+1} (\ln k)^{1/2+\delta} \mathbb{E}\left[\Delta_k | \mathcal{F}_{k-1}\right] + \sum_{k=1}^{n-1} \kappa_{n,k} \frac{\tau_{k+1}}{\sqrt{\gamma_{k+1}}} \frac{1}{k+1} (\ln k)^{1/2+\delta} \Xi_k$$

with $\Xi_k = \Delta_k - \mathbb{E}\left[\Delta_k | \mathcal{F}_{k-1}\right]$. Remark that (Ξ_{n+1}) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$ defined for all n by $\mathcal{F}_n = \sigma\left((X_1, Z_1), \ldots, (X_n, Z_n)\right)$. Thanks to inequality (12) coupled with lemma A.1,

$$\sum_{k=1}^{n-1} \kappa_{n,k} \frac{\tau_{k+1}}{\sqrt{\gamma_{k+1}}} \frac{1}{k+1} (\ln k)^{1/2+\delta} \mathbb{E}\left[\Delta_k | \mathcal{F}_{k-1}\right] = o\left(\frac{(\ln n)^{1/2+\delta}}{n^{1-\gamma/2}} \right) \quad a.s.$$

Let us now consider $\alpha \in (1/2, \tau)$, and $V_n = n^{2\alpha} \left(\sum_{k=1}^{n-1} \kappa_{n,k} \frac{\tau_{k+1}}{\sqrt{\gamma_{k+1}}} \frac{1}{k+1} (\ln k)^{1/2+\delta} \Xi_k \right)^2$. Then,

$$\mathbb{E}\left[V_n | \mathcal{F}_{n-1}\right] = |1 - \tau_n|^2 \left(\frac{n}{n-1} \right)^{2\alpha} V_{n-1} + n^{2\alpha} \frac{\tau_n^2}{\gamma_n} \frac{(\ln(n-1))^{1+\delta}}{n^2} \mathbb{E}\left[\Delta_n^2 | \mathcal{F}_{n-1}\right]$$

Since

$$|1 - \tau_n|^2 \left(\frac{n}{n-1} \right)^{2\alpha} = 1 - 2\frac{\tau - \alpha}{n} + o\left(\frac{1}{n} \right),$$

thanks to inequality (13) and applying Robbins-Siegmund Theorem,

$$\left(\sum_{k=1}^{n-1} \kappa_{n,k} \frac{\tau_{k+1}}{\sqrt{\gamma_{k+1}}} \frac{1}{k+1} (\ln k)^{1/2+\delta} \Xi_k \right)^2 = O\left(\frac{1}{n^{2\alpha}} \right) \quad a.s.$$

which concludes the proof.

A.7 Proof of Theorem 5.1

Verifying (A1). First, remark that for all $h \in \mathbb{R}^d$, $\|\nabla_h g(X,Y,h)\| \leq X$. Then, since X admits a second order moment, Assumption **(A1b)** is verified. Furthermore, we have

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

and **(A1a)** is so verified. Furthermore, since X admits a second order moment and since the functional π is continuous, the functional

$$\Sigma : h \longmapsto \mathbb{E}\left[\nabla_h g(X,Y,h)\nabla_h g(X,Y,h)^T\right] = \mathbb{E}\left[\left(Y - \pi\left(X^T h\right)\right)^2 XX^T\right]$$

is continuous on $\mathbb{R}^d$, and in particular at θ. Then, **(A1c)** is verified.

Verifying (A2a) and (A2b). First, remark that **(A2a)** is verified thanks to hypothesis. Furthermore, note that for all $h \in \mathbb{R}^d$,

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

and by continuity of π and since X admits a second order moment, G is twice continuously differentiable. Furthermore, $\left\|\nabla^2 G(h)\right\|_{op} \leq \frac{1}{4}\mathbb{E}\left[\|X\|^2\right]$ and **(A2a)** is so verified.

