

An efficient stochastic Newton algorithm for parameter estimation in logistic regressions

Bernard Bercu, Antoine Godichon, Bruno Portier

▶ **To cite this version:**

Bernard Bercu, Antoine Godichon, Bruno Portier. An efficient stochastic Newton algorithm for parameter estimation in logistic regressions. SIAM Journal on Control and Optimization, 2020, 58 (1), pp.348-367. 10.1137/19M1261717 . hal-02103041

HAL Id: hal-02103041
https://hal.science/hal-02103041

Submitted on 18 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

AN EFFICIENT STOCHASTIC NEWTON ALGORITHM FOR PARAMETER ESTIMATION IN LOGISTIC REGRESSIONS

BERNARD BERCU, ANTOINE GODICHON, AND BRUNO PORTIER

ABSTRACT. Logistic regression is a well-known statistical model which is commonly used in the situation where the output is a binary random variable. It has a wide range of applications including machine learning, public health, social sciences, ecology and econometry. In order to estimate the unknown parameters of logistic regression with data streams arriving sequentially and at high speed, we focus our attention on a recursive stochastic algorithm. More precisely, we investigate the asymptotic behavior of a new stochastic Newton algorithm. It enables to easily update the estimates when the data arrive sequentially and to have research steps in all directions. We establish the almost sure convergence of our stochastic Newton algorithm as well as its asymptotic normality. All our theoretical results are illustrated by numerical experiments.

1. INTRODUCTION

Logistic regression is a well-known statistical model which is commonly used in the situation where the output is a binary random variable. It has a wide range of applications including machine learning [1], public health [11], social sciences, ecology [9] and econometry [17]. In what follows, we will consider a sequence (X_n, Y_n) of random variables taking values in $\mathbb{R}^d \times \{0,1\}$, and we will assume that (X_n) is a sequence of independent and identically distributed random vectors such that, that for all $n \geqslant 1$, the conditional distribution of Y_n knowing X_n is a Bernoulli distribution [8]. More precisely, let $\theta = (\theta_0, \ldots, \theta_d)^T$ be the unknown parameter belonging to $\mathbb{R}^{d+1}$ of the logistic regression. For all $n \geqslant 1$, we denote $\Phi_n = (1, X_n^T)^T$ and we assume that

$$\mathcal{L}\big(Y_n|\Phi_n\big) = \mathcal{B}\big(\pi(\theta^T\Phi_n)\big) \qquad \text{where} \qquad \pi(x) = \frac{\exp(x)}{1+\exp(x)}.$$

Our goal is the estimation of the vector of parameters θ. For that purpose, let G be the convex positive function defined, for all $h \in \mathbb{R}^{d+1}$, by

$$\begin{aligned} G(h) &= \mathbb{E}\Big[-\log\big(\pi(h^T\Phi)^Y\big(1-\pi(h^T\Phi)\big)^{1-Y}\big)\Big], \\ &= \mathbb{E}\Big[\log\big(1+\exp(h^T\Phi)\big) - h^T\Phi Y\Big] \end{aligned}$$

where $\mathcal{L}(Y|\Phi) = \mathcal{B}\big(\pi(\theta^T\Phi)\big)$ and Φ shares the same distribution as Φ_1. We clearly have $\mathbb{E}\left[Y|\Phi\right] = \pi(\theta^T\Phi)$. Hence, one can easily check that the unknown parameter θ satisfies

$$\nabla G(\theta) = \mathbb{E}\Big[\Phi(\pi(\theta^T\Phi) - Y)\Big] = 0. \tag{1.1}$$

Consequently, under some standard convexity assumptions on G,

$$\theta = \arg\min_{h\in\mathbb{R}^{d+1}} G(h). \tag{1.2}$$

Since there is no explicit solution of the equation $\nabla G(h) = 0$, it is necessary to make use of an approximation algorithm in order to estimate θ. Usually, when the sample size is fixed, we approximate the solution with the help of a Newton root-finding numerical algorithm. However, when the data streams arrive sequentially and at high speed, it is much more appropriate and efficient to treat them with the help of stochastic gradient algorithms. We refer the reader to the seminal paper [15] and to its averaged version [14, 16], as well as to the more recent contributions on the logistic regression [1, 6, 5]. One can observe that in these last references, the conditional distribution $\mathcal{L}(Y|\Phi) = \mathcal{R}\big(\pi(\theta^T\Phi)\big)$ is the Rademacher distribution, instead of the usual Bernoulli $\mathcal{B}\big(\pi(\theta^T\Phi)\big)$ one.

In this paper, we propose an alternative strategy to stochastic gradient algorithms, in the spirit of the Newton algorithm, in the sense that the step sequence of stochastic gradient algorithms is replaced by recursive estimates of the inverse of the Hessian matrix of the function we are minimizing. This strategy enables us to properly deal with the situation where the Hessian matrix has eigenvalues with significantly different absolute values. Indeed, in that case, it can be necessary to adapt automatically the step of the algorithm in all directions. To be more precise, we propose to estimate the unknown parameter θ with the help of a stochastic Newton algorithm given, for all $n \geqslant 1$, by

$$\begin{aligned} a_n &= \pi(\theta_{n-1}^T\Phi_n)\left(1 - \pi(\theta_{n-1}^T\Phi_n)\right), \\ S_n^{-1} &= S_{n-1}^{-1} - a_n(1 + a_n\Phi_n^T S_{n-1}^{-1}\Phi_n)^{-1} S_{n-1}^{-1}\Phi_n\Phi_n^T S_{n-1}^{-1}, \\ \theta_n &= \theta_{n-1} + S_n^{-1}\Phi_n\left(Y_n - \pi(\theta_{n-1}^T\Phi_n)\right) \end{aligned}$$

where the initial value θ_0 is a bounded vector of $\mathbb{R}^{d+1}$ which can be arbitrarily chosen and S_0 is a positive definite and deterministic matrix. For the sake of simplicity and in all the sequel, we take $S_0 = I_{d+1}$ where I_{d+1} stands for the identity matrix of order

$(d+1)$. One can observe that

$$S_n = \sum_{k=1}^{n} a_k \Phi_k \Phi_k^T + I_{d+1}$$

Moreover, S_n^{-1} is updated recursively, thanks to Riccati's equation ([4], page 96) which enables us to avoid the useless inversion of a matrix at each iteration of the algorithm. Furthermore, the matrix $n^{-1}S_n$ is an estimate of the Hessian matrix $\nabla^2 G(\theta)$ at the unknown value θ. In order to ensure the convergence of the stochastic Newton algorithm, a modified version of this algorithm is provided. We shall prove its asymptotic efficiency by establishing its almost sure convergence and its asymptotic normality.

This algorithm is closely related to the iterative one used to estimate the unknown vector θ of a linear regression model satisfying, for all $n \geqslant 1$, $Y_n = \theta^T \Phi_n + \varepsilon_n$. As a matter of fact, the updating of the least squares estimator of the parameter θ is given by

$$\begin{aligned} S_n^{-1} &= S_{n-1}^{-1} - (1 + \Phi_n^T S_{n-1}^{-1} \Phi_n)^{-1} S_{n-1}^{-1} \Phi_n \Phi_n^T S_{n-1}^{-1} \\ \theta_n &= \theta_{n-1} + S_n^{-1} \Phi_n \left(Y_n - \theta_{n-1}^T \Phi_n \right) \end{aligned}$$

where the initial value θ_0 can be arbitrarily chosen and S_0 is a positive definite deterministic matrix. This algorithm can be considered as a Newton stochastic algorithm since the matrix $n^{-1}S_n$ is an estimate of the Hessian matrix of the least squares criterion $\mathbb{E}\left[(Y - \theta^T \Phi)^2\right]/2$.

To the best of our knowledge and apart from the least squares estimate mentioned above, stochastic Newton algorithms are hardly ever used and studied since they often require the inversion of a matrix at each step, which can be very expensive in term of time calculation. An alternative to the stochastic Newton algorithm is the BFGS algorithm [12, 10, 2] based on the recursive estimation of a matrix whose behavior is closed to the one of the inverse of the Hessian matrix. Nevertheless, this last estimate does not converge to the exact inverse of the Hessian matrix. Consequently, the estimation of the unknown vector θ is not satisfactory.

The paper is organized as follows. Section 2 describes the framework and our main assumptions. In Section 3, we introduce our new stochastic Newton algorithm. Section 4 is devoted to its almost sure convergence as well as its asymptotic normality. Our theoretical results are illustrated by numerical experiments in Section 5. Finally, all technical proofs are postponed to Sections 6 and 7.

