On the rates of convergence of Parallelized Averaged Stochastic Gradient Algorithms

Antoine Godichon-Baggioni*, Sofiane Saadane**
* Laboratoire de Mathématiques de l'INSA de Rouen,
INSA de Rouen, 76800 Saint-Etienne-du-Rouvray, France
** Institut de Mathématiques de Toulouse,
INSA de Toulouse, 31000 Toulouse, France
email: * antoine.godichon@insa-rouen.fr, ** saadane@insa-toulouse.fr,

Abstract

The growing interest for high dimensional and functional data analysis led in the last decade to an important research developing a consequent amount of techniques. Parallelized algorithms, which consist in distributing and treat the data into different machines, for example, are a good answer to deal with large samples taking values in high dimensional spaces. We introduce here a parallelized averaged stochastic gradient algorithm, which enables to treat efficiently and recursively the data, and so, without taking care if the distribution of the data into the machines is uniform. The rate of convergence in quadratic mean as well as the asymptotic normality of the parallelized estimates are given, for strongly and locally strongly convex objectives.

Keywords: Stochastic Gradient Descent, Averaging, Distributed estimation, Central Limit Theorem, Asynchronous parallel optimization.

1 Introduction

The growing interest for high dimensional and functional data analysis led in the last decade to many research papers developing a consequent amount of techniques. Data that have to face statisticians can now be extremely large so that it creates a need for economic calculation techniques. Parallelized algorithms are now a good answer to this challenge and authors are using various techniques and software (multicore processors for example). For instance, [27] deals with gradient descent for least square type functions (strongly convex objective functions) by using a global averaging technique meaning that each machine carries out a gradient descent and the final outcome of the procedure is an averaging of all these results. In a recent work, [24] also deals with stochastic gradient by proposing a rewriting procedure where each processor can rewrite the data of an other. This paper offers good numerical results proving its efficiency and showing in particular that the rewriting technique is sparse (meaning it does affect a too much important part of the data). We can

also point out the work of [19] where two implementations of stochastic gradient are coupled : one is over a computer network and the other one is on a shared memory system. Finally, [3] introduces a parallelized averaged stochastic gradient algorithm and establishes some convergence properties, such as the asymptotic normality. Those four examples show in particular that the literature assumes (most of the time) that each machine, when working with a parallelized algorithm, receive the same amount of data. The reality of course is far from this setup. Indeed, in many practical cases, data are acquired and treated by different machines (see the nice example of software architecture given in [4]). For this reason, it is interesting to investigate a parallelized algorithm when the distribution of the data into the machines is not uniform. Moreover, data are often acquired sequentially, then, it is important to have an algorithm which enables to simply update the estimates. We so focus on the parallelization of averaged stochastic gradient algorithms.

Stochastic Gradient Descents (SGD for short) are usually used for estimating the minimizer of a convex function and are commonly fast, do not need to store all the data into memory, and are recursive, which enables to simply update the estimates when the data arrive sequentially [25, 9, 18]. In order to improve the convergence, [26] and [23] introduced the Averaged Stochastic Gradient Descent (ASGD for short). Among the studied cases, two of them attract a lot of attention : globally strongly convex objectives [2, 11], and locally strongly convex ones [21, 22, 13, 14]. Indeed, in those cases the theoretical study can be pushed very far due to the quite nice structure of the objective function.

In this paper, we introduce Parallelized Stochastic Gradient (PASG) algorithm, which consists in running p samples of ASGD with sample sizes n_i, $i \in \{1, ..., p\}$. After a run, the results are centralized using an averaging step, *i.e.* taking the arithmetic mean of all the estimates obtained with each SGD (or equivalently taking the weighted mean of those obtained with each ASGD). The interest of this procedure is its ability to deal with large samples not necessarily with the same size (we can suppose $n_i \neq n_j$ for any $i \neq j$). We then establish the efficiency of the algorithms by proving that they have a quadratic convergence rate of $O\left(\frac{1}{n}\right)$, which is the optimal one for stochastic approximation. In a second time, we establish the asymptotic normality of the estimates and see that it has an optimal asymptotic variance [22].

The paper is organized as follows: Section 2, some recalls on SGD and its averaged version are done and the general framework as well as the PASG algorithm are introduced. The two contexts (globally and locally convex objective) are introduced in Section 3 as well as the rate of convergence in quadratic mean and the asymptotic normality of the estimates obtained with the PASG algortihm. Section 4, a simulation study for the estimation of the geometric median shows the efficiency of the method. Finally, the proofs are postponed in Section 5.

2 Framework and algorithms

2.1 General framework and usual averaged stochastic gradient algorithm

Let X be a random variable taking values in a space $\mathcal{X}$, and let H be a separable Hilbert space, not necessarily of finite dimension, such as $\mathbb{R}^d$ or $L^2(I)$, for some close interval $I \subset \mathbb{R}$. In what follows, we denote by $\langle .,. \rangle$ its inner product, and by $\|.\|$ the associated norm. Let $G : H \longrightarrow \mathbb{R}$ be the function we would like to minimize, defined for all $h \in H$ by

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

where $g : \mathcal{X} \times H \longrightarrow \mathbb{R}$. We consider from now that the functional G is convex, and that for almost every $x \in \mathcal{X}$, the functional $g\left(x, .\right)$ is Fréchet-differentiable for the second variable and we denote by $\nabla_h g\left(x, .\right)$ its gradient. Let $X_1, ..., X_n, ...$ be random variables with the same law as X, the stochastic gradient descent (SGD for short) is defined recursively for all $n \geq 1$ by ([25])

$$m_{n+1} = m_n - \gamma_n \nabla_h g\left(X_{n+1}, m_n\right), \tag{2}$$

with m_1 bounded, and a step sequence $\left(\gamma_n\right)_{n \geq 1}$ of the form $\gamma_n := c_\gamma n^{-\alpha}$, with $c_\gamma > 0$ and $\alpha \in \left(\frac{1}{2}, 1\right)$. Remark that it is possible to take a step sequence of the form $\gamma_n = \frac{c}{n}$, but it necessitates to have some information on the smallest eigenvalue of the Hessian of the functional G at m ([21]). In order to improve the convergence, [26] (see also [23] for first results) introduced the averaged stochastic gradient descent (ASGD for short), defined recursively for all $n \geq 1$ by

