Lp and almost sure rates of convergence of averaged stochastic gradient algorithms: locally strongly convex objective

Antoine Godichon-Baggioni

▶ **To cite this version:**

Antoine Godichon-Baggioni. Lp and almost sure rates of convergence of averaged stochastic gradient algorithms: locally strongly convex objective. ESAIM: Probability and Statistics, 2019, 23, pp.841-873. 10.1051/ps/2019011 . hal-01678852

HAL Id: hal-01678852
https://hal.science/hal-01678852

Submitted on 30 Jan 2024

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

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

ESAIM: PS 23 (2019) 841–873
https://doi.org/10.1051/ps/2019011

ESAIM: Probability and Statistics
www.esaim-ps.org

L^p AND ALMOST SURE RATES OF CONVERGENCE OF AVERAGED STOCHASTIC GRADIENT ALGORITHMS: LOCALLY STRONGLY CONVEX OBJECTIVE

Antoine Godichon-Baggioni*

Abstract. An usual problem in statistics consists in estimating the minimizer of a convex function. When we have to deal with large samples taking values in high dimensional spaces, stochastic gradient algorithms and their averaged versions are efficient candidates. Indeed, (1) they do not need too much computational efforts, (2) they do not need to store all the data, which is crucial when we deal with big data, (3) they allow to simply update the estimates, which is important when data arrive sequentially. The aim of this work is to give asymptotic and non asymptotic rates of convergence of stochastic gradient estimates as well as of their averaged versions when the function we would like to minimize is only locally strongly convex.

Mathematics Subject Classification. 62L12, 62G35, 62L20.

Received September 7, 2018. Accepted June 14, 2019.

1. Introduction

With the development of automatic sensors, it is more and more important to think about methods able to deal with large samples of observations taking values in high dimensional spaces such as functional spaces. We focus here on an usual stochastic optimization problem which consists in estimating

$$m := \arg\min_{h \in H} \mathbb{E}\left[g(X,h)\right], \tag{1.1}$$

where H is a Hilbert space and X is a random variable supposed to be taking value in a space $\mathcal{X}$ and $g : \mathcal{X} \times H \longrightarrow \mathbb{R}$. One usual method, given a sample $X_1, \ldots, X_n$, is to consider the empirical problem generated by this sample, *i.e.* to consider the M-estimates (see the books of [18, 21] among others)

$$\widehat{m}_n := \arg\min_{h \in H} \sum_{k=1}^{n} g\left(X_k, h\right),$$

and to approximate $\widehat{m}_n$ using deterministic optimization methods (see [3] for instance). Nevertheless, one of the most important problem of such methods is that they become computationally expensive when we deal with

Keywords and phrases: Stochastic optimization, Stochastic gradient algorithm, averaging, Robust statistics.

Institut de Mathématiques de Toulouse, Université Paul Sabatier, Toulouse, France.

* Corresponding author: antoine.godichon_baggioni@upmc.fr

© The author. Published by EDP Sciences, SMAI 2019

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

large samples taking values in high dimensional spaces. Thus, in order to overcome this, stochastic gradient algorithms introduced by [27] are efficient candidates. Indeed, they do not need too much computational efforts, do not require to store all the data and can be simply updated, which represents a real interest when the data arrive sequentially.

The literature is very large on this domain (see the books of [12, 20] among others) and on the method to improve their convergence which consists in averaging the Robbins-Monro estimates, which was introduced by [29] and whose first convergence results were given by [26]. Many asymptotic results exist in the literature when data lies in finite dimensional spaces (see [12, 23, 24] for instance) but the proofs can not be directly adapted for infinite dimensional spaces. Moreover, an asymptotic result such as a Central Limit Theorem does not give any clue of how far the distribution of the estimate is from its asymptotic law for a fixed sample size n. Then, non asymptotic properties are always desirable for statisticians who deal with real data (see the nice arguments of [28] for example). As a consequence, these last few years, statisticans have more and more focused on non asymptotic rates of convergence. For example, [22] and [1] give some general conditions to get the rate of convergence in quadratic mean of averaged stochastic gradient algorithms, while [13], for instance, focus on non asymptotic rates for strongly convex stochastic composite optimization.

The aim of this work is to seek inspiration in the demonstration methods introduced by [6] and improved by [4, 14] to give convergence results for stochastic gradient algorithms and their averaged versions when the function we would like to minimize is only locally strongly convex. First, we establish almost sure rates of convergence of the estimates in general Hilbert spaces. Furthermore, as mentioned above, asymptotic results are often non sufficient, and L^p rates of convergence of the algorithms are so given.

The paper is organized as follows. Section 2 introduces the framework, assumptions, the algorithms and some convexity properties on the function we would like to minimize. Two examples of application are given in Section 3: we first focus on the estimation of geometric quantiles, which are a generalization of the real quantiles introduced by [8]. They are robust indicators which can be useful in statistical depth and outliers detection (see [30], [9] or [17]). In a second time, we focus on the estimation of generalized p-means [2, 25], used in several domains such that computer vision [31] or medical imaging [15]. In a third time, stochastic gradient algorithms can be applied in several regressions [1, 10] and we focus on robust logistic regression. In Section 4, the almost sure and L^p rates of convergence of the estimates are given. Our theoretical results are illustrated by numerical experiments in Section 5. Finally, the proofs are postponed in Section 6 and in Appendix.

2. The algorithms and assumptions

2.1. Assumptions and general framework

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

$$G(h) := \mathbb{E}\left[g(X, h)\right], \tag{2.1}$$

where $g : \mathcal{X} \times H \longrightarrow \mathbb{R}$. Moreover, let us suppose that the functional G is convex. We consider from now that the following assumptions are fulfilled:

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

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

(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 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 $h \in \mathcal{B}(m, \epsilon)$, 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) Let $f : \mathcal{X} \times H \longrightarrow \mathbb{R}_+$ and let C be a positive constant such that for almost every $x \in \mathcal{X}$ and for all $h \in H$, $\|\nabla_h g(x, h)\| \leq f(x, h) + C\|h - m\|$ almost surely.

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

$$\mathbb{E}\left[f(X, h)^2\right] \leq L_1.$$

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

$$\mathbb{E}\left[f(X, h)^{2q}\right] \leq L_q.$$

Note that for the sake of simplicity, we often denote by the same way the different constants. We now make some comments on the assumptions. First, note that no convexity assumption on the functional g is required.

Assumptions **(A2)** and **(A3)** give some properties on the spectrum of the Hessian and ensure that the functional G is locally strongly convex. Note that assumption **(A3)** can be resumed as $\lambda_{\min}(\Gamma_m) > 0$, where $\lambda_{\min}(.)$ is the function which gives the smallest eigenvalue (or the lim inf of the eigenvalues in infinite dimensional spaces) of a linear operator, if the functional $h \mapsto \lambda_{\min}(\Gamma_h)$ is continuous on a neighborhood of m.

Moreover, assumption **(A4)** allows to bound the remainder term in the Taylor's expansion of the gradient. Note that since the functional G is twice continuously differentiable and since $\Phi(m) = 0$, it comes $\Phi(h) = \int_0^1 \Gamma_{m+t(h-m)}(h-m)\mathrm{d}t$, and in a particular case, $\Phi(h) - \Gamma_m(h-m) = \int_0^1 \left(\Gamma_{m+t(h-m)}(h-m) - \Gamma_m(h-m)\right)\mathrm{d}t$. Thus, assumption **(A4)** can be verified by giving a neighborhood of m for each there is a positive constant C_ϵ such for all h in this neighborhood, if we consider the functional $\varphi_h : [0,1] \longrightarrow H$ defined for all $t \in [0,1]$ by $\varphi_h(t) := \Gamma_{m+t(h-m)}(h - m)$, then for all $t \in [0,1]$,

$$\|\varphi_h'(t)\| \leq C_\epsilon \|h - m\|^2.$$

Assumption **(A5)** enables us to bound the gradient under conditions on the functional f. More precisely, **(A5a)** is sufficient to get the almost sure rates of convergence while we need to assume **(A5b)** to obtain the L^p rates of convergence. This still represents a significant relaxation of the usual conditions needed to get non asymptotic results. For example, a main difference with [1] and [14] is that, instead of having a bounded gradient, we split this bound into two parts: one which admits qth moments, and one which depends on the estimation error. Moreover, note that it is possible to replace assumption **(A5)** by

(A5a′) There is a positive constant L^1 such that for all $h \in H$,

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

(A5b′) For all integer q, there is a positive constant L_q such that for all $h \in H$,

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

Remark 2.1. These assumptions are analogous to the usual ones in finite dimension ([23], [24]) but in our case, the proofs remain true in infinite dimension.

Remark 2.2. Note that the Hessian of the functional G is not supposed to be compact. Then, if $H = \mathbb{R}^d$, its smallest eigenvalue $\lambda_{\min}\left(\Gamma_m\right)$ does not necessarily converge to 0 when the dimension d tends to infinity.

2.2. The algorithms

Let $X_1, \ldots, X_n, \ldots$ be independent random variables with the same law as X. The stochastic gradient algorithm is defined recursively by

$$\begin{aligned} Z_{n+1} &= Z_n - \gamma_n \nabla_h g\left(X_{n+1}, Z_n\right) \\ &=: Z_n - \gamma_n U_{n+1}, \end{aligned} \tag{2.2}$$

where Z_1 is chosen bounded and $U_{n+1} := \nabla_h g\left(X_{n+1}, Z_n\right)$. The step sequence (γ_n) is a decreasing sequence of positive real numbers which verifies the following usual assumptions (see [12])

$$\sum_{n\geq 1} \gamma_n = \infty, \qquad \sum_{n\geq 1} \gamma_n^2 < \infty.$$

The term U_{n+1} can be considered as a random perturbation of the gradient Φ at Z_n. Indeed, let $(\mathcal{F}_n)$ be the sequence of σ-algebra defined for all $n \geq 1$ by $\mathcal{F}_n := \sigma\left(X_1, \ldots, X_n\right) = \sigma\left(Z_1, \ldots, Z_n\right)$, then

$$\mathbb{E}\left[U_{n+1} | \mathcal{F}_n\right] = \nabla G(Z_n) =: \Phi\left(Z_n\right).$$

In order to improve the convergence, we now introduce the averaged algorithm ([29], [26]) defined recursively by

$$\overline{Z}_{n+1} = \overline{Z}_n + \frac{1}{n+1}\left(Z_{n+1} - \overline{Z}_n\right), \tag{2.3}$$

with $\overline{Z}_1 = Z_1$. This can also be written as follows

$$\overline{Z}_n = \frac{1}{n}\sum_{k=1}^{n} Z_k.$$

2.3. Some convexity properties

We now give some convexity properties of the functional G. The proofs are given in Appendix. First, since $\nabla G(m) = 0$ and since G is twice continuously differentiable, note that for all $h \in H$,

$$\nabla G(h) = \nabla G(h) - \nabla G(m) = \int_0^1 \Gamma_{m+t(h-m)}(h-m)\mathrm{d}t.$$

The first proposition gives the local strong convexity of the functional G.

Proposition 2.3. *Assume **(A1)** to **(A3)** and **(A5a)** hold. For all positive constant A and for all $h \in \mathcal{B}(m, A)$,*

$$\langle \nabla G(h), h - m \rangle \geq c_A \|h - m\|^2,$$

with $c_A := \min\left\{\frac{\lambda_{\min}}{2}, \frac{\lambda_{\min}\epsilon}{2A}\right\}$. Moreover, there is a positive constant C such that for all $h \in H$,

$$|\langle \nabla G(h), h - m \rangle| \leq C \|h - m\|^2.$$

*This result remains true replacing assumption **(A5a)** by **(A5a′)**.*

The following corollary ensures that m is the unique solution of the problem defined by (1.1).

Corollary 2.4. *Assume **(A1)** to **(A3)** and **(A5a)** hold. Then, m is the unique solution of the equation*

$$\nabla G(h) = 0,$$

and in a particular case, m is the unique minimizer of the functional G.

Remark 2.5. Assumption **(A3)** and Proposition 2.3 enable us to invert the Hessian at m and to have a control on the "loss" of strong convexity. More precisely, assumption **(A3)** could be replaced by

(A3′) There is a basis composed of eigenvectors of Γ_m and its smallest eigenvalue $\lambda_{\min}$ (or the lim inf of the eigenvalues in the case of infinite dimensional spaces) is positive. Moreover there are positive constant c, c' such that for all $A > 0$ and for all $h \in \mathcal{B}(m, A)$,

$$\langle \nabla G(h), h - m \rangle \geq \min\left\{c, \frac{c'}{A}\right\} \|h - m\|^2.$$

Finally, the last proposition gives an uniform bound of the remainder term in the Taylor's expansion of the gradient.