2. FRAMEWORK

In what follows, we shall consider a couple of random variables (X, Y) taking values in $\mathbb{R}^d \times \{0, 1\}$ where d is a positive integer, and such that

$$\mathcal{L}\big(Y|\Phi\big) = \mathcal{B}\big(\pi(\theta^T\Phi)\big) \qquad \text{where} \qquad \pi(x) = \frac{\exp(x)}{1+\exp(x)}$$

with $\Phi = (1, X^T)^T$ and where $\theta = (\theta_0, \theta_1, ..., \theta_d)$ is the unknown parameter to estimate. We recall that θ is a minimizer of the convex function G defined, for all $h \in \mathbb{R}^{d+1}$, by

$$G(h) = \mathbb{E}\Big[-\log\big(\pi(h^T\Phi)^Y\big(1-\pi(h^T\Phi)\big)^{1-Y}\big)\Big] = \mathbb{E}\big[g\left(\Phi, Y, h\right)\big]. \tag{2.1}$$

In all the sequel, we assume that the following assumptions are satisfied.

- **(A1)** The vector Φ has a finite moment of order 2 and the matrix $\mathbb{E}[\Phi\Phi^T]$ is positive definite.
- **(A2)** The Hessian matrix $\nabla^2 G\left(\theta\right)$ is positive definite.

These assumptions ensure that θ is the unique minimizer of the functional G. Assumption **(A1)** enables us to find a first lower bound for the smallest eigenvalue of the estimates of the Hessian matrix, while assumptions **(A1)** and **(A2)** give the unicity of the minimizer of G and ensure that the functional G is twice continuously differentiable. More precisely, for all $h \in \mathbb{R}^{d+1}$, we have

$$\nabla G(h) = \mathbb{E}\left[\nabla_h g\left(\Phi, Y, h\right)\right] = \mathbb{E}\left[\frac{\exp\left(h^T\Phi\right)}{1+\exp\left(h^T\Phi\right)}\Phi\right] - \mathbb{E}\left[Y\Phi\right], \tag{2.2}$$

$$\nabla^2 G\left(h\right) = \mathbb{E}\left[\nabla_h^2 g\left(\Phi, Y, h\right)\right] = \frac{1}{4}\,\mathbb{E}\left[\frac{1}{\left(\cosh\left(h^T\Phi/2\right)\right)^2}\Phi\Phi^T\right]. \tag{2.3}$$

Remark 2.1. *In the previous literature, it is more usual to consider a variable Y taking values in $\{-1, 1\}$, which means that $\mathcal{L}(Y|\Phi) = \mathcal{R}\big(\pi(\theta^T\Phi)\big)$ is the Rademacher distribution* [1, 6, 5]. *In this context, θ is a minimizer of the functional G defined, for all $h \in \mathbb{R}^{d+1}$, by*

$$G(h) = \mathbb{E}\Big[\ln\Big(1+\exp\Big(-Yh^T\Phi\Big)\Big)\Big].$$

Under assumptions, the functional G is twice continuously differentiable and, for all $h \in \mathbb{R}^{d+1}$,

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

$$\nabla^2 G\left(h\right) = \frac{1}{4}\,\mathbb{E}\left[\frac{1}{\left(\cosh\left(h^T\Phi/2\right)\right)^2}\Phi\Phi^T\right].$$

One can observe that the Hessian matrix remains the same. It ensures that the algorithm introduced in Section 3 can be adapted to this functional and that all the results given in Section 4 still hold in this case.

3. Stochastic Newton algorithm

In order to deal with massive data acquired online, let us recall that the stochastic Newton algorithm presented in the introduction is given, for all $n \geqslant 1$, by

$$a_n = \pi(\theta_{n-1}^T \Phi_n)\left(1 - \pi(\theta_{n-1}^T \Phi_n)\right), \tag{3.1}$$

$$S_n^{-1} = S_{n-1}^{-1} - a_n(1 + a_n \Phi_n^T S_{n-1}^{-1} \Phi_n)^{-1} S_{n-1}^{-1} \Phi_n \Phi_n^T S_{n-1}^{-1}, \tag{3.2}$$

$$\theta_n = \theta_{n-1} + S_n^{-1} \Phi_n \left(Y_n - \pi(\theta_{n-1}^T \Phi_n)\right) \tag{3.3}$$

where the initial value θ_0 is a bounded vector of $\mathbb{R}^{d+1}$ which can be arbitrarily chosen and $S_0 = I_{d+1}$. Unfortunately, we were not able to prove that $n^{-1} S_n$ converges almost surely to the Hessian matrix $\nabla^2 G(\theta)$, as well as to establish the almost sure convergence of θ_n to θ. This is the reason why we slightly modify our strategy by proposing a truncated version of previous estimates given, for all $n \geqslant 1$, by

$$\widehat{a}_n = \pi(\widehat{\theta}_{n-1}^T \Phi_n)\left(1 - \pi(\widehat{\theta}_{n-1}^T \Phi_n)\right) \tag{3.4}$$

$$\widehat{\theta}_n = \widehat{\theta}_{n-1} + S_{n-1}^{-1} \Phi_n \left(Y_n - \pi(\widehat{\theta}_{n-1}^T \Phi_n)\right) \tag{3.5}$$

$$S_n^{-1} = S_{n-1}^{-1} - \alpha_n(1 + \alpha_n \Phi_n^T S_{n-1}^{-1} \Phi_n)^{-1} S_{n-1}^{-1} \Phi_n \Phi_n^T S_{n-1}^{-1} \tag{3.6}$$

where the initial value $\widehat{\theta}_0$ is a bounded vector of $\mathbb{R}^{d+1}$ which can be arbitrarily chosen, $S_0 = I_{d+1}$ and (α_n) is a sequence of random variable defined, for some positive constant c_α, by

$$\alpha_n = \max\left\{\widehat{a}_n, \frac{c_\alpha}{n^\beta}\right\} = \max\left\{\frac{1}{4\left(\cosh\left(\Phi_n^T \widehat{\theta}_{n-1}/2\right)\right)^2}, \frac{c_\alpha}{n^\beta}\right\} \tag{3.7}$$

with $\beta \in]0, 1/2[$. From now on and for the sake of simplicity, we assume that $c_\alpha \leqslant 1/4$. It immediately implies that, for all $n \geqslant 1$, $\alpha_n \leqslant 1/4$. However, the proofs remains true for any $c_\alpha > 0$. We already saw in Section 1 that S_n^{-1} coincides with the exact inverse of the weighted matrix S_n given, for all $n \geqslant 1$, by

$$S_n = \sum_{k=1}^{n} \alpha_k \Phi_k \Phi_k^T + I_{d+1}. \tag{3.8}$$

Moreover, we will see in Section 4 that, even with this truncation of the estimate of the Hessian matrix, $n^{-1} S_n$ converges almost surely to the Hessian matrix $\nabla^2 G(\theta)$. Consequently, we will still have an optimal asymptotic behavior of the estimator $\widehat{\theta}_n$ of θ.

4. Main results

Our first result deals with the almost sure convergence of our estimates of θ and the Hessian matrix $\nabla^2 G(\theta)$. For all $n \geqslant 1$, denote

$$\overline{S}_n = \frac{1}{n} S_n.$$

Theorem 4.1. *Assume that* ***(A1)*** *and* ***(A2)*** *are satisfied. Then, we have the almost sure convergences*

$$\lim_{n\to\infty} \widehat{\theta}_n = \theta \qquad a.s \tag{4.1}$$

$$\lim_{n\to\infty} \overline{S}_n = \nabla^2 G(\theta) \qquad a.s \tag{4.2}$$

We now focus on the almost sure rates of convergence of our estimate of θ.

Theorem 4.2. *Assume that* ***(A1)*** *and* ***(A2)*** *are satisfied. Then, we have for all* $\gamma > 0$,

$$\left\|\widehat{\theta}_n - \theta\right\|^2 = o\left(\frac{(\log n)^{1+\gamma}}{n}\right) \qquad a.s \tag{4.3}$$

Moreover, suppose the random vector Φ *has a finite moment of order* > 2. *Then, we have*

$$\left\|\widehat{\theta}_n - \theta\right\|^2 = O\left(\frac{\log n}{n}\right) \qquad a.s \tag{4.4}$$

The almost sure rates of convergence of our estimate of the Hessian matrix $\nabla^2 G(\theta)$ and its inverse are as follows.