$$\overline{m}_{n+1} = \overline{m}_n + \frac{1}{n+1}\left(m_{n+1} - \overline{m}_n\right), \tag{3}$$

with $\overline{m}_1 = m_1$. We speak about averaging since it can be written as

$$\overline{m}_n = \frac{1}{n}\sum_{k=1}^{n} m_k. \tag{4}$$

2.2 The parallelized averaged stochastic gradient algorithm

We consider from now a set $\{1, ..., p\}$ of machines. The data are spread over the machines, *i.e.* each entity i receives sequentially a sequence of independent random variables $X_{i,1}, ..., X_{i,k}, ...$ Then, each entity $i = 1, ..., p$ will compute the SGD and its averaged version defined recursively for all $k \geq 1$ by

$$\left\{ \begin{array}{ll} m_{i,k+1} & = m_{i,k} - \gamma_k \nabla_h g\left(X_{i,k+1}, m_{i,k}\right), \\ \overline{m}_{i,k+1} & = \overline{m}_{i,k} + \frac{1}{k+1}\left(m_{i,k+1} - \overline{m}_{i,k}\right), \end{array} \right.$$

with $m_{1,1} = \overline{m}_{1,1}, ..., m_{p,1} = \overline{m}_{p,1}$ bounded and $\left(\gamma_k\right)_{k \geq 1}$ a step sequence of the form $\gamma_k = c_\gamma k^{-\alpha}$, with $c_\gamma > 0$ and $\alpha \in \left(\frac{1}{2}, 1\right)$. Let $n = n_1 + ... + n_p$, the parallelized averaged stochastic

gradient estimate at time n is defined by

$$\widehat{m}_n := \frac{1}{\sum_{i=1}^{p} n_i} \sum_{i=1}^{p} n_i \overline{m}_{i,n_i}, \tag{5}$$

which can be written as

$$\widehat{m}_n = \frac{1}{\sum_{i=1}^{p} n_i} \sum_{i=1}^{p} \sum_{k=1}^{n_i} m_{i,k}. \tag{6}$$

Moreover, this algorithm can be written recursively. Indeed, setting $n' = n'_1 + ... + n'_p$, such that for all $i = 1, ..., p$ we have $n'_i \geq n_i$,

$$\widehat{m}_{n'} = \frac{\sum_{i=1}^{p} n_i}{\sum_{i=1}^{p} n'_i} \widehat{m}_n + \frac{1}{\sum_{i=1}^{p} n'_i} \sum_{i=1}^{p} \left(n'_i \overline{m}_{i,n'_i} - n_i \overline{m}_{i,n_i} \right). \tag{7}$$

3 Convergence results

3.1 Strongly convex objective

We now introduce sufficient conditions which ensures the convergence of stochastic gradient algorithms and of the PASG-algorithm when the functional G is strongly convex.

(H1) The functional G is differentiable and denoting by Φ its gradient, there exists $m \in H$ such that

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

(H2) 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.

(H3) There is a positive constant C_m such that for all $h \in H$,

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

(H4) There are positive constants L_1, L_2 such that for all $h \in H$,

$$\mathbb{E}\left[\|\nabla_h g(X, h)\|^2 \right] \leq L_1 \left(1 + \|h - m\|^2 \right),$$
$$\mathbb{E}\left[\|\nabla_h g(X, h)\|^4 \right] \leq L_2 \left(1 + \|h - m\|^4 \right).$$

(H5) The functional G is μ-strongly convex: for all $h, h' \in H$,

$$G(h) \geq G(h') + \langle \nabla G(h), h' - h \rangle + \mu \left\| h - h' \right\|^2.$$

We now make some comments on the assumptions: **(H1)** simply states the existence of a local minimum which is a necessary condition for our work. Assumptions **(H2)** to **(H5)** are smoothness properties stating that G is μ-strongly convex, coercive and have at most quadratic growth. **(H5)** is still standard and is the most favorable case when dealing with convex optimization problems, leading to the best possible achievable rates. Remark that the literature is very large on the rate of convergence of stochastic gradient algorithms in the case of strongly convex objective (see [2] among others) and one can check that under assumptions **(H1)** to **(H5)**, there are positive constants C_1, C_2 such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\| m_n - m \right\|^2\right] \leq \frac{C_1}{n^\alpha}, \tag{8}$$

$$\mathbb{E}\left[\left\| m_n - m \right\|^4\right] \leq \frac{C_2}{n^{2\alpha}}. \tag{9}$$

The following theorem gives the rates of convergence in quadratic mean of the PASG-algorithm.

Theorem 3.1. *Suppose assumptions **(H1)** to **(H5)** hold. Then, for all $n = \sum_{i=1}^p n_i$,*

$$\mathbb{E}\left[\left\| \widehat{m}_n - m \right\|^2\right] \leq \frac{L_1 \lambda_{\min}^{-2}}{\sum_{i=1}^p n_i} + \sum_{j=1}^{5} \lambda_{\min}^{-2} A_{j,p,n}^2 + \sum_{j,j'=1, j \neq j'}^{6} \lambda_{\min}^{-2} A_{j,p,n} A_{j',p,n},$$

where $\lambda_{\min} \geq \mu$ is the smallest (or limit inf for infinite dimensional spaces) eigenvalue of Γ_m and

$$A_{1,p,n}^2 = A_{2,p,n}^2 := \frac{p^2 C_1 c_\gamma^{-2}}{\left(\sum_{i=1}^p n_i\right)^2}, \qquad A_{3,p,n}^2 := \frac{4p^{2-\alpha} \alpha c_\gamma^{-2} C_1}{\left(\sum_{i=1}^p n_i\right)^{2-\alpha}}, \qquad A_{4,p,n}^2 := \frac{C_m^2 C_2 \left(1-\alpha\right)^{-2} p^{2\alpha}}{\left(\sum_{i=1}^p n_i\right)^{2\alpha}},$$

$$A_{5,p,n}^2 := \frac{L_1 C_1 \left(1-\alpha\right)^{-1} p^\alpha}{\left(\sum_{i=1}^p n_i\right)^{1+\alpha}}, \qquad A_{6,p,n}^2 := \frac{L_1}{\sum_{i=1}^p n_i}.$$

More precisely, we have

$$\mathbb{E}\left[\left\| \widehat{m}_n - m \right\|^2\right] \leq \frac{L_1 \lambda_{\min}^{-2}}{\sum_{i=1}^p n_i} + o\left(\frac{1}{\sum_{i=1}^p n_i}\right).$$

Remark:

- One can note that the rate of convergence is the optimal one for strongly convex function (see [20] for instance) and that the choice $\alpha \in (1/2, 1)$ is crucial to obtain this bound. Indeed, when $\alpha \in (0, 1/2)$, the result is different mainly because remainder terms play a preponderant role while our choice is justified by the central limit theorem. Indeed, the rate of convergence can be considered as optimal in our case since it perfectly reflects the asymptotic normality (see Theorem 3.2).

