WORKING PAPERS

N° 856

February 2022

“Optimal non-asymptotic analysis of the Ruppert-Polyak averaging stochastic algorithm”

Sébastien Gadat and Fabien Panloup

Toulouse School of Economics

Optimal non-asymptotic analysis of the Ruppert-Polyak averaging stochastic algorithm

Sébastien Gadat, Fabien Panloup

February 8, 2022

Abstract

This paper is devoted to the non-asymptotic analysis of the Ruppert-Polyak averaging method introduced in [26] and [28] for the minimization of a smooth function f with a stochastic algorithm. We first establish a general non-asymptotic optimal bound: if $\hat{\theta}_n$ is the position of the algorithm at step n, we prove that

$$\mathbb{E}|\hat{\theta}_n - \arg\min(f)|^2 \leqslant \frac{\mathrm{Tr}(\Sigma^\star)}{n} + C_{d,f} n^{-r_\beta},$$

where $\Sigma^\star$ is the limiting covariance matrix of the CLT demonstrated in [26] and $C_{d,f}n^{-r_\beta}$ is a new state-of-the-art second order term that translates the effect of the dimension. We also identify the optimal gain of the baseline SGD $\gamma_n = \gamma n^{-3/4}$, leading to a second-order term with $r_{3/4} = 5/4$. Second, we show that this result holds under some Kurdyka-Łojiasewicz-type condition [21, 22] for function f, which is far more general than the standard uniformly strongly convex case. In particular, it makes it possible to handle some pathological examples such as on-line learning for logistic regression and recursive quantile estimation.

1 Introduction

We consider the problem of minimizing $f : \mathbb{R}^d \to \mathbb{R}$ when $f \in \mathcal{C}^2(\mathbb{R}^d, \mathbb{R})$, $\lim_{|\theta| \to +\infty} f(\theta) = +\infty$ and $\theta^\star$ is the unique critical point of f, so that $\theta^\star = \mathrm{argmin}(f)$. Let us assume that ∇f admits the following representation: a measurable function $\Lambda : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^d$ and a random variable Z with values in $\mathbb{R}^p$ exist such that:

$$\forall \theta \in \mathbb{R}^d, \quad \nabla f(\theta) = \mathbb{E}_{Z\sim\mu}[\Lambda(\theta, Z)]. \tag{1}$$

Without loss of generality, we assume in the paper that $f(\theta^\star) = 0$.

1.1 Averaging principle for stochastic algorithms

Stochastic Gradient Descent The Robbins-Monro procedure (see [27]) is built with an i.i.d. sequence of observations $(Z_i)_{i\geqslant 1}$ distributed according to μ. Under some mild assumptions, the minimizers of f can be approximated with a stochastic gradient descent (SGD) $(\theta_n)_{n\geqslant 0}$ defined by: $\theta_0 \in \mathbb{R}^d$ and

$$\forall n \geqslant 0, \quad \theta_{n+1} = \theta_n - \gamma_{n+1}\Lambda(\theta_n, Z_{n+1}), \tag{2}$$

where $(\gamma_n)_{n\geqslant 1}$ is a non-increasing gain sequence of positive numbers such that:

$$\gamma_n = \gamma n^{-\beta} \quad \text{with}\, \beta \in [1/2, 1) \quad \text{and} \quad \Gamma_n = \sum_{k=1}^n \gamma_k \sim \frac{\gamma}{1-\beta} n^{1-\beta}.$$

Equation (2) is sometimes written as a noisy gradient descent:

$$\forall n \geqslant 0 : \quad \theta_{n+1} = \theta_n - \gamma_{n+1}\nabla f(\theta_n) + \gamma_{n+1}\Delta\mathcal{M}_{n+1}, \tag{3}$$

where $(\Delta\mathcal{M}_{n+1})_{n\geqslant 0}$ stands for a sequence of noises (martingale increments), *i.e.* $\forall n \geqslant 1$, $\mathbb{E}\left[\Delta\mathcal{M}_{n+1}|\mathcal{F}_n\right] = 0$, where $(\mathcal{F}_n)_{n\geqslant 0}$ is the filtration defined by $\mathcal{F}_n = \sigma(Z_1, \ldots, Z_n)$ for $n \geqslant 1$, $\mathcal{F}_0$ is the trivial σ-field and for a given σ-field $\mathcal{G}$, $\mathbb{E}[\,.\,|\mathcal{G}]$ stands for the related conditional expectation.

Averaging The Ruppert-Polyak averaging procedure (referred to as RP below) consists in introducing a Cesaro average over the past iterations of the SGD:

$$\hat{\theta}_n = \frac{1}{n}\sum_{k=1}^{n}\theta_k, \quad n \geqslant 1.$$

This averaging procedure is a way to improve the convergence properties of the original SGD $(\theta_n)_{n\geqslant 1}$. We recall the CLT associated with $(\hat{\theta}_n)_{n\geqslant 0}$, the statement is adapted from [26][1] in the strongly convex situation $SC(\alpha)$:

$$SC(\alpha) := \left\{ f \in \mathcal{C}^2(\mathbb{R}^d) \, : \, D^2 f - \alpha I_d \geqslant 0 \right\} \tag{4}$$

where $D^2 f$ stands for the Hessian matrix of f and inequality $A \geqslant 0$ for any matrix A has to be understood in the sense of quadratic forms.

Theorem 1 (Ruppert-Polyak CLT)**.** *Assume $f \in SC(\alpha)$, $\|D^2 f\|_\infty < \infty$ and $\lim_n \mathbb{E}[\Delta\mathcal{M}_{n+1}\Delta\mathcal{M}_{n+1}^T|\mathcal{F}_n] = S^\star$ in probability, then:*

$$\sqrt{n}(\hat{\theta}_n - \theta^\star) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma^\star) \, with \, \Sigma^\star = \{D^2 f(\theta^\star)\}^{-1} S^\star \{D^2 f(\theta^\star)\}^{-1}. \tag{5}$$

This result is achieved *asymptotically* in the situation where f is assumed to be *strongly uniformly convex* (For the sake of simplicity, we will only write *strongly convex* in the rest of the paper). We refer to [26] for the initial asymptotic description and to [20] for some more general results. In [6], a non-asymptotic result is obtained in the strongly convex situation under restrictive moment assumptions on the noisy gradients. The problem is also tackled non asymptotically in some specific cases when the strong convexity property fails (on-line logistic regression [3], recursive median estimation [11, 16] for example). But a non-asymptotic result for a more general class of functions that preserves a sharp optimal $O(n^{-1})$ rate of the $\mathbb{L}^2$-risk is missing yet.

$\mathbb{L}^p$ rates Beyond the $\mathbb{L}^2$-risk of the original SGD and of the averaged sequence, a popular alternative is also to study some more general $\mathbb{L}^p$-risk for a general $p \geqslant 2$. Of course, such results are interesting by themselves and we refer to [17] for a specific study of the geometric median estimation problem, and to [18] for a more general study in locally strongly convex problems. But $\mathbb{L}^p$-risks represent also a common intermediary step to derive some $\mathbb{L}^2$-risk results for the averaged sequence with the help of a linearization of the drift term induced by averaging. This is for example the case when looking for either asymptotic results (see *e.g.* [25]) or non-asymptotic ones in specific situations (in the case of the logistic regression, we refer to [5] for instance).

Optimality and dimensional effect The bias-variance decomposition of the mean square error (M.S.E.) associated with Theorem 1 induces that we cannot expect a behaviour of the M.S.E. lower than $\text{Tr}(\Sigma^\star)n^{-1}$, which is the variance brought by the Gaussian limit. Therefore, we will refer to a *non-asymptotic optimal* M.S.E. upper bound as soon as we obtain an upper bound that holds *for any n* such that the first order term is $\text{Tr}(\Sigma^\star)n^{-1}$:

$$\mathbb{E}[|\hat{\theta}_n - \theta^\star|^2] \leqslant \text{Tr}(\Sigma^\star)n^{-1} + a_2 n^{-\rho}. \tag{6}$$

We emphasize that the leading term $\text{Tr}(\Sigma^\star)n^{-1}$ corresponds to the Cramer-Rao lower-bound in some specific cases of statistical models, so that it is also commonly admitted that the RP averaging cannot be improved to obtain a lower variance (asymptotically or not) with any other estimation method.

Finally, we observe that $\text{Tr}(\Sigma^\star)$ generally grows with the dimension of the ambient space d (of course it depends on the nature of $\Sigma^\star$), and so is expected the second order term with a_2 in (6). As a common nowadays statistical paradigm, we will pay a specific attention to the effect of d on the second order term in (6).

Below, we will obtain an *optimal* upper bound with the desired and unimprovable $\text{Tr}(\Sigma^\star)n^{-1}$ leading term. Nevertheless, we do not know at this stage whether the second order term essentially parametrized by a_2 and $\rho > 1$ is also optimal or not.

[1] In [26], the result is stated in a more general framework with the help of a Lyapunov function. We have chosen to simplify the statement for the sake of readability.

1.2 Main contribution of the paper

We will prove a non-asymptotic result on a large set of functions that satisfy a Kurdyka-Łojasiewicz inequality:

Global KL inequality $(\mathbf{H^r_{KL}})$ The function f is $\mathcal{C}^2(\mathbb{R}^d, \mathbb{R})$ with $D^2 f$ bounded and Lipschitz, $D^2 f(\theta^\star)$ invertible and for $r \in [0, 1/2]$:

$$\liminf_{|x| \longrightarrow +\infty} f^{-r}|\nabla f| > 0 \qquad \text{and} \qquad \limsup_{|x| \longrightarrow +\infty} f^{-r}|\nabla f| < +\infty \tag{7}$$

We establish the next result (whose statement will be more general later on):

Theorem 2. *Assume* $(\mathbf{H^r_{KL}})$. *Suppose that the covariance of the martingale increment is Lipschitz continuous and that* $\mathbb{E}[|\Delta \mathcal{M}_{n+1}|^6 e^{(1+|\Delta \mathcal{M}_{n+1}|^4)^{1/2-r}}|\mathcal{F}_n] < +\infty$ *a.s.. Then, a constant* C_r *exists such that:*

$$\mathbb{E}(|\hat{\theta}_n - \theta^\star|^2) \leqslant \frac{\mathrm{Tr}(\Sigma^\star)}{n} + C_r \left(\frac{\sqrt{d}}{\underline{\mu}} \right)^3 n^{-(\beta+1/2)\wedge(2-\beta)},$$

where $\underline{\mu}$ *is the lowest eigenvalue of* $D^2 f(\theta^\star)$.

Our main result is therefore an optimal non-asymptotic bound of the $\mathbb{L}^2$-risk for the RP-algorithm under some very general assumptions beyond the traditional convexity point of view. Our bound is optimal at the first order since it attains the Cramer-Rao lower bound (*i.e.* rate in $O(\mathrm{Tr}(\Sigma^\star) n^{-1})$ with the lowest possible variance) and provides a second order term which is better than other results of the literature (see Table 1 for details).

Our proof strategy will be splitted into two steps. In a first stage, we obtain a general theorem under a so-called *consistency* assumption on the original SGD $(\theta_n)_{n \geqslant 0}$ (see Section 2.2). In a second stage, we show that this consistency assumption holds in the strongly convex case but also under the *Kurdyka- Łojasiewicz inequality* $(\mathbf{H^r_{KL}})$ (see [21, 22]), which is a much weaker situation than the traditionnal strongly convex settings. This second part leads to some considerable improvements of state of the art results since important applications are not tackled by the strongly convex setting: typically on-line logistic regression or recursive quantile approximation (among others). A range of applications are listed in the next table, enriched by a comparison with existing results in the literature:

	Setting	Cramer-Rao	2^{nd} order υ_n	$\gamma_n = \gamma_1 n^{-\beta}$	Anytime
Our work	Strong. Convex Convex (Smooth KL) Logist. Reg. (KL) Recurs. Quantile (KL)	Yes : $\frac{\mathrm{Tr}(\Sigma^\star)}{n}$	$n^{-(\beta+\frac{1}{2})\wedge(2-\beta)}$, $\upsilon_n^\star = O(n^{-\frac{5}{4}})$	$\beta \in (1/2, 1)$ $\beta^\star = 3/4$	Yes
BM(11) [6]	Strong. Convex	Yes : $\frac{\mathrm{Tr}(\Sigma^\star)}{n}$	$n^{-(\beta+\frac{1}{2})\wedge(\frac{3}{2}-\beta)}$, $\upsilon_n^\star = O(n^{-\frac{7}{6}})$	$\beta \in (1/2, 1)$ $\beta^\star = 2/3$	Yes
BM(11) [6]	Convex Logist. Reg. Recurs. Quantile	No: $O(n^{-1/2})$ No: $O(n^{-1/2})$ $\varnothing$	$\varnothing$	$\beta = 1/2$	Yes
B(14) [3]	Logist. Reg.	No: $O\left(\frac{1}{n\lambda^2_{min}\{D^2 f(\theta^\star)\}}\right)$	$\varnothing$	$\beta = 1/2$	No
CCGB(17) [16]	Recurs. Quantile	No: $O\left(\frac{1}{n}\right)$	$n^{-(\beta+\frac{1}{2})\wedge(\frac{3}{2}-\beta)}$, $\upsilon_n^\star = O(n^{-\frac{7}{6}})$	$\beta \in (1/2, 1)$ $\beta^\star = 2/3$	Yes

Table 1: Overview of our results and comparisons with the literature. $\upsilon_n^\star$ refers to the optimal (smallest) size of the second-order term when β is chosen equal to $\beta^\star$.

2 Main results

This section presents our main notations and our precise statements.

2.1 Notations

For any vector $y \in \mathbb{R}^d$, y^T is the transpose of y and $|y|$ is the Euclidean norm. The set $\mathcal{M}_d(\mathbb{R})$ refers to the set of squared real matrices of size $d \times d$ and the tensor product $\otimes 2$ is used to refer to the following quadratic form:

$$\forall M \in \mathcal{M}_d(\mathbb{R}) \quad \forall y \in \mathbb{R}^d \qquad My^{\otimes 2} = y^T M y.$$

I_d is the identity matrix and $\mathcal{O}_d(\mathbb{R})$ denotes the set of orthonormal matrices:

$$\mathcal{O}_d(\mathbb{R}) := \left\{ Q \in \mathcal{M}_d(\mathbb{R}) \,:\, Q^T Q = I_d \right\}.$$

Finally, the notation $\| \, . \, \|$ corresponds to the operatorial norm on $\mathcal{M}_d(\mathbb{R})$:

$$\|A\| = \sqrt{\rho(A^T A)},$$

where $\rho(A^T A)$ refers to the largest eigenvalue of $A^T A$. In the meantime, for any twice differentiable function f, we introduce the following notation:

$$\rho_\infty(f) := \sup_{x \in \mathbb{R}^d} \|D^2 f(x)\|,$$

which is the largest eigenvalue of $D^2 f$ over the state space. We also define $\underline{\mu} = \min(1, Sp(D^2 f(\theta^\star)))$ and the Lipschitz constant:

$$\|D^2 f\|_{\text{Lip}} := \inf\{c \geqslant 1 \,:\, \forall (x,y) \in \mathbb{R}^d \quad \|D^2 f(x) - D^2 f(y)\| \leqslant c\,|x-y|\}.$$

For two positive sequences $(a_n)_{n \geqslant 1}$ and $(b_n)_{n \geqslant 1}$, the notation $a_n \lesssim_{id} b_n$ refers to a domination relationship, *i.e.* $a_n \leqslant c\, b_n$ **where $c > 0$ is independent of n and of the dimension of the ambient space d.** The binary relationship $a_n = \mathcal{O}_{id}(b_n)$ then holds if and only if $|a_n| \lesssim_{id} |b_n|$.

2.2 Non asymptotic adaptive and optimal inequality

We state our main general result (Theorem 3) under some general assumptions on the noise part and on the behavior of the L^p-norm of the **SGD procedure** $(\theta_n)_{n \geqslant 1}$ (*(L^p,$\sqrt{\gamma_n}$)-consistency*). We introduce the next property:

Definition 1 ((L^p,$\sqrt{\gamma_n}$)-consistency). *A SGD sequence $(\theta_n)_{n \geqslant 1}$ satisfies the (L^p,$\sqrt{\gamma_n}$)-consistency if:*

$$\exists\, c_p \geqslant 1, \quad \forall n \geqslant 1 \qquad \mathbb{E}|\theta_n|^p \leqslant c_p \{\gamma_n\}^{\frac{p}{2}}.$$

Note that according to the Jensen inequality, the (L^p,$\sqrt{\gamma_n}$)-consistency implies the (L^q,$\sqrt{\gamma_n}$)-consistency for any $0 < q < p$ with $c_q \leqslant \{c_p\}^{q/p}$.

The above definition refers to the behaviour of the SGD $(\theta_n)_{n \geqslant 1}$ defined by Equation (2). We will prove that it is a key property to derive sharp non-asymptotic bounds for the RP-algorithm $(\hat{\theta}_n)_{n \geqslant 1}$ (see Theorem 3 below).

We introduce an assumption on the covariance of the martingale increment:

Assumption ($\mathbf{H_S}$) *The covariance of the martingale* (3) *satisfies:*

$$\mathbb{E}\left[\Delta \mathcal{M}_{n+1} \Delta \mathcal{M}_{n+1}^t | \mathcal{F}_n\right] = S(\theta_n) \qquad a.s.$$

where $S : \mathbb{R}^d \to \mathcal{M}_d(\mathbb{R})$ is a Lipschitz continuous function:

$$\exists L > 0 \quad \forall (\theta_1, \theta_2) \in \mathbb{R}^d \qquad \|S(\theta_1) - S(\theta_2)\| \leqslant L|\theta_1 - \theta_2|.$$

The smallest value of L is denoted by $\|S\|_{\text{Lip}}$.

When compared to Theorem 1, Assumption $(\mathbf{H_S})$ is more restrictive but in fact corresponds to the usual framework. Under additional technicalities, this assumption may be relaxed to a local Lipschitz behaviour of S. For reasons of clarity, we preferred to reduce our purpose to this reasonable setting.

Theorem 3 (Optimal non-asymptotic bound). *If $(\theta_n)_{n\geqslant 1}$ is $(L^4, \sqrt{\gamma_n})$-consistent, if $(\mathbf{H}_S)$ holds and $D^2 f(\theta^\star)$ is positive-definite, then for any n:*

$$\mathbb{E}|\hat{\theta}_n - \theta^\star|^2 \leqslant \frac{\text{Tr}(\Sigma^\star)}{n} + C_\beta(c_4, f, S)\left(\frac{\sqrt{d}}{\underline{\mu}}\right)^3 n^{-r_\beta} \textit{ with} \, r_\beta = \left(\beta + \frac{1}{2}\right) \wedge (2 - \beta), \qquad (8)$$

and $\Sigma^\star$ is defined in Equation (5) (with $S^\star = S(\theta^\star)$). In particular, $r_\beta > 1$ for all $\beta \in (1/2, 1)$ and $\beta \longmapsto r_\beta$ attains its maximum for $\beta = 3/4$ and $r_{3/4} = 5/4$.

The quantity $C_\beta(c_4, f, S)$ is made precise in Proposition 7 . Theorem 3 deserves several remarks.

• *Sharpness of the first order term*: we obtain the exact optimal rate $O(n^{-1})$ with the sharp constant $\text{Tr}(\Sigma^\star)$ as shown by Theorem 1. At the first order, Theorem 3 shows that the averaging is minimax optimal with respect to the Cramer-Rao lower bound. The result is adaptive with respect to the value of the Hessian $D^2 f(\theta^\star)$: any sequence $\gamma_n = \gamma n^{-\beta}$ with $\beta \in (1/2, 1)$ and $\gamma > 0$, regardless the value of β or γ, produces the result of Theorem 3. Such an adaptive property does not hold for the initial sequence $(\theta_n)_{n\geqslant 1}$ as proved by the CLT satisfied by the SGD $(\theta_n)_{n\geqslant 1}$ (see [13] for example).

• *Second order term*: Even though any value of $\beta \in (1/2, 1)$ yields a $\frac{\text{Tr}(\Sigma^\star)}{n}$ leading term, the "optimal" choice of β remains unclear. In [6] and [16], $\beta = 2/3$ is motivated by the optimization of the second order term. In particular, [6] obtains in the strongly convex case an upper bound of the order $\frac{\text{Tr}(\Sigma^\star)}{n} + O(n^{-7/6})$: Theorem 3 of [6] ensures that:

$$\sqrt{\mathbb{E}|\hat{\theta}_n - \theta^\star|^2} \leqslant \sqrt{\frac{\text{Tr}(\Sigma^\star)}{n}} + Cn^{-2/3},$$

which in turn implies that $\mathbb{E}|\hat{\theta}_n - \theta^\star|^2 \leqslant \frac{\text{Tr}(\Sigma^\star)}{n} + 2C\sqrt{\text{Tr}(\Sigma^\star)}\frac{n^{-2/3}}{\sqrt{n}} + C^2 n^{-4/3} = \frac{\text{Tr}(\Sigma^\star)}{n} + O\left(n^{-7/6}\right)$.

