Convergence in quadratic mean of averaged stochastic gradient algorithms without strong convexity nor bounded gradient

Antoine Godichon-Baggioni

Convergence in quadratic mean of averaged stochastic gradient algorithms without strong convexity nor bounded gradient

Antoine Godichon-Baggioni, antoine.godichon_baggioni@sorbonne-universite.fr
Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne-Université, 75005 Paris, France

Abstract

Online averaged stochastic gradient algorithms are more and more studied since (i) they can deal quickly with large sample taking values in high dimensional spaces, (ii) they enable to treat data sequentially, (iii) they are known to be asymptotically efficient. In this paper, we focus on giving explicit bounds of the quadratic mean error of the estimates, and this, with very weak assumptions, i.e without supposing that the function we would like to minimize is strongly convex or admits a bounded gradient.

1 Introduction

A usual problem in stochastic optimization and machine learning is, considering a random variable X, to estimate the minimizer of a convex function G of the form

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

where h lies in a separable Hilbert space $\mathcal{H}$. This problem is encountered when we estimate, for instance, the parameters of logistic regressions (Bach, 2014; Cohen et al., 2017), the geometric median and quantiles (Cardot et al., 2013; Godichon-Baggioni, 2016; Cardot et al., 2017), or superquantiles (Bercu et al., 2020; Costa and Gadat, 2020). Since the gradient or the Hessian of G cannot be explicitly calculated, one cannot apply usual optimization methods such that gradient or Newton algorithms to approximate the minimizer. A solution to overcome this problem, considering n i.i.d copies $X_1, \ldots, X_n$ of X, is to approximate the solution of the empirical function

$$G_n(h) = \frac{1}{n}\sum_{k=1}^{n} g(X_k,h).$$

Nevertheless, this often necessitates high computational costs when the dimension of $\mathcal{H}$ and the sample size are both large. In order to partially overcome this cost problem, one way is to focus on mini-batch gradient algorithms, i.e to consider iterative estimates of the

form

$$m_{t+1} = m_t - \gamma_t \sum_{i \in S_t} \nabla_h g\left(X_i, m_t\right)$$

where $S_t \subset \{1, \ldots, n\}$ is the mini-batch considered at time t (Konečnỳ et al., 2015; Alfarra et al., 2020). Nevertheless, these kinds of methods necessitate to store all the data into memory and do not enable to easily update the estimates if the data arrive sequentially. In order to address these problems, the online stochastic gradient algorithm introduced by Robbins and Monro (1951) should be preferred. Nevertheless, as mentioned in Pelletier (1998), the estimates obtained with this algorithm hardly ever attain the asymptotic efficiency. Fortunately, one can consider its averaged version introduced by Ruppert (1988) and Polyak and Juditsky (1992) which is known to be asymptotically efficient (Pelletier, 2000). In this paper, we focus on non asymptotic analysis of such estimates.

1.1 Related works

The rate of convergence in quadratic mean of averaged stochastic gradient algorithms in the case where G is strongly convex was given in Bach and Moulines (2013). Nevertheless, the loss of strong convexity generates several technical problems and makes the obtaining of non asymptotic results much more difficult. In recent works, Bach (2014) and Gadat and Panloup (2017) succeeded in obtaining the L^2 rates of convergence of the estimates but supposed for this that the gradient of g is bounded, which can be considered as restrictive. For instance, this is not verified in most of regressions if the eplicative variable is not bounded, or in the case of the recursive estimation of p means with $p \in (1, 2)$ (Godichon-Baggioni, 2019b). In Godichon-Baggioni (2019a), the gradient of g was not supposed to be bounded anymore, but it was assumed that it admits moments of any order. Furthermore, the upper bounds of the quadratic mean errors of the estimates at time n were not explicitly given. In addition, in Cardot et al. (2017), non asymptotic confidence balls were given in the case of the recursive estimation of the geometric median, but these balls where only available from a non calculated rank. Recently, Costa and Gadat (2020) focus on the use of stochastic gradient algorithms for superquantiles estimation and give uniform bounds of the quadratic mean error of the estimates. Nevertheless, here again, the bound depends on non calculated constants. Finally, in a recent work, Défossez et al. (2020) give simple proof for obtaining convergence results for some adaptive stochastic gradient methods.

1.2 Contribution

In this work, the aim is to give a very weak framework for each we are able to obtain explicit L^2 rates of convergence of stochastic gradient estimates and their averaged version. First, we replace usual strong convexity assumption by strict (or locally strong) convexity. Second we do not assume that the gradient of g is bounded or admits moments of any order, but we only suppose that it admits a fourth order moment. Finally, under weak assumptions,

we give explicit bounds of the quadratic mean errors of the estimates and prove that, up to a calculated rest term, the averaged estimates achieve the Cramer-Rao bound.

1.3 Notations

In this paper, we denote by $\|.\|$ the euclidean norm on $\mathcal{H}$, $\langle .,. \rangle$ the associated inner product, and $\|.\|_{op}$ the spectral norm of operators on $\mathcal{H}$. Remark that given $h, h' \in \mathcal{H}$, we will also write $\langle h, h' \rangle = h^T h'$. Furthermore, for all $h \in \mathcal{H}$ and $r > 0$, $\mathcal{B}(h, r) := \{h' \in \mathcal{H}, \|h - h'\| \leq r\}$. Finally, for any $x \in \mathbb{R}$, $\lceil x \rceil$ gives the superior integer part of x.

1.4 Paper organization

The paper is organized as follows: first the framework and assumptions are given and discussed in Section 2. The rate of convergence in quadratic mean of the stochastic gradient estimates are introduced in Section 3 while the ones for their averaged version are given in Section 4. Finally, the proofs of the convergence results for gradient estimates and their averaged version are respectively postponed in Sections 5 and 6.

2 Framework

In what follows, we consider a random variable X taking values in a measurable space $\mathcal{X}$ and let $\mathcal{H}$ be a separable Hilbert space (not necessarily of finite dimension). We focus on the estimation of the minimizer θ of the convex function $G : \mathcal{H} \longrightarrow \mathbb{R}$ defined for all $h \in \mathcal{H}$ by

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

with $g : \mathcal{X} \times \mathcal{H} \longrightarrow \mathbb{R}$. Throughout the suite, we will suppose that the following assumptions are fulfilled:

(A1) For almost every $x \in \mathcal{X}$, the functional $g(x, .)$ is differentiable on $\mathcal{H}$ and there are non-negative constants C_1, C_1', C_2, C_2' such that for all $h \in \mathcal{H}$,

$$\mathbb{E}\left[\|\nabla_h g\left(X, h\right)\|^2\right] \leq C_1 + C_2\left(G(h) - G(\theta)\right), \qquad \mathbb{E}\left[\|\nabla_h g\left(X, h\right)\|^4\right] \leq C_1' + C_2'\left(G(h) - G(\theta)\right)^2$$

(A2) The functional G is twice continuously differentiable and $\lambda_{\min} := \lambda_{\min}\left(\nabla^2 G(\theta)\right) > 0$.

(A3) The Hessian of G is uniformly bounded on $\mathcal{H}$, i.e there is a positive constant $L_{\nabla G}$ such that for all $h \in \mathcal{H}$,

$$\left\|\nabla^2 G(h)\right\|_{op} \leq L_{\nabla G}.$$

(A4) There are positive constants λ_0, r_{λ_0} and a non-negative constant C_{λ_0} such that $\forall h \in \mathcal{B}\left(\theta, r_{\lambda_0}\right)$,

$$\lambda_{\min}\left(\nabla^2 G\left(h\right)\right) \geq \lambda_0 \qquad \text{and} \qquad \left\|\nabla G(h) - \nabla^2 G(\theta)\left(h - \theta\right)\right\| \leq C_{\lambda_0}\|h - \theta\|^2$$

Remark that Assumption **(A1)** ensures that the functional G is differentiable. One of the main difference with Bach and Moulines (2013) and Gadat and Panloup (2017) is that they suppose that the gradient of g is uniformly bounded. Moreover, an important difference with Godichon-Baggioni (2019a) is that we only suppose that the moment of order four of the gradient exists instead of each moments. In addition, Assumption **(A2)** leads the functional G to be strictly convex, so that θ is its unique minimizer. Furthermore, Assumption **(A3)** ensures that the gradient of G is $L_{\nabla G}$-lipschitz. Finally, Assumption **(A4)** just means that there is a neighborhood of θ on each we have both locally strong convexity of G and a locally quadratic increasing of the rest term in the Taylor's expansion of the gradient (which is verified as soon as the Hessian of G is lipschitz on a neighborhood of θ). Remark that if $\mathcal{H}$ is a finite dimensional space, the local strong convexity was already given by **(A2)**. As a conclusion, these assumptions can be considered as weak compare to the existing ones in the literature on non-asymptotic results.

3 The stochastic gradient algorithm

In what follows, let us consider $X_1, \ldots, X_n, X_{n+1}, \ldots$ be i.i.d copies of X. The stochastic gradient algorithm is defined recursively for all $n \geq 0$ by (Robbins and Monro, 1951)

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \nabla_h g\left(X_{n+1}, \theta\right), \tag{1}$$

with θ_0 bounded. We consider from now a stepsequence (γ_n) of the form $\gamma_n = c_\gamma n^{-\alpha}$, where $c_\gamma > 0$ and $\alpha \in (1/2, 1)$.

3.1 Case with unbounded gradient

In this section, we focus on the case where $C_2 \neq 0$ or $C_2' \neq 0$. We first give the rate of convergence in quadratic mean of $G\left(\theta_n\right)$.

Lemma 3.1. *Suppose Assumptions* ***(A1)*** *to* ***(A4)*** *hold. Then,*

$$\mathbb{E}\left[\left(G\left(\theta_n\right) - G(\theta)\right)^2\right] \leq e^{-\frac{1}{4}c_\gamma a_0 n^{1-\alpha}} e^{2a_1 c_\gamma^2 \frac{2\alpha}{2\alpha-1} + 2a_2 c_\gamma^3 \frac{3\alpha}{3\alpha-1}} \left(u_0 + \sigma^2 c_\gamma^3 \frac{3\alpha}{3\alpha-1}\right) + \frac{2^{1+4\alpha}\sigma^2 c_\gamma^2}{a_0} n^{-2\alpha}$$

with $u_0 = \mathbb{E}\left[\left(G\left(\theta_0\right) - G(\theta)\right)^2\right]$, $a_0 = \frac{\lambda_0^2 \min\left\{1, r_{\lambda_0}^2\right\}}{L_{\nabla G}}$, $a_1 = \max\left\{\frac{\lambda_0^4}{4L_{\nabla G}^2}, C_2\left(4L_{\nabla G}+1\right)\right\}$, $a_2 = \frac{1}{2}L_{\nabla G}^2 C_2'$, *and* $\sigma^2 = \frac{C_1^2(4L_{\nabla G}+1)^2 L_{\nabla G}}{12\lambda_0^2 \min\left\{1, r_{\lambda_0}^2\right\}} + \frac{c_\gamma L_{\nabla G}^2 C_1'}{2}$.

In a simple way, this lemma ensures that we have the usual rate of convergence $\mathbb{E}\left[G\left(\theta_n\right)\right] - G(\theta) = O\left(n^{-\alpha}\right)$. This result is crucial to give the following rate of convergence in quadratic mean of the estimates θ_n.

Theorem 3.1. *Suppose assumptions* ***(A1)*** *to* ***(A4)*** *hold. Then,*

$$\mathbb{E}\left[\|\theta_n-\theta\|^2\right] \leq Ae^{-\frac{1}{4}\lambda_{\min}c_\gamma n^{1-\alpha}} + c_1\frac{2L_\delta^2}{\lambda_{\min}^2}e^{-\frac{1}{8}a_0c_\gamma n^{1-\alpha}} + \frac{2^{2+8\alpha}\sigma^2c_\gamma^2}{a_0}\frac{L_\delta^2}{\lambda_{\min}^2}n^{-2\alpha} + \frac{2^{1+\alpha}C_1}{\lambda_{\min}}c_\gamma n^{-\alpha}$$

with a_0, a_1, a_2, σ^2 *defined in Lemma 3.1,* $v_0 = \mathbb{E}\left[\|\theta_0-\theta\|^2\right]$, $L_\delta = \max\left\{\frac{2C_{\lambda_0}}{\lambda_0}, \frac{2L_{\nabla G}}{\lambda_0 r_{\lambda_0}}\right\}$, $b_1 = \frac{L_{\nabla G}}{2}\max\left\{C_2, \frac{\lambda_{\min}^2}{2L_{\nabla G}}\right\}$, $c_1 = \exp\left(2a_1c_\gamma^2\frac{2\alpha}{2\alpha-1} + 2a_2c_\gamma^3\frac{3\alpha}{3\alpha-1}\right)\left(v_0 + \sigma^2c_\gamma^3\frac{3\alpha}{3\alpha-1}\right)$ *and*

$$A = e^{2b_1c_\gamma^2\frac{2\alpha}{2\alpha-1}}\left(v_0 + \frac{2\alpha c_\gamma^2C_1}{2\alpha-1} + 2\frac{L_\delta^2}{\lambda_{\min}}\left(u_0c_\gamma + c_1 + \frac{4c_1}{a_0(1-\alpha)}e^{-\frac{1}{4}a_0c_\gamma} + \frac{2^{1+4\alpha}\sigma^2c_\gamma^3}{a_0}\frac{3\alpha}{3\alpha-1}\right)\right).$$

In other words, we get the usual rate of convergence $\mathbb{E}\left[\|\theta_n-\theta\|^2\right] = O\left(n^{-\alpha}\right)$ (Bach and Moulines, 2013; Gadat and Panloup, 2017; Godichon-Baggioni, 2019a) and so, with weak assumptions. Moreover, contrary to Gadat and Panloup (2017) and Godichon-Baggioni (2019a), we give an explicit boud of the quadratic mean error. Finally, note that for the main term, i.e $\frac{2^{1+\alpha}C_1}{\lambda_{\min}}c_\gamma n^{-\alpha}$, we succeed in obtaining a term analogous to the one in the strongly convex case given by Bach and Moulines (2013). Let us now discuss about the rest terms. The term $Ae^{-\frac{1}{4}\lambda_{\min}c_\gamma n^{1-\alpha}}$ can be seen as a quantification of the error due to the initialization while the term $c_1\frac{2L_\delta^2}{\lambda_{\min}^2}e^{-\frac{1}{8}a_0c_\gamma n^{1-\alpha}} + \frac{2^{2+8\alpha}\sigma^2c_\gamma^2}{a_0}\frac{L_\delta^2}{\lambda_{\min}^2}n^{-2\alpha}$ comes from the error approximation of $\nabla^2G(\theta)\left(\theta_n-\theta\right)$ by $\nabla G\left(\theta_n\right)$. Remark that in the particular case of the linear regression, $C_{\lambda_0} = 0$ for any r_{λ_0}. Moreover, one can take $r_{\lambda_0} = +\infty$ and $\lambda_0 = \lambda_{\min}$, which leads to $L_\delta = 0$ and to a bound analogous to the one in Bach and Moulines (2013).