Theorem 4.3. *Assume that* ***(A1)*** *and* ***(A2)*** *are satisfied and that the random vector* Φ *has a finite moment of order* 4. *Then, we have for all* $0 < \beta < 1/2$,

$$\left\|\overline{S}_n - \nabla^2 G(\theta)\right\|^2 = O\left(\frac{1}{n^{2\beta}}\right) \qquad a.s \tag{4.5}$$

In addition, we also have

$$\left\|\overline{S}_n^{-1} - \left(\nabla^2 G(\theta)\right)^{-1}\right\|^2 = O\left(\frac{1}{n^{2\beta}}\right) \qquad a.s \tag{4.6}$$

Proof. The proofs of Theorems 4.1, 4.2 and 4.3 are given in Section 6. □

Remark 4.1. *One can observe that we do not obtain the parametric rate* $1/n$ *for these estimates. This is due to the truncation* α_n *which slightly modifies our estimation procedure. However, without this truncation, we were not able to establish the almost sure convergence of any estimate. Finally, the last result* (4.6) *ensures that our estimation procedure performs pretty well and that the estimator* $\widehat{\theta}_n$ *has an optimal asymptotic behavior.*

Theorem 4.4. *Assume that* ***(A1)*** *and* ***(A2)*** *are satisfied and that the random vector* Φ *has a finite moment of order* 4*. Then, we have the asymptotic normality*

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow[\mathrm{n}\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \left(\nabla^2 G\left(\theta\right)\right)^{-1}\right). \tag{4.7}$$

Proof. The proof of Theorem 4.4 is given in Section 7. □

Remark 4.2. *We deduce from* (4.2) *and* (4.7) *that*

$$\left(\widehat{\theta}_n - \theta\right)^T S_n \left(\widehat{\theta}_n - \theta\right) \xrightarrow[\mathrm{n}\to\infty]{\mathcal{L}} \chi^2(d+1). \tag{4.8}$$

Convergence (4.8) *allows us to build confidence regions for the parameter* θ*. Moreover, for any vector* $w \in \mathbb{R}^{d+1}$ *different from zero, we also have*

$$\frac{w^T\left(\widehat{\theta}_n - \theta\right)}{\sqrt{w^T S_n^{-1} w}} \xrightarrow[\mathrm{n}\to\infty]{\mathcal{L}} \mathcal{N}(0,1). \tag{4.9}$$

Confidence intervals and significance tests for the components of θ *can be designed from* (4.9)*. One can observe that our stochastic Newton algorithm has the same asymptotic behavior as the averaged version of a stochastic gradient algorithm* [5, 7, 13].

5. Numerical experiments

The goal of this section is to illustrate the asymptotic behavior of the truncated stochastic Newton algorithm (TSN) defined by equation (3.5). For that purpose, we will focus on the model introduced in [3] and used for comparing several gradient algorithms. We shall compare the numerical performances of the TSN algorithm with those obtained with three different algorithms : the Stochastic Newton (SN) algorithm given by equation (3.3), the stochastic gradient algorithm (SG), and the averaged stochastic gradient algorithm (ASG). Let us mention that simulations were carried out using the statistical software R.

5.1. Experiment model.

We focus on the model introduced in [3], defined by

$$\mathcal{L}(Y|\Phi) = \mathcal{B}\big(\pi(\theta^T\Phi)\big)$$

where $\Phi = (1, X^T)^T$ and X is a random vector of $\mathbb{R}^d$ with $d = 10$ with independent coordinates uniformly distributed on the interval $[0,1]$. Moreover the unknown parameter $\theta = (-9, 0, 3, -9, 4, -9, 15, 0, -7, 1, 0)^T$. This model is particularly interresting since it leads to a Hessian matrix $\nabla^2 G(\theta)$ with eigenvalues of different order sizes. Indeed, one can see in Table 1 that the smallest eigenvalue of $\nabla^2 G(\theta)$ is close to 4.422 10^{-4} while its largest eigenvalue is close to 0.1239.

0.1239	$2.832\ 10^{-3}$	$2.822\ 10^{-3}$	$2.816\ 10^{-3}$	$2.778\ 10^{-3}$	$2.806\ 10^{-3}$
$2.651\ 10^{-3}$	$2.517\ 10^{-3}$	$2.1567\ 10^{-3}$	$9.012\ 10^{-4}$	$4.422\ 10^{-4}$	

TABLE 1. Estimated eigenvalues of $\nabla^2 G(\theta)$ arranged in decreasing order.

5.2. **Comparison of the different algorithms.** Our comparisons are based on the mean squared error (MSE) defined, for all estimate $\widehat{\theta}_n$ of θ, by

$$\mathbb{E}\left[\left\|\widehat{\theta}_n - \theta\right\|^2\right].$$

We simulate $N = 400$ samples wth a maximum number of iterations $n = 5\,000$. For each sample, we estimate the unknown parameter θ using the four algorithms (TSN, SN, SG, ASG) which are initialized identically by choosing the initial value $\widehat{\theta}_0$ uniformly in a compact subset containing the true value θ. For the TSN and SN algorithms, we take $S_0 = I_{d+1}$. In addition, for the TSN algorithm, we choose the truncation term defined by $c_\alpha = 10^{-10}$ and $\beta = 0.49$. Finally, to be fairplay, we choose the best step sequence for the SG algorithm with the help of a cross validation method. Figure 1 shows the decreasing behavior of the MSE, calculated for the four algorithms, as the number of iterations n grows from 1 to 5 000.

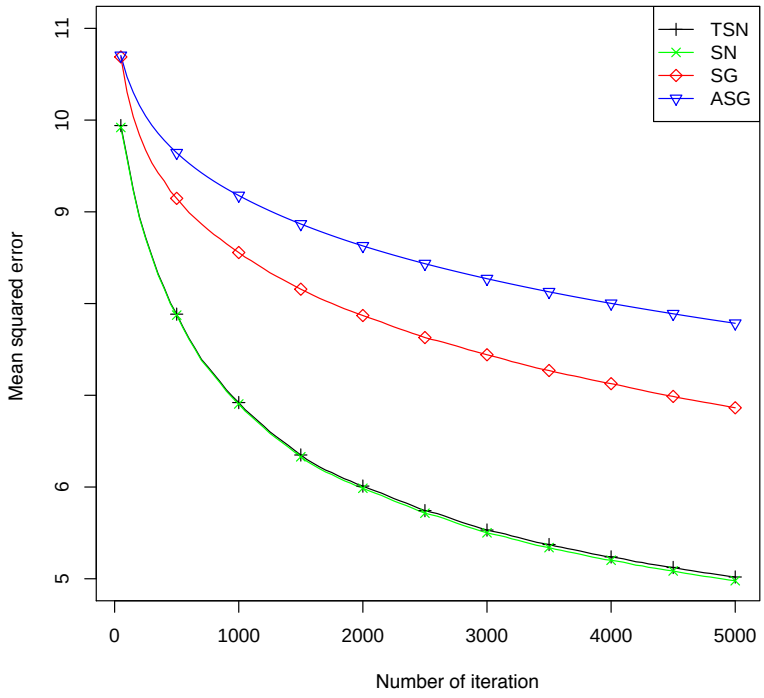


FIGURE 1. Mean squared error of the four algorithms.

It is clear that the stochastic Newton algorithms perform much more better than the stochastic gradient algorithms The bad behavior of the stochastic gradient algorithms is certainly due to the fact that the eigenvalues of the Hessian matrix $\nabla^2 G(\theta)$ are at different scales. One can also observe that it is quite useless to average the SG algorithm.

On can find in Figure 2 the boxplots of the $N = 400$ values of the squared error $\|\widehat{\theta}_n - \theta\|^2$ computed for the TSN and SN algorithms, as well as for the deterministic Newton-Raphson algorithm (NR).

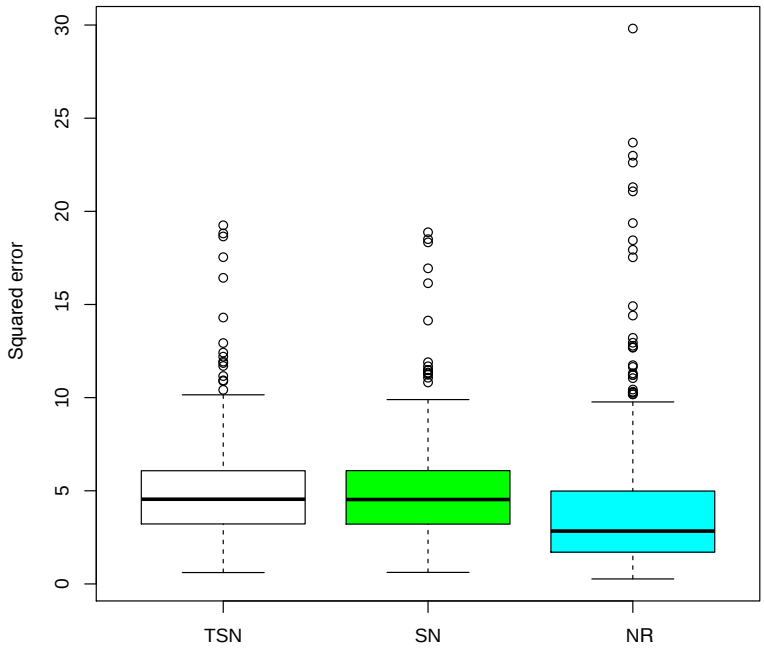


FIGURE 2. Boxplots of the squared error for the TSN, SN and the NR algorithms.