Our Theorem 3 improves this second order term: $\beta = 3/4$ leads to an upper bound of the order $\frac{\text{Tr}(\Sigma^\star)}{n} + O(n^{-5/4})$. Moreover, for any value of $\beta \in (1/2, 1)$, the second order term $O(n^{-(\beta+1/2)\wedge(2-\beta)})$ in Theorem 3 is always better than $O(n^{-(\beta+1/2)\wedge(3/2-\beta)})$, the one of [6]. For further comments on this topic (including the particular case of null third derivatives), we refer to Section 3.3.

Our result is also related to some recent works on some Berry-Esseen upper bounds derived for the CLT stated in Theorem 1. Corollary 5 of [1] shows that for any Lipschitz and twice differentiable function h:

$$\left|\mathbb{E}h[\sqrt{n}(\hat{\theta}_n - \theta^\star)] - \mathbb{E}h(Z)\right| \leqslant C\frac{d^2}{\sqrt{n}},$$

where Z is a multivariate Gaussian random variable $\mathcal{N}(0, \Sigma^\star)$. However, this result is not exactly of the same nature. Actually, in order to derive M.S.E. bounds (or at least L^1-bounds), this would involve to apply the above result to $h(z) = |z|^2$ (or to $h(z) = |z|$ for the L^1-error). But this is not possible since $h(z) = |z|^2$ is not Lipschitz and $h(z) = |z|$ is not differentiable everywhere. Furthermore, the bounds obtained in [1] seem to strongly depend on these assumptions and it is not clear that technical extensions would allow to include such types of functions. Thus, even though these techniques lead to sharp bounds, they seem to not be adapted to derive second-order bounds for the M.S.E. of the L^1-error.

Finally, we point out that the effect of the dimension d and of the lowest eigenvalue $\underline{\mu}$ is sligthly stronger on the second order term (proportional to n^{-r_β} up to $d^{3/2}\underline{\mu}^{-3}$) than on the first order one (proportional to

n^{-1} up to $d\underline{\mu}^{-2}$). To the best of our knowlegde, Theorem 3 is the first non-asymptotic regret analysis of the Ruppert-Polyak algorithm for a large classe of functions and identifies a dimension-dependent upper bound $d^{3/2}\underline{\mu}^{-3}n^{-r_\beta}$.

• *Idea of the proof*: the proof of Theorem 3 is achieved through a spectral analysis of the (non-homogeneous) second-order Markov chain induced by $(\theta_n, \hat{\theta}_n)_{n\geqslant 1}$. This spectral analysis requires a preliminary linearization step of the drift from $\hat{\theta}_n$ to $\hat{\theta}_{n+1}$. The cost of this linearization is absorbed by a preliminary control of the initial sequence $(\theta_n)_{n\geqslant 1}$, obtained with the $(L^p, \sqrt{\gamma_n})$-consistency for $p = 4$ (see Proposition 1 and Theorem 5). We emphasize that this linearization applies regardless the global assumptions on the objective function: we only impose a local curvature near $\theta^\star$.

• *Anytime strategy*: an important feature of on-line optimization algorithm is the *anytime property*, *i.e.*, the ability of the algorithm to produce an optimal performance regardless the choice of the stopping iteration time since in many situations the final number of iterations is not known in advance. In general, such an anytime property fails when the step-size sequence depends on the final horizon. One way to bypass this issue is to use the doubling trick strategy (see, *e.g.* [12]) , which produces an anytime algorithm and that degrades the final rate with a multiplicative log term. However, for the RP algorithm, such a doubling trick on the initial SGD sequence is questionable: there is no recursive expression for the RP averaging associated with the doubling trick strategy.

As indicated in our Theorem 3, in [6] (for strongly convex function) and [16] (quantile estimation), the sequence $(\gamma_n)_{n\geqslant 1}$ is chosen independently of the final horizon time, and the procedure is therefore anytime. Oppositely, the step-size sequence proposed in [3] highly depends on the final number of iterations (the proposed sequence is proportional to $\frac{1}{2R^2\sqrt{n}}$ where n is the stopping time: *i.e.* the strategy of [3] for on-line logistic regression is not anytime).

2.3 $(L^p,\sqrt{\gamma_n})$-consistency

In Theorem 3, the control of the moments of the SGD sequence $(\theta_n)_{n\geqslant 1}$ is fundamental to derive (8). We first deal with the strongly convex case.

2.3.1 $(L^p,\sqrt{\gamma_n})$-consistency with strong convexity

Here, we introduce an additional condition on the noise, denoted by $(\mathbf{H}^{\mathbf{SC}}_{\mathbf{\Sigma_p}})$.
Assumption $(\mathbf{H}^{\mathbf{SC}}_{\mathbf{\Sigma_p}})$ *For a given $p \in \mathbb{N}^\star$, a constant Σ_p exists such that*

$$\forall n \geqslant 0 \qquad \mathbb{E}[|\Delta\mathcal{M}_{n+1}|^{2p}|\mathcal{F}_n] \leqslant \Sigma_p(1 + (f(\theta_n))^p \qquad a.s.$$

We emphasize that even though $SC(\alpha)$ is a potentially restrictive assumption on f, $(\mathbf{H}^{\mathbf{SC}}_{\mathbf{\Sigma_p}})$ is not restrictive and allows a polynomial dependency in $f(\theta_n)$ of the moments of $\Delta\mathcal{M}_n$, which is much weaker than the bounded increments used in [6]. For example, such an assumption holds in the case of the recursive linear least square problem. In that case, we retrieve the baseline assumption introduced in [13] that only provides an almost sure convergence of $(\theta_n)_{n\geqslant 1}$ towards $\theta^\star$ without any rate. In this setting, we can state the following proposition.

Proposition 1. *Assume that f is $SC(\alpha)$, $x \longmapsto D^2f(x)$ is Lipschitz bounded and $(\Delta\mathcal{M}_n)_{n\geqslant 1}$ satisfies $(\mathbf{H}^{\mathbf{SC}}_{\mathbf{\Sigma_p}})$. Then $(\theta_n)_{n\geqslant 1}$ is $(L^p,\sqrt{\gamma_n})$-consistent, i.e.*

$$\forall p \geqslant 1 \quad \exists\, c_p > 0 \qquad \mathbb{E}|\theta_n - \theta^\star|^p \leqslant c_p\{\gamma_n\}^{p/2}.$$

The proof works using an induction on p, initialized at $p = 1$ for integer values of p, and then may be generalized to any $p \geqslant 1$ thanks to the Jensen inequality. The proof of this result is well known in the strongly convex situation and is left to the reader. Up to some minor modifications, the main arguments are also contained in the more general result stated in Theorem 11 whose proof is given in Section 4.2. A direct consequence of Proposition 1 and Theorem 3 is the next corollary.

Corollary 4. *If* $(\mathbf{H_S})$ *and the assumptions of Proposition 1 hold, then:*

$$\forall n \in \mathbb{N}^\star \qquad \mathbb{E}\left[|\hat{\theta}_n - \theta^\star|^2\right] \leqslant \frac{\mathrm{Tr}(\Sigma^\star)}{n} + C_\beta(c_4, f, S)\left(\frac{\sqrt{d}}{\underline{\mu}}\right)^3 n^{-r_\beta}$$

where r_β *is defined in Theorem 3 and* $C_\beta(c_4, f, S)$ *in Proposition 7.*

2.3.2 $(L^p, \sqrt{\gamma_n})$-consistency without strong convexity

In some interesting cases, the latter $SC(\alpha)$ is not suitable because the repelling effect towards $\theta^\star$ of $\nabla f(x)$ is not strong enough for large values of $|x|$: this is the case for the logistic regression and the recursive quantile where the function ∇f is asymptotically flat for large values of $|x|$. Motivated by these examples, we generalize the class of functions f for which the $(L^p, \sqrt{\gamma_n})$-consistency property holds. For this purpose, we define Assumption $(\mathbf{H}_\phi)$ by:

Assumption $(\mathbf{H}_\phi)$ *$D^2 f$ is bounded and Lipschitz, $D^2 f(\theta^\star)$ invertible and:*

- *i)* ϕ *is* $\mathcal{C}^2(\mathbb{R}_+, \mathbb{R}_+)$ *non-decreasing and* $\exists\, x_0 \geqslant 0 \,:\, \forall x \geqslant x_0,\ \phi''(x) \leqslant 0.$
- *ii) Two positive numbers* m *and* M *exist such that* $\forall x \in \mathbb{R}^d \backslash \{\theta^\star\}$*:*

$$0 < m \leqslant \phi'(f(x))|\nabla f(x)|^2 + \frac{|\nabla f(x)|^2}{f(x)} \leqslant M. \tag{9}$$

Roughly speaking, ϕ quantifies the deficit of convexity far from $\theta^\star$.
When $\phi \equiv 1$, we recover the previous case: $SC(\alpha) \implies (\mathbf{H}_\phi)$ with $\phi \equiv 1$. Actually, in this case, $\alpha_1 > 0$ and $\alpha_2 > 0$ exist such that

$$\frac{\alpha_1}{2}|x - \theta^\star|^2 \leqslant f(x) \leqslant \frac{\alpha_2}{2}|x - \theta^\star|^2, \quad \text{and} \quad \alpha_1|x - \theta^\star| \leqslant |\nabla f(x)| \leqslant \alpha_2|x - \theta^\star|.$$

But $(\mathbf{H}_\phi)$ is more general since it can be true even when $D^2 f$ vanishes.
The opposite case is $\phi(x) = x$. In this setting, $(\mathbf{H}_\phi)$ is satisfied when $m \leqslant |\nabla f(x)|^2 \leqslant M$ with some positive m and M. Note that this framework includes the logistic regression and the recursive quantile (see Subsection 2.5).

For practical purposes, we introduced in Section 1.2 a parametric version of Assumption $(\mathbf{H}_\phi)$ denoted by $(\mathbf{H^r_{KL}})$, which may be seen as a global *Kurdyka-Lojasiewicz gradient inequality* (see, *e.g.* [21, 22] and Subsection 2.4 for details). $(\mathbf{H}_\phi)$ and $(\mathbf{H^r_{KL}})$ are linked by the following proposition.

Proposition 2. $(\mathbf{H^r_{KL}}) \implies (\mathbf{H}_\phi)$ *for any non-decreasing* $\mathcal{C}^2$*-function* $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ *such that* $\phi(u) = u^{1-2r}$ *on* $[1, +\infty)$*. Furthermore,*

$$\liminf_{|x| \to +\infty} f(x)|x|^{-\frac{1}{1-r}} > 0. \tag{10}$$

The implication is easy to prove: near $\theta^\star$, $f(x) \lesssim_{id} |x - \theta^\star|^2$ and $|x - \theta^\star| \lesssim |\nabla f(x)|$ since $\nabla f(\theta^\star) = 0$ and $D^2 f(\theta^\star) > 0$ (in the sense of symmetric matrices) so that $x \mapsto \frac{|\nabla f(x)|^2}{f(x)}$ is lower-bounded by a positive constant near $\theta^\star$. Since $\theta^\star$ is the unique critical point of f, we can also repeat the same argument on any compact set of $\mathbb{R}^d$ since f is twice differentiable and ∇f a continuous function. For large values of $|x|$, the lower-bound of (9) is a direct consequence of (7). Finally, the upper-bound is a consequence of the fact that $\|D^2 f\|_\infty < +\infty$ and from (7) again. The proof of (10) is postponed to Appendix 4.2. Note that this property will be important to derive the $(L^p, \sqrt{\gamma_n})$-consistency (see Theorem 5). Further comments are postponed to Subsection 2.4 and the rest of this paragraph is devoted to the main consequences of $(\mathbf{H}_\phi)$ and $(\mathbf{H^r_{KL}})$.

As in $SC(\alpha)$, Assumptions $(\mathbf{H}_\phi)$ and $(\mathbf{H^r_{KL}})$ need to be combined with some (more stringent) assumption on the martingale increment:

Assumption $(\mathbf{H}^{\phi}_{\bar{\mathbf{\Sigma}}_{\mathbf{p}}})$ A constant $\bar{\Sigma}_p$ exists such that:

$$\forall n \geqslant 0, \qquad \mathbb{E}[|\Delta\mathcal{M}_{n+1}|^{2p+2} e^{\phi(\gamma_1 |\Delta\mathcal{M}_{n+1}|^2)} | \mathcal{F}_n] \leqslant \bar{\Sigma}_p \qquad \text{a.s.} \tag{11}$$

Remark 1. *The general form of this assumption can be roughly explained as follows: the main idea of Theorem 5 below is to use the function $x \mapsto f^p(x) e^{\phi(f(x))}$ to obtain a contraction property. When $(\Delta\mathcal{M}_n)_{n\geqslant 1}$ is bounded, $(\mathbf{H}^{\phi}_{\bar{\mathbf{\Sigma}}_{\mathbf{p}}})$ is automatically satisfied (this is the case for the recursive quantile and for the logistic regression of bounded variables: see Subsection 2.5). In some cases, Assumption $(\mathbf{H}^{\phi}_{\bar{\mathbf{\Sigma}}_{\mathbf{p}}})$ may appear a little bit restrictive since it asks for some exponential moment on the noise $\Delta\mathcal{M}_{n+1}$ that applies at each iteration of the algorithm. However, note that in [Bach, 2014], the assumption is clearly stronger since the work requires that the noisy gradients are bounded almost surely. In particular, the assumption of [Bach, 2014] relies on $\nabla f(\theta_n) + \Delta\mathcal{M}_{n+1}$ and not simply on $\Delta\mathcal{M}_{n+1}$. Implicitely, it introduces a kind of boundedness assumption on the sequence $(\theta_n)_{n\geqslant 1}$ itself. Secondly, our assumption $(\mathbf{H}^{\phi}_{\bar{\mathbf{\Sigma}}_{\mathbf{p}}})$ introduces a kind of continuum effect between strongly and weakly convex cases through the effect of the function ϕ, which typically evolves like $\phi(u) = u^{1-2r}$ with r between 0 (very weakly convex case) and $1/2$ (strongly convex case).*

We state the main result for a potentially non-convex function f.

Theorem 5. *For any $p \geqslant 1$, if f satisfies $(\mathbf{H}_{\phi})$ and $(\mathbf{H}^{\phi}_{\bar{\mathbf{\Sigma}}_{\mathbf{p}}})$ holds, then:*

i) A constant c_p exists such that:

$$\mathbb{E}[f^p(\theta_n) e^{\phi(f(\theta_n))}] \leqslant c_p \{\gamma_n\}^p.$$

ii) If $\liminf_{|x| \to +\infty} |x|^{-2p} f^p(x) e^{\phi(f(x))} > 0$, then $(\theta_n)_{n\geqslant 1}$ is $(L^{2p}, \sqrt{\gamma_n})$-consistent:

$$\mathbb{E}|\theta_n - \theta^\star|^{2p} \leqslant c_p \{\gamma_n\}^p.$$

iii) If $(\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}})$ holds, $(\theta_n)_{n\geqslant 1}$ is $(L^{2p}, \sqrt{\gamma_n})$-consistent.

Proof. The proof of *i)* is postponed to Section 4 and is stated in Theorem 11.
ii) is a consequence of *i)*: actually, we only need to prove that the function $\tau(x) = f^p(x) e^{\phi(f(x))}$, $x \in \mathbb{R}^d$, satisfies $\inf_{x \in \mathbb{R}^d \setminus \{0\}} \tau(x) |x - \theta^\star|^{-2p} > 0$. Near $\theta^\star$, $D^2 f(\theta^\star)$ is positive-definite and $x \mapsto \tau(x)|x - \theta^\star|^{-2p}$ is lower-bounded by a positive constant. Since τ is positive on $\mathbb{R}^d$, the result follows from the additional assumption $\liminf_{|x| \to +\infty} \tau(x)|x|^{-2p} > 0$.
Finally, for *iii)*, we have to prove that the additional assumption of *ii)* holds under $(\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}})$. It is a consequence of (10) and $\phi(x) = (1 + |x|^2)^{\frac{1-2r}{2}}$. ▫ □

Theorem 3 allows to derive non-asymptotic bounds under $(\mathbf{H}_{\phi})$.

Corollary 6. *Assume $(\mathbf{H_S})$, $(\mathbf{H}_{\phi})$ and $(\mathbf{H}^{\phi}_{\bar{\mathbf{\Sigma}}_{\mathbf{p}}})$ with $p = 2$, then:*

$$\forall n \in \mathbb{N}^\star \qquad \mathbb{E}\left[|\hat{\theta}_n - \theta^\star|^2\right] \leqslant \frac{\mathrm{Tr}(\Sigma^\star)}{n} + C_\beta(c_4, f, S) \left(\frac{\sqrt{d}}{\underline{\mu}}\right)^3 n^{-r_\beta},$$

where c_4 is given in Theorem 11 and r_β is defined in Theorem 3.

At first sight, the result brought by Corollary 6 may appear surprising: we obtain a $O(1/n)$ rate for the mean-squared error of the averaged sequence towards $\theta^\star$ *without strong convexity*, including, for example, some situations where $f(x) \sim |x|$ as $|x| \to +\infty$. However, this result does not contradict the minimax rate of convergence $O(1/\sqrt{n})$ for stochastic optimization problems in the simple convex case (see, *e.g.* [2] or [23]). The above minimax result $O(1/\sqrt{n})$ simply refers to the worst situation in the class of convex functions *that are* *not necessarily differentiable*, whereas $(\mathbf{H}_\phi)$ describes a set of functions that are not necessarily strongly convex or even simply convex, but all these functions belong to $\mathcal{C}^2(\mathbb{R}^d,\mathbb{R})$ and have a positive curvature around $\theta^\star$. In particular, the worst case is attained in [2] through linear combinations of shifted piecewise affine functions $x \longmapsto |x+1/2|$ and $x \longmapsto |x-1/2|$, functions for which Assumption $(\mathbf{H}_\phi)$ is obviously not satisfied. According to the results in Appendix H of [4], the local curvature near $\theta^\star$ makes it possible to obtain a $O(n^{-1})$ rate whereas the smoothness assumption allows to obtain a precise constant, leading to the Cramer-Rao lower bound in the specific setting of [4].

2.4 Comments on $(\mathbf{H}_\phi)$ and link with the Kurdyka-Łojasiewicz inequality

To the best of our knowledge, this is the first work that uses this in stochastic optimization and it thus deserves several comments.

f does not necessarily need to be convex It is important to notice that the function f itself is not necessarily assumed to be convex. The minimal requirement is that f possesses a unique critical point. Our analysis will be based on a descent lemma for the SGD $(\theta_n)_{n\geqslant 0}$. We will use a Lyapunov analysis that will involve $f^p e^{\phi(f)}$ instead of f itself for the sequence $(\theta_n)_{n\geqslant 0}$. The descent property will then be derived from Equation (9) in *ii)* of $(\mathbf{H}_\phi)$. Thereafter, we will be able to exploit a spectral analysis of the dynamical system that governs $(\hat{\theta}_n)_{n\geqslant 0}$. We stress that usually the results without any convexity assumption are usually limited to almost sure convergence with the help of the Robbins-Siegmund Lemma (see, *e.g.* [13]). As will be shown later on, $(\mathbf{H}_\phi)$ will be sufficient to derive efficient convergence rates for the averaged sequence $(\hat{\theta}_n)_{n\geqslant 0}$ *without any strong convexity.*

f is necessarily sub-quadratic and L-smooth $(\mathbf{H}_\phi)$ entails an a priori upper bound for f that cannot increase faster than a quadratic form. We have:

$$\begin{aligned}\forall x\in\mathbb{R}^d \qquad \frac{|\nabla f(x)|^2}{f(x)} \leqslant M \quad &\Longrightarrow \quad |\nabla(\sqrt{f})| \leqslant \frac{\sqrt{M}}{2} \\ &\Longrightarrow \quad f(x) \leqslant \frac{M}{4}\|x-\theta^\star\|^2.\end{aligned}$$

However, we also need a slightly stronger condition with $D^2 f$ bounded over $\mathbb{R}^d$, meaning that f is L-smooth for a suitable value of L (with an L-Lipschitz gradient). We refer to [24] for an introduction to this class of functions. Even in deterministic settings, the L-smooth property is a minimal requirement for good convergence rates in smooth optimization problems (see, *e.g.* [7]).