- Investigating the case $\alpha = 1$ is a tricky question discussed before, and not necessary since optimality is already obtained. However, we point out the work of ([2] and [10]) where the specificity of this case is discussed with accurate computations.
- Remark that the remainders terms are negligible since $p = o\left(n^{\max\left\{\frac{2\alpha-1}{2\alpha}, \frac{1-\alpha}{2-\alpha}\right\}}\right)$. For example, for $\alpha = \frac{2}{3}$, they are negligible since $p = o\left(\sqrt{n}\right)$.
- Among the classical examples, one can think about least-square regression, where the objective function is of the form $\mathbb{E}\left[\left(\langle X, h\rangle - Y\right)^2\right]$, for $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$ (see [8, 7] for instance).

In order to establish a Central Limit Theorem, let us now introduce a new assumption:

(H6) Let $\|.\|_F$ be the Frobenius norm for linear operators,

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

where for all $h, h', h'' \in H$, $h \otimes h'(h'') = \langle h, h''\rangle h'$.

We can now give the asymptotic normality of $(\widehat{m}_n)$.

Theorem 3.2. *Suppose assumptions **(H1)** to **(H6)** hold. Then, let $n = \sum_{i=1}^{p} n_i$,*

$$\lim_{n \to \infty} \sqrt{\sum_{i=1}^{p} n_i} \left(\widehat{m}_n - m\right) \sim \mathcal{N}\left(0, \Gamma_m^{-1} \Sigma \Gamma_m^{-1}\right),$$

where

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

3.2 Locally strongly convex objective

We now focus on the framework introduced by [13] and [14] when G is only locally strongly convex:

(H7) There exists a positive constant ϵ such that for all $h \in \mathcal{B}\left(m, \epsilon\right)$, there is a 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 $h \in \mathcal{B}\left(m, \epsilon\right)$, and for all eigenvalue λ_h of Γ_h, we have $\lambda_h \geq \frac{\lambda_{\min}}{2} > 0$.

(H8) For all integer q, there is a positive constant L_q such that for all $h \in H$,

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

The main difference with previous framework is that we just have to assume the local strong convexity of the functional we would like to minimize, and in return, we have to assume

the existence of the q-th moments of the gradient. Note that under assumptions **(H1)** to **(H4)** and **(H7)**, it was proven that for all positive constant δ,

$$\|m_n - m\|^2 = o\left(\frac{(\ln n)^\delta}{n^\alpha}\right). \tag{10}$$

Moreover, suppose assumption **(H8)** holds too, it was proven that for all positive integer p, there is a positive constant C_p such that for all $n \geq 1$,

$$\mathbb{E}\left[\|m_n - m\|^{2p}\right] \leq \frac{C_p}{n^{p\alpha}}. \tag{11}$$

Then, we can now give the rate of convergence in quadratic mean of the PASG-algorithm for locally strongly convex objectives.

Theorem 3.3. *Suppose assumptions **(H1)** to **(H3)** and **(H7)**, **(H8)** hold. Then, for all $n = \sum_{i=1}^p n_i$,*

$$\mathbb{E}\left[\|\widehat{m}_n - m\|^2\right] \leq \frac{L_1 \lambda_{\min}^{-2}}{\sum_{i=1}^p n_i} + \sum_{j=1}^{5} \lambda_{\min}^{-2} A_{j,p,n}^2 + \sum_{j,j'=1, j\neq j'}^{6} \lambda_{\min}^{-2} A_{j,p,n} A_{j',p,n},$$

where

$$A_{1,p,n}^2 = A_{2,p,n}^2 := \frac{p^2 C_1 c_\gamma^{-2}}{\left(\sum_{i=1}^p n_i\right)^2}, \qquad A_{3,p,n}^2 := \frac{4p^{2-\alpha}\alpha c_\gamma^{-2} C_1}{\left(\sum_{i=1}^p n_i\right)^{2-\alpha}}, \qquad A_{4,p,n}^2 := \frac{C_m^2 C_2 (1-\alpha)^{-2} p^{2\alpha}}{\left(\sum_{i=1}^p n_i\right)^{2\alpha}},$$

$$A_{5,p,n}^2 := \frac{L_1 C_1 (1-\alpha)^{-1} p^\alpha}{\left(\sum_{i=1}^p n_i\right)^{1+\alpha}}, \qquad A_{6,p,n}^2 := \frac{L_1}{\sum_{i=1}^p n_i}.$$

More precisely, we have

$$\mathbb{E}\left[\|\widehat{m}_n - m\|^2\right] \leq \frac{L_1 \lambda_{\min}^{-2}}{\sum_{i=1}^p n_i} + o\left(\frac{1}{\sum_{i=1}^p n_i}\right).$$

Finally, we also establish the asymptotic normality of $(\widehat{m}_n)$.

Theorem 3.4. *Suppose assumptions **(H1)** to **(H4)** and **(H6)**, **(H7)** hold. Then, let $n = \sum_{i=1}^p n_i$,*

$$\lim_{n\to\infty} \sqrt{\sum_{i=1}^{p} n_i} \left(\widehat{m}_n - m\right) \sim \mathcal{N}\left(0, \Gamma_m^{-1} \Sigma \Gamma_m^{-1}\right),$$

where

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

Remark:

- Among the classical examples, one can think about logistic regression [1], which leads to minimize $\mathbb{E}\left[\ln\left(1 + \exp\left(-Y\left\langle X, h\right\rangle\right)\right)\right]$, where $Y \in \{-1, 1\}$ and $X \in \mathbb{R}^d$. One can also think about the estimation of the geometric median (see [15, 6, 5] among others), where the objective function is $\mathbb{E}\left[\|X - h\| - \|X\|\right]$.

