Online estimation of the asymptotic variance for averaged stochastic gradient algorithms

Antoine Godichon

Version of Record: https://www.sciencedirect.com/science/article/pii/S0378375819300011
Manuscript_d0425abd491bd75e2d7de9c93bad2b9f

Online estimation of the asymptotic variance for averaged stochastic gradient algorithms

Antoine Godichon-Baggioni
Institut de Mathématiques de Toulouse,
Université Paul Sabatier, 31000 Toulouse, France
email: godichon@insa-toulouse.fr

October 16, 2017

Abstract

Stochastic gradient algorithms are more and more studied since they can deal efficiently and online with large samples in high dimensional spaces. In this paper, we first establish a Central Limit Theorem for these estimates as well as for their averaged version in general Hilbert spaces. Moreover, since having the asymptotic normality of estimates is often unusable without an estimation of the asymptotic variance, we introduce a new recursive algorithm for estimating this last one, and we establish its almost sure rate of convergence as well as its rate of convergence in quadratic mean. Finally, two examples consisting in estimating the parameters of the logistic regression and estimating geometric quantiles are given.

Keywords: Stochastic Gradient Algorithm, Averaging, Central Limit Theorem, Asymptotic Variance.

1 Introduction

High Dimensional and Functional Data Analysis are interesting domains which do not have stopped growing for many years. To consider these kinds of data, it is more and more important to think about methods which take into account the high dimension as well as the possibility of having large samples. In this paper, we focus on an usual stochastic optimization problem which consists in estimating

$$m := \arg\min_{h \in H} \mathbb{E}\left[g\left(X, h\right)\right],$$

where X is a random variable taking values in a space $\mathcal{X}$ and $g : \mathcal{X} \times H \longrightarrow \mathbb{R}$, where H is a separable Hilbert space. In order to build an estimator of m, an usual method was to consider the solver of the problem generated by the sample, i.e to consider M-estimates (see Huber and Ronchetti (2009) and Maronna et al. (2006) among others). In order to build these estimates, deterministic convex optimization algorithms (see Boyd and Vandenberghe

(2004)) are often used (see Vardi and Zhang (2000), Oja and Niinimaa (1985) in the case of the median), and these methods are really efficient in small dimensional spaces.

Nevertheless, in a context of high dimensional spaces, this kind of method can encounter many computational problems. The main ones are that it needs to store all the data, which can be expensive in term of memory and that they cannot deal online with the data. In order to overcome this, stochastic gradient algorithms (Robbins and Monro (1951)) are efficient candidates since they do not need to store the data into memory, and they can be easily updated, which is crucial if the data arrive sequentially (see Duflo (1996), Duflo (1997), Kushner and Yin (2003a) or Nemirovski et al. (2009) among others). In order to improve the convergence, Ruppert (1988) and Polyak and Juditsky (1992) introduced its averaged version (see also Dippon and Renz (1997) for a weighted version). These algorithms have become crucial to statistics and modern machine learning (Bach and Moulines (2013), Bach (2014), Juditsky et al. (2014)). There are already many results on these algorithms in the literature, that we can split into two parts: asymptotic results, such as almost sure rates of convergence (Schwabe and Walk, 1996; Duflo, 1997; Walk, 1992; Pelletier, 1998, 2000), and non asymptotic ones, such as rates of convergence in quadratic mean (Cardot et al., 2017; Godichon-Baggioni, 2016a; Bach and Moulines, 2013; Bach, 2014; Nemirovski et al., 2009).

In a recent work, Godichon-Baggioni (2016b) introduces a new framework, with only locally strongly convexity assumptions, in general Hilbert spaces, which allows to obtain almost sure and L^p rates of convergence. In keeping with it, and in order to have a deeper study of the stochastic gradient algorithm as well as of its averaged version (up to a new assumption), we first give the asymptotic normality of the estimates. In a second time, since a Central Limit Theorem is often unusable without an estimation of the variance, we introduce a recursive algorithm, inspired by Gahbiche and Pelletier (2000), to estimate the asymptotic variance of the averaged estimator and we establish its rates of convergence. As far as we know, there was not yet an efficient and recursive estimate of the asymptotic variance in the literature. Finally, two examples of application are given. The first usual one consists in estimating the parameters of the logistic regression (Bach, 2014) while the second one consists in estimating geometric quantiles (see Chaudhuri (1996) and Chakraborty and Chaudhuri (2014)), which are useful robust indicators in statistics. Indeed, they are often used in data depth and outliers detection (Serfling (2006), Hallin and Paindaveine (2006)), as well as for robust estimation of the mean and variance (see Minsker et al. (2014)), or for Robust Principal Component Analysis (Gervini (2008), Kraus and Panaretos (2012), Cardot and Godichon-Baggioni (2017)).

The paper is organized as follows: Section 2 recalls the framework introduced by Godichon-Baggioni (2016b) before giving two new assumptions which allow to get the rate of convergence of the estimators of the asymptotic variance. In section 3, the stochastic gradient algorithm as well as its averaged version are introduced and their asymptotic normality are given. The recursive estimator of the asymptotic variance is given in Section 4 and its almost sure as well as its quadratic mean rates of convergence are established. Applications, consisting in estimating the logistic regression parameters and in the recursive estimation

of geometric quantiles, are given in Section 5 as well as a short simulation study. Finally, the proofs are postponed in Section 6 and in a Supplementary file.

2 Assumptions

Let H be a separable Hilbert space such as $\mathbb{R}^d$ or $L^2(I)$ (for some closed interval $I \subset \mathbb{R}$), we denote by $\langle .,. \rangle$ its inner product and by $\|.\|$ the associated norm. Let X be a random variable taking values in a space $\mathcal{X}$, and let $G : H \longrightarrow \mathbb{R}$ be the function we would like to minimize, defined for all $h \in H$ by

$$G(h) := \mathbb{E}\left[g(X,h)\right], \tag{1}$$

where $g : \mathcal{X} \times H \longrightarrow \mathbb{R}$. Moreover, let us suppose that the functional G is convex. Finally, let us introduce the space of linear operators on H, denoted by $\mathcal{S}(H)$, equipped with the Frobenius (or Hilbert-Schmidt) inner product, which is defined by

$$\langle A, B \rangle_F := \sum_{j \in J} \left\langle A(e_j), B(e_j) \right\rangle, \quad \forall A, B \in \mathcal{S}(H),$$

where $\left(e_j\right)_{j \in J}$ is an orthonormal basis of H. We denote by $\|.\|_F$ the associated norm, and $\mathcal{S}(H)$ is then a separable Hilbert space. Let us recall the framework introduced by Godichon-Baggioni (2016b):

(A1) The functional g is Frechet-differentiable for the second variable almost everywhere. Moreover, G is differentiable and there exists $m \in H$ such that

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

(A2) The functional G is twice continuously differentiable almost everywhere and for all positive constant A, there is a positive constant C_A such that for all $h \in \mathcal{B}(m, A)$,

$$\|\Gamma_h\|_{op} \leq C_A,$$

where Γ_h is the Hessian of the functional G at h and $\|.\|_{op}$ is the usual spectral norm for linear operators.

(A3) There exists a positive constant ϵ such that for all $h \in \mathcal{B}(m, \epsilon)$, there is an orthonormal basis of H composed of eigenvectors of Γ_h. Moreover, let us denote by $\lambda_{\min}$ the limit inf of the eigenvalues of Γ_m, then $\lambda_{\min}$ is positive. Finally, for all $h \in \mathcal{B}(m, \epsilon)$, and for all eigenvalue λ_h of Γ_h, we have $\lambda_h \geq \frac{\lambda_{\min}}{2} > 0$.

(A4) There are positive constants ϵ, C_ϵ such that for all $h \in \mathcal{B}(m, \epsilon)$,

$$\|\nabla G(h) - \Gamma_m(h - m)\| \leq C_\epsilon \|h - m\|^2.$$

(A5) **(a)** There is a positive constant L_1 such that for all $h \in H$,

$$\mathbb{E}\left[\left\|\nabla_h g\left(X,h\right)\right\|^2\right] \leq L_1\left(1+\left\|h-m\right\|^2\right).$$

(a') There is a positive constant L_2 such that for all $h \in H$,

$$\mathbb{E}\left[\left\|\nabla_h g\left(X,h\right)\right\|^4\right] \leq L_2\left(1+\left\|h-m\right\|^4\right).$$

(b) For all integer q, there is a positive constant L_q such that for all $h \in H$,

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

Let us now make some comments on assumptions. First, Assumption **(A1)** ensures the existence of a solution and enables to use a stochastic gradient descent, while **(A2)** gives some smoothness properties on the objective function. Assumption **(A3)** ensures the uniqueness of the minimizer of G, and **(A4)**,**(A5)** give bounds of the gradient and of the remainder term of its Taylor's expansion. The main difference between this framework and the usual one for strongly convex objective is that we just assume the local strong convexity of the objective function, and in return, p-th moments of the gradient of the functional g have to be bounded. Note also that the Hessian of the functional G is not supposed to be compact, so that its smallest eigenvalue does not necessarily converge to 0 when the dimension tends to infinity (a counter example is given in Section 5). Remark that assumptions **(A1)** to **(A5b)** are deeply discussed in Godichon-Baggioni (2016b). Let us now introduce two new assumptions.

(A6) Let $\varphi : H \longrightarrow \mathcal{S}(H)$ be the functional defined for all $h \in H$ by

$$\varphi\left(h\right) := \mathbb{E}\left[\nabla_h g\left(X,h\right) \otimes \nabla_h g\left(X,h\right)\right].$$

(a) The functional φ is continuous at m with respect to the Frobenius norm:

$$\lim_{h \to m} \left\|\mathbb{E}\left[\nabla_h g\left(X,m\right) \otimes \nabla_h g\left(X,m\right)\right] - \mathbb{E}\left[\nabla_h g\left(X,h\right) \otimes \nabla_h g\left(X,h\right)\right]\right\|_F = 0.$$

(b) The functional φ is locally lipschitz on a neighborhood of m: there are positive constants ϵ, C'_ϵ, such that for all $h \in \mathcal{B}\left(m,\epsilon\right)$,

$$\left\|\mathbb{E}\left[\nabla_h g\left(X,m\right) \otimes \nabla_h g\left(X,m\right) - \nabla_h g\left(X,h\right) \otimes \nabla_h g\left(X,h\right)\right]\right\|_F \leq C'_\epsilon \left\|h-m\right\|.$$

Assumption **(A6a)** enables to establish the asymptotic normality of the stochastic gradient descent as well as of its averaged version. Note that under **(A5a)**, the functional φ is bounded, and more precisely

$$\left\|\mathbb{E}\left[\nabla_h g\left(X,h\right) \otimes \nabla_h g\left(X,h\right)\right]\right\|_F \leq \mathbb{E}\left[\left\|\nabla_h g\left(X,h\right)\right\|^2\right] \leq L_1\left(1+\left\|h-m\right\|^2\right).$$

Assumption **(A6b)** can be verified by giving a bound, on a neighborhood of m, of the derivative of the functional φ. This last assumption allows to give the rate of convergence of the estimators of the asymptotic variance. An example is given for the special case of the geometric median in a supplementary file.