About the Kurdyka-Łojasiewicz inequality $(\mathbf{H}_\phi)$ should be related to the KL inequalities. The Łojasiewicz gradient inequality [22] with exponent r is:

$$\exists m>0 \quad \exists\, r\in[0,1) \quad \forall x\in\mathbb{R}^d \qquad f(x)^{-r}|\nabla f(x)| \geqslant m, \tag{12}$$

while a generalization (see, *e.g.*, [21]) is governed by the existence of a concave increasing "desingularizing" function ψ such that: $|\nabla(\psi\circ f)| \geqslant 1$. The Łojasiewicz gradient inequality is then just a particular case of the previous inequality while choosing $\psi(t)=ct^{1-r}$. We refer to [8] that characterizes some large families of functions f such that a generalized KL-inequality holds.

In this paper, the KL-type gradient inequality appears through $(\mathbf{H}_{\mathbf{KL}}^{\mathbf{r}})$ with $r\in[0,1/2]$, which implies $(\mathbf{H}_\phi)$ (see Proposition 2). However, it should be noticed that $(\mathbf{H}_{\mathbf{KL}}^{\mathbf{r}})$ is slightly different from (12) since we only enforce the function $f^{-r}|\nabla f|$ to be *asymptotically* lower-bounded by a positive constant.

In fact, in our setting where f has only one critical point and where $D^2 f(\theta^\star)>0$, it is easy to prove that $(\mathbf{H}_{\mathbf{KL}}^{\mathbf{r}})$ implies (12) everywhere: around $\theta^\star$, $D^2 f(\theta^\star)$ is positive definite so that we could choose $r=1/2$ and then satisfy the Łojasiewicz gradient inequality (12) near $\theta^\star$ so that the link between $(\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}})$ given in (7) and (12) has to be understood for large values of $|x|$.

Moreover, Proposition 2 states that the classical Łojasiewicz gradient inequality (12) associated with the assumption of **local** invertibility of $D^2 f(\theta^\star)$ implies $(\mathbf{H}_\phi)$. The choice $r = 1/2$ in Equation (12) corresponds to the strongly-convex case with $\phi = 1$ and $\psi(t) = \sqrt{t}$. Conversely, the Łojasiewicz exponent $r = 0$ corresponds to the weak repelling force $|\nabla f(x)|^2 \propto 1$ as $|x| \to +\infty$ and $\phi(t) = \sqrt{1+t^2}$, leading to $\psi(t) = t$.

Finally, the interest of $(\mathbf{H}_\phi)$ in the stochastic framework is related to the behavior of the algorithm when $(\theta_n)_{n\geqslant 1}$ is far from $\theta^\star$, whereas in the deterministic framework, the main interest of the desingularizing function ψ is used around $\theta^\star$ to derive fast linear rates even in non strongly convex situations (see *e.g.* [9]). The difficulty to assert some good properties of stochastic algorithms is not the same as the one for deterministic problems: it is more difficult to control the time for a stochastic algorithm to come back far from $\theta^\star$ than for a deterministic method with a weakly reverting effect of $-\nabla f$ because of the noise on the algorithm. In contrast, the rate of a deterministic method crucially depends on the local behavior of ∇f around $\theta^\star$ (see, *e.g.* [9]).

Dissipative condition We also observe from Proposition 2 that $(\mathbf{H}_\phi)$ has no prior link with a dissipativity condition standardly used in theory of P.D.E. and stochastic processes for assessing trend to equilibrium of dynamical systems

$$f(x) \geqslant \alpha |x|^2 - \beta.$$

Consider $f(x) = |x|^\rho$, we verify easily that dissipativity holds for $\rho \geqslant 2$ whereas $(\mathbf{H}_\phi)$ is verified when $\rho \leqslant 2$.

Counter-examples of the global KL inequality Finally, we should have in mind what kind of functions do not satisfy the global Łojasiewicz inequality (12). Since we assumed f to have a unique minimizer $\theta^\star$ with $D^2 f(\theta^\star)$ invertible, $f^{-r}|\nabla f| \geqslant m > 0$ should only fail asymptotically. From Equation (10) of Proposition 2, we know that $|x| \lesssim_{id} f(x)$ for large values of $|x|$. As a consequence, any function f with logarithmic growth or comparable to $|x|^r$ growth with $r \in (0,1)$ at infinity can not be managed by this assumption. Another counter-example occurs when f exhibits an infinite sequence of oscillations in the values of $f' \geqslant 0$ with longer and longer areas near $f' = 0$ when $|x|$ is increasing. We refer to [9] for the following function that does not satisfy KL for any $r \geqslant 2$: $f(x) = x^{2r}[2 + \cos(x^{-1})]$ if $x \neq 0$ and $f(0) = 0$.

2.5 Applications

Strongly convex situation First, Corollary 4 provides a very tractable criterion to assess the non-asymptotic first-order optimality of the averaging procedure since $SC(\alpha)$ is easy to check. For example, considering the **recursive mean square estimation** problem (see, *i.e.*, [13]), $\theta \longrightarrow f(\theta)$ is quadratic. In that case, the problem is strongly convex, and the noise increment satisfies:

$$\mathbb{E}[|\Delta \mathcal{M}_n|^{2p} | \mathcal{F}_n] \leqslant \Sigma_p (1 + (f(\theta_n))^p \qquad \text{a.s.}$$

Then Proposition 1 yields the $(L^p, \sqrt{\gamma_n})$ consistency rate of $(\theta_n)_{n\geqslant 1}$, which implies a first-order optimal excess risk for $(\hat{\theta}_n)_{n\geqslant 1}$ with a $O(n^{-5/4})$ second-order term. We stress that [6] also proves a sharp non-asymptotic $O(1/n)$ rate of convergence with a $O(n^{-7/6})$ second-order term and a more restrictive assumption on $\Delta\mathcal{M}_n$. Hence, Corollary 4 yields a stronger result in that case.

Assumptions $(\mathbf{H}_\phi)$ *and* $(\mathbf{H}^{\phi}_{\mathbf{\Sigma_p}})$ *hold in many situations*

- *Semi-algebraic case* Before explicit examples, an argument relies on the statement of Theorem 2 of [8]: a coercive convex proper and semi-algebraic continuous function f (see [8] for some details), satisfies the KL inequality.
- *On-line logistic regression* The logistic regression corresponds to:

$$f(\theta) := \mathbb{E}\left[\log\left(1 + e^{-Y<X,\theta>}\right)\right] \tag{13}$$

where X is a $\mathbb{R}^d$ random variable and $Y|X$ takes its value in $\{-1,1\}$ with:

$$P[Y=1\,|X=x]=\frac{1}{1+e^{-<x,\theta^\star>}}. \tag{14}$$

We then observe a sequence of i.i.d. replications (X_i,Y_i) and the SGD is:

$$\theta_{n+1}=\theta_n+\gamma_{n+1}\frac{Y_nX_n}{1+e^{Y_n<\theta_n,X_n>}}=\theta_n-\gamma_{n+1}\nabla f(\theta_n)+\gamma_{n+1}\Delta\mathcal{M}_{n+1}. \tag{15}$$

We state the following result:

Proposition 3. *If the law of X is compactly supported and elliptic: for any $e\in\mathcal{S}^{d-1}(\mathbb{R}^d)$, $Var(<X,e>)>0$. Then*

i) f defined in (13) is convex with D^2f bounded and Lipschitz continous, $D^2f(\theta^\star)$ is invertible and f satisfies $(\mathbf{H^r_{KL}})$ with $r=0$.

ii) If $\Sigma^\star$ is defined in (5), the averaged sequence $(\hat{\theta}_n)_{n\geqslant 1}$ satisfies:

$$\exists\, C_d>0\quad\forall n\geqslant 1\qquad \mathbb{E}|\hat{\theta}_n-\theta^\star|^2\leqslant\frac{\mathrm{Tr}(\Sigma^\star)}{n}+C_dn^{-5/4}.$$

Proof: We study $i)$. Some straightforward computations yield:

$$\nabla f(\theta)=\mathbb{E}\left[\frac{X\left[e^{<X,\theta>}-e^{<X,\theta^\star>}\right]}{\left[1+e^{<X,\theta>}\right]\left[1+e^{<X,\theta^\star>}\right]}\right]\quad\text{and}\quad D^2f(\theta)_{k,l}=\mathbb{E}\left[\frac{X_kX_le^{<X,\theta>}}{(1+e^{<X,\theta>})^2}\right]$$

We deduce that $\nabla f(\theta^\star)=0$ and that (see [3] for example) f is convex with

$$<\theta-\theta^\star,\nabla f(\theta)>=\mathbb{E}\left[\frac{\left[<X,\theta>-<X,\theta^\star>\right]\left[e^{<X,\theta>}-e^{<X,\theta^\star>}\right]}{\left[1+e^{<X,\theta^\star>}\right]\left[1+e^{<X,\theta>}\right]}\right]\geqslant 0,$$

because $(x-y)[e^x-e^y]>0$ for every pair (x,y) such that $x\neq y$. It implies that $\theta^\star$ is the unique minimizer of f. Moreover, $D^2f(\theta^\star)=\mathbb{E}\left[XX^T\frac{e^{<X,\theta^\star>}}{(1+e^{<X,\theta^\star>}}\right]$ is invertible as soon as the design matrix is invertible. This property easily follows from the ellipticity condition on the distribution of the design:

$$\forall e\in\mathcal{S}^{d-1}(\mathbb{R}^d)\qquad Var(<X,e>)=e^T\mathbb{E}[XX^T]e>0,$$

which proves that the Hessian $D^2f(\theta^\star)$ is invertible. Regarding now the asymptotic norm of $|\nabla f(\theta)|$, the Lebesgue Theorem yields, $\forall e\in\mathcal{S}^{d-1}(\mathbb{R}^d)$:

$$\begin{aligned}
\lim_{t\longrightarrow+\infty}|\nabla f(te)| &= \left|\mathbb{E}\left[\frac{X\mathbf{1}_{<X,e>\geqslant 0}-Xe^{<X,\theta^\star>}\mathbf{1}_{<X,e><0}}{1+e^{<X,\theta^\star>}}\right]\right|\\
&\geqslant \left|\left\langle\mathbb{E}\left[\frac{X\mathbf{1}_{<X,e>\geqslant 0}-Xe^{<X,\theta^\star>}\mathbf{1}_{<X,e><0}}{1+e^{<X,\theta^\star>}}\right],e\right\rangle\right|\\
&\geqslant \left|\mathbb{E}\left[\frac{<X,e>\mathbf{1}_{<X,e>\geqslant 0}-<X,e>e^{<X,\theta^\star>}\mathbf{1}_{<X,e><0}}{1+e^{<X,\theta^\star>}}\right]\right|\\
&\geqslant \left|\mathbb{E}\left[\frac{<X,e>\mathbf{1}_{<X,e>\geqslant 0}}{1+e^{<X,\theta^\star>}}\right]\right|\wedge\left|\mathbb{E}\left[\frac{<X,-e>e^{<X,\theta^\star>}\mathbf{1}_{<X,-e>\geqslant 0}}{1+e^{<X,\theta^\star>}}\right]\right|
\end{aligned}$$

where we used the orthogonal decomposition on e and $e^{\perp}$. Hence for any e, $\lim_{t \longrightarrow +\infty} |\nabla f(te)| > 0$. The Cauchy-Schwarz inequality $| < X, \theta^{\star} > | \leqslant |X||\theta^{\star}|$ yields:

$$\frac{< X, e > \mathbf{1}_{<X,e> \geqslant 0}}{1 + e^{<X,\theta^{\star}>}} \geqslant < X, e > e^{-|X||\theta^{\star}|} \frac{\mathbf{1}_{<X,e> \geqslant 0}}{2}.$$

In a same way, we also observe that

$$\frac{< X, -e > e^{<X,\theta^{\star}>} \mathbf{1}_{<X,-e> \geqslant 0}}{1 + e^{<X,\theta^{\star}>}} = \frac{< X, -e > \mathbf{1}_{<X,-e> \geqslant 0}}{1 + e^{-<X,\theta^{\star}>}} \geqslant < X, -e > e^{-|X||\theta^{\star}|} \frac{\mathbf{1}_{<X,-e> \geqslant 0}}{2}.$$

The assumption on the ellipticity of the design X yields $\forall e \in \mathcal{S}^{d-1}(\mathbb{R}^d)$:

$$\left| \mathbb{E} \left[\frac{< X, e > \mathbf{1}_{<X,e> \geqslant 0}}{1 + e^{<X,\theta^{\star}>}} \right] \right| > 0 \qquad \text{and} \qquad \left| \mathbb{E} \left[\frac{< X, -e > e^{<X,\theta^{\star}>} \mathbf{1}_{<X,-e> \geqslant 0}}{1 + e^{<X,\theta^{\star}>}} \right] \right| > 0.$$

Since $\mathcal{S}^{d-1}(\mathbb{R}^d)$ is a compact space and that $e \longmapsto \left| \mathbb{E} \left[\frac{<X,e> \mathbf{1}_{<X,e> \geqslant 0}}{1+e^{<X,\theta^{\star}>}} \right] \right| \wedge \left| \mathbb{E} \left[\frac{<X,-e> e^{<X,\theta^{\star}>} \mathbf{1}_{<X,-e> \geqslant 0}}{1+e^{<X,\theta^{\star}>}} \right] \right|$ is a continuous function (by the Lebesgue continuity theorem), we thus obtain that:

$$\inf_{e \in \mathcal{S}^{d-1}(\mathbb{R}^d)} \left| \mathbb{E} \left[\frac{< X, e > \mathbf{1}_{<X,e> \geqslant 0}}{1 + e^{<X,\theta^{\star}>}} \right] \right| \wedge \left| \mathbb{E} \left[\frac{< X, -e > e^{<X,\theta^{\star}>} \mathbf{1}_{<X,-e> \geqslant 0}}{1 + e^{<X,\theta^{\star}>}} \right] \right| > 0.$$

We then deduce that:

$$\liminf_{|\theta| \longrightarrow +\infty} |\nabla f(\theta)| \geqslant \frac{1}{2} \inf_{e \in \mathcal{S}^{d-1}(\mathbb{R}^d)} \mathbb{E} \left[< X, e >_{+} e^{-|X||\theta^{\star}|} \right] > 0.$$

It is straightforward to check that $\limsup_{|\theta| \longrightarrow +\infty} |\nabla f(\theta)| < +\infty$, which concludes the proof of i).

We now prove ii) and apply Corollary 6. In that case, Assumption $(\mathbf{H}^{\mathbf{r}}_{\mathbf{KL}})$ holds with $r = 0$. Regarding Assumption $(\mathbf{H}^{\phi}_{\mathbf{\Sigma_p}})$, we can observe that the martingale increments are *bounded* owing to the boundedness of X (see [3], for example) and Inequality (11) is satisfied. Hence, Corollary 6 implies that $(\theta_n)_{n \geqslant 1}$ is a L^p-$\{\sqrt{\gamma_n}\}$ consistent sequence for any $p \geqslant 2$. We can therefore apply Theorem 3 for the averaging procedure $(\hat{\theta}_n)_{n \geqslant 1}$, with $\Sigma^{\star}$ given in (5). This ends the proof. □

Recursive quantile The recursive quantile estimation problem is a standard example that may be stated as follows (see, *e.g.* [13] for details). For a given cumulative distribution G defined over $\mathbb{R}$, the problem is to find q_α such that $G(q_\alpha) = 1 - \alpha$. We assume that we observe a sequence of i.i.d. $(X_i)_{i \geqslant 1}$ distributed with a cumulative distribution G. The recursive quantile is then:

$$\theta_{n+1} = \theta_n - \gamma_{n+1} \left[\mathbf{1}_{X_n \leqslant \theta_n} - (1 - \alpha) \right] = \theta_n - \gamma_{n+1} [G(\theta_n) - (1 - \alpha)] + \gamma_{n+1} \Delta \mathcal{M}_{n+1},$$

In that situation, the function f' is defined by:

$$f'(\theta) = \int_{q_\alpha}^{\theta} p(s) \mathrm{d}s = G(\theta) - G(q_\alpha),$$

where p is the density with respect to the Lebesgue measure such that $G(q) = \displaystyle\int_{-\infty}^{q} p$. Below, we consider the case where p is a Lipschitz continuous function with $p(q_\alpha) > 0$. To satisfy $f(q_\alpha) = 0$, we define f by: $f(\theta) := \int_{q_\alpha}^{\theta} \int_{q_\alpha}^{u} p(s) \mathrm{d}s \mathrm{d}u$, whose minimum is 0 and is attained when $\theta = q_\alpha$. It can immediately be checked that $f''(q_\alpha) \neq 0$ as soon as $p(q_\alpha) > 0$ and $f'(\theta) \longrightarrow 1 - \alpha$ when $\theta \longrightarrow +\infty$ while $f'(\theta) \longrightarrow -\alpha$ when $\theta \longrightarrow -\infty$.

Therefore, f satisfies $(\mathbf{H}_\phi)$ since $(\mathbf{H}^r_{KL})$ and Equation (12) hold with $r = 0$ and $\phi(t) = t$. Again, regarding Assumption $(\mathbf{H}^{\phi}_{\mathbf{\Sigma_p}})$, we can observe that the martingale increments are *bounded*. Therefore, Inequality (11) is obviously satisfied since ϕ is a monotone increasing function. Corollary 6 implies that $(\hat{\theta}_n)_{n\geqslant 1}$ satisfies:

$$\forall n \geqslant 1 \qquad \mathbb{E}|\hat{\theta}_n - q_\alpha|^2 \leqslant \frac{\alpha(1-\alpha)}{p(q_\alpha)\, n} + \mathcal{O}\left(\frac{n^{-5/4}}{p(q_\alpha)^3}\right)$$

2.6 Organization of the paper

The rest of the paper is dedicated to the proofs, organized as follows.
In Section 3, we detail our spectral analysis of the behavior of $(\hat{\theta}_n)_{n\geqslant 1}$ and we provide the main tools for the proof of Theorem 3, that we conclude in Section 3.1. In particular, Proposition 7 provides the main argument to derive the sharp exact first-order rate of convergence, and the results postponed below in Section 3 only represent technical lemmas that are useful for the proof of Proposition 7. Section 4 is dedicated to the proof of the $(L^p,\sqrt{\gamma_n})$-consistency under $(\mathbf{H}_\phi)$ (proof of Theorem 5 *i*)). The generalization to the stronger situation of strong convexity (Proposition 1) is left to the reader (it only requires slight changes).

3 Non asymptotic optimal averaging procedure

We first assume without loss of generality that $\theta^\star = 0$ and that $f(\theta^\star) = 0$. Our proof relies on a spectral strategy developed by [15] for the study of the Heavy Ball with Friction stochastic algorithm. For the sake of convenience, we assume below that $\gamma = 1$, which means that $\gamma_n = n^{-\beta}$.

3.1 Proof of Theorem 3

The starting point is to exhibit the coupled dynamics of $(\theta_n, \hat{\theta}_n)$. For this purpose, we introduce the notation for the drift at time n:

$$\Lambda_n := \int_0^1 D^2 f(t\theta_n)\mathrm{d}t \quad \text{so that} \quad \Lambda_n \theta_n = \nabla f(\theta_n), \tag{16}$$

using the Taylor formula and the fact that $\nabla f(\theta^\star) = 0$. The recursive evolution of $(\theta_n, \hat{\theta}_n)$ is then precised in the next proposition.

Proposition 4. *If $Z_n = (\theta_n, \hat{\theta}_n)$, then:*

$$Z_{n+1} = \begin{pmatrix} I_d - \gamma_{n+1}\Lambda_n & 0 \\ \frac{1}{n+1}(I_d - \gamma_{n+1}\Lambda_n) & (1 - \frac{1}{n+1})I_d \end{pmatrix} Z_n + \gamma_{n+1} \begin{pmatrix} \Delta\mathcal{M}_{n+1} \\ \frac{\Delta\mathcal{M}_{n+1}}{n+1} \end{pmatrix}. \tag{17}$$

<u>Proof:</u> We start from $\hat{\theta}_{n+1} = \hat{\theta}_n + \frac{1}{n+1}\left(\theta_{n+1} - \hat{\theta}_n\right)$. Now, Equation (2) yields:

$$\forall n \in \mathbb{N} \qquad \begin{cases} \theta_{n+1} = \theta_n - \gamma_{n+1}\nabla f(\theta_n) + \gamma_{n+1}\Delta\mathcal{M}_{n+1} \\ \hat{\theta}_{n+1} = \hat{\theta}_n(1 - \frac{1}{n+1}) + \frac{1}{n+1}\left(\theta_n - \gamma_{n+1}\nabla f(\theta_n) + \gamma_{n+1}\Delta\mathcal{M}_{n+1}\right). \end{cases}$$

The result then follows from (16). □

The next result describes the linearization (Λ_n is replaced by $\Lambda^\star := D^2 f(\theta^\star)$).