Verifying (H1'). Only the mains ideas are given since a detailed analogous proof is available in [2] (proof of Theorem 4.1). Remark that thanks to Riccati's formula, we have

$$\overline{S}_n = \frac{1}{n+1}\left(S_0 + \sum_{k=1}^{n} a_k X_k X_k^T\right)$$

and that by definition, $a_k \geq \frac{c_\beta}{k^\beta}$. Then, $\lambda_{\min}\left(\overline{S}_n\right) \geq \frac{1}{n+1}\lambda_{\min}\left(\sum_{k=1}^{n}\frac{c_\beta}{k^\beta}X_k X_k^T\right)$, and one can easily check that

$$\frac{1}{\sum_{k=1}^{n}\frac{c_\beta}{k^\beta}}\sum_{k=1}^{n}\frac{c_\beta}{k^\beta}X_k X_k^T \xrightarrow[n\to+\infty]{p.s} \mathbb{E}\left[XX^T\right] \tag{33}$$

which is supposed to be positive (since $\nabla^2 G(\theta)$ is). Then, one can easily check that

$$\lambda_{\max}\left(\overline{S}_n^{-1}\right) = O\left(n^\beta\right) \quad a.s.$$

In a same way, since $a_k \leq \frac{1}{4}$, one can easily check that

$$\lambda_{\max}\left(\overline{S}_n^{-1}\right) = O(1) \quad a.s$$

and **(H1')** is so verified.

Conclusion 1. Since Assumptions **(A1)**, **(A2a)**, **(A2b)** as well as **(H1')** are verified, Theorem 4.1 holds, i.e $\tilde{\theta}_n$ and $\theta_{\tau,n}$ converge almost surely to θ.

Verifying (H2a'). Only the mains ideas are given since a detailed analogous proof is available in [2] (proof of Theorem 4.1). Remark that $\overline{S}_n$ can be written as

$$\overline{S}_n = \frac{1}{n+1}\left(\overline{S}_0 + \sum_{k=1}^{n} \overline{a}_k X_k X_k^T + \sum_{k=1}^{n} (a_k - \overline{a}_k) X_k X_k^T\right)$$

and since $a_k - \overline{a}_k \neq 0$ if and only if $a_k > \overline{a}_k$, it comes thanks to equation (33),

$$\begin{aligned}\left\| \frac{1}{n+1}\overline{S}_0 + \frac{1}{n+1}\sum_{k=1}^{n} a_k X_k X_k^T \right\|_{op} &= \frac{1}{n+1}\left\|\overline{S}_0\right\|_{op} + \frac{1}{n+1}\left\|\sum_{k=1}^{n} \frac{c_\beta}{k^\beta} X_k X_k^T\right\|_{op} \\ &= O\left(n^{-\beta}\right) \quad a.s.\end{aligned}$$

Furthermore, as in the proof of Theorem 4.1 in [2], one can check, since $\theta_{\tau,n}$ converges to θ, that

$$\frac{1}{n}\sum_{k=1}^{n} \overline{a}_k X_k X_k^T = \frac{1}{n}\sum_{k=1}^{n} \pi\left(\theta_{\tau,k-1}^T X_k\right)\left(1 - \pi\left(\theta_{\tau,k-1}^T X_k\right)\right) X_k X_k^T \xrightarrow[n\to+\infty]{a.s} \nabla^2 G(\theta)$$

and **(H2a')** is so verified.

Conclusion 2. Since Assumptions **(A1)**, **(A2a)**, **(A2b)**, **(H1')** and **(H2a')** are verified, if X admits a moment of order $2+2\eta$ with $\eta > \frac{1}{\alpha} - 1$, Theorem 4.2 holds, i.e

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

Verifying (A2c). Thanks to Lemma 6.2 in [2], we have for all $h_1, h_2 \in \mathbb{R}^d$,

$$\left|\pi\left(h_1^T X\right)\left(1 - \pi\left(h_1^T X\right)\right) - \pi\left(h_2^T X\right)\left(1 - \pi\left(h_2^T X\right)\right)\right| \leq \frac{1}{12\sqrt{3}}\|X\|\,\|h_1 - h_2\|$$

and in a particular case, it comes

$$\left\|\nabla^2 G\left(h_1\right)-\nabla^2 G\left(h_2\right)\right\|_{op} \leq \frac{1}{12\sqrt{3}}\mathbb{E}\left[\|X\|^3\right]\|h_1-h_2\|$$

and **(A2c)** is verified since X admits a third order moment.