Proposition 2.6. *Assume **(A1)**, **(A2)**, **(A4)** and **(A5a)** hold. Then, there is a positive constant C_m such that for all $h \in H$,*

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

*This result remains true replacing assumption **(A5a)** by **(A5a′)**.*

3. Applications

3.1. Applications in general separable Hilbert spaces

In this section, let us consider a separable Hilbert space H and let X be a random variable taking values in H.
Estimating geometric quantiles: The geometric quantile m^v of X corresponding to a direction v, where $v \in H$ and $\|v\| < 1$, is defined by

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

Note that if $v = 0$, the geometric quantile m^0 corresponds to the geometric median [16, 19]. Let G_v be the function we would like to minimize, defined for all $h \in H$ by $G_v(h) := \mathbb{E}\left[\|X - h\| + \langle X - h, v \rangle\right]$. Since $\|v\| < 1$,

it comes

$$\lim_{\|h\|\to\infty} G_v(h) = +\infty,$$

and G_v admits so a minimizer m^v, which is also a solution of the following equation

$$\nabla G_v(h) = -\mathbb{E}\left[\frac{X-h}{\|X-h\|}\right] - v = 0.$$

Then, assumption **(A1)** is fulfilled and the stochastic gradient algorithm and its averaged version are defined recursively for all $n \geq 1$ by

$$m_{n+1}^v = m_n^v + \gamma_n \left(\frac{X_{n+1} - m_n^v}{\|X_{n+1} - m_n^v\|} + v\right),$$
$$\overline{m}_{n+1}^v = \overline{m}_n^v + \frac{1}{n+1}\left(m_{n+1}^v - \overline{m}_n^v\right),$$

with $m_1^v = \overline{m}_1^v$ chosen bounded (choosing a positive constant M, one can take m_1^v of the form $m_1^v := X_1 \, \mathbb{1}_{\|X_1\| \leq M}$ for example). In order to ensure the uniqueness of the geometric quantiles and the convergence of these estimates, we consider from now that the following assumptions are fulfilled:

(B1) The random variable X is not concentrated on a straight line: for all $h \in H$, there is $h' \in H$ such that $\langle h, h' \rangle = 0$ and

$$\mathrm{Var}\left(\langle X, h' \rangle\right) > 0.$$

(B2) The random variable X is not concentrated around single points: for all positive constant A, there is a positive constant C_A such that for all $h \in \mathcal{B}(m^v, A)$,

$$\mathbb{E}\left[\frac{1}{\|X-h\|}\right] \leq C_A, \qquad \mathbb{E}\left[\frac{1}{\|X-h\|^2}\right] \leq C_A.$$

Note that assumption **(B2)** is not restrictive when we deal with a high dimensional space. For example, if $H = \mathbb{R}^d$ with $d \geq 3$, as discussed in [5, 7], this condition is satisfied since X admits a density which is bounded on every compact subset of $\mathbb{R}^d$. Finally, this assumption ensures the existence of the Hessian of G_v, which is defined for all $h \in H$ by

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

where for all $h, h', h'' \in H$, $h \otimes h'(h'') := \langle h, h'' \rangle \, h'$. Moreover, Corollary 2.1 in [6] ensures that if assumptions **(B1)** and **(B2)** are fulfilled, assumptions **(A2)** and **(A3)** are verified, while Lemma 5.1 in [6] ensures that assumption **(A4)** is fulfilled. Finally, for all positive integer $p \geq 1$ and for all $h \in H$,

$$\mathbb{E}\left[\left\|\frac{X-h}{\|X-h\|} + v\right\|^{2p}\right] \leq 2^{2p},$$

and assumptions **(A5a)** and **(A5b)** are so verified.

Estimating p-means: Les $p \in (1,2)$, then, the p-mean of X is defined by

$$m^{(p)} = \arg\min_{h\in H} \mathbb{E}\left[\|X-h\|^p\right]^{\frac{1}{p}} = \arg\min_{h\in H} \frac{1}{p}\mathbb{E}\left[\|X-h\|^p\right] \tag{3.1}$$

Note that the cases $p=1$ and $p=2$ correspond respectively to the geometric median and the usual mean. Let G_p be the function we would like to minimize defined for all $h \in H$ by $G_p(h) = \frac{1}{p}\mathbb{E}\left[\|X-h\|^p\right]$. This function is convex and

$$\lim_{\|h\|\to\infty} G_p(h) = +\infty,$$

and G_p admits so a minimizer $m^{(p)}$, which is also a solution of the following equation

$$\nabla G_p(h) = -\mathbb{E}\left[(X-h)\|X-h\|^{p-2}\right] = 0.$$

Then, assumption **(A1)** is fulfilled and the stochastic gradient algorithm and its averaged version are defined recursively for all $n \geq 1$ by

$$\begin{aligned} m_{n+1}^{(p)} &= m_n^{(p)} + \gamma_n \left(X_{n+1} - m_n^{(p)}\right) \left\|X_{n+1} - m_n^{(p)}\right\|^{p-2} \\ \overline{m}_{n+1}^{(p)} &= \overline{m}_n^{(p)} + \frac{1}{n+1}\left(m_{n+1}^{(p)} - \overline{m}_n^{(p)}\right). \end{aligned}$$

In order to ensure some differentiability properties and the convergence of the estimates, les us now introduce some assumptions:

(B1a′) The random variable X admits a moment of order $2p-2$.

(B1b′) For all positive integer q, the random variable X admits a moment of order q.

(B2′) The random variable X is not concentrated around single points: for all positive constant A, there is a positive constant C_A such that for all $h \in \mathcal{B}\left(m^{(p)}, A\right)$,

$$\mathbb{E}\left[\|X-h\|^{p-2}\right] \leq C_A \qquad\qquad \mathbb{E}\left[\|X-h\|^{p-3}\right] \leq C_A$$

Assumption **(B1a′)** ensures that the gradient of G_p is well defined and that assumption **(A5a)** is fulfilled while assumption **(B1b′)** ensures that **(A5b)** is fulfilled. Indeed, for all $h \in H$,

$$\begin{aligned} \|\nabla_h g_p(X,h)\| = \|X-h\|^{p-1} &\leq 2^{p-1}\left(\left\|X-m^{(p)}\right\|^{p-1} + \left\|m^{(p)}-h\right\|^{p-1}\right) \\ &\leq 2^{p-1}\left(\left\|X-m^{(p)}\right\|^{p-1} + 1 + \left\|m^{(p)}-h\right\|\right) \end{aligned}$$

Remark that this example can not be treated thanks to the theoretical tools of [14] and [1]. Indeed, in these previous papers, uniform bounds of the gradient are needed while in this example, the gradient is bounded by a term with finite moments and a term depending on the estimation errors. Finally, assumption **(B2′)** ensures that the function we would like to minimize is twice continuously differentiable and

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

Since $p \in (1,2)$, $\lambda_{\min}\left(\nabla^2 G(m)\right) \geq (p-1)\mathbb{E}\left[\frac{1}{\|X-m\|}\right] > 0$ and assumption **(A3)** is so fulfilled. Finally, thanks to **(B2′)**, assumptions **(A2)** and **(A4)** are also fulfilled.

3.2. An application in a finite dimensional space: a robust logistic regression

Let $d \geq 1$ and $H = \mathbb{R}^d$. Let (X, Y) be a couple of random variables taking values in $H \times \{-1, 1\}$. The aim is to minimize the functional G_r defined for all $h \in \mathbb{R}^d$ by (see [1])

$$G_r(h) := \mathbb{E}\left[\log\left(\cosh\left(Y - \langle X, h\rangle\right)\right)\right].$$

In order to ensure the existence and uniqueness of the solution, we consider from now that the following assumptions are fulfilled:

(B1″) There exists m^r such that $\nabla G_r(m^r) = 0$.
(B2″) The Hessian of the functional G_r at m^r is positive.
(B3a″) The random variable X admits a 2-nd moment.
(B3b″) For all integer p, the random variable X admits a pth moment.

Assumption **(B1″)** ensures the existence of a solution while **(B2′)** gives its uniqueness. Assumption **(B3a″)** ensures that the functional G_r is twice Fréchet-differentiable and its gradient and Hessian are defined for all $h \in \mathbb{R}^d$ by

$$\nabla G_r(h) = \mathbb{E}\left[\frac{-\sinh\left(Y - \langle X, h\rangle\right)}{\cosh\left(Y - \langle X, h\rangle\right)} X\right],$$
$$\nabla^2 G_r(h) = \mathbb{E}\left[\frac{1}{\left(\cosh\left(Y - \langle X, h\rangle\right)\right)^2} X \otimes X\right].$$

Note that assumption **(B2″)** is verified, for example, since there are positive constants M, M' such that the matrix $\mathbb{E}\left[X \otimes X \mathbb{1}_{\{\|X\| \leq M\}} \mathbb{1}_{\{\|Y\| \leq M'\}}\right]$ is positive. Then, the solution m^r can be estimated recursively as follows:

$$m^r_{n+1} = m^r_n + \gamma_n \frac{\sinh\left(Y_{n+1} - \langle X_{n+1}, m^r_n\rangle\right)}{\cosh\left(Y_{n+1} - \langle X_{n+1}, m^r_n\rangle\right)} X_{n+1},$$
$$\overline{m}^r_{n+1} = \overline{m}^r_n + \frac{1}{n+1}\left(m^r_{n+1} - \overline{m}^r_n\right),$$

with $\overline{m}^r_1 = m^r_1$ bounded. Under assumptions **(B1″)** to **(B3a″)**, hypothesis **(A1)** to **(A5a)** are satisfied, while under additional assumption **(B3b″)**, hypothesis **(A5b)** is satisfied. Remark that this example is already treated in [1], but only for a bounded gradient, *i.e.* under the existence of a positive constant R such that

$$\frac{\left|\sinh\left(Y - \langle X, h\rangle\right)\right|}{\cosh\left(Y - \langle X, h\rangle\right)} \|X\| \leq R,$$

i.e. only in the case where X is bounded.

Remark 3.1. Remark that these results remain true for several cases of regression. For example, one can consider the logistic regression

$$m^l := \arg\min_{h \in \mathbb{R}^d} \mathbb{E}\left[\log\left(1 + \exp\left(-Y\langle X, h\rangle\right)\right)\right],$$

with (X, Y) taking values in $\mathrm{R}^d \times \{-1, 1\}$. Then, one can consider estimates of the form

$$m_{n+1}^l = m_n^l + \gamma_n \frac{\exp\left(-Y_{n+1}\left\langle X_{n+1}, m_n^l\right\rangle\right)}{1+\exp\left(-Y_{n+1}\left\langle X_{n+1}, m_n^l\right\rangle\right)} Y_{n+1} X_{n+1},$$
$$\overline{m}_{n+1}^l = \overline{m}_n^l + \frac{1}{n+1}\left(m_{n+1}^l - \overline{m}_n^l\right).$$

4. Rates of convergence

In this section, we consider a learning rate sequence $(\gamma_n)_{n\geq 1}$ of the form $\gamma_n := c_\gamma n^{-\alpha}$ with $c_\gamma > 0$ and $\alpha \in (1/2, 1)$. Note that taking $\alpha = 1$ could be possible with a good choice of the value of the constant c_γ (taking $c_\gamma > \frac{1}{\lambda_{\min}}$ for instance). Nevertheless, the averaging step enables us to get the optimal rate of convergence with a smaller variance than the stochastic gradient algorithm with a fastly decreasing step sequence $\gamma_n = c_\gamma n^{-1}$ (see [23, 24, 26] for more details).

4.1. Almost sure rates of convergence

In this section, we focus on the almost sure rates of convergence of the algorithms defined in (2.2) and (2.3). First, the following theorem gives the consistency of the algorithms.

Theorem 4.1. *Suppose* ***(A1)*** *to* ***(A3)*** *and* ***(A5a)*** *hold. Then,*

$$\lim_{n\to\infty} \|Z_n - m\| = 0 \quad a.s,$$
$$\lim_{n\to\infty} \left\|\overline{Z}_n - m\right\| = 0 \quad a.s.$$

This result remains true replacing assumptions ***(A3)*** *and/or* ***(A5a)*** *by* ***(A3′)*** *and/or* ***(A5a′)****.*

The following theorem gives the almost sure rates of convergence of the stochastic gradient algorithm as well as of its averaged version under the additional assumption **(A4)**.

Theorem 4.2. *Suppose* ***(A1)*** *to* ***(A5a)*** *hold. For all* $\delta, \delta' > 0$*,*

$$\|Z_n - m\|^2 = o\left(\frac{(\ln n)^\delta}{n^\alpha}\right) \quad a.s,$$
$$\left\|\overline{Z}_n - m\right\|^2 = o\left(\frac{(\ln n)^{1+\delta'}}{n}\right) \quad a.s.$$

This result remains true replacing assumptions ***(A3)*** *and/or* ***(A5a)*** *by* ***(A3′)*** *and/or* ***(A5a′)****.*

Note that similar results are given in [23], but only in finite dimension. More precisely, the given proofs cannot be directly extended to the case where H is an infinite dimensional space. For example, these methods rely on the fact that the Hessian of the functional G admits finite dimensional eigenspaces, which is not necessarily true for general Hilbert spaces. Another problem is that norms are not equivalent in infinite dimensional spaces, and consequently, the Hilbert-Schmidt (or Frobenius) norm for linear operators is not necessarily finite even if the

spectral norm is. For example, under assumption **(A3)**, if H is an infinite dimensional space,

$$\|\Gamma_m\|_{op} \leq C_{\|m\|}, \qquad \text{and} \qquad \|\Gamma_m\|_{H-S} = +\infty,$$

where $\|.\|_{H-S}$ is the Hilbert-Schmidt norm.

4.2. L^p rates of convergence

In this section, we focus on the L^p rates of convergence of the algorithms. The proofs are postponed in Section 6. The idea is to give non asymptotic results without focusing only on the rate of convergence in quadratic mean. Indeed, recent works (see [4, 14] for instance), confirm that having L^p rates of convergence can be very useful to establish rates of convergence of more complex estimates.

Theorem 4.3. *Assume **(A1)** to **(A5b)** hold. Then, for all integer p, there is a positive constant K_p such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \leq \frac{K_p}{n^{p\alpha}}. \tag{4.1}$$

*This result remains true replacing assumptions **(A3)** and/or **(A5b)** by **(A3′)** and/or **(A5b′)**.*

Finally, the last theorem gives the L^p rates of convergence of the averaged estimates.

Theorem 4.4. *Assume **(A1)** to **(A5b)** hold. Then, for all integer p, there is a positive constant K'_p such that for all $n \geq 1$,*

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

*This result remains true replacing assumptions **(A3)** and/or **(A5b)** by **(A3′)** and/or **(A5b′)**.*

As done in [6, 14], one can check that, under assumptions, these rates of convergence are the optimal ones for Robbins-Monro algorithms and their averaged versions, *i.e.* one can prove that there are positive constants c, c' such that for all $n \geq 1$,

$$\mathbb{E}\left[\|Z_n - m\|^2\right] \geq \frac{c}{n^\alpha}, \qquad \mathbb{E}\left[\left\|\overline{Z}_n - m\right\|^2\right] \geq \frac{c'}{n}.$$

Remark 4.5. One can obtain the same L^p and almost sure rates of convergence for the stochastic gradient algorithm replacing assumption **(A4)** by

(A4′) There are positive constants $\epsilon > 0$ and $\beta \in (1, 2]$ such that for all $h \in \mathcal{B}(m, \epsilon)$

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

Moreover, one can get the same L^p and almost sure rates of convergence for the averaged algorithm replacing **(A4)** by **(A4′)** and taking a step sequence of the form $\gamma_n := c_\gamma n^{-\alpha}$ with $\alpha \in (\beta^{-1}, 1)$.