Proposition 5. *$Q \in \mathcal{O}_d(\mathbb{R})$ exists such that $\check{Z}_n = \begin{pmatrix} Q & 0 \\ 0 & Q \end{pmatrix} Z_n$ satisfies:*

$$\check{Z}_{n+1} = A_n \check{Z}_n + \gamma_{n+1} \begin{pmatrix} Q\Delta\mathcal{M}_{n+1} \\ \frac{Q\Delta\mathcal{M}_{n+1}}{n+1} \end{pmatrix} + \underbrace{\gamma_{n+1} \begin{pmatrix} Q(\Lambda^\star - \Lambda_n)\theta_n \\ Q(\Lambda_n - \Lambda^\star)\frac{\theta_n}{n+1} \end{pmatrix}}_{:=\check{v}_n}, \tag{18}$$

where $D^\star$ is the diagonal matrix associated to the eigenvalues of $\Lambda^\star$ and

$$A_n := \begin{pmatrix} I_d - \gamma_{n+1} D^\star & 0 \\ \frac{1}{n+1}(I_d - \gamma_{n+1} D^\star) & (1 - \frac{1}{n+1}) I_d \end{pmatrix}. \tag{19}$$

<u>Proof:</u> We write $\Lambda_n = \underbrace{D^2 f(\theta^\star)}_{:=\Lambda^\star} + (\Lambda_n - D^2 f(\theta^\star))$ and use the spectrum of $\Lambda^\star$.

$$Z_{n+1} = \begin{pmatrix} I_d - \gamma_{n+1}\Lambda^\star & 0 \\ \frac{1}{n+1}(I_d - \gamma_{n+1}\Lambda^\star) & (1 - \frac{1}{n+1}) I_d \end{pmatrix} Z_n + \gamma_{n+1} \begin{pmatrix} \Delta\mathcal{M}_{n+1} \\ \frac{\Delta\mathcal{M}_{n+1}}{n+1} \end{pmatrix} + \upsilon_n, \tag{20}$$

where the term υ_n will be shown to be negligible and is defined by

$$\upsilon_n := \gamma_{n+1} \begin{pmatrix} (\Lambda^\star - \Lambda_n)\theta_n \\ (\Lambda_n - \Lambda^\star)\frac{\theta_n}{n+1} \end{pmatrix}.$$

The matrix $\Lambda^\star$ is the Hessian of f at $\theta^\star$ and is a symmetric positive matrix, which may be reduced into a diagonal matrix $D^\star = Diag(\mu_1^\star, \ldots, \mu_d^\star)$ with positive eigenvalues in an orthonormal basis: $\exists\, Q \in \mathcal{O}_d(\mathbb{R})\, \Lambda^\star = Q^T D^\star Q$ with $Q^T = Q^{-1}$. The new sequence adapted to the spectral decomposition of $\Lambda^\star$ is:

$$\check{Z}_n = \begin{pmatrix} Q & 0 \\ 0 & Q \end{pmatrix} Z_n = \begin{pmatrix} Q\theta_n \\ Q\hat{\theta}_n \end{pmatrix}. \tag{21}$$

Using $Q\Lambda^\star = D^\star Q$, we obtain the equality described in Equation (18). □

An important feature about $(\check{Z}_n)_{n\geqslant 1})$ is the blockwise structure of A_n:

$$A_n = \left(\begin{array}{cc} \begin{bmatrix} 1-\gamma_{n+1}\mu_1^\star & 0 & \ldots & 0 \\ 0 & 1-\gamma_{n+1}\mu_2^\star & \ldots & \vdots \\ \vdots & \ldots & \ddots & \vdots \\ 0 & \ldots & 0 & 1-\gamma_{n+1}\mu_d^\star \end{bmatrix} & \mathbf{0_d} \\ \begin{bmatrix} \frac{1-\gamma_{n+1}\mu_1^\star}{n+1} & 0 & \ldots & 0 \\ 0 & \frac{1-\gamma_{n+1}\mu_2^\star}{n+1} & \ldots & \vdots \\ \vdots & \ldots & \ddots & \vdots \\ 0 & \ldots & 0 & \frac{1-\gamma_{n+1}\mu_d^\star}{n+1} \end{bmatrix} & (1-\frac{1}{n+1})\mathbf{I_d} \end{array} \right). \tag{22}$$

The matrices made of components (i,i) $(i,d+i)$, $(d+i,i)$ and $(d+i,d+i)$ have a similar form, which is the object of the next proposition.

Proposition 6. *For $\mu \in \mathbb{R}$ and $n \geqslant 1$, set $E_{\mu,n} := \begin{pmatrix} 1-\gamma_{n+1}\mu & 0 \\ \frac{1-\mu\gamma_{n+1}}{n+1} & 1-\frac{1}{n+1} \end{pmatrix}$.*

- *If $1 - \mu\gamma_{n+1}(n+1) \neq 0$, define $\epsilon_{\mu,n+1}$ by:*

$$\epsilon_{\mu,n+1} := \frac{1-\mu\gamma_{n+1}}{1-\mu\gamma_{n+1}(n+1)}, \tag{23}$$

The eigenvalues of $E_{\mu,n}$ are then given by $Sp(E_{\mu,n}) = \left\{1-\mu\gamma_{n+1}, 1-\frac{1}{n+1}\right\}$, whereas the associated eigenvectors are:

$$u_{\mu,n} = \begin{pmatrix} 1 \\ \epsilon_{\mu,n+1} \end{pmatrix} \quad \textit{and} \quad v = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

- *If $1 - \mu\gamma_{n+1}(n+1) = 0$, $E_{\mu,n}$ is not diagonalizable in $\mathbb{R}$.*

At this stage, we point out that the eigenvectors are modified from one iteration to another in our spectral analysis of $(\hat{\theta}_n)_{n\geqslant 1}$ (see Lemma 12).

Remark 2. *The spectral decomposition of $E_{\mu,n}$ will be important below.*

- *$E_{\mu,n}$ (and A_n) is not symmetric (see Equation (22)), leading to a non-orthonormal change of basis and some difficulties for the study of $(\check{Z}_n)_{n\geqslant 1}$.*
- *To a lesser extent, it is also interesting to point out that this "no self-adjointness" property of A_n is a new example of acceleration of convergence rates with the help of non symmetric dynamical systems. (see [29, 14, 10, 15]).*
- *The first eigenvalue of $E_{\mu,n}$ is $1-\mu\gamma_{n+1}$, and essentially acts on the component θ_n of the vector Z_n. We recover a standard contraction on the SGD.*
- *Interestingly, the second eigenvalue of $E_{\mu,n}$ is $1-(n+1)^{-1}$, which is* **independent** *of the value of μ. This eigenvalue acts on the component brought by $\hat{\theta}_n$ in the vector Z_n, and is at the core of our study of $(\hat{\theta}_n)_{n\geqslant 1}$.*

From the factorization $E_{\mu,n} = \begin{pmatrix} 1 & 0 \\ \epsilon_{\mu,n+1} & 1 \end{pmatrix}\begin{pmatrix} 1-\mu\gamma_{n+1} & 0 \\ 0 & 1-\frac{1}{n+1} \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -\epsilon_{\mu,n+1} & 1 \end{pmatrix}$, we define the diagonal matrix $\mathcal{E}_{n,D^\star}$ by:

$$\mathcal{E}_{n,D^\star} = Diag(\epsilon_{\mu_1^\star,n+1},\ldots,\epsilon_{\mu_d^\star,n+1}), \tag{24}$$

we then deduce the spectral decomposition of A_n:

$$A_n = \begin{pmatrix} I_d & 0 \\ \mathcal{E}_{n,D^\star} & I_d \end{pmatrix}\begin{pmatrix} I_d-\gamma_{n+1}D^\star & 0 \\ 0 & (1-\frac{1}{n+1})I_d \end{pmatrix}\begin{pmatrix} I_d & 0 \\ -\mathcal{E}_{n,D^\star} & I_d \end{pmatrix}. \tag{25}$$

We introduce the last change of basis as:

$$\widetilde{Z}_n := \begin{pmatrix} I_d & 0 \\ -\mathcal{E}_{n,D^\star} & I_d \end{pmatrix}\check{Z}_n =: \begin{pmatrix} \widetilde{Z}_n^{(1)} \\ \widetilde{Z}_n^{(2)} \end{pmatrix}. \tag{26}$$

For the sequence $(\widetilde{Z}_n)_{n\geqslant 1}$, the following proposition holds:

Proposition 7. *Assume $(\mathbf{H_S})$ and $(\theta_n)_{n\geqslant 1}$ is a $(L^4,\sqrt{\gamma_n})$-consistent:*

- *i) A constant c_4 exists such that: $\forall n\geqslant 1 \qquad \mathbb{E}\left|\widetilde{Z}_n^{(1)}\right|^4 \leqslant c_4\gamma_n^2$.*
- *ii) Let $\beta\in[1/2,1]$, $r_\beta = \{(\beta+1/2)\wedge(2-\beta)\} > 1$ and n_1 be the positive integer defined in Lemma 13. For any $n\geqslant n_1$:*
$$\mathbb{E}\left|\widetilde{Z}_n^{(2)}\right|^2 \leqslant \frac{\mathrm{Tr}(\Sigma^\star)}{n} + C_\beta(c_4,f,S)\left(\frac{\sqrt{d}}{\underline{\mu}}\right)^3 \mathcal{O}_{id}\left(n^{-r_\beta}\right).$$
where $C_\beta(c_4,f,S) = c_4(1-\beta)^{-1}C_{f,S}$ with $C_{f,S} = \|D^2 f\|_{\mathrm{Lip}} + \|S^\star\| + \|S\|_{\mathrm{Lip}}$.

We are driven to the "optimal" choice $\beta = 3/4$, which in turns implies that:

$$\forall n\in\mathbb{N}^\star \qquad \mathbb{E}|\widetilde{Z}_n^{(2)}|^2 \leqslant \frac{\mathrm{Tr}(\Sigma^\star)}{n} + c_4 C_{f,S}\left(\frac{\sqrt{d}}{\underline{\mu}}\right)^3 \mathcal{O}_{id}\left(n^{-5/4}\right).$$

Proof. Proof of *i*): By Equations (21) and (26), $\widetilde{Z}_n^{(1)} = Q\theta_n$. The $(L^p, \sqrt{\gamma_n})$-consistency of $(\widetilde{Z}_n^{(1)})_{n\geqslant 1}$ then comes from the one of $(\theta_n)_{n\geqslant 1}$.

Proof of *ii*): We pick n_0 such that $\forall n \geqslant n_0 : \epsilon_{\mu,n} < 0$ for any $\mu \in Sp(\Lambda^\star)$.

Step 1: Recursion formula. In order to study the behavior of the L^2-norm of $(\widetilde{Z}_n^{(2)})_{n\geqslant 0}$, we first precise the relationship between $\widetilde{Z}_n$ and $\widetilde{Z}_{n+1}$. Equations (18) and (20) combined with definitions (21) and (26) yield:

$$\begin{aligned}
\widetilde{Z}_{n+1} &= \begin{pmatrix} I_d & 0 \\ -\mathcal{E}_{n+1,D^\star} & I_d \end{pmatrix} \check{Z}_{n+1} \\
&= \begin{pmatrix} I_d & 0 \\ -\mathcal{E}_{n+1,D^\star} & I_d \end{pmatrix} \left(A_n \check{Z}_n + \gamma_{n+1} \begin{pmatrix} Q\Delta\mathcal{M}_{n+1} \\ \frac{Q\Delta\mathcal{M}_{n+1}}{n+1} \end{pmatrix} + \check{\upsilon}_n \right) \\
&= \begin{pmatrix} I_d & 0 \\ -\mathcal{E}_{n+1,D^\star} & I_d \end{pmatrix} \begin{pmatrix} I_d & 0 \\ \mathcal{E}_{n,D^\star} & I_d \end{pmatrix} \begin{pmatrix} I_d - \gamma_{n+1} D^\star & 0 \\ 0 & (1 - \frac{1}{n+1}) I_d \end{pmatrix} \widetilde{Z}_n \\
&+ \gamma_{n+1} \left[\begin{pmatrix} Q\Delta\mathcal{M}_{n+1} \\ (-\mathcal{E}_{n+1,D^\star} + \frac{I_d}{n+1}) Q\Delta\mathcal{M}_{n+1} \end{pmatrix} + \begin{pmatrix} Q(\Lambda^\star - \Lambda_n)\theta_n \\ (\mathcal{E}_{n+1,D^\star} - \frac{I_d}{n+1}) Q(\Lambda^\star - \Lambda_n)\theta_n \end{pmatrix} \right],
\end{aligned}$$

where we used the eigenvalues of A_n in (25). $D^2 f$ is Lipschitz so that:

$$\|\Lambda^\star - \Lambda_n\| \leqslant \int_0^1 \|D^2 f(t\theta_n) - D^2 f(0)\| \mathrm{d}t \leqslant \frac{1}{2} \|D^2 f\|_{\mathrm{Lip}} |\theta_n|.$$

Then, we deduce that:

$$\begin{cases} \widetilde{Z}_{n+1}^{(1)} = (I_d - \gamma_{n+1} D^\star) \widetilde{Z}_n^{(1)} + \gamma_{n+1} \left(Q\Delta\mathcal{M}_{n+1} + \mathcal{O}_{id}\left(\|D^2 f\|_{\mathrm{Lip}} |\theta_n|^2 \right) \right) \\ \widetilde{Z}_{n+1}^{(2)} = (1 - \frac{1}{n+1}) \widetilde{Z}_n^{(2)} + \Omega_n \widetilde{Z}_n^{(1)} + \gamma_{n+1} \Upsilon_n \left(Q\Delta\mathcal{M}_{n+1} + \mathcal{O}_{id}\left(\|D^2 f\|_{\mathrm{Lip}} |\theta_n|^2 \right) \right), \end{cases} \tag{27}$$

with $\Omega_n = (\mathcal{E}_{n,D^\star} - \mathcal{E}_{n+1,D^\star})(I_d - \gamma_{n+1} D^\star)$ and $\Upsilon_n = \mathcal{E}_{n+1,D^\star} - \frac{I_d}{n+1}$.

Step 2: $\mathbb{E}[|\tilde{\mathbf{Z}}_{\mathbf{n}}^{(\mathbf{2})}|^{\mathbf{2}}] = \mathcal{O}_{\mathbf{id}}(\mathbf{n}^{-\mathbf{1}})$ We introduce the covariance:

$$\forall i \in \{1, \ldots, d\} \qquad \omega_n(i) = \mathbb{E}[(\widetilde{Z}_n)_i (\widetilde{Z}_n)_{d+i}] = \mathbb{E}[(\widetilde{Z}_n^{(1)})_i (\widetilde{Z}_n^{(2)})_i], \tag{28}$$

and the useful coefficient:

$$\forall i \in \{1, \ldots, d\} \qquad \alpha_n^i = 2\left(1 - \frac{1}{n+1}\right) \{\Omega_n\}_{i,i}. \tag{29}$$

We use the Young inequality $ab \leqslant \frac{\epsilon}{2} a^2 + \frac{1}{2\epsilon} b^2$ with $\epsilon = n^{\beta - \frac{1}{2}}$.

$$\frac{\mathbb{E}[|\theta_n|^2 |\tilde{Z}_n^{(2)}|]}{n} \leqslant \frac{n^{\beta - \frac{1}{2}}}{2n} \mathbb{E}[|\theta_n|^4] + \frac{n^{-\beta + \frac{1}{2}}}{2n} \mathbb{E}[|\tilde{Z}_n^{(2)}|^2] \leqslant \frac{c_4}{2} \gamma_n \left(n^{-\frac{3}{2}} + n^{-\frac{1}{2}} \mathbb{E}[|\tilde{Z}_n^{(2)}|^2] \right).$$

Second, Lemma 13 implies that $i \in \{1, \ldots, d\} : |\alpha_n^i| \lesssim_{id} \underline{\mu}^{-1} n^{\beta - 2}$. Hence

$$\begin{aligned}
\textstyle\sum_{i=1}^d |\alpha_n^i \omega_n(i)| &\leqslant \frac{1}{\underline{\mu} \gamma_n n^2} \left((\underline{\mu} n^{\frac{1}{2} - \frac{\beta}{2}})^{-1} \mathbb{E}|\widetilde{Z}_n^{(1)}|^2 + \underline{\mu} n^{\frac{1}{2} - \frac{\beta}{2}} \mathbb{E}|\tilde{Z}_n^{(2)}|^2 \right) \\
&\leqslant \frac{c_2 n^{-\frac{5}{2} + \frac{\beta}{2}}}{\underline{\mu}^2} + n^{\frac{\beta}{2} - \frac{3}{2}} \mathbb{E}|\tilde{Z}_n^{(2)}|^2.
\end{aligned}$$

We use this inequality into Lemma 13 *ii*), an integer n_1 exists (see Lemma 13):

$$\begin{aligned}\forall n \geqslant n_1 \qquad \mathbb{E}[|\widetilde{Z}^{(2)}_{n+1}|^2] &\leqslant \left(\left(1-\frac{1}{n+1}\right)^2 + \left[n^{-\beta-\frac{1}{2}} + n^{\frac{\beta}{2}-\frac{3}{2}}\right]\right)\mathbb{E}[|\widetilde{Z}^{(2)}_n|^2] \\ &+ \frac{\mathrm{Tr}(\Sigma^\star)}{(n+1)^2} + \mathcal{O}_{id}\left(\frac{c_2 n^{-\frac{5}{2}+\frac{\beta}{2}}}{\underline{\mu}^2} + \frac{c_4 d}{\underline{\mu}^3} C_S\left(n^{-\frac{3}{2}-\beta} \vee n^{-3+\beta}\right)\right) \\ &\leqslant \left(\left(1-\frac{1}{n+1}\right)^2 + C_1 n^{-r}\right)\mathbb{E}[|\widetilde{Z}^{(2)}_n|^2] + \frac{\mathrm{Tr}(\Sigma^\star)}{(n+1)^2} + C_2 n^{-q},\end{aligned}$$

where C_S is defined in Lemma 13, $C_1 := 2$, $r = (\beta + 1/2) \wedge (3/2 - \beta/2)$, $q = (3/2 + \beta) \wedge (5/2 - \beta/2)$ and $C_2 = \mathcal{O}_{id}\left(\frac{c_4 d}{\underline{\mu}^3} C_S\right)$. Setting $N = n_1$ and $u_n = \mathbb{E}[|\widetilde{Z}^{(2)}_n|^2]$, we apply Lemma 16 and deduce that:

$$\forall n \geqslant N \qquad \mathbb{E}[|\widetilde{Z}^{(2)}_n|^2] \leqslant \frac{\mathrm{Tr}(\Sigma^\star)}{n} + \mathcal{O}_{id}\left(\frac{u_{n_1} n_1^2}{n^2} + \mathrm{Tr}(\Sigma^\star) n^{-r} + C_2 n^{-q}\right).$$

Using the arguments of (50), $\mathrm{Tr}(\Sigma^\star) \leqslant d\underline{\mu}^{-2}\|S^\star\|$. The definition of C_S yields:

$$\forall n \geqslant n_1, \qquad \mathbb{E}[|\widetilde{Z}^{(2)}_n|^2] \lesssim_{id} \frac{c_4 d}{\underline{\mu}^3} C_S n^{-1} + \frac{u_{n_1} n_1}{n}.$$

Remark that Lemma 15 entails $u_{n_1} \lesssim_{id} \frac{c_2}{1-\beta} n_1^{-\beta}$ so that:

$$\begin{aligned} u_{n_1} n_1 &\lesssim_{id} \frac{c_2}{1-\beta} n_1^{1-\beta} \lesssim_{id} \frac{c_2}{1-\beta}\left(n_0^{1-\beta} + c_2^{\frac{1-\beta}{\beta}} + \|D^2 f\|_{\mathrm{Lip}}^{1-\beta}\right) \\ &\lesssim_{id} \frac{c_2}{(1-\beta)\underline{\mu}} + \frac{c_2^2}{1-\beta} + \frac{\|D^2 f\|_{\mathrm{Lip}}}{1-\beta} \lesssim_{id} \frac{c_4 d}{(1-\beta)\underline{\mu}^3} C_S,\end{aligned}$$

using in particular that $c_2 \leqslant \sqrt{c_4}$. As a conclusion, we finally get:

$$\forall n \geqslant n_1, \qquad \mathbb{E}[|\widetilde{Z}^{(2)}_n|^2] \lesssim_{id} \frac{c_4 d}{(1-\beta)\underline{\mu}^3} C_S n^{-1}. \tag{30}$$