4 Numerical experiments

In this section, in order to illustrate our theoretical work, we propose some numerical experiments. We consider from now a Gaussian vector X taking values in $\mathbb{R}^d$, with mean $\mathbb{E}[X] = 0$ and variance $\mathbb{E}[X \otimes X] = I_d$. We focus on the estimation of the Geometric Median of X, which is defined by [15, 17]

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

Note that in this case, $m = 0$, and the SGD is defined for all $i = 1, \ldots, p$ and $k \geq 1$ by [6]

$$m_{i,k+1} = m_{i,k} + \gamma_k \frac{X_{i,k+1} - m_{i,k}}{\|X_{i,k} - m_{i,k}\|}.$$

We now consider a step sequence $\gamma_k = k^{-2/3}$ and numbers of machines equal to $1, 10, 50, 200, 500$, and we assume that data are uniformly distributed into the different machines. Note that in this case, our algorithm is the same as the one introduced by [3]. Moreover, when $p = 1$, this corresponds to the usual ASG algorithm. Finally, we also consider a case with $p = 10$ but where we have a Non-Equal Distribution ("NED" for short) between the machines, and the following vector gives the percentage of data per machine:

$$v_p = (0.05, 0.45, 1.5, 3, 8, 10, 10, 17, 20, 30)$$

In Figure 1, one can see that the number of used machines (with $p \ll \sqrt{n}$) does not seem to have a strong impact on the quadratic error of the estimates, which tends to confirm the results given by Theorems 3.1 and 3.3. Moreover, it seems to confirm that taking into account the numbers of data per machine during the parallelization step enables to take care of estimation coming from strongly different and inhomogeneous sources.

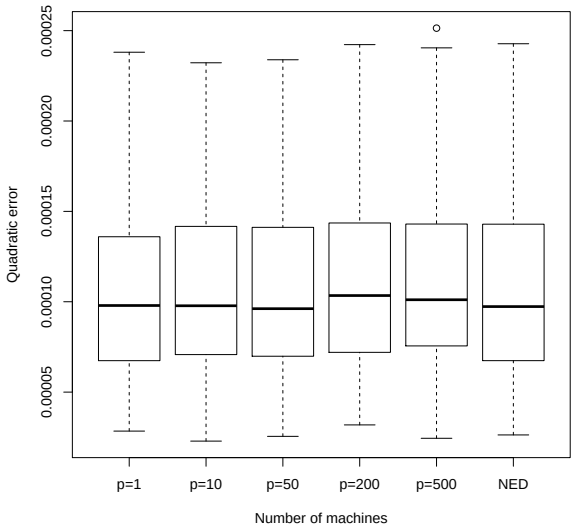

Figure 1: Quadratic error obtained with the PASGD for a sample size $n = 10^5$ and for different numbers of machines ($p = 1, 10, 50, 200, 500$) and for a Non-Equal Distribution ("NED") for $p = 10$.

Figure 2 tends to confirm Theorems 3.2 and 3.4, i.e it tends to confirm that the asymptotic behavior of the PASG algorithm does not depend on the number of machines or on the homogeneity of the data distribution.

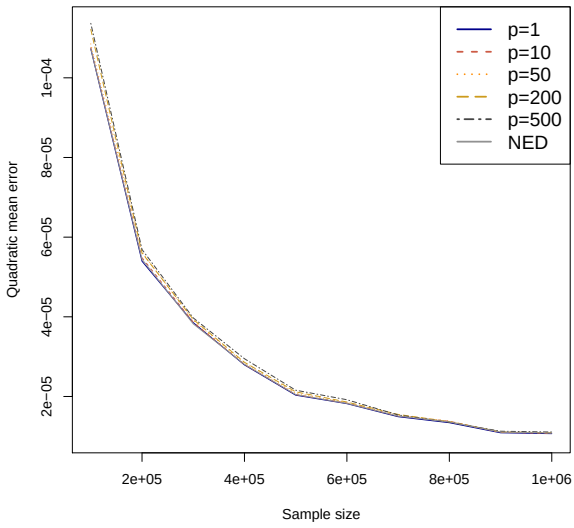

Figure 2: Evolution of the quadratic mean error compare to the sample size n for different numbers of machines ($p = 1, 10, 50, 200, 500$) and for a Non-Equal Distribution ("NED") for $p = 10$.

In a recent work, [3] proposes to parallelize ASGD but for an uniform distribution of the data between the machines, and so that without taking the number of data per machine into account. In Figure 3, we show the significant improvement represented by our algorithm compare to the previous one. Indeed, although the algorithm proposed by [3] stay quite efficient in the context of Non-Equal Distribution, it can be very less accurate than the PSGD algorithm.

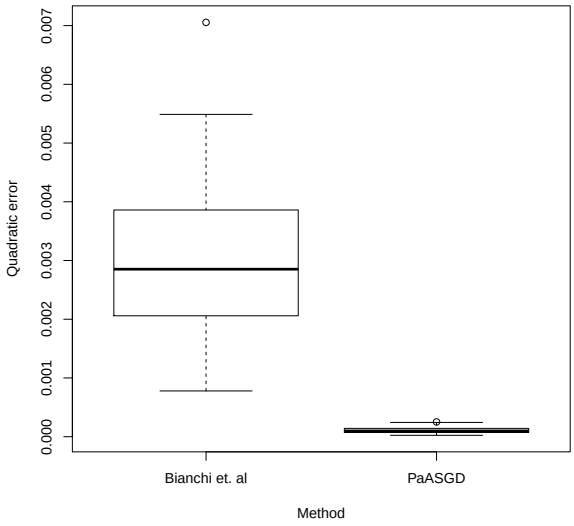

Figure 3: Comparison between the quadratic errors obtained with the PASG algorithm and the ones obtained with the algorithm introduced by Bianchi et. al, for $p = 10$ machines and with a Non-Equal Distribution.