Remark 4.6. Let p be a positive integer, it is possible to get the L^{2p} rates of convergence of the Robbins-Monro algorithm just supposing that there is a positive integer q such that $q > 2p + 2$ and a positive constant L_q such that $\mathbb{E}\left[f(X,h)^{2q}\right] \leq L_q$ (or such that $\mathbb{E}\left[\nabla_h g(X,h)\right] \leq L_q\left(1 + \|h - m\|^{2q}\right)$) and taking a step sequence of the form $\gamma_n := c_\gamma n^{-\alpha}$ with $\alpha \in \left(\frac{1}{2}, \frac{q}{p+2+q}\right)$.

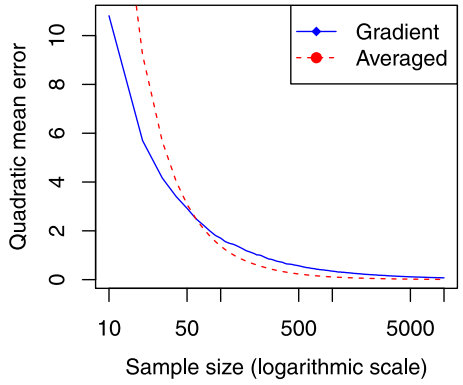

FIGURE 1. Comparison of the evolution of the quadratic mean error of gradient estimates (in blue) and of their averaged version (in red) in relation to the sample size.

5. Simulation study

In this section, we consider a random gaussian vector $X \sim \mathcal{N}(0, I_{100})$ taking values in $\mathbb{R}^{100}$, and we aim to estimate the p-mean $m^{(p)}$ of X with $p = 1.5$. Note that in this case, $m^{(p)} = 0_{\mathbb{R}^{100}}$. We now consider q samples $X_{1,1}, \ldots, X_{1,n}, \ldots, X_{q,1}, \ldots, X_{q,n}$ with a size n. In order to compare the different estimates, for a fixed sample size n, we will consider the empirical quadratic mean error of the estimates, *i.e.* given an estimate $\hat{m}$ of m and the associate estimates $\hat{m}_{1,n}, \ldots, \hat{m}_{q,n}$, we will consider

$$\mathrm{QME}(\hat{m}, m) = \frac{1}{q} \sum_{i=1}^{q} \|\hat{m}_{i,n} - m\|^2 .$$

In order to initialize the algorithms, we take the first data, *i.e.* $m_{i,1}^{(p)} = X_{i,1}$. In Figure 1, we consider a step sequence $\gamma_n = c_\gamma n^{-\alpha}$ with $c = 2$ and $\alpha = 0.66$. One can check that the averaged algorithm converges faster than the gradient and become better after having dealt with a small number of data (about 50). This quite bad behavior on the first step can be explained by a quite bad initialization of the gradient algorithm which so spend some time before turning around the target. In Figure 2, we study the impact of the choice of α on the performance of the estimates for a fixed constant $c_\gamma = 2$. The case where $\alpha = 1$ is not considered since it needs to have informations on the smallest eigenvalue of the Hessian of the functional we would like to minimize, informations that are usually unknown. Without any surprise (in view of Thms. 4.2 an 4.3), gradient estimates seems to converge faster when α increases. Inversely, for small sample size, the averaged version seems to converge faster when α decreases for small sample size, before having analogous behaviors for $n = 1000$. This can be explained by the fact that the less important is α, the more the gradient estimates will "move", and the more they have a chance to turn around the target quickly.

Finally, in Table 1, we study the impact on the estimates of the choices of α and c_γ for a moderate sample size $n = 10^4$. As expected, one can see that averaged estimates are globally better than gradient ones and are more stable in relation to the choice of the step sequence. The quite critical choices of step sequence for the averaged algorithm are when we both take c_γ small and α close to 1. This is not surprising because here again, the gradient steps need too much data before turning around the target, since, for example,

$$\sum_{i=1}^{10^4} i^{-0.9} \simeq 15.7.$$

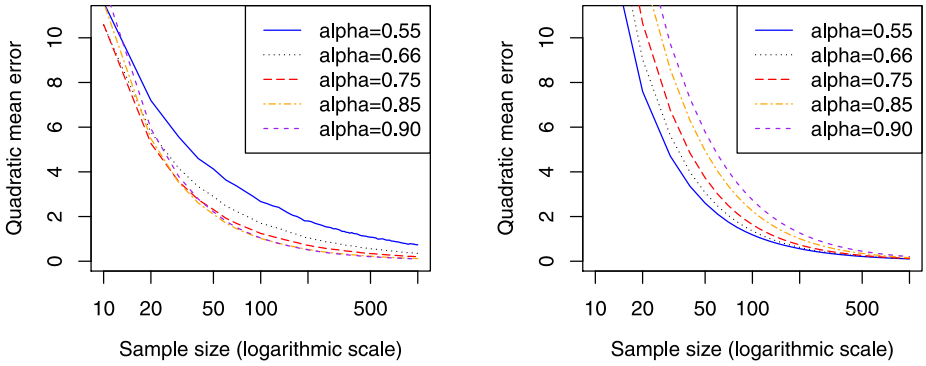

FIGURE 2. Comparison of the quadratic mean error of gradient estimates (*on the left*) and of their averaged version (*on the right*) in relation to the sample size for $\alpha = 0.55, 0.66, 0.75, 0.85, 0.9$.

TABLE 1. Quadratic mean errors ($.10^{-2}$) of the gradient estimates (*on the left*) and of averaged estimates (*on the right*) for a sample size $n = 10\,000$ for different α and c_γ.

		Gradient estimates						**Averaged estimates**				
		α						α				
		0.55	0.66	0.75	0.85	0.9		0.55	0.66	0.75	0.85	0.9
c_γ	1	9.95	3.76	1.84	1.07	2.33	c_γ	1.05	1.12	1.41	4.02	12.05
	2	20.29	7.39	3.40	1.52	1.15		1.00	1.05	1.11	1.34	1.63
	5	50.34	17.80	8.08	3.37	2.20		1.01	1.01	1.03	1.08	1.13
	10	101.75	36.40	15.79	6.54	4.18		1.01	1.00	1.02	1.05	1.06
	20	209.05	73.62	31.32	12.87	7.94		0.99	1.00	0.99	1.04	1.03

6. PROOFS

6.1. Some decompositions of the algorithms

In order to simplify the proofs thereafter, we introduce some usual decompositions of the algorithms. First, let us recall that the Robbins-Monro algorithm is defined by

$$Z_{n+1} = Z_n - \gamma_n U_{n+1}, \tag{6.1}$$

with $U_{n+1} := \nabla_h g\left(X_{n+1}, Z_n\right)$. Then, let $\xi_{n+1} := \Phi(Z_n) - U_{n+1}$, equality (6.1) can be written as

$$Z_{n+1} - m = Z_n - m - \gamma_n \Phi(Z_n) + \gamma_n \xi_{n+1}. \tag{6.2}$$

Note that (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$. Furthermore, linearizing the gradient, equation (6.2) can be written as

$$Z_{n+1} - m = \left(I_H - \gamma_n \Gamma_m\right)\left(Z_n - m\right) + \gamma_n \xi_{n+1} - \gamma_n \delta_n, \tag{6.3}$$

where $\delta_n := \Phi(Z_n) - \Gamma_m\left(Z_n - m\right)$ is the remainder term in the Taylor's expansion of the gradient. Note that thanks to Proposition 2.6, there is a positive constant C_m such that for all $n \geq 1$, $\|\delta_n\| \leq C_m \left\|Z_n - m\right\|^2$. Finally,

by induction, we have the following usual decomposition

$$Z_n - m = \beta_{n-1}\left(Z_1 - m\right) + \beta_{n-1}M_n - \beta_{n-1}R_n, \tag{6.4}$$

with

$$\beta_{n-1} := \prod_{k=1}^{n-1}\left(I_H - \gamma_k\Gamma_m\right), \qquad M_n := \sum_{k=1}^{n-1}\gamma_k\beta_k^{-1}\xi_{k+1},$$
$$\beta_0 := I_H, \qquad R_n := \sum_{k=1}^{n-1}\gamma_k\beta_k^{-1}\delta_k.$$

In the same way, in order to get the rates of convergence, we need to exhibit a new decomposition of the averaged algorithm. In this aim, equality (6.3) can be written as

$$\Gamma_m\left(Z_n - m\right) = \frac{Z_n - m}{\gamma_n} - \frac{Z_{n+1} - m}{\gamma_n} + \xi_{n+1} - \delta_n.$$

As in [24], summing these equalities, applying Abel's transform and dividing by n, we have

$$\Gamma_m\left(\overline{Z}_n - m\right) = \frac{1}{n}\left(\frac{Z_1 - m}{\gamma_1} - \frac{Z_{n+1} - m}{\gamma_n} + \sum_{k=2}^{n}\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)\left(Z_k - m\right) - \sum_{k=1}^{n}\delta_k\right) + \frac{1}{n}\sum_{k=1}^{n}\xi_{k+1}. \tag{6.5}$$

6.2. Proof of Section 4.1

Proof of Theorem 4.1. Using decomposition (6.2) and since (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$,

$$\mathbb{E}\left[\left\|Z_{n+1} - m\right\|^2 |\mathcal{F}_n\right] = \left\|Z_n - m\right\|^2 - 2\gamma_n\left\langle Z_n - m, \Phi(Z_n)\right\rangle + \gamma_n^2\left\|\Phi(Z_n)\right\|^2 + \gamma_n^2\mathbb{E}\left[\left\|\xi_{n+1}\right\|^2 |\mathcal{F}_n\right].$$

Moreover, with Assumption **(A5a)**,

$$\begin{aligned}\mathbb{E}\left[\left\|\xi_{n+1}\right\|^2 |\mathcal{F}_n\right] &= \mathbb{E}\left[\left\|U_{n+1}\right\|^2 |\mathcal{F}_n\right] - 2\left\langle\mathbb{E}\left[U_{n+1}|\mathcal{F}_n\right], \Phi(Z_n)\right\rangle + \left\|\Phi(Z_n)\right\|^2 \\ &\le \mathbb{E}\left[\left(f(X_{n+1}, Z_n) + C\left\|Z_n - m\right\|\right)^2 |\mathcal{F}_n\right] - \left\|\Phi(Z_n)\right\|^2 \\ &\le 2L_1 + 2C^2\left\|Z_n - m\right\|^2 - \left\|\Phi(Z_n)\right\|^2.\end{aligned}$$

Thus,

$$\mathbb{E}\left[\left\|Z_{n+1} - m\right\|^2 |\mathcal{F}_n\right] \le \left(1 + 2C^2\gamma_n^2\right)\left\|Z_n - m\right\|^2 - 2\gamma_n\left\langle\Phi(Z_n), Z_n - m\right\rangle + 2\gamma_n^2 L_1.$$

Since $\langle\Phi(Z_n), Z_n - m\rangle \ge 0$ and $\sum_{n\ge1}\gamma_n^2 < +\infty$, Robbins-Siegmund theorem (see Thm. E.1) ensures that $\|Z_n - m\|$ converges almost surely to a finite random variable and that

$$\sum_{n\ge1}\gamma_n\left\langle\Phi(Z_n), Z_n - m\right\rangle < +\infty \quad a.s.$$

Moreover, since $\langle \Phi(Z_n), Z_n - m \rangle \geq 0$, by induction, there is a positive constant M such that for all $n \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1}-m\|^2\right] &\leq \left(1+2C^2\gamma_n^2\right)\mathbb{E}\left[\|Z_n-m\|^2\right] + 2\gamma_n^2 L_1 \\ &\leq \left(\prod_{k\geq 1}\left(1+2C^2\gamma_k^2\right)\right)\mathbb{E}\left[\|Z_1-m\|^2\right] + 2L_1\left(\prod_{k\geq 1}\left(1+2C^2\gamma_k^2\right)\right)\sum_{k\geq 1}\gamma_k^2 \\ &\leq M.\end{aligned}$$

Thus, one can conclude the proof in the same way as in the proof of Theorem 3.1 in [5] for instance. Finally, one can apply Toeplitz's lemma (see [12], Lem. 2.2.13) to get the strong consistency of the averaged algorithm. □

In order to get the almost sure rates of convergence of the Robbins-Monro algorithm, we now introduce a technical lemma which gives the rate of convergence of the martingale term $\beta_{n-1}M_n$ in decomposition (6.4).