5.3. **Some comments concerning the truncation.** To close this section, let us make some comments concerning the truncation term introduced in the TSN algorithm. This short numerical experiment tends to show that the use of the truncation is artificial and useless. Indeed, one can take the constant c_α in (3.7) as small as possible and see that the TSN and SN algorithms match. Finally, an inappropriate choice of c_α can lead to a poor numerical behavior of the TSN algorithm.

6. Proofs of the almost sure convergence results

6.1. **Two technical lemmas.** We start the proofs of the almost sure convergence results with two technical lemmas.

Lemma 6.1. *Assume that the random vector Φ has a finite moment of order 2. Then, we have the almost sure convergence for all $0 < \beta < 1$,*

$$\lim_{n\to\infty} \frac{1}{\sum_{k=1}^{n} k^{-\beta}} \sum_{k=1}^{n} k^{-\beta} \Phi_k \Phi_k^T = \mathbb{E}\left[\Phi\Phi^T\right] \qquad a.s \tag{6.1}$$

Remark 6.1. *We obtain from (3.7) together with (3.8) that for all $n \geqslant 1$,*

$$S_n \geqslant c_\alpha \sum_{k=1}^{n} k^{-\beta} \Phi_k \Phi_k^T. \tag{6.2}$$

Denote by $\lambda_{min}(S_n)$ the minimum eigenvalue of the positive definite matrix S_n. We immediately obtain from (6.1) and (6.2) that for n large enough

$$\lambda_{min}(S_n) \geqslant \frac{\lambda c_\alpha}{2} \sum_{k=1}^{n} k^{-\beta} \geqslant \frac{\lambda c_\alpha}{2} n^{1-\beta} \qquad a.s$$

*where λ stands for the minimum eigenvalue of the positive definite deterministic matrix $\mathbb{E}\left[\Phi\Phi^T\right]$. Consequently, we have under assumption **(A1)** that for n large enough,*

$$\left(\lambda_{min}(S_n)\right)^{-2} \leqslant \frac{4}{\lambda^2 c_\alpha^2} \frac{1}{n^{2(1-\beta)}} \qquad a.s$$

Therefore, as soon as $\beta \in]0, 1/2[$,

$$\sum_{n=1}^{\infty} \left(\lambda_{\min}\left(S_n\right)\right)^{-2} < \infty \qquad a.s \tag{6.3}$$

Proof. It follows from a straightforward Abel transform calculation that

$$\begin{aligned} \sum_{k=1}^{n} k^{-\beta} \Phi_k \Phi_k^T &= \sum_{k=1}^{n} k^{-\beta} \left(\Sigma_k - \Sigma_{k-1}\right), \\ &= n^{-\beta} \Sigma_n + \sum_{k=1}^{n-1} \left(k^{-\beta} - (k+1)^{-\beta}\right) \Sigma_k, \\ &= n^{-\beta} \Sigma_n + \sum_{k=1}^{n-1} b_k k^{-1} \Sigma_k \end{aligned} \tag{6.4}$$

where $\Sigma_0 = 0$ and for all $n \geqslant 1$,

$$\Sigma_n = \sum_{k=1}^{n} \Phi_k \Phi_k^T \qquad \text{and} \qquad b_n = n\left(n^{-\beta} - (n+1)^{-\beta}\right).$$

On the one hand, we obtain from the standard strong law of large numbers that

$$\lim_{n\to\infty} \frac{1}{n}\Sigma_n = \mathbb{E}\left[\Phi\Phi^T\right] \qquad \text{a.s} \tag{6.5}$$

On the other hand,

$$\sum_{k=1}^{n} b_k = \sum_{k=1}^{n} k^{-\beta} - n^{1-\beta}$$

which implies that

$$\lim_{n\to\infty} \frac{1}{n^{1-\beta}} \sum_{k=1}^{n} b_k = \frac{\beta}{1-\beta}.$$

Then, we deduce from Toeplitz's lemma given e.g. in ([4], page 54) that

$$\lim_{n\to\infty} \frac{1}{n^{1-\beta}} \sum_{k=1}^{n-1} b_k k^{-1}\Sigma_k = \frac{\beta}{1-\beta}\mathbb{E}\left[\Phi\Phi^T\right] \qquad \text{a.s} \tag{6.6}$$

Consequently, we obtain from (6.4) together with (6.5) and (6.6) that

$$\lim_{n\to\infty} \frac{1}{n^{1-\beta}} \sum_{k=1}^{n} k^{-\beta}\Phi_k\Phi_k^T = \frac{1}{1-\beta}\mathbb{E}\left[\Phi\Phi^T\right] \qquad \text{a.s}$$

which immediately leads to (6.1). □

Our second lemma concerns a useful Lipschitz property of the function α defined, for all $h, \ell \in \mathbb{R}^{d+1}$, by

$$\alpha\left(h,\ell\right) = \pi\big(h^T\ell\big)\big(1-\pi\big(h^T\ell\big)\big) = \frac{1}{4\left(\cosh\left(h^T\ell/2\right)\right)^2}. \tag{6.7}$$

Lemma 6.2. *For all $h, \ell, \ell' \in \mathbb{R}^{d+1}$, we have*

$$\left|\alpha\left(h,\ell\right) - \alpha\left(h,\ell'\right)\right| \leqslant \frac{1}{12\sqrt{3}} \left\|h\right\| \left\|\ell - \ell'\right\|. \tag{6.8}$$

Proof. Let φ be the function defined, for all $x \in \mathbb{R}$, by

$$\varphi(x) = \pi(x)(1-\pi(x)) = \frac{\exp(x)}{(1+\exp(x))^2} = \frac{1}{4(\cosh(x/2))^2}.$$

We clearly have

$$\begin{aligned}
\varphi'(x) &= \frac{\exp(x)(1-\exp(x))}{(1+\exp(x))^3}, \\
\varphi''(x) &= \frac{\exp(x)((\exp(x))^2 - 4\exp(x) + 1)}{(1+\exp(x))^4}.
\end{aligned}$$

It is not hard to see that for all $x \in \mathbb{R}$,

$$|\varphi'(x)| \leqslant \frac{1}{6\sqrt{3}}. \tag{6.9}$$

Hence, it follows from (6.9) together with the mean value theorem that for all $x, y \in \mathbb{R}$,

$$|\varphi(x) - \varphi(y)| \leqslant \frac{1}{6\sqrt{3}}|x - y|. \tag{6.10}$$

Consequently, we obtain from (6.10) that for all $h, \ell, \ell' \in \mathbb{R}^{d+1}$,

$$\left|\alpha\left(h, \ell\right) - \alpha\left(h, \ell'\right)\right| \leqslant \frac{1}{12\sqrt{3}}\left|h^T(\ell - \ell')\right| \leqslant \frac{1}{12\sqrt{3}}\left\|h\right\|\left\|\ell - \ell'\right\|$$

which completes the proof of Lemma 6.2. □

6.2. **Proof of Theorem 4.1.** We are now in the position to proceed to the proof of the almost sure convergence (4.1). By a Taylor expansion of the twice continuously differentiable functional G, there exists $\xi_n \in \mathbb{R}^{d+1}$ such that

$$G\big(\widehat{\theta}_{n+1}\big) = G\big(\widehat{\theta}_n\big) + \nabla G\big(\widehat{\theta}_n\big)^T\big(\widehat{\theta}_{n+1} - \widehat{\theta}_n\big) + \frac{1}{2}\big(\widehat{\theta}_{n+1} - \widehat{\theta}_n\big)^T\nabla^2 G(\xi_n)\big(\widehat{\theta}_{n+1} - \widehat{\theta}_n\big). \tag{6.11}$$

We clearly have from (2.3) that

$$\left\|\nabla^2 G(\xi_n)\right\| \leqslant \frac{1}{4}\mathbb{E}\left[\|\Phi\|^2\right].$$