Step 3: Control of the covariance Inequality (30) yields for $n \geqslant n_1$:

$$\mathbb{E}[|\theta_n|^2 |\tilde{Z}^{(2)}_n|] \leqslant \sqrt{\mathbb{E}[|\theta_n|^4]}\sqrt{\mathbb{E}[|\tilde{Z}^{(2)}_n|^2]} = c_4 \sqrt{\frac{dC_S}{(1-\beta)\underline{\mu}^3}}\,\mathcal{O}_{id}\left(\frac{\gamma_n}{\sqrt{n}}\right). \tag{31}$$

Plugging this control into Lemma 13 *i*), we obtain that for all $i \in \{1, \ldots, d\}$:

$$\left|\omega_{n+1}(i) - (1 - \gamma_{n+1}\mu_i^\star)\frac{n}{n+1}\omega_n(i)\right| \lesssim_{id} C_\omega \frac{\gamma_n}{n} + c_4\|D^2 f\|_{\mathrm{Lip}}\sqrt{\frac{dC_S}{(1-\beta)\underline{\mu}^3}}\frac{\gamma_n^2}{\sqrt{n}}.$$

Now, remark that $\gamma_n \lesssim_{id} n^{-1/2}$ so that we conclude that $\mathbb{E}[|\theta_n|^2 |\tilde{Z}^{(2)}_n|]$ shall be neglected in $(\omega_n(i))_{n \geqslant 1}$. Now, set

$$C_{f,S} = \|D^2 f\|_{\mathrm{Lip}} + C_S + 1.$$

We have

$$\forall n \geqslant 1 \qquad \left|\omega_{n+1}(i) - (1 - \gamma_{n+1}\mu_i^\star)\frac{n}{n+1}\omega_n(i)\right| \leqslant \mathcal{O}_{id}\left(\frac{c_4\sqrt{d}C_{f,S}}{\underline{\mu}^{3/2}}\frac{\gamma_n}{n}\right).$$

From Lemma 14 stated in Appendix 4.2, we conclude that:

$$\forall i \in \{1, \ldots, d\} \qquad |\omega_n(i)| \leqslant \mathcal{O}_{id}\left(\frac{c_4\sqrt{d}C_{f,S}}{\underline{\mu}^{3/2}}\frac{1}{n}\right). \tag{32}$$

Step 4: Conclusion of the proof From (32) and (31), we have:

$$\begin{aligned} \textstyle\sum_{i=1}^d \alpha_n^i \omega_n(i) &= \frac{c_4 d^{3/2} C_{f,S}}{\underline{\mu}^{5/2}} \mathcal{O}_{id}\left(\frac{1}{n^3\gamma_n}\right) \text{ and} \\ \frac{\mathbb{E}[|\theta_n|^2 |\tilde{Z}_n^{(2)}|]}{n} &= c_4 \frac{\sqrt{d}}{\underline{\mu}^{3/2} C_{f,S}} \mathcal{O}_{id}\left(\frac{\gamma_n}{n^{3/2}}\right). \end{aligned}$$

We use these bounds in the statement of Lemma 13 *ii*) and deduce that:

$$\begin{aligned} \mathbb{E}[|\widetilde{Z}_{n+1}^{(2)}|^2] &\leqslant \left(1 - \frac{1}{n+1}\right)^2 \mathbb{E}[|\widetilde{Z}_n^{(2)}|^2] + \frac{\mathrm{Tr}(\Sigma^\star)}{(n+1)^2} \\ &\quad + \frac{c_4 d^{3/2} C_{f,S}}{\underline{\mu}^3} \mathcal{O}_{id}\left(n^{-3+\beta} \vee n^{-3/2-\beta}\right), \end{aligned}$$

where we used that $\gamma_n = n^{-\beta}$ so that $\sqrt{\gamma_n} n^{-2} = o(\gamma_n n^{-3/2})$ regardless the value of $\beta \in (1/2, 1)$. Applying again Lemma 16 with $C_1 = 0$ and $q_\beta = (\frac{3}{2} + \beta) \wedge (3 - \beta)$, we obtain the desired result. □

□

3.2 End of the proof of Theorem 3

To end the study of $(\hat{\theta}_n)_{n \geqslant 1}$, we first remark that for $n \geqslant n_1$:

$$|\hat{\theta}_n|^2 \leqslant 2\left(|\tilde{Z}_n^{(2)}|^2 + \rho\left(\mathcal{E}_{n,D^\star}\right)^2 |\check{Z}_n^{(1)}|^2\right),$$

which in turn implies the desired inequality when $n \geqslant n_1$. When $n \leqslant n_1$, we deduce from Lemma 15 that

$$\mathbb{E}|\hat{\theta}_n|^2 \leqslant \frac{c_2 n^{-\beta}}{1-\beta} \leqslant \frac{c_2}{1-\beta} n_1^{r_\beta - \beta} n^{-r_\beta}.$$

Since $r_\beta - \beta = \frac{1}{2} \wedge \{2(1-\beta)\}$ and:

$$n_1 \lesssim_{id} \underline{\mu}^{-\frac{1}{1-\beta}} + c_2^{\frac{1}{\beta}} + \|D^2 f\|_\infty,$$

we get

$$c_2 n_1^{r_\beta - \beta} \lesssim_{id} \frac{c_2}{1-\beta}\left(\underline{\mu}^{-2} + c_2 + \|D^2 f\|_\infty\right).$$

Up to a universal constant, this last upper bound is smaller than $\frac{c_4 d^{3/2} C_{f,S}}{(1-\beta)\underline{\mu}^3}$. □

3.3 Further remarks on the second order term

When $x \longmapsto D^2 f(x)$ is constant (or also when the function f to minimize is $\mathcal{C}^3$ with third partial derivatives Lipschitz and null at $\theta^\star$), we remark that $\Lambda_n - \Lambda^\star = O(|\theta_n|^2)$. Following the proof of Lemma 13, the error term is replaced by $n^{-1}\mathcal{O}_{id}(\mathbb{E}[|\theta_n|^3 |\tilde{Z}_n^{(2)}|]) \lesssim_{id} (n^{-1}\gamma_n)^{\frac{3}{2}}$ if the $(L^6, \sqrt{\gamma_n})$-consistency holds. Hence, we obtain:

$$\mathbb{E}[|\widetilde{Z}_{n+1}^{(2)}|^2] \leqslant \left(1 - \frac{1}{n+1}\right)^2 \mathbb{E}[|\widetilde{Z}_n^{(2)}|^2] + \mathcal{O}(n^{-3}\gamma_n^{-1}) + \mathcal{O}\left(\frac{\sqrt{\gamma_n}}{n^2}\right),$$

which is a better upper bound (from the point of view of the exponent on n only) comparing to the recursion obtained in the end of the previous proof. The rate is then optimized with $\beta = 2/3$ and $r_\beta = \frac{4}{3}$.

The previous remark shows that we may obtain a different size of the second order terms when f is locally symmetric around $\theta^\star$ (which occurs when $D^3 f(\theta^\star) = 0$) whereas when f is not locally symmetric, Theorem 3 proves that this second order term may be fixed of size $O(n^{-5/4})$. We have computed (with a Monte-Carlo approximation) $n \mapsto n^\rho \left(\mathbb{E}[|\hat{\theta}_n - \theta^\star|^2] - \frac{\mathrm{Tr}(\Sigma^\star)}{n}\right)$ with $\rho = \frac{5}{4}$ and $\beta = \frac{3}{4}$ for a locally non-symmetric f_1 around $\theta^\star$ and $n \mapsto n^\rho \left(\mathbb{E}[|\hat{\theta}_n - \theta^\star|^2] - \frac{\mathrm{Tr}(\Sigma^\star)}{n}\right)$ with $\rho = \frac{4}{3}$ and $\beta = \frac{2}{3}$ for a locally symmetric f_2. We have used $f_1(x) = \frac{x^2}{2} e^{-\arctan(x)}$ and $f_2(x) = \frac{x^2}{2}$, which trivially fall into the two different cases (see Figure 1): our simulations confirm that the second-order terms are of the right sizes and cannot be improved.

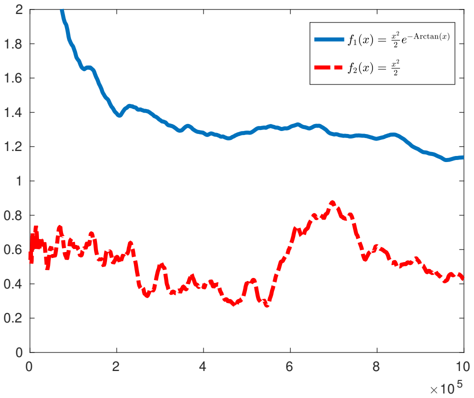

Figure 1: $n \mapsto n^\rho \left(\mathbb{E}[|\hat{\theta}_n - \theta^\star|^2] - \frac{\mathrm{Tr}(\Sigma^\star)}{n}\right)$. Blue curve: $\rho = \frac{5}{4}$ and $\beta = \frac{3}{4}$ for a non locally symmetric function f_1. Red curve: $\rho = \frac{4}{3}$ and $\beta = \frac{2}{3}$ for a locally symmetric function f_2.

4 Proof of the $(L^p, \sqrt{\gamma_n})$-consistency - (Theorem 5)

The main objective of this section is to prove Theorem 5. Our analysis is based on a Lyapunov-type approach with the help of $V_p : \mathbb{R}^d \to \mathbb{R}$ defined as:

$$\forall p \geqslant 1 \qquad V_p(x) = f^p(x) \exp(\phi(f(x))).$$

4.1 Taylor's expansion on V_p

To prove our main result (Theorem 11), we need to establish some technical results related to ϕ and V_p. The first result is a simple sub-additive property on ϕ that essentially relies on the concavity property on $[x_0, +\infty)$.

Lemma 7. *If ϕ satisfies $(\mathbf{H}_\phi)(i)$, then a constant c_ϕ exists such that:*

$$\forall (x, y) \in \mathbb{R}_+ \qquad \phi(x + y) \leqslant \phi(x) + \phi(y) + c_\phi.$$

Proof: Since $\phi'' \leqslant 0$ on $[x_0, +\infty)$, the function ϕ is concave on $[x_0, +\infty)$. Hence, the function $x \mapsto \phi(x+y) - \phi(x)$ is decreasing on $[x_0, +\infty)$ so that:

$$\forall x \geqslant x_0 \quad \phi(x + y) \leqslant \phi(x) + \phi(x_0 + y) - \phi(x_0).$$

Since ϕ' is decreasing on $[x_0, +\infty)$, then ϕ' is upper-bounded and a constant $C > 0$ exists such that $\phi(y + x_0) \leqslant \phi(y) + Cx_0$. We then deduce that:

$$\forall x \geqslant x_0 \quad \forall y \geqslant 0 \qquad \phi(x + y) \leqslant \phi(x) + \phi(y) + Cx_0 - \phi(x_0). \tag{33}$$

In the other situation when $x \leqslant x_0$, the fact that ϕ is non-decreasing yields and Equation (33) applied at point x_0 yields:

$$\phi(x + y) \leqslant \phi(x_0 + y) \leqslant \phi(y) + Cx_0 \leqslant \phi(x) + \phi(y) + Cx_0.$$

We then obtain the desired inequality for any value of x and y in $\mathbb{R}_+$. □

The next result is a straightforward computation left to the reader.

Lemma 8. *For any $p \in \mathbb{N}^\star$ and any $x \in \mathbb{R}^d \backslash \{\theta^\star\}$, we have:*

i)

$$\nabla V_p(x) = V_p(x) \left(p \frac{\nabla f(x)}{f(x)} + \phi'(f(x)) \nabla f(x) \right).$$

ii)

$$D^2 V_p(x) = V_p(x) \left[\psi_1(x) \nabla f(x) \otimes \nabla f(x) + \psi_2(x) D^2 f(x) \right],$$

where ψ_1 and ψ_2 are given by:

$$\psi_1(x) := \left(\frac{p}{f(x)} + \phi'(f(x)) \right)^2 - \frac{p}{f^2(x)} + \phi''(f(x)) \quad \textit{and} \quad \psi_2(x) := \frac{p}{f(x)} + \phi'(f(x)).$$

The next lemma translates the effect of the drift of the algorithm on the exponential function introduced in the definition of V_p.

Lemma 9. *If f satisfies $(\mathbf{H}_\phi)$: $0 < m \leqslant \phi'(f)|\nabla f|^2 + \frac{|\nabla f|^2}{f} \leqslant M$, then*

i)

$$\forall x \in \mathbb{R}^d \qquad \langle \nabla V_p(x), \nabla f(x) \rangle \geqslant m V_p(x).$$

ii)

$$\forall \xi \in \mathbb{R}^d \qquad \rho(D^2 V_p(\xi)) \lesssim_{id} (1 + M^2 + \rho_\infty(f)) \left(V_{p-1}(\xi) + \frac{V_p(\xi)}{1 + |\nabla f(\xi)|^2} \right).$$

Proof: *i)* We apply Lemma 8 *i)* and obtain that:

$$\forall x \in \mathbb{R}^d \backslash \{\theta^\star\} \qquad \frac{\langle \nabla V_p(x), \nabla f(x) \rangle}{V_p(x)} = p \frac{|\nabla f(x)|^2}{f(x)} + \phi'(f(x)) |\nabla f(x)|^2.$$

The result then follows from Assumption $(\mathbf{H}_\phi)$ *ii)*.
ii) We apply Lemma 8 *ii)*. We have that $\forall y \in \mathbb{R}^d$:

$$\begin{aligned}
\frac{\langle y, D^2 V_p(\xi) y \rangle}{\|y\|^2} &= \frac{V_p(\xi)}{\|y\|^2} \left[\psi_1(\xi) \langle y, \nabla f(\xi) \otimes \nabla f(\xi) y \rangle + \psi_2(\xi) \langle y, D^2 f(\xi) y \rangle \right] \\
&\leqslant V_p(\xi) \left(\left[\frac{2p^2}{f^2(\xi)} + 2\{\phi'(f(\xi))\}^2 - \frac{p}{f^2(\xi)} + \phi''(f(\xi)) \right] |\nabla f(\xi)|^2 + \left[\frac{p}{f(\xi)} + \phi'(f(\xi)) \right] \rho_\infty^2(f) \right).
\end{aligned}$$

We now use the constant M involved in Assumption $(\mathbf{H}_\phi)$:

$$\frac{|\nabla f(\xi)|^2}{f^2(\xi)} \leqslant \frac{M}{f(\xi)} \qquad \text{and} \qquad \phi'(f(\xi))^2 |\nabla f(\xi)|^2 \leqslant M \phi'(f(\xi)).$$

Using the definition of M, we deduce that:

$$\begin{aligned}
&\frac{\langle y, D^2V_p(\xi)y\rangle}{\|y\|^2}\\
&\lesssim_{id} \quad MV_p(\xi)\left[\frac{1}{f(\xi)} + \phi'(f(\xi)) + \phi''(f(\xi))f(\xi)\right] + r\rho_\infty(f)V_p(\xi)\left[\frac{1}{f(\xi)} + \phi'(f(\xi))\right]\\
&\lesssim_{id} \quad (M+\rho_\infty(f))V_{p-1}(\xi) + MV_p(\xi)\left(\phi'(f(\xi)) + \phi''(f(\xi))f(\xi)\right).
\end{aligned}$$

$\mathbf{H}_\phi$ implies that ϕ'' is negative for $u \geqslant x_0$ so that ϕ' is bounded (it is a non-negative function and non-increasing on $[x_0, +\infty)$). Now, $\mathbf{H}_\phi$(ii) yields

$$\begin{aligned}
&\sup_{\xi\in\mathbb{R}^d}\left(\phi'(f(\xi)) + \phi''(f(\xi))f(\xi)\right)(1+|\nabla f(\xi)|^2)\\
&\qquad \leqslant \quad (\|\phi'\|_\infty + M + \sup_{x\in[0,x_0]}\phi''(x)(x_0(1+Mx_0))\\
&\qquad \lesssim_{id} \quad 1+M.
\end{aligned}$$

We then deduce that

$$\forall y \in \mathbb{R}^d \qquad \frac{\langle y, D^2V_p(\xi)y\rangle}{\|y\|^2}\lesssim_{id}(1+M^2+\rho_\infty(f))\left(V_{p-1}(\xi) + \frac{V_p(\xi)}{1+\|\nabla f(\xi)\|^2}\right).$$

The second assertion follows. □

Lemma 10. *Suppose that* $\mathbf{H}_\phi$ *holds and consider* $r \in [0,1]$. *For any* $\delta > 0$, $\varepsilon > 0$ *define* $\xi_{\delta,\varepsilon,x,\ell} = x + \ell\delta\left(-\nabla f(x) + \varepsilon\right)$ *with* $\ell \in [0,1]$. *Then,*

i) Assume that $\delta > 0$ *is such that* $\rho_\infty(f)\delta \leqslant 1/2$, *then:*

$$f(\xi_{\delta,\varepsilon,x,\ell}) \leqslant f(x) + \delta|\varepsilon|^2.$$

ii) Assume that $\rho_\infty(f)\delta \leqslant 1/2$, *then* $q_p(\phi)$ *exists such that for all* $\forall x \in \mathbb{R}^d$:

$$\begin{aligned}
D^2V_p(\xi_{\delta,\varepsilon,x,\ell})(-\nabla f(x)+\varepsilon)^{\otimes 2} \leqslant\ & q_p(\phi)\bar{\Sigma}_p^{-1}(1+\|\varepsilon\|^{2(p+1)})e^{\phi(\delta\|\varepsilon\|^2)}\\
&\times (V_{p-1}(x) + V_p(x) + \delta^{p-1}),
\end{aligned}$$

with $q_p(\phi) = \mathcal{O}_{id}(1+M^3+\rho_\infty(f))\bar{\Sigma}_p$.

iii) When $\phi = 0$, *if* $\rho_\infty(f)\delta \leqslant 1/2$, *then a* $q_p(\phi)$ *exists such that* $\forall x \in \mathbb{R}^d$,

$$\begin{aligned}
D^2f^p(\xi_{\delta,\varepsilon,x,\ell})(-\nabla f(x)+\varepsilon)^{\otimes 2} \leqslant\ & q_p(\phi)(1+\Sigma_p)^{-1}\\
&\times u\left(f^p(x) + f^{p-1}(x)|\varepsilon|^2 + \delta^{p-1}|\varepsilon|^{2p}\right).
\end{aligned}$$

Furthermore, $q_p(\phi) = \mathcal{O}_{id}\left((1+M^2+\rho_\infty(f))(1+\Sigma_p)\right)$.

Proof:
Proof of *i)* Using the Taylor formula, $\tilde{\xi} \in [x, \xi_{\delta,\varepsilon,x,\ell}]$ exists such that:

$$f(\xi_{\delta,\varepsilon,x,\ell}) = f(x) - \ell\delta\|\nabla f(x)\|^2 + \ell\delta\langle\nabla f(x), \varepsilon\rangle + \frac{\ell^2\delta^2}{2}D^2f(\tilde{\xi})\left(-\nabla f(x)+\varepsilon\right)^{\otimes 2}.$$

From $\|a+b\|^2 \leqslant 2(\|a\|^2 + \|b\|^2)$ and the definition of $\rho_\infty(f)$, we get:

$$D^2f(\tilde{\xi})\left(-\nabla f(x)+\varepsilon\right)^{\otimes 2} \leqslant 2\rho_\infty(f)\left(\|\nabla f(x)\|^2 + \|\varepsilon\|^2\right).$$

The elementary inequality $|\langle u,v\rangle| \leqslant \frac{1}{2}(\|u\|^2+\|v\|^2)$ yields:

$$\begin{aligned} f(\xi_{\delta,\varepsilon,x,\ell}) &\leqslant f(x) - \ell\delta\|\nabla f(x)\|^2 + \ell\delta\langle \nabla f(x),\varepsilon\rangle + \ell^2\delta^2\rho_\infty(f)\left(\|\nabla f(x)\|^2+\|\varepsilon\|^2\right) \\ &\leqslant f(x) + \ell\delta\left[-\frac{1}{2}+\ell\delta\rho_\infty(f)\right]\|\nabla f(x)\|^2 + \left[\frac{\ell\delta}{2}+\rho_\infty(f)\ell^2\delta^2\right]\|\varepsilon\|^2 \qquad (34) \\ &\leqslant f(x)+\ell\delta\|\varepsilon\|^2 \leqslant f(x)+\delta\|\varepsilon\|^2, \end{aligned}$$

where in the last line we use that $\ell \leqslant 1$ and the condition $\delta\rho_\infty(f) \leqslant 1/2$.

Proof of *ii)* We divide the proof into 3 steps.

