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

Sébastien Gadat, Fabien Panloup

May 26, 2020

Abstract

This paper is devoted to the non-asymptotic analysis of the Ruppert-Polyak averaging method introduced in [22] and [24] 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

$$E|\hat{\theta}_n - \arg\min_p f(p)|^2 \leq \frac{\text{Tr}(\Sigma^*)}{n} + C_{d, f} n^{-r\beta},$$

where $\Sigma^*$ is the limiting covariance matrix of the CLT demonstrated in [22] 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 $r = 3/4$. Second, we show that this result holds under some Kurdyka-Lojasiewicz-type condition [18, 19] 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 $\theta^* = \arg\min_p f(p)$. Let us assume that $\nabla f$ 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:

$$\forall \theta \in \mathbb{R}^d, \quad \nabla f(\theta) = \mathbb{E}_Z \sim_n [\Lambda(\theta, Z)].$$

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

1.1 Averaging principle for stochastic algorithms

Stochastic Gradient Descent The Robbins-Monro procedure (see [23]) is built with an i.i.d. sequence of observations $(Z_i)_{i \geq 1}$ distributed according to $\mu$. Under some mild assumptions, the minimizers of $f$ can be approximated with a stochastic gradient descent (SGD) $(\theta_n)_{n \geq 0}$ defined by: $\theta_0 \in \mathbb{R}^d$ and

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

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

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

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

$$\forall n \geq 0 : \quad \theta_{n+1} = \theta_n - \gamma_{n+1} \nabla f(\theta_n) + \gamma_{n+1} \Delta M_{n+1},$$

where $(\Delta M_{n+1})_{n \geq 0}$ stands for a sequence of noises (martingale increments), i.e. $\forall n \geq 1, \quad \mathbb{E} [\Delta M_{n+1} | \mathcal{F}_n] = 0$, where $(\mathcal{F}_n)_{n \geq 0}$ is the filtration defined by $\mathcal{F}_n = \sigma(Z_1, \ldots, Z_n)$ for $n \geq 1$, $\mathcal{F}_0$ is the trivial $\sigma$-field and for a given $\sigma$-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:

\[ \hat{\theta}_n = \frac{1}{n} \sum_{k=1}^{n} \theta_k, \quad n \geq 1. \]

This averaging procedure is a way to improve the convergence properties of the original SGD \((\theta_n)_{n \geq 1}\). We recall the CLT associated with \((\hat{\theta}_n)_{n \geq 0}\), the statement is adapted from [22] in the strongly convex situation \(SC(\alpha)\):

\[ SC(\alpha) := \{ f \in C^2(\mathbb{R}^d) : D^2f - \alpha I_d \succeq 0 \} \]

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

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

\[ \sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{n \to +\infty} \mathcal{N}(0, \Sigma^*) \quad \text{with} \quad \Sigma^* = \left( D^2f(\theta^*) \right)^{-1} S^* \left( D^2f(\theta^*) \right)^{-1}. \]

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 [22] for the initial asymptotic description and to [17] for some more general results. In [5], 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 (online logistic regression [3], recursive median estimation [10, 15] for example). But a non-asymptotic result for a more general class of functions that preserves a sharp optimal \(O(n^{-\frac{1}{2}})\) rate of the \(L^2\)-risk is missing yet.

**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 \(\text{Tr}(\Sigma^*)n^{-\frac{1}{2}}\), 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 \(\text{Tr}(\Sigma^*)n^{-1}\):

\[ E[|\hat{\theta}_n - \theta^*|^2] \leq \text{Tr}(\Sigma^*)n^{-1} + a_2 n^{-\rho}. \]

We emphasize that the leading term \(\text{Tr}(\Sigma^*)n^{-1}\) 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 \(\text{Tr}(\Sigma^*)\) naturally grows with the dimension of the ambient space \(d\), 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 \(\text{Tr}(\Sigma^*)n^{-1}\) 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.

### 1.2 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** (**H**\(_{\text{KL}}\)). The function \(f\) is \(C^2(\mathbb{R}^d, \mathbb{R})\) with \(D^2f\) bounded and Lipschitz, \(D^2f(\theta^*)\) invertible and for \(r \in [0, 1/2]\):

\[ \liminf_{|x| \to +\infty} f^{-r} |\nabla f| > 0 \quad \text{and} \quad \limsup_{|x| \to +\infty} f^{-r} |\nabla f| < +\infty \]