Hence, we obtain from (3.5) together with (6.11) that

$$\begin{aligned} G\big(\widehat{\theta}_{n+1}\big) &\leqslant G\big(\widehat{\theta}_n\big) + \nabla G\big(\widehat{\theta}_n\big)^T\big(\widehat{\theta}_{n+1} - \widehat{\theta}_n\big) + \frac{1}{8}\mathbb{E}\left[\|\Phi\|^2\right]\left\|\widehat{\theta}_{n+1} - \widehat{\theta}_n\right\|^2, \\ &= G\big(\widehat{\theta}_n\big) - \nabla G\big(\widehat{\theta}_n\big)^T S_n^{-1} Z_{n+1} + \frac{1}{8}\mathbb{E}\left[\|\Phi\|^2\right]\left\|S_n^{-1} Z_{n+1}\right\|^2, \\ &\leqslant G\big(\widehat{\theta}_n\big) - \nabla G\big(\widehat{\theta}_n\big)^T S_n^{-1} Z_{n+1} + \frac{1}{8}\mathbb{E}\left[\|\Phi\|^2\right]\left(\lambda_{\min}\left(S_n\right)\right)^{-2}\left\|Z_{n+1}\right\|^2 \end{aligned}$$

where $Z_{n+1} = \nabla_h g\big(\Phi_{n+1}, Y_{n+1}, \widehat{\theta}_n\big)$. Since $\|Z_{n+1}\| \leqslant \|\Phi_{n+1}\|$, it implies that

$$G\big(\widehat{\theta}_{n+1}\big) \leqslant G\big(\widehat{\theta}_n\big) - \nabla G\big(\widehat{\theta}_n\big)^T S_n^{-1} Z_{n+1} + \frac{1}{8}\mathbb{E}\left[\|\Phi\|^2\right]\left(\lambda_{\min}\left(S_n\right)\right)^{-2}\left\|\Phi_{n+1}\right\|^2. \tag{6.12}$$

Let $\mathbb{F} = (\mathcal{F}_n)$ be the filtration given, for all $n \geqslant 1$, by $\mathcal{F}_n = \sigma\left(\left(\Phi_1, Y_1\right), \ldots, \left(\Phi_n, Y_n\right)\right)$. We clearly have $\mathbb{E}\left[Z_{n+1}|\mathcal{F}_n\right] = \nabla G\big(\widehat{\theta}_n\big)$. Consequently, we obtain from (6.12) that

$$\mathbb{E}\big[G\big(\widehat{\theta}_{n+1}\big)|\mathcal{F}_n\big] \leqslant G\big(\widehat{\theta}_n\big) - \nabla G\big(\widehat{\theta}_n\big)^T S_n^{-1}\nabla G\big(\widehat{\theta}_n\big) + \frac{1}{8}\big(\mathbb{E}\left[\|\Phi\|^2\right]\big)^2\big(\lambda_{\min}\left(S_n\right)\big)^{-2} \quad \text{a.s.} \tag{6.13}$$

Our goal is now to apply the Robbins-Siegmund theorem ([4], page 18) to the three positive sequences (V_n), (A_n) and (B_n) given by $V_n = G\big(\widehat{\theta}_n\big)$,

$$A_n = \frac{1}{8}\big(\mathbb{E}\left[\|\Phi\|^2\right]\big)^2\big(\lambda_{\min}\left(S_n\right)\big)^{-2} \qquad \text{and} \qquad B_n = \left\|S_n^{-1/2}\nabla G\big(\widehat{\theta}_n\big)\right\|^2.$$

It clearly follows from (6.13) that

$$\mathbb{E}\big[V_{n+1}|\mathcal{F}_n\big] \leqslant V_n + A_n - B_n \qquad \text{a.s.}$$

Moreover, we already saw from (6.3) that

$$\sum_{n=1}^{\infty} A_n < \infty \qquad \text{a.s} \tag{6.14}$$

Consequently, we can deduce from the Robbins-Siegmund theorem that (V_n) convergences almost surely to a finite random variable and

$$\sum_{n=1}^{\infty} B_n < \infty \qquad \text{a.s} \tag{6.15}$$

Furthermore, since $B_n \geqslant \left(\lambda_{\max}\left(S_n\right)\right)^{-1} \left\|\nabla G\left(\widehat{\theta}_n\right)\right\|^2$, we get from (6.15) that

$$\sum_{n=1}^{\infty} \left(\lambda_{\max}\left(S_n\right)\right)^{-1} \left\|\nabla G\left(\widehat{\theta}_n\right)\right\|^2 < \infty \qquad \text{a.s} \tag{6.16}$$

In addition, we obtain from (3.7) together with (3.8) that

$$\lambda_{\max}\left(S_n\right) \leqslant 1 + \frac{1}{4}\lambda_{\max}\Big(\sum_{k=1}^{n} \Phi_k \Phi_k^T\Big)$$

since, for all $n \geqslant 1$, $\alpha_n \leqslant 1/4$. Therefore, (6.5) ensures that for n large enough

$$\lambda_{max}(S_n) \leqslant \Lambda n \qquad \text{a.s}$$

where Λ is for the maximum eigenvalue of the positive definite deterministic matrix $\mathbb{E}\big[\Phi\Phi^T\big]$. It implies that

$$\sum_{n=1}^{\infty} \left(\lambda_{\max}\left(S_n\right)\right)^{-1} = +\infty \qquad \text{a.s} \tag{6.17}$$

Hence, it follows from the conjunction of (6.16) and (6.17) that $\nabla G\big(\widehat{\theta}_n\big)$ converges to 0 almost surely. It means that $\widehat{\theta}_n$ converges almost surely to the unique zero θ of the gradient, which is exactly what we wanted to prove. It remains to prove the almost sure convergence (4.2). We infer from (3.8) that

$$\overline{S}_n = \frac{1}{n}\sum_{k=1}^{n}\big(\alpha_k - \alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big)\big)\Phi_k\Phi_k^T + \frac{1}{n}\sum_{k=1}^{n}\alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big)\Phi_k\Phi_k^T + \frac{1}{n}I_{d+1}. \tag{6.18}$$

We now give the convergence of the two terms on the right-hand side of (6.18). For the first one, one can observe that $\alpha_n = \alpha\big(\Phi_n, \widehat{\theta}_{n-1}\big)$ as soon as $\alpha\big(\Phi_n, \widehat{\theta}_{n-1}\big) \geqslant c_\alpha n^{-\beta}$.

Consequently,

$$\frac{1}{n}\sum_{k=1}^{n}\big(\alpha_k - \alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big)\big)\Phi_k\Phi_k^T = \frac{1}{n}\sum_{k=1}^{n}\big(\alpha_k - \alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big)\big)\Phi_k\Phi_k^T \, \mathrm{I}_{\left\{\alpha(\Phi_k, \widehat{\theta}_{k-1}) \leqslant c_\alpha k^{-\beta}\right\}}.$$

It implies that

$$\left\|\frac{1}{n}\sum_{k=1}^{n}\big(\alpha_k - \alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big)\big)\Phi_k\Phi_k^T\right\| \leqslant \frac{c_\alpha}{n}\sum_{k=1}^{n}\frac{1}{k^\beta}\big\|\Phi_k\big\|^2.$$

However, one can easily check from Lemma 6.1 that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\frac{1}{k^\beta}\big\|\Phi_k\big\|^2 = 0 \qquad \text{a.s}$$

It means that the first term on the right-hand side of (6.18) goes to 0 almost surely. We now study the convergence of the second term on the right-hand side of (6.18) which can be rewritten as

$$\frac{1}{n}\sum_{k=1}^{n}\alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big)\Phi_k\Phi_k^T = \frac{1}{n}\sum_{k=1}^{n}\alpha\big(\Phi_k, \theta\big)\Phi_k\Phi_k^T. + \frac{1}{n}\sum_{k=1}^{n}\big(\alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big) - \alpha\big(\Phi_k, \theta\big)\big)\Phi_k\Phi_k^T$$

On the one hand, thanks to the standard strong law of large numbers, we clearly have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\alpha\big(\Phi_k, \theta\big)\Phi_k\Phi_k^T = \mathbb{E}\big[\alpha\big(\Phi, \theta\big)\Phi\Phi^T\big] = \nabla^2 G(\theta) \qquad \text{a.s} \tag{6.19}$$

On the other hand, denote by R_n the remainder

$$R_n = \sum_{k=1}^{n}\big(\alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big) - \alpha\big(\Phi_k, \theta\big)\big)\Phi_k\Phi_k^T.$$

We can split R_n into two terms $R_n = P_n + Q_n$ where, for some positive constant M,

$$\begin{aligned} P_n &= \sum_{k=1}^{n}\big(\alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big) - \alpha\big(\Phi_k, \theta\big)\big)\Phi_k\Phi_k^T \, \mathrm{I}_{\left\{\|\Phi_k\| \leqslant M\right\}} \\ Q_n &= \sum_{k=1}^{n}\big(\alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big) - \alpha\big(\Phi_k, \theta\big)\big)\Phi_k\Phi_k^T \, \mathrm{I}_{\left\{\|\Phi_k\| > M\right\}}. \end{aligned}$$