3.2 Case with $\|\nabla G(.)\|$ bounded

Since in several cases such as logistic regression, softmax regression or the estimation of the geometric median one has $C_2 = C_2' = 0$, we now focus on this case to have more precise bounds. We first give the rate of convergence in quadratic mean of $G\left(\theta_n\right)$.

Lemma 3.2. *Suppose assumptions* ***(A1)*** *to* ***(A4)*** *hold. Then, for all* $n \geq 1$,

$$\mathbb{E}\left[\left(G\left(\theta_n\right)-G(\theta)\right)^2\right] \leq c_{n_0'}\exp\left(-\frac{1}{2}a_0c_\gamma n^{1-\alpha}\right) + \sigma^2M_0c_\gamma^2n^{-2\alpha}$$

with $n_0' = \inf\left\{n, a_0\gamma_{n+1} \leq 1\right\}$, $c_{n_0'} := \sigma^2\left(\exp\left(\frac{1}{2}a_0c_\gamma\left(n_0'+1\right)^{1-\alpha}\right)\gamma_{n_0'}^3 + c_\gamma^3\frac{3\alpha}{3\alpha-1}\right)$, $M_0 := \max\left\{\frac{2^{4\alpha}}{a_0}, c_\gamma\right\}$ *and* a_0, σ^2 *defined in Lemma 3.1.*

We can now give the rate of convergence in quadratic mean of θ_n in the particular case where $C_2 = C_2' = 0$.

Theorem 3.2. *Suppose Assumptions* ***(A1)*** *to* ***(A4)*** *hold. Then*

$$\mathbb{E}\left[\|\theta_n-\theta\|^2\right] \leq A'e^{-\lambda_{\min}c_\gamma n^{1-\alpha}} + \frac{c_{n_0'}L_\delta^2}{\lambda_{\min}^2}e^{-\frac{1}{4}a_0c_\gamma n^{1-\alpha}} + \frac{L_\delta^2c_\gamma^2\sigma^2}{\lambda_{\min}^2}M_0n^{-2\alpha} + \frac{2^\alpha C_1c_\gamma}{\lambda_{\min}}n^{-\alpha}$$

with $n_1' = \min\{n, \lambda_{\min}\gamma_{n+1} \leq 1\}$, a_0, σ^2 *defined in Lemma 3.1,* $c_{n_0'}, M_0$ *defined in Lemma 3.2, and*

$$A' = e^{\lambda_{\min} c_\gamma (n_1'+1)^{1-\alpha}} \left(C_1 c_\gamma^2 \frac{2\alpha}{2\alpha - 1} + c_{n_0'} + c_\gamma u_0 + \frac{2c_{n_0'}}{a_0(1-\alpha)} e^{-\frac{1}{2}a_0 c_\gamma} + \sigma^2 c_\gamma^3 M_0 \frac{3\alpha}{3\alpha - 1} \right).$$

Remark that here again, without surprise, the main term $\frac{2^\alpha C_1 c_\gamma}{\lambda_{\min}} n^{-\alpha}$ is analogous to the one for the strongly convex case given by Bach and Moulines (2013).

4 The averaged algorithm

Let us recall that the averaged algorithm introduced by Ruppert (1988) and Polyak and Juditsky (1992) is defined for all $n \geq 0$ by

$$\overline{\theta}_n = \frac{1}{n+1} \sum_{k=0}^{n} \theta_k,$$

which can be written recursively as

$$\overline{\theta}_{n+1} = \overline{\theta}_n + \frac{1}{n+2} \left(\theta_{n+1} - \overline{\theta}_n \right).$$

4.1 Case with unbounded gradient

In this section, we focus on the case where $C_2 \neq 0$ or $C_2' \neq 0$. The following theorem gives a first rate of convergence of the averaged estimates.

Theorem 4.1. *Suppose Assumptions* ***(A1)*** *to* ***(A4)*** *hold. Then*

$$\begin{aligned}
\lambda_{\min} \sqrt{\mathbb{E}\left[\left\| \overline{\theta}_n - \theta \right\|^2 \right]} \leq{}& \frac{\sqrt{C_1}}{\sqrt{n+1}} + \frac{L_\delta 2^{1/2+2\alpha} \sigma c_\gamma}{\sqrt{a_0}(1-\alpha)} \frac{1}{(n+1)^\alpha} + \frac{2^{\frac{1+\alpha}{2}} 5 \sqrt{C_1}}{\sqrt{c_\gamma}\sqrt{\lambda_{\min}}} \frac{1}{(n+1)^{1-\alpha/2}} \\
&+ \frac{\sqrt{C_2} 2^{1/4+\alpha} \sqrt{\sigma}\sqrt{c_\gamma}}{a_0^{1/4}\sqrt{1-\alpha}(n+1)^{1/2+\alpha/2}} + \frac{2^{1+4\alpha} \sigma L_\delta}{\sqrt{a_0}\lambda_{\min}} \frac{\ln(n+1)}{n+1} + \frac{A_\infty + D_\infty + L_\delta B_\infty + \sqrt{C_2}\sqrt{B_\infty} + c_\gamma^{-1/2} v_0}{n+1} \\
&+ \frac{\sqrt{A}}{c_\gamma} \frac{e^{-\frac{1}{8}\lambda_{\min} c_\gamma n^{1-\alpha}}}{(n+1)^{1-\alpha}} + \frac{\sqrt{2}\sqrt{c_1} L_\delta}{c_\gamma \lambda_{\min}} \frac{e^{-\frac{1}{16} a_0 c_\gamma n^{1-\alpha}}}{(n+1)^{1-\alpha}}
\end{aligned}$$

with $A_\infty := \frac{\sqrt{A}}{c_\gamma} \sum_{n=0}^{+\infty} e^{-\frac{1}{8}\lambda_{\min} c_\gamma n^{1-\alpha}}$, $B_\infty := \sum_{n=0}^{+\infty} e^{-\frac{1}{8} c_\gamma a_0 n^{1-\alpha}} e^{a_1 c_\gamma^2 \frac{2\alpha}{2\alpha-1} + a_2 c_\gamma^3 \frac{3\alpha}{3\alpha-1}} \left(\sqrt{u_0} + \sigma c_\gamma^{3/2} \sqrt{\frac{3\alpha}{3\alpha-1}} \right)$, *and* $D_\infty := \frac{\sqrt{2}\sqrt{c_1} L_\delta}{\lambda_{\min} c_\gamma} \sum_{n=0}^{+\infty} e^{-\frac{1}{16} a_0 c_\gamma n^{1-\alpha}}$.

The main conclusion of this theorem is that we achieve the usual rate of convergence $\frac{\sqrt{C_1}}{\sqrt{n+1}}$ while the two main rest terms converge at rates $\frac{1}{(n+1)^\alpha}$ and $\frac{1}{(n+1)^{1-\alpha/2}}$ which seems to suggest that the best choice of α could be $\alpha = 2/3$. Nevertheless, in a recent work and in the special case where ∇g is uniformly bounded, Gadat and Panloup (2017) give upper bound for each the best rate of convergence should be achieve for $\alpha = 3/4$. Furthermore, in the particular case of linear regression for which L_δ can be chosen equal to 0 and the two main

rest terms are so of order $\frac{1}{(n+1)^{1-\alpha/2}}$ and $\frac{1}{(n+1)^{1/2+\alpha/2}}$ which suggests to take α close to $\frac{1}{2}$. Nevertheless, our bounds as the ones given in Gadat and Panloup (2017) or Bach and Moulines (2013) can be considered as quite rough, that complicates to answer definitely and generally on the best choice of α.

In order to get a (quasi) optimal rate of convergence, let us suppose from now that the variance of the gradient of g is lipschitz, i.e that the following assumption is fulfilled:

(A5) The functional $\Sigma : h \longmapsto \Sigma(h) = \mathbb{E}\left[\nabla_h g(X,h)\, \nabla_h g(X,h)^T\right]$ is L_Σ lipschitz with respect to the spectral norm.

Remark that this assumption is already present in Godichon-Baggioni (2019b) and is analogous to Assumption (H_S) in Gadat and Panloup (2017). The following theorem ensures that, up to rest terms, the averaged estimates achieve the "Cramer-Rao bound".

Theorem 4.2. *Suppose Assumptions **(A1)** to **(A5)** hold. Then,*

$$\sqrt{\mathbb{E}\left[\left\|\overline{\theta}_n - \theta\right\|^2\right]} \leq \frac{\sqrt{Tr\left(H^{-1}\Sigma H^{-1}\right)}}{\sqrt{n+1}} + \frac{L_\delta 2^{1/2+2\alpha}\sigma c_\gamma}{\sqrt{a_0}(1-\alpha)}\frac{1}{\lambda_{\min}(n+1)^\alpha} + \frac{2^{\frac{1+\alpha}{2}}5\sqrt{C_1}}{\sqrt{c_\gamma}\lambda_{\min}^{3/2}}\frac{1}{(n+1)^{1-\alpha/2}}$$
$$+ \frac{2^{1/2+\alpha/2}\sqrt{C_1}\sqrt{L_\Sigma}\sqrt{c_\gamma}}{\lambda_{\min}^{3/2}\sqrt{1-\alpha}(n+1)^{1/2+\alpha/2}} + \frac{2^{1+4\alpha}\sigma L_\delta}{\sqrt{a_0}\lambda_{\min}^2}\frac{\ln(n+1)}{n+1}$$
$$+ \frac{A_\infty + D_\infty + L_\delta B_\infty + \left(\sqrt{L_\Sigma} + c_\gamma^{-1/2}\right)\sqrt{v_0} + \sqrt{L_\Sigma}c_\gamma A_\infty + \sqrt{L_\Sigma}c_\gamma D_\infty + \frac{2^{1+4\alpha}\sqrt{L_\Sigma}\sigma c_\gamma L_\delta\sqrt{2\alpha}a_0^{-1/2}}{\lambda_{\min}\sqrt{2\alpha-1}}}{\lambda_{\min}(n+1)}$$
$$+ \frac{\sqrt{A}}{c_\gamma}\frac{e^{-\frac{1}{8}\lambda_{\min}c_\gamma n^{1-\alpha}}}{\lambda_{\min}(n+1)^{1-\alpha}} + \frac{\sqrt{2}\sqrt{c_1}L_\delta}{c_\gamma}\frac{e^{-\frac{1}{16}a_0 c_\gamma n^{1-\alpha}}}{\lambda_{\min}^2(n+1)^{1-\alpha}}$$

Remark 4.1. *Note that we speak about Cramer Rao bound in the sens that under regularity assumptions, any estimate $\tilde{\theta}_n$ should verify for almost any $\theta \in \mathcal{H}$,*

$$\liminf_n n\mathbb{E}\left[\left(\tilde{\theta}_n - \theta\right)^2\right] \geq Tr\left(H^{-1}\Sigma(\theta)H^{-1}\right)$$

4.2 Case where $\|\nabla G\|$ is bounded

We now focus on the case where $C_2 = C_2' = 0$. The following theorem gives the rate of convergence of averaged estimates in this case.

Theorem 4.3. *Suppose Assumptions **(A1)** to **(A4)** hold and that $C_2 = C_2' = 0$. Then,*

$$\lambda_{\min}\sqrt{\mathbb{E}\left[\left\|\overline{\theta}_n - \theta\right\|^2\right]} \leq \frac{\sqrt{C_1}}{\sqrt{n+1}} + \frac{L_\delta\sigma c_\gamma\sqrt{M_0}}{(1-\alpha)}\frac{1}{(n+1)^\alpha} + \frac{2^{\frac{\alpha}{2}}5\sqrt{C_1}}{\sqrt{c_\gamma}\sqrt{\lambda_{\min}}}\frac{1}{(n+1)^{1-\alpha/2}}$$
$$+ \frac{\sigma L_\delta\sqrt{M_0}}{\lambda_{\min}(n+1)}\frac{\ln(n+1)}{n+1} + \frac{\sigma L_\delta\sqrt{M_0}\lambda_{\min}^{-1} + A_\infty' + D_\infty' + L_\delta B_\infty'}{n+1}$$
$$+ \frac{\sqrt{A'}}{c_\gamma}\frac{e^{-\frac{1}{2}\lambda_{\min}c_\gamma n^{1-\alpha}}}{(n+1)^{1-\alpha}} + \frac{\sqrt{c_{n_0'}}L_\delta}{c_\gamma\lambda_{\min}}\frac{e^{-\frac{1}{8}a_0 c_\gamma n^{1-\alpha}}}{(n+1)^{1-\alpha}}$$

with $A'_\infty := \frac{\sqrt{A'}}{c_\gamma}\sum_{n=0}^{+\infty} e^{-\frac{1}{2}\lambda_{\min} c_\gamma n^{1-\alpha}}$, $B'_\infty = \left(\sqrt{c_{n'_0}} + \sqrt{u_0}\right)\sum_{n\geq 0}\exp\left(-\frac{1}{4}a_0 c_\gamma n^{1-\alpha}\right)$ *and* $D'_\infty := \frac{\sqrt{c_{n'_0}}L_\delta}{\lambda_{\min} c_\gamma}\sum_{n=0}^{+\infty} e^{-\frac{1}{8}a_0 c_\gamma n^{1-\alpha}}$.