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

\[ ^1 \text{In [22], the result is stated in a more general framework with the help of a Lyapunov function. We have chosen to simplify the statement for the sake of readability.} \]
Theorem 2. Assume $\mathcal{H}_{\text{KL}}^*$. Suppose that the covariance of the martingale increment is Lipschitz continuous and that $\mathbb{E}[|\Delta M_{n+1}[6] e^{(1+|\Delta M_{n+1}|)^{1/2-r} |F_n]} < +\infty$ a.s.. Then, a constant $C_r$ exists such that:

$$
\mathbb{E}(\hat{\theta}_n - \theta^* |^2) \leq \frac{\text{Tr}(\Sigma^*)}{n} + C_r \left( \frac{\nu d}{\mu} \right)^3 n^{-(\beta+1/2) \cdot (2-\beta)},
$$

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

Our main result is therefore an optimal non-asymptotic bound of the $L^2$-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 $O(\text{Tr}(\Sigma^*) n^{-1})$ 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 $\theta_n$ (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-Lojasiewicz inequality $\mathcal{H}_{\text{KL}}^*$ (see [18, 19]), which is a much weaker situation than the traditional 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 approximations (among others). A range of applications are listed in the next table, enriched by a comparison with existing results in the literature:

<table>
<thead>
<tr>
<th>Setting</th>
<th>Cramer-Rao</th>
<th>2nd order $v_n$</th>
<th>$\gamma_n = \gamma_1 n^{-\beta}$</th>
<th>Anytime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our work</td>
<td>Yes : $\text{Tr}(\Sigma^*)$</td>
<td>$n^{-(\beta+1/2) \cdot (2-\beta)}$, $v_n^* = O(n^{-\frac{3}{4}})$</td>
<td>$\beta \in (1/2, 1)$, $\beta^* = 3/4$</td>
<td>Yes</td>
</tr>
<tr>
<td>BM(11) [5]</td>
<td>Strong. Convex</td>
<td>$n^{-(\beta+1/2) \cdot (2-\beta)}$, $v_n^* = O(n^{-\frac{3}{4}})$</td>
<td>$\beta \in (1/2, 1)$, $\beta^* = 2/3$</td>
<td>Yes</td>
</tr>
<tr>
<td>BM(11) [5]</td>
<td>Convex</td>
<td>No: $O(n^{-1/2})$</td>
<td>$\emptyset$</td>
<td>$\beta = 1/2$</td>
</tr>
<tr>
<td>BM(11) [5]</td>
<td>Logist. Reg.</td>
<td>No: $O(n^{-1/2})$</td>
<td>$\emptyset$</td>
<td>$\beta = 1/2$</td>
</tr>
<tr>
<td>BM(11) [5]</td>
<td>Recurs. Quantile</td>
<td>No: $O\left(\frac{1}{n^{\gamma_1 (1+D^2 f(\theta^*))}}\right)$</td>
<td>$\emptyset$</td>
<td>$\beta = 1/2$</td>
</tr>
<tr>
<td>CCGB(17) [15]</td>
<td>Recurs. Quantile</td>
<td>No: $O\left(\frac{1}{n}\right)$</td>
<td>$n^{-(\beta+1/2) \cdot (2-\beta)}$, $v_n^* = O(n^{-\frac{7}{4}})$</td>
<td>$\beta \in (1/2, 1)$, $\beta^* = 2/3$</td>
</tr>
</tbody>
</table>

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

## 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 $\otimes$ is used to refer to the following quadratic form:

$$
\forall M \in \mathcal{M}_d(\mathbb{R}) \hspace{1em} \forall y \in \mathbb{R}^d \hspace{1em} My \otimes y^T = y^T M y.
$$
$I_d$ is the identity matrix and $\mathcal{O}_d(\mathbb{R})$ denotes the set of orthonormal matrices:

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

Finally, the notation $\| \cdot \|$ corresponds to the operator 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:

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

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

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

For two positive sequences $(a_n)_{n \geq 1}$ and $(b_n)_{n \geq 1}$, the notation $a_n \leq c b_n$ refers to a domination relationship, i.e. $a_n \leq c b_n$ 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 $|a_n| \leq c |b_n|$.

### 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 $(\theta_n)_{n \geq 1}$ ($(L^p, \sqrt{\gamma_n})$-consistency). We introduce the next property:

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

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

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 $c_q \leq c_p^{q/p}$.