- Step 1: Comparison between $V_r(\xi_{\delta,\varepsilon,x,\ell})$ and $V_r(x)$. We consider $r \geqslant 0$ and write $\xi = \xi_{\delta,\varepsilon,x,\ell}$ for the sake of convenience. Since ϕ is non-decreasing, one first deduces from (i) that:

$$V_r(\xi) \leqslant (f(x)+\delta\|\varepsilon\|^2)^r \exp\left(\phi(f(x)+\delta\|\varepsilon\|^2)\right).$$

Lemma 7 and $(|a|+|b|)^r \leqslant 2^{r-1}(|a|^r+|b|^r)$ yields:

$$V_r(\xi) \leqslant 2^{r-1}\left(f^r(x)+\delta^r\|\varepsilon\|^{2r}\right)e^{\phi(f(x))+\phi(\delta\|\varepsilon\|^2)+c_\phi}.$$

Setting $T_{\varepsilon,\gamma,r} = (1+\|\varepsilon\|^{2r})\exp(\phi(\delta\|\varepsilon\|^2)$, and using that $V_0 = e^{\phi(f)}$:

$$\begin{aligned} \forall r \geqslant 0 \quad \exists\, C_r > 0 \qquad V_r(\xi) &\lesssim_{id} \exp(\phi(\delta\|\varepsilon\|^2))\left[V_r(x)+\delta^r\|\varepsilon\|^{2r}V_0(x)\right] \\ &\lesssim_{id} \exp(\phi(\delta\|\varepsilon\|^2))\left[(1+\|\varepsilon\|^{2r})V_r(x)+\delta^r\|\varepsilon\|^{2r}\right] \\ &\lesssim_{id} T_{\varepsilon,\gamma,r}\left[V_r(x)+\delta^r\right]. \qquad (35) \end{aligned}$$

where in the second line, we used that $V_0 \leqslant e^{\phi(1)}+V_r$.

- Step 2: Upper bound of $\rho(D^2V_p(\xi)).|\nabla f(x)|^2$. We apply Lemma 9 *ii)*. Setting $q_1 = 1+M^2+\rho_\infty(f)$, we get:

$$\begin{aligned} \rho(D^2V_p(\xi)).|\nabla f(x)|^2 &\lesssim_{id} q_1\left(V_{p-1}(\xi)+\frac{V_p(\xi)}{1+\|\nabla f(\xi)\|^2}\right)|\nabla f(x)|^2 \\ &\lesssim_{id} q_1\left(T_{\epsilon,\delta,p-1}[V_{p-1}(x)+\delta^{p-1}]+\frac{T_{\varepsilon,\delta,p}[V_p(x)+\delta^p]}{1+\|\nabla f(\xi)\|^2}\right)|\nabla f(x)|^2 \\ &\lesssim_{id} q_1\Big(T_{\epsilon,\delta,p-1}[V_{p-1}(x)|\nabla f(x)|^2+\delta^{p-1}|\nabla f(x)|^2] \\ &\quad + T_{\varepsilon,\delta,p}\frac{|\nabla f(x)|^2}{1+\|\nabla f(\xi)\|^2}[V_p(x)+\delta^p]\Big). \end{aligned}$$

Under Assumption $(\mathbf{H}_\phi)$, $V_{p-1}(x)|\nabla f(x)|^2 \leqslant MV_p(x)$ and $\delta^{p-1}|\nabla f(x)|^2 \leqslant M\delta^{p-1}f(x) \leqslant M\delta^{p-1}(1+V_p(x))$. Inequality $T_{\epsilon,\delta,p-1} \leqslant 2T_{\epsilon,\delta,p}$ leads to:

$$\begin{aligned} &\|D^2V_p(\xi)\|.|\nabla f(x)|^2 \\ &\quad\lesssim_{id} \quad q_1MT_{\epsilon,\delta,p-1}[V_p(x)+\delta^{p-1}]+q_1\frac{T_{\epsilon,\delta,p}[V_p(x)+\delta^p]|\nabla f(x)|^2}{1+\|\nabla f(\xi)\|^2} \\ &\quad\lesssim_{id} \quad (1+M^3+\rho_\infty(f))T_{\epsilon,\delta,p}\left[[V_p(x)+\delta^{p-1}]+\frac{[V_p(x)+\delta^p]|\nabla f(x)|^2}{1+\|\nabla f(\xi)\|^2}\right]. \qquad (36) \end{aligned}$$

To handle Equation (36), we are driven to derive an upper bound of $\frac{|\nabla f(x)|^2}{1+|\nabla f(\xi)|^2}$. According to the Taylor formula, a ξ' exists in $[x,\xi]$ such that:

$$\nabla f(x) = \nabla f(\xi) - \ell\delta D^2 f(\xi')\left(-\nabla f(x)+\varepsilon\right),$$

and the triangle inequality associated with $\ell \in [0,1]$ yields:

$$|\nabla f(x)| \leqslant |\nabla f(\xi)| + \rho_\infty(f)\delta(|\nabla f(x)| + \|\varepsilon\|).$$

Gathering all the terms with $|\nabla f(x)|$ on the left hand side and using $\rho_\infty(f)\delta \leqslant 1/2$, we obtain that:

$$|\nabla f(x)| \leqslant (1-\rho_\infty(f)\delta)^{-1}\left(|\nabla f(\xi)| + \|\varepsilon\|\right) \leqslant 2\left(|\nabla f(\xi)| + \|\varepsilon\|\right).$$

The elementary inequality $(u+v)^2 \leqslant 2(u^2+v^2)$ leads to $|\nabla f(x)|^2 \leqslant 8(\|\nabla f(\xi)\|^2 + \|\varepsilon\|^2)$. As a consequence,

$$\frac{|\nabla f(x)|^2}{1+\|\nabla f(\xi)\|^2} \leqslant 8(1+\|\varepsilon\|^2).$$

We use this last inequality into (36) and obtain that:

$$\rho\left(V_p(\xi)\right).|\nabla f(x)|^2 \lesssim_{id} (1+M^3+\rho_\infty(f))T_{\epsilon,\delta,p}\left(\delta^{p-1} + [V_p(x)+\delta^p](1+\|\varepsilon\|^2)\right).$$

Finally, since $T_{\varepsilon,\delta,p}(1+\|\varepsilon\|^2) \leqslant 3T_{\varepsilon,\delta,p+1}$, we conclude that:

$$\rho\left(V_p(\xi)\right).|\nabla f(x)|^2 \lesssim_{id} (1+M^3+\rho_\infty(f))T_{\epsilon,\delta,p+1}\left(\delta^{p-1} + V_p(x)\right), \tag{37}$$

- Step 3: Upper bound of $\|D^2V_p(\xi_{\delta,\varepsilon,x,\ell})\|_s.\|\varepsilon\|^2$. (35) and Lemma 9 *ii)* yield:

$$\rho\left(D^2V_p(\xi_{\delta,\varepsilon,x,\ell})\right).\|\varepsilon\|^2 \lesssim_{id} q_1 T_{\varepsilon,\delta,p+1}\left(V_{p-1}(x) + V_p(x) + \delta^{p-1}\right). \tag{38}$$

The result then follows from the combination of Equations (37) and (38).

Proof of *iii)* Finally, let us consider the particular case $\phi = 0$ that corresponds to the situation where $SC(\alpha)$. Going back to Lemma 8 *ii)* and noting that $(\mathbf{H}_\phi)$ in this case reads $mf(x) \leqslant |\nabla f(x)|^2 \leqslant Mf(x)$, we deduce that:

$$\rho(D^2 f^p(\xi)) \leqslant (M+\rho_\infty(f))f^{p-1}(\xi). \tag{39}$$

Using Lemma 10 *i)*, we have, if $\delta\rho_\infty(f) \leqslant 1/2$:

$$\rho(D^2f^p(\xi)) \lesssim_{id} (M+\rho_\infty(f))\left(f^{p-1}(x) + \delta^{p-1}|\varepsilon|^{2(p-1)}\right),$$

so that:

$$\begin{aligned}\rho(D^2f^p(\xi))\left(|\nabla f(x)|^2 + |\varepsilon|^2\right) \lesssim_{id} (M+\rho_\infty(f))\Big(&Mf^p(x) + f^{p-1}(x)|\varepsilon|^2 \\ &+\delta^{p-1}(Mf(x)|\varepsilon|^{2(p-1)} + |\varepsilon|^{2p})\Big).\end{aligned}$$

The inequality follows when $p=1$. When $p>1$, we deduce from the Young inequality that $f(x)|\varepsilon|^{2(p-1)} \lesssim_{id} f^p(x) + |\varepsilon|^{2p}$ and deduce the result. □

4.2 Main result

We have the following result on the convergence rate of the SGD $(\theta_n)_{n\geqslant 1}$.

Theorem 11. *Let $p \geqslant 1$, if $(\mathbf{H}_\phi)$ and $(\mathbf{H}^{\phi}_{\mathbf{\Sigma_p}})$ when $\phi \neq 0$ (or $(\mathbf{H}_\phi)$ and $(\mathbf{H}^{\mathbf{SC}}_{\mathbf{\Sigma_p}})$ if $\phi = 0$) hold. Consider $q_p(\phi)$ from Lemma 10 and:*

$$n_0 := \inf\left\{n \,:\, \gamma_{n+1}q_p(\phi) \leqslant \frac{m}{4} \text{ and } \forall q \leqslant p : \left(\frac{\gamma_n^q - \gamma_{n+1}^q}{m\gamma_{n+1}^{q+1}}\right) \leqslant \frac{1}{8}\right\}.$$

i) For all $n \geqslant n_0$*,*

$$\mathbb{E}[V_p(\theta_{n+1})] \leqslant \left(1 - \tfrac{3m}{4}\gamma_{n+1}\right)\mathbb{E}[V_p(\theta_n)] + \mathcal{O}_{id}\left(q_p(\phi)\left(\mathbb{E}[V_{p-1}(\theta_n)]\gamma_{n+1}^2 + \gamma_{n+1}^{p+1}\right)\right). \tag{40}$$

ii) $(\bar{C}_p)_{p\geqslant 1}$ *exists such that for all* $n \geqslant n_0$, $\mathbb{E}\left[V_p(\theta_n)\right] \leqslant \bar{C}_p\{\gamma_n\}^p$ *with*

$$\bar{C}_1 = \mathbb{E}[V_1(\theta_{n_0})]\frac{q_1(\phi)}{m} + \mathcal{O}_{id}\left(\frac{q_1(\phi)}{m}\right)$$

and for every integer $p \geqslant 2$*,*

$$\bar{C}_p = \mathbb{E}[V_p(\theta_{n_0})]\left(\frac{q_p(\phi)}{m}\right)^p + \mathcal{O}_{id}\left((\bar{C}_{p-1}+1)\frac{q_p(\phi)}{m}\right). \tag{41}$$

Proof of Theorem 11 *i)*: We apply the Taylor formula to V_p and obtain that:

$$\begin{aligned} V_p(\theta_{n+1}) &= V_p(\theta_n) - \gamma_{n+1}\langle \nabla V_p(\theta_n), \nabla f(\theta_n)\rangle + \gamma_{n+1}\langle V_p(\theta_n), \Delta\mathcal{M}_{n+1}\rangle \\ &\quad + \frac{\gamma_{n+1}^2}{2} D^2 V_p(\xi_{n+1})(-\nabla f(\theta_n) + \Delta\mathcal{M}_{n+1})^{\otimes 2}, \end{aligned}$$

where $\xi_{n+1} = \theta_n + \ell_n \Delta\theta_{n+1}$, where $\ell_n \in [0,1]$. Using Lemma 9 *i)*, we get

$$\forall n \in \mathbb{N}^\star \qquad V_p(\theta_n) - \gamma_{n+1}\langle \nabla V_p(\theta_n), \nabla f(\theta_n)\rangle \leqslant V_p(\theta_n)(1 - m\gamma_{n+1}). \tag{42}$$

Now, we need to consider separately the cases $\phi = 0$ and $\phi \neq 0$.
• Case $\phi \neq 0$: Since $\gamma_n \rho_\infty(f) \leqslant \gamma_n q_p(\phi) \leqslant 1/2$ for all $n \geqslant n_0$, $(\mathbf{H}^{\phi}_{\bar{\boldsymbol{\Sigma}}_{\mathbf{p}}})$ yields

$$\mathbb{E}[(1 + |\Delta\mathcal{M}_{n+1}|^{2(p+1)})\exp(\phi(\gamma|\Delta\mathcal{M}_{n+1}|^2))|\mathcal{F}_n] \leqslant \bar{\Sigma}_p.$$

Thus, we deduce from Lemma 10 *ii)* that for every $n \geqslant n_1$

$$\begin{aligned} &\mathbb{E}\left[V_p(\theta_{n+1})\,|\,\mathcal{F}_n\right] \\ &\quad \leqslant (1 - m\gamma_{n+1})V_p(\theta_n) + q_p(\phi)\left(\gamma_{n+1}^2(V_p(\theta_n) + V_{p-1}(\theta_n)\{\gamma_{n+1}\}^{p+1}\right). \end{aligned}$$

This yields

$$\begin{aligned} &\mathbb{E}\left[V_p(\theta_{n+1})\,|\,\mathcal{F}_n\right] \\ &\quad \leqslant \left(1 - m\gamma_{n+1} + q_p(\phi)\gamma_{n+1}^2\right)V_p(\theta_n) + q_p(\phi)(\gamma_{n+1}^2 V_{p-1}(\theta_n) + \{\gamma_{n+1}\}^{p+1}). \end{aligned} \tag{43}$$

The result follows since $n \geqslant n_0$ so that $1 - m\gamma_{n+1} + q_p(\phi)\gamma_{n+1}^2 \leqslant 1 - \frac{3m}{4}\gamma_{n+1}$.
• Case $\phi = 0$: By Lemma 10 *iii)* and Assumption $(\mathbf{H}^{\mathbf{SC}}_{\boldsymbol{\Sigma}_{\mathbf{p}}})$, we have for all $n \geqslant n_0$,

$$\begin{aligned} &\mathbb{E}\left[D^2 f^p(\xi_{n+1})(-\nabla f(\theta_n) + \Delta\mathcal{M}_{n+1})^{\otimes 2}|\mathcal{F}_n\right] \\ &\leqslant \quad q_p(\phi)(1+\Sigma_p)^{-1}\left(f^p(\theta_n) + f^{p-1}(\theta_n)\Sigma_1(1 + f(\theta_n)) + \gamma_{n+1}^{p-1}\Sigma_p(1 + f^p(\theta_n))\right) \\ &\leqslant \quad q_p(\phi)\frac{1 + \Sigma_1 + \Sigma_p}{1 + \Sigma_p}(f^p(\theta_n) + f^{p-1}(\theta_n)) + \gamma_{n+1}^{p+1}) \end{aligned}$$

Since $|x|^2 \leqslant 1 + |x|^{2p}$ for any $p \geqslant 1$, $\Sigma_1 \leqslant 1 + \Sigma_p$. Thus, at the price of replacing $q_p(\phi)$ by $2q_p(\phi)$ ($q_p(\phi)$ is defined up to a universal constant), we get

$$\begin{aligned} &\gamma_{n+1}^2\mathbb{E}\left[D^2 f^p(\xi_{n+1})(-\nabla f(\theta_n) + \Delta\mathcal{M}_{n+1})^{\otimes 2}|\mathcal{F}_n\right] \\ &\quad \leqslant q_p(\phi)\left(\gamma_{n+1}^2(f^p(\theta_n) + f^{p-1}(\theta_n)) + \gamma_{n+1}^{p+1}\right). \end{aligned}$$

Now, the initial Taylor formula with the previous inequality ends the proof.

Proof of Theorem 11 *ii)*: This result is obtained by an induction on p. We preliminary consider the situation where p is an integer greater than 1. Then, a general result is deduced for any $p \geqslant 1$ using the Jensen inequality:

$$\mathbb{E}[|X|^p] \leqslant \left(\mathbb{E}[|X|^{p'}]\right)^{p/p'},$$

where p' is an integer larger than p since $t \mapsto |t|^{p/p'}$ is a concave function.

- We first consider the case where $p = 1$, we use the elementary inequality $V_0 \leqslant 1 + V_1$ and obtain that

$$\forall n \geqslant n_0, \quad \mathbb{E}[V_1(\theta_{n+1})] \leqslant (1 - \frac{3m}{4}\gamma_{n+1} + q_1(\phi)\gamma_{n+1}^2)\mathbb{E}[V_1(\theta_n)] + 2q_1(\phi)\gamma_{n+1}^2.$$

According to our choice on n, we deduce that

$$\forall n \geqslant n_0, \quad \mathbb{E}[V_1(\theta_{n+1})] \leqslant (1 - \frac{m}{2}\gamma_{n+1})\mathbb{E}[V_1(\theta_n)] + 2q_1(\phi)\gamma_{n+1}^2.$$

Set $v_n = \gamma_n^{-1}\mathbb{E}[V_1(\theta_n)]$. We obtain

$$\forall n \geqslant n_0, \quad v_{n+1} \leqslant (1 - \frac{m}{2}\gamma_{n+1})v_n\frac{\gamma_{n+1}^{-1}}{\gamma_n^{-1}} + q_1(\phi)\gamma_{n+1}.$$

According to the construction of n_0, we can check that for all $n \geqslant n_0$,

$$\left(\frac{\gamma_n}{\gamma_{n+1}}\right)^p \leqslant 1 + \frac{m}{4}\gamma_{n+1}.$$

We then obtain

$$\forall n \geqslant n_0, \quad v_{n+1} \leqslant \left(1 - \frac{m}{2}\gamma_{n+1}\right) v_n \left(1 + \frac{m}{4}\gamma_{n+1}\right) + q_1(\phi)\gamma_{n+1}. \tag{44}$$

Since $(1 - \frac{m}{2}\gamma_{n+1})(1 + \frac{m}{4}\gamma_{n+1}) \leqslant 1 - \frac{m}{4}\gamma_{n+1}$, we deduce that for all $n \geqslant n_0$,

$$v_n \leqslant v_{n_0}\prod_{k=n_0+1}^{n}(1 - \frac{m}{4}\gamma_k) + q_1(\phi)\textstyle\sum_{k=n_0+1}^{n}\gamma_k\displaystyle\prod_{\ell=k}^{n-1}(1 - \frac{m}{4}\gamma_\ell)$$

with the convention $\prod_{\varnothing} = 1$. By the elementary inequality $\log(1+x) \leqslant x$ for $x > -1$, this yields

$$v_n \leqslant v_{n_0}e^{-\frac{m}{4}(\Gamma_n - \Gamma_k)} + q_1(\phi)\textstyle\sum_{k=n_0+1}^{n}\gamma_k e^{-\frac{m}{4}(\Gamma_{n-1} - \Gamma_{k-1})}.$$

On the one hand, $e^{-\frac{m}{4}(\Gamma_n - \Gamma_k)} \leqslant 1$, on the other hand, a series/integral comparison yields

$$\begin{aligned}\textstyle\sum_{k=n_0+1}^{n}\gamma_k e^{-\frac{m}{4}(\Gamma_{n-1} - \Gamma_{k-1})} &\leqslant e^{-\frac{m}{4}\Gamma_{n-1}}\int_0^{\Gamma_n} e^{\frac{m}{4}x}\mathrm{d}x \\ &\leqslant \frac{4}{m}e^{-\frac{m}{4}\Gamma_{n-1}}\left(e^{-\frac{m}{4}\Gamma_n} - 1\right) \leqslant \frac{4e^{\frac{m}{4}\gamma_{n_0}}}{m} \leqslant \frac{5}{m}.\end{aligned}$$

where we used in the last equality that $m\gamma_{n_0} \leqslant \gamma_{n_0}q_1(\phi) \leqslant 1/2$.

- Let us now assume that p is an integer greater than 2 and assume that $\bar{C}_{p-1}$ is finite. Then, for all $n \geqslant n_0$, $\mathbb{E}[V^{p-1}(\theta_n)] \leqslant \bar{C}_{p-1}\gamma_n^{p-1}$ and hence, by (40), we deduce that

$$\forall n \geqslant n_0, \quad \mathbb{E}[V_p(\theta_{n+1})] \leqslant \left(1 - \frac{3m}{4}\gamma_{n+1}\right)\mathbb{E}[V_p(\theta_n)] + q_p(\phi)\left(\bar{C}_{p-1}\gamma_{n+1}^p + \gamma_{n+1}^{p+1}\right).$$

As a consequence, by setting $v_n = \gamma_n^{-p}\mathbb{E}[V_p(\theta_n)]$ and dividing the above inequality by γ_{n+1}^p, we obtain

$$\forall n \geqslant n_0, \quad v_{n+1} \leqslant (1 - \frac{3m}{4}\gamma_{n+1})v_n \frac{\gamma_{n+1}^{-p}}{\gamma_n^{-p}} + q_p(\phi)(1+\bar{C}_{p-1})\gamma_{n+1}.$$

As in the case $p = 1$, the definition of n_0 implies that for all $n \geqslant n_0$,

$$\forall n \geqslant n_0, \quad v_{n+1} \leqslant (1 - \frac{m}{4}\gamma_{n+1})v_n + q_p(\phi)(1+\bar{C}_{p-1})\gamma_{n+1}.$$

The end of the proof is identical to the case $p = 1$. □

Acknowledgements The authors gratefully acknowledge Jérôme Bolte and Gersende Fort for stimulating discussions on the Kurdyka-Łojasiewicz inequality and averaged stochastic optimization algorithms. We also warmly thank Francis Bach for several constructive comments on an earlier version of our work and pointing out some useful recent references.