Considering from now that Assumption **(A5)** is fulfilled, we can now prove that the averaged estimates also achieve, unsurprisingly, the "Cramer-Rao bound" in the case where the gradient of G is bounded.

Theorem 4.4. *Suppose Assumptions* ***(A1)*** *to* ***(A5)*** *hold and that* $C_2 = C'_2 = 0$. *Then,*

$$\sqrt{\mathbb{E}\left[\left\|\overline{\theta}_n - \theta\right\|^2\right]} \leq \frac{\sqrt{Tr\left(H^{-1}\Sigma H^{-1}\right)}}{\sqrt{n+1}} + \frac{L_\delta \sigma c_\gamma \sqrt{M_0}}{(1-\alpha)}\frac{1}{\lambda_{\min}(n+1)^\alpha} + \frac{2^{\frac{\alpha}{2}}5\sqrt{C_1}}{\sqrt{c_\gamma}}\frac{1}{\lambda_{\min}^{3/2}(n+1)^{1-\alpha/2}}$$
$$+ \frac{\sqrt{L_\Sigma}2^{\alpha/2}\sqrt{C_1}}{\sqrt{1-\alpha}}\frac{1}{\lambda_{\min}^{3/2}(n+1)^{1/2+\alpha/2}} + \frac{\sigma L_\delta\sqrt{M_0}}{\lambda_{\min}(n+1)}\frac{\ln(n+1)}{(n+1)\lambda_{\min}}$$
$$+ \frac{\left(\sigma + \sqrt{L_\Sigma}c_\gamma\sqrt{\frac{2\alpha}{2\alpha-1}}\right)L_\delta\sqrt{M_0}\lambda_{\min}^{-1} + A'_\infty + D'_\infty + L_\delta B'_\infty + \sqrt{L_\Sigma}\sqrt{v_0} + \sqrt{L_\Sigma}c_\gamma A'_\infty + \sqrt{L_\Sigma}c_\gamma D'_\infty}{(n+1)\lambda_{\min}}$$
$$+ \frac{\sqrt{A'}}{c_\gamma}\frac{e^{-\frac{1}{2}\lambda_{\min}n^{1-\alpha}}}{\lambda_{\min}(n+1)^{1-\alpha}} + \frac{\sqrt{c_{n'_0}}L_\delta}{c_\gamma}\frac{e^{-\frac{1}{8}a_0 c_\gamma n^{1-\alpha}}}{\lambda_{\min}^2(n+1)^{1-\alpha}}$$

Conclusion

In this paper, we provide explicit upper bounds of the quadratic mean error of the online stochastic gradient estimates as well as of their averaged version, and so under very weak assumptions. A first extension of this work could be the obtaining of precise (via concentration inequalities) and calculable confidence balls or ellipse for θ with the help of averaged estimates. A second extension of this work could be to focus on the non-asymptotic rate of convergence of online adaptive stochastic gradient algorithms, such that Adagrad (Duchi et al., 2011), or stochastic Newton algorithms (Boyer and Godichon-Baggioni, 2020). Finally since the averaged estimates are known to be sensitive to a bad initialization, a last perspective could be to extend this work to the Weighted Averaged Stochastic Gradient estimates (Mokkadem and Pelletier, 2011).

Acknowledgments

The author would like to thank Pierre Tarrago for the many fruitful discussions that enable to deeply improve this work.

5 Proofs of Section 3

5.1 Some properties on the functionnal G

First remark that with the help of a Taylor's expansion of G, for all $h \in \mathcal{H}$,

$$G(h) = G(\theta) + (h-\theta)^T\int_0^1 (1-t)\nabla^2 G(\theta + t(h-\theta))dt(h-\theta).$$

Then, thanks to Assumption **(A3)**,

$$G(h) - G(\theta) \leq \frac{1}{2} L_{\nabla G} \|h - \theta\|^2 . \tag{2}$$

Furthermore, thanks to Assumption **(A4)**, for all $h \in \mathcal{B}(\theta, r_{\lambda_0})$,

$$(h-\theta)^T \int_0^1 (1-t) \nabla^2 G(\theta + t(h-\theta)) dt (h-\theta)^T \geq \frac{1}{2} \lambda_0 \|h - \theta\|^2 .$$

If $h \notin \mathcal{B}(\theta, r_{\lambda_0})$, i.e if $\|h - \theta\| > r_{\lambda_0}$, one has

$$\begin{aligned}(h-\theta)^T \int_0^1 (1-t) \nabla^2 G(\theta + t(h-\theta)) dt (h-\theta)^T &\geq (h-\theta)^T \int_0^{\frac{r_{\lambda_0}}{\|h-\theta\|}} (1-t) \nabla^2 G(\theta + t(h-\theta)) dt (h-\theta) \\ &\geq \frac{1}{2} \lambda_0 r_{\lambda_0} \|h - \theta\| .\end{aligned}$$

Then,

$$G(h) - G(\theta) \geq \frac{\lambda_0}{2} \|h - \theta\|^2 \mathbf{1}_{\|h-\theta\| \leq r_{\lambda_0}} + \frac{\lambda_0}{2} r_{\lambda_0} \|h - \theta\| \mathbf{1}_{\|h-\theta\| > r_{\lambda_0}} \tag{3}$$

5.2 Proof of Lemma 3.1

First, thanks to a Taylor's decomposition of G coupled with assumption **(A3)**, we have

$$\begin{aligned}G(\theta_{n+1}) - G(\theta) &= G(\theta_n) - G(\theta) + \langle \nabla G(\theta_n), \theta_{n+1} - \theta_n \rangle \\ &\quad + (\theta_{n+1} - \theta_n)^T \int_0^1 (1-t) \nabla^2 G(\theta_{n+1} + t(\theta_n - \theta_{n+1})) dt (\theta_{n+1} - \theta_n) \\ &\leq G(\theta_n) - G(\theta) - \gamma_{n+1} \langle \nabla G(\theta_n), \nabla_h g(X_{n+1}, \theta_n) \rangle + \frac{1}{2} \gamma_{n+1}^2 L_{\nabla G} \|\nabla_h g(X_{n+1}, \theta_n)\|^2 .\end{aligned}$$

Denoting $V_n := G(\theta_n) - G(\theta)$ and $g'_{n+1} = \nabla_h g(X_{n+1}, \theta_n)$, and thanks to Cauchy-Schwartz inequality, it comes

$$\begin{aligned}V_{n+1}^2 \leq V_n^2 &+ \gamma_{n+1}^2 \|\nabla G(\theta_n)\|^2 \left\|g'_{n+1}\right\|^2 + \frac{1}{4} \gamma_{n+1}^4 L_{\nabla G}^2 \left\|g'_{n+1}\right\|^4 + \gamma_{n+1}^3 L_{\nabla G} \left\|g'_{n+1}\right\|^3 \|\nabla G(\theta_n)\| \\ &+ \gamma_{n+1}^2 V_n \left\|g'_{n+1}\right\|^2 - 2\gamma_{n+1} V_n \left\langle \nabla G(\theta_n), g'_{n+1} \right\rangle\end{aligned}$$

Then, since

$$\left\|g'_{n+1}\right\|^3 \|\nabla G(\theta_n)\| \leq \frac{L_{\nabla G} \gamma_{n+1}}{4} \left\|g'_{n+1}\right\|^4 + \frac{1}{L_{\nabla G} \gamma_{n+1}} \left\|g'_{n+1}\right\|^2 \|\nabla G(\theta_n)\|^2$$

it comes

$$\begin{aligned}V_{n+1}^2 \leq V_n^2 &+ 2\gamma_{n+1}^2 \|\nabla G(\theta_n)\|^2 \left\|g'_{n+1}\right\|^2 + \frac{1}{2} \gamma_{n+1}^4 L_{\nabla G}^2 \left\|g'_{n+1}\right\|^4 + \gamma_{n+1}^2 V_n \left\|g'_{n+1}\right\|^2 \\ &- 2\gamma_{n+1} V_n \left\langle \nabla G(\theta_n), g'_{n+1} \right\rangle .\end{aligned}$$

Taking the conditional expectation and thanks to assumption **(A2)**,

$$\mathbb{E}\left[V_{n+1}^2|\mathcal{F}_n\right] \leq V_n^2 + 2\gamma_{n+1}^2 \left\|\nabla G\left(\theta_n\right)\right\|^2 \left(C_1 + C_2 V_n\right) + \frac{1}{2}\gamma_{n+1}^4 L_{\nabla G}^2 \left(C_1' + C_2' V_n^2\right) + \gamma_{n+1}^2 \left(C_1 + C_2 V_n\right) V_n - 2\gamma_{n+1} \left\|\nabla G\left(\theta_n\right)\right\|^2 V_n \tag{4}$$

Remark that thanks to Assumption **(A3)**,

$$\left\|\nabla G\left(\theta_n\right)\right\|^2 \leq 2L_{\nabla G}\left(G\left(\theta_n\right) - G(\theta)\right)$$

Then, one can rewrite inequality (4) as

$$\mathbb{E}\left[V_{n+1}^2|\mathcal{F}_n\right] \leq \left(1 + C_2\left(4L_{\nabla G} + 1\right)\gamma_{n+1}^2 + \frac{1}{2}\gamma_{n+1}^4 L_{\nabla G}^2 C_2'\right) V_n^2 + C_1\left(4L_{\nabla G} + 1\right)\gamma_{n+1}^2 V_n - 2\gamma_{n+1}\left\|\nabla G\left(\theta_n\right)\right\|^2 V_n + \frac{1}{2}\gamma_{n+1}^4 L_{\nabla G}^2 C_1'. \tag{5}$$

Let us now give a lower bound of $\left\|\nabla G\left(\theta_n\right)\right\|^2$. Thanks to a Taylor's decomposition of the gradient,

$$\left\|\nabla G\left(\theta_n\right)\right\|^2 \geq \left(\int_0^1 \lambda_{\min}\left(\nabla^2 G\left(\theta + t\left(\theta_n - \theta\right)\right)\right) dt\right)^2 \left\|\theta_n - \theta\right\|^2$$

Let us denote $\eta_n := \sqrt{\frac{2}{L_{\nabla G}} \frac{G(\theta_n) - G(\theta)}{\left\|\theta_n - \theta\right\|^2}} \min\left\{1, r_{\lambda_0}\right\}$. Thanks to inequality (2), $\eta_n \leq \min\left\{1, r_{\lambda_0}\right\}$, so that, with the help of Assumption **(A4)**, it comes

$$\left\|\nabla G\left(\theta_n\right)\right\|^2 \geq \left(\int_0^{\eta_n} \lambda_{\min}\left(\nabla^2 G\left(\theta + t\left(\theta_n - \theta\right)\right)\right) dt\right)^2 \left\|\theta_n - \theta\right\|^2 \geq \frac{2\lambda_0^2}{L_{\nabla G}} \min\left\{1, r_{\lambda_0}^2\right\} V_n \tag{6}$$

and one can rewrite inequality (5) as

$$\mathbb{E}\left[V_{n+1}^2|\mathcal{F}_n\right] \leq \left(1 - \frac{4\lambda_0^2}{L_{\nabla G}} \min\left\{1, r_{\lambda_0}^2\right\} \gamma_{n+1} + C_2\left(4L_{\nabla G} + 1\right)\gamma_{n+1}^2 + \frac{L_{\nabla G}^2 C_2'}{2}\gamma_{n+1}^4\right) V_n^2 + C_1\left(4L_{\nabla G} + 1\right)\gamma_{n+1}^2 V_n + \frac{1}{2}\gamma_{n+1}^4 L_{\nabla G}^2 C_1' \tag{7}$$

Finally, since

$$\gamma_{n+1}^2 C_1\left(4L_{\nabla G} + 1\right) V_n \leq \frac{3\lambda_0^2}{L_{\nabla G}} \min\left\{1, r_{\lambda_0}^2\right\} \gamma_{n+1} V_n^2 + \gamma_{n+1}^3 \frac{C_1^2\left(4L_{\nabla G} + 1\right)^2 L_{\nabla G}}{12\lambda_0^2 \min\left\{1, r_{\lambda_0}^2\right\}},$$

one can rewrite inequality (7) as

$$\mathbb{E}\left[V_{n+1}^2|\mathcal{F}_n\right] \leq \left(1-\gamma_{n+1}\frac{\lambda_0^2}{L_{\nabla G}}\min\left\{1,r_{\lambda_0}^2\right\}+C_2\left(4L_{\nabla G}+1\right)\gamma_{n+1}^2+\frac{L_{\nabla G}^2C_2'}{2}\gamma_{n+1}^4\right)V_n^2$$
$$+\gamma_{n+1}^3\frac{C_1^2\left(4L_{\nabla G}+1\right)^2L_{\nabla G}}{12\lambda_0^2\min\left\{1,r_{\lambda_0}^2\right\}}+\frac{1}{2}\gamma_{n+1}^4L_{\nabla G}^2C_1' \tag{8}$$