Lemma 6.1. *Suppose assumptions **(A1)** to **(A3)** and **(A5a)** hold. Then, for all $\delta > 0$,*

$$\|\beta_{n-1}M_n\|^2 = o\left(\frac{(\ln n)^\delta}{n^\alpha}\right) \quad a.s.$$

Proof of Lemma 6.1. Since (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$, and since $M_{n+1} = M_n + \gamma_n\beta_n^{-1}\xi_{n+1}$,

$$\begin{aligned}\mathbb{E}\left[\|\beta_n M_{n+1}\|^2 |\mathcal{F}_n\right] &= \|\beta_n M_n\|^2 + 2\gamma_n\langle \beta_n M_n, \mathbb{E}\left[\xi_{n+1}|\mathcal{F}_n\right]\rangle + \gamma_n^2\mathbb{E}\left[\|\xi_{n+1}\|^2|\mathcal{F}_n\right] \\ &= \|\beta_n M_n\|^2 + \gamma_n^2\mathbb{E}\left[\|\xi_{n+1}\|^2|\mathcal{F}_n\right] \\ &\leq \|I_H - \gamma_n\Gamma_m\|_{op}^2\|\beta_{n-1}M_n\|^2 + \gamma_n^2\mathbb{E}\left[\|\xi_{n+1}\|^2|\mathcal{F}_n\right].\end{aligned}$$

Since each eigenvalue λ of Γ_m verifies $0 < \lambda_{\min} \leq \lambda \leq C$ and since (γ_n) converges to 0, there is a rank n_0 such that for all $n \geq n_0$, $\|I_H - \gamma_n\Gamma_m\|_{op} \leq 1 - \lambda_{\min}\gamma_n$. Thus, for all $n \geq n_0$,

$$\mathbb{E}\left[\|\beta_n M_{n+1}\|^2|\mathcal{F}_n\right] \leq (1-\lambda_{\min}\gamma_n)^2\|\beta_{n-1}M_n\|^2 + \gamma_n^2\mathbb{E}\left[\|\xi_{n+1}\|^2|\mathcal{F}_n\right].$$

Let $\delta > 0$, for all $n \geq 1$, let $V_n := \frac{n^{2\alpha-1}}{(\ln n)^{1+\delta}}\|\beta_{n-1}M_n\|^2$, then for all $n \geq n_0$,

$$\begin{aligned}\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] &\leq (1-\lambda_{\min}\gamma_n)^2\frac{(n+1)^{2\alpha-1}}{(\ln(n+1))^{1+\delta}}\|\beta_{n-1}M_n\|^2 + \frac{(n+1)^{2\alpha-1}}{(\ln(n+1))^{1+\delta}}\gamma_n^2\mathbb{E}\left[\|\xi_{n+1}\|^2|\mathcal{F}_n\right] \\ &= (1-\lambda_{\min}\gamma_n)^2\left(\frac{n+1}{n}\right)^{2\alpha-1}\left(\frac{\ln n}{\ln(n+1)}\right)^{1+\delta}V_n + \frac{(n+1)^{2\alpha-1}}{(\ln(n+1))^{1+\delta}}\gamma_n^2\mathbb{E}\left[\|\xi_{n+1}\|^2|\mathcal{F}_n\right].\end{aligned}$$

Moreover, there are a positive constant c and a rank n_0' (let us take $n_0' \geq n_0$) such that for all $n \geq n_0'$,

$$\left(1-\lambda_{\min}c_\gamma n^{-\alpha}\right)\left(\frac{n+1}{n}\right)^{2\alpha-1}\left(\frac{\ln n}{\ln(n+1)}\right)^{1+\delta} \leq 1 - cn^{-\alpha}.$$

Furthermore, $cn^{-\alpha}V_n = c\frac{n^{\alpha-1}}{(\ln n)^{1+\delta}}\left\|\beta_{n-1}M_n\right\|^2$. Thus, for all $n \geq n_0'$,

$$\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq V_n + \frac{(n+1)^{2\alpha-1}}{(\ln(n+1))^{1+\delta}}\gamma_n^2\mathbb{E}\left[\left\|\xi_{n+1}\right\|^2|\mathcal{F}_n\right] - c\frac{n^{\alpha-1}}{(\ln n)^{1+\delta}}\left\|\beta_{n-1}M_n\right\|^2. \tag{6.6}$$

Finally, since $\mathbb{E}\left[\left\|\xi_{n+1}\right\|^2|\mathcal{F}_n\right] \leq 2L_1 + 2C\left\|Z_n - m\right\|^2$ and since $\|Z_n - m\|$ converges almost surely to 0, the application of the Robbins-Siegmund theorem (see Theorem E.1) ensures that (V_n) converges almost surely to a finite random variable and ensures that

$$\sum_{n\geq n_0'}\frac{n^{\alpha-1}}{(\ln n)^{1+\delta}}\left\|\beta_{n-1}M_n\right\|^2 < \infty \quad a.s.$$

Previous inequality can also be written as

$$\sum_{n\geq n_0'}\frac{1}{n\ln n}\left(\frac{n^{\alpha}}{(\ln n)^{\delta}}\left\|\beta_{n-1}M_n\right\|^2\right) < \infty \quad a.s,$$

so that we necessarily have, applying Toeplitz's lemma,

$$\frac{n^{\alpha}}{(\ln n)^{\delta}}\left\|\beta_{n-1}M_n\right\|^2 \xrightarrow[n\to\infty]{a.s} 0. \tag{6.7}$$

□

Remark 6.2. Note that this proof is the main difference with [24]. Indeed, in order to prove the same result, many methods were used but they cannot be directly applied if H is a infinite dimensional space. For example, it is based on the fact that the Hessian of the function we would like to minimize admits finite dimensional eigenspaces, which is not automatically verified in our case.

Proof of Theorem 4.2. **R**ate of convergence of the Robbins-Monro algorithm: Applying decomposition (6.4), as in [23], let

$$\Delta_n = \beta_{n-1}\left(Z_1 - m\right) - \beta_{n-1}R_n = \left(Z_n - m\right) - \beta_{n-1}M_n.$$

We have

$$\begin{aligned}\Delta_{n+1} &= Z_{n+1} - m - \beta_n M_{n+1}\\ &= \left(I_H - \gamma_n\Gamma_m\right)\left(Z_n - m\right) + \gamma_n\xi_{n+1} - \gamma_n\delta_n - \gamma_n\xi_{n+1} - \left(I_H - \gamma_n\Gamma_m\right)\beta_{n-1}M_n\\ &= \left(I_H - \gamma_n\Gamma_m\right)\Delta_n - \gamma_n\delta_n.\end{aligned}$$

Thus, applying a lemma of stabilization (see [11] Lem. 4.1.1 for instance), and since $\|\delta_n\| \leq C_m\left\|Z_n - m\right\|^2$,

$$\|\Delta_n\| = O\left(\|\delta_n\|\right) = O\left(\left\|Z_n - m\right\|^2\right) \quad a.s.$$

Finally, since (Z_n) converges almost surely to m, $\|\Delta_n\| = o(\|Z_n - m\|)$ almost surely and

$$\begin{aligned}\|Z_n - m\| &\leq \|\beta_{n-1} M_n\| + \|\Delta_n\| \\ &= o\left(\frac{(\ln n)^{\delta/2}}{n^{\alpha/2}}\right) + o(\|Z_n - m\|) \quad a.s,\end{aligned}$$

which concludes the proof.

Rate of convergence of the averaged algorithm: With the help of decomposition (6.5),

$$\begin{aligned}\left\|\overline{Z}_n - m\right\|^2 \leq \frac{5}{\lambda_{\min}^2 n^2}\frac{\|Z_1 - m\|^2}{\gamma_1^2} + \frac{5}{\lambda_{\min}^2 n^2}\frac{\|Z_{n+1} - m\|^2}{\gamma_n^2} + \frac{5}{\lambda_{\min}^2 n^2}\left\|\sum_{k=1}^{n}\delta_k\right\|^2 \\ + \frac{5}{\lambda_{\min}^2 n^2}\left\|\sum_{k=2}^{n}(Z_k - m)\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)\right\|^2 + \frac{5}{\lambda_{\min}^2 n^2}\left\|\sum_{k=1}^{n}\xi_{k+1}\right\|^2.\end{aligned}$$

As in [14], thanks to the almost sure rate of convergence of the Robbins-Monro algorithm, one can check that

$$\begin{aligned}\frac{1}{n^2}\frac{\|Z_1 - m\|}{\gamma_1} &= o\left(\frac{1}{n}\right) \quad a.s, \\ \frac{1}{n^2}\frac{\|Z_{n+1} - m\|^2}{\gamma_n^2} &= o\left(\frac{1}{n}\right) \quad a.s, \\ \frac{1}{n^2}\left\|\sum_{k=2}^{n}(Z_k - m)\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)\right\|^2 &= o\left(\frac{1}{n}\right) \quad a.s, \\ \frac{1}{n^2}\left\|\sum_{k=1}^{n}\delta_k\right\|^2 &= o\left(\frac{1}{n}\right) \quad a.s.\end{aligned}$$

Let $\delta > 0$ and $M_n' := \frac{\sqrt{n}}{\sqrt{(\ln n)^{1+\delta}}}\left\|\frac{1}{n}\sum_{k=1}^{n}\xi_{k+1}\right\| = \frac{1}{\sqrt{n(\ln n)^{1+\delta}}}\left\|\sum_{k=1}^{n}\xi_{k+1}\right\|$. Since (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$, and since

$$\begin{aligned}\mathbb{E}\left[\|\xi_{n+2}\|^2 |\mathcal{F}_{n+1}\right] &\leq 2\mathbb{E}\left[f(X_{n+2}, Z_{n+1})^2 |\mathcal{F}_{n+1}\right] + 2C^2\|Z_{n+1} - m\|^2 \\ &\leq 2L_1 + 2C^2\|Z_{n+1} - m\|^2,\end{aligned}$$

we have

$$\begin{aligned}\mathbb{E}\left[M_{n+1}'^2 |\mathcal{F}_{n+1}\right] &= \frac{n(\ln n)^{1+\delta}}{(n+1)(\ln(n+1))^{1+\delta}}M_n'^2 + \frac{1}{(n+1)(\ln(n+1))^{1+\delta}}\mathbb{E}\left[\|\xi_{n+2}\|^2 |\mathcal{F}_{n+1}\right] \\ &\leq M_n'^2 + \frac{1}{(n+1)(\ln(n+1))^{1+\delta}}\left(2L_1 + 2C^2\|Z_{n+1} - m\|^2\right).\end{aligned}$$

Since $\|Z_{n+1} - m\|$ converges almost surely to 0, applying Robbins-Siegmund theorem (see Thm. E.1), $M_n'^2$ converges almost surely to a finite random variable, which concludes the proof. □

6.3. Proof of Theorem 4.3

In order to prove Theorem 4.3 with the help of a strong induction on p, we have to introduce some technical lemmas (the proofs are given in Appendix). Note that these lemmas remain true replacing assumptions **(A3)** and/or **(A5b)** by **(A3′)** and/or **(A5b′)** but the proofs are only given for the first assumptions.

The first lemma gives a bound of the $2p$th moment when inequality (4.1) is verified for all integer from 0 to $p-1$.

Lemma 6.3. *Assume **(A1)** to **(A5b)** hold. Let p be a positive integer, and suppose that for all $k \leq p-1$, there is a positive constant K_k such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|Z_n - m\|^{2k}\right] \leq \frac{K_k}{n^{k\alpha}}. \tag{6.8}$$

Then, there are positive constants c_0, C_1, C_2 and a rank n_α such that for all $n \geq n_\alpha$,

$$\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right] \leq (1 - c_0\gamma_n)\,\mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{C_1}{n^{(p+1)\alpha}} + C_2\gamma_n\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right].$$

Then, the second lemma gives an upper bound of the $(2p+2)$th moment when inequality (4.1) is verified for all integer from 0 to $p-1$.

Lemma 6.4. *Assume **(A1)** to **(A3)** and **(A5b)** hold. Let p be a positive integer, and suppose that for all $k \leq p-1$, there is a positive constant K_k such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|Z_n - m\|^{2k}\right] \leq \frac{K_k}{n^{k\alpha}}.$$

Then, there are positive constants C_1', C_2' and a rank n_α such that for all $n \geq n_\alpha$,

$$\mathbb{E}\left[\|Z_{n+1} - m\|^{2p+2}\right] \leq \left(1 - \frac{2}{n}\right)^{p+1}\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{C_1'}{n^{(p+2)\alpha}} + C_2'\gamma_n^2\mathbb{E}\left[\|Z_n - m\|^{2p}\right].$$

Finally, the last lemma enables us to give a bound of the probability for the Robbins-Monro algorithm to go far away from m, which is crucial in order to prove Lemma 6.4.

Lemma 6.5. *Assume **(A1)** to **(A3)** and **(A5b)** hold. Then, for all integer $p \geq 1$, there is a positive constant M_p such that for all $n \geq 1$,*

$$\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \leq M_p.$$

Proof of Theorem 4.3. As in [14], we will prove with the help of a strong induction that for all integer $p \geq 1$, and for all $\beta \in \left(\alpha, \frac{p+2}{p}\alpha - \frac{1}{p}\right)$, there are positive constants $K_p, C_{\beta,p}$ such that for all $n \geq 1$,

$$\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \leq \frac{K_p}{n^{p\alpha}},$$
$$\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] \leq \frac{C_{\beta,p}}{n^{\beta p}}.$$

Applying Lemma 6.5, Lemma 6.3 and Lemma 6.4, as soon as the initialization is satisfied, the proof is strictly analogous to the proof of Theorem 4.1 in [14]. Thus, we will just prove that for $p = 1$ and for all $\beta \in (\alpha, 3\alpha - 1)$,

there are positive constants $K_1', C_{\beta,1}'$ such that for all $n \geq 1$,

$$\mathbb{E}\left[\|Z_n - m\|^2\right] \leq \frac{K_1'}{n^\alpha},$$
$$\mathbb{E}\left[\|Z_n - m\|^4\right] \leq \frac{C_{\beta,1}'}{n^\beta}.$$

We now split the end of the proof into two steps.