It follows from the Lipschitz property of the function α given is Lemma 6.2 that

$$\left\|\frac{1}{n}P_n\right\| \leqslant \frac{M}{12\sqrt{3}}\frac{1}{n}\sum_{k=1}^{n}\big\|\widehat{\theta}_{k-1} - \theta\big\| \, \big\|\Phi_k\big\|^2.$$

Hence, we deduce from (4.1) together with (6.5) that

$$\lim_{n\to\infty}\frac{1}{n}P_n = 0 \qquad \text{a.s} \tag{6.20}$$

Furthermore, we also have

$$\left\|\frac{1}{n}Q_n\right\| \leqslant \frac{1}{2n}\sum_{k=1}^{n}\|\Phi_k\|^2\,\mathrm{I}_{\left\{\|\Phi_k\|>M\right\}}.$$

We deduce once again from the strong law of large numbers that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\|\Phi_k\|^2\,\mathrm{I}_{\left\{\|\Phi_k\|>M\right\}} = \mathbb{E}\left[\|\Phi\|^2\,\mathrm{I}_{\left\{\|\Phi\|>M\right\}}\right] \qquad \text{a.s}$$

which implies via (6.20) that for any positive constant M,

$$\limsup_{n\to\infty}\left\|\frac{1}{n}R_n\right\| \leqslant \frac{1}{2}\mathbb{E}\left[\|\Phi\|^2\,\mathrm{I}_{\left\{\|\Phi\|>M\right\}}\right] \qquad \text{a.s} \tag{6.21}$$

Nonetheless, we obtain from the Lebesgue dominated convergence theorem that

$$\lim_{M\to\infty}\mathbb{E}\left[\|\Phi\|^2\,\mathrm{I}_{\left\{\|\Phi\|>M\right\}}\right] = 0.$$

Consequently, we find from (6.21)

$$\lim_{n\to\infty}\frac{1}{n}R_n = 0 \qquad \text{a.s}$$

Finally, (4.2) follows from (6.18) and (6.19), which achieves the proof of Theorem 4.1. □

6.3. **Proof of Theorem 4.2.** It follows from equation (3.5) that for all $n \geqslant 1$,

$$\widehat{\theta}_{n+1} - \theta = \widehat{\theta}_n - \theta - \frac{1}{n}\left(\overline{S}_n^{-1} - S^{-1}\right)Z_{n+1} - \frac{1}{n}S^{-1}Z_{n+1}$$

where $Z_{n+1} = \nabla_h g\left(\Phi_{n+1}, Y_{n+1}, \widehat{\theta}_n\right)$ and $S = \nabla^2 G(\theta)$. Consequently,

$$\widehat{\theta}_{n+1} - \theta = \widehat{\theta}_n - \theta - \frac{1}{n}\left(\overline{S}_n^{-1} - S^{-1}\right)Z_{n+1} - \frac{1}{n}S^{-1}\left(\nabla G\left(\widehat{\theta}_n\right) + \varepsilon_{n+1}\right) \tag{6.22}$$

where $\varepsilon_{n+1} = Z_{n+1} - \nabla G\left(\widehat{\theta}_n\right)$. We already saw that $\mathbb{E}\left[Z_{n+1}|\mathcal{F}_n\right] = \nabla G\left(\widehat{\theta}_n\right)$ which clearly implies that (ε_n) is a martingale difference sequence, $\mathbb{E}\left[\varepsilon_{n+1}|\mathcal{F}_n\right] = 0$. Denote by δ_n the remainder of the Taylor's expansion of the gradient

$$\delta_n = \nabla G\left(\widehat{\theta}_n\right) - \nabla^2 G(\theta)\left(\widehat{\theta}_n - \theta\right) = \nabla G\left(\widehat{\theta}_n\right) - S\left(\widehat{\theta}_n - \theta\right).$$

We deduce from (6.22) that for all $n \geqslant 1$,

$$\widehat{\theta}_{n+1} - \theta = \left(1 - \frac{1}{n}\right)\left(\widehat{\theta}_n - \theta\right) - \frac{1}{n}\left(\overline{S}_n^{-1} - S^{-1}\right)Z_{n+1} - \frac{1}{n}S^{-1}\left(\delta_n + \varepsilon_{n+1}\right), \tag{6.23}$$

which leads to

$$\begin{aligned}
\widehat{\theta}_{n+1} - \theta &= -\frac{1}{n}\sum_{k=1}^{n}\Big(\overline{S}_k^{-1} - S^{-1}\Big)Z_{k+1} - \frac{1}{n}S^{-1}\sum_{k=1}^{n}\Big(\delta_k + \varepsilon_{k+1}\Big), \\
&= -\frac{1}{n}\sum_{k=1}^{n}\Big(\overline{S}_k^{-1} - S^{-1}\Big)\Big(\varepsilon_{k+1} + \nabla G\big(\widehat{\theta}_k\big)\Big) - \frac{1}{n}S^{-1}\sum_{k=1}^{n}\Big(\delta_k + \varepsilon_{k+1}\Big), \\
&= -\frac{1}{n}M_{n+1} - \Delta_n \qquad (6.24)
\end{aligned}$$

where

$$M_{n+1} = \sum_{k=1}^{n}\overline{S}_k^{-1}\varepsilon_{k+1}$$

and

$$\begin{aligned}
\Delta_n &= \frac{1}{n}\sum_{k=1}^{n}\Big(\overline{S}_k^{-1} - S^{-1}\Big)\nabla G\big(\widehat{\theta}_k\big) + \frac{1}{n}S^{-1}\sum_{k=1}^{n}\delta_k, \\
&= \frac{1}{n}\sum_{k=1}^{n}\Big(\overline{S}_k^{-1} - S^{-1}\Big)S\Big(\widehat{\theta}_k - \theta\Big) + \frac{1}{n}\sum_{k=1}^{n}\overline{S}_k^{-1}\delta_k. \qquad (6.25)
\end{aligned}$$

We claim that the remainder δ_n is negligeable. As a matter of fact,

$$\begin{aligned}
\|\delta_n\| &= \Big\|\int_0^1 \nabla^2 G\big(\theta + t\big(\widehat{\theta}_n - \theta\big)\big)\big(\widehat{\theta}_n - \theta\big)dt - \nabla^2 G(\theta)\big(\widehat{\theta}_n - \theta\big)\Big\|, \\
&\leqslant \int_0^1 \Big\|\nabla^2 G\big(\theta + t\big(\widehat{\theta}_n - \theta\big)\big) - \nabla^2 G(\theta)\Big\|dt\big\|\widehat{\theta}_n - \theta\big\|.
\end{aligned}$$

However, the functional G is twice continuously differentiable and $\widehat{\theta}_n$ converges almost surely to θ, which ensures that

$$\|\delta_n\| = o\big(\big\|\widehat{\theta}_n - \theta\big\|\big) \qquad \text{a.s} \qquad (6.26)$$

Then, we obtain from (4.2), (6.25) and (6.26) that it exists a constant $0 < c < 1/2$ and a finite positive random variable D such that for all $n \geqslant 1$,

$$\|\Delta_n\| \leqslant cL_n + \frac{1}{n}D \qquad \text{a.s} \qquad (6.27)$$

where

$$L_n = \frac{1}{n}\sum_{k=1}^{n}\|\widehat{\theta}_k - \theta\|.$$

Therefore, we deduce from (6.24) and (6.27) that for all $n \geqslant 1$,

$$\begin{aligned} L_{n+1} &= \Big(1 - \frac{1}{n+1}\Big) L_n + \frac{1}{n+1} \|\widehat{\theta}_{n+1} - \theta\|, \\ &\leqslant \Big(1 - \frac{1}{n+1}\Big) L_n + \frac{1}{n+1} \Big(\frac{1}{n}\|M_{n+1}\| + \|\Delta_n\|\Big) \qquad \text{a.s} \\ &\leqslant \Big(1 - \frac{d}{n+1}\Big) L_n + \frac{1}{n(n+1)} \Big(\|M_{n+1}\| + D\Big) \qquad \text{a.s} \end{aligned}$$

where $d = 1 - c$. It clearly implies by induction that for all $n \geqslant 1$,

$$L_n \leqslant \prod_{k=2}^{n} \Big(1 - \frac{d}{k}\Big) L_1 + \sum_{k=2}^{n} \prod_{i=k+1}^{n} \Big(1 - \frac{d}{i}\Big) \frac{1}{k(k+1)} \Big(\|M_{k+1}\| + D\Big) \qquad \text{a.s} \tag{6.28}$$