Verifying inequalities (12) **and** (13)**.** Remark that for all $n \geq 1$,

$$\left\|a_n X_n X_n^T\right\| \leq \max\left\{\frac{1}{4}, c_\beta\right\}\|X\|^2 =: C_{a,\beta}\|X\|^2$$

so that, if X admits a fourth order moment,

$$\mathbb{E}\left[\left\|a_n X_n X_n^T\right\|\right] \leq C_{a,\beta}\mathbb{E}\left[\|X\|^2\right] \qquad \text{and} \qquad \mathbb{E}\left[\left\|a_n X_n X_n^T\right\|^2\right] \leq C_{a,\beta}^2\mathbb{E}\left[\|X\|^4\right]$$

and inequalities (12) and (13) are so verified.

Conclusion 3. Theorem 4.4 holds, meaning that

$$\|\theta_{\tau,n}-\theta\|^2 = O\left(\frac{\ln n}{n}\right) \quad a.s \qquad \text{and} \qquad \sqrt{n}\left(\theta_{\tau,n}-\theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\right).$$

Convergence of $\overline{S}_n$**.** First, let us recall

$$\overline{S}_n = \frac{1}{n+1}\left(\overline{S}_0 + \underbrace{\sum_{k=1}^{n}\overline{a}_k X_k X_k^T}_{A_n} + \sum_{k=1}^{n}\left(a_k-\overline{a}_k\right)X_k X_k^T\right)$$

and that

$$\left\|\frac{1}{n+1}\left(\overline{S}_0 + \sum_{k=1}^{n}\left(a_k-\overline{a}_k\right)X_k X_k^T\right)\right\|^2 = O\left(\frac{1}{n^{2\beta}}\right) \quad a.s.$$

Furthermore, let us split A_n into two terms, i.e

$$A_n = \sum_{k=1}^{n}\nabla^2 G\left(\theta_{\tau,k-1}\right) + \sum_{k=1}^{n}\Xi_k$$

with $\Xi_k = \overline{a}_k X_k X_k^T - \nabla^2 G\left(\theta_{\tau,k-1}\right)$. Since (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$ and since $\mathbb{E}\left[\|\Xi_k\|^2 \,|\mathcal{F}_{k-1}\right] \leq \frac{1}{16}\mathbb{E}\left[\|X\|^4\right]$, we have (see Theorem 6.2 in [6]) for all $\delta > 0$

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

Furthermore, since X admits a third order moment, the Hessian is $\frac{1}{12\sqrt{3}}\mathbb{E}\left[\|X\|^3\right]$-lipschitz and for all $\delta > 0$, by Toeplitz Lemma,

$$\left\| \frac{1}{n}\sum_{k=1}^{n} \nabla^2 G\left(\theta_{\tau,k-1}\right) - \nabla^2 G(\theta) \right\| \leq \frac{\mathbb{E}\left[\|X\|^3\right]}{12\sqrt{3}n} \sum_{k=1}^{n} \|\theta_{\tau,k-1} - \tau_k\| = o\left(\frac{(\ln n)^{1/2+\delta/2}}{\sqrt{n}}\right) \quad a.s,$$

which concludes the proof.

A.8 Proofs of Theorems 5.2 and 5.3

In what follows, let us recall that we consider the functional $G : \mathbb{R}^p \times \ldots \times \mathbb{R}^p \longrightarrow \mathbb{R}$ defined for all $h = (h_1, \ldots, h_K)$ by

$$G(h) := \mathbb{E}\left[\log\left(\frac{e^{h_Y^T X}}{\sum_{k=1}^{K} e^{h_k^T X}}\right)\right]$$