Step 1: Calibration of the constants. In order to simplify the demonstration thereafter, we now introduce some notations. Let $K_1', C_{\beta,1}'$ be positive constants such that $K_1' \geq 2^{1+\alpha} C_1 c_0^{-1} c_\gamma^{-1}$, ($c_0, C_1$ are defined in Lem. 6.3), and $2K_1' \geq C_{\beta,1}' \geq K_1' \geq 1$. By definition of β, there is a rank $n_\beta \geq n_\alpha$ (n_α is defined in Lem. 6.3 and in Lem. 6.4) such that for all $n \geq n_\beta$,

$$(1 - c_0\gamma_n)\left(\frac{n+1}{n}\right)^\alpha + \frac{1}{2}c_0\gamma_n + \frac{2^{\alpha+\beta+1}c_\gamma C_2}{(n+1)^\beta} \leq 1,$$
$$\left(1 - \frac{2}{n}\right)^2\left(\frac{n+1}{n}\right)^\beta + \left(C_1' + C_2' c_\gamma^2\right) 2^{3\alpha}\frac{1}{(n+1)^{3\alpha-\beta}} \leq 1,$$

with C_2 defined in Lemma 6.3 and C_1', C_2' defined in Lemma 6.4. The rank n_β exists because since $\beta > \alpha$,

$$\begin{aligned}(1 - c_0\gamma_n)\left(\frac{n+1}{n}\right)^\alpha + \frac{1}{2}c_0\gamma_n + \frac{2^{\alpha+\beta+1}c_\gamma C_2}{(n+1)^\beta} &= 1 - c_0\gamma_n + \frac{\alpha}{n} + \frac{1}{2}c_0\gamma_n + O\left(\frac{1}{n^\beta}\right)\\ &= 1 - \frac{1}{2}c_0\gamma_n + o\left(\frac{1}{n^\alpha}\right).\end{aligned}$$

Moreover, since $\beta < 3\alpha - 1$, we have $\beta < 2$, and

$$\begin{aligned}\left(1 - \frac{2}{n}\right)^2\left(\frac{n+1}{n}\right)^\beta + \left(C_1' + C_2' c_\gamma^2\right) 2^{3\alpha}\frac{1}{(n+1)^{3\alpha-\beta}} &= 1 - (4-2\beta)\frac{1}{n} + o\left(\frac{1}{n}\right) + O\left(\frac{1}{n^{3\alpha-\beta}}\right)\\ &= 1 - (4-2\beta)\frac{1}{n} + o\left(\frac{1}{n}\right).\end{aligned}$$

Step 2: The induction on n. Let us take $K_1' \geq \max_{1\leq k\leq n_\beta}\left\{k^\alpha \mathbb{E}\left[\|Z_k - m\|^2\right]\right\}$ and $C_{\beta,1}' \geq \max_{1\leq k\leq n_\beta}\left\{k^\beta \mathbb{E}\left[\|Z_k - m\|^4\right]\right\}$. We now prove by induction that for all $n \geq n_\beta$,

$$\mathbb{E}\left[\|Z_n - m\|^2\right] \leq \frac{K_1'}{n^\alpha},$$
$$\mathbb{E}\left[\|Z_n - m\|^4\right] \leq \frac{C_{\beta,1}'}{n^\beta}.$$

Applying Lemma 6.3 and by induction, since $2K_1' \geq C_{\beta,1}' \geq K_1' \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1} - m\|^2\right] &\leq (1 - c_0\gamma_n)\,\mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{C_1}{n^{2\alpha}} + C_2\gamma_n \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right]\\ &\leq (1 - c_0\gamma_n)\frac{K_1'}{n^\alpha} + \frac{C_1}{n^{2\alpha}} + 2C_2\gamma_n\frac{K_1'}{n^\beta}.\end{aligned}$$

Factorizing by $\frac{K_1'}{(n+1)^\alpha}$,

$$\mathbb{E}\left[\|Z_{n+1}-m\|^2\right] \leq (1-c_0\gamma_n)\left(\frac{n+1}{n}\right)^\alpha \frac{K_1'}{(n+1)^\alpha} + \frac{2^\alpha C_1 c_\gamma^{-1}\gamma_n}{(n+1)^\alpha} + \frac{2^{\alpha+\beta+1}c_\gamma C_2}{(n+1)^\beta}\frac{K_1'}{(n+1)^\alpha}.$$

Taking $K_1' \geq 2^{1+\alpha}C_1 c_\gamma^{-1}c_0^{-1}$,

$$\mathbb{E}\left[\|Z_{n+1}-m\|^2\right] \leq \left((1-c_0\gamma_n)\left(\frac{n+1}{n}\right)^\alpha + \frac{1}{2}c_0\gamma_n + \frac{2^{\alpha+\beta+1}c_\gamma C_2}{(n+1)^\beta}\right)\frac{K_1'}{(n+1)^\alpha}.$$

By definition of n_β,

$$\mathbb{E}\left[\|Z_{n+1}-m\|^2\right] \leq \frac{K_1'}{(n+1)^\alpha}. \tag{6.9}$$

In the same way, one can check by induction and applying Lemma 6.4 that

$$\mathbb{E}\left[\|Z_{n+1}-m\|^4\right] \leq \left(\left(1-\frac{2}{n}\right)^2\left(\frac{n+1}{n}\right)^\beta + 2^{3\alpha}\frac{C_1'+C_2'c_\gamma^2}{(n+1)^{3\alpha-\beta}}\right)\frac{C_{\beta,1}'}{(n+1)^\beta}.$$

By definition of n_β,

$$\mathbb{E}\left[\|Z_{n+1}-m\|^4\right] \leq \frac{C_{\beta,1}'}{n^\beta}, \tag{6.10}$$

which concludes the induction on n, and one can conclude the induction on p and the proof in a similar way as in [14]. □

6.4. Proof of Theorem 4.4

Proof of Theorem 4.4. Let $\lambda_{\min}$ be the smallest eigenvalue of Γ_m, with the help of decomposition (6.5), for all integer $p \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{Z}_n-m\right\|^{2p}\right] \leq{}& \frac{5^{2p-1}}{\lambda_{\min}^{2p}n^{2p}}\frac{\mathbb{E}\left[\|Z_1-m\|^{2p}\right]}{\gamma_1^{2p}} + \frac{5^{2p-1}}{\lambda_{\min}^{2p}n^{2p}}\frac{\mathbb{E}\left[\|Z_{n+1}-m\|^{2p}\right]}{\gamma_n^{2p}} + \frac{5^{2p-1}}{\lambda_{\min}^{2p}n^{2p}}\mathbb{E}\left[\left\|\sum_{k=1}^n \delta_k\right\|^{2p}\right] \\ &+ \frac{5^{2p-1}}{\lambda_{\min}^{2p}n^{2p}}\mathbb{E}\left[\left\|\sum_{k=2}^n (Z_k-m)\left(\frac{1}{\gamma_k}-\frac{1}{\gamma_{k-1}}\right)\right\|^{2p}\right] + \frac{5^{2p-1}}{\lambda_{\min}^{2p}n^{2p}}\mathbb{E}\left[\left\|\sum_{k=1}^n \xi_{k+1}\right\|^{2p}\right].\end{aligned}$$

As in [14], applying Theorem 4.3 and Lemma 4.1 in [14], one can check that there are positive constants $R_{1,p}, R_{2,p}, R_{3,p}, R_{4,p}$ such that for all $n \geq 1$,

$$\frac{1}{n^{2p}} \frac{\mathbb{E}\left[\|Z_1 - m\|^{2p}\right]}{\gamma_1^{2p}} \leq \frac{R_{1,p}}{n^{2p}},$$
$$\frac{1}{n^{2p}} \frac{\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right]}{\gamma_n^{2p}} \leq \frac{R_{2,p}}{n^{(2-\alpha)p}},$$
$$\frac{1}{n^{2p}} \mathbb{E}\left[\left\|\sum_{k=2}^{n} (Z_k - m)\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)\right\|^{2p}\right] \leq \frac{R_{3,p}}{n^{(2-\alpha)p}},$$
$$\frac{1}{n^{2p}} \mathbb{E}\left[\left\|\sum_{k=1}^{n} \delta_k\right\|^{2p}\right] \leq \frac{R_{4,p}}{n^{2\alpha p}}.$$

We now prove with the help of a strong induction that for all integer $p \geq 1$, there is a positive constant C_p such that

$$\mathbb{E}\left[\left\|\sum_{k=1}^{n} \xi_{k+1}\right\|^{2p}\right] \leq C_p n^p.$$

Step 1: Initialization of the induction. Since (ξ_n) is martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$,

$$\mathbb{E}\left[\left\|\sum_{k=1}^{n} \xi_{k+1}\right\|^{2}\right] = \sum_{k=1}^{n} \mathbb{E}\left[\|\xi_{k+1}\|^2\right] + 2\sum_{k=1}^{n}\sum_{k'=k+1}^{n} \mathbb{E}\left[\langle \xi_{k+1}, \xi_{k'+1}\rangle\right] = \sum_{k=1}^{n} \mathbb{E}\left[\|\xi_{k+1}\|^2\right].$$

Moreover, since $\mathbb{E}\left[\|\xi_{n+1}\|^2 |\mathcal{F}_n\right] \leq \mathbb{E}\left[\|U_{n+1}\|^2 |\mathcal{F}_n\right] \leq 2\mathbb{E}\left[f(X_{n+1}, Z_n)^2 |\mathcal{F}_n\right] + 2C^2 \|Z_n - m\|^2$, applying Theorem 4.3, there is a positive constant C_1 such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\|\sum_{k=1}^{n} \xi_{k+1}\right\|^{2}\right] \leq 2\sum_{k=1}^{n} \mathbb{E}\left[f(X_{k+1}, Z_k)^2 |\mathcal{F}_k\right] + 2C^2 \sum_{k=1}^{n} \mathbb{E}\left[\|Z_k - m\|^2\right] \leq C_1 n.$$

Step 2: The induction. Let $p \geq 2$, we suppose from now that for all $p' \leq p-1$, there is a positive constant $C_{p'}$ such that for all $n \geq 1$,

$$\mathbb{E}\left[\left\|\sum_{k=1}^{n} \xi_{k+1}\right\|^{2p'}\right] \leq C_{p'} n^{p'}.$$

First, note that

$$\left\|\sum_{k=1}^{n+1} \xi_{k+1}\right\|^2 = \left\|\sum_{k=1}^{n} \xi_{k+1}\right\|^2 + 2\left\langle \sum_{k=1}^{n} \xi_{k+1}, \xi_{n+2}\right\rangle + \|\xi_{n+2}\|^2.$$

Thus, let $M_n := \sum_{k=1}^{n} \xi_{k+1}$, with the help of previous equality and applying Cauchy-Schwarz's inequality,

$$\begin{aligned}\|M_{n+1}\|^{2p} \leq \left(\|M_n\|^2 + \|\xi_{n+2}\|^2\right)^p + 2\left\langle M_n, \xi_{n+2}\right\rangle \left(\|M_n\|^2 + \|\xi_{n+2}\|^2\right)^{p-1} \\ + \sum_{k=2}^{p} \binom{p}{k} 2^k \|M_n\|^k \|\xi_{n+2}\|^k \left(\|M_n\|^2 + \|\xi_{n+2}\|^2\right)^{p-k}.\end{aligned}$$

We now bound the expectation of the three terms on the right-hand side of previous inequality. First, since

$$\begin{aligned}\|U_{n+1}\| &\leq f\left(X_{n+1}, Z_n\right) + C\|Z_n - m\|, \\ \|\Phi(Z_n)\| &\leq \sqrt{L_1} + C\|Z_n - m\|,\end{aligned}$$

we have

$$\begin{aligned}\mathbb{E}\left[\|\xi_{n+2}\|^{2k} |\mathcal{F}_{n+1}\right] &\leq 3^{2k-1}\left(\mathbb{E}\left[f(X_{n+2}, Z_n)^{2k} |\mathcal{F}_{n+1}\right] + 2^{2k} C^{2k} \|Z_{n+1} - m\|^{2k} + L_1^k\right) \\ &\leq 3^{2k-1}\left(L_k + L_1^k + 2^{2k} C^{2k} \|Z_{n+1} - m\|^{2k}\right).\end{aligned}$$

Then, since M_n is F_{n+1}-measurable,

$$\begin{aligned}\mathbb{E}\left[\left(\|M_n\|^2 + \|\xi_{n+2}\|^2\right)^p\right] &\leq \mathbb{E}\left[\|M_n\|^{2p}\right] + \sum_{k=1}^{p} \binom{p}{k} \mathbb{E}\left[\mathbb{E}\left[\|\xi_{n+2}\|^{2k} |\mathcal{F}_n\right] \|M_n\|^{2p-2k}\right] \\ &\leq \mathbb{E}\left[\|M_n\|^{2p}\right] + \sum_{k=1}^{p} \binom{p}{k} 3^{2k-1}\left(L_k + L_1^k\right) \mathbb{E}\left[\|M_n\|^{2p-2k}\right] \\ &\quad + \sum_{k=1}^{p} \binom{p}{k} 3^{2k-1} 2^{2k} C^{2k} \mathbb{E}\left[\|Z_{n+1} - m\|^{2k} \|M_n\|^{2p-2k}\right]\end{aligned}$$

By induction,

$$\sum_{k=1}^{p} \binom{p}{k} 3^{2k-1}\left(L_k + L_1^k\right) \mathbb{E}\left[\|M_n\|^{2p-2k}\right] \leq \sum_{k=1}^{p} \binom{p}{k} 3^{2k-1}\left(L_k + L_1^k\right) C_{p-k} n^{p-k} = O\left(n^{p-1}\right).$$

Moreover, since for all positive real number a and for all positive integer q, $a \leq 1 + a^q$, applying Hölder's inequality and by induction, let

$$\begin{aligned}(\star) :&= \sum_{k=1}^{p} \binom{p}{k} 3^{2k-1} 2^{2k} C^{2k} \mathbb{E}\left[\|Z_{n+1} - m\|^{2k} \|M_n\|^{2p-2k}\right] \\ &\leq \sum_{k=1}^{p} \binom{p}{k} 3^{2k-1} 2^{2k} C^{2k} \mathbb{E}\left[\|M_n\|^{2p-2k}\right] + \sum_{k=1}^{p} \binom{p}{k} 3^{2k-1} 2^{2k} C^{2k} \mathbb{E}\left[\|Z_{n+1} - m\|^{2qk} \|M_n\|^{2p-2k}\right] \\ &\leq \sum_{k=1}^{p} \binom{p}{k} 3^{2k-1} 2^{2k} C^{2k} \left(\mathbb{E}\left[\|Z_{n+1} - m\|^{2qp}\right]\right)^{\frac{k}{p}} \left(\mathbb{E}\left[\|M_n\|^{2p}\right]\right)^{\frac{2p-2k}{2p}} + O\left(n^{p-1}\right).\end{aligned}$$