Remark 2.1. *For all $h \in H$ and $A > 0$,*

$$\mathcal{B}(h, A) = \left\{ h' \in H, \quad \left\| h - h' \right\| < A \right\}.$$

Remark 2.2. *Let $h, h' \in H$, the linear operator $h \otimes h' : H \longrightarrow H$ is defined for all $h'' \in H$ by $h \otimes h'(h'') := \langle h, h'' \rangle h'$. Moreover,*

$$\left\| h \otimes h' \right\|_F = \|h\| \left\| h' \right\|. \tag{2}$$

3 The stochastic gradient algorithm and its averaged version

3.1 The Robbins-Monro algorithm

In what follows, let $X_1, ..., X_n$ be independent random variables with the same law as X. The stochastic gradient algorithm is defined recursively for all $n \geq 1$ by

$$m_{n+1} = m_n - \gamma_n \nabla_h g\left(X_{n+1}, m_n\right), \tag{3}$$

with m_1 bounded and (γ_n) is a step sequence of the form $\gamma_n := c_\gamma n^{-\alpha}$, with $c_\gamma > 0$ and $\alpha \in \left(\frac{1}{2}, 1\right)$. Moreover, let $(\mathcal{F}_n)_{n \geq 1}$ be the sequence of σ-algebras defined for all $n \geq 1$ by $\mathcal{F}_n := \sigma(X_1, ..., X_n)$. Then, the algorithm can be considered as a noisy (or stochastic) gradient algorithm since it can be written as

$$m_{n+1} = m_n - \gamma_n \Phi(m_n) + \gamma_n \xi_{n+1}, \tag{4}$$

where $\Phi(m_n) := \nabla G(m_n)$, and (ξ_n), defined for all $n \geq 1$ by $\xi_{n+1} := \Phi(m_n) - \nabla_h g\left(X_{n+1}, m_n\right)$, is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$. Finally, note that under assumptions **(A1)** to **(A5a)**, it was proven in Godichon-Baggioni (2016b) that for all positive constant δ,

$$\left\| m_n - m \right\|^2 = o\left(\frac{(\ln n)^\delta}{n^\alpha} \right) \quad a.s. \tag{5}$$

Moreover, assuming that **(A5b)** is also fulfilled, for all positive integer p, there is a constant C_p such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\| m_n - m \right\|^{2p} \right] \leq \frac{C_p}{n^{p\alpha}}. \tag{6}$$

In order to get a deeper study of this estimate, we now give its asymptotic normality.

Theorem 3.1. *Suppose assumptions **(A1)** to **(A5a')** and **(A6a)** hold. Then, we have the convergence*

in law

$$\lim_{n\to\infty} \frac{1}{\sqrt{\gamma_n}} (m_n - m) \sim \mathcal{N}(0, \Sigma_{RM}),$$

with

$$\Sigma_{RM} := \int_0^{+\infty} e^{-s\Gamma_m} \Sigma' e^{-s\Gamma_m} ds, \qquad \text{and} \qquad \Sigma' := \mathbb{E}\left[\nabla_h g(X, m) \otimes \nabla_h g(X, m)\right].$$

The proof is given in a Supplementary file. Note that the variance Σ_{RM} does not depend on the step sequence (γ_n), but Theorem 3.1 could be written as

$$\lim_{n\to\infty} n^{\alpha/2} (m_n - m) \sim \mathcal{N}(0, c_\gamma \Sigma_{RM}),$$

Remark 3.1. *Let M be a squared matrix, e^M is defined by (see Horn and Johnson (2012) among others)*

$$e^M = \sum_{k=0}^{\infty} \frac{1}{k!} M^k.$$

*Thanks to assumptions **(A2)**,**(A3)**, $0 < \lambda_{\min}(\Gamma_m) \leq \lambda_{\max}(\Gamma_m) < \infty$, while under **(A5a)** and by dominated convergence,*

$$\|\Sigma_{RM}\|_F \leq \int_0^{+\infty} \left\|e^{-s\Gamma_m}\right\|_{op}^2 \|\Sigma'\|_F ds \leq \int_0^{+\infty} e^{-2s\lambda_{\min}} \|\Sigma'\|_F ds \leq \frac{L_1}{2\lambda_{\min}},$$

and Σ_{RM} is so well defined.

Remark 3.2. *Note that analogous results are given by (Fabian, 1968; Pelletier, 1998) in the particular case of finite dimensional spaces while, for analogous results in Banach and Hilbert spaces, one can also see Walk (1992), Ljung et al. (2012), Kushner and Yin (2003b).*

Remark 3.3. *Note that taking a step sequence of the form $\gamma_n = \frac{c}{n}$ with $c > \frac{2}{\lambda_{\min}}$ is possible, and one can obtain the following asymptotic normality (see Pelletier (2000) among others for the case of finite dimensional spaces)*

$$\lim_{n\to\infty} \sqrt{n} (m_n - m) \sim \mathcal{N}(0, c\Sigma').$$

Nevertheless, it does not only necessitate to have some information on the Hessian Γ_m, but $c\Sigma'$ is also not the optimal variance (see Duflo (1997) and Pelletier (2000) for instance).

3.2 The averaged algorithm

As mentioned in Remark 3.3, having the parametric rate of convergence ($O\left(\frac{1}{n}\right)$) with the Robbins-Monro algorithm is possible taking a good choice of step sequence (γ_n). Nevertheless, this choice is often complicated and the asymptotic variance which is obtained is not optimal. Then, in order to improve the convergence, let us now introduce the averaged algorithm (see Ruppert (1988) and Polyak and Juditsky (1992)) defined for all $n \geq 1$ by

$$\overline{m}_n = \frac{1}{n} \sum_{k=1}^{n} m_k.$$

This can be written recursively for all $n \geq 1$ as

$$\overline{m}_{n+1} = \overline{m}_n + \frac{1}{n+1}\left(m_{n+1} - \overline{m}_n\right). \tag{7}$$

It was proven in Godichon-Baggioni (2016b) that under assumptions **(A1)** to **(A5a)**, for all $\delta > 0$,

$$\left\|\overline{m}_n - m\right\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) \quad a.s. \tag{8}$$

Suppose assumption **(A5b)** is also fulfilled, for all positive integer p, there is a positive constant C'_p such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\|\overline{m}_n - m\right\|^{2p}\right] \leq \frac{C'_p}{n^p}. \tag{9}$$

Finally, in order to have a deeper study of this estimate, we now give its asymptotic normality.

Theorem 3.2. *Suppose assumptions **(A1)** to **(A5a')** and **(A6a)** are verified. Then, we have the convergence in law*

$$\lim_{n\to\infty} \sqrt{n}\left(\overline{m}_n - m\right) \sim \mathcal{N}\left(0, \Sigma\right),$$

with $\Sigma := \Gamma_m^{-1}\Sigma'\Gamma_m^{-1}$, *and* $\Sigma' := \mathbb{E}\left[\nabla_h g\left(X, m\right) \otimes \nabla_h g\left(X, m\right)\right]$.

The proof is given in Section 6. For analogous results, one can also see Schwabe and Walk (1996), Pelletier (2000), Dippon and Walk (2006).

4 Recursive estimation of the asymptotic variance

4.1 Some existing estimators

A first naive method to estimate the asymptotic variance could be to estimate the Hessian Γ_m and the variance Σ' as follows

$$\Gamma_m^{(n+1)} = \Gamma_m^{(n)} + \frac{1}{n+1}\left(\nabla_h^2 g\left(X_{n+1}, \overline{m}_n\right) - \Gamma_m^{(n)}\right),$$
$$\Sigma'_{n+1} = \Sigma'_n + \frac{1}{n+1}\left(\nabla_h g\left(X_{n+1}, \overline{m}_n\right) \otimes \nabla_h g\left(X_{n+1}, \overline{m}_n\right) - \Sigma'_n\right),$$

but the main problem is that under assumptions **(A2)**, **(A3)** and **(A5a)**, if H is an infinite dimensional space, then

$$\left\|\Gamma_m\right\|_F = \infty, \qquad \text{while} \qquad \left\|\Gamma_m^{-1}\Sigma'\Gamma_m^{-1}\right\|_F \leq \frac{L_1}{\lambda_{\min}^2}.$$

Another problem is that, in order to get a recursive estimator of the asymptotic variance, it needs to invert a matrix at each iteration, which costs much calculus time in high dimensional spaces. A second estimator of the asymptotic variance was introduced in Pelletier

(2000), defined for all $n \geq 1$ by

$$\widehat{\Sigma}_n = \frac{1}{\ln n} \sum_{k=1}^{n} (m_k - \overline{m}_n) \otimes (m_k - \overline{m}_n), \tag{10}$$

and under **(A1)** to **(A6b)**,

$$\mathbb{E}\left[\left\| \widehat{\Sigma}_n - \Sigma \right\|_F^2 \right] = O\left(\frac{1}{\ln n} \right).$$

Thus, this estimator faces two main problems: it is not recursive and it converges very slowly. Finally, in order to solve the second problem, a faster algorithm was introduced by Gahbiche and Pelletier (2000), defined for all $n \geq 1$ by

$$\tilde{\Sigma}_n := \frac{1-\delta}{n^{1-\delta}} \sum_{k=1}^{n} \frac{1}{k^{\delta+s+\mu}} \exp\left(-\frac{k^{1-s}}{1-s} \right) \left(\sum_{j=1}^{k} j^{\mu/2} e^{\frac{j^{1-s}}{2(1-s)}} (m_j - \overline{m}_n) \right) \otimes \left(\sum_{j=1}^{k} j^{\mu/2} e^{\frac{j^{1-s}}{2(1-s)}} (m_j - \overline{m}_n) \right), \tag{11}$$

with $(1+\alpha)/2 < s < 1$, $\mu \geq 0$ and $s/2 < \delta < (1+s)/2$. This algorithm is first based on an usual decomposition of the stochastic gradient algorithm (see equation (18)) which enables to make appear a martingale term which carries the convergence rate (see equation (27)). In a second time, the objective is to find step sequences which enable to improve the rate of convergence of the variance estimate (see Gahbiche and Pelletier (2000) for technical details on assumptions on the step sequences). In the case of finite dimensional spaces, the following convergence in probability is given (under some assumptions)

$$\frac{n^{1/2-s/2}}{(\ln \ln n)^c} \left\| \tilde{\Sigma}_n - \Sigma \right\|_{op} \xrightarrow[n\to\infty]{\mathbb{P}} 0,$$

with $c > 0$. A first technical problem is that only the convergence in probability is given, in the case of finite dimensional spaces, and for the usual spectral norm. A second one is that it is not recursive and it cannot be easily updated.

4.2 A recursive and fast estimate

We now give a recursive version of the algorithm defined by (11) to estimate the asymptotic variance in separable Hilbert spaces, before establishing its rates of convergence (almost sure and in quadratic mean). This algorithm is defined by

$$\Sigma_n := \frac{1-\delta}{n^{1-\delta}} \sum_{k=1}^{n} \frac{1}{k^{\delta+s+\mu}} \exp\left(-\frac{k^{1-s}}{1-s} \right) \left(\sum_{j=1}^{k} j^{\mu/2} e^{\frac{j^{1-s}}{2(1-s)}} (m_j - \overline{m}_j) \right) \otimes \left(\sum_{j=1}^{k} j^{\mu/2} e^{\frac{j^{1-s}}{2(1-s)}} (m_j - \overline{m}_j) \right), \tag{12}$$

with

$$(1+\alpha)/2 < s < 1, \qquad \mu \geq 0, \qquad \text{and} \qquad s/2 < \delta < (1+s)/2. \tag{13}$$

The difference with previous algorithm is the replacement of $\overline{m}_n$ by $\overline{m}_j$, which enables the estimates to be written recursively for all $n \geq 1$ as

$$V_{n+1} = V_n + (n+1)^{\mu/2} \exp\left(\frac{(n+1)^{1-s}}{2(1-s)}\right)(m_{n+1} - \overline{m}_{n+1}),$$
$$\Sigma_{n+1} = \left(\frac{n}{n+1}\right)^{1-\delta} \Sigma_n + \frac{1-\delta}{(n+1)^{\delta+s+\mu}} \exp\left(-\frac{(n+1)^{1-s}}{1-s}\right) V_{n+1} \otimes V_{n+1},$$

with $V_1 = \Sigma_1 = 0$. Then, contrary to previous algorithms, this one does not need to store all the estimations into memory and can be easily updated. Finally, the following theorem ensures that it is quite fast.