Let us denote $a_0 = \frac{\lambda_0^2\min\left\{1,r_{\lambda_0}^2\right\}}{L_{\nabla G}}$, $a_1 = \max\left\{\frac{\lambda_0^4}{4L_{\nabla G}^2}, C_2\left(4L_{\nabla G}+1\right)\right\}$, $a_2 = \frac{1}{2}L_{\nabla G}^2C_2'$, $\sigma^2 = \frac{C_1^2(4L_{\nabla G}+1)^2L_{\nabla G}}{12\lambda_0^2\min\left\{1,r_{\lambda_0}^2\right\}}+\frac{c_\gamma L_{\nabla G}^2C_1'}{2}$, and $u_n = \mathbb{E}\left[V_n^2\right]$, one can rewrite inequality (8) as

$$u_{n+1} \leq \left(1-a_0\gamma_{n+1}+a_1\gamma_{n+1}^2+a_2\gamma_{n+1}^3\right)u_n+\sigma^2\gamma_{n+1}^3$$

Let $n_0 = \inf\left\{n, a_0 \geq 2a_1\gamma_{n+1}+2a_2\gamma_{n+1}^2\right\}$. Then, one can rewrite inequality (8) as

$$u_{n+1} = \begin{cases} \left(1+a_1\gamma_{n+1}^2+a_2\gamma_{n+1}^3\right)u_n+\sigma^2\gamma_{n+1}^3 & \text{if } n < n_0 \\ \left(1-\frac{1}{2}a_0\gamma_{n+1}\right)u_n+\sigma^2\gamma_{n+1}^3 & \text{if } n \geq n_0 \end{cases}$$

Remark that if $n \geq n_0$, by definition of a_1,

$$\frac{1}{2}a_0\gamma_{n+1} \leq \frac{\lambda_0^4}{4a_1L_{\nabla G}^2} \leq 1. \tag{9}$$

We now consider two distinct cases: $n \leq n_0$ and $n > n_0$.
Case where $n \leq n_0$: With the help of an induction, one can check that for all $n \leq n_0$,

$$u_n \leq \underbrace{\prod_{i=1}^n\left(1+a_1\gamma_i^2+a_2\gamma_i^3\right)u_0}_{=:U_{1,n}}+\underbrace{\sum_{k=1}^n\prod_{i=k+1}^n\left(1+a_1\gamma_i^2+a_2\gamma_i^3\right)\sigma^2\gamma_n^3}_{=:U_{2,n}}$$

As in Bach and Moulines (2013), remark that by definition of n_0 and since $1+x \leq e^x$, for all $n \leq n_0$,

$$U_{1,n} \leq u_0\exp\left(\sum_{k=1}^n a_1\gamma_k^2+a_2\gamma_k^3\right) \leq u_0\exp\left(-\frac{1}{2}a_0\sum_{k=1}^n\gamma_k\right)\exp\left(2\sum_{k=1}^n a_1\gamma_k^2+a_2\gamma_k^3\right) \tag{10}$$

In a same way, one can check that for all $n \leq n_0$,

$$\begin{aligned} U_{2,n} &\leq \prod_{k=1}^n\left(1+a_1\gamma_k^2+a_2\gamma_k^3\right)\sum_{k=1}^n\sigma^2\gamma_k^3 \\ &\leq \exp\left(\sum_{k=1}^n a_1\gamma_k^2+a_2\gamma_k^2\right)\sum_{k=1}^n\sigma^2\gamma_k^3 \\ &\leq \exp\left(-\frac{1}{2}a_0\sum_{k=1}^n\gamma_k\right)\exp\left(2\sum_{k=1}^n a_1\gamma_k^2+a_2\gamma_k^3\right)\sum_{k=1}^n\sigma^2\gamma_k^3 \end{aligned} \tag{11}$$

Case where $n > n_0$: With the help of an induction, one can check that for all $n > n_0$,

$$u_n \leq \underbrace{\prod_{i=n_0+1}^{n}\left(1-\frac{1}{2}a_0\gamma_i\right)u_{n_0}}_{=:U_{3,n}} + \underbrace{\sum_{k=n_0+1}^{n}\prod_{i=k+1}^{n}\left(1-\frac{1}{2}a_0\gamma_i\right)\sigma^2\gamma_k^3}_{=:U_{4,n}}$$

Furthermore, since

$$u_{n_0} \leq U_{1,n_0} + U_{2,n_0} \leq \exp\left(-\frac{1}{2}\sum_{k=1}^{n_0}\gamma_k\right)\exp\left(2\sum_{k=1}^{n_0}a_1\gamma_k^2 + a_2\gamma_k^3\right)\left(u_0 + \sigma^2\sum_{k=1}^{n_0}\gamma_k^3\right)$$

one can obtain

$$U_{3,n} \leq \exp\left(-\frac{1}{2}a_0\sum_{k=1}^{n}\gamma_k\right)\exp\left(2\sum_{k=1}^{n}a_1\gamma_k^2 + a_2\gamma_k^3\right)\left(u_0 + \sigma^2\sum_{k=1}^{n_0}\gamma_k^3\right) \tag{12}$$

Let us now bound $U_{4,n}$ and differentiate two cases: $n_0 < \lceil n/2\rceil - 1$ and $n_0 \geq \lceil n/2\rceil - 1$.
Case where $n > n_0 \geq \lceil n/2\rceil - 1$: Since γ_k is decreasing,

$$\begin{aligned}
U_{4,n} &\leq \sigma^2\gamma_{n_0+1}^2\sum_{k=n_0+1}^{n}\prod_{i=k+1}^{n}\left(1-\frac{1}{2}a_0\gamma_i\right)\gamma_k \\
&= \frac{2\sigma^2}{a_0}\gamma_{n_0+1}^2\sum_{k=n_0+1}^{n}\prod_{i=k+1}^{n}\left(1-\frac{1}{2}a_0\gamma_i\right) - \prod_{i=k}^{n}\left(1-\frac{1}{2}a_0\gamma_i\right) \\
&\leq \frac{2\sigma^2}{a_0}\gamma_{n_0+1}^2\left(1-\prod_{i=n_0+1}^{n}\left(1-\frac{1}{2}a_0\gamma_i\right)\right)
\end{aligned} \tag{13}$$

and thanks to inequality (9) and since γ_k is decreasing,

$$U_{4,n} \leq \frac{2\sigma^2}{a_0}\gamma_{n_0+1}^2 \leq \frac{2\sigma^2}{a_0}\gamma_{\lceil n/2\rceil}^2$$

Case where $n_0 < \lceil n/2\rceil - 1$: As in Bach and Moulines (2013), for all $m = n_0+1, \ldots, n$, one has

$$U_{4,n} \leq \exp\left(-\frac{1}{2}a_0\sum_{k=m+1}^{n}\gamma_k\right)\sum_{k=n_0+1}^{m}\sigma^2\gamma_k^3 + \frac{2\sigma^2}{a_0}\gamma_m^2.$$

Taking $m = \lceil n/2\rceil - 1$, leads to

$$U_{4,n} \leq \exp\left(-\frac{1}{2}a_0\sum_{k=\lceil n/2\rceil}^{n}\gamma_k\right)\sum_{k=n_0+1}^{\lceil n/2\rceil}\sigma^2\gamma_k^3 + \frac{2\sigma^2}{a_0}\gamma_{\lceil n/2\rceil-1}^2.$$

Final bound of $U_{4,n}$: Since γ_k is decreasing,

$$U_{4,n} \leq \exp\left(-\frac{1}{2}a_0\sum_{k=\lceil n/2\rceil}^{n}\gamma_k\right)\sum_{k=n_0+1}^{\lceil n/2\rceil}\sigma^2\gamma_k^3 + \frac{2\sigma^2}{a_0}\gamma_{\lceil n/2\rceil-1}^2. \tag{14}$$

Lower bound of $\sum_{k=1}^{n}\gamma_k$: Remark that since γ_k is decreasing, for all $n \geq 1$,

$$\sum_{k=1}^{n}\gamma_k \geq \sum_{k=\lceil n/2\rceil}^{n}\gamma_k \geq \frac{n}{2}\gamma_n = \frac{c_\gamma}{2}n^{1-\alpha}.$$

Conclusion: Thanks to inequalities (10) to (14), it comes

$$u_n \leq \exp\left(-\frac{1}{2}a_0\sum_{k=\lceil n/2\rceil}^{n}\gamma_k\right)\exp\left(2\sum_{k=1}^{n}a_1\gamma_k^2+a_2\gamma_k^3\right)\left(u_0+\sum_{k=1}^{n}\sigma^2\gamma_k^3\right)+\frac{2\sigma^2}{a_0}\gamma_n' \tag{15}$$

with

$$\gamma_n' = \begin{cases} \gamma_{\lceil n/2\rceil-1}^2 & \text{if } \lceil n/2\rceil > n_0+1 \\ \gamma_{\lceil n/2\rceil}^2 & \text{if } \lceil n/2\rceil \leq n_0+1 \text{ and } n \geq n_0+1 \\ 0 & \text{else} \end{cases}$$

Then, using integral tests for convergence,

$$u_n \leq \underbrace{\exp\left(-\frac{1}{4}c_\gamma a_0 n^{1-\alpha}\right)\exp\left(2a_1c_\gamma^2\frac{2\alpha}{2\alpha-1}+2a_2c_\gamma^3\frac{3\alpha}{3\alpha-1}\right)\left(u_0+\sigma^2c_\gamma^3\frac{3\alpha}{3\alpha-1}\right)+\frac{2^{1+4\alpha}\sigma^2c_\gamma^2}{a_0}n^{-2\alpha}}_{=:v_n} \tag{16}$$

5.3 Proof of Theorem 3.1

We have, since θ_n is $\mathcal{F}_n$-measurable,

$$\mathbb{E}\left[\|\theta_{n+1}-\theta\|^2|\mathcal{F}_n\right] = \|\theta_n-\theta\|^2 - 2\gamma_{n+1}\langle\theta_n-\theta,\nabla G(\theta_n)\rangle + \gamma_{n+1}^2\mathbb{E}\left[\|\nabla_h g(X_{n+1},\theta_n)\|^2|\mathcal{F}_n\right].$$

Then, linearizing the gradient, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\theta_{n+1}-\theta\|^2|\mathcal{F}_n\right] &= \|\theta_n-\theta\|^2 - 2\gamma_{n+1}\langle\theta_n-\theta,H(\theta_n-\theta)\rangle + 2\gamma_{n+1}\langle\theta_n-\theta,\delta_n\rangle \\ &\quad + \gamma_{n+1}^2\mathbb{E}\left[\|\nabla_h g(X_{n+1},\theta_n)\|^2|\mathcal{F}_n\right].\end{aligned}$$

with $\delta_n = H(\theta_n-\theta)-\nabla G(\theta_n)$. Thanks to Assumption **(A1)** and **(A2)** as well as Cauchy-Schwarz inequality,

$$\mathbb{E}\left[\|\theta_{n+1}-\theta\|^2|\mathcal{F}_n\right] \leq (1-\gamma_{n+1}\lambda_{\min})\|\theta_n-\theta\|^2 + \gamma_{n+1}^2C_1 + \frac{\gamma_{n+1}}{\lambda_{\min}}\|\delta_n\|^2 + \gamma_{n+1}^2C_2(G(\theta_n)-G(\theta))$$

leading, thanks to inequality (2), to

$$\mathbb{E}\left[\|\theta_{n+1}-\theta\|^2\right] \leq \left(1-\gamma_{n+1}\lambda_{\min}+\frac{1}{2}\gamma_{n+1}^2C_2L_{\nabla G}\right)\mathbb{E}\left[\|\theta_n-\theta\|^2\right] + \gamma_{n+1}^2C_1 + \frac{\gamma_{n+1}}{\lambda_{\min}}\mathbb{E}\left[\|\delta_n\|^2\right] \tag{17}$$

Remark that in order to have a usual induction relation on the quadratic mean error, we need to have a rate of convergence of $\mathbb{E}\left[\|\delta_n\|^2\right]$. Here is the main difference with Godichon-Baggioni (2019a): remarking that thanks to assumption **(A3)**, $\|\delta_n\| \leq L_{\nabla G}\|\theta_n-\theta\|$, with the help of

(A4), it comes

$$\|\delta_n\| = \|\delta_n\| \mathbf{1}_{\|\theta_n-\theta\|\leq r_{\lambda_0}} + \|\delta_n\| \mathbf{1}_{\|\theta_n-\theta\|>r_{\lambda_0}} \leq C_{\lambda_0} \|\theta_n-\theta\|^2 \mathbf{1}_{\|\theta_n-\theta\|\leq r_{\lambda_0}} + L_{\nabla G} \|\theta_n-\theta\| \mathbf{1}_{\|\theta_n-\theta\|>r_{\lambda_0}}$$

Then, thanks to inequality (3), it comes

$$\|\delta_n\| \leq \frac{2C_{\lambda_0}}{\lambda_0} \left(G\left(\theta_n\right) - G(\theta)\right) \mathbf{1}_{\|\theta_n-\theta\|\leq r_{\lambda_0}} + \frac{2L_{\nabla G}}{\lambda_0 r_{\lambda_0}} \left(G\left(\theta_n\right) - G(\theta)\right) \mathbf{1}_{\|\theta_n-\theta\|>r_{\lambda_0}} \leq L_\delta \left(G\left(\theta_n\right) - G(\theta)\right) \tag{18}$$

with $L_\delta = \max\left\{\frac{2C_{\lambda_0}}{\lambda_0}, \frac{2L_{\nabla G}}{\lambda_0 r_{\lambda_0}}\right\}$. Then, one can rewrite inequality (17) as

$$\mathbb{E}\left[\|\theta_{n+1}-\theta\|^2\right] \leq \left(1 - \gamma_{n+1}\lambda_{\min} + \frac{1}{2}\gamma_{n+1}^2 C_2 L_{\nabla G}\right) \mathbb{E}\left[\|\theta_n-\theta\|^2\right] + \gamma_{n+1}^2 C_1 + \frac{\gamma_{n+1}}{\lambda_{\min}} L_\delta^2 v_n \tag{19}$$

with v_n defined in equation (16). Let us denote $b_1 = \frac{L_{\nabla G}}{2} \max\left\{C_2, \frac{\lambda_{\min}^2}{2L_{\nabla G}}\right\}$, and let $n_1 = \inf\left\{n, \lambda_{\min} \geq 2\gamma_{n+1} b_1\right\}$. Then, denoting $w_n = \mathbb{E}\left[\|\theta_n-\theta\|^2\right]$, one can rewrite inequality (19) as

$$w_{n+1} \leq \begin{cases} \left(1 + b_1\gamma_{n+1}^2\right) w_n + C_1\gamma_{n+1}^2 + \frac{\gamma_{n+1}}{\lambda_{\min}} L_\delta^2 v_n & \text{if } n < n_1 \\ \left(1 - \frac{1}{2}\lambda_{\min}\gamma_{n+1}\right) w_n + C_1\gamma_{n+1}^2 + \frac{\gamma_{n+1}}{\lambda_{\min}} L_\delta^2 v_n & \text{if } n \geq n_1 \end{cases}$$