A.8.1 Some results on the functional G

Verifying assumption (A1). First remark that

$$\|\nabla_h g\left(X, Y, h\right)\| \leq \sqrt{K}\,\|X\|\,.$$

and if X admits a second order moment, **(A1b)** is verified. Furthermore it is evident that **(A1a)** is verified. Indeed, we have

$$\nabla G(\theta) = \mathbb{E}\left[\mathbb{E}\left[\nabla_h g(X, Y, \theta)|X\right]\right] = \mathbb{E}\left[\begin{pmatrix} X\left(\frac{e^{\theta_1^T X}}{\sum_{k=1}^{K} e^{\theta_k^T X}} - \mathbb{E}\left[\mathbf{1}_{Y=1}|X\right]\right) \\ \vdots \\ X\left(\frac{e^{\theta_K^T X}}{\sum_{k=1}^{K} e^{\theta_k^T X}} - \mathbb{E}\left[\mathbf{1}_{Y=K}|X\right]\right) \end{pmatrix}\right] = 0.$$

Furthermore, for all h,

$$\mathbb{E}\left[\nabla_h g(X, Y, h)\nabla_h g\left(X, Y, h\right)^T\right] - \mathbb{E}\left[\nabla_h g(X, Y, \theta)\nabla_h g\left(X, Y, \theta\right)^T\right] = \mathbb{E}\left[\left(\sigma(X, h) - \sigma(X, \theta)\right)\left(\sigma(X, h) - \sigma(X, \theta)\right)^T \otimes XX^T\right]. \tag{34}$$

and since the functional σ is bounded and continuous, by dominated convergence, since X admits a second order moment, **(A1c)** is verified.

Verifying assumption (A2). First, remark that for all h, since $\|\sigma(.,.)\|$ is bounded by $\sqrt{K}$,

$$\left\|\nabla^2 G(h)\right\|_{op} \leq \mathbb{E}\left[\|X\|^2 \|\sigma(X,h)\|\right] \leq \sqrt{K}\mathbb{E}\left[\|X\|^2\right].$$

Then, if X admits a second order moment, assumption (A2a) is verified. Furthermore, (A2b) is verified by hypothesis. Finally, let us consider the functional $F_{k'} : [0,1] \longrightarrow \mathbb{R}$ defined for all t by

$$F_{k'}(t) = \frac{e^{(\theta_{k'}+t(h_{k'}-\theta_{k'}))^T X}}{\sum_{k=1}^K e^{(\theta_k+t(h_k-\theta_k))^T X}}$$

Then, for all $t \in [0,1]$,

$$F'(t) = \frac{(h_{k'}-\theta_{k'})^T X e^{(\theta_{k'}+t(h_{k'}-\theta_{k'}))^T X}}{\sum_{k=1}^K e^{(\theta_k+t(h_k-\theta_k))^T X}} - \underbrace{\frac{e^{(\theta_{k'}+(h_{k'}-\theta_{k'}))^T X}\sum_{k=1}^K t\,(h_k-\theta_k)^T X e^{(\theta_k+t(h_k-\theta_k))^T X}}{\left(\sum_{k=1}^K e^{(\theta_k+t(h_k-\theta_k))^T X}\right)^2}}_{(*)}$$

First, one can check that for all $t \in [0,1]$,

$$\left|\frac{(h_{k'}-\theta_{k'})^T X e^{(\theta_{k'}+t(h_{k'}-\theta_{k'}))^T X}}{\sum_{k=1}^K e^{(\theta_k+t(h_k-\theta_k))^T X}}\right| \leq \|h_{k'}-\theta_{k'}\|\,\|X\|.$$

Furthermore, applying Cauchy-Schwarz inequality,

$$\begin{aligned}(*) &\leq \|X\| \frac{e^{(\theta_{k'}+t(h_{k'}-\theta_{k'}))^T X}}{\left(\sum_{k=1}^K e^{(\theta_k+t(h_k-\theta_k))^T X}\right)^2}\sqrt{\sum_{k=1}^K \|h_k-\theta_k\|^2}\sqrt{\sum_{k=1}^K e^{2(\theta_k+t(h_k-\theta_k))^T X}} \\ &\leq \|X\|\,\|\theta-h\|.\end{aligned}$$

Then,

$$\left|\frac{e^{\theta_{k'}^T X}}{\sum_{k=1}^K e^{\theta_k^T X}} - \frac{e^{h_{k'}^T X}}{\sum_{k=1}^K e^{h_k^T X}}\right| \leq \|X\|\left(\|h_{k'}-\theta_{k'}\| + \|\theta-h\|\right). \tag{35}$$