Hereafter, we shall proceed to the evaluation of the right-hand side term in (6.28). The sequence (M_n) is a locally square-integrable multi-dimensional martingale with predictable quadratic variation given by

$$\langle M \rangle_n = \sum_{k=2}^{n} \overline{S}_{k-1}^{-1} \mathbb{E}\big[\varepsilon_k \varepsilon_k^T | \mathcal{F}_{k-1}\big] \overline{S}_{k-1}^{-1}. \tag{6.29}$$

However, for all $n \geqslant 1$,

$$\begin{aligned} \mathbb{E}\big[\varepsilon_{n+1}\varepsilon_{n+1}^T | \mathcal{F}_n\big] &= \mathbb{E}\Big[Z_{n+1} Z_{n+1}^T | \mathcal{F}_n\Big] - \nabla G\big(\widehat{\theta}_n\big) \nabla G\big(\widehat{\theta}_n\big)^T \\ &= \mathbb{E}\Big[\Big(\pi\big(\widehat{\theta}_n^T \Phi_{n+1}\big) - Y_{n+1}\Big)^2 \Phi_{n+1}\Phi_{n+1}^T | \mathcal{F}_n\Big] - \nabla G\big(\widehat{\theta}_n\big) \nabla G\big(\widehat{\theta}_n\big)^T \\ &= \mathbb{E}\Big[\Big(\pi\big(\widehat{\theta}_n^T \Phi_{n+1}\big) - \pi\big(\theta^T \Phi_{n+1}\big) + \pi\big(\theta^T \Phi_{n+1}\big) - Y_{n+1}\Big)^2 \Phi_{n+1}\Phi_{n+1}^T | \mathcal{F}_n\Big] \\ &- \nabla G\big(\widehat{\theta}_n\big) \nabla G\big(\widehat{\theta}_n\big)^T. \end{aligned}$$

Since $\mathbb{E}\big[Y_{n+1} | \Phi_{n+1}\big] = \pi\left(\theta^T \Phi_{n+1}\right)$, we obtain that for all $n \geqslant 1$,

$$\begin{aligned} \mathbb{E}\big[\varepsilon_{n+1}\varepsilon_{n+1}^T | \mathcal{F}_n\big] &= \mathbb{E}\Big[\Big(\pi\big(\widehat{\theta}_n^T \Phi_{n+1}\big) - \pi\big(\theta^T \Phi_{n+1}\big)\Big)^2 \Phi_{n+1}\Phi_{n+1}^T | \mathcal{F}_n\Big] \\ &+ \mathbb{E}\Big[\Big(\pi\big(\theta^T \Phi_{n+1}\big) - Y_{n+1}\big)\Big)^2 \Phi_{n+1}\Phi_{n+1}^T | \mathcal{F}_n\Big] - \nabla G\big(\widehat{\theta}_n\big) \nabla G\big(\widehat{\theta}_n\big)^T \\ &= \mathbb{E}\Big[\Big(\pi\big(\widehat{\theta}_n^T \Phi_{n+1}\big) - \pi\big(\theta^T \Phi_{n+1}\big)\Big)^2 \Phi_{n+1}\Phi_{n+1}^T | \mathcal{F}_n\Big] + \nabla^2 G(\theta) \\ &- \nabla G\big(\widehat{\theta}_n\big) \nabla G\big(\widehat{\theta}_n\big)^T. \end{aligned}$$

By continuity together with (4.1), we have the almost sure convergences

$$\lim_{n \to \infty} \nabla G\big(\widehat{\theta}_n\big) \nabla G\big(\widehat{\theta}_n\big)^T = 0 \qquad \text{a.s}$$

and

$$\lim_{n\to\infty} \mathbb{E}\Big[\Big(\pi\big(\widehat{\theta}_n^T\Phi_{n+1}\big) - \pi\big(\theta^T\Phi_{n+1}\big)\Big)^2 \Phi_{n+1}\Phi_{n+1}^T|\mathcal{F}_n\Big] = 0 \qquad \text{a.s}$$

Therefore, we obtain from (4.2) and (6.29) that

$$\lim_{n\to\infty} \frac{1}{n}\langle M\rangle_n = \big(\nabla^2 G\left(\theta\right)\big)^{-1} \qquad \text{a.s} \tag{6.30}$$

Hence, it follows from the strong law of large numbers for multi-dimensional martingales given e.g. by Theorem 4.13.16 in [4] that for any $\gamma > 0$,

$$\big\|M_n\big\|^2 = o\big(n\big(\log n\big)^{1+\gamma}\big) \qquad \text{a.s} \tag{6.31}$$

Moreover, if the random vector Φ has a finite moment of order > 2, we also have the more precise almost sure rate of convergence

$$\big\|M_n\big\|^2 = O\big(n\log n\big) \qquad \text{a.s} \tag{6.32}$$

We will prove (4.4) inasmuch as the proof for (4.3) follows essentially the same lines. We deduce from (6.32) that it exists a finite positive random variable C such that for all $n \geqslant 1$

$$\big\|M_{n+1}\big\| \leqslant C\sqrt{n\log n} \qquad \text{a.s} \tag{6.33}$$

We are now in position to find an upper-bound for inequality (6.28). Via the elementary $1 - x \leqslant \exp(-x)$, we clearly have

$$\prod_{k=2}^{n}\Big(1 - \frac{d}{k}\Big) \leqslant \Big(\frac{2}{n+1}\Big)^d \qquad \text{and} \qquad \prod_{i=k+1}^{n}\Big(1 - \frac{d}{i}\Big) \leqslant \Big(\frac{k+1}{n+1}\Big)^d.$$

Consequently, we obtain from (6.28) and (6.33) that for all $n \geqslant 1$,

$$L_n \leqslant \Big(\frac{2}{n+1}\Big)^d L_1 + \sum_{k=2}^{n}\Big(\frac{k+1}{n+1}\Big)^d \frac{1}{k(k+1)}\Big(C\sqrt{k\log k} + D\Big) \qquad \text{a.s}$$

leading to

$$L_n \leqslant \Big(\frac{2}{n}\Big)^d L_1 + \frac{A\big(\log n\big)^{1/2}}{n^d}\sum_{k=2}^{n}\frac{1}{k^a} \qquad \text{a.s} \tag{6.34}$$

where $A = \max(C, D)$ and $a = 3/2 - d = 1/2 + c$. Hereafter, we recall that the positive constant c has been chosen such that $c < 1/2$ which means that $0 < a < 1$. Hence, we

find from (6.34) that for all $n \geqslant 1$,

$$\begin{aligned} L_n &\leqslant \Big(\frac{2}{n}\Big)^d L_1 + \frac{A\big(\log n\big)^{1/2}}{n^{d+a-1}} \qquad \text{a.s} \\ &\leqslant \Big(\frac{2}{n}\Big)^d L_1 + A\Big(\frac{\log n}{n}\Big)^{1/2} \qquad \text{a.s} \end{aligned}$$

Since $d > 1/2$, it immediately implies that

$$L_n^2 = O\Big(\frac{\log n}{n}\Big) \qquad \text{a.s} \tag{6.35}$$

Then, it follows from the conjunction of (6.27) and (6.35) that

$$\|\Delta_n\|^2 = O\Big(\frac{\log n}{n}\Big) \qquad \text{a.s}$$

It ensures, via (6.24) and (6.32), that

$$\|\widehat{\theta}_n - \theta\|^2 = O\Big(\frac{\log n}{n}\Big) \qquad \text{a.s}$$

which is exactly what we wanted to prove. □

6.4. **Proof of Theorem 4.3.** First of all, it follows from (3.8) that $\overline{S}_n$ can be splitted into two terms

$$\begin{aligned} \overline{S}_n &= \frac{1}{n}\sum_{k=1}^{n} \alpha_k \Phi_k \Phi_k^T + \frac{1}{n} I_{d+1} \\ &= \frac{1}{n} T_n + \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\big[\alpha_k \Phi_k \Phi_k^T | \mathcal{F}_{k-1}\big] + \frac{1}{n} I_{d+1} \end{aligned} \tag{6.36}$$

where

$$T_n = \sum_{k=1}^{n} \alpha_k \Phi_k \Phi_k^T - \mathbb{E}\big[\alpha_k \Phi_k \Phi_k^T | \mathcal{F}_{k-1}\big].$$

The sequence (T_n) is a locally square-integrable multi-dimensional martingale. Since the random vector Φ has a finite moment of order 4 and for all $n \geqslant 1$, $\alpha_n \leqslant 1/4$, we obtain from the strong law of large numbers for multi-dimensional martingales given e.g. by Theorem 4.13.16 in [4] that for any $\gamma > 0$,

$$\big\|T_n\big\|^2 = o\big(n\big(\log n\big)^{1+\gamma}\big) \qquad \text{a.s} \tag{6.37}$$