Note that $\left(\mathbb{E}\left[\|M_n\|^{2p}\right]\right)^{\frac{2p-2k}{2p}} \leq 1+\mathbb{E}\left[\|M_n\|^{2p}\right]$. Thus, taking $q \geq 2$ and applying Theorem 4.3, there are positive constants C_0, C_1' such that

$$\begin{aligned}(\star) &\leq \sum_{k=1}^{p}\binom{p}{k} 3^{2k-1}2^{2k}C^{2k}\left(K_{qp}\right)^{\frac{k}{p}}\frac{1}{n^{qk\alpha}}\left(1+\mathbb{E}\left[\|M_n\|^{2p}\right]\right)+O\left(n^{p-1}\right)\\ &\leq C_0\gamma_n^2\mathbb{E}\left[\|M_n\|^{2p}\right]+C_1'n^{p-1}.\end{aligned}$$

Finally, there are positive constants C_0, C_1 such that

$$\mathbb{E}\left[\left(\|M_n\|^2+\|\xi_{n+2}\|^2\right)^p\right] \leq \left(1+C_0\gamma_n^2\right)\mathbb{E}\left[\|M_n\|^{2p}\right]+C_1n^{p-1}. \tag{6.11}$$

Moreover, since (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$ and applying Lemma E.2,

$$\begin{aligned}2\mathbb{E}\left[\langle M_n,\xi_{n+2}\rangle\left(\|M_n\|^2+\|\xi_{n+2}\|^2\right)^{p-1}\right] &= 2\sum_{k=1}^{p-1}\binom{p-1}{k}\mathbb{E}\left[\langle M_n,\xi_{n+2}\rangle\|\xi_{n+2}\|^{2k}\|M_n\|^{2p-2-2k}\right]\\ &\leq \sum_{k=1}^{p-1}\binom{p-1}{k}\mathbb{E}\left[\|\xi_{n+2}\|^{2k+2}\|M_n\|^{2p-2-2k}\right]\\ &\quad+\sum_{k=1}^{p-1}\binom{p-1}{k}\mathbb{E}\left[\|\xi_{n+2}\|^{2k}\|M_n\|^{2p-2k}\right]\end{aligned}$$

Since $p \geq 2$ and by induction, as for $(\star)$, one can check that there are positive constants C_0', C_1' such that for all $n \geq 1$,

$$2\mathbb{E}\left[\langle M_n,\xi_{n+2}\rangle\left(\|M_n\|^2+\|\xi_{n+2}\|^2\right)^{p-1}\right] \leq C_0'\gamma_n^2\mathbb{E}\left[\|M_n\|^{2p}\right]+C_1'n^{p-1}. \tag{6.12}$$

Moreover, let

$$\begin{aligned}(\star\star) :&= \sum_{k=2}^{p}\binom{p}{k}2^k\mathbb{E}\left[\|M_n\|^k\|\xi_{n+2}\|^k\left(\|M_n\|^2+\|\xi_{n+2}\|^2\right)^{p-k}\right]\\ &\leq \sum_{k=2}^{p}\binom{p}{k}2^{p-1}\mathbb{E}\left[\|\xi_{n+2}\|^k\|M_n\|^{2p-k}\right]+\sum_{k=2}^{p}\binom{p}{k}2^{p-1}\mathbb{E}\left[\|M_n\|^k\|\xi_{n+2}\|^{2p-k}\right].\end{aligned}$$

We now bound the two terms on the right-hand side of previous inequality. First, let

$$\begin{aligned}(\star\star') &:= \sum_{k=2}^{p}\binom{p}{k}2^{p-1}\mathbb{E}\left[\|M_n\|^k\|\xi_{n+2}\|^{2p-k}\right]\\ &\leq \sum_{k=2}^{p}\binom{p}{k}2^{p-3}\mathbb{E}\left[\left(\|M_n\|^2+\|M_n\|^{2k-2}\right)\left(\|\xi_{n+2}\|^{2p-2k+2}+\|\xi_{n+2}\|^{2p-2}\right)\right]\end{aligned}$$

As for $(\star)$, one can check that there are positive constants C_0'', C_1'' such that for all $n \geq 1$,

$$(\star\star') \leq C_0'' \gamma_n^2 \mathbb{E}\left[\|M_n\|^{2p}\right] + C_1'' n^{p-1}.$$

In the same way, let

$$\begin{aligned}(\star\star'') &:= \sum_{k=2}^{p} \binom{p}{k} 2^{p-1} \mathbb{E}\left[\|\xi_{n+2}\|^k \|M_n\|^{2p-k}\right] \\ &\leq \sum_{k=2}^{p} \binom{p}{k} 2^{p-3} \mathbb{E}\left[\left(\|\xi_{n+2}\|^2 + \|\xi_{n+2}\|^{2k-2}\right)\left(\|M_n\|^{2p-2k+2} + \|M_n\|^2\right)\right]\end{aligned}$$

As for $(\star)$, there are positive constants C_0''', C_1''' such that

$$(\star\star'') \leq C_0''' \gamma_n^2 \mathbb{E}\left[\|M_n\|^{2p}\right] + C_1''' n^{p-1},$$

and in a particular case

$$(\star\star) \leq \left(C_0'' + C_0'''\right) \gamma_n^2 \mathbb{E}\left[\|M_n\|^{2p}\right] + \left(C_1'' + C_1'''\right) n^{p-1}. \tag{6.13}$$

Thus, thanks to inequalities (6.11) to (6.13), there are positive constants B_0, B_1 such that for all $n \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\|M_{n+1}\|^{2p}\right] &\leq \left(1 + B_0\gamma_n^2\right) \mathbb{E}\left[\|M_n\|^{2p}\right] + B_1 n^{p-1} \\ &\leq \left(\prod_{k=1}^{\infty}\left(1 + B_0\gamma_k^2\right)\right) \mathbb{E}\left[\|M_1\|^{2p}\right] + \left(\prod_{k=1}^{\infty}\left(1 + B_0\gamma_k^2\right)\right) \sum_{k=1}^{n} B_1 k^{p-1} \\ &\leq \left(\prod_{k=1}^{\infty}\left(1 + B_0\gamma_k^2\right)\right) \mathbb{E}\left[\|M_1\|^{2p}\right] + \left(\prod_{k=1}^{\infty}\left(1 + B_0\gamma_k^2\right)\right) B_1 n^p,\end{aligned}$$

which concludes the induction and the proof. □

Appendix A. Proofs of Propositions 2.3 and 2.6

Proof of Proposition 2.3. If $h \in \mathcal{B}(m, \epsilon)$, under assumptions **(A2)** and **(A3)** and by dominated convergence,

$$\langle \Phi(h), h - m \rangle = \left\langle \int_0^1 \Gamma_{m+t(h-m)}(h-m)\mathrm{d}t, h - m \right\rangle \geq \frac{\lambda_{\min}}{2} \|h - m\|^2.$$

In the same way, if $\|h - m\| > \epsilon$, since G is convex, under assumptions **(A2)** and **(A3)** and by dominated convergence,

$$\begin{aligned}\langle \Phi(h), h - m \rangle = \left\langle \int_0^1 \Gamma_{m+t(h-m)}(h-m)\mathrm{d}t, h - m \right\rangle &\geq \int_0^{\frac{\epsilon}{\|h-m\|}} \left\langle \Gamma_{m+t(h-m)}(h-m), h-m \right\rangle \mathrm{d}t \\ &\geq \frac{\lambda_{\min}\epsilon}{2} \|h - m\|.\end{aligned}$$

Thus, let A be a positive constant and $h \in \mathcal{B}(m, A)$,

$$\langle \Phi(h), h - m \rangle \geq c_A \|h - m\|^2,$$

with $c_A := \min\left\{\frac{\lambda_{\min}}{2}, \frac{\lambda_{\min}\epsilon}{2A}\right\}$. We now give an upper bound of this term. First, thanks to assumption **(A2)**, let A be a positive constant, for all $h \in \mathcal{B}(m, A)$,

$$\langle \Phi(h), h - m \rangle = \int_0^1 \langle \Gamma_{m+t(h-m)}(h - m), h - m \rangle \, \mathrm{d}t \leq \int_0^1 \left\|\Gamma_{m+t(h-m)}\right\|_{op} \|h - m\|^2 \, \mathrm{d}t \leq C_A \|h - m\|^2.$$

Moreover, applying Cauchy-Schwarz's inequality and thanks to assumption **(A5a)**, for all $h \in H$ such that $\|h - m\| \geq A$,

$$|\langle \Phi(h), h - m \rangle| \leq \sqrt{L_1} \|h - m\| + C \|h - m\|^2 \leq \left(\frac{\sqrt{L_1}}{A} + C\right) \|h - m\|^2,$$

which concludes the proof. □

Proof of Proposition 2.6. Let us recall that there are positive constants ϵ, C_ϵ such that for all $h \in \mathcal{B}(m, \epsilon)$,

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

Let $h \in H$ such that $\|h - m\| \geq \epsilon$. Then, thanks to assumptions **(A2)** and **(A3)**,

$$\begin{aligned} \|\Phi(h) - \Gamma_m(h - m)\| &\leq \|\Phi(h)\| + \|\Gamma_m\|_{op} \|h - m\| \\ &\leq (\mathbb{E}[f(X, h)] + C \|h - m\|) + C_0 \|h - m\| \\ &\leq \left(\frac{\sqrt{L_1}}{\epsilon^2} + \frac{C}{\epsilon} + \frac{C_0}{\epsilon}\right) \|h - m\|^2, \end{aligned}$$

which concludes the proof. □

Appendix B. Proof of Lemma 6.5

We propose here a not detailed proof. For analogous and more detailed calculus, one can see the proof of Lemma 5.3 in [6].

Proof of Lemma 6.5. We prove Lemma 6.5 with the help of a strong induction on p. The case $p = 1$ is already done in the proof of Theorem 3.1. We suppose from now that $p \geq 2$ and that for all $k \leq p - 1$, there is a positive constant M_k such that for all $n \geq 1$,

$$\mathbb{E}\left[\|Z_n - m\|^{2k}\right] \leq M_k.$$

Let $V_n := Z_n - m - \gamma_n \Phi(Z_n)$, and with the help of decomposition (6.2)

$$\begin{aligned} \|Z_{n+1} - m\|^2 &= \|V_n\|^2 + \gamma_n^2 \|\xi_{n+1}\|^2 + 2\gamma_n \langle V_n, \xi_{n+1} \rangle \\ &\leq \|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2 + 2\gamma_n \langle Z_n - m, \xi_{n+1} \rangle. \end{aligned}$$

Thus, applying Cauchy-Schwarz's inequaltiy

$$\|Z_{n+1}-m\|^{2p} \leq \left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^p + 2p\gamma_n\langle Z_n-m,\xi_{n+1}\rangle\left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p-1} + \sum_{k=2}^{p}\binom{p}{k}2^k\gamma_n^k\|Z_n-m\|^k\|\xi_{n+1}\|^k\left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p-k}. \tag{B.1}$$

Applying Lemma E.2, for all positive integer k,

$$\|U_{n+1}\|^k \leq 2^{k-1}f(X_{n+1},Z_n)+2^{k-1}C^k\|Z_n-m\|^k \quad a.s, \tag{B.2}$$

$$\|\xi_{n+1}\|^k \leq 3^{k-1}f(X_{n+1},Z_n)^k+3^{k-1}2^kC^k\|Z_n-m\|^k+3^{k-1}L_1^{\frac{k}{2}} \quad a.s. \tag{B.3}$$

Moreover, since $\langle\Phi(Z_n),Z_n-m\rangle\geq 0$,

$$\|V_n\|^2 \leq \left(1+2C^2\gamma_n^2\right)\|Z_n-m\|^2+2\gamma_n^2L_1 \tag{B.4}$$

We now bound each term on the right-hand side of inequality (B.1).

Bounding $(*):=\mathbb{E}\left[\left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^p\right]$**.** Applying inequality (B.2),

$$(*) \leq \mathbb{E}\left[\|V_n\|^{2p}\right]+\sum_{k=1}^{p}\binom{p}{k}\gamma_n^{2p}2^{2p-2}L_1^{p-k}\mathbb{E}\left[\mathbb{E}\left[f(X_{n+1},Z_n)^{2k}|\mathcal{F}_n\right]+C^{2k}\|Z_n-m\|^{2k}\right] + \sum_{k=1}^{p}\binom{p}{k}\gamma_n^{2k}2^{p+k-2}\left(1+2C^2c_\gamma^2\right)^{p-k}\mathbb{E}\left[\|Z_n-m\|^{2p-2k}\left(\mathbb{E}\left[f(X_{n+1},Z_n)^{2k}|\mathcal{F}_n\right]+C^{2k}\|Z_n-m\|^{2k}\right)\right].$$

Moreover, thanks to assumption **(A5b)** and by induction, there are positive constants A_0,A_1 such that

$$(*) \leq \mathbb{E}\left[\|V_n\|^{2p}\right]+A_0\gamma_n^2\mathbb{E}\left[\|Z_n-m\|^{2p}\right]+A_1\gamma_n^2. \tag{B.5}$$

Thanks to inequality (B.4) and by induction,

$$\begin{aligned}\mathbb{E}\left[\|V_n\|^{2p}\right] &\leq \left(1+2C^2\gamma_n^2\right)^p\mathbb{E}\left[\|Z_n-m\|^{2p}\right]+\sum_{k=1}^{p}\binom{p}{k}\left(1+2C^2\gamma_n^2\right)^{p-k}2^kL_1^k\gamma_n^{2k}\mathbb{E}\left[\|Z_n-m\|^{2p-2k}\right]\\ &\leq \left(1+2C^2\gamma_n^2\right)^p\mathbb{E}\left[\|Z_n-m\|^{2p}\right]+O\left(\gamma_n^2\right).\end{aligned}$$

Then, there are positive constants A_2,A_3 such that

$$(*) \leq \left(1+A_2\gamma_n^2\right)\mathbb{E}\left[\|Z_n-m\|^{2p}\right]+A_3\gamma_n^2. \tag{B.6}$$

Bounding $(**):=2p\gamma_n\mathbb{E}\left[\langle\xi_{n+1},Z_n-m\rangle\left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p-1}\right]$**.** Since (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$, and since V_n is $\mathcal{F}_n$-measurable, and applying inequalities (B.2) to (B.4), and by induction, one can check that there are positive constants A_1',A_2' such that

$$(**) \leq A_1'\gamma_n^3\mathbb{E}\left[\|Z_n-m\|^{2p}\right]+A_2'\gamma_n^3. \tag{B.7}$$

Bounding $(***) := \sum_{k=2}^{p} \binom{p}{k} 2^k \gamma_n^k \mathbb{E}\left[\|Z_n - m\|^k \|\xi_{n+1}\|^k \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-k}\right]$. Applying Lemma E.2, and inequalities (B.2) to (B.4) and by induction, one can check that there are positive constants A_1'', A_2'' such that

$$(***) \leq A_1'' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + A_2'' \gamma_n^2. \tag{B.8}$$

Conclusion. Applying inequalities (B.6) to (B.8) and by induction, there are positive constants B_1, B_2 such that

$$\begin{aligned}
\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right] &\leq \left(1 + B_1\gamma_n^2\right) \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + B_2\gamma_n^2 \\
&\leq \left(\prod_{k=1}^{\infty} \left(1 + B_1\gamma_k^2\right)\right) \mathbb{E}\left[\|Z_1 - m\|^{2p}\right] + B_2 \left(\prod_{k=1}^{\infty} \left(1 + B_1\gamma_k^2\right)\right) \sum_{k=1}^{\infty} \gamma_k^2 \\
&\leq M_p,
\end{aligned}$$

which concludes the induction and the proof.