Theorem 4.1. *Suppose assumptions **(A1)** to **(A5a')** and **(A6b)** hold. Then, the sequence (Σ_n) defined by (12) verifies for all positive constant γ,*

$$\|\Sigma_n - \Sigma\|_F^2 = o\left(\frac{(\ln n)^\gamma}{n^{1-s}}\right) \quad a.s.$$

*Moreover, suppose **(A5b)** holds too, there is a positive constant C such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|\Sigma_n - \Sigma\|_F^2\right] \leq \frac{C}{n^{1-s}}$$

The proof is given in Section 6.

Corollary 4.1. *Suppose assumptions **(A1)** to **(A5a')** and **(A6b)** hold. Then, for all positive constant γ,*

$$\|\tilde{\Sigma}_n - \Sigma\|_F^2 = o\left(\frac{(\ln n)^\gamma}{n^{1-s}}\right) \quad a.s.$$

*Moreover, suppose **(A5b)** holds too, there is a positive constant C such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|\tilde{\Sigma}_n - \Sigma\|_F^2\right] \leq \frac{C}{n^{1-s}}$$

Remark 4.1. *The constant C in Theorem 4.1 depends on the constants introduced in assumptions, on the initialization of the stochastic gradient descent, and on $\alpha, \delta, \mu, s, c_\gamma$.*

Remark 4.2. *Estimating recursively the asymptotic variance coupled with Theorem 3.2 can be useful to build online asymptotic confidence balls. Moreover, in the recent literature, non asymptotic convergence rates are often given under the form*

$$\mathbb{E}\left[\|\overline{m}_n - m\|^2\right] \leq \frac{\|\Sigma\|_F}{n} + R_n,$$

where R_n is a rest term. Then, using the recursive variance estimates could enable to have, in practice, a precise bound of the quadratic mean error, and in the short term, it could allow to get precise non asymptotic confidence balls.

Remark 4.3. *In order to get a faster algorithm (in term of computational time), one can consider a parallelized version of previous estimates. This consists in splitting the sample into p parts, and to*

run the algorithm on each subsample to get p estimates $\Sigma_{n/p,i}$, before taking the mean of these p last ones.

5 Applications

5.1 Application to the logistic regression

Let d be a positive integer, and let $Y \in \{-1, 1\}$ and $X \in \mathbb{R}^d$ be random variables. In order to get the parameter $m^l \in \mathbb{R}^d$ of the logistic regression, the aim is to minimize the functional G_l defined for all $h \in \mathbb{R}^d$ by

$$G_l(h) := \mathbb{E}\left[\log\left(1 + \exp\left(-Y\left\langle X, h\right\rangle\right)\right)\right]. \tag{14}$$

Under usual assumptions (see Bach (2014) among others), the functional G_l is locally strongly convex and twice Fréchet differentiable with for all $h \in \mathbb{R}^d$,

$$\nabla G_l(h) = -\mathbb{E}\left[\frac{\exp\left(-Y\left\langle X, h\right\rangle\right)}{1 + \exp\left(-Y\left\langle X, h\right\rangle\right)} YX\right], \quad \nabla^2 G_l(h) = \mathbb{E}\left[\frac{\exp\left(-Y\left\langle X, h\right\rangle\right)}{\left(1 + \exp\left(-Y\left\langle X, h\right\rangle\right)\right)^2} X \otimes X\right].$$

Then, the parameters of the logistic regression and the asymptotic variance can be estimated simultaneously as:

$$\begin{aligned}
m^l_{n+1} &= m^l_n + \gamma_n \frac{\exp\left(-Y_{n+1}\left\langle X_{n+1}, m^l_n\right\rangle\right)}{1 + \exp\left(-Y_{n+1}\left\langle X_{n+1}, m^l_n\right\rangle\right)} Y_{n+1} X_{n+1}, \\
\overline{m}^l_{n+1} &= \overline{m}^l_n + \frac{1}{n+1}\left(m^l_{n+1} - \overline{m}^l_n\right), \\
V^l_{n+1} &= V^l_n + (n+1)^{\mu/2} \exp\left(\frac{(n+1)^{1-s}}{2(1-s)}\right)\left(m^l_{n+1} - \overline{m}^l_{n+1}\right), \\
\Sigma^l_{n+1} &= \left(\frac{n}{n+1}\right)^{1-\delta} \Sigma^l_n + \frac{1-\delta}{(n+1)^{\delta+s+\mu}} \exp\left(-\frac{(n+1)^{1-s}}{1-s}\right) V^l_{n+1} \otimes V^l_{n+1}.
\end{aligned}$$

5.2 Application to the geometric median and geometric quantiles

Let H be a separable Hilbert space and let X be a random variable taking values in H. Let $v \in H$ such that $\|v\| < 1$, the geometric quantile m^v corresponding to the direction v (see Chaudhuri (1996)) is defined by

$$m^v := \arg\min_{h \in H} \mathbb{E}\left[\|X - h\| - \|X\|\right] - \langle h, v\rangle, \tag{15}$$

and in a particular case, the geometric median m (see Haldane (1948)) corresponds to the case where $v = 0$. Under usual assumptions (see Kemperman (1987) and Cardot et al. (2013) among others), the functional G_v is locally strongly convex and twice Fréchet-differentiable

with for all $h \in H$,

$$\nabla G^v(h) = -\mathbb{E}\left[\frac{X-h}{\|X-h\|} + v\right], \quad \nabla^2 G^v(h) = \mathbb{E}\left[\frac{1}{\|X-h\|}\left(I_H - \frac{(X-h)\otimes(X-h)}{\|X-h\|^2}\right)\right].$$

Then, it is possible to estimate simultaneously and recursively the geometric quantile m^v as well as the asymptotic variance of the averaged estimator as follows:

$$\begin{aligned}
m^v_{n+1} &= m^v_n + \gamma_n\left(\frac{X_{n+1}-m^v_n}{\|X_{n+1}-m^v_n\|} + v\right),\\
\overline{m}^v_{n+1} &= \overline{m}^v_n + \frac{1}{n+1}\left(m^v_{n+1} - \overline{m}^v_n\right),\\
V^v_{n+1} &= V^v_n + (n+1)^{\mu/2}\exp\left(\frac{(n+1)^{1-s}}{2(1-s)}\right)\left(m^v_{n+1} - \overline{m}^v_{n+1}\right),\\
\Sigma^v_{n+1} &= \left(\frac{n}{n+1}\right)^{1-\delta}\Sigma_n + \frac{1-\delta}{(n+1)^{\delta+s+\mu}}\exp\left(-\frac{(n+1)^{1-s}}{1-s}\right)V^v_{n+1}\otimes V^v_{n+1}.
\end{aligned}$$

Note that under usual assumptions, the asymptotic variance obtained is the same as the one obtained with non-recursive estimates (Maronna et al., 2006; Gervini, 2008) in the special case of the geometric median.

5.3 A short simulation study

We focus here on the estimation of the geometric median. We consider from now that X is a random variable taking values in $\mathbb{R}^d$, with $d \geq 3$, and following a uniform law on the unit sphere $\mathcal{S}^d$. Then, the geometric median m is equal to 0 and the Hessian of the functional G_0 at m verifies

$$\Gamma_m = \mathbb{E}\left[\frac{1}{\|X\|}\left(I_d - \frac{X}{\|X\|}\otimes\frac{X}{\|X\|}\right)\right] = I_d - \mathbb{E}\left[X\otimes X\right] = \frac{d-1}{d}I_d.$$

Note that assumptions **(A1)** and **(A6b)** are then verified (see Section 3 in Godichon-Baggioni (2016b), Lemma A.1 in Godichon-Baggioni et al. (2017) and the supplementary file to be convinced). Finally, the asymptotic variance of the stochastic gradient estimate and of its averaged version verify

$$\begin{aligned}
\Sigma_{RM} &= \int_0^\infty e^{-s\Gamma_m}\mathbb{E}\left[\frac{X}{\|X\|}\otimes\frac{X}{\|X\|}\right]e^{-s\Gamma_m}ds = \frac{1}{2(d-1)}I_d,\\
\Sigma &= \Gamma_m^{-1}\mathbb{E}\left[\frac{X}{\|X\|}\otimes\frac{X}{\|X\|}\right]\Gamma_m^{-1} = \frac{d}{(d-1)^2}I_d.
\end{aligned}$$

First, let us consider a stepsequence $\gamma_n = n^{-2/3}$ and let us study the quality of the Gaussian approximation of Q_n, Q'_n, where

$$Q_n := \sqrt{2(d-1)}n^{1/3}\left(m_n - m\right), \qquad \text{and} \qquad Q'_n := \sqrt{n}\frac{d-1}{\sqrt{d}}\left(\overline{m}_n - m\right).$$

Figure 1 (respectively Figure 2) seems to confirm Theorem 3.1 (respectively Theorem 3.2) since we can see that the estimated density of a component of Q_n (respectively Q'_n) is close to the density of $\mathcal{N}(0,1)$, and so, even for small sample sizes ($n = 200$), which is also confirmed by a Kolmogorov-Smirnov test.

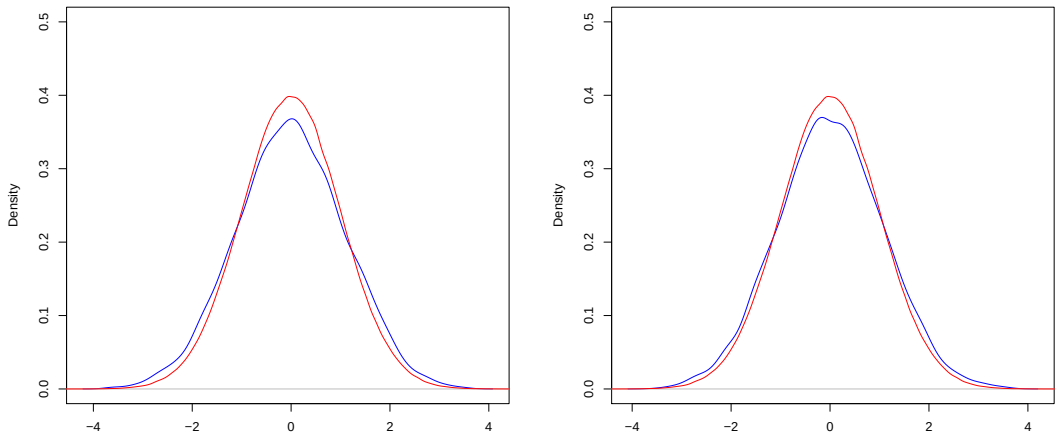

Figure 1: Estimated density of a component of Q_n (in blue) compared to the standard gaussian density (in red), with $n = 200$ (on the left) and $n = 5000$ (on the right).

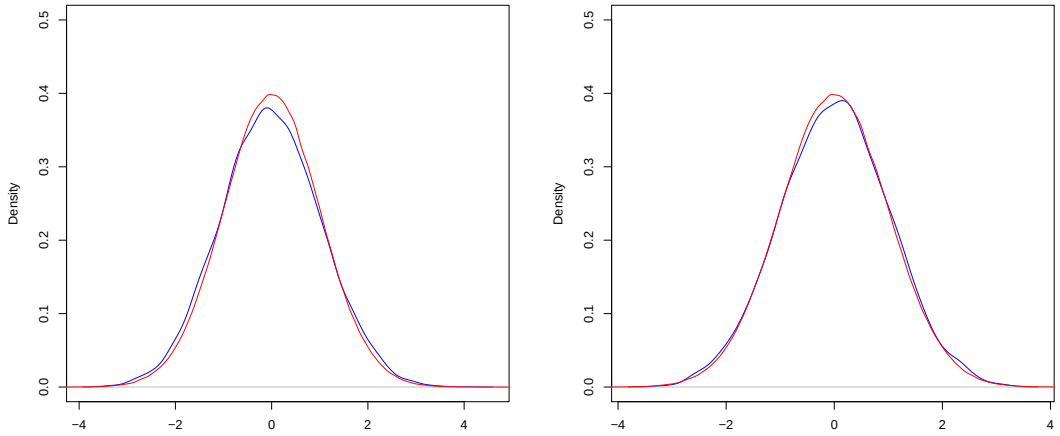

Figure 2: Estimated density of a component of Q'_n (in blue) compared to the standard gaussian density (in red), with $n = 200$ (on the left) and $n = 5000$ (on the right).