Furthermore, by definition of b_1, remark that for all $n \geq n_1$,

$$\frac{1}{2}\lambda_{\min}\gamma_{n+1} \leq \frac{\lambda_{\min}^2}{4b_1} \leq 1. \tag{20}$$

Case where $n \leq n_1$: With the help of an induction, one can check that for all $n \leq n_1$,

$$w_n \leq \underbrace{\prod_{i=1}^n \left(1 + \gamma_i^2 b_1\right)}_{=:A_{1,n}} w_0 + \underbrace{\sum_{k=1}^n \prod_{i=k+1}^n \left(1 + b_1\gamma_i^2\right) \left(C_1\gamma_k + \frac{L_\delta^2}{\lambda_{\min}}\gamma_k v_{k-1}\right)}_{=:B_{1,n}}$$

Remark that by definition of n_1 and since $1 + x \leq e^x$,

$$A_{1,n} \leq \exp\left(\sum_{k=1}^n b_1\gamma_k^2\right) \leq \exp\left(-\frac{1}{2}\lambda_{\min}\sum_{k=1}^n \gamma_k\right) \exp\left(2b_1\sum_{k=1}^n \gamma_k^2\right)$$

Furthermore, by definition of n_1, one can check that

$$\begin{aligned} B_{1,n} &\leq \prod_{k=1}^n \left(1 + b_1\gamma_k^2\right) \sum_{k=1}^n \left(C_1\gamma_k^2 + \frac{L_\delta^2}{\lambda_{\min}}\gamma_k v_{k-1}\right) \\ &\leq \exp\left(b_1\sum_{k=1}^n \gamma_k^2\right) \sum_{k=1}^n \left(C_1\gamma_k^2 + \frac{L_\delta^2}{\lambda_{\min}}\gamma_k v_{k-1}\right) \\ &\leq \exp\left(-\frac{1}{2}\lambda_{\min}\sum_{k=1}^n \gamma_k\right) \exp\left(2b_1\sum_{k=1}^n \gamma_k^2\right) \sum_{k=1}^n \left(C_1\gamma_k^2 + \frac{L_\delta^2}{\lambda_{\min}}\gamma_k v_{k-1}\right) \end{aligned}$$

Then, if $n \leq n_1$, one have

$$w_n \leq \exp\left(-\frac{1}{2}\lambda_{\min}\sum_{k=1}^{n}\gamma_k\right)\exp\left(2b_1\sum_{k=1}^{n}\gamma_k^2\right)\left(w_0+\sum_{k=1}^{n}\left(C_1\gamma_k^2+\frac{L_\delta^2}{\lambda_{\min}}\gamma_k v_{k-1}\right)\right) \tag{21}$$

Case where $n > n_1$: With the help of an induction, one can check that for all $n > n_1$,

$$w_n = \underbrace{\prod_{i=n_1+1}^{n}\left(1-\frac{1}{2}\lambda_{\min}\gamma_i\right)w_{n_1}}_{=:A_{2,n}} + \underbrace{\sum_{k=n_1+1}^{n}\prod_{i=k+1}^{n}\left(1-\frac{1}{2}\lambda_{\min}\gamma_i\right)\left(\gamma_k^2 C_1+\frac{L_\delta^2}{\lambda_{\min}}\gamma_k v_{k-1}\right)}_{=:B_{2,n}}$$

Thanks to inequality (20), one has $\prod_{i=n_1+1}^{n}\left(1-\frac{1}{2}\lambda_{\min}\gamma_i\right) \leq \exp\left(-\frac{1}{2}\lambda_{\min}\sum_{i=n_1+1}^{n}\gamma_i\right)$, and with the help of inequality (21), it comes

$$A_{2,n} \leq \exp\left(-\frac{1}{2}\lambda_{\min}\sum_{k=1}^{n}\gamma_k\right)\exp\left(2b_1\sum_{k=1}^{n_1}\gamma_k^2\right)\left(w_0+\sum_{k=1}^{n_1}\left(\gamma_k^2 C_1+\frac{L_\delta^2}{\lambda_{\min}}\gamma_k v_{k-1}\right)\right)$$

Let us now bound $B_{2,n}$ and differentiate two cases: $\lceil n/2\rceil - 1 > n_1$ and $\lceil n/2\rceil - 1 \leq n_1$.
Case where $n > n_1 \geq \lceil n/2\rceil - 1$: Since γ_k and v_k are decreasing, and since

$$\begin{aligned}
B_{2,n} &\leq \left(\gamma_{n_1+1}C_1+\frac{L_\delta^2}{\lambda_{\min}}v_{n_1}\right)\sum_{k=n_1+1}^{n}\prod_{i=k+1}^{n}\left(1-\frac{1}{2}\lambda_{\min}\gamma_i\right)\gamma_k \\
&= \left(\gamma_{n_1+1}C_1+\frac{L_\delta^2}{\lambda_{\min}}v_{n_1}\right)\frac{2}{\lambda_{\min}}\sum_{k=n_1+1}^{n}\prod_{i=k+1}^{n}\left(1-\frac{1}{2}\lambda_{\min}\gamma_i\right)-\prod_{i=k}^{n}\left(1-\frac{1}{2}\lambda_{\min}\gamma_i\right)
\end{aligned}$$

With the help of inequality (20) and since γ_k and v_k are decreasing,

$$\begin{aligned}
B_{2,n} &\leq \left(\gamma_{n_1+1}C_1+\frac{L_\delta^2}{\lambda_{\min}}v_{n_1}\right)\frac{2}{\lambda_{\min}}\left(1-\prod_{i=n_1+1}^{n}\left(1-\frac{1}{2}\lambda_{\min}\gamma_i\right)\right) \\
&\leq \left(\gamma_{\lceil n/2\rceil}C_1+\frac{L_\delta^2}{\lambda_{\min}}v_{\lceil n/2\rceil-1}\right)\frac{2}{\lambda_{\min}}
\end{aligned}$$

Case where $n_1 < \lceil n/2\rceil - 1$: As in Bach and Moulines (2013), since γ_k and v_k are decreasing, one can check that for all $m = n_1+1,\ldots,n$

$$B_{2,n} \leq \exp\left(-\frac{1}{2}\lambda_{\min}\sum_{k=m+1}^{n}\gamma_k\right)\sum_{k=n_1+1}^{m}\left(\gamma_k^2 C_1+\frac{L_\delta^2}{\lambda_{\min}}\gamma_k v_{k-1}\right)+\gamma_m\frac{2C_1}{\lambda_{\min}}+\frac{2L_\delta^2}{\lambda_{\min}^2}v_{m-1}$$

Taking $m = \lceil n/2\rceil - 1$, it comes by definition of n_1,

$$B_{2,n} \leq \exp\left(-\frac{1}{2}\lambda_{\min}\sum_{k=\lceil n/2\rceil}^{n}\gamma_k\right)\sum_{k=n_1+1}^{\lceil n/2\rceil}\left(\gamma_k^2 C_1+\frac{L_\delta^2}{\lambda_{\min}}\gamma_k v_{k-1}\right)+\gamma_{\lceil n/2\rceil-1}\frac{2C_1}{\lambda_{\min}}+\frac{2L_\delta^2}{\lambda_{\min}^2}v_{\lceil n/2\rceil-2}$$

Final bound of $B_{2,n}$: For all $n \geq n_1$,

$$B_{2,n} \leq \exp\left(-\frac{1}{2}\lambda_{\min}\sum_{k=\lceil n/2\rceil}^{n}\gamma_k\right)\sum_{k=n_1+1}^{\lceil n/2\rceil}\left(\gamma_k^2 C_1 + \frac{L_\delta^2}{\lambda_{\min}}\gamma_k v_{k-1}\right) + r_n$$

with

$$r_n = \begin{cases} \gamma_{\lceil n/2\rceil-1}\frac{2C_1}{\lambda_{\min}} + \frac{2L_\delta^2}{\lambda_{\min}^2}v_{\lceil n/2\rceil-2} & \text{if } n_1 < \lceil n/2\rceil - 1 \\ \gamma_{\lceil n/2\rceil}\frac{2C_1}{\lambda_{\min}} + \frac{2L_\delta^2}{\lambda_{\min}^2}v_{\lceil n/2\rceil-1} & \text{if } \lceil n/2\rceil - 1 \leq n_1 \text{ and } n > n_1 \\ 0 & \text{else} \end{cases}$$

Final bound of w_n: Let us recall that $\sum_{k=1}^{n}\gamma_k \geq \sum_{k=\lceil n/2\rceil}^{n}\gamma_k \geq \frac{n}{2}\gamma_n = \frac{c_\gamma}{2}n^{1-\alpha}$, so that, with the help of integral tests for convergence,

$$w_n \leq \exp\left(-\frac{1}{4}\lambda_{\min}c_\gamma n^{1-\alpha}\right)\exp\left(2b_1c_\gamma^2\frac{2\alpha}{2\alpha-1}\right)\left(w_0 + c_\gamma^2 C_1\frac{2\alpha}{2\alpha-1} + 2\frac{L_\delta^2}{\lambda_{\min}}\sum_{k=1}^{n}\gamma_k v_{k-1}\right) + r_n$$

Let us recall that for all $n \geq 1$,

$$v_n \leq \exp\left(-\frac{1}{4}a_0c_\gamma n^{1-\alpha}\right)\underbrace{\exp\left(2a_1c_\gamma^2\frac{2\alpha}{2\alpha-1} + 2a_2c_\gamma^3\frac{3\alpha}{3\alpha-1}\right)\left(v_0 + \sigma^2c_\gamma^3\frac{3\alpha}{3\alpha-1}\right)}_{=:c_1} + \frac{2^{1+4\alpha}\sigma^2c_\gamma^2}{a_0}n^{-2\alpha}$$

With the help of integral tests for convergence,

$$\begin{aligned}\sum_{k=1}^{n}\gamma_k v_{k-1} &\leq u_0c_\gamma + c_1 + c_1\int_1^n c_\gamma t^{-\alpha}\exp\left(-\frac{1}{4}a_0c_\gamma t^{1-\alpha}\right)dt + \frac{2^{1+4\alpha}\sigma^2c_\gamma^3}{a_0}\frac{3\alpha}{3\alpha-1} \\ &\leq u_0c_\gamma + c_1 - \frac{4c_1}{a_0(1-\alpha)}\left[\exp\left(-\frac{1}{4}a_0c_\gamma t^{1-\alpha}\right)\right]_1^n + \frac{2^{1+4\alpha}\sigma^2c_\gamma^3}{a_0}\frac{3\alpha}{3\alpha-1} \\ &\leq u_0c_\gamma + c_1 + \frac{4c_1}{a_0(1-\alpha)}\exp\left(-\frac{1}{4}a_0c_\gamma\right) + \frac{2^{1+4\alpha}\sigma^2c_\gamma^3}{a_0}\frac{3\alpha}{3\alpha-1}\end{aligned}$$

Finally, thanks to inequality (15)

$$r_n \leq \left(c_1\exp\left(-\frac{1}{8}a_0c_\gamma n^{1-\alpha}\right) + \frac{2^{1+8\alpha}\sigma^2c_\gamma^2}{a_0}n^{-2\alpha}\right)\frac{2L_\delta^2}{\lambda_{\min}^2} + \frac{2^{1+\alpha}C_1}{\lambda_{\min}}c_\gamma n^{-\alpha}$$

it comes

$$\begin{aligned}w_n \leq\ & e^{-\frac{1}{4}\lambda_{\min}c_\gamma n^{1-\alpha}}e^{2b_1c_\gamma^2\frac{2\alpha}{2\alpha-1}}\left(w_0 + c_\gamma^2C_1\frac{2\alpha}{2\alpha-1} + \frac{2L_\delta^2}{\lambda_{\min}}\left(u_0c_\gamma + c_1 + \frac{4c_1}{a_0(1-\alpha)}e^{-\frac{1}{4}a_0c_\gamma} + \frac{2^{1+4\alpha}\sigma^2c_\gamma^2}{a_0}\frac{3\alpha}{3\alpha-1}\right)\right) \\ &+ \left(c_1\exp\left(-\frac{1}{8}a_0c_\gamma n^{1-\alpha}\right) + \frac{2^{1+8\alpha}\sigma^2c_\gamma^2}{a_0}n^{-2\alpha}\right)\frac{2L_\delta^2}{\lambda_{\min}^2} + \frac{2^{1+\alpha}C_1}{\lambda_{\min}}c_\gamma n^{-\alpha}\end{aligned}$$

i.e

$$\mathbb{E}\left[\|\theta_n-\theta\|^2\right] \leq Ae^{-\frac{1}{4}\lambda_{\min}c_\gamma n^{1-\alpha}} + \left(c_1 e^{-\frac{1}{8}a_0c_\gamma n^{1-\alpha}} + \frac{2^{1+8\alpha}\sigma^2c_\gamma^2}{a_0}n^{-2\alpha}\right)\frac{2L_\delta^2}{\lambda_{\min}^2} + \frac{2^{1+\alpha}C_1}{\lambda_{\min}}c_\gamma n^{-\alpha}$$

with

$$A = e^{2b_1c_\gamma^2\frac{2\alpha}{2\alpha-1}}\left(w_0 + \frac{2\alpha c_\gamma^2 C_1}{2\alpha-1} + 2\frac{L_\delta^2}{\lambda_{\min}}\left(u_0c_\gamma + c_1 + \frac{4c_1}{a_0(1-\alpha)}e^{-\frac{1}{4}a_0c_\gamma} + \frac{2^{1+4\alpha}\sigma^2c_\gamma^3}{a_0}\frac{3\alpha}{3\alpha-1}\right)\right)$$