5 Proofs

5.1 Some decompositions of the algorithms

We first recall some usual decompositions of the algorithms, which will be useful in the proofs. First, for all $k \geq 1$, let us introduce the sequences $(\xi_{i,k})$, defined, for all $i = 1, ..., p$ and for all $k \geq 1$, by $\xi_{i,k+1} := \nabla G(m_{i,k}) - \nabla_h g(X_{i,k+1}, m_{i,k})$. Moreover, let us introduce the sequences of σ-algebras $(\mathcal{F}_{1,k}), ..., (\mathcal{F}_{p,k})$ defined for all $i = 1, ..., p$ and $k \geq 1$ by $\mathcal{F}_{i,k} := \sigma(X_{i,1}, ..., X_{i,k})$. Then, for all $i = 1, ..., p$, $(\xi_{i,k})_k$ is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_{i,k})_k$. Moreover, the SGD can be written, for all $i = 1, ..., p$ and for all $k \geq 1$, as

$$m_{i,k+1} = m_{i,k} - \gamma_k \nabla G(m_{i,k}) + \gamma_k \xi_{i,k+1}. \tag{12}$$

Moreover, linearizing the gradient, the SGD can be decomposed as

$$m_{i,k+1} - m = (I_H - \gamma_k \Gamma_m)(m_{i,k} - m) + \gamma_k \xi_{i,k+1} - \gamma_k \delta_{i,k}, \tag{13}$$

where $\delta_{i,k} := \nabla G(m_{i,k}) - \Gamma_m(m_{i,k} - m)$ is the remainder term in the Taylor's expansion of the gradient. Finally, summing these equalities, applying an Abel's transform and dividing by n_i, one can obtain (see [22])

$$\Gamma_m[\overline{m}_{i,n_i} - m] = \frac{m_{i,1} - m}{n_i\gamma_1} - \frac{m_{i,n_i+1} - m}{n_i\gamma_{n_i}} + \frac{1}{n_i}\sum_{k=2}^{n_i}(m_{i,k} - m)\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right) - \frac{1}{n_i}\sum_{k=1}^{n_i}\delta_{i,k} + \frac{1}{n_i}\sum_{k=1}^{n}\xi_{i,k+1}. \tag{14}$$

Finally, by linearity, the PASG algorithm can be written as

$$\Gamma_m[\widehat{m}_n - m] = \frac{1}{\sum_{i=1}^p n_i}\sum_{i=1}^p \frac{m_{i,1} - m}{\gamma_1} - \frac{1}{\sum_{i=1}^p n_i}\sum_{i=1}^p \frac{m_{i,n_i+1} - m}{\gamma_{n_i}} + \frac{1}{\sum_{i=1}^p n_i}\sum_{i=1}^p\sum_{k=2}^{n_i}(m_{i,k} - m)\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right) - \frac{1}{\sum_{i=1}^p n_i}\sum_{i=1}^p\sum_{k=1}^{n_i}\delta_{i,k} + \frac{1}{\sum_{i=1}^p n_i}\sum_{i=1}^p\sum_{k=1}^{n}\xi_{i,k+1}. \tag{15}$$

5.2 Proof of Theorems 3.1 and 3.3

In order to prove Theorems 3.1 and 3.3, we just have to bound each term on the right-hand side of (15).

Bounding $\mathbb{E}\left[\left\|\frac{1}{\sum_{i=1}^p n_i}\sum_{i=1}^p \frac{m_{i,1}-m}{\gamma_1}\right\|^2\right]$. With the help of Cauchy-Schwarz inequality and of inequality (11) or (8), one can check that

$$\mathbb{E}\left[\left\|\frac{1}{\sum_{i=1}^p n_i}\sum_{i=1}^p \frac{m_{i,1} - m}{\gamma_1}\right\|^2\right] \leq \frac{p}{\left(\sum_{i=1}^p n_i\right)^2}\sum_{i=1}^p \mathbb{E}\left[\left\|\frac{m_{i,1} - m}{\gamma_1}\right\|^2\right] \leq \frac{p^2 C_1 c_\gamma^{-2}}{\left(\sum_{i=1}^p n_i\right)^2} := A_{1,p}^2. \tag{16}$$

Bounding $\mathbb{E}\left[\left\|\frac{1}{\sum_{i=1}^p n_i}\sum_{i=1}^p \frac{m_{i,n_i+1}-m}{\gamma_{n_i}}\right\|^2\right]$. In the same way, with the help of Cauchy-Schwarz inequality and of inequality (11) or (8),

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{\sum_{i=1}^p n_i}\sum_{i=1}^p \frac{m_{i,n_i+1} - m}{\gamma_{n_i}}\right\|^2\right] &\leq \frac{p}{\left(\sum_{i=1}^p n_i\right)^2}\sum_{i=1}^p \mathbb{E}\left[\left\|\frac{m_{i,n_i+1} - m}{\gamma_{n_i}}\right\|^2\right] \\ &\leq \frac{p}{\left(\sum_{i=1}^p n_i\right)^2}\sum_{i=1}^p C_1 c_\gamma^{-2}\left(\frac{n_i}{n_i+1}\right)^\alpha.\end{aligned}$$

This yields,

$$\mathbb{E}\left[\left\|\frac{1}{\sum_{i=1}^p n_i}\sum_{i=1}^p \frac{m_{i,n_i+1} - m}{\gamma_{n_i}}\right\|^2\right] \leq \frac{p^2 c_\gamma^{-2} C_1}{\left(\sum_{i=1}^p n_i\right)^2} := A_{2,p}^2. \tag{17}$$

Bounding $\frac{1}{\sum_{i=1}^p n_i}\sum_{i=1}^p\sum_{k=2}^{n_i}(m_{i,k} - m)\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)$. In the same way, applying Lemma 4.3

in [12],

$$\mathbb{E}\left[\left\|\frac{1}{\sum_{i=1}^{p} n_i}\sum_{i=1}^{p}\sum_{k=2}^{n_i}(m_{i,k}-m)\left(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}\right)\right\|^2\right] \leq \frac{p}{\left(\sum_{i=1}^{p} n_i\right)^2}\sum_{i=1}^{p}\mathbb{E}\left[\left\|\sum_{k=2}^{n_i}(m_{i,k}-m)\left(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}\right)\right\|^2\right]$$
$$\leq \frac{p}{\left(\sum_{i=1}^{p} n_i\right)^2}\sum_{i=1}^{p}\left(\sum_{k=2}^{n_i}\sqrt{\mathbb{E}\left[\|m_{i,k}-m\|^2\right]}\left|\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}\right|\right)^2.$$