In Figure 3, we consider the evolution of the quadratic mean error, with respect to the Frobenius norm, of the estimates (Σ_n) of Σ defined by (12), with regard to the sample size. For this, we generate 100 samples, and use the parallelized version of the algorithms. Figure 3 tends to confirm that for small dimensional spaces ($d = 10$), the estimates of the

asymptotic variance converge quite quickly and that it is still the case for moderate dimensional spaces ($d = 5000$).

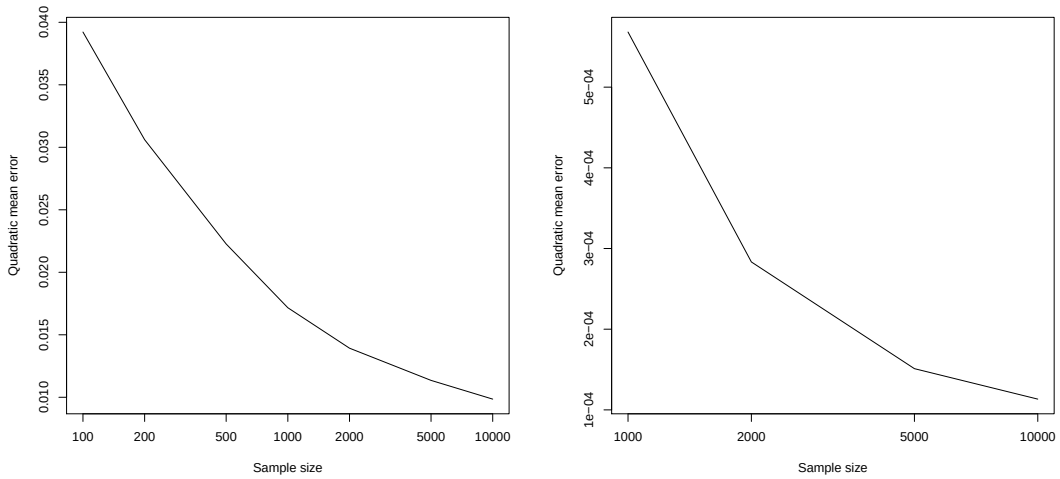

Figure 3: Evolution of the quadratic mean error of the estimation of the asymptotic variance Σ with respect to the Frobenius norm for $d = 10$ (on the left) and $d = 5000$ (on the right).

6 Proofs

6.1 Some decompositions of the algorithms

In order to simplify the proofs, let us now give some decompositions of the algorithms.

6.1.1 The Robbins-Monro algorithm

Let us recall that the stochastic gradient algorithm can be written as

$$m_{n+1} - m = m_n - m - \gamma_n \Phi(m_n) + \gamma_n \xi_{n+1}.$$

Linearizing the gradient, it comes

$$m_{n+1} - m = (I_H - \gamma_n \Gamma_m)(m_n - m) + \gamma_n \xi_{n+1} - \gamma_n \delta_n, \tag{16}$$

where $\delta_n := \Gamma_m (m_n - m) - \Phi(m_n)$ is the remainder term in the Taylor's expansion of the gradient. Thanks to previous decomposition and with the help of an induction (see Duflo (1996) or Duflo (1997) for instance), one can check that for all $n \geq 1$,

$$m_n - m = \beta_{n-1}(m_1 - m) - \beta_{n-1} \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \delta_k + \beta_{n-1} \sum_{k=1}^{n-1} \gamma_k \beta_k^{-1} \xi_{k+1}, \tag{17}$$

with $\beta_n := \prod_{k=1}^{n} (I_H - \gamma_k \Gamma_m)$ for all $n \geq 1$ and $\beta_0 := I_H$. Finally, the asymptotic variance can be seen as the almost sure limit of the sequence of random variables $\left(\Gamma_m^{-1} \xi_n \otimes \Gamma_m^{-1} \xi_n\right)_n$

(see the proof of Theorem 3.2). Then, in order to prove the convergence of the estimates, we need to exhibit this sequence. In this aim, one can rewrite equation (16) as

$$m_n - m = \frac{T_n}{\gamma_n} - \frac{T_{n+1}}{\gamma_n} + \Xi_{n+1} - \Delta_n, \tag{18}$$

with

$$T_n := \Gamma_m^{-1}(m_n - m), \qquad \Xi_{n+1} := \Gamma_m^{-1}(\xi_{n+1}), \qquad \Delta_n := \Gamma_m^{-1}(\delta_n).$$

6.1.2 The averaged algorithm

Summing equalities (18) and dividing by n, we obtain the following decomposition of the averaged estimator

$$\overline{m}_n - m = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{T_k}{\gamma_k} - \frac{T_{k+1}}{\gamma_k}\right) - \frac{1}{n}\sum_{k=1}^{n}\Delta_k + \frac{1}{n}\sum_{k=1}^{n}\Xi_{k+1}. \tag{19}$$

Finally, by linearity and applying an Abel's transform to the first term on the right-hand side of previous equality (see Delyon and Juditsky (1992) or Delyon and Juditsky (1993) for instance),

$$\begin{aligned}\Gamma_m(\overline{m}_n - m) = {} & \frac{m_1 - m}{n\gamma_1} - \frac{m_{n+1} - m}{n\gamma_n} + \frac{1}{n}\sum_{k=2}^{n}\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)(m_k - m) - \frac{1}{n}\sum_{k=1}^{n}\delta_k \\ & + \frac{1}{n}\sum_{k=1}^{n}\xi_{k+1}.\end{aligned} \tag{20}$$

6.1.3 The recursive estimator of the asymptotic variance

In order to simplify the proof of Theorem 4.1, we will introduce a new estimator of the variance. In this aim, let us now introduce the sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ defined for all $n \geq 1$ by $a_n := \exp\left(\frac{n^{1-s}}{2(1-s)}\right)$ and $b_n := \sum_{k=1}^{n} a_k^2$. Then, thanks to decomposition (18), let

$$\begin{aligned}\overline{T}_n &:= \frac{1}{\sqrt{b_n}}\sum_{k=1}^{n} a_k(m_k - m) \\ &= \frac{1}{\sqrt{b_n}}\left(\sum_{k=1}^{n}\frac{a_k}{\gamma_k}(T_k - T_{k+1}) + \sum_{k=1}^{n} a_k\Delta_k + \sum_{k=1}^{n} a_k\Xi_{k+1}\right) \\ &=: \frac{1}{\sqrt{b_n}}(A_{1,n} + A_{2,n} + M_{n+1}).\end{aligned} \tag{21}$$

In order to simplify several proofs, we now give L^p upper bounds of the terms on the right-hand side of previous equality.

Lemma 6.1. *Suppose assumptions **(A1)** to **(A5b)** hold. Then, for all positive integer p,*

$$\mathbb{E}\left[\left\|\sum_{k=1}^{n}\frac{a_k}{\gamma_k}\left(T_k - T_{k+1}\right)\right\|^{2p}\right] = O\left(\exp\left(\frac{pn^{1-s}}{1-s}\right)n^{p\alpha}\right),$$
$$\mathbb{E}\left[\left\|\sum_{k=1}^{n}a_k\Delta_k\right\|^{2p}\right] = O\left(\exp\left(\frac{pn^{1-s}}{1-s}\right)n^{p(s-\alpha}\right),$$
$$\mathbb{E}\left[\left\|\sum_{k=1}^{n}a_k\Xi_{k+1}\right\|^{2p}\right] = O\left(\exp\left(\frac{pn^{1-s}}{1-s}\right)n^{ps}\right)$$

The proof of this lemma as well as an analogous lemma which gives the asymptotic almost sure behavior of these terms are given in a Supplementary file. We can now introduce the following estimator

$$\overline{\Sigma}_n = \frac{1}{\sum_{k=1}^{n}k^{-\delta}}\sum_{k=1}^{n}\frac{1}{k^{\delta}}T_k\otimes T_k, \tag{22}$$

and one can decompose Σ_n as follows:

$$\begin{aligned}\Sigma_n - \Sigma = \Sigma_n &- \frac{1-\delta}{n^{1-\delta}}\sum_{k=1}^{n}\frac{1}{k^{\delta+s}}\exp\left(-\frac{k^{1-s}}{1-s}\right)\left(\sum_{j=1}^{k}e^{\frac{j^{1-s}}{2(1-s)}}\left(m_j-m\right)\right)\otimes\left(\sum_{j=1}^{k}e^{\frac{j^{1-s}}{2(1-s)}}\left(m_j-m\right)\right)\\
&+\frac{1-\delta}{n^{1-\delta}}\sum_{k=1}^{n}\frac{1}{k^{\delta+s}}\exp\left(-\frac{k^{1-s}}{1-s}\right)\left(\sum_{j=1}^{k}e^{\frac{j^{1-s}}{2(1-s)}}\left(m_j-m\right)\right)\otimes\left(\sum_{j=1}^{k}e^{\frac{j^{1-s}}{2(1-s)}}\left(m_j-m\right)\right)-\overline{\Sigma}_n\\
&+\overline{\Sigma}_n-\Sigma.\end{aligned}$$

6.2 Proof of Theorem 3.2

Proof of Theorem 3.2. Let us recall that the averaged algorithm can be written as

$$\begin{aligned}\Gamma_m\left(\overline{m}_n - m\right) = &\frac{m_1-m}{n\gamma_1} - \frac{m_{n+1}-m}{n\gamma_n} + \frac{1}{n}\sum_{k=2}^{n}\left(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}\right)\left(m_k-m\right) - \frac{1}{n}\sum_{k=1}^{n}\delta_k\\
&+\frac{1}{n}\sum_{k=1}^{n}\xi_{k+1}.\end{aligned}$$

It is proven in Godichon-Baggioni (2016b) that

$$\begin{aligned}\frac{\left\|m_1-m\right\|}{\sqrt{n}\gamma_1} &= o\left(1\right) \quad a.s,\\
\frac{\left\|m_{n+1}-m\right\|}{\sqrt{n}\gamma_n} &= o(1) \quad a.s,\\
\frac{1}{\sqrt{n}}\left\|\sum_{k=2}^{n}\left(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}\right)\left(m_k-m\right)\right\| &= o(1) \quad a.s,\\
\frac{1}{\sqrt{n}}\left\|\sum_{k=1}^{n}\delta_k\right\| &= o(1) \quad a.s.\end{aligned}$$

In order get the asymptotic normality of the martingale term $\left(\frac{1}{n}\sum_{k=1}^{n}\xi_{k+1}\right)$, let us check that assumptions of Theorem 5.1 in Jakubowski (1988) are fulfilled, i.e let $(e_i)_{i\in I}$ be an orthonormal basis of H and $\psi_{i,j} := \langle \Sigma' e_i, e_j\rangle$ for all $i,j \in I$, we have to verify

$$\forall \eta > 0, \quad \lim_{n\to\infty} \mathbb{P}\left(\sup_{1\le k\le n} \frac{1}{\sqrt{n}} \|\xi_{k+1}\| > \eta\right) = 0, \tag{23}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \langle \xi_{k+1}, e_i\rangle \langle \xi_{k+1}, e_j\rangle = \psi_{i,j} \quad a.s, \quad \forall i,j \in I, \tag{24}$$

$$\forall \epsilon > 0, \quad \lim_{N\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n}\sum_{j=N}^{\infty} \langle \xi_{k+1}, e_j\rangle^2 > \epsilon\right) = 0. \tag{25}$$