Then,

$$\|\sigma(X,h)-\sigma(X,\theta)\| \leq 2\sqrt{K}\,\|X\|\,\|\theta-h\| \tag{36}$$

and

$$\|\mathrm{diag}\left(\sigma(X,\theta)\right) - \mathrm{diag}\left(\sigma(X,h)\right)\| \leq 2\sqrt{K}\,\|X\|\,\|\theta-h\|$$

Then,

$$\left\|\nabla^2 G(\theta) - \nabla^2 G(h)\right\| \leq 6\sqrt{K}\mathbb{E}\left[\|X\|^3\right]\|h-\theta\|.$$

Finally, if X admits a third order moment, the Hessian is $6\sqrt{K}\mathbb{E}\left[\|X\|^3\right]$-Lipschitz and Assumption **(A2c)** is thus verified.

A.8.2 Proof of Theorems 5.2 and 5.3

Proof of Theorem 5.2.

Verifying (H1). Remark that

$$\lambda_{\min}\left(H_n\right) \geq \lambda_{\min}\left(\sum_{k=1}^{n} \beta_k Z_k Z_k^T\right)$$

and that, since $\beta \in (0, 1/2)$,

$$\frac{1-\beta}{c_\beta^{-1} n^{1-\beta}} \sum_{k=1}^{n} \beta_k Z_k Z_k^T \xrightarrow[n\to+\infty]{a.s} I_{pK}.$$

Then,

$$\left\|\overline{H}_n^{-1}\right\| = O\left(n^\beta\right) \quad a.s.$$

Furthermore, since $\|\nabla_h g(X, Y, h)\| \leq \|X\|$, and thanks to assumption **(HS1a)**,

$$\frac{1}{n+1}\left\|\sum_{k=1}^{n} \nabla_h g\left(X_k, Y_k, \theta_{k-1}\right) \nabla_h g\left(X_k, Y_k, \theta_{k-1}\right)^T\right\| \leq \frac{1}{n}\sum_{k=1}^{n}\|X_k\|^2 \xrightarrow[n\to+\infty]{a.s} \mathbb{E}\left[\|X\|^2\right]$$

ans assumption **(H1)** is so verified.

Conclusion 1. Since Assumptions **(A1a)**, **(A1b)**,**(A2a)**, **(A2b)** and **(H1)** are fulfilled, Theorem 3.1 holds, i.e θ_n converges almost surely to θ.

Verifying (H2a). First, let us rewrite $\overline{H}_n$ as

$$\overline{H}_n = \frac{1}{n+1}\left(\overline{H}_0^{-1} + \underbrace{\sum_{k=1}^{n} \nabla_h g\left(X_k, Y_k, \theta_{k-1}\right) \nabla_h g\left(X_k, Y_k, \theta\right)^T}_{:=A_n} + \sum_{k=1}^{n} \frac{c_\beta}{k^\beta} Z_k Z_k^T\right)$$

and we have already proven that

$$\left\| \frac{1}{n+1} \left(\overline{H}_0^{-1} + \sum_{k=1}^{n} \frac{c_\beta}{k^\beta} Z_k Z_k^T \right) \right\|^2 = O\left(\frac{1}{n^{2\beta}} \right) \quad p.s.$$

Furthermore, one can rewrite A_n as

$$A_n = \sum_{k=1}^{n} \mathbb{E}\left[\nabla_h g\left(X_k, Y_k, \theta_{k-1}\right) \nabla_h g\left(X_k, Y_k, \theta_{k-1}\right)^T | \mathcal{F}_{k-1} \right] + \sum_{k=1}^{n} \Xi_k$$

with

$$\begin{aligned} \Xi_k :=& \nabla_h g\left(X_k, Y_k, \theta_{k-1}\right) \nabla_h g\left(X_k, Y_k, \theta_{k-1}\right)^T \\ & - \mathbb{E}\left[\nabla_h g\left(X_k, Y_k, \theta_{k-1}\right) \nabla_h g\left(X_k, Y_k, \theta_{k-1}\right)^T | `\mathcal{F}_{k-1} \right]. \end{aligned}$$