Since

$$\left|\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}\right| \leq \alpha\, c_\gamma^{-1}(k-1)^{\alpha-1},$$

and applying inequality (11) or (8)

$$\mathbb{E}\left[\left\|\frac{1}{\sum_{i=1}^{p} n_i}\sum_{i=1}^{p}\sum_{k=2}^{n_i}(m_{i,k}-m)\left(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}\right)\right\|^2\right] \leq \frac{p\alpha^2 c_\gamma^{-2}C_1}{\left(\sum_{i=1}^{p} n_i\right)^2}\sum_{i=1}^{p}\left(\sum_{k=2}^{n_i}\frac{(k-1)^{\alpha-1}}{k^{\alpha/2}}\right)^2$$
$$\leq \frac{p\alpha^2 c_\gamma^{-2}C_1}{\left(\sum_{i=1}^{p} n_i\right)^2}\sum_{i=1}^{p}\left(\sum_{k=2}^{n_i}\frac{(k-1)^{\alpha-1}}{(k-1)^{\alpha/2}}\right)^2$$
$$\leq \frac{p\alpha^2 c_\gamma^{-2}C_1}{\left(\sum_{i=1}^{p} n_i\right)^2}\sum_{i=1}^{p}\left(\sum_{k=2}^{n_i}(k-1)^{\alpha/2-1}\right)^2.$$

Then, since $\alpha < 1$, with the help of an integral test for convergence and thanks to Hölder's inequality,

$$\mathbb{E}\left[\left\|\frac{1}{\sum_{i=1}^{p} n_i}\sum_{i=1}^{p}\sum_{k=2}^{n_i}(m_{i,k}-m)\left(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}\right)\right\|^2\right] \leq \frac{4p\alpha c_\gamma^{-2}C_1}{\left(\sum_{i=1}^{p} n_i\right)^2}\sum_{i=1}^{p}(n_i-1)^\alpha$$
$$\leq \frac{4p^{2-\alpha}\alpha c_\gamma^{-2}C_1}{\left(\sum_{i=1}^{p} n_i\right)^2}\left(\sum_{i=1}^{p}(n_i-1)\right)^\alpha$$
$$\leq \frac{4p^{2-\alpha}\alpha c_\gamma^{-2}C_1}{\left(\sum_{i=1}^{p} n_i\right)^{2-\alpha}} := A_{3,p}^2. \tag{18}$$

Bounding $\mathbb{E}\left[\left\|\frac{1}{\sum_{i=1}^{p} n_i}\sum_{i=1}^{p}\sum_{k=1}^{n_i}\delta_{i,k}\right\|^2\right]$**.** First, let us recall that there is a positive constant C_m such that for all $i = 1, ..., p$, and for all integer k,

$$\|\delta_{i,k}\| \leq C_m\,\|m_{i,k}-m\|^2. \tag{19}$$

Moreover, thanks to Lemma 4.1 in [12],

$$\mathbb{E}\left[\left\|\frac{1}{\sum_{i=1}^{p} n_i}\sum_{i=1}^{p}\sum_{k=1}^{n_i}\delta_{i,k}\right\|^2\right] \leq \frac{p}{\left(\sum_{i=1}^{p} n_i\right)^2}\sum_{i=1}^{p}\mathbb{E}\left[\left\|\sum_{k=1}^{n_i}\delta_{i,k}\right\|^2\right] \leq \frac{p}{\left(\sum_{i=1}^{p} n_i\right)^2}\sum_{i=1}^{p}\left(\sum_{k=1}^{n_i}\sqrt{\mathbb{E}\left[\|\delta_{i,k}\|^2\right]}\right)^2.$$

Then, applying inequalities (19) and (11) or (9),

$$\mathbb{E}\left[\left\|\frac{1}{\sum_{i=1}^{p} n_i}\sum_{i=1}^{p}\sum_{k=1}^{n_i}\delta_{i,k}\right\|^2\right] \leq \frac{C_m^2 p}{\left(\sum_{i=1}^{p} n_i\right)^2}\sum_{i=1}^{p}\left(\sum_{k=1}^{n_i}\sqrt{\mathbb{E}\left[\|m_{i,k}-m\|^4\right]}\right)^2 \leq \frac{C_m^2 C_2 p}{\left(\sum_{i=1}^{p} n_i\right)^2}\sum_{i=1}^{p}\left(\sum_{k=1}^{n_i} k^{-\alpha}\right)^2.$$

Thus, since $1/2 < \alpha < 1$, with the help of an integral test for convergence and thanks to Hölder's inequality,

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{\sum_{i=1}^{p} n_i}\sum_{i=1}^{p}\sum_{k=1}^{n_i}\delta_{i,k}\right\|^2\right] &\leq \frac{C_m^2 C_2 (1-\alpha)^{-2} p}{\left(\sum_{i=1}^{p} n_i\right)^2}\sum_{i=1}^{p} n_i^{2-2\alpha} \leq \frac{C_m^2 C_2 (1-\alpha)^{-2} p^{2\alpha}}{\left(\sum_{i=1}^{p} n_i\right)^2}\left(\sum_{i=1}^{p} n_i\right)^{2-2\alpha} \\ &\leq \frac{C_m^2 C_2 (1-\alpha)^{-2} p^{2\alpha}}{\left(\sum_{i=1}^{p} n_i\right)^{2\alpha}} := A_{4,p}^2. \end{aligned} \tag{20}$$

Bounding $\mathbb{E}\left[\left\|\frac{1}{\sum_{i=1}^{p} n_i}\sum_{i=1}^{p}\sum_{k=1}^{n}\xi_{i,k+1}\right\|^2\right]$. First, by definition of the sequences $(\xi_{i,k})$ and thanks to assumption **(H4)** or **(H8)**, for all $i = 1, ..., p$ and for all positive integer k,

$$\begin{aligned}\mathbb{E}\left[\|\xi_{i,k+1}\|^2\right] &= \mathbb{E}\left[\|\nabla_h g(X_{i,+1}, m_{i,k})\|^2\right] - \mathbb{E}\left[\|\nabla G(m_{i,k})\|^2\right] \\ &\leq L_1 + L_1\mathbb{E}\left[\|m_{i,k}-m\|^2\right]. \end{aligned} \tag{21}$$