Proof of (23) Let $\eta > 0$, applying Markov's inequality,

$$\begin{aligned}\mathbb{P}\left(\sup_{1\le k\le n} \frac{1}{\sqrt{n}} \|\xi_{k+1}\| > \eta\right) &\le \sum_{k=1}^{n} \mathbb{P}\left(\frac{1}{\sqrt{n}} \|\xi_{k+1}\| > \eta\right) \\ &\le \frac{1}{n^2\eta^4}\sum_{k=1}^{n} \mathbb{E}\left[\|\xi_{k+1}\|^4\right].\end{aligned}$$

Then, applying Lemma H.1, there is a positive constant C such that

$$\mathbb{P}\left(\sup_{1\le k\le n} \frac{1}{\sqrt{n}} \|\xi_{k+1}\| > \eta\right) \le \frac{1}{n^2\eta^4}\sum_{k=1}^{n} C = \frac{C}{n\eta^4}.$$

Proof of (24). First, note that

$$\frac{1}{n}\sum_{k=1}^{n} \xi_{k+1}\otimes\xi_{k+1} = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[\xi_{k+1}\otimes\xi_{k+1}|\mathcal{F}_k\right] + \frac{1}{n}\sum_{k=1}^{n} \epsilon_{k+1},$$

with $\epsilon_{k+1} := \xi_{k+1}\otimes\xi_{k+1} - \mathbb{E}\left[\xi_{k+1}\otimes\xi_{k+1}|\mathcal{F}_k\right]$. Remark that (ϵ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$, and one can check that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \epsilon_{k+1} = 0 \quad a.s.$$

Let us now prove that the sequence of operators $\left(\mathbb{E}\left[\xi_{k+1}\otimes\xi_{k+1}|\mathcal{F}_k\right]\right)$ converges almost surely to Σ', with respect to the Frobenius norm. Note that

$$\begin{aligned}\left\|\mathbb{E}\left[\xi_{k+1}\otimes\xi_{k+1}|\mathcal{F}_k\right] - \Sigma'\right\| &= \left\|\mathbb{E}\left[\nabla_h g\left(X_{k+1}, m_k\right)\otimes\nabla_h g\left(X_{k+1}, m_k\right)|\mathcal{F}_k\right] - \Sigma' - \Phi\left(m_k\right)\otimes\Phi\left(m_k\right)\right\|_F \\ &\le \left\|\mathbb{E}\left[\nabla_h g\left(X_{k+1}, m_k\right)\otimes\nabla_h g\left(X_{k+1}, m_k\right)|\mathcal{F}_k\right] - \Sigma'\right\|_F + \left\|\Phi\left(m_k\right)\otimes\Phi\left(m_k\right)\right\|_F.\end{aligned}$$

Then, thanks to assumption **(A6a)**, since $\|\Phi(m_k)\| \le C\|m_k - m\|$ and since (m_k) converges

to m almost surely (see Godichon-Baggioni (2016b)),

$$\lim_{k\to\infty} \left\| \mathbb{E}\left[\nabla_h g\left(X_{k+1}, m_k\right) \otimes \nabla_h g\left(X_{k+1}, m_k\right) | \mathcal{F}_k\right] - \Sigma' \right\|_F = 0 \quad a.s,$$
$$\lim_{k\to\infty} \left\| \Phi\left(m_k\right) \otimes \Phi\left(m_k\right) \right\|_F = \lim_{n\to\infty} \left\| \Phi(m_k) \right\|^2 = 0 \quad a.s.$$

In a particular case, for all $i, j \in I$,

$$\lim_{k\to\infty} \left\langle \mathbb{E}\left[\xi_{k+1} \otimes \xi_{k+1} | \mathcal{F}_k\right](e_i), e_j \right\rangle = \psi_{i,j} := \left\langle \Sigma'(e_i), e_j \right\rangle \quad a.s.$$

Thus, applying Toeplitz's lemma,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \left\langle \mathbb{E}\left[\xi_{k+1} \otimes \xi_{k+1} | \mathcal{F}_k\right](e_i), e_j \right\rangle = \psi_{i,j} \quad a.s.$$

Finally, for all $i, j \in I$,

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \left\langle \xi_{k+1}, e_i \right\rangle \left\langle \xi_{k+1}, e_j \right\rangle &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \left\langle \xi_{k+1} \otimes \xi_{k+1}(e_i), e_j \right\rangle \\ &= \psi_{i,j} \quad a.s. \end{aligned}$$

Proof of (25). Let $\epsilon > 0$, applying Markov's inequality,

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n} \sum_{k=1}^{n} \sum_{j=N}^{\infty} \left\langle \xi_{k+1}, e_j \right\rangle > \epsilon \right) &\leq \frac{1}{n\epsilon^2} \sum_{k=1}^{n} \sum_{j=N}^{\infty} \mathbb{E}\left[\left\langle \xi_{k+1}, e_j \right\rangle^2 \right] \\ &= \frac{1}{n\epsilon^2} \sum_{k=1}^{n} \sum_{j=N}^{\infty} \mathbb{E}\left[\mathbb{E}\left[\left\langle \xi_{k+1}, e_j \right\rangle^2 | \mathcal{F}_k \right] \right]. \end{aligned}$$

Since for all $j \in I$, $\left\langle \xi_{k+1}, e_j \right\rangle^2 = \left\langle \xi_{k+1} \otimes \xi_{k+1}(e_j, e_j \right\rangle$, and by linearity

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n} \sum_{k=1}^{n} \sum_{j=N}^{\infty} \left\langle \xi_{k+1}, e_j \right\rangle > \epsilon \right) &\leq \frac{1}{\epsilon^2} \sum_{j=N}^{\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}\left[\mathbb{E}\left[\left\langle \xi_{k+1} \otimes \xi_{k+1}(e_j), e_j \right\rangle | \mathcal{F}_k \right] \right] \\ &= \frac{1}{\epsilon^2} \sum_{j=N}^{\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}\left[\left\langle \mathbb{E}\left[\xi_{k+1} \otimes \xi_{k+1} | \mathcal{F}_k\right](e_j), e_j \right\rangle \right]. \end{aligned}$$

Since $\mathbb{E}\left[\xi_{k+1} \otimes \xi_{k+1} | \mathcal{F}_k\right]$ converges almost surely to Σ' and by dominated convergence,

$$\limsup_n \mathbb{P}\left(\frac{1}{n} \sum_{k=1}^{n} \sum_{j=N}^{\infty} \left\langle \xi_{k+1}, e_j \right\rangle > \epsilon \right) \leq \frac{1}{\epsilon} \sum_{j=N}^{\infty} \left\langle \Sigma'(e_j), e_j \right\rangle.$$

Moreover, since $\Sigma' = \mathbb{E}\left[\nabla_h g\left(X, m\right) \otimes \nabla_h g\left(X, m\right)\right]$, thanks to assumption **(A5a)**,

$$\sum_{j=1}^{\infty} \left\langle \Sigma'(e_j), e_j \right\rangle = \left\| \mathbb{E}\left[\nabla_h g\left(X, m\right) \otimes \nabla_h g\left(X, m\right)\right] \right\|_F \leq \mathbb{E}\left[\left\| \nabla_h g\left(X, m\right) \right\|^2 \right] \leq L_1.$$

Thus, since for all $j \in I$, $\langle \Sigma'(e_j), e_j \rangle \geq 0$,

$$\lim_{N \to \infty} \sum_{j=N}^{\infty} \langle \Sigma'(e_j), e_j \rangle = 0,$$

which concludes the proof. □

6.3 Proof of Theorem 4.1

For the sake of simplicity, the proof is given for $mu = 0$ (the case where $\mu > 0$ is strictly analogous). Let us recall that equation (12) can be written as

$$\begin{aligned}\Sigma_n - \Sigma = \Sigma_n &- \frac{1-\delta}{n^{1-\delta}} \sum_{k=1}^{n} \frac{1}{k^{\delta+s}} \exp\left(-\frac{k^{1-s}}{1-s}\right) \left(\sum_{j=1}^{k} e^{\frac{j^{1-s}}{2(1-s)}} (m_j - m)\right) \otimes \left(\sum_{j=1}^{k} e^{\frac{j^{1-s}}{2(1-s)}} (m_j - m)\right) \\ &+ \frac{1-\delta}{n^{1-\delta}} \sum_{k=1}^{n} \frac{1}{k^{\delta+s}} \exp\left(-\frac{k^{1-s}}{1-s}\right) \left(\sum_{j=1}^{k} e^{\frac{j^{1-s}}{2(1-s)}} (m_j - m)\right) \otimes \left(\sum_{j=1}^{k} e^{\frac{j^{1-s}}{2(1-s)}} (m_j - m)\right) - \overline{\Sigma}_n \\ &+ \overline{\Sigma}_n - \Sigma. \end{aligned} \tag{26}$$

In order to prove Theorem 4.1, we just have to give the rates of convergence of the terms on the right-hand side of previous equality. The following lemma gives the almost sure and the rate of convergence in quadratic mean of the first term on the right-hand side of previous equality.

Lemma 6.2. *Suppose assumptions* ***(A1)*** *to* ***(A5a')*** *and* ***(A6b)*** *hold. Then, for all* $\gamma > 0$,

$$\left\| \Sigma_n - \frac{1-\delta}{n^{1-\delta}} \sum_{k=1}^{n} \frac{1}{k^{\delta+s}} e^{-\frac{k^{1-s}}{1-s}} \left(\sum_{j=1}^{k} e^{\frac{j^{1-s}}{2(1-s)}} (m_j - m)\right) \otimes \left(\sum_{j=1}^{k} e^{\frac{j^{1-s}}{2(1-s)}} (m_j - m)\right) \right\|_F^2 = o\left(\frac{(\ln n)^\gamma}{n^{1-s}}\right) \quad a.s.$$

Moreover, suppose assumption ***(A5b)*** *holds too. Then,*

$$\mathbb{E}\left[\left\| \Sigma_n - \frac{1-\delta}{n^{1-\delta}} \sum_{k=1}^{n} \frac{1}{k^{\delta+s}} e^{-\frac{k^{1-s}}{1-s}} \left(\sum_{j=1}^{k} e^{\frac{j^{1-s}}{2(1-s)}} (m_j - m)\right) \otimes \left(\sum_{j=1}^{k} e^{\frac{j^{1-s}}{2(1-s)}} (m_j - m)\right) \right\|_F^2 \right] = O\left(\frac{1}{n^{1-s}}\right).$$

The proof is given in a Supplementary file. The following lemma gives the almost sure and the rate of convergence in quadratic mean of the second term on the right-hand side of equality (26).

Lemma 6.3. *Suppose assumptions* ***(A1)*** *to* ***(A5a')*** *and* ***(A6b)*** *hold. Then, for all* $\gamma > 0$,

$$\left\| \frac{1-\delta}{n^{1-\delta}} \sum_{k=1}^{n} \frac{1}{k^{\delta+s}} e^{-\frac{k^{1-s}}{1-s}} \left(\sum_{j=1}^{k} e^{\frac{j^{1-s}}{2(1-s)}} (m_j - m)\right) \otimes \left(\sum_{j=1}^{k} e^{\frac{j^{1-s}}{2(1-s)}} (m_j - m)\right) - \overline{\Sigma}_n \right\|_F^2 = o\left(\frac{(\ln n)^\gamma}{n^{2(1-s)}}\right) \quad a.s.$$

Moreover, suppose assumption ***(A5b)*** *holds too. Then*

$$\mathbb{E}\left[\left\|\frac{1-\delta}{n^{1-\delta}}\sum_{k=1}^{n}\frac{1}{k^{\delta+s}}e^{-\frac{k^{1-s}}{1-s}}\left(\sum_{j=1}^{k}e^{\frac{j^{1-s}}{2(1-s)}}\left(m_j-m\right)\right)\otimes\left(\sum_{j=1}^{k}e^{\frac{j^{1-s}}{2(1-s)}}\left(m_j-m\right)\right)-\overline{\Sigma}_n\right\|_F^2\right]=O\left(\frac{1}{n^{2(1-s)}}\right).$$

The proof is given in a Supplementary file. Finally, the following Proposition gives the almost sure and the rate of convergence in quadratic mean of the last term on the right-hand side of equality (26).