First, note that if X admits a fourth order moment, $\mathbb{E}\left[\left\| \nabla_h g\left(X_k, Y_k, \theta_{k-1}\right) \right\|^4 | \mathcal{F}_{k-1} \right] \leq K^2 \mathbb{E}\left[\|X\|^4 \right]$, so that applying Theorem 6.2 in [6], for all $\delta > 0$,

$$\left\| \frac{1}{n+1} \sum_{k=0}^{n} \Xi_k \right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n} \right) \quad a.s.$$

Furthermore, we have proven that if X admits a second order moment, **(A1c)** is fulfilled. Then, since θ_n converges almost surely to θ and by continuity

$$\begin{aligned} \frac{1}{n+1} \sum_{k=1}^{n} & \mathbb{E}\left[\nabla_h g\left(X_k, Y_k, \theta_{k-1}\right) \nabla_h g\left(X_k, Y_k, \theta_{k-1}\right)^T | \mathcal{F}_{k-1} \right] \\ & \xrightarrow[n \to +\infty]{a.s} \mathbb{E}\left[\nabla_h g\left(X, Y, \theta\right) \nabla_h g\left(X, Y, \theta\right) \right] = H, \end{aligned}$$

and Assumption **(H2a)** is so fulfilled.

Conclusion 2. If X admits a second order moment, Assumptions **(A1)**, **(A2)**, **(H1)**, **(H2a)** are verified and Theorem 3.2 holds, i.e

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

Verifying (H2b). Note that thanks to inequalities (34) and (36),

$$\begin{aligned} \Big\| \mathbb{E}\big[\nabla_h g\left(X, Y, h\right) \nabla_h g\left(X, Y, h\right) - \nabla_h g\left(X, Y, \theta\right) \nabla_h \left(X, Y, \theta\right)^T \big] \Big\| & \\ \leq \mathbb{E}\left[\|\sigma(X, h) - \sigma(X, \theta)\|^2 \|X\|^2 \right] & \\ \leq 4K \mathbb{E}\left[\|X\|^4 \right] \|h - \theta\|^2. & \end{aligned}$$

Then, thanks to the Toeplitz lemma, it comes that for all $\delta > 0$,

$$\left\| \frac{1}{n+1} \sum_{k=1}^{n} \mathbb{E}\left[\nabla_h g\left(X_k, Y_k, \theta_{k-1}\right) \nabla_h g\left(X_k, Y_k, \theta_{k-1}\right)^T | \mathcal{F}_{k-1} \right] - H \right\|$$
$$\leq \frac{\|H\|}{n+1} + \frac{4K\mathbb{E}\left[\|X\|^4\right]}{n+1} \sum_{k=1}^{n} \|\theta_{k-1} - \theta\|^2$$
$$= o\left(\frac{(\ln n)^{1+\delta}}{n} \right) \quad a.s.$$

Then,

$$\left\|\overline{H}_n - H\right\|^2 = O\left(\frac{1}{n^{2\beta}}\right) \quad a.s,$$

and **(H2b)** is so verified.

Conclusion 3. If X admits a fourth order moment, Assumptions **(A1)**, **(A2)**, **(H1)**, **(H2a)** and **(H2b)** are fulfilled, so that Theorem 3.3 holds. Since

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

we so have

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

□

Proof of Theorem 5.3.

Verifying (H1') and conclusion 1. If X admits a second order moment, with the same calculus as in the proof of Theorem 5.2, up to take $\beta < \gamma - 1/2$ instead of $\beta < 1/2$, one can check that Assumption **(H1')** is verified. Then, $\tilde{\theta}_n$ and $\theta_{\tau,n}$ converge almost surely to θ.