Moreover, since for all $i = 1, ..., p$, $(\xi_{i,k})$ is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_{i,k})$ and since for all $i = 1, ..., p$ and $j = 1, ..., p$ such taht $i \neq j$ the sequences $(\xi_{i,k})$ and $(\xi_{j,k})$ are independent,

$$\mathbb{E}\left[\sum_{k=1}^{n_i}\xi_{i,k}\right] = 0 \quad \forall i = 1, ..., p, \quad \text{and} \quad \mathbb{E}\left[\left\langle\sum_{k=1}^{n_i}\xi_{i,k}, \sum_{k=1}^{n_j}\xi_{j,k}\right\rangle\right] = 0, \quad \forall i, j = 1, ..., p \quad \text{s.t} \quad i \neq j.$$

Then, applying inequality (21),

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{\sum_{i=1}^{p} n_i}\sum_{i=1}^{p}\sum_{k=1}^{n_i}\xi_{i,k+1}\right\|^2\right] &= \frac{1}{\left(\sum_{i=1}^{p} n_i\right)^2}\sum_{i=1}^{p}\mathbb{E}\left[\left\|\sum_{k=1}^{n_i}\xi_{i,k+1}\right\|^2\right] \\ &= \frac{1}{\left(\sum_{i=1}^{p} n_i\right)^2}\sum_{i=1}^{p}\sum_{k=1}^{n_i}\mathbb{E}\left[\|\xi_{i,k+1}\|^2\right] \\ &\leq \frac{1}{\left(\sum_{i=1}^{p} n_i\right)^2}\sum_{i=1}^{p}\sum_{k=1}^{n_i}\left(L_1 + L_1\mathbb{E}\left[\|m_{i,k}-m\|^2\right]\right)\end{aligned}$$

Then, thanks to inequality (11) or (8), and with the help of an integral test for convergence,

$$\mathbb{E}\left[\left\|\frac{1}{\sum_{i=1}^{p} n_i}\sum_{i=1}^{p}\sum_{k=1}^{n_i}\xi_{i,k+1}\right\|^2\right] \leq \frac{L_1}{\sum_{i=1}^{p} n_i} + \frac{L_1 C_1}{\left(\sum_{i=1}^{p} n_i\right)^2}\sum_{i=1}^{p}\sum_{k=1}^{n_i} k^{-\alpha}$$
$$\leq \frac{L_1}{\sum_{i=1}^{p} n_i} + \frac{L_1 C_1 (1-\alpha)^{-1}}{\left(\sum_{i=1}^{p} n_i\right)^2}\sum_{i=1}^{p} n_i^{1-\alpha}.$$

Finally, applying Hölder's inequality and since $\alpha < 1$,

$$\mathbb{E}\left[\left\|\frac{1}{\sum_{i=1}^{p} n_i}\sum_{i=1}^{p}\sum_{k=1}^{n_i}\xi_{i,k+1}\right\|^2\right] \leq \frac{L_1}{\sum_{i=1}^{p} n_i} + \frac{L_1 C_1 (1-\alpha)^{-1} p^{\alpha}}{\left(\sum_{i=1}^{p} n_i\right)^{1+\alpha}} := A_{6,p}^2 + A_{5,p}^2. \tag{22}$$

Conclusion. Since the smallest eigenvalue (or the limit inf of the eignevalues for infinite dimensional spaces) of Γ_m denoted by $\lambda_{\min}$ is positive, let $n = \sum_{i=1}^{p} n_i$,

$$\mathbb{E}\left[\|\widehat{m}_n - m\|^2\right] \leq \frac{1}{\lambda_{\min}^2}\mathbb{E}\left[\|\Gamma_m(\widehat{m}_n - m)\|^2\right].$$

Then, applying Cauchy-Schwarz's inequality as well as inequalities (16) to (22), one can check that

$$\mathbb{E}\left[\|\widehat{m}_n - m\|^2\right] \leq \sum_{j=1}^{6}\lambda_{\min}^{-2} A_{j,p,n}^2 + \sum_{j,j'=1,\ j\neq j'}^{6}\lambda_{\min}^{-2} A_{j,p,n} A_{j',p,n},$$

which concludes the proof.

5.3 Proof of Theorem 3.2 and 3.4

First, one can check that the first term on the right-hand side of equality (15) are negligeable, i.e

$$\frac{1}{\sqrt{\sum_{i=1}^{p} n_i}}\sum_{i=1}^{p}\frac{m_{i,1} - m}{\gamma_1} = o(1) \qquad \mathbb{P},$$
$$\frac{1}{\sqrt{\sum_{i=1}^{p} n_i}}\sum_{i=1}^{p}\frac{m_{i,n_i+1} - m}{\gamma_{n_i}} = o(1) \qquad \mathbb{P},$$
$$\frac{1}{\sqrt{\sum_{i=1}^{p} n_i}}\sum_{i=1}^{p}\sum_{k=2}^{n_i}(m_{i,k} - m)\left(\frac{1}{\gamma_{k+1}} - \frac{1}{\gamma_k}\right) = o(1) \qquad \mathbb{P},$$
$$\frac{1}{\sqrt{\sum_{i=1}^{p} n_i}}\sum_{i=1}^{p}\sum_{k=1}^{n_i}\delta_{i,k} = o(1) \qquad \mathbb{P}.$$

Indeed, it is a direct application of inequalities (16) to (22) when assumptions **(H1)** to **(H5)** are verified and a direct application of Theorem 4.1 in [13] when assumptions **(H1)** to **(H4)** and **(H7)** are verified. In order to get the asymptotic normality of the term $\left(\sum_{i=1}^{p}\sum_{k=1}^{n_i}\xi_{i,k+1}\right)$,

let us first remark that with a good choice of index, this term can be seen as a sum of martingale differences term. Then, we just have to check that assumptions of Theorem 5.1 in [16] are fulfilled, i.e let $(e_j)_{j\in J}$ be an orthonormal basis of H and $\psi_{j,j'} := \langle \Sigma e_j, e_{j'} \rangle$ for all $j, j' \in J$, we have to verify:

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

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

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

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

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

Moreover, note that thanks to Assumption **(H4)** or **(H8)**, for all $i = 1, ..., p$ and $1 \le k \le n_i$,

$$\mathbb{E}\left[\|\xi_{i,k+1}\|^4 \right] \le 2^4 \mathbb{E}\left[\|\nabla_h g(X_{i,k+1}, m_{i,k})\|^4 \right] \le 2^4 L_2 \left(1 + \mathbb{E}\left[\|m_{i,k}\|^4 \right] \right) \le 2^4 L_2 (1 + C_2).$$