5.4 Proof of Lemma 3.2

If $C_2 = C_2' = 0$, one can rewrite inequality (8) as

$$u_{n+1} \leq (1-a_0\gamma_{n+1})\,u_n + \sigma^2\gamma_{n+1}^3$$

with $u_n = \mathbb{E}\left[V_n^2\right]$, $a_0 = \frac{\lambda_0^2\min\left\{1,r_{\lambda_0}^2\right\}}{L_{\nabla G}}$, and $\sigma^2 = \frac{C_1^2(4L_{\nabla G}+1)^2L_{\nabla G}}{12\lambda_0^2\min\left\{1,r_{\lambda_0}^2\right\}} + \frac{c_\gamma L_{\nabla G}^2C_1'}{2}$. Let $n_0' = \inf\{n, a_0\gamma_{n+1} \leq 1\}$. One can rewrite previous inequality as

$$u_{n+1} \leq \begin{cases} \sigma^2\gamma_{n+1}^3 & \text{if } n < n_0' \\ (1-a_0\gamma_{n+1})\,u_n + \sigma^2\gamma_{n+1}^3 & \text{if } n \geq n_0'. \end{cases}$$

Then, we just have to study the case where $n > n_0'$. With the help of an induction, one has

$$u_n \leq \underbrace{\prod_{i=n_0'+1}^{n}(1-a_0\gamma_i)\,u_{n_0}}_{=:U_{3,n}'} + \underbrace{\sum_{k=n_0'+1}^{n}\prod_{i=k+1}^{n}(1-a_0\gamma_i)\,\sigma^2\gamma_k^3}_{=:U_{4,n}'}$$

We now bound each term on the right-hand side of previous inequality.
Bounding $U_{3,n}'$: By definition of n_0', and since $1+x \leq e^x$,

$$U_{3,n}' \leq \exp\left(-a_0\sum_{k=n_0'+1}^{n}\gamma_k\right)\sigma^2\gamma_{n_0'}^3$$

With the help of an integral test for convergence,

$$\begin{aligned} U_{3,n}' &\leq \exp\left(-a_0c_\gamma\int_{n_0'+1}^{n}t^{-\alpha}dt\right)\sigma^2\gamma_{n_0'}^3 \\ &= \exp\left(-a_0c_\gamma\frac{1}{1-\alpha}\left((n+1)^{1-\alpha} - \left(n_0'+1\right)^{1-\alpha}\right)\right)\sigma^2\gamma_{n_0'}^3 \\ &\leq \exp\left(-\frac{1}{2}a_0c_\gamma\left((n+1)^{1-\alpha} - \left(n_0'+1\right)^{1-\alpha}\right)\right)\sigma^2\gamma_{n_0'}^3 \end{aligned} \tag{22}$$

Bounding $U_{4,n}'$: As in the proof of Lemma 3.1, we will consider two cases: $n_0' < \lceil n/2\rceil - 1$ and $n_0' \geq \lceil n/2\rceil - 1$.

Case where $n_0' \geq \lceil n/2 \rceil - 1$: With calculus analogous to (13), one can obtain

$$U'_{4,n} \leq \frac{\sigma^2}{a_0} \gamma^2_{\lceil n/2 \rceil}. \tag{23}$$

Case where $n_0' < \lceil n/2 \rceil - 1$: As in Bach and Moulines (2013), for all $m = n_0' + 1, \ldots, n$,

$$U'_{4,n} \leq \exp\left(-a_0 \sum_{k=m+1}^{n} \gamma_k\right) \sum_{k=n_0+1}^{m} \sigma^2 \gamma_k^3 + \frac{\sigma^2}{a_0} \gamma_m^2.$$

Taking $m = \lceil n/2 \rceil - 1$ and with the help of an integral test for convergence, it comes

$$\begin{aligned} U'_{4,n} &\leq \exp\left(-a_0 \sum_{k=\lceil n/2 \rceil}^{n} \gamma_k\right) \sum_{k=n_0+1}^{\lceil n/2 \rceil - 1} \sigma^2 \gamma_k^3 + \frac{\sigma^2}{a_0} \gamma^2_{\lceil n/2 \rceil - 1} \\ &\leq \exp\left(-\frac{1}{2} a_0 c_\gamma n^{1-\alpha}\right) \sigma^2 c_\gamma^3 \frac{3\alpha}{3\alpha - 1} + \frac{\sigma^2}{a_0} \gamma^2_{\lceil n/2 \rceil - 1} \end{aligned} \tag{24}$$

Bounding u_n: Thanks to inequalities (22),(23) and (24), we have

$$u_n \leq \exp\left(-\frac{1}{2} a_0 c_\gamma n^{1-\alpha}\right) \sigma^2 \max\left\{\exp\left(\frac{1}{2} a_0 c_\gamma \left(n_0' + 1\right)^{1-\alpha}\right) \gamma^3_{n_0'}, c_\gamma^3 \frac{3\alpha}{3\alpha - 1}\right\} + r'_n$$

with

$$r'_n = \begin{cases} \sigma^2 \gamma_n^3 & \text{if } n \leq n_0' \\ \frac{\sigma^2}{a_0} \gamma^2_{\lceil n/2 \rceil} & \text{if } n > n_0' \geq \lceil n/2 \rceil - 1 \\ \frac{\sigma^2}{a_0} \gamma^2_{\lceil n/2 \rceil - 1} & \text{else} \end{cases}$$

which can be also writen as

$$u_n \leq \begin{cases} \sigma^2 \gamma_n^3 & \text{if } n \leq n_0' \\ e^{-\frac{1}{2} a_0 c_\gamma n^{1-\alpha}} e^{\frac{1}{2} a_0 c_\gamma (n_0'+1)^{1-\alpha}} \sigma^2 \gamma^3_{n_0'} + \frac{\sigma^2}{a_0} \gamma_{\lceil n/2 \rceil} & \text{if } n > n_0' \geq \lceil n/2 \rceil - 1 \\ e^{-\frac{1}{2} a_0 c_\gamma n^{1-\alpha}} \sigma^2 c_\gamma^3 \frac{3\alpha}{3\alpha - 1} + \frac{\sigma^2}{a_0} \gamma^2_{\lceil n/2 \rceil - 1} & \text{else} \end{cases}$$

or as

$$u_n \leq \underbrace{\exp\left(-\frac{1}{2} a_0 c_\gamma n^{1-\alpha}\right) \sigma^2 \left(\exp\left(\frac{1}{2} a_0 c_\gamma \left(n_0' + 1\right)^{1-\alpha}\right) \gamma^3_{n_0'} + c_\gamma^3 \frac{3\alpha}{3\alpha - 1}\right) + \sigma^2 M_0 c_\gamma^2 n^{-2\alpha}}_{=: v'_n} \tag{25}$$

with $M_0 = \max\left\{\frac{2^{4\alpha}}{a_0}, c_\gamma\right\}$.

5.5 Proof of Theorem 3.2

If $C_2 = 0$, by definition of v'_n (see equation (25)), one can rewrite inequality (19) as

$$\mathbb{E}\left[\|\theta_{n+1} - \theta\|^2\right] \leq (1 - \gamma_{n+1} \lambda_{\min}) \, \mathbb{E}\left[\|\theta_n - \theta\|^2\right] + \gamma_{n+1}^2 C_1 + \frac{\gamma_{n+1}}{\lambda_{\min}} L_\delta^2 v'_n. \tag{26}$$

Let us denote $n_1' = \min\{n, \lambda_{\min}\gamma_{n+1} \geq 1\}$. One can rewrite inequality (26) as

$$w_{n+1} \leq \begin{cases} \gamma_{n+1}^2 C_1 + \frac{\gamma_{n+1}}{\lambda_{\min}} L_\delta^2 v_n' & \text{if } n < n_1' \\ (1 - \gamma_{n+1}\lambda_{\min})\, w_n + \gamma_{n+1}^2 C_1 + \frac{\gamma_{n+1}}{\lambda_{\min}} L_\delta^2 v_n' & \text{if } n \geq n_1 \end{cases}$$

with $\mathbb{E}\left[\|\theta_n - \theta\|^2\right]$. We now focus on the case where $n > n_1$. First, remark that with the help of an induction, one can obtain

$$w_n \leq \underbrace{\prod_{i=n_1'+1}^{n} (1 - \lambda_{\min}\gamma_i)\, w_{n_1'}}_{=:A_{1,n}'} + \underbrace{\sum_{k=n_1'+1}^{n} \prod_{i=k+1}^{n} (1 - \lambda_{\min}\gamma_i) \left(\gamma_k^2 C_1 + \frac{\gamma_k}{\lambda_{\min}} L_\delta^2 v_{k-1}'\right)}_{=:A_{2,n}'}$$

and we now bound each term on the right-hand side of previous inequality.

Bounding $A_{1,n}'$: By definition of n_1' and with the help of an integral test for convergence, one can check that

$$A_{1,n}' \leq \exp\left(-\lambda_{\min} c_\gamma \left((n+1)^{1-\alpha} - \left(n_1'+1\right)^{1-\alpha}\right)\right) \left(\gamma_{n_1'}^2 C_1 + \frac{\gamma_{n_1'}}{\lambda_{\min}} L_\delta^2 v_{n_1'-1}'\right) \tag{27}$$

Bounding $A_{2,n}'$: As we did in previous calculus, since γ_k and v_k' are decreasing, one can check that if $n_1' \geq \lceil n/2 \rceil - 1$,

$$A_{2,n}' \leq \frac{C_1}{\lambda_{\min}} \gamma_{\lceil n/2 \rceil} + \frac{L_\delta^2}{\lambda_{\min}^2} v_{\lceil n/2 \rceil - 1}'$$

and if $n_1' < \lceil n/2 \rceil - 1$,

$$A_{2,n}' \leq \exp\left(-\lambda_{\min} \sum_{k=\lceil n/2 \rceil}^{n} \gamma_k\right) \left(C_1 c_\gamma^2 \frac{2\alpha}{2\alpha - 1} + \frac{L_\delta^2}{\lambda_{\min}} \sum_{k=1}^{n} \gamma_k v_{k-1}'\right) + \frac{C_1}{\lambda_{\min}} \gamma_{\lceil n/2 \rceil - 1} + \frac{L_\delta^2}{\lambda_{\min}^2} v_{\lceil n/2 \rceil - 2}'$$

Then,

$$A_{2,n}' \leq \exp\left(-\lambda_{\min} \sum_{k=\lceil n/2 \rceil}^{n} \gamma_k\right) \left(C_1 c_\gamma^2 \frac{2\alpha}{2\alpha - 1} + \frac{L_\delta^2}{\lambda_{\min}} \sum_{k=1}^{n} \gamma_k v_{k-1}'\right) + r_n'' \tag{28}$$

with

$$r_n'' = \begin{cases} \frac{C_1}{\lambda_{\min}} \gamma_{\lceil n/2 \rceil - 1} + \frac{L_\delta^2}{\lambda_{\min}^2} v_{\lceil n/2 \rceil - 2}' & \text{if } n_{1'} < \lceil n/2 \rceil - 1 \\ \frac{C_1}{\lambda_{\min}} \gamma_{\lceil n/2 \rceil} + \frac{L_\delta^2}{\lambda_{\min}^2} v_{\lceil n/2 \rceil - 1}' & \text{if } \lceil n/2 \rceil - 1 \leq n_1 \text{ and } n > n_1 \\ 0 & \text{else} \end{cases}$$

Let us denote $c_{n_0'} := \sigma^2 \left(\exp\left(\frac{1}{2} a_0 c_\gamma \left(n_0' + 1\right)^{1-\alpha}\right) \gamma_{n_0'}^3 + c_\gamma^3 \frac{3\alpha}{3\alpha - 1}\right)$, i.e one can bound v_n' as (with v_n' defined in (25))

$$v_n' \leq c_{n_0'} \exp\left(-\frac{1}{2} a_0 c_\gamma n^{1-\alpha}\right) + \sigma^2 M_0 c_\gamma^2 n^{-2\alpha}$$

Then, with the help of an integral test for convergence, one can check that

$$\sum_{k=1}^{n} \gamma_k v'_{k-1} \leq c_{n'_0} + c_\gamma u_0 + \frac{2c_{n'_0}}{a_0(1-\alpha)} \exp\left(-\frac{1}{2}a_0 c_\gamma\right) + \sigma^2 c_\gamma^3 M_0 \frac{3\alpha}{3\alpha - 1}$$

Furthermore, one can check that

$$r''_n \leq \left(c_{n'_0} \exp\left(-\frac{1}{4}a_0 c_\gamma n^{1-\alpha}\right) + \sigma^2 c_\gamma^2 M_0 n^{-2\alpha}\right) \frac{L_\delta^2}{\lambda_{\min}^2} + \frac{2^\alpha C_1}{\lambda_{\min}} c_\gamma n^{-\alpha}. \tag{29}$$