Proposition 6.1. *Suppose assumptions* ***(A1)*** *to* ***(A5a')*** *and* ***(A6b)*** *hold. Then, there is a positive constant* γ *such that*

$$\left\|\overline{\Sigma}_n-\Sigma\right\|_F^2=o\left(\frac{(\ln n)^\delta}{n^{1-s}}\right)\quad a.s.$$

Suppose assumption ***(A5b)*** *holds too. Then, there is a positive constant* C *such that for all* $n\geq 1$,

$$\mathbb{E}\left[\left\|\overline{\Sigma}_n-\Sigma\right\|_F^2\right]\leq\frac{C}{n^{1-s}}.$$

Proof of Proposition 6.1. Applying equality (2), one can check that

$$\begin{aligned}\left\|\overline{\Sigma}_n-\Sigma\right\|_F\leq{}&\frac{1}{\sum_{k=1}^n k^{-\delta}}\sum_{k=1}^{n}\frac{1}{k^\delta b_k}\left\|A_{1,k}\right\|^2+\frac{1}{\sum_{k=1}^n k^{-\delta}}\sum_{k=1}^{n}\frac{1}{k^\delta b_k}\left\|A_{2,k}\right\|^2\\&+2\frac{1}{\sum_{k=1}^n k^{-\delta}}\sum_{k=1}^{n}\frac{1}{k^\delta b_k}\left\|A_{1,k}\right\|\left\|A_{2,k}\right\|+2\frac{1}{\sum_{k=1}^n k^{-\delta}}\sum_{k=1}^{n}\frac{1}{k^\delta b_k}\left\|A_{1,k}\right\|\left\|M_{k+1}\right\|\\&+2\frac{1}{\sum_{k=1}^n k^{-\delta}}\sum_{k=1}^{n}\frac{1}{k^\delta b_k}\left\|A_{2,k}\right\|\left\|M_{k+1}\right\|+\left\|\frac{1}{\sum_{k=1}^n k^{-\delta}}\sum_{k=1}^{n}\frac{1}{k^\delta}\left(\frac{1}{b_k}M_{k+1}\otimes M_{k+1}-\Sigma\right)\right\|_F,\end{aligned}\tag{27}$$

where $A_{1,k}, A_{2,k}, M_{k+1}$ are defined in (21). The following Lemma gives the rate of convergence in quadratic mean of the first terms on the right-hand side of previous inequality.

Lemma 6.4. *Suppose Assumptions* ***(A1)*** *to* ***(A6b)*** *hold. Then, for all* $i,j\in\{1,2\}$,

$$\mathbb{E}\left[\left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\sum_{k=1}^{n}\frac{1}{k^\delta b_k}\left\|A_{i,k}\right\|\left\|A_{j,k}\right\|\right)^2\right]=o\left(\frac{1}{n^{1-s}}\right),$$

$$\mathbb{E}\left[\left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\sum_{k=1}^{n}\frac{1}{k^\delta b_k}\left\|A_{i,k}\right\|\left\|M_{k+1}\right\|\right)^2\right]=o\left(\frac{1}{n^{1-s}}\right).$$

The proof of this lemma as well as its "almost sure version" are given in a Supplementary file.

Then, we just have to bound the last term on the right-hand side of inequality (27). First

let us decompose $M_{k+1} \otimes M_{k+1}$ as

$$M_{k+1} \otimes M_{k+1} = \sum_{j=1}^{k} a_j^2 \Xi_{j+1} \otimes \Xi_{j+1} + \sum_{j=1}^{k} a_j \Xi_{j+1} \otimes M_j + \sum_{j=1}^{k} a_j \Xi_{j+1} \otimes \left(M_{k+1} - M_{j+1}\right) + \sum_{j=1}^{k} a_j M_j \otimes \Xi_{j+1} + \sum_{j=1}^{k} a_j \left(M_{k+1} - M_{j+1}\right) \otimes \Xi_{j+1}.$$

Note that for all j, M_j is $\mathcal{F}_j$-measurable and $\mathbb{E}\left[\Xi_{j+1} \otimes M_j | \mathcal{F}_j\right] = 0$. Moreover,

$$\begin{aligned}\frac{1}{\sum_{k=1}^{n} k^{-\delta}} \sum_{k=1}^{n} \frac{1}{k^\delta} \left(\frac{1}{b_k} M_{k+1} \otimes M_{k+1} - \Sigma\right) &= \frac{1}{\sum_{k=1}^{n} k^{-\delta}} \sum_{k=1}^{n} \frac{1}{k^\delta} \frac{1}{b_k} \sum_{j=1}^{k} a_j \left(\Xi_{j+1} \otimes \Xi_{j+1} - \Sigma\right) \\ &+ \frac{1}{\sum_{k=1}^{n} k^{-\delta}} \sum_{k=1}^{n} \frac{1}{k^\delta} \frac{1}{b_k} \sum_{j=1}^{k} a_j \Xi_{j+1} \otimes M_j + \frac{1}{\sum_{k=1}^{n} k^{-\delta}} \sum_{k=1}^{n} \frac{1}{k^\delta} \frac{1}{b_k} \sum_{j=1}^{k} a_j \xi_{j+1} \otimes \left(M_{k+1} - M_{j+1}\right) \\ &+ \frac{1}{\sum_{k=1}^{n} k^{-\delta}} \sum_{k=1}^{n} \frac{1}{k^\delta} \frac{1}{b_k} \sum_{j=1}^{k} a_j M_j \otimes \Xi_{j+1} + \frac{1}{\sum_{k=1}^{n} k^{-\delta}} \sum_{k=1}^{n} \frac{1}{k^\delta} \frac{1}{b_k} \sum_{j=1}^{k} a_j \left(M_{k+1} - M_{j+1}\right) \otimes \Xi_{j+1}.\end{aligned}$$

The end of the proof consists in giving a bound of the quadratic mean of each term on the right-hand side of previous equality. Note that the almost sure rates of convergence are not proven since it is quite analogous.

Bounding $\mathbb{E}\left[\left\|\frac{1}{\sum_{k=1}^{n} k^{-\delta}} \sum_{k=1}^{n} \frac{1}{k^\delta} \frac{1}{b_k} \sum_{j=1}^{k} a_j \Xi_{j+1} \otimes M_j\right\|_F^2\right]$. First, note that

$$\frac{1}{\sum_{k=1}^{n} k^{-\delta}} \sum_{k=1}^{n} \frac{1}{k^\delta} \frac{1}{b_k} \sum_{j=1}^{k} a_j \Xi_{j+1} \otimes M_j = \frac{1}{\sum_{k=1}^{n} k^{-\delta}} \sum_{k=1}^{n} \left(\sum_{j=k}^{n} \frac{1}{k^\delta} \frac{1}{b_k}\right) a_k \Xi_{k+1} \otimes M_k.$$

Moreover, with the help of an integral test for convergence, one can check that there is a positive constant C such that for all positive integers $k \leq n$,

$$\sum_{j=k}^{n} \frac{1}{k^\delta} \frac{1}{b_k} \leq \frac{C}{k^\delta} \exp\left(-\frac{k^{1-s}}{(1-s)}\right). \tag{28}$$

Furthermore, since $\left(\Xi_{j+1} \otimes M_j\right)_j$ is a sequence of martingale differences adapted to the filtration $\left(\mathcal{F}_j\right)$, let

$$\begin{aligned}(*) &:= \mathbb{E}\left[\left\|\frac{1}{\sum_{k=1}^{n} k^{-\delta}} \sum_{k=1}^{n} \frac{1}{k^\delta} \frac{1}{b_k} \sum_{j=1}^{k} a_j \Xi_{j+1} \otimes M_j\right\|_F^2\right] \\ &= \mathbb{E}\left[\left\|\frac{1}{\sum_{k=1}^{n} k^{-\delta}} \sum_{k=1}^{n} \left(\sum_{j=k}^{n} \frac{1}{k^\delta} \frac{1}{b_k}\right) a_k \Xi_{k+1} \otimes M_k\right\|_F^2\right] \\ &= \left(\frac{1}{\sum_{k=1}^{n} k^{-\delta}}\right)^2 \sum_{k=1}^{n} \left(\sum_{j=k}^{n} \frac{1}{k^\delta} \frac{1}{b_k}\right)^2 a_k^2 \mathbb{E}\left[\|\Xi_{k+1} \otimes M_k\|_F^2\right]\end{aligned}$$

Then, applying equality (2) and Cauchy-Schwarz's inequality,

$$(*) \leq \left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2 \sum_{k=1}^n \left(\sum_{j=k}^n \frac{1}{k^\delta}\frac{1}{b_k}\right)^2 a_k^2 \mathbb{E}\left[\|\Xi_{k+1}\|^2 \|M_k\|^2\right]$$
$$\leq \left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2 \sum_{k=1}^n \left(\sum_{j=k}^n \frac{1}{k^\delta}\frac{1}{b_k}\right)^2 a_k^2 \sqrt{\mathbb{E}\left[\|\Xi_{k+1}\|^4\right] \mathbb{E}\left[\|M_k\|^4\right]}.$$

Finally, applying Lemmas 6.1 and H.1 as well as inequality (28),

$$(*) = O\left(\left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2 \sum_{k=1}^n \frac{1}{k^{2\delta - s}}\right) = O\left(\frac{1}{n^{1-s}}\right).$$

With analogous calculus, one can check

$$\mathbb{E}\left[\left\|\frac{1}{\sum_{k=1}^n k^{-\delta}} \sum_{k=1}^n \frac{1}{k^\delta}\frac{1}{b_k} \sum_{j=1}^k a_j M_j \otimes \Xi_{j+1}\right\|_F^2\right] = O\left(\frac{1}{n^{1-s}}\right).$$

Bounding $\mathbb{E}\left[\left\|\frac{1}{\sum_{k=1}^n k^{-\delta}} \sum_{k=1}^n \frac{1}{k^\delta}\frac{1}{b_k} \sum_{j=1}^k a_j \Xi_{j+1} \otimes \left(M_{k+1} - M_{j+1}\right)\right\|_F^2\right]$. First, note that

$$\sum_{j=1}^k a_j \Xi_{j+1} \otimes \left(M_{k+1} - M_j\right) = \sum_{j=1}^k \sum_{j'=j+1}^k a_j a_{j'} \Xi_{j+1} \otimes \Xi_{j'+1}$$
$$= \sum_{j'=2}^k \sum_{j=1}^{j'-1} a_j a_{j'} \Xi_{j+1} \otimes \Xi_{j'+1}.$$