Verifying (H2a') and conclusion 2. If X admits a fourth order moment, with the same calculus as in the proof of Theorem 5.2, up to take $\beta < \gamma - 1/2$ instead of $\beta < 1/2$, one can check that Assumption **(H2a')** is verified. Then,

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

Furthermore, let us recall that

$$\theta_{\tau,n} - \theta = \kappa_{n,0}\left(\theta_{0,\tau} - \theta\right) + \sum_{k=0}^{n-1} \kappa_{n,k} \tau_{k+1} \left(\tilde{\theta}_k - \theta\right)$$

and that $|\kappa_{n,0}| = O\left(n^{-\tau}\right)$. Furthermore, applying Lemma A.1, for all $\delta > 0$,

$$\sum_{k=0}^{n-1} \kappa_{n,k}\tau_{k+1} \left\|\tilde{\theta}_k - \theta\right\| = o\left(\frac{(\ln n)^{1/2+\delta/2}}{n^{\gamma/2}}\right) \quad a.s$$

leading, since $\tau > 1/2$, to

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

Verifying (H2b). Note that by inequalities (34) and (36),

$$\begin{aligned}\left\|\mathbb{E}\left[\nabla_h g\left(X,Y,h\right)\nabla_h g\left(X,Y,h\right) - \nabla_h g\left(X,Y,\theta\right)\nabla_h\left(X,Y,\theta\right)^T\right]\right\| \\ \leq \mathbb{E}\left[\|\sigma(X,h) - \sigma(X,\theta)\|^2 \|X\|^2\right] \\ \leq 4K\mathbb{E}\left[\|X\|^4\right]\|h-\theta\|^2.\end{aligned}$$

Then, thanks to the Toeplitz lemma, it comes that for all $\delta > 0$,

$$\begin{aligned}&\left\|\frac{1}{n+1}\sum_{k=1}^{n}\mathbb{E}\left[\nabla_h g\left(X_k,Y_k,\theta_{\tau,k-1}\right)\nabla_h g\left(X_k,Y_k,\theta_{\tau,k-1}\right)^T|\mathcal{F}_{k-1}\right] - H\right\| \\ &\leq \frac{\|H\|}{n+1} + \frac{4K\mathbb{E}\left[\|X\|^4\right]}{n+1}\sum_{k=1}^{n}\|\theta_{\tau,k-1}-\theta\|^2 \\ &= o\left(\frac{(\ln n)^{1+\delta}}{n^{\gamma}}\right) \quad a.s.\end{aligned}$$

Then, since $\beta < \gamma - 1/2$,

$$\left\|\overline{H}_n - H\right\|^2 = O\left(\frac{1}{n^{2\beta}}\right) \quad a.s,$$

and **(H2b)** is so verified.

Conclusion 3. If X admits a fourth order moment, Assumptions **(A1)**, **(A2)**, **(H1')**, **(H2a')** and **(H2b')** are fulfilled, and Theorem 4.3 holds, i.e

$$\|\theta_{\tau,n} - \theta\| = O\left(\frac{\ln n}{n}\right) \quad a.s \qquad \text{and} \qquad \sqrt{n}\left(\theta_{\tau,n} - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, H^{-1}\right)$$

A.9 An usuful technical Lemma

Let us now give an useful technical lemma introduced (as Lemma 4) in [13].

Lemma A.1. *Let* $(a_n) \in \mathcal{GS}(a^*)$ *with* $a^* > 0$ *and* $na_n \xrightarrow[n\to+\infty]{} \psi \geq 0$. *Let* $m > 0$ *and* $(u_n) \in \mathcal{GS}(u^*)$ *with* u^* *such that* $m + u^*\psi > 0$ *and* α_n *such that* $\alpha_n \xrightarrow[n\to+\infty]{} 0$. *Then*

$$u_n^{-1} \prod_{i=1}^{n} (1 - a_j)^m \xrightarrow[n\to+\infty]{} 0.$$

$$u_n^{-1} \sum_{k=1}^{n} \prod_{j=k+1}^{n} (1 - a_j)^m a_k u_k \xrightarrow[n\to+\infty]{} (m + u^*\psi)^{-1}.$$

$$u_n^{-1} \sum_{k=1}^{n} \prod_{j=k+1}^{n} (1 - a_j)^m a_k u_k \alpha_k \xrightarrow[n\to+\infty]{}.$$