Let us now give the rate of convergence of the second term on the right-hand side of (6.36). On the one hand, we have the decomposition

$$\sum_{k=1}^{n} \alpha_k \Phi_k \Phi_k^T = \sum_{k=1}^{n} \big(\alpha_k - \alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big)\big)\Phi_k \Phi_k^T + \sum_{k=1}^{n} \alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big)\Phi_k \Phi_k^T.$$

We already saw in the proof of Theorem 4.1 that

$$\sum_{k=1}^{n} \big(\alpha_k - \alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big)\big)\Phi_k\Phi_k^T \leqslant c_\alpha \sum_{k=1}^{n} \frac{1}{k^\beta}\Phi_k\Phi_k^T \tag{6.38}$$

Hence, by taking the conditionnal expectation on both sides of (6.38), we obtain that

$$\begin{aligned} \Big\| \sum_{k=1}^{n} \mathbb{E}\Big[\big(\alpha_k - \alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big)\big)\Phi_k\Phi_k^T | \mathcal{F}_{k-1}\Big] \Big\| &\leqslant c_\alpha \sum_{k=1}^{n} \frac{1}{k^\beta}\mathbb{E}\big[\|\Phi_k\|^2 | \mathcal{F}_{k-1}\big] \\ &\leqslant \frac{c_\alpha \mathbb{E}\big[\|\Phi\|^2\big]}{1-\beta} n^{1-\beta}. \end{aligned} \tag{6.39}$$

On the other hand,

$$\sum_{k=1}^{n} \mathbb{E}\Big[\alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big)\Phi_k\Phi_k^T | \mathcal{F}_{k-1}\Big] = \sum_{k=1}^{n} \Big(\nabla^2 G\big(\widehat{\theta}_{k-1}\big) - \nabla^2 G\big(\theta\big)\Big) + n\nabla^2 G\big(\theta\big).$$

Consequently, we immediately deduce from inequality (6.8) that

$$\begin{aligned} \Big\| \sum_{k=1}^{n} \mathbb{E}\Big[\alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big)\Phi_k\Phi_k^T | \mathcal{F}_{k-1}\Big] - n\nabla^2 G\big(\theta\big) \Big\| &\leqslant \sum_{k=1}^{n} \Big\| \nabla^2 G\big(\widehat{\theta}_{k-1}\big) - \nabla^2 G\big(\theta\big) \Big\| \\ &\leqslant \frac{1}{12\sqrt{3}} \sum_{k=1}^{n} \big\|\widehat{\theta}_{k-1} - \theta\big\| \mathbb{E}\big[\|\Phi_k\|^3 | \mathcal{F}_{k-1}\big] \end{aligned}$$

which implies that

$$\Big\| \sum_{k=1}^{n} \mathbb{E}\Big[\alpha\big(\Phi_k, \widehat{\theta}_{k-1}\big)\Phi_k\Phi_k^T | \mathcal{F}_{k-1}\Big] - n\nabla^2 G\big(\theta\big) \Big\| \leqslant \frac{\mathbb{E}\big[\|\Phi\|^3\big]}{12\sqrt{3}} \sum_{k=1}^{n} \big\|\widehat{\theta}_{k-1} - \theta\big\| \tag{6.40}$$

Finally, it follows from the conjunction of (6.37), (6.39) and (6.40) together with (6.35) that for all $0 < \beta < 1/2$,

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

which achieves the proof of (4.5). Moreover, we obtain (4.6) from (4.5) via the identity

$$\overline{S}_n^{-1} - \Big(\nabla^2 G\left(\theta\right)\Big)^{-1} = \overline{S}_n^{-1}\Big(\nabla^2 G\left(\theta\right) - \overline{S}_n\Big)\Big(\nabla^2 G\left(\theta\right)\Big)^{-1}.$$

7. Proofs of the asymptotic normality result

We are now in the position to proceed to the proof of the asymptotic normality (4.7). We clearly have from (6.24) that

$$\sqrt{n}\big(\widehat{\theta}_{n+1} - \theta\big) = -\frac{1}{\sqrt{n}} M_{n+1} - R_n \tag{7.1}$$

where the remainder $R_n = \sqrt{n}\Delta_n$. First of all, we claim that

$$\lim_{n\to\infty} R_n = 0 \qquad \text{a.s} \tag{7.2}$$

As a matter of fact, it follows from (6.25) that $R_n = P_n + Q_n$ where

$$\begin{aligned} P_n &= \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \left(\overline{S}_k^{-1} - S^{-1}\right) S\left(\widehat{\theta}_k - \theta\right), \\ Q_n &= \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \overline{S}_k^{-1} \delta_k. \end{aligned}$$

We have from (4.4) together with (4.6) that

$$\|P_n\| = O\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{1}{k^\beta} \frac{\sqrt{\log k}}{\sqrt{k}}\right) = O\left(\frac{\sqrt{\log n}}{n^\beta}\right) \qquad \text{a.s}$$

which implies that

$$\lim_{n\to\infty} P_n = 0 \qquad \text{a.s} \tag{7.3}$$

Moreover, we obtain from inequality (6.8) and (4.4) that

$$\|Q_n\| = O\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \left\|\widehat{\theta}_k - \theta\right\|^2\right) = O\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{\log k}{k}\right) = O\left(\frac{(\log n)^2}{\sqrt{n}}\right) \qquad \text{a.s}$$

which also implies that

$$\lim_{n\to\infty} Q_n = 0 \qquad \text{a.s} \tag{7.4}$$

Consequently, (7.3) and (7.4) clearly lead to convergence (7.2). Hereafter, it only remains to study the asymptotic behavior of the martingale term M_n. We already saw from (6.30) that its predictable quadratic variation $\langle M \rangle_n$ satisfies

$$\lim_{n\to\infty} \frac{1}{n} \langle M \rangle_n = \left(\nabla^2 G\left(\theta\right)\right)^{-1} \qquad \text{a.s}$$

In addtion, as $\varepsilon_{n+1} = \nabla_h g\left(\Phi_{n+1}, Y_{n+1}, \widehat{\theta}_n\right) - \nabla G\left(\widehat{\theta}_n\right)$, we clearly have the very simple upper-bound

$$\left\|\varepsilon_{n+1}\right\| \leqslant \left\|\Phi_{n+1}\right\| + \mathbb{E}\left[\left\|\Phi\right\|\right].$$

Hence, since Φ has a finite moment of order 4,

$$\sup_{n\geqslant 1} \mathbb{E}\left[\left\|\varepsilon_n\right\|^4\right] < \infty \tag{7.5}$$

Therefore, we immediately obtain from (7.5) that (M_n) satisfies Lindeberg's condition. Finally, we deduce from the central limit theorem for martingales given by Corollary

2.1.10 in [4] that

$$\frac{1}{\sqrt{n}} M_n \xrightarrow[\mathrm{n}\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \left(\nabla^2 G\left(\theta\right)\right)^{-1}\right)$$

which, via (7.1) and (7.2), completes the proof of Theorem 4.4. □

REFERENCES

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

[2] R. H. Byrd, S. L. Hansen, J. Nocedal, and Y. Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

[3] S. Clémençon, P. Bertail, E. Chautru, and G. Papa. Survey schemes for stochastic gradient descent with applications to m-estimation. *arXiv preprint arXiv:1501.02218*, 2015.

[4] M. 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.

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

[6] A. Godichon-Baggioni. Lp and almost sure rates of convergence of averaged stochastic gradient algorithms with applications to online robust estimation. *arXiv preprint arXiv:1609.05479*, 2016.

[7] A. Godichon-Baggioni. Online estimation of the asymptotic variance for averaged stochastic gradient algorithms. *arXiv preprint arXiv:1702.00931*, 2017.

[8] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[9] O. Komori, S. Eguchi, S. Ikeda, H. Okamura, M. Ichinokawa, and S. Nakayama. An asymmetric logistic regression model for ecological data. *Methods in Ecology and Evolution*, 7(2):249–260, 2016.

[10] A. Lucchi, B. McWilliams, and T. Hofmann. A variance reduced stochastic newton method. *arXiv preprint arXiv:1503.08316*, 2015.

[11] J. Merlo, P. Wagner, N. Ghith, and G. Leckie. An original stepwise multilevel logistic regression analysis of discriminatory accuracy: the case of neighbourhoods and health. *PLoS One*, 11(4):e0153778, 2016.

[12] A. Mokhtari and A. Ribeiro. Res: Regularized stochastic bfgs algorithm. *IEEE Transactions on Signal Processing*, 62(23):6089–6104, 2014.

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

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

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

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

[17] H. R. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28, 2014.