Note that $\left(\sum_{j=1}^{j'-1} a_j a_{j'} \Xi_{j+1} \otimes \Xi_{j'+1}\right)_{j'}$ is a sequence of martingale differences adapted to the filtration $\left(\mathcal{F}_{j'}\right)$. Furthermore,

$$\mathbb{E}\left[\left\|\frac{1}{\sum_{k=1}^n k^{-\delta}} \sum_{k=1}^n \frac{1}{k^\delta}\frac{1}{b_k} \sum_{j=1}^k a_j \Xi_{j+1} \otimes \left(M_{k+1} - M_j\right)\right\|_F^2\right]$$
$$= \left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2 \sum_{k=1}^n \frac{1}{k^{2\delta}}\frac{1}{b_k^2} \mathbb{E}\left[\left\|\sum_{j'=2}^k \sum_{j=1}^{j'-1} a_j a_{j'} \Xi_{j+1} \otimes \Xi_{j'+1}\right\|_F^2\right]$$
$$+ \left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2 \mathbb{E}\left[\sum_{k=2}^n \sum_{j=1}^{k-1} b_k^{-1} k^{-\delta} b_j^{-1} j^{-\delta} \left\langle \sum_{j''=2}^j \sum_{j'=1}^{j''-1} a_{j'} a_{j''} \Xi_{j'+1} \otimes \Xi_{j''+1}, \sum_{i''=2}^k \sum_{i'=1}^{i''-1} a_{i'} a_{i''} \Xi_{i'+1} \otimes \Xi_{i''+1} \right\rangle\right].$$

Then end of the proof consists in bounding the two terms on the right-hand side of previous equality. First, since $\left(\sum_{j=1}^{j'-1} a_j a_{j'} \Xi_{j+1} \otimes \Xi_{j'+1}\right)_{j'}$ is a sequence of martingale differences

adapted to the filtration $(\mathcal{F}_{j'})$, let

$$(\star) := \left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2 \sum_{k=1}^n \frac{1}{k^{2\delta}} \frac{1}{b_k^2} \mathbb{E}\left[\left\| \sum_{j'=2}^k \sum_{j=1}^{j'-1} a_j a_{j'} \Xi_{j+1} \otimes \Xi_{j'+1} \right\|_F^2\right]$$
$$= \left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2 \sum_{k=1}^n \frac{1}{k^{2\delta}} \frac{1}{b_k^2} \sum_{j'=2}^k \mathbb{E}\left[\left\| \sum_{j=1}^{j'-1} a_j a_{j'} \Xi_{j+1} \otimes \Xi_{j'+1} \right\|_F^2\right].$$

Then, applying equality (2) and Cauchy-Schwarz's inequality,

$$(\star) = \left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2 \sum_{k=1}^n \frac{1}{k^{2\delta}} \frac{1}{b_k^2} \sum_{j'=2}^k a_{j'}^2 \mathbb{E}\left[\left\| \sum_{j=1}^{j'-1} a_j \Xi_{j+1} \right\|^2 \left\| \Xi_{j'+1} \right\|^2\right]$$
$$\leq \left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2 \sum_{k=1}^n \frac{1}{k^{2\delta}} \frac{1}{b_k^2} \sum_{j'=2}^k a_{j'}^2 \sqrt{\mathbb{E}\left[\left\| \Xi_{j'+1} \right\|^4\right]} \sqrt{\mathbb{E}\left[\left\| \sum_{j=1}^{j'-1} a_j \Xi_{j+1} \right\|^4\right]}$$

Finally, applying Lemma H.1, H.2 and 6.1,

$$(\star) = O\left(\left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2 \sum_{k=1}^n \frac{1}{k^{2\delta}} \frac{1}{b_k^2} \sum_{j'=2}^k a_{j'}^4 j'^s\right) = O\left(\left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2 \sum_{k=1}^n \frac{1}{k^{2\delta}} \frac{1}{b_k^2} a_k^4 k^{2s}\right) = O\left(\frac{1}{n^{\min\{2-2\delta,1\}}}\right).$$

Then, since $\delta < (1+s)/2$,

$$\left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2 \sum_{k=1}^n \frac{1}{k^{2\delta}} \frac{1}{b_k^2} \mathbb{E}\left[\left\| \sum_{j'=2}^k \sum_{j=1}^{j'-1} a_j a_{j'} \Xi_{j+1} \otimes \Xi_{j'+1} \right\|_F^2\right] = o\left(\frac{1}{n^{1-s}}\right).$$

In the same way, by linearity, let

$$(\star\star) := \left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2 \mathbb{E}\left[\sum_{k=2}^n \sum_{j=1}^{k-1} b_k^{-1} k^{-\delta} b_j^{-1} j^{-\delta} \left\langle \sum_{j''=2}^j \sum_{j'=1}^{j''-1} a_{j'} a_{j''} \Xi_{j'+1} \otimes \Xi_{j''+1}, \sum_{i''=2}^k \sum_{i'=1}^{i''-1} a_{i'} a_{i''} \Xi_{i'+1} \otimes \Xi_{i''+1} \right\rangle\right]$$
$$= \left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2 \sum_{k=2}^n \sum_{j=1}^{k-1} b_k^{-1} k^{-\delta} b_j^{-1} j^{-\delta} \mathbb{E}\left[\left\langle \sum_{j''=2}^j \sum_{j'=1}^{j''-1} a_{j'} a_{j''} \Xi_{j'+1} \otimes \Xi_{j''+1}, \sum_{i''=2}^j \sum_{i'=1}^{i''-1} a_{i'} a_{i''} \Xi_{i'+1} \otimes \Xi_{i''+1} \right\rangle_F\right]$$
$$+ \left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2 \sum_{k=2}^n \sum_{j=1}^{k-1} b_k^{-1} k^{-\delta} b_j^{-1} j^{-\delta} \mathbb{E}\left[\left\langle \sum_{j''=2}^j \sum_{j'=1}^{j''-1} a_{j'} a_{j''} \Xi_{j'+1} \otimes \Xi_{j''+1}, \sum_{i''=j+1}^k \sum_{i'=1}^{i''-1} a_{i'} a_{i''} \Xi_{i'+1} \otimes \Xi_{i''+1} \right\rangle_F\right].$$

Since $(\Xi_{i''})$ is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_{i''})$,

$$\begin{aligned}
&\sum_{k=2}^{n}\sum_{j=1}^{k-1} b_k^{-1}k^{-\delta}b_j^{-1}j^{-\delta}\mathbb{E}\left[\left\langle \sum_{j''=2}^{j}\sum_{j'=1}^{j''-1} a_{j'}a_{j''}\Xi_{j'+1}\otimes\Xi_{j''+1}, \sum_{i''=j+1}^{k}\sum_{i'=1}^{i''-1} a_{i'}a_{i''}\Xi_{i'+1}\otimes\Xi_{i''+1}\right\rangle_F\right] \\
&\quad = \sum_{k=2}^{n}\sum_{j=1}^{k-1} b_k^{-1}k^{-\delta}b_j^{-1}j^{-\delta}\sum_{j''=2}^{j}\sum_{j'=1}^{j''-1}\sum_{i''=j+1}^{k}\sum_{i'=1}^{i''-1} a_{i'}a_{i''}a_{j'}a_{j''}\mathbb{E}\left[\left\langle \Xi_{j'+1}\otimes\Xi_{j''+1}, \Xi_{i'+1}\otimes\Xi_{i''+1}\right\rangle_F\right] \\
&\quad = \sum_{k=2}^{n}\sum_{j=1}^{k-1} b_k^{-1}k^{-\delta}b_j^{-1}j^{-\delta}\sum_{j''=2}^{j}\sum_{j'=1}^{j''-1}\sum_{i''=j+1}^{k}\sum_{i'=1}^{i''-1} a_{i'}a_{i''}a_{j'}a_{j''}\mathbb{E}\left[\left\langle \Xi_{j'+1}\otimes\Xi_{j''+1}, \Xi_{i'+1}\otimes\mathbb{E}\left[\Xi_{i''+1}|\mathcal{F}_{i''}\right]\right\rangle_F\right] \\
&\quad = 0.
\end{aligned}$$

Furthermore, since $\left(\sum_{j''=2}^{j}\sum_{j'=1}^{j''-1} a_{j'}a_{j''}\Xi_{j'+1}\otimes\Xi_{j''+1}\right)_{j''}$ is a sequence of martingale differences adapted to the filtration $\left(\mathcal{F}_{j''}\right)$ and applying equality (2),

$$\begin{aligned}
(\star\star) &= \left(\frac{1}{\sum_{k=1}^{n}k^{-\delta}}\right)^2\sum_{k=2}^{n}\sum_{j=1}^{k} b_k^{-1}k^{-\delta}b_j^{-1}j^{-\delta}\mathbb{E}\left[\left\|\sum_{j''=1}^{j}\sum_{j'=1}^{j''-1} a_{j'}a_{j''}\Xi_{j'+1}\otimes\Xi_{j''+1}\right\|_F^2\right] \\
&= \left(\frac{1}{\sum_{k=1}^{n}k^{-\delta}}\right)^2\sum_{k=2}^{n}\sum_{j=1}^{k} b_k^{-1}k^{-\delta}b_j^{-1}j^{-\delta}\sum_{j''=1}^{j}\mathbb{E}\left[\left\|\sum_{j'=1}^{j''-1} a_{j'}a_{j''}\Xi_{j'+1}\otimes\Xi_{j''+1}\right\|_F^2\right] \\
&= \left(\frac{1}{\sum_{k=1}^{n}k^{-\delta}}\right)^2\sum_{k=2}^{n}\sum_{j=1}^{k} b_k^{-1}k^{-\delta}b_j^{-1}j^{-\delta}\sum_{j''=1}^{j}a_{j''}^2\mathbb{E}\left[\left\|\sum_{j'=1}^{j''-1} a_{j'}\Xi_{j'+1}\right\|^2\left\|\Xi_{j''+1}\right\|^2\right].
\end{aligned}$$

Applying Cauchy-Schwarz's inequality as well as Lemmas H.1 and 6.1,

$$\begin{aligned}
(\star\star) &\leq \left(\frac{1}{\sum_{k=1}^{n}k^{-\delta}}\right)^2\sum_{k=1}^{n}\sum_{j=1}^{k} b_k^{-1}k^{-\delta}b_j^{-1}j^{-\delta}\sum_{j''=1}^{j}a_{j''}^2\sqrt{\mathbb{E}\left[\left\|\sum_{j'=1}^{j''-1} a_{j'}\Xi_{j'+1}\right\|_F^4\right]\mathbb{E}\left[\left\|\Xi_{j''+1}\right\|_F^4\right]} \\
&= O\left(\left(\frac{1}{\sum_{k=1}^{n}k^{-\delta}}\right)^2\sum_{k=1}^{n}\sum_{j=1}^{k} b_k^{-1}k^{-\delta}b_j^{-1}j^{-\delta}\sum_{j''=1}^{j}a_{j''}^4j''^s\right).
\end{aligned}$$

Finally, applying Lemma H.2,

$$\begin{aligned}
(\star\star) &= O\left(\left(\frac{1}{\sum_{k=1}^{n}k^{-\delta}}\right)^2\sum_{k=1}^{n}\sum_{j=1}^{k} b_k^{-1}k^{-\delta}b_j^{-1}j^{-\delta}a_j^4j^{2s}\right) \\
&= O\left(\left(\frac{1}{\sum_{k=1}^{n}k^{-\delta}}\right)^2\sum_{k=1}^{n} b_k^{-1}k^{-2\delta}k^{2s}a_k^2\right) \\
&= O\left(\frac{1}{n^{1-s}}\right).
\end{aligned}$$

Thus,

$$\mathbb{E}\left[\left\|\frac{1}{\sum_{k=1}^{n}k^{-\delta}}\sum_{k=1}^{n}\frac{1}{k^\delta}\frac{1}{b_k}\sum_{j=1}^{k}a_j\Xi_{j+1}\otimes\left(M_{k+1}-M_{j+1}\right)\right\|_F^2\right] = O\left(\frac{1}{n^{1-s}}\right).$$