□

Appendix C. Proof of Lemma 6.4

We propose here a not detailed proof. For analogous and more detailed calculus, one can see the proof of Lemma 4.2 in [14].

Proof of Lemma 6.4. Let $p \geq 1$, we suppose from now that for all integer $k < p$, there is a positive constant K_k such that for all $n \geq 1$,

$$\mathbb{E}\left[\|Z_n - m\|^{2k}\right] \leq \frac{K_k}{n^{k\alpha}}. \tag{C.1}$$

As in the previous proof, let us recall that

$$\begin{aligned}
\|Z_{n+1} - m\|^{2p+2} \leq \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p+1} &+ 2(p+1)\gamma_n \langle Z_n - m, \xi_{n+1}\rangle \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p} \\
&+ \sum_{k=2}^{p+1} \binom{p+1}{k} 2^k \gamma_n^k \|Z_n - m\|^k \|\xi_{n+1}\|^k \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p+1-k}.
\end{aligned} \tag{C.2}$$

We now bound the expectation of each term on the right-hand side of previous inequality.

Bounding $(**) := \mathbb{E}\left[2(p+1)\gamma_n \langle Z_n - m, \xi_{n+1}\rangle \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p}\right]$. Since (ξ_n) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$, and applying inequalities (B.2) and (B.3), and thanks to asumtpion **(A5b)** as well as inequality (C.1), one can check that there are positive constants A_1, A_2, A_3 such that

$$(**) \leq A_1 \gamma_n^3 \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + A_2 \gamma_n^3 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{A_3}{n^{(p+2)\alpha}}. \tag{C.3}$$

Bounding $(***) := \sum_{k=2}^{p+1} \binom{p+1}{k} 2^k \gamma_n^k \mathbb{E}\left[\|Z_n - m\|^k \|\xi_{n+1}\|^k \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p+1-k}\right]$. First, thanks to inequality (B.3) and Lemma E.2, one can check that there are positive constants A_1', A_2', A_3' such that

$$(***) \leq A_1' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + A_2' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{A_3'}{n^{(p+2)\alpha}}. \tag{C.4}$$

Thus, applying inequalities (C.4) to (C.6), there are positive constants B_0, B_1, B_2 such that

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1} - m\|^{2p+2}\right] \leq{}& \mathbb{E}\left[\left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p+1}\right] + B_0 \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] \\ &+ B_1 \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{B_2}{n^{(p+2)\alpha}}.\end{aligned} \tag{C.5}$$

Bounding $(*) := \mathbb{E}\left[\left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p+1}\right]$. As in the proof of Lemma 6.5, and thanks to induction inequality (C.1), there are positive constants A_0, A_0', A_0'' such that

$$(*) \leq \mathbb{E}\left[\|V_n\|^{2p+2}\right] + A_0 \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + A_0' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{A_0''}{n^{(p+2)\alpha}}. \tag{C.6}$$

Then, in order to conclude the proof, we just have to bound $\mathbb{E}\left[\|V_n\|^{2p}\right]$. Applying Proposition 2.1, one can check that there is a positive constant c and a rank n_α' such that for all $n \geq n_\alpha'$,

$$C \|Z_n - m\|^2 \mathbb{1}_{\{\|Z_n - m\| \leq c n^{1-\alpha}\}} \geq \langle \Phi(Z_n), Z_n - m \rangle \mathbb{1}_{\{\|Z_n - m\| \leq c n^{1-\alpha}\}} \geq \frac{4}{c_\gamma n^{1-\alpha}} \|Z_n - m\|^2 \mathbb{1}_{\{\|Z_n - m\| \leq c n^{1-\alpha}\}}.$$

Then, since $\|\Phi(Z_n)\|^2 \leq 2C^2 \|Z_n - m\|^2 + 2L_1 \gamma_n^2$, there is a rank n_α'' such that for all $n \geq n_\alpha''$,

$$\|Z_n - m - \gamma_n \Phi(Z_n)\|^2 \mathbb{1}_{\{\|Z_n - m\| \leq c n^{1-\alpha}\}} \leq \left(1 - \frac{3}{n}\right) \|Z_n - m\|^2 \mathbb{1}_{\{\|Z_n - m\| \leq c n^{1-\alpha}\}} + 2L_1 \gamma_n^2.$$

Then, one can check that there are positive constants A_1''', A_2''' such that

$$\begin{aligned}&\mathbb{E}\left[\|Z_n - m - \gamma_n \Phi(Z_n)\|^{2p+2} \mathbb{1}_{\{\|Z_n - m\| \leq c n^{1-\alpha}\}}\right] \\ &\leq \left(1 - \frac{3}{n}\right)^{p+1} \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + A_1''' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{A_2'''}{n^{(p+2)\alpha}}.\end{aligned}$$

Moreover, applying Cauchy-Schwarz's inequality, Markov's inequality and Lemma 6.5, for all positive integer q,

$$\begin{aligned}\mathbb{E}\left[\|Z_n - m\|^{2p+2} \mathbb{1}_{\{\|Z_n - m\| \geq c n^{1-\alpha}\}}\right] &\leq \sqrt{\mathbb{E}\left[\|Z_n - m\|^{4p+4}\right]} \sqrt{\mathbb{P}\left[\|Z_n - m\| \geq c n^{1-\alpha}\right]} \\ &\leq \sqrt{M_{2p+2}} \frac{\sqrt{M_q}}{c^q n^{q(1-\alpha)}},\end{aligned}$$

and one can conclude the proof applying inequality (C.5), taking $q \geq \frac{(p+2)\alpha}{1-\alpha}$ and taking a rank n_α such that for all $n \geq n_\alpha$, $\left(1 - \frac{3}{n}\right)^{p+1} + \left(B_0 + A_1'''\right) \gamma_n^2 \leq \left(1 - \frac{2}{n}\right)^{p+1}$. □

Remark C.1. Note that in order to get the rate of convergence in quadratic mean of the Robbins-Monro algorithm, *i.e.* in the case where $p = 1$, we just have to suppose that there are a positive integer $q \geq \frac{3\alpha}{1-\alpha}$ and a positive constant L_q such that for all $h \in H$, $\mathbb{E}\left[f(X,h)^{2q}\right] \leq L_q$.

Appendix D. Proof of Lemma 6.3

We propose here a not detailed proof. For analogous and more detailed calculus, one can see Lemma 4.1 in [14].

Proof of Lemma 5.2. Let $p \geq 1$, we suppose from now that for all integer $k < p$, there is a positive constant K_k such that for all $n \geq 1$,

$$\mathbb{E}\left[\|Z_n - m\|^{2k}\right] \leq \frac{K_k}{n^{k\alpha}}. \tag{D.1}$$

Using decomposition (6.3) and Cauchy-Schwarz's inequality, there are a positive constant c' and a rank n'_α such that for all $n \geq n'_\alpha$,

$$\|Z_{n+1} - m\|^2 \leq (1 - c'\gamma_n)\|Z_n - m\|^2 + \gamma_n^2\|U_{n+1}\|^2 + 2\gamma_n\langle Z_n - m, \xi_{n+1}\rangle + 2\gamma_n\|Z_n - m\|\|\delta_n\|.$$

If $p = 1$, thanks to Proposition 2.6, we have

$$2\|\delta_n\|\|Z_n - m\| \leq \frac{c'}{2}\gamma_n\|Z_n - m\|^2 + 2\frac{C_m^2}{c'}\|Z_n - m\|^4,$$

and since (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$, applying inequality (B.2), for all $n \geq n'_\alpha$,

$$\mathbb{E}\left[\|Z_{n+1} - m\|^2\right] \leq \left(1 - \frac{c'}{2}\gamma_n + 2C^2\gamma_n^2\right)\mathbb{E}\left[\|Z_n - m\|^2\right] + 2\gamma_n^2 L_1 + 2\gamma_n\frac{C_m^2}{c'}\mathbb{E}\left[\|Z_n - m\|^4\right],$$

and one can conclude the proof for $p = 1$ taking a rank n_α and a positive constant c such that for all $n \geq n_\alpha$, $1 - \frac{c'}{2}\gamma_n + 2C^2\gamma_n^2 \leq 1 - c\gamma_n$.

We suppose from now that $p \geq 2$. For all $n \geq n'_\alpha$,

$$\begin{aligned}\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right] \leq{}& (1 - c'\gamma_n)\,\mathbb{E}\left[\|Z_n - m\|^2\|Z_{n+1} - m\|^{2p-2}\right] + 2\gamma_n\mathbb{E}\left[\|Z_n - m\|\|\delta_n\|\|Z_{n+1} - m\|^{2p-2}\right] \\ &+ \gamma_n^2\mathbb{E}\left[\|U_{n+1}\|^2\|Z_{n+1} - m\|^{2p-2}\right] + 2\gamma_n\mathbb{E}\left[\langle Z_n - m, \xi_{n+1}\rangle\|Z_{n+1} - m\|^{2p-2}\right].\end{aligned} \tag{D.2}$$

We now bound each term which appear on the right-hand side of inequality (D.2) when we replace $\|Z_{n+1} - m\|^{2p-2}$ by the bound given by inequality (C.2).

Bounding $(\star) := (1 - c'\gamma_n)\,\mathbb{E}\left[\|Z_n - m\|^2\|Z_{n+1} - m\|^{2p-2}\right]$**.** First, applying inequality (B.4)

$$\begin{aligned}(*) :&= (1 - c'\gamma_n)\,\mathbb{E}\left[\|Z_n - m\|^2\left(\|V_n\|^2 + \gamma_n^2\|U_{n+1}\|^2\right)^{p-1}\right] \\ &\leq (1 - c'\gamma_n)\left(1 + 2C^2\gamma_n^2\right)^{p-1}\mathbb{E}\left[\|Z_n - m\|^{2p}\right] \\ &\quad + \sum_{k=0}^{p-2}\binom{p-1}{k}(1 - c'\gamma_n)\,\gamma_n^{2(p-1-k)}\left(1 + 2C^2c_\gamma^2\right)^k\mathbb{E}\left[\|Z_n - m\|^{2k+2}\left(2L_1 + \|U_{n+1}\|^2\right)^{p-1-k}\right].\end{aligned}$$

Applying inequalities (B.2) and (D.1), thanks to assumption **(A5b)** and since for all $n \geq n_\alpha$ we have $1 - c'\gamma_n \leq 1$, and for all $k \leq p-2$, we have $2p-1-k \ \geq p+1$, one can check that there is a positive constant A_1 such that

$$(*) \leq \left(1 - c'\gamma_n + A_1\gamma_n^2\right) \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + O\left(\frac{1}{n^{(p+1)\alpha}}\right). \tag{D.3}$$

In the same way, since (ξ_n) is a sequence of martingale differences adapted to $(\mathcal{F}_n)$, applying Cauchy-Schwarz's inequality, as well as inequalities (D.1), (B.2) to (B.4), one can check that there is a positive constant A_2 such that

$$\begin{aligned}(*)' :&= 2(p-1)\left(1 - c'\gamma_n\right)\gamma_n \mathbb{E}\left[\|Z_n - m\|^2 \langle Z_n - m, \xi_{n+1}\rangle \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-2}\right] \\ &\leq A_2\gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + O\left(\frac{1}{n^{(p+1)\alpha}}\right). \end{aligned} \tag{D.4}$$

In the same way, applying inequalities (B.4) and (D.1), with analogous calculus to the previous ones, one can check that there are positive constants A_3, A_4 such that

$$\begin{aligned}(*)'' :&= \left(1 - c'\gamma_n\right) \sum_{k=2}^{p-1} \binom{p-1}{k} 2^k \gamma_n^k \mathbb{E}\left[\|Z_n - m\|^k \|\xi_{n+1}\|^k \|Z_n - m\|^2 \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-1-k}\right] \\ &\leq A_3\gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + A_4\gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + O\left(\frac{1}{n^{(p+1)\alpha}}\right). \end{aligned} \tag{D.5}$$

Finally, applying inequalities (D.3) to (D.5), there are positive constants B_0, B_1, B_2 such that

$$(\star) \leq \left(1 - c'\gamma_n + B_0\gamma_n^2\right) \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + B_2\gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{B_1}{n^{(p+1)\alpha}}.$$

Bounding $(\star\star) := 2\gamma_n \mathbb{E}\left[\|Z_n - m\| \|\delta_n\| \|Z_{n+1} - m\|^{2p-2}\right]$**.** First, let

$$\begin{aligned}(*) :&= 2\gamma_n \mathbb{E}\left[\|Z_n - m\| \|\delta_n\| \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-1}\right] \\ &\leq 2^{p-1}\gamma_n \mathbb{E}\left[\|Z_n - m\| \|\delta_n\| \|V_n\|^{2p-2}\right] + 2^{p-1}\gamma_n^{2p-1} \mathbb{E}\left[\|Z_n - m\| \|\delta_n\| \|U_{n+1}\|^{2p-2}\right]. \end{aligned}$$