Final bound of $\mathbb{E}\left[\|\theta_n - \theta\|^2\right]$: As a conclusion, thanks to inequalities (27), (28) and (29),

$$\mathbb{E}\left[\|\theta_n - \theta\|^2\right] \leq A' e^{-\lambda_{\min} c_\gamma n^{1-\alpha}} + \left(c_{n'_0} e^{-\frac{1}{4}a_0 c_\gamma n^{1-\alpha}} + \sigma^2 c_\gamma^2 M_0 n^{-2\alpha}\right) \frac{L_\delta^2}{\lambda_{\min}^2} + \frac{2^\alpha C_1 c_\gamma}{\lambda_{\min}} n^{-\alpha}$$

with

$$A' = e^{\lambda_{\min} c_\gamma (n'_1+1)^{1-\alpha}} \left(C_1 c_\gamma^2 \frac{2\alpha}{2\alpha - 1} + c_{n'_0} + c_\gamma u_0 + \frac{2c_{n'_0}}{a_0(1-\alpha)} e^{-\frac{1}{2}a_0 c_\gamma} + \sigma^2 c_\gamma^3 M_0 \frac{3\alpha}{3\alpha - 1}\right).$$

6 Proofs of Section 4

In order to prove theorems of Section 4, let us first give some usual decompositions of the estimates. First, remark that one can rewrite θ_{n+1} as

$$\theta_{n+1} - \theta = \theta_n - \theta - \gamma_{n+1} \nabla G(\theta_n) + \gamma_{n+1} \xi_{n+1} \tag{30}$$

where $\xi_{n+1} := \nabla G(\theta_n) - \nabla_h g(X_{n+1}, \theta)$ is a martingale difference adapted to $\mathcal{F}_n$. Furthermore, denoting $H = \nabla^2 G(\theta)$ and linearizing the gradient, one has

$$\theta_{n+1} - \theta = (I_d - \gamma_{n+1} H)(\theta_n - \theta) + \gamma_{n+1} \xi_{n+1} - \gamma_{n+1} \delta_n \tag{31}$$

where $\delta_n := \nabla G(\theta_n) - H(\theta_n - \theta)$ is the remainder term in the Taylor's expansion of the gradient. This inequality can be rewrite as

$$H(\theta_n - \theta) = \frac{\theta_n - \theta_{n+1}}{\gamma_{n+1}} + \xi_{n+1} - \delta_n.$$

Summing these equalities, dividing by $n+1$ and applying an Abel's transform (see Pelletier (2000) for more details), it comes

$$\begin{aligned} H(\overline{\theta}_n - \theta) = {} & \frac{\theta_0 - \theta}{\gamma_1(n+1)} - \frac{\theta_{n+1} - \theta}{\gamma_{n+1}(n+1)} + \frac{1}{n+1} \sum_{k=1}^{n} (\theta_k - \theta) \left(\frac{1}{\gamma_{k+1}} - \frac{1}{\gamma_k}\right) + \frac{1}{n+1} \sum_{k=0}^{n} \xi_{k+1} \\ & - \frac{1}{n+1} \sum_{k=0}^{n} \delta_k \end{aligned} \tag{32}$$

6.1 Proof of Theorem 4.1

In order to prove Theorem 4.1, let us bound each term on the right hand-side of equality (32).

Bounding $\sqrt{\mathbb{E}\left[\left\|\frac{\theta_{n+1}-\theta}{\gamma_{n+1}(n+1)}\right\|^2\right]}$: Thanks to Theorem 3.1, one has

$$\sqrt{\mathbb{E}\left[\left\|\frac{\theta_{n+1}-\theta}{\gamma_{n+1}(n+1)}\right\|^2\right]} \leq \frac{\sqrt{A}e^{-\frac{1}{8}\lambda_{\min}c_\gamma n^{1-\alpha}}}{c_\gamma(n+1)^{1-\alpha}} + \frac{\sqrt{2}\sqrt{c_1}L_\delta}{\lambda_{\min}c_\gamma(n+1)^{1-\alpha}}\exp\left(-\frac{1}{16}a_0c_\gamma n^{1-\alpha}\right) + \frac{2^{1+4\alpha}\sigma}{\sqrt{a_0}}\frac{L_\delta}{\lambda_{\min}(n+1)} + \frac{2^{\frac{1+\alpha}{2}}\sqrt{C_1}}{\sqrt{\lambda_{\min}}}\frac{1}{\sqrt{c_\gamma}(n+1)^{1-\alpha/2}} \tag{33}$$

Bounding $R_n := \frac{1}{n+1}\sqrt{\mathbb{E}\left[\left\|\sum_{k=1}^n(\theta_k-\theta)\left(\frac{1}{\gamma_{k+1}}-\frac{1}{\gamma_k}\right)\right\|^2\right]}$. First remark that $\left|\frac{1}{\gamma_{k+1}}-\frac{1}{\gamma_k}\right| \leq \alpha c_\gamma^{-1}k^{\alpha-1} \leq c_\gamma^{-1}k^{\alpha-1}$, so that, thanks to Minkowski's inequality,

$$R_n \leq \frac{1}{n+1}\sum_{k=1}^n\sqrt{\mathbb{E}\left[\|\theta_k-\theta\|^2\right]}\left|\frac{1}{\gamma_{k+1}}-\frac{1}{\gamma_k}\right| \leq \frac{\alpha c_\gamma^{-1}}{n+1}\sum_{k=1}^n\sqrt{\mathbb{E}\left[\|\theta_k-\theta\|^2\right]}k^{\alpha-1}$$

Denoting

$$A_\infty := \frac{\sqrt{A}}{c_\gamma}\sum_{n=0}^{+\infty}e^{-\frac{1}{8}\lambda_{\min}c_\gamma n^{1-\alpha}} \qquad \text{and} \qquad D_\infty := \frac{\sqrt{2}\sqrt{c_1}L_\delta}{\lambda_{\min}c_\gamma}\sum_{n=0}^{+\infty}e^{-\frac{1}{16}a_0c_\gamma n^{1-\alpha}}$$

it comes

$$\begin{aligned} R_n &\leq \frac{A_\infty+D_\infty}{n+1} + \frac{2^{1+4\alpha}\sigma L_\delta}{\sqrt{a_0}\lambda_{\min}}\frac{1}{n+1}\sum_{k=1}^n k^{-1} + \frac{2^{\frac{1+\alpha}{2}}\sqrt{C_1}}{\sqrt{c_\gamma}\sqrt{\lambda_{\min}}}\frac{\alpha}{n+1}\sum_{k=1}^n k^{\alpha/2-1} \\ &\leq \frac{A_\infty+D_\infty}{n+1} + \frac{2^{1+4\alpha}\sigma L_\delta}{\sqrt{a_0}\lambda_{\min}}\frac{\ln(n+1)}{n+1} + \frac{2^{\frac{3+\alpha}{2}}\sqrt{C_1}}{\sqrt{c_\gamma}\sqrt{\lambda_{\min}}}\frac{1}{(n+1)^{1-\alpha/2}} \end{aligned} \tag{34}$$

Bounding $R'_n = \frac{1}{n+1}\sqrt{\mathbb{E}\left[\|\sum_{k=0}^n\delta_k\|^2\right]}$. First remark that thanks to Minkowski's inequality coupled with inequality (18), one has

$$R'_n \leq \frac{1}{n+1}\sum_{k=0}^n\sqrt{\mathbb{E}\left[\|\delta_n\|^2\right]} \leq \frac{L_\delta\sqrt{u_0}}{n+1} + \frac{L_\delta}{n+1}\sum_{k=1}^n\sqrt{\mathbb{E}\left[\left(G\left(\theta_n\right)-G(\theta)\right)^2\right]}$$

Then, applying Lemma 3.1 and denoting

$$B_\infty := \sum_{n=0}^{+\infty}e^{-\frac{1}{8}c_\gamma a_0 n^{1-\alpha}}e^{a_1c_\gamma^2\frac{2\alpha}{2\alpha-1}+a_2c_\gamma^3\frac{3\alpha}{3\alpha-1}}\left(\sqrt{u_0}+\sigma c_\gamma^{3/2}\sqrt{\frac{3\alpha}{3\alpha-1}}\right),$$

one has

$$R'_n \leq \overbrace{\frac{L_\delta\sqrt{u_0}}{n+1} + \frac{L_\delta}{n+1}\sum_{k=1}^{n} e^{-\frac{1}{8}c_\gamma a_0 k^{1-\alpha}} e^{a_1 c_\gamma^2 \frac{2\alpha}{2\alpha-1} + a_2 c_\gamma^3 \frac{3\alpha}{3\alpha-1}} \left(\sqrt{u_0} + \sigma c_\gamma^{3/2}\sqrt{\frac{3\alpha}{3\alpha-1}}\right)}^{\leq \frac{L_\delta B_\infty}{n+1}} + \frac{L_\delta 2^{1/2+2\alpha}\sigma c_\gamma}{\sqrt{a_0}} \frac{1}{n+1}\sum_{k=1}^{n} k^{-\alpha}$$
$$\leq \frac{L_\delta B_\infty}{n+1} + \frac{L_\delta 2^{1/2+2\alpha}\sigma c_\gamma}{\sqrt{a_0}(1-\alpha)} \frac{1}{(n+1)^\alpha} \tag{35}$$

Bounding $M_n := \frac{1}{n+1}\sqrt{\mathbb{E}\left[\left\|\sum_{k=0}^{n}\xi_{k+1}\right\|^2\right]}$. Remark that by definition of ξ_{n+1} and thanks to Assumption **(A1)**, one has

$$\mathbb{E}\left[\|\xi_{n+1}\|^2 |\mathcal{F}_n\right] = \mathbb{E}\left[\|\nabla_h g(X_{n+1},\theta_n)\|^2 |\mathcal{F}_n\right] - \|\nabla G(\theta_n)\|^2 \leq \mathbb{E}\left[\|\nabla_h g(X_{n+1},\theta_n)\|^2 |\mathcal{F}_n\right]$$
$$\leq C_1 + C_2\left(G(\theta_n) - G(\theta)\right)$$

Furthermore, since (ξ_{n+1}) is a sequence of martingale differences adapted to the filtration $(\mathcal{F}_n)$ and applying Hölder inequality, one has

$$M_n = \frac{1}{n+1}\sqrt{\sum_{k=0}^{n}\mathbb{E}\left[\|\xi_{k+1}\|^2\right]} \leq \frac{\sqrt{C_1}}{\sqrt{n+1}} + \frac{\sqrt{C_2}}{n+1}\sqrt{\sqrt{u_0} + \sum_{k=1}^{n}\sqrt{\mathbb{E}\left[\left(G(\theta_n) - G(\theta)\right)^2\right]}}.$$

Thanks to Lemma 3.1, it comes

$$M_n \leq \frac{\sqrt{C_1}}{\sqrt{n+1}} + \frac{\sqrt{C_2}}{n+1}\overbrace{\sqrt{\sqrt{u_0} + \sum_{k=1}^{n} e^{-\frac{1}{8}c_\gamma a_0 n^{1-\alpha}} e^{a_1 c_\gamma^2 \frac{2\alpha}{2\alpha-1} + a_2 c_\gamma^3 \frac{3\alpha}{3\alpha-1}} \left(\sqrt{u_0} + \sigma c_\gamma^{3/2}\sqrt{\frac{3\alpha}{3\alpha-1}}\right)}}^{\leq\sqrt{B_\infty}}$$
$$+ \frac{\sqrt{C_2}}{n+1}\sqrt{\sum_{k=1}^{n}\frac{2^{1/2+2\alpha}\sigma c_\gamma}{\sqrt{a_0}}k^{-\alpha}}$$

Finally, it comes

$$M_n \leq \frac{\sqrt{C_1}}{\sqrt{n+1}} + \frac{\sqrt{C_2}\sqrt{B_\infty}}{n+1} + \frac{\sqrt{C_2}2^{1/4+\alpha}\sqrt{\sigma}\sqrt{c_\gamma}}{a_0^{1/4}\sqrt{1-\alpha}} \frac{1}{(n+1)^{1/2+\alpha/2}} \tag{36}$$

Conclusion: Thanks to inequalities (33) to (35), one has

$$\lambda_{\min}\sqrt{\mathbb{E}\left[\left\|\overline{\theta}_n - \theta\right\|^2\right]} \leq \frac{\sqrt{C_1}}{\sqrt{n+1}} + \frac{L_\delta 2^{1/2+2\alpha}\sigma c_\gamma}{\sqrt{a_0}(1-\alpha)} \frac{1}{(n+1)^\alpha} + \frac{2^{\frac{1+\alpha}{2}}5\sqrt{C_1}}{\sqrt{c_\gamma}\sqrt{\lambda_{\min}}} \frac{1}{(n+1)^{1-\alpha/2}}$$
$$+ \frac{\sqrt{C_2}2^{1/4+\alpha}\sqrt{\sigma}\sqrt{c_\gamma}}{a_0^{1/4}\sqrt{1-\alpha}(n+1)^{1/2+\alpha/2}} + \frac{2^{1+4\alpha}\sigma L_\delta}{\sqrt{a_0}\lambda_{\min}} \frac{\ln(n+1)}{n+1} + \frac{A_\infty + D_\infty + L_\delta B_\infty + \sqrt{C_2}\sqrt{B_\infty} + c_\gamma^{-1/2}\sqrt{v_0}}{n+1}$$

6.2 Proof of Theorem 4.2

In order to prove Theorem 4.2, we just have to give a better bound of the martingale term $\frac{1}{n+1}\sum_{k=0}^{n} H^{-1}\xi_k$. First, let us recall that

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{k=0}^{n} H^{-1}\xi_{k+1}\right\|^2\right] &\leq \sum_{k=0}^{n}\mathbb{E}\left[\left\|H^{-1}\nabla_h g\left(X_{k+1},\theta_k\right)\right\|^2\right] \\
&\leq \sum_{k=0}^{n}\mathbb{E}\left[\mathrm{Tr}\left(H^{-1}\nabla_h g\left(X_{k+1},\theta_k\right)\nabla_h g\left(X_{k+1},\theta_k\right)^T H^{-1}\right)\right] \\
&= \sum_{k=0}^{n}\mathbb{E}\left[\mathrm{Tr}\left(H^{-1}\underbrace{\mathbb{E}\left[\nabla_h g\left(X_{k+1},\theta_k\right)\nabla_h g\left(X_{k+1},\theta_k\right)^T|\mathcal{F}_k\right]}_{=\Sigma(\theta_k)}H^{-1}\right)\right]
\end{aligned}$$