Moreover, with analogous calculus, one can check

$$\mathbb{E}\left[\left\|\frac{1}{\sum_{k=1}^n k^{-\delta}}\sum_{k=1}^n \frac{1}{k^\delta}\frac{1}{b_k}\sum_{j=1}^k a_j\left(M_{k+1}-M_{j+1}\right)\otimes\Xi_{j+1}\right\|_F^2\right] = O\left(\frac{1}{n^{1-s}}\right).$$

Bounding $\frac{1}{\sum_{k=1}^n k^{-\delta}}\sum_{k=1}^n \frac{1}{k^\delta b_k}\sum_{j=1}^k a_k^2\left(\Xi_{k+1}\otimes\Xi_{k+1}-\Sigma\right)$**.** First , note that

$$\begin{aligned}\frac{1}{\sum_{k=1}^n k^{-\delta}}\sum_{k=1}^n \frac{1}{k^\delta b_k}\sum_{j=1}^k a_k^2\left(\Xi_{k+1}\otimes\Xi_{k+1}-\Sigma\right) &= \frac{1}{\sum_{k=1}^n k^{-\delta}}\sum_{k=1}^n \frac{1}{k^\delta b_k}\sum_{j=1}^k a_k^2\left(\mathbb{E}\left[\Xi_{k+1}\otimes\Xi_{k+1}|\mathcal{F}_k\right]-\Sigma\right)\\ &+\frac{1}{\sum_{k=1}^n k^{-\delta}}\sum_{k=1}^n \frac{1}{k^\delta b_k}\sum_{j=1}^k a_k^2\left(\Xi_{k+1}\otimes\Xi_{k+1}-\mathbb{E}\left[\Xi_{k+1}\otimes\Xi_{k+1}|\mathcal{F}_k\right]\right)\end{aligned}$$

The end of the proof consists in bounding the quadratic mean of the terms on the right-hand side of previous equality. First, applying Lemma H.4, let

$$\begin{aligned}(\star) &:= \mathbb{E}\left[\left\|\frac{1}{\sum_{k=1}^n k^{-\delta}}\sum_{k=1}^n \frac{1}{k^\delta b_k}\sum_{j=1}^k a_j^2\left(\mathbb{E}\left[\Xi_{k+1}\otimes\Xi_{k+1}|\mathcal{F}_k\right]-\Sigma\right)\right\|_F^2\right]\\ &\leq \left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2\left(\sum_{k=1}^n \frac{1}{k^\delta b_k}\sqrt{\mathbb{E}\left[\left\|\sum_{j=1}^k a_j^2\left(\mathbb{E}\left[\Xi_{k+1}\otimes\Xi_{k+1}|\mathcal{F}_k\right]-\Sigma\right)\right\|_F^2\right]}\right)^2\\ &\leq \left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2\left(\sum_{k=1}^n \frac{1}{k^\delta b_k}\sum_{j=1}^k a_j^2\sqrt{\mathbb{E}\left[\left\|\mathbb{E}\left[\Xi_{k+1}\otimes\Xi_{k+1}|\mathcal{F}_k\right]-\Sigma\right\|_F^2\right]}\right)^2\end{aligned}$$

Then, applying inequality (6) and Corollary H.1,

$$\begin{aligned}(\star) &= O\left(\left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2\left(\sum_{k=1}^n \frac{1}{k^\delta b_k}\sum_{j=1}^k a_j^2\sqrt{\mathbb{E}\left[\left\|m_n-m\right\|^2\right]}\right)^2\right)\\ &= O\left(\left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2\left(\sum_{k=1}^n \frac{1}{k^\delta b_k}\sum_{j=1}^k a_j^2 j^{-\alpha/2}\right)^2\right).\end{aligned}$$

Furthermore, thanks to Lemma H.2,

$$(\star) = O\left(\left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2\left(\sum_{k=1}^n \frac{1}{k^\delta b_k}a_k^2 k^{s-\alpha/2}\right)^2\right) = O\left(\left(\frac{1}{\sum_{k=1}^n k^{-\delta}}\right)^2 n^{2-2\delta-\alpha}\right) = O\left(\frac{1}{n^\alpha}\right).$$

Thus, since $\alpha > 1/2$,

$$\mathbb{E}\left[\left\|\frac{1}{\sum_{k=1}^n k^{-\delta}}\sum_{k=1}^n \frac{1}{k^\delta b_k}\sum_{j=1}^k a_j^2\left(\mathbb{E}\left[\Xi_{k+1}\otimes\Xi_{k+1}|\mathcal{F}_k\right]-\Sigma\right)\right\|_F^2\right] = o\left(\frac{1}{n^{1-s}}\right).$$

Moreover, applying Lemma H.4, let

$$(\star\star) := \mathbb{E}\left[\left\|\frac{1}{\sum_{k=1}^{n} k^{-\delta}}\sum_{k=1}^{n}\frac{1}{k^{\delta}b_k}\sum_{j=1}^{k} a_j^2\left(\Xi_{k+1}\otimes\Xi_{k+1} - \mathbb{E}\left[\Xi_{k+1}\otimes\Xi_{k+1}|\mathcal{F}_k\right]\right)\right\|_F^2\right]$$
$$\leq \left(\frac{1}{\sum_{k=1}^{n} k^{-\delta}}\right)^2\left(\sum_{k=1}^{n}\frac{1}{k^{\delta}b_k}\sqrt{\mathbb{E}\left[\left\|\sum_{j=1}^{k} a_j^2\left(\Xi_{k+1}\otimes\Xi_{k+1} - \mathbb{E}\left[\Xi_{k+1}\otimes\Xi_{k+1}|\mathcal{F}_k\right]\right)\right\|_F^2\right]}\right)^2.$$

Furthermore, since $\left(\mathbb{E}\left[\Xi_{k+1}\otimes\Xi_{k+1}|\mathcal{F}_k\right] - \Xi_{k+1}\otimes\Xi_{k+1}\right)$ is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_k)$ and applying Lemma H.1,

$$(\star\star) \leq \left(\frac{1}{\sum_{k=1}^{n} k^{-\delta}}\right)^2\left(\sum_{k=1}^{n}\frac{1}{k^{\delta}b_k}\sqrt{\sum_{j=1}^{k} a_j^4\mathbb{E}\left[\left\|\left(\Xi_{k+1}\otimes\Xi_{k+1} - \mathbb{E}\left[\Xi_{k+1}\otimes\Xi_{k+1}|\mathcal{F}_k\right]\right)\right\|_F^2\right]}\right)^2$$
$$= O\left(\left(\frac{1}{\sum_{k=1}^{n} k^{-\delta}}\right)^2\left(\sum_{k=1}^{n}\frac{1}{k^{\delta}b_k}\sqrt{\sum_{j=1}^{k} a_j^4}\right)^2\right).$$

Then, applying Lemma H.2,

$$(\star\star) = O\left(\left(\frac{1}{\sum_{k=1}^{n} k^{-\delta}}\right)^2\left(\sum_{k=1}^{n}\frac{1}{k^{\delta}b_k}a_k^2 k^{s/2}\right)^2\right) = O\left(\left(\frac{1}{\sum_{k=1}^{n} k^{-\delta}}\right)^2\left(\sum_{k=1}^{n} k^{-\delta-s/2}\right)^2\right) = O\left(\frac{1}{n^{2-s}}\right).$$

Finally,

$$\mathbb{E}\left[\left\|\frac{1}{\sum_{k=1}^{n} k^{-\delta}}\sum_{k=1}^{n}\frac{1}{k^{\delta}b_k}\sum_{j=1}^{k} a_j^2\left(\Xi_{k+1}\otimes\Xi_{k+1} - \mathbb{E}\left[\Xi_{k+1}\otimes\Xi_{k+1}|\mathcal{F}_k\right]\right)\right\|_F^2\right] = o\left(\frac{1}{n^{1-s}}\right),$$

which concludes the proof. □

References

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

Bach, F. and Moulines, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate O (1/n). In *Advances in Neural Information Processing Systems*, pages 773–781.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

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

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

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

Chakraborty, A. and Chaudhuri, P. (2014). The spatial distribution in infinite dimensional spaces and related quantiles and depths. *The Annals of Statistics*, 42:1203–1231.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.*, 91(434):862–872.

Delyon, B. and Juditsky, A. (1992). Stochastic optimization with averaging of trajectories. *Stochastics: An International Journal of Probability and Stochastic Processes*, 39(2-3):107–118.

Delyon, B. and Juditsky, A. (1993). Accelerated stochastic approximation. *SIAM Journal on Optimization*, 3(4):868–881.

Dippon, J. and Renz, J. (1997). Weighted means in stochastic approximation of minima. *SIAM Journal on Control and Optimization*, 35(5):1811–1827.

Dippon, J. and Walk, H. (2006). The averaged robbins–monro method for linear problems in a banach space. *Journal of Theoretical Probability*, 19(1):166–189.

Duflo, M. (1996). *Algorithmes stochastiques*. Springer Berlin.

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

Fabian, V. (1968). On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, pages 1327–1332.

Gahbiche, M. and Pelletier, M. (2000). On the estimation of the asymptotic covariance matrix for the averaged robbins–monro algorithm. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 331(3):255–260.

Gervini, D. (2008). Robust functional estimation using the median and spherical principal components. *Biometrika*, 95(3):587–600.

Godichon-Baggioni, A. (2016a). Estimating the geometric median in hilbert spaces with stochastic gradient algorithms: Lp and almost sure rates of convergence. *Journal of Multivariate Analysis*, 146:209–222.

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

Godichon-Baggioni, A., Portier, B., et al. (2017). An averaged projected robbins-monro algorithm for estimating the parameters of a truncated spherical distribution. *Electronic Journal of Statistics*, 11(1):1890–1927.

Haldane, J. B. S. (1948). Note on the median of a multivariate distribution. *Biometrika*, 35(3-4):414–417.

Hallin, M. and Paindaveine, D. (2006). Semiparametrically efficient rank-based inference for shape. i. optimal rank-based tests for sphericity. *The Annals of Statistics*, 34(6):2707–2756.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Huber, P. and Ronchetti, E. (2009). *Robust Statistics*. John Wiley and Sons, second edition.

Jakubowski, A. (1988). Tightness criteria for random measures with application to the principle of conditioning in Hilbert spaces. *Probab. Math. Statist.*, 9(1):95–114.

Juditsky, A., Nesterov, Y., et al. (2014). Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44–80.

Kemperman, J. (1987). The median of a finite measure on a Banach space. In *Statistical data analysis based on the L_1-norm and related methods (Neuchâtel, 1987)*, pages 217–230. North-Holland, Amsterdam.

Kraus, D. and Panaretos, V. M. (2012). Dispersion operators and resistant second-order functional data analysis. *Biometrika*, 99:813–832.

Kushner, H. J. and Yin, G. (2003a). *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media.

Kushner, H. J. and Yin, G. G. (2003b). *Stochastic approximation and recursive algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition. Stochastic Modelling and Applied Probability.

Ljung, L., Pflug, G. C., and Walk, H. (2012). *Stochastic approximation and optimization of random systems*, volume 17. Birkhäuser.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester. Theory and methods.

Minsker, S., Srivastava, S., Lin, L., and Dunson, D. (2014). Scalable and robust bayesian inference via the median posterior. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1656–1664.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

Oja, H. and Niinimaa, A. (1985). Asymptotic properties of the generalized median in the case of multivariate normality. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 372–377.

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

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

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

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

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

Schwabe, R. and Walk, H. (1996). On a stochastic approximation procedure based on averaging. *Metrika*, 44(1):165–180.

Serfling, R. (2006). Depth functions in nonparametric multivariate inference. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 72:1.

Vardi, Y. and Zhang, C.-H. (2000). The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA*, 97(4):1423–1426.

Walk, H. (1992). Foundations of stochastic approximation. In *Stochastic Approximation and Optimization of Random Systems*, pages 1–51. Springer.