Moreover, thanks to Proposition 2.6 and inequalities (B.4), (B.2), and (D.1), one can check that there are positive constants A_1, A_2, A_3 such that

$$(*) \leq \left(\frac{c'}{4}\gamma_n + A_1\gamma_n^2\right) \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + A_2\gamma_n \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{A_3}{n^{(p+1)\alpha}}. \tag{D.6}$$

Since (ξ_n) is a sequence of martingale differences, let

$$(*)' := 4(p-1)\gamma_n^2\mathbb{E}\left[\|Z_n-m\|\,\|\delta_n\|\,\langle Z_n-m,\xi_{n+1}\rangle\left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p-2}\right]$$
$$= 4(p-1)\sum_{k=1}^{p-2}\binom{p-2}{k}\gamma_n^{2k+2}\mathbb{E}\left[\|Z_n-m\|\,\|\delta_n\|\,\langle Z_n-m,\xi_{n+1}\rangle\,\|V_n\|^{2(p-2-k)}\,\|U_{n+1}\|^{2k}\right].$$

Thanks to Proposition 2.6 and inequalities (D.1), (B.2) and (B.3), one can check that there are positive constants A_1', A_2', A_3' such that

$$(*)' \leq A_1'\gamma_n^2\mathbb{E}\left[\|Z_n-m\|^{2p}\right]+A_2'\gamma_n^2\mathbb{E}\left[\|Z_n-m\|^{2p+2}\right]+\frac{A_3'}{n^{(p+1)\alpha}}. \tag{D.7}$$

Finally, let

$$(*)'' := 2\gamma_n\mathbb{E}\left[\|Z_n-m\|\,\|\delta_n\|\sum_{k=2}^{p-1}\binom{p-1}{k}2^k\gamma_n^k\,\|Z_n-m\|^k\,\|\xi_{n+1}\|^k\left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p-1-k}\right].$$

With similar calculus, applying inequalities (D.1), (B.2) and (B.3), one can check that there are positive constants A_0'', A_1'', A_2'' such that

$$(*)'' \leq A_0''\gamma_n^2\mathbb{E}\left[\|Z_n-m\|^{2p}\right]+A_1''\gamma_n^2\mathbb{E}\left[\|Z_n-m\|^{2p+2}\right]+\frac{A_2''}{n^{(p+1)\alpha}}. \tag{D.8}$$

Finally, applying inequalities (D.6) to (D.8), there are positive constants B_0', B_1', B_2' such that

$$(\star\star) \leq \left(\frac{1}{4}c'\gamma_n+B_0'\gamma_n^2\right)\mathbb{E}\left[\|Z_n-m\|^{2p}\right]+B_1'\gamma_n\mathbb{E}\left[\|Z_n-m\|^{2p+2}\right]+\frac{B_2'}{n^{(p+1)\alpha}}.$$

Bounding $\gamma_n^2\mathbb{E}\left[\|U_{n+1}\|^2\,\|Z_{n+1}-m\|^{2p-2}\right]$. First, applying inequalities (D.1), (B.4), (B.2) and (B.3), there are positive constants A_0, A_1 such that

$$\gamma_n^2\mathbb{E}\left[\|U_{n+1}\|^2\left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p-1}\right] \leq A_0\gamma_n^2\mathbb{E}\left[\|Z_n-m\|^{2p}\right]+\frac{A_1}{n^{(p+1)\alpha}}. \tag{D.9}$$

Applying inequalities (D.1), (B.2) and (B.3), one can check that there are positive constants A_0', A_1' such that

$$\left|2(p-1)\gamma_n^3\mathbb{E}\left[\|U_{n+1}\|^2\,\langle Z_n-m,\xi_{n+1}\rangle\left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p-2}\right]\right| \leq A_0'\gamma_n^2\mathbb{E}\left[\|Z_n-m\|^{2p}\right]+\frac{A_1'}{n^{(p+1)\alpha}}. \tag{D.10}$$

Finally, let

$$(*)' := \sum_{k=2}^{p-1}\binom{p-1}{k}2^k\gamma_n^{k+2}\mathbb{E}\left[\|U_{n+1}\|^2\,\|Z_n-m\|^k\,\|\xi_{n+1}\|^k\left(\|V_n\|^2+\gamma_n^2\|U_{n+1}\|^2\right)^{p-1-k}\right]$$

Applying inequalities (D.1), (B.2) and (B.3), there are positive constants A_0'', A_1'' such that

$$(*)' \leq A_0'' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{A_1''}{n^{(p+1)\alpha}}. \tag{D.11}$$

Thus, applying inequalities (D.10) and (D.11), there are positive constants B_0'', B_1'' such that

$$\gamma_n^2 \mathbb{E}\left[\|U_{n+1}\|^2 \|Z_{n+1} - m\|^{2p-2}\right] \leq B_0'' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{B_1''}{n^{(p+1)\alpha}}.$$

Bounding $2\gamma_n \mathbb{E}\left[\langle Z_n - m, \xi_{n+1}\rangle \|Z_{n+1} - m\|^{2p-2}\right]$**.** First, since (ξ_n) is a martingale differences sequence adapted to the filtration $(\mathcal{F}_n)$, let

$$\begin{aligned}(*) :&= 2\gamma_n \mathbb{E}\left[\langle \xi_{n+1}, Z_n - m\rangle \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-1}\right] \\ &\leq \sum_{k=1}^{p-1} \binom{p-1}{k} \gamma_n^{2k+1} \mathbb{E}\left[\left(\|\xi_{n+1}\|^2 + \|Z_n - m\|^2\right) \|V_n\|^{2(p-1-k)} \|U_{n+1}\|^{2k}\right].\end{aligned}$$

Thus, applying inequalities (D.1), (B.2) and (B.3), one can check that there are positive constants A_0, A_1 such that

$$2\gamma_n \mathbb{E}\left[\langle \xi_{n+1}, Z_n - m\rangle \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-1}\right] \leq A_0 \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{A_1}{n^{(p+1)\alpha}}. \tag{D.12}$$

In the same way, since $p \geq 2$, applying inequalities (D.1), (B.2) and (B.3), one can check that there are positive constants A_0', A_1' such that

$$4(p-1)\gamma_n^2 \mathbb{E}\left[\langle Z_n - m, \xi_{n+1}\rangle^2 \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-2}\right] \leq A_0' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + \frac{A_1'}{n^{(p+1)\alpha}}. \tag{D.13}$$

Finally, applying inequalities (D.1), (B.2) and (B.3), one can check that there are positive constants A_0'', A_1'', A_2'' such that

$$\begin{aligned}(*)'' :&= 2\sum_{k=2}^{p-1} \binom{p-1}{k} \gamma_n^{k+1} \mathbb{E}\left[\langle Z_n - m, \xi_{n+1}\rangle \|Z_n - m\|^k \|\xi_{n+1}\|^k \left(\|V_n\|^2 + \gamma_n^2 \|U_{n+1}\|^2\right)^{p-k}\right] \\ &\leq A_0'' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + A_1'' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{A_2''}{n^{(p+1)\alpha}}.\end{aligned} \tag{D.14}$$

Then, applying inequalities (D.12) and (D.14), there are positive constants B_0''', B_1''', B_2''' such that

$$2\gamma_n \mathbb{E}\left[\langle Z_n - m, \xi_{n+1}\rangle \|Z_{n+1} - m\|^{2p-2}\right] \leq B_0''' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + B_1''' \gamma_n^2 \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{B_2'''}{n^{(p+1)\alpha}}.$$

Conclusion

We have proved that there are positive constants c_0, C_1, C_2 such that for all $n \geq n_\alpha'$;

$$\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right] \leq \left(1 - \frac{c'}{2}\gamma_n + c_0 \gamma_n^2\right) \mathbb{E}\left[\|Z_n - m\|^{2p}\right] + C_1 \gamma_n \mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{C_2}{n^{(p+1)\alpha}}.$$

Then, there are a positive constant c and a rank $n_\alpha \geq n'_\alpha$ such that for all $n \geq n_\alpha$, $1 - \frac{c'}{2}\gamma_n + c_0\,\gamma_n^2 \leq 1 - c\gamma_n$, and in a particular case, for all $n \geq n_\alpha$,

$$\mathbb{E}\left[\|Z_{n+1} - m\|^{2p}\right] \leq (1 - c\gamma_n)\,\mathbb{E}\left[\|Z_n - m\|^{2p}\right] + C_1\gamma_n\mathbb{E}\left[\|Z_n - m\|^{2p+2}\right] + \frac{C_2}{n^{(p+1)\alpha}}. \tag{D.15}$$

□

Appendix E. Some useful existing results

Let us recall Robbins-Siegmund theorem (see [12] for instance):

Theorem E.1. *[Robbins-Siegmund theorem] Let $(V_n), (A_n), (B_n), (C_n)$ be non negative random variables adapted to a filtration $(\mathcal{F}_n)$ such that*

$$\mathbb{E}\left[V_{n+1}|\mathcal{F}_n\right] \leq V_n\,(1 + A_n) + B_n - C_n.$$

Then, on $\Gamma = \left\{\sum_{n\geq 1} A_n < +\infty \text{ and } \sum_{n\geq 1} B_n < +\infty\right\}$, (V_n) converges almost surely to a finite random variable V_∞ and $\sum_{n\geq 1} C_n < +\infty$ almost surely.

Let us now recall Lemma A.1 in [14]:

Lemma E.2. *Let p, n be two positive integers and let $a_1, \ldots, a_n$ be positive constants. Then,*

$$\left(\sum_{j=1}^{n} a_j\right)^p = n^{p-1}\sum_{j=1}^{n} a_j^p.$$

References

[1] F. Bach, Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J. Mach. Learn. Res.* **15** (2014) 595–627.

[2] J.M. Borwein and P.B. Borwein, Pi and the AGM. Wiley, New York (1987).

[3] S. Boyd and L. Vandenberghe, Convex optimization. Cambridge University Press (2004).

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

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

[6] H. Cardot, P. Cénac, A. Godichon-Baggioni *et al.*, Online estimation of the geometric median in hilbert spaces: non asymptotic confidence balls. *Ann. Stat.* **45** (2017) 591–614.

[7] P. Chaudhuri, Multivariate location estimation using extension of R-estimates through U-statistics type approach. *Ann. Statist.* **20** (1992) 897–916.

[8] P. Chaudhuri, On a geometric notion of quantiles for multivariate data. *J. Am. Stat. Assoc.* **91** (1996) 862–872.

[9] Y. Chen, X. Dang, H. Peng and H.L. Bart, Outlier detection with the kernelized spatial depth function. *IEEE Trans. Pattern Anal. Mach. Intel.* **31** (2009) 288–305.

[10] K. Cohen, A. Nedic and R. Srikant, On projected stochastic gradient descent algorithm with weighted averaging for least squares regression. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016) 2314–2318.

[11] M. Duflo, Algorithmes stochastiques. Springer, Berlin (1996).

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

[13] S. Ghadimi and G. Lan, Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: a generic algorithmic framework. *SIAM J. Optim.* **22** (2012) 1469–1492.

[14] A. Godichon-Baggioni, Estimating the geometric median in hilbert spaces with stochastic gradient algorithms: Lp and almost sure rates of convergence. *J. Multivar. Anal.* **146** (2016) 209–222.

[15] A. Goh, C. Lenglet, P.M. Thompson and R. Vidal, A nonparametric Riemannian framework for processing high angular resolution diffusion images (HARDI). *IEEE Conference on Computer Vision and Pattern Recognition* (2009) 2496–2503.

[16] J.B.S. Haldane, Note on the median of a multivariate distribution. *Biometrika* **35** (1948) 414–417.
[17] M. Hallin and D. Paindaveine, Semiparametrically efficient rank-based inference for shape. I. Optimal rank-based tests for sphericity. *Ann. Stat.* **34** (2006) 2707–2756.
[18] P. Huber and E. Ronchetti, Robust Statistics, 2nd edn. John Wiley and Sons (2009).
[19] J. 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). North-Holland, Amsterdam (1987) 217–230.
[20] H.J. Kushner and G.G. Yin, Stochastic approximation and recursive algorithms and applications, Stochastic Modelling and Applied Probability. 2nd edn. Vol. 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York (2003).
[21] R.A. Maronna, R.D. Martin and V.J. Yohai, Robust statistics. Theory and methods. *Wiley Series in Probability and Statistics*. John Wiley & Sons, Ltd., Chichester (2006).
[22] E. Moulines and F.R. Bach, Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In Advances in Neural Information Processing Systems (2011) 451–459.
[23] M. Pelletier, On the almost sure asymptotic behaviour of stochastic algorithms. *Stoch. Process. Appl.* **78** (1998) 217–244.
[24] M. Pelletier, Asymptotic almost sure efficiency of averaged stochastic algorithms. *SIAM J. Control Optim.* **39** (2000) 49–72.
[25] G.H. Polya, G. Harold and Littlewood, Inequalities. University Press (1952).
[26] B. Polyak and A. Juditsky, Acceleration of stochastic approximation. *SIAM J. Cont. Optim.* **30** (1992) 838–855.
[27] H. Robbins and S. Monro, A stochastic approximation method. *Ann. Math. Stat.* (1951) 400–407.
[28] M. Rudelson, Recent developments in non-asymptotic theory of random matrices. Modern Aspects of Random Matrix Theory. In Vol. 72 of *Proceedings of Symposia in Applied Mathematics* (2014) 83–120.
[29] D. Ruppert, Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering (1988).
[30] R. Serfling, Depth functions in nonparametric multivariate inference. In Vol. 72 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* **72** (2006).
[31] P. Turaga, A. Veeraraghavan and R. Chellappa, Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision. *IEEE Conference on Computer Vision and Pattern Recognition* (2008).