References

[1] A. Anastasiou and K. Balasubramanian and M.A. Erdogdu, Normal Approximation for Stochastic Gradient Descent via Non-Asymptotic Rates of Martingale CLT, Proceedings of the Thirty-Second Conference on Learning Theory, 99, 115-137 (2019)

[2] A. Agarwal and P. L. Bartlett and P. Ravikumar and M. J. Wainwright, Information-Theoretic Lower Bounds on the Oracle Complexity of Stochastic Convex Optimization, IEEE Transactions on Information Theory, 58(5), 3235-3249 (2012)

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

[4] N. Flammarion and F. Bach, From Averaging to Acceleration, There is Only a Step-size, Proceedings of the International Conference on Learning Theory (COLT) (2015)

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

[6] F. Bach and E. Moulines, Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning, Advances in Neural Information Processing Systems (2011)

[7] D.P. Bertsekas, Nonlinear programming, Athena Scientific Optimization and Computation Series, Belmont, MA, xiv+777 (1999)

[8] J. Bolte and A. Daniilidis and O. Ley and L. Mazet, Characterizations of Lojasiewicz inequalities: subgradient flows, talweg, convexity, Transactions of the American Mathematical Society, 362(6), 3319-3363 (2010)

[9] J. Bolte and P. Nguyen and J. Peypouquet and B. W. Suter, From error bounds to the complexity of first-order descent methods for convex functions, Math. Program. (A), 165(2), 471-507 (2017)

[10] A. Cabot and H. Engler and S. Gadat, On the long time behavior of second order differential equations with asymptotically small dissipation, Trans. Amer. Math. Soc., 361(11), 5983-6017 (2009)

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

[12] N. Cesa-Bianchi and G. Lugosi, Prediction, learning, and games, Cambridge University Press, Cambridge, xii+394 (2006).

[13] M. Duflo, Random Iterative Models, Adaptive algorithms and stochastic approximations, Springer-Verlag, New-York, Applications of Mathematics, (1997)

[14] S. Gadat and L. Miclo, Spectral decompositions and L2-operator norms of toy hypocoercive semi-groups, Kinetic and Related Models, 6(2), 317-372 (2013)

[15] S. Gadat and F. Panloup and S. Saadane, Stochastic Heavy Ball, Electronic Journal of Statistics, 12(1), 461-529 (2018)

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

[17] A. Godichon-Baggioni, Estimating the geometric median in Hilbert spaces with stochastic gradient algorithms: L^p and almost sure rates of convergence, J. Multivar. Anal. 146 209?222 (2016)

[18] A. Godichon-Baggioni, L^p and almost sure rates of convergence of averaged stochastic gradient algorithms: locally strongly convex objective. ESAIM: Probability and Statistics, 23:841?873 (2019)

[19] P. Huber, Robust Estimation of a Location Parameter. The Annals of Statistics. 53 (1): 73-101 (1964)

[20] G. Fort, Central limit theorems for stochastic approximation with controlled Markov chain dynamics, ESAIM. Probability and Statistics, 19, 60-80 (2015)

[21] K. Kurdyka, On gradients of functions definable in o-minimal structures, Ann. Inst. Fourier (Grenoble), 48(3), 769-783 (1988)

[22] S. Lojasiewicz, Une propriété topologique des sous-ensembles analytiques réels, Editions du CNRS, Paris, Les Équations aux Dérivées Partielles, 87-89 (1963)

[23] A. Nemirovski and D. Yudin, Problem complexity and method efficiency in optimization, Wiley-Interscience Series in Discrete Mathematics (1983)

[24] Y. Nesterov, Introductory Lectures on Convex Optimization. A basic course, Kluwer Academic Publishers, Series: Applied Optimization, Boston, MA (2004)

[25] M. Pelletier, Asymptotic almost sure efficiency of averaged stochastic algorithms. SIAM J. Control Optim. 39 49?72 (2000)

[26] B. T. Polyak and A. Juditsky, Acceleration of Stochastic Approximation by Averaging, SIAM Journal on Control and Optimization, 30(4), 838-855 (1992)

[27] H. Robbins and S. Monro, A Stochastic Approximation Method, Annals of Mathematical Statistics, 22, 400-407 (1951)

[28] D. Ruppert, Technical Report, 781, Cornell University Operations Research and Industrial Engineering (1988)

[29] C. Villani, Hypocoercivity, Mem. Amer. Math. Soc., 202(950), (2009)

Appendix A: Technical lemmas for Theorem 2

In the next lemma, we study some properties of $(\epsilon_{\mu,n})_{n\geqslant 1}$ involved in the change of basis related to the evolution of $(\hat{\theta}_n)_{n\geqslant 1}$ (see Proposition 6). Roughly speaking, we quantify the effect and variability of this change of basis.

Without loss of generality, we assume in the following proofs that $\underline{\mu} \leqslant 1$, $\gamma = 1$ et $c_2 \geqslant 1$, $\|D^2 f\|_{\text{Lip}} \geqslant 1$.

Lemma 12. *Assume that $\gamma_n = \gamma n^{-\beta}$ with $\beta \in (0,1)$. Let $\underline{\mu} > 0$. For any $\mu \geqslant \underline{\mu}$,*

$$\forall n \geqslant n_0 := \left\lceil \left(\frac{2}{\underline{\mu}}\right)^{1/(1-\beta)} \right\rceil, \qquad |\epsilon_{\mu,n} - \epsilon_{\mu,n+1}| \lesssim_{id} \frac{1}{\underline{\mu}} n^{\beta-2} \tag{45}$$

and $|\epsilon_{\mu,n}| \lesssim_{id} (\underline{\mu}\gamma_n n)^{-1}$.

<u>Proof:</u> First, remark that for $n \geqslant n_0$, $\mu\gamma_n n - 1 \geqslant \frac{1}{2}\mu\gamma_n n$, so that $\epsilon_{\mu,n}$ is well-defined for any $n \geqslant n_0$ and,

$$|\epsilon_{\mu,n}| \leqslant \frac{2}{\underline{\mu}\gamma_n n} + \frac{2\mu\gamma_n}{\mu\gamma_n n} \leqslant \frac{2}{n}\left(\frac{1}{\underline{\mu}\gamma_n} + 1\right) \lesssim_{id} \frac{1}{\underline{\mu}\gamma_n n},$$

since $\underline{\mu}\gamma_n \leqslant 1$ for every $n \geqslant 1$. As concerns (45), we observe that for $n \geqslant n_0$,

$$\begin{aligned}
|\epsilon_{\mu,n} - \epsilon_{\mu,n+1}| &= \left| \frac{1-\mu\gamma_n}{1-\mu\gamma_n n} - \frac{1-\mu\gamma_{n+1}}{1-\mu\gamma_{n+1}(n+1)} \right| \\
&= \left| \frac{(1-\mu\gamma_n)(1-\mu\gamma_{n+1}(n+1)) - (1-\mu\gamma_{n+1})(1-\mu\gamma_n n)}{(1-\mu\gamma_n n)(1-\mu\gamma_{n+1}(n+1))} \right| \\
&\leqslant \mu \frac{(\gamma_n - \gamma_{n+1}) + |(n+1)\gamma_{n+1} - n\gamma_n| + \mu\gamma_n\gamma_{n+1}}{(\mu\gamma_n n - 1)(\mu\gamma_{n+1}(n+1) - 1)} \\
&\leqslant \mu(2+\mu)\frac{n^{-\beta}}{n\mu\gamma_n(n+1)\mu\gamma_{n+1}} \lesssim_{id} \left(1 + \frac{1}{\underline{\mu}}\right) n^{2-\beta},
\end{aligned}$$

which yields the result since $\underline{\mu} \leqslant 1$.

□

Lemma 13. *Set $n_1 := n_0 \vee \lceil c_2^{1/\beta} \rceil \vee \lceil \|D^2 f\|_{\mathrm{Lip}} \rceil$. Under the assumptions of Proposition 7, we have:*
i) For any $i \in \{1,\ldots,d\}$, $\omega_n(i) = \mathbb{E}[(\widetilde{Z}_n^{(1)})_i(\widetilde{Z}_n^{(2)})_i]$ satisfies $\forall n \geqslant n_1$,

$$\left| \omega_{n+1}(i) - (1-\gamma_{n+1}\mu_i^\star)\left(1 - \frac{1}{n+1}\right)\omega_n(i) \right| \lesssim_{id} C_\omega \frac{\gamma_n}{n} + \gamma_{n+1}\|D^2 f\|_{\mathrm{Lip}}\mathbb{E}[|\theta_n|^2|\widetilde{Z}_n^{(2)}|].$$

where $C_\omega = \frac{d(\|S^*\| + \|S\|_{\mathrm{Lip}}) + c_4\|D^2 f\|_{\mathrm{Lip}})}{\underline{\mu}}$.
ii) Set

$$|\widehat{\Delta Z_{n+1}^2}| := \left| \mathbb{E}[|\widetilde{Z}_{n+1}^{(2)}|^2] - \left(1 - \frac{1}{n+1}\right)^2 \mathbb{E}[|\widetilde{Z}_n^{(2)}|^2] - \textstyle\sum_{i=1}^d \alpha_n^i \omega_n(i) - \dfrac{\mathrm{Tr}(\Sigma^\star)}{(n+1)^2} \right|.$$

We have:

$$|\widehat{\Delta Z_{n+1}^2}| \lesssim_{id} n^{-\frac{1}{2}-\beta}\mathbb{E}[|\widetilde{Z}_n^{(2)}|^2 + \frac{c_4 d}{\underline{\mu}^3} C_S \mathcal{O}_{id}\left(n^{-\frac{3}{2}-\beta} \vee n^{-3+\beta}\right)$$

where $C_S = \|S^*\| + \|S\|_{\mathrm{Lip}} + 1$ and α_n^i, defined by (29), satisfies $|\alpha_n^i| \lesssim_{id} (\underline{\mu})^{-1} n^{\beta-2}$, $i = 1,\ldots,d$.

<u>Proof:</u> First, remark that under the definition of n_1, we have for all $n \geqslant n_1$, $\mu\gamma_n n - 1 \geqslant \frac{1}{2}\mu\gamma_n n$, $c_2\gamma_n \leqslant 1$ and $\|D^2 f\|_{\mathrm{Lip}} n^{-1} \leqslant 1$. Then, for all $n \geqslant n_0$, Υ_n and Ω_n are well-defined deterministic matrices and by Lemma 12, we can verify that

$$\gamma_{n+1}\|\Upsilon_n\| \lesssim_{id} \frac{1}{n\underline{\mu}} \text{ and } \gamma_{n+1}\|\Omega_n\| \leqslant \gamma_{n+1}\|\mathcal{E}_{n,D^\star} - \mathcal{E}_{n+1,D^\star}\|\|I_d - \gamma_{n+1}D^\star\| \leqslant \frac{1}{\underline{\mu}n^2}. \tag{46}$$

$i)$ Now, let us prove the first statement and let $n \geqslant n_1$. Using (27), we have

$$\begin{aligned}\left|\omega_{n+1}(i) - (1-\gamma_{n+1}\mu_i^\star)\left(1-\frac{1}{n+1}\right)\omega_n(i)\right| &\leqslant \gamma_{n+1}\|D^2 f\|_{\mathrm{Lip}}\mathbb{E}[|\theta_n|^2|\widetilde{Z}_n^{(2)}|]. \\ &+ \gamma_{n+1}^2\mathbb{E}[\{Q\Delta\mathcal{M}_{n+1}\}_i\{\Upsilon_n Q\Delta\mathcal{M}_{n+1}\}_i] + \gamma_{n+1}r_n^{(1)},\end{aligned}$$

where,

$$\begin{aligned}|r_n^{(1)}| \lesssim_{id} \|\Omega_n\| &\left(\frac{\mathbb{E}|\widetilde{Z}_n^{(1)}|^2}{\gamma_{n+1}} + \mathbb{E}[\|D^2 f\|_{\mathrm{Lip}}|\theta_n|^2|\widetilde{Z}_n^{(1)}|]\right) \\ &+ \|\Upsilon_n\|\left(\|D^2 f\|_{\mathrm{Lip}}\mathbb{E}[|\widetilde{Z}_n^{(1)}|.|\theta_n|^2] + \gamma_{n+1}\|D^2 f\|_{\mathrm{Lip}}^2\mathbb{E}|\theta_n|^4\right).\end{aligned}$$

The Cauchy-Schwarz inequality and $|\widetilde{Z}_n^{(1)}| = |\theta_n|$ yield

$$\mathbb{E}[|\theta_n|^2|\widetilde{Z}_n^{(1)}|] \leqslant \left\{\mathbb{E}[|\theta_n|^4|\right\}^{1/2}\left\{\mathbb{E}[|\widetilde{Z}_n^{(1)}|^2]\right\}^{1/2} \leqslant \sqrt{c_2 c_4}\gamma_{n+1}^{3/2} \leqslant c_4\gamma_{n+1}^{3/2}.$$

Therefore, using $1 \leqslant c_2 \leqslant \sqrt{c_4} \leqslant c_4$ and $\|D^2 f\|_{\mathrm{Lip}}n^{-1} \leqslant 1$, (46) implies:

$$\begin{aligned}\gamma_{n+1}r_n^{(1)} &\lesssim_{id}\frac{1}{n^2\underline{\mu}}\left(c_2 + c_4\|D^2 f\|_{\mathrm{Lip}}\gamma_{n+1}^{\frac{3}{2}}\right) + \frac{c_4\|D^2 f\|_{\mathrm{Lip}}}{n\underline{\mu}}\left(\gamma_{n+1}^{\frac{3}{2}} + \|D^2 f\|_{\mathrm{Lip}}\gamma_{n+1}^3\right) \\ &\lesssim_{id}\frac{c_4\|D^2 f\|_{\mathrm{Lip}}}{\underline{\mu}}\left(\frac{\gamma_n^{\frac{3}{2}}}{n} + \gamma_n^3\right) \lesssim_{id}\frac{c_4\|D^2 f\|_{\mathrm{Lip}}}{\underline{\mu}}\frac{\gamma_n}{n},\end{aligned}$$

where we used that $\beta \geqslant 1/2$. In the meantime, under $(\mathbf{H_S})$ and because $Q \in O_d(\mathbb{R})$ and $c_2\gamma_n < 1$ when $n \geqslant n_1$, we have $\forall i \in \{1,\dots,d\}$,

$$\begin{aligned}\left|\mathbb{E}[\{Q\Delta\mathcal{M}_{n+1}\}_i\{\Upsilon_n Q\Delta\mathcal{M}_{n+1}\}_i]\right| &\leqslant \|\Upsilon_n\|\mathbb{E}[|\Delta\mathcal{M}_{n+1}|^2] \leqslant \|\Upsilon_n\|\mathbb{E}[\mathrm{Tr}(S(\theta_n))] \\ &\leqslant \|\Upsilon_n\|(d\mathbb{E}[\|S(\theta_n)\|]) \leqslant d\|\Upsilon_n\|(\|S(\theta^*)\| + \|S\|_{\mathrm{Lip}}\mathbb{E}|\theta_n|) \\ &\lesssim_{id} d\|\Upsilon_n\|(\|S^*\| + \|S\|_{\mathrm{Lip}}).\end{aligned}$$

since $c_2\gamma_n \leqslant 1$ for $n \geqslant n_1$. We therefore deduce from (46) and from the previous lines that

$$\forall i \in \{1,\dots,d\} \qquad \gamma_{n+1}^2\left|\mathbb{E}[\{Q\Delta\mathcal{M}_{n+1}\}_i\{\Upsilon_n Q\Delta\mathcal{M}_{n+1}\}_i]\right| \leqslant \frac{d(\|S^*\| + \|S\|_{\mathrm{Lip}})\gamma_n}{n\underline{\mu}}.$$

A compilation of the previous bounds (taking into accounts only non-universal constants) leads to

$$\begin{aligned}\left|\omega_{n+1}(i) - (1-\gamma_{n+1}\mu_i^\star)\left(1-\frac{1}{n+1}\right)\omega_n(i)\right| &\lesssim_{id}\gamma_{n+1}\|D^2 f\|_{\mathrm{Lip}}\mathbb{E}[|\theta_n|^2|\widetilde{Z}_n^{(2)}|] \\ &+ \frac{d(\|S^*\| + \|S\|_{\mathrm{Lip}}) + c_4\|D^2 f\|_{\mathrm{Lip}}}{\underline{\mu}}\frac{\gamma_n}{n}.\end{aligned}$$

$ii)$ We set $\Delta N_{n+1} = \Upsilon_n Q\Delta\mathcal{M}_{n+1}$ and recall that α_n^i is defined in (29) by $\alpha_n^i = 2(1-(n+1)^{-1})(\Omega_n)_{i,i}$. Starting from (27) and $|\widetilde{Z}_n^{(1)}| = |\theta_n|$ with a conditional expectation argument, we use that Ω_n is diagonal to obtain

$$\begin{aligned}\mathbb{E}[|\widetilde{Z}_{n+1}^{(2)}|^2] = &\left(1-\frac{1}{n+1}\right)^2\mathbb{E}[|\widetilde{Z}_n^{(2)}|^2] + \sum_{i=1}^d\alpha_n^i\omega_n(i) + \gamma_{n+1}^2\mathbb{E}|\Delta N_{n+1}|^2 \\ &+ \mathbb{E}[|\Omega_n\widetilde{Z}_n^{(1)}|^2] + \mathcal{O}_{id}\left(\gamma_{n+1}\|\Upsilon_n\|\mathbb{E}[|\theta_n|^2|\widetilde{Z}_n^{(2)}|]\right) \\ &+\mathcal{O}_{id}(\gamma_{n+1}\|D^2 f\|_{\mathrm{Lip}}\|\Upsilon_n\|\|\Omega_n\|\mathbb{E}[|\widetilde{Z}_n^{(1)}||\theta_n|^2]) + \mathcal{O}_{id}(\gamma_{n+1}^2\|D^2 f\|_{\mathrm{Lip}}^2\|\Upsilon_n\|^2\mathbb{E}[|\theta_n|^4]).\end{aligned} \tag{47}$$

First, by (46),

$$\mathbb{E}[|\Omega_n \widetilde{Z}_n^{(1)}|^2] \leqslant \|\Omega_n\|^2 \mathbb{E}[|\widetilde{Z}_n^{(1)}|^2] \leqslant c_2\gamma_n\|\Omega_n\|^2 \leqslant \frac{c_2 n^{-4}}{\underline{\mu}^2\gamma_n^2} = \frac{c_4}{\underline{\mu}^2} O\left(n^{-4+\beta}\right) \tag{I1}$$

In the meantime, (46) yields

$$\begin{aligned}\gamma_{n+1}\|\Upsilon_n\|\mathbb{E}[|\theta_n|^2|\widetilde{Z}_n^{(2)}|] &= \frac{1}{n\underline{\mu}}\mathcal{O}_{id}\left(\mathbb{E}[|\theta_n|^2|\widetilde{Z}_n^{(2)}|]\right)\\ &= \frac{1}{n\underline{\mu}}\mathcal{O}_{id}\left(\frac{\mathbb{E}[|\theta_n|^4\upsilon_n^{-1}+\upsilon_n\mathbb{E}[|\widetilde{Z}_n^{(2)}|^2]}{2}\right).\end{aligned}$$

Choosing $\upsilon_n = \underline{\mu}n^{1/2-\beta}$, we obtain

$$O\left(\gamma_{n+1}\|\Upsilon_n\|\mathbb{E}[|\theta_n|^2|\widetilde{Z}_n^{(2)}|\right) = \frac{c_4}{\underline{\mu}^2}\mathcal{O}_{id}\left(n^{-\frac{3}{2}-\beta}\right) + n^{-\frac{1}{2}-\beta}\mathbb{E}[|\widetilde{Z}_n^{(2)}|^2]. \tag{I2}$$

The $(L^p, \sqrt{\gamma_n})$ consistency property associated to the Cauchy-Schwarz inequality and the fact that $c_2 \leqslant \sqrt{c_4} \leqslant c_4$ imply that

$$\gamma_{n+1}\|\Upsilon_n\|\|D^2 f\|_{\text{Lip}}\|\Omega_n\|\mathbb{E}[|\widetilde{Z}_n^{(1)}||\theta_n|^2] = \frac{c_4\|D^2 f\|_{\text{Lip}}}{\underline{\mu}^2}\mathcal{O}_{id}\left(n^{-3-\beta/2}\right) \tag{I3}$$