Since the functional $\Sigma(.)$ is L_Σ-lipschitz and denoting $\Sigma = \Sigma(\theta)$, one has

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{k=0}^{n} H^{-1}\xi_{k+1}\right\|^2\right] &= (n+1)\mathrm{Tr}\left(H^{-1}\Sigma H^{-1}\right) + \sum_{k=0}^{n}\mathbb{E}\left[\mathrm{Tr}\left(H^{-1}\left(\Sigma\left(\theta_k\right)-\Sigma\left(\theta\right)\right)H^{-1}\right)\right] \\
&\leq (n+1)\mathrm{Tr}\left(H^{-1}\Sigma H^{-1}\right) + \frac{L_\Sigma}{\lambda_{\min}^2}\sum_{k=0}^{n}\mathbb{E}\left[\left\|\theta_k-\theta\right\|^2\right] \qquad (37)
\end{aligned}$$

Then, thanks to Theorem 3.1, it comes

$$\begin{aligned}
\sqrt{\mathbb{E}\left[\left\|\sum_{k=0}^{n} H^{-1}\xi_{k+1}\right\|^2\right]} &\leq \sqrt{\mathrm{Tr}\left(H^{-1}\Sigma H^{-1}\right)}\sqrt{n+1} + \frac{\sqrt{L_\Sigma}\sqrt{v_0}}{\lambda_{\min}} + \frac{\sqrt{L_\Sigma}}{\lambda_{\min}}\sqrt{A\sum_{k=1}^{n} e^{-\frac{1}{4}\lambda_{\min}c_\gamma k^{1-\alpha}}} \\
&+ \frac{\sqrt{2}\sqrt{L_\Sigma}\sqrt{c_1}L_\delta}{\lambda_{\min}^2}\sqrt{\sum_{k=1}^{n}\exp\left(-\frac{1}{8}a_0c_\gamma k^{1-\alpha}\right)} + \frac{2^{1+4\alpha}\sqrt{L_\Sigma}\sigma c_\gamma L_\delta}{\sqrt{a_0}\lambda_{\min}^2}\sqrt{\sum_{k=1}^{n}k^{-2\alpha}} \\
&+ \frac{2^{1/2+\alpha/2}\sqrt{C_1}\sqrt{L_\Sigma}\sqrt{c_\gamma}}{\lambda_{\min}^{3/2}}\sqrt{\sum_{k=1}^{n}k^{-\alpha}}
\end{aligned}$$

Then, thanks to Minkovski's inequality and by definition of A_∞ and D_∞,

$$\begin{aligned}
\frac{1}{n+1}\sqrt{\mathbb{E}\left[\left\|\sum_{k=0}^{n} H^{-1}\xi_{k+1}\right\|^2\right]} &\leq \frac{\sqrt{\mathrm{Tr}\left(H^{-1}\Sigma H^{-1}\right)}}{\sqrt{n+1}} + \frac{\sqrt{L_\Sigma}\sqrt{v_0}}{\lambda_{\min}(n+1)} + \frac{\sqrt{L_\Sigma}c_\gamma A_\infty}{\lambda_{\min}(n+1)} \\
&+ \frac{\sqrt{L_\Sigma}c_\gamma D_\infty}{\lambda_{\min}(n+1)} + \frac{2^{1+4\alpha}\sqrt{L_\Sigma}\sigma c_\gamma L_\delta\sqrt{2\alpha}}{\sqrt{a_0}\lambda_{\min}^2\sqrt{2\alpha-1}(n+1)} \\
&+ \frac{2^{1/2+\alpha/2}\sqrt{C_1}\sqrt{L_\Sigma}\sqrt{c_\gamma}}{\lambda_{\min}^{3/2}\sqrt{1-\alpha}(n+1)^{1/2+\alpha/2}},
\end{aligned}$$

which concludes the proof.

6.3 Proof of Theorem 4.3

In order to prove Theorem 4.3, let us bound each term on the right hand-side of equality (32).

Bounding $\sqrt{\mathbb{E}\left[\left\|\frac{\theta_{n+1}-\theta}{\gamma_{n+1}(n+1)}\right\|^2\right]}$**:** Thanks to Theorem 3.2, one has

$$\sqrt{\mathbb{E}\left[\left\|\frac{\theta_{n+1}-\theta}{\gamma_{n+1}(n+1)}\right\|^2\right]} \leq \frac{\sqrt{A'}e^{-\frac{1}{2}\lambda_{\min}c_\gamma n^{1-\alpha}}}{c_\gamma(n+1)^{1-\alpha}} + \frac{\sqrt{c_{n_0'}}L_\delta}{c_\gamma\lambda_{\min}(n+1)^{1-\alpha}}\exp\left(-\frac{1}{8}a_0c_\gamma n^{1-\alpha}\right) + \frac{\sigma L_\delta\sqrt{M_0}}{\lambda_{\min}(n+1)} + \frac{2^{\frac{\alpha}{2}}\sqrt{C_1}}{\sqrt{\lambda_{\min}}}\frac{1}{\sqrt{c_\gamma}(n+1)^{1-\alpha/2}}$$

Bounding $R_n := \frac{1}{n+1}\sqrt{\mathbb{E}\left[\left\|\sum_{k=1}^n(\theta_k-\theta)\left(\frac{1}{\gamma_{k+1}}-\frac{1}{\gamma_k}\right)\right\|^2\right]}$**.** Recalling that

$$R_n \leq \frac{c_\gamma^{-1}}{n+1}\sum_{k=1}^n\sqrt{\mathbb{E}\left[\|\theta_k-\theta\|^2\right]}k^{\alpha-1}$$

Denoting

$$A'_\infty := \frac{\sqrt{A'}}{c_\gamma}\sum_{n=0}^{+\infty}e^{-\frac{1}{2}\lambda_{\min}c_\gamma n^{1-\alpha}} \qquad \text{and} \qquad D'_\infty := \frac{\sqrt{c_{n_0'}}L_\delta}{\lambda_{\min}c_\gamma}\sum_{n=0}^{+\infty}e^{-\frac{1}{8}a_0c_\gamma n^{1-\alpha}}$$

it comes

$$\begin{aligned}R_n &\leq \frac{A'_\infty+D'_\infty}{n+1} + \frac{\sigma L_\delta\sqrt{M_0}}{\lambda_{\min}(n+1)}\sum_{k=1}^n k^{-1} + \frac{2^{\frac{\alpha}{2}}\sqrt{C_1}}{\sqrt{\lambda_{\min}}\sqrt{c_\gamma}}\frac{1}{n+1}\sum_{k=1}^n k^{\alpha/2-1}\\ &\leq \frac{A'_\infty+D'_\infty}{n+1} + \frac{\sigma L_\delta\sqrt{M_0}}{\lambda_{\min}(n+1)}\frac{\ln(n+1)}{n+1} + \frac{2^{1+\frac{\alpha}{2}}\sqrt{C_1}}{\alpha\sqrt{\lambda_{\min}}\sqrt{c_\gamma}}\frac{1}{(n+1)^{1-\alpha/2}}\end{aligned}$$

Bounding $R'_n = \frac{1}{n+1}\sqrt{\mathbb{E}\left[\|\sum_{k=0}^n\delta_k\|^2\right]}$**.** Let us recall that

$$R'_n \leq \frac{1}{n+1}\sum_{k=0}^n\sqrt{\mathbb{E}\left[\|\delta_n\|^2\right]} \leq \frac{L_\delta\sqrt{u_0}}{n+1} + \frac{L_\delta}{n+1}\sum_{k=1}^n\sqrt{\mathbb{E}\left[\left(G\left(\theta_n\right)-G(\theta)\right)^2\right]}.$$

Furthermore, denoting

$$B'_\infty = \left(\sqrt{c_{n_0'}}+\sqrt{u_0}\right)\sum_{n\geq 0}\exp\left(-\frac{1}{4}a_0c_\gamma n^{1-\alpha}\right)$$

and with the help of Lemma 3.2, one has

$$R'_n \leq \overbrace{\frac{L_\delta\sqrt{u_0}}{n+1} + \frac{L_\delta\sqrt{c_{n'_0}}}{n+1}\sum_{k=1}^{n}\exp\left(-\frac{1}{4}a_0c_\gamma k^{1-\alpha}\right)}^{\leq \frac{L_\delta B'_\infty}{n+1}} + L_\delta\sigma c_\gamma\sqrt{M_0}\frac{1}{n+1}\sum_{k=1}^{n}k^{-\alpha}$$
$$\leq \frac{L_\delta B'_\infty}{n+1} + \frac{L_\delta\sigma c_\gamma\sqrt{M_0}}{(1-\alpha)(n+1)^\alpha}$$

Bounding M_n: Recalling that

$$M_n \leq \frac{\sqrt{C_1}}{\sqrt{n+1}} + \frac{\sqrt{C_2}}{n+1}\sqrt{\sqrt{u_0} + \sum_{k=1}^{n}\sqrt{\mathbb{E}\left[\left(G\left(\theta_n\right) - G(\theta)\right)^2\right]}}.$$

and since $C_2 = 0$, one has

$$M_n \leq \frac{\sqrt{C_1}}{\sqrt{n+1}}$$

which concludes the proof.

6.4 Proof of Theorem 4.4

In order to prove Theorem 4.4, we just have to give a better bound of the martingale term $\frac{1}{n+1}\sum_{k=0}^{n}H^{-1}\xi_{k+1}$. Thanks to inequality (37) couple with Theorem 3.2, it comes

$$\sqrt{\mathbb{E}\left[\left\|\sum_{k=0}^{n}H^{-1}\xi_{k+1}\right\|^2\right]} \leq \sqrt{\mathrm{Tr}\left(H^{-1}\Sigma H^{-1}\right)}\sqrt{n+1} + \frac{\sqrt{L_\Sigma}\sqrt{v_0}}{\lambda_{\min}} + \frac{\sqrt{L_\Sigma}}{\lambda_{\min}}\sqrt{A'\sum_{k=1}^{n}e^{-\lambda_{\min}c_\gamma n^{1-\alpha}}}$$
$$+ \frac{\sqrt{L_\Sigma}\sqrt{c_{n'_0}}L_\delta}{\lambda_{\min}^2}\sqrt{\sum_{k=1}^{n}e^{-\frac{1}{8}a_0c_\gamma n^{1-\alpha}}} + \frac{\sqrt{L_\Sigma}L_\delta c_\gamma}{\lambda_{\min}^2}\sqrt{M_0}\sqrt{\sum_{k=1}^{n}k^{-2\alpha}}$$
$$+ \frac{\sqrt{L_\Sigma}2^{\frac{\alpha}{2}}\sqrt{C_1}}{\lambda_{\min}^{3/2}}\sqrt{\sum_{k=1}^{n}k^{-\alpha}}.$$

Then, by Minkowski's inequality, it comes

$$\sqrt{\mathbb{E}\left[\left\|\sum_{k=0}^{n}H^{-1}\xi_{k+1}\right\|^2\right]} \leq \sqrt{\mathrm{Tr}\left(H^{-1}\Sigma H^{-1}\right)}\sqrt{n+1} + \frac{\sqrt{L_\Sigma}\sqrt{v_0}}{\lambda_{\min}} + \frac{\sqrt{L_\Sigma}}{\lambda_{\min}}c_\gamma A'_\infty$$
$$+ \frac{\sqrt{L_\Sigma}}{\lambda_{\min}}c_\gamma D'_\infty + \frac{\sqrt{L_\Sigma}L_\delta c_\gamma}{\lambda_{\min}^2}\sqrt{M_0}\sqrt{\frac{2\alpha}{2\alpha-1}}$$
$$+ \frac{\sqrt{L_\Sigma}2^{\frac{\alpha}{2}}\sqrt{C_1}}{\lambda_{\min}^{3/2}}\frac{1}{\sqrt{1-\alpha}}(n+1)^{\frac{1-\alpha}{2}}.$$

References

Alfarra, M., Hanzely, S., Albasyoni, A., Ghanem, B., and Richtarik, P. (2020). Adaptive learning of the optimal mini-batch size of sgd. *arXiv preprint arXiv:2005.01097*.

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

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

Bercu, B., Costa, M., and Gadat, S. (2020). Stochastic approximation algorithms for superquantiles estimation. *arXiv preprint arXiv:2007.14659*.

Boyer, C. and Godichon-Baggioni, A. (2020). On the asymptotic rate of convergence of stochastic newton algorithms and their weighted averaged versions. *arXiv preprint arXiv:2011.09706*.

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

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

Cohen, K., Nedić, A., and Srikant, R. (2017). On projected stochastic gradient descent algorithm with weighted averaging for least squares regression. *IEEE Transactions on Automatic Control*, 62(11):5974–5981.

Costa, M. and Gadat, S. (2020). Non asymptotic controls on a recursive superquantile approximation.

Défossez, A., Bottou, L., Bach, F., and Usunier, N. (2020). A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

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

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

Godichon-Baggioni, A. (2019a). Lp and almost sure rates of convergence of averaged stochastic gradient algorithms: locally strongly convex objective. *ESAIM: Probability and Statistics*, 23:841–873.

Godichon-Baggioni, A. (2019b). Online estimation of the asymptotic variance for averaged stochastic gradient algorithms. *Journal of Statistical Planning and Inference*, 203:1–19.

Konečnỳ, J., Liu, J., Richtárik, P., and Takáč, M. (2015). Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):242–255.

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

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

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

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

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

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