□

B Simulations for the softmax regression

The considered multinomial regression model is defined in the case of three-label classification in dimension $d = 3$, for all $k = 1, 2, 3$, by

$$\mathbb{P}[Y = k|X] = \frac{e^{\theta_k^T X}}{\sum_{k'=1}^{3} e^{\theta_{k'}^T X}}$$

with $\theta = \left(\theta_1^T, \theta_2^T, \theta_3^T\right)^T$ chosen randomly on the unit sphereof $\mathbb{R}^9$. In Figure 6, we display the evolution of the quadratic mean error of the different estimates, for three different initializations, for correlated Gaussian variables $X \sim \mathcal{N}\left(0, A\text{diag}\left(\frac{i^2}{d^2}\right)_{i=1,\dots,d} A^T\right)$ where A is an orthogonal matrix randomly generated. Results in the case of heteroscedastic Gaussian variables $X \sim \mathcal{N}\left(0, \text{diag}\left(\frac{i^2}{d^2}\right)_{i=1,\dots,d}\right)$ are very similar and can be found in Figure 7. In Figure 6, one can see that again the averaged versions (weighted or not) converge faster in the case of bad initial point. The improvement over the Adagrad algorithm is made clearer as the initial point is chosen further from the optimum.

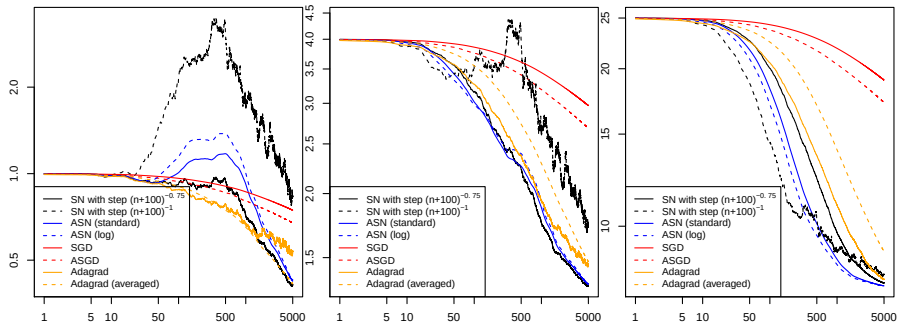

Figure 6: (Softmax regression with correlated Gaussian variables) Mean-squared error of the distance to the optimum θ with respect to the sample size for different initializations: $\theta_0 = \theta + r_0 U$, where U is a uniform random variable on the unit sphere of $\mathbb{R}^d$, with $r_0 = 1$ (left), $r_0 = 2$ (middle) or $r_0 = 5$ (right). Each curve is obtained by an average over 50 different samples of size $n = 5000$ (drawing a different initial point each time).

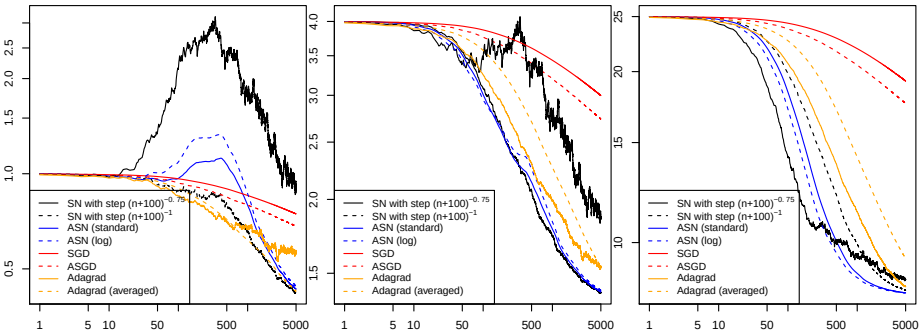

Figure 7: (Softmax regression with heteroscedastic Gaussian variables) Mean-squared error of the distance to the optimum θ with respect to the sample size for different initializations: $\theta_0 = \theta + r_0 U$, where U is a uniform random variable on the unit sphere of $\mathbb{R}^d$, with $r_0 = 1$ (left), $r_0 = 2$ (middle) or $r_0 = 5$ (right). Each curve is obtained by an average over 50 different samples of size $n = 5000$ (drawing a different initial point each time).