Then,

$$\mathbb{P}\left(\sup_{1\le k\le n_i, i=1,\dots p} \frac{1}{\sqrt{\sum_{i=1}^p n_i}} \|\xi_{i,k+1}\| > \eta \right) \le \frac{\eta^{-4} 2^4 L_2 (1 + C_2)}{\sum_{i=1}^p n_i}.$$

Proof of (24): First, let $\otimes$ be the bilinear application defined for all $h, h', h'' \in H$ by $(h \otimes h')(h'') = \langle h, h'' \rangle h'$. Note that

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

with $\epsilon_{i,k+1} := \xi_{i,k+1} \otimes \xi_{i,k+1} - \mathbb{E}\left[\xi_{i,k+1} \otimes \xi_{i,k+1} | \mathcal{F}_{i,k} \right]$. Remark that we a good choice of index, $(\epsilon_{i,k})_{i,k}$ can be seen as a sequence of martingale differences, and one can check that

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

Let us now prove that the sequence of operators $\left(\mathbb{E}\left[\xi_{i,k+1} \otimes \xi_{i,k+1} | \mathcal{F}_{i,k} \right] \right)_{i,k}$ converges almost

surely to Σ with respect to the Frobenius norm when k goes to infinity. First, note that

$$\begin{aligned}
&\left\|\mathbb{E}\left[\xi_{i,k+1}\otimes\xi_{i,k+1}|\mathcal{F}_k\right]-\Sigma\right\|_F\\
&=\left\|\mathbb{E}\left[\nabla_h g\left(X_{i,k+1},m_{i,k}\right)\otimes\nabla_h g\left(X_{i,k+1},m_{i,k}\right)|\mathcal{F}_{i,k}\right]-\Sigma-\nabla G\left(m_{i,k}\right)\otimes\nabla G\left(m_{i,k}\right)\right\|_F\\
&\leq\left\|\mathbb{E}\left[\nabla_h g\left(X_{i,k+1},m_{i,k}\right)\otimes\nabla_h g\left(X_{i,k+1},m_{i,k}\right)|\mathcal{F}_{i,k}\right]-\Sigma\right\|_F+\left\|\nabla G\left(m_{i,k}\right)\otimes\nabla G\left(m_{i,k}\right)\right\|_F.
\end{aligned}$$

Then, thanks to Assumption **(H6)**, since $\|\nabla G\left(m_{i,k}\right)\|\leq C\left\|m_{i,k}-m\right\|$ for all i,k (see [13]), and since $(m_{i,k})$ converges almost surely to m,

$$\begin{aligned}
&\lim_{k\to\infty}\left\|\mathbb{E}\left[\nabla_h g\left(X_{i,k+1},m_{i,k}\right)\otimes\nabla_h g\left(X_{i,k+1},m_{i,k}\right)|\mathcal{F}_{i,k}\right]-\Sigma\right\|_F=0\quad a.s,\quad\forall i=1,...,p,\\
&\lim_{k\to\infty}\left\|\nabla G\left(m_{i,k}\right)\otimes\nabla G\left(m_{i,k}\right)\right\|_F=\lim_{k\to\infty}\left\|\nabla G\left(m_{i,k}\right)\right\|_F^2=0\quad a.s,\quad\forall i=1,...,p.
\end{aligned}$$

Then, the sequences $\left(\mathbb{E}\left[\xi_{i,k+1}\otimes\xi_{i,k+1}|\mathcal{F}_k\right]\right)_{k\geq1}$ converges almost surely to Σ with respect to the Frobenius norm and as a consequence, for all $j,j'\in J$,

$$\lim_{k\to\infty}\left\langle\mathbb{E}\left(\xi_{i,k+1}\otimes\xi_{i,k+1}|\mathcal{F}_{i,k}\right]\left(e_j\right),e_{j'}\right\rangle=\psi_{j,j'}\quad a.s,\quad\forall i=1,...,p.$$

Thus, applying Toeplitz's lemma, for all $j,j'\in J$,

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

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

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

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

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

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

$$\limsup_n\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^p\sum_{k=1}^{n_i}\sum_{j=N}^{\infty}\left\langle\xi_{i,k+1},e_j\right\rangle>\epsilon\right)\leq\frac{1}{\epsilon^2}\sum_{j=N}^{\infty}\left\langle\Sigma(e_j),e_j\right\rangle,$$

and one can conclude as in [14].

References

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

[3] Pascal Bianchi, Gersende Fort, and Walid Hachem. Performance of a distributed stochastic approximation algorithm. *IEEE Transactions on Information Theory*, 59(11):7405–7418, 2013.

[4] Gérard Biau and Ryad Zenine. Online asynchronous distributed regression. *arXiv:1407.4373*, 2014.

[5] Hervé Cardot, Peggy Cénac, Antoine Godichon-Baggioni, et al. Online estimation of the geometric median in hilbert spaces: Nonasymptotic confidence balls. *The Annals of Statistics*, 45(2):591–614, 2017.

[6] Hervé Cardot, Peggy Cénac, and Pierre-André Zitt. Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19(1):18–43, 2013.

[7] Kobi Cohen, Angelia Nedic, and R Srikant. On projected stochastic gradient descent algorithm with weighted averaging for least squares regression. *IEEE Transactions on Automatic Control*, 2017.

[8] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *arXiv preprint arXiv:1602.05419*, 2016.

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

[10] S. Gadat, F. Panloup, and S. Saadane. Stochastic heavy ball. 2016.

[11] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[12] Antoine Godichon-Baggioni. 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, 2016.

[13] Antoine 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.

[14] Antoine Godichon-Baggioni. A central limit theorem for averaged stochastic gradient algorithms in hilbert spaces and online estimation of the asymptotic variance. application to the geometric median and quantiles. *arXiv preprint arXiv:1702.00931*, 2017.

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

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

[17] Johannes Kemperman. 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, 1987.

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

[19] Ji Liu, Stephen J Wright, Christopher Ré, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. *Journal of Machine Learning Research*, 16(285-322):1–5, 2015.

[20] A.S. Nemirovsky and D.B. Yudin. Problem complexity and method efficiency in optimization. 1983.

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

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

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

[24] Benjamin Recht and Christopher Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

[25] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

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

[27] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010.