Finally, we also obtain that

$$\begin{aligned}\gamma_{n+1}^2\|\Upsilon_n\|^2\|D^2 f\|^2_{\text{Lip}}\mathbb{E}[|\theta_n|^4]) &= \frac{c_4\|D^2 f\|^2_{\text{Lip}}}{\underline{\mu}^2}\mathcal{O}_{id}\left(n^{-2\beta-2}\right)\\ &= \frac{c_4\|D^2 f\|_{\text{Lip}}}{\underline{\mu}^2}\mathcal{O}_{id}\left(n^{-2\beta-1}\right),\end{aligned} \tag{I4}$$

where we used that $\|D^2 f\|_{\text{Lip}} n^{-1} \leqslant 1$.
To achieve the proof, it remains to study $\gamma_{n+1}^2\mathbb{E}|\Delta N_{n+1}|^2$. First, set $B_n = Q^T\Upsilon_n^2 Q$. Using that Υ_n is a diagonal matrix, we have

$$\begin{aligned}\gamma_{n+1}^2|\Delta N_{n+1}|^2 = \gamma_{n+1}^2\text{Tr}(|\Delta N_{n+1}|^2) &= \gamma_{n+1}^2\text{Tr}(\Delta N_{n+1}^T\Delta N_{n+1})\\ &= \gamma_{n+1}^2\text{Tr}(\Delta\mathcal{M}_{n+1}^T B_n\Delta\mathcal{M}_{n+1})\\ &= \gamma_{n+1}^2\text{Tr}(B_n\Delta\mathcal{M}_{n+1}\Delta\mathcal{M}_{n+1}^T)\end{aligned}$$

Since the trace is a linear application and B_n is a deterministic matrix,

$$\gamma_{n+1}^2\mathbb{E}[|\Delta N_{n+1}\}|^2|\mathcal{F}_n] = \gamma_{n+1}^2\text{Tr}(B_n\mathbb{E}[\Delta\mathcal{M}_{n+1}\Delta\mathcal{M}_{n+1}^T|\mathcal{F}_n]) = \gamma_{n+1}^2\text{Tr}(B_n S(\theta_n)) \tag{48}$$

where we applied Assumption $(\mathbf{H_S})$. For B_n, we first remark that

$$-\gamma_{n+1}\Upsilon_n = (n+1)^{-1}\{D^\star\}^{-1} + \Delta_{n+1}$$

where $(\Delta_n)_{n\geqslant 0}$ is a sequence of matrices defined by:

$$\Delta_n = \text{Diag}\left\{\frac{1+(n+1)\{\mu_i^\star\}^2\gamma_{n+1}^2}{(n+1)\mu_i^\star((n+1)\gamma_{n+1}\mu_i^\star-1)} + \frac{\gamma_{n+1}}{n+1},\ i=1,\ldots,d\right\}.$$

For $n \geqslant n_1$ and every $i \in \{1, \dots, d\}$, $\mu_i^\star \gamma_n n - 1 \geqslant \frac{1}{2}\mu_i^\star \gamma_n n$ (by the beginning of the proof of Lemma 12) so that

$$\frac{1 + (n+1)\{\mu_i^\star\}^2\gamma_{n+1}^2}{(n+1)\mu_i^\star((n+1)\gamma_{n+1}\mu_i^\star - 1)} \leqslant \frac{2}{(n+1)^2\{\mu_i^\star\}^2\gamma_{n+1}} + \frac{2\gamma_{n+1}}{n+1}.$$

Using the diagonal structure of Δ_n, we get for $n \geqslant n_1$,

$$\|\Delta_n\| \lesssim_{id} \frac{2}{n^2\gamma_n\underline{\mu}^2} + \frac{\gamma_{n+1}}{n+1} \lesssim_{id} \frac{1}{n^2\gamma_n\underline{\mu}^2}, \tag{49}$$

since $\beta > 1/2$. Then, using that $B_n = Q^T\Upsilon_n^2 Q$ and that $Q^T\{D^\star\}^{-2}Q = \{\Lambda^\star\}^{-1}$, it follows from (48) that

$$\begin{aligned}
\gamma_{n+1}^2\mathbb{E}[|\Delta N_{n+1}\}|^2] &= \gamma_{n+1}^2\mathbb{E}[\mathrm{Tr}(B_n S(\theta_n))] \\
&= \gamma_{n+1}^2\mathbb{E}\left[\mathrm{Tr}\left(Q^T(\{D^\star\}^{-1} + \Delta_{n+1})^2 Q S(\theta_n)\right)\right] \\
= \frac{1}{(n+1)^2}\mathrm{Tr}(\{\Lambda^\star\}^{-2}QS(\theta^\star)) + \sum_{i=1}^{3}\mathcal{R}_n^i &= \frac{\mathrm{Tr}(\Sigma^\star)}{(n+1)^2} + \sum_{i=1}^{3}\mathbb{E}[\mathcal{R}_n^i]
\end{aligned}$$

with,

$$\begin{aligned}
&\mathcal{R}_n^1 = \frac{\mathrm{Tr}\left(\{\Lambda^\star\}^{-2}(S(\theta_n) - S(\theta^\star))\right)}{(n+1)^2}, \ \mathcal{R}_n^2 = \frac{2}{n+1}\mathrm{Tr}\left(Q^T\{D^\star\}^{-1}\Delta_{n+1}QS(\theta_n)\right), \\
&\text{and} \quad \mathcal{R}_n^3 = \mathrm{Tr}\left((Q^T\{\Delta_{n+1}\}^2QS(\theta_n)\right).
\end{aligned}$$

Note that for $\mathcal{R}_n^2$, we used that $\{D^\star\}^{-1}$ and Δ_{n+1} commute. It remains to bound the remainder terms $\mathcal{R}_n^i$, $i = 1, 2, 3$. To this end, let us denote by $\|\,.\,\|_F$ the Frobenius norm defined for a square matrix A by $\|A\|_F = \sqrt{\mathrm{Tr}(A^TA)}$. Owing to the sub-multiplicativity of this norm and to the fact that $\|A\|_F \leqslant \sqrt{d}\|A\|$ (where $\|A\| = \sqrt{\rho(A^TA)}$), we obtain:

$$|\mathbb{E}[\mathcal{R}_n^1]| \leqslant \|\{\Lambda^\star\}^{-2}\|_F\|S(\theta_n) - S(\theta^\star)\|_F \leqslant \frac{d\|S\|_{\mathrm{Lip}}\mathbb{E}[|\theta_n|]}{\underline{\mu}^2 n^2} \leqslant \frac{d\|S\|_{\mathrm{Lip}}\sqrt{c_2\gamma_n}}{\underline{\mu}^2 n^2}, \tag{50}$$

Using (49), one can check that

$$|\mathbb{E}[\mathcal{R}_n^2]| \leqslant \frac{2}{n+1}\|\{D^\star\}^{-1}\Delta_{n+1}\|_F\mathbb{E}[\|S(\theta_n)\|_F] \leqslant \frac{2d(\|S^*\| + \|S\|_{\mathrm{Lip}})}{n^3\gamma_n\underline{\mu}^3}.$$

Finally,

$$|\mathbb{E}[\mathcal{R}_n^3]| \leqslant \|\Delta_{n+1}^2\|_F\mathbb{E}[\|S(\theta_n)\|_F] \leqslant \frac{2d(\|S^*\| + \|S\|_{\mathrm{Lip}})}{n^4\gamma_n^2\underline{\mu}^4} \lesssim_{id} \frac{d(\|S^*\| + \|S\|_{\mathrm{Lip}})}{n^3\gamma_n\underline{\mu}^3},$$

where in the second inequality, we used that $n^{1-\beta} \geqslant 2\underline{\mu}^{-1}$ for $n \geqslant n_0$.

A combination of the above upper bounds of $\mathcal{R}_n^i$, $i = 1, 2, 3$ yields:

$$\gamma_{n+1}^2\mathbb{E}[|\Delta N_{n+1}\}|^2] = \frac{\mathrm{Tr}(\Sigma^\star)}{(n+1)^2} + c_4\frac{d(\|S^*\| + \|S\|_{\mathrm{Lip}})}{\gamma\underline{\mu}^3}\mathcal{O}_{id}\left(n^{-(2+\beta/2)} \vee n^{-3+\beta}\right). \tag{I5}$$

Keeping in mind the expansion (47), we now compare the above control with (I1), (I2), (I3) and (I4). First, we can omit (I1) which is controlled by the above *r.h.s* since $4 - \beta > 3 - \beta$. Second, we compare the first term of the r.h.s. of (I2) with the terms involved in (I3), (I4) and (I5) and remark that

$$\forall \beta \in (1/2, 1) \qquad n^{-3-\beta/2} \vee n^{-2\beta-1} \vee n^{-\beta/2-2} = \mathcal{O}_{id}(n^{-\frac{3}{2}-\beta}).$$

Considering the worst constant of each term, we obtain that:

$$|\widehat{\Delta Z^2_{n+1}}| = \frac{c_4 d}{\underline{\mu}^3} C_{f,S}\mathcal{O}_{id}\left(n^{-\frac{3}{2}-\beta} \vee n^{-3+\beta}\right) + n^{-\frac{1}{2}-\beta}\mathbb{E}|\widetilde{Z}^{(2)}_n|^2.$$

where $C_{f,S}$ is defined in the statement of the lemma. □

Lemma 14. *Assume that $(u_n)_{n\geqslant 0}$ is a real sequence that satisfies for all $n \geqslant n_0$ and for a given $\mu > 0$:*

$$u_{n+1} = (1-\gamma_{n+1}\mu)\,\frac{n}{n+1}u_n + \beta_{n+1},$$

with $\beta_n \leqslant \frac{\square\gamma_n}{n}$. Then, a constant C independent on $\square$ exists such that

$$u_n \leqslant \frac{\square}{n}\left(n_0 u_{n_0} + \mu^{-1}\right).$$

<u>Proof:</u> With the convention $\prod_{\varnothing} = 1$ and $\sum_{\varnothing} = 0$, we have for every $n \geqslant n_0$:

$$u_n = \left(\prod_{k=n_0+1}^{n}(1-\gamma_k\mu)\frac{k}{k+1}\right)u_{n_0} + \sum\nolimits_{k=n_0+1}^{n}\beta_k \prod_{\ell=k+1}^{n}(1-\gamma_\ell\mu)\frac{\ell}{\ell+1}.$$

Using that for any $x > -1$, $\log(1+x) \leqslant x$, we obtain for every $n \geqslant n_0+1$

$$\prod_{k=n_0+1}^{n}(1-\gamma_k\mu)\frac{k}{k+1} \leqslant \frac{n_0}{n+1}e^{-\mu(\Gamma_n - \Gamma_{n_0})} \leqslant \frac{n_0}{n+1},$$

where $\Gamma_n = \sum\nolimits_{k=0}^{n}\gamma_k$. Concerning the second term, we have

$$\sum\nolimits_{k=n_0+1}^{n}\beta_k \prod_{\ell=k+1}^{n}(1-\gamma_\ell\mu)\frac{\ell}{\ell+1} \leqslant \frac{1}{n+1}\left(e^{-\mu\Gamma_n}\sum\nolimits_{k=n_0+1}^{n}\beta_k(k+1)e^{\mu\Gamma_k}\right).$$

Since $\beta_k(k+1) \leqslant \square\gamma_{k+1}$, the monotonicity of $x \longmapsto xe^{\mu x}$ yields:

$$\sum\nolimits_{k=n_0+1}^{n}\beta_k(k+1)e^{\mu\Gamma_k} \leqslant \square\sum\nolimits_{k=n_0+1}^{n}\gamma_{k+1}e^{\mu\Gamma_k} \leqslant \square\int_{\Gamma_{n_0+1}}^{\Gamma_{n+1}} e^{\mu x}dx.$$

We deduce that:

$$\frac{1}{n+1}\left(e^{-\mu\Gamma_n}\sum\nolimits_{k=n_0+1}^{n}\beta_k(k+1)e^{\mu\Gamma_k}\right) \leqslant \frac{\square}{\mu(n+1)}.$$

□

Remark 3. *Using $\log(1+x) = x + c(x)x^2$ where c is bounded on $[-1/2, 1/2]$, a modification of the proof leads to $\liminf_{n\to+\infty} nu_n > 0$ when $\sum\gamma_k^2 < +\infty$.*

Lemma 15. *Assume that (θ_n) is $(L^2, \sqrt{\gamma_n})$ consistent. Then, for all $n \geqslant 0$,*

$$\mathbb{E}|\hat{\theta}_n|^2 \lesssim_{id} \frac{c_2}{1-\beta}n^{-\beta}. \tag{51}$$

Let n_1 be defined in Lemma 13. Under the assumptions of Proposition 7:

$$\mathbb{E}|\tilde{Z}^{(2)}_{n_1}|^2 \lesssim_{id} \frac{c_2}{1-\beta}n_1^{-\beta}. \tag{52}$$

Proof: Under the assumption, $\mathbb{E}|\theta_n|^2 \leqslant c_2\gamma_n$. Keeping in mind that $\theta^\star = 0$, we deduce from the Jensen inequality that

$$\mathbb{E}\left[|\hat{\theta}_n|^2\right] \leqslant \frac{c_2}{n}\sum_{k=1}^{n} \gamma_k \lesssim_{id} \frac{c_2}{1-\beta} n^{-\beta}.$$

For the second part of the proof, we use that:

$$\tilde{Z}_n^{(2)} = -\mathcal{E}_{n,D^\star}\check{Z}_n^{(1)} + \check{Z}_n^{(2)}.$$

Thus,

$$\begin{aligned} |\tilde{Z}_n^{(2)}|_2^2 &\leqslant \rho\left(\mathcal{E}_{n,D^\star}\right)^2 |\check{Z}_n^{(1)}|^2 + |\check{Z}_n^{(2)}|^2 \\ &\leqslant |\epsilon_{\mu,n}|^2|\theta_n|^2 + |\hat{\theta}_n|^2, \end{aligned}$$

where in the second line, we used that Q is an orthogonal matrix. By Lemma 12, for every $n \geqslant n_1 \geqslant n_0$, $\forall \mu \in \mathrm{Sp}(D^\star)$, $|\epsilon_{n,\mu}| \lesssim_{id} \frac{n^{-1+\beta}}{\underline{\mu}}$ so that

$$\mathbb{E}|\tilde{Z}_{n_1}^{(2)}|^2 \lesssim_{id} \left[\frac{n_1^{-1+\beta}}{\underline{\mu}}\right]^2 \frac{c_2}{1-\beta} n_1^{-\beta} + \frac{c_2}{1-\beta} n_1^{-\beta},$$

where in the second line, we used that $n_1^{-1+\beta} \leqslant n_0^{-1+\beta} \leqslant \underline{\mu}/2$. □

Lemma 16. *Let N be a positive integer and $(u_n)_{n\geqslant 0}$ be a sequence which satisfies*

$$\forall n \geqslant N \qquad u_{n+1} \leqslant u_n \left[\left(1 - \frac{1}{n+1}\right)^2 + C_1 n^{-r}\right] + \frac{V}{(n+1)^2} + C_2 n^{-q},$$

with $r \in (1,2]$ and $q \in (2,3]$. Assume that (C_1, N) satisfies: $C_1 N^{1-r} \leqslant 1$, Then,

$$\begin{aligned} \forall n \geqslant N \qquad u_n &\leqslant \frac{V}{n} + \mathcal{O}_{id}\left(\frac{u_N N^2}{n^2} + C_1 V n^{-r} + C_2 n^{-(q-1)}\right) \\ &\leqslant \frac{V}{n} + n^{-r\wedge(q-1)}\mathcal{O}_{id}\left(u_N N^{r\wedge(q-1)} + C_1 V + C_2\right). \end{aligned}$$

Proof: For the sake of simplicity, in whole the proof, we will denote by c any universal constant ($i.e$ independent of whole the parameters of the problem). An iteration of the inequality yields for all $n \geqslant N$

$$u_n \leqslant u_N \prod_{k=N+1}^{n} \Upsilon_k + \sum_{k=N+1}^{n}\left(\frac{V}{k^2} + C_2 k^{-q}\right)\prod_{\ell=k+1}^{n}\Upsilon_\ell \tag{53}$$

with $\Upsilon_\ell = (1-1/\ell)^2 + C_1(\ell-1)^{-r}$ and the conventions $\sum_\varnothing = 0$ and $\prod_\varnothing = 1$. Remark that

$$\begin{aligned} \prod_{\ell=k+1}^{n}\Upsilon_\ell &= \frac{k^2}{n^2}\prod_{\ell=k+1}^{n}\left(1 + C_1\left(\frac{\ell}{\ell-1}\right)^2(\ell-1)^{-r}\right) \\ &\leqslant \frac{k^2}{n^2}\exp\left(2C_1\sum_{\ell=k}^{n-1}\ell^{-r}\right) \leqslant \frac{k^2}{n^2}\exp\left(\frac{2C_1}{r-1}(k-1)^{1-r}\right), \end{aligned}$$

where in the last line, we used the inequality $\log(1+x) \leqslant x$ for $x > -1$ and a comparison between series and integrals. Now, a constant c exists such that $\exp(x) \leqslant 1 + cx$ on $[0, 2]$ and with the condition $C_1 N^{1-r} \leqslant 1$, we get for every $k \in \{N+1, \ldots, n-1\}$,

$$\prod_{\ell=k+1}^{n} \Upsilon_\ell \leqslant \frac{k^2}{n^2}\left(1 + cC_1 k^{1-r}\right) \leqslant c\frac{k^2}{n^2}.$$

Plugging this inequality into (53) leads to: for all $n \geqslant N$,

$$\begin{aligned} u_n &\leqslant c\frac{u_N N^2}{n^2} + \frac{V(n-N)}{n^2} + c\frac{C_1 V}{n^2}\sum_{k=N+1}^{n} k^{1-r} + c\frac{C_2}{n^2}\sum_{k=N+1}^{n} k^{2-q} \\ &\leqslant \frac{V}{n} + c\left(\frac{u_N N^2}{n^2} + C_1 V n^{-r} + C_2 n^{-(q-1)}\right). \end{aligned}$$

This yields the first inequality. The second one follows easily.

□

Appendix B: Growth at infinity under the KL gradient inequality

In this section, we prove the property (10) of Proposition 2. Without loss of generality, we can assume that $\theta^\star = f(\theta^\star) = 0$.

Proof: Consider $0 \leqslant t \leqslant s$ and $x \in \mathbb{R}^d$. We then associate the solution of the differential equation associated to the flow $-\nabla f$ initialized at x:

$$\chi_x(0) = x \qquad \text{and} \qquad \dot{\chi}_x = -\nabla f(\chi_x).$$

The length of the curve $L(\chi_x, t, s)$ is defined by

$$L(\chi_x, t, s) = \int_t^s \|\dot{\chi}_x(\tau)\| \mathrm{d}\tau.$$

Under Assumption $(\mathbf{H_{KL}^r})$, we can consider $\varphi(a) = \frac{a^{1-r}}{1-r}$ and we have that

$$\varphi'(f(x))\|\nabla f(x)\| \geqslant m > 0.$$

We now observe that $e : s \longmapsto \varphi(f(\chi_x(s)))$ satisfies:

$$\begin{aligned} e'(\tau) &= \varphi'(f(\chi_x(\tau)))\langle \nabla f(\chi_x(\tau)), \dot{\chi}_x(\tau)\rangle \\ &= -\varphi'(f(\chi_x(\tau)))\|\nabla f(\chi_x(\tau))\|^2 \\ &\leqslant -m\|\dot{\chi}_x(\tau)\| \end{aligned}$$

We deduce that:

$$e(t) - e(s) = \int_s^t e'(\tau)\mathrm{d}\tau \geqslant m\int_t^s \|\dot{\chi}_x(\tau)\|\mathrm{d}\tau \geqslant mL(\chi_x, t, s) \tag{54}$$

Now choosing $t = 0$ and $s \longrightarrow +\infty$, we have $e(0) - \lim_{s\longrightarrow+\infty} e(s) = \varphi(f(x)) - \varphi(\min f) = \varphi(f(x))$, and Equation (54) yields

$$\varphi(f(x)) \geqslant mL(\chi_x, 0, +\infty) \geqslant m\|x\|$$

because $\chi_x(+\infty) = \arg\min f = 0$. We deduce that

$$f(x) \geqslant \varphi^{-1}(m\|x\|) = \{m(1-r)\}^{\frac{1}{1-r}} \|x\|^{\frac{1}{1-r}}.$$

which is the desired conclusion. □