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

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

**Assumption (H8)** The covariance of the martingale (3) satisfies:

$$\mathbb{E} \left[ \Delta M_{n+1} \Delta M_{n+1}^T \mid \mathcal{F}_n \right] = S(\theta_n) \quad \text{a.s.}$$

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

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

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

When compared to Theorem 1, Assumption (H8) 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 \((\theta_n)_{n \geq 1}\) is \((L^4, \sqrt{n})\)-consistent, if \((H_S)\) holds and \(D^2 f(\theta^*)\) is positive-definite, then for any \(n\):

\[
\mathbb{E} [\hat{\theta}_n - \theta^*]^2 \leq \frac{\text{Tr}(\Sigma^*)}{n} + C_\beta (c_4, f, S) \left( \frac{\sqrt{d}}{\mu} \right)^3 n^{-r_\beta} \text{ with } r_\beta = \left( \beta + \frac{1}{2} \right) \wedge (2 - \beta),
\]

and \(\Sigma^*\) is defined in Equation (5) (with \(S^* = S(\theta^*)\)). In particular, \(r_\beta > 1\) for all \(\beta \in (1/2, 1)\) and \(\beta \mapsto r_\beta\) 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 \(\text{Tr}(\Sigma^*)\) 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 \(D^2 f(\theta^*)\): any sequence \(\gamma_n = \gamma n^{-\beta}\) with \(\beta \in (1/2, 1)\) and \(\gamma > 0\), regardless the value of \(\beta\) or \(\gamma\), produces the result of Theorem 3. Such an adaptive property does not hold for the initial sequence \((\theta_n)_{n \geq 1}\) as proved by the CLT satisfied by the SGD \((\theta_n)_{n \geq 1}\) (see [12] for example).

- **Second order term:** Even though any value of \(\beta \in (1/2, 1)\) yields a \(\frac{\text{Tr}(\Sigma^*)}{n}\) leading term, the “optimal” choice of \(\beta\) remains unclear. In [5] and [15], \(\beta = 2/3\) is motivated by the optimization of the second order term. In particular, [5] obtains in the strongly convex case an upper bound of the order \(\frac{\text{Tr}(\Sigma^*)}{n} + O(n^{-7/6})\): Theorem 3 of [5] ensures that:

\[
\sqrt{\mathbb{E} [\hat{\theta}_n - \theta^*]^2} \leq \sqrt{\frac{\text{Tr}(\Sigma^*)}{n}} + C n^{-2/3},
\]

which in turn implies that \(\mathbb{E} [\hat{\theta}_n - \theta^*]^2 \leq \frac{\text{Tr}(\Sigma^*)}{n} + 2C \sqrt{\text{Tr}(\Sigma^*)} \frac{n^{-2/3}}{\sqrt{n}} + C^2 n^{-4/3} = \frac{\text{Tr}(\Sigma^*)}{n} + O(n^{-7/6})\).

Our Theorem 3 improves this second order term: \(\beta = 3/4\) leads to an upper bound of the order \(\frac{\text{Tr}(\Sigma^*)}{n} + O(n^{-5/4})\). 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 [5]. 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 \(h\):

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

where \(Z\) is a multivariate Gaussian random variable \(\mathcal{N}(0, \Sigma^*)\). 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 \(\mu\) is slightly stronger on the second order term (proportional to \(n^{-r_\beta}\) up to \(d^{3/2} \mu^{-3}\)) than on the first order one (proportional to \(n^{-1}\) up to \(d \mu^{-2}\)). To the best of our knowledge, Theorem 3 is the first non-asymptotic regret analysis of the Ruppert-Polyak algorithm for a large class of functions and identifies a dimension-dependent upper bound \(d^{3/2} \mu^{-3} n^{-r_\beta}\).

- **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 \((\theta_n, \hat{\theta}_n)_{n \geq 1}\). This spectral analysis requires a preliminary linearization step of the drift from \(\hat{\theta}_n\) to \(\theta_{n+1}\). The cost of this linearization is absorbed by a preliminary control of the initial sequence \((\theta_n)_{n \geq 1}\), obtained with the \((L^p, \sqrt{n})\)-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 \( \theta^* \).

- **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. [11]), 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 [5] (for strongly convex function) and [15] (quantile estimation), the sequence \((\gamma_n)_{n \geq 1}\) 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 \(\frac{1}{2^{\frac{n}{2}}}\), 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 \((\theta_n)_{n \geq 1}\) 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 \((H_{\Sigma^p}^{SC})\).

**Assumption \((H_{\Sigma^p}^{SC})\)** For a given \(p \in \mathbb{N}^*\), a constant \(\Sigma_p\) exists such that

\[
\forall n \geq 0 \quad \mathbb{E}[|\Delta M_{n+1}|^2p | F_n] \leq \Sigma_p (1 + (f(\theta_n))^p) \quad a.s.
\]

We emphasize that even though \(SC(\alpha)\) is a potentially restrictive assumption on \(f\), \((H_{\Sigma^p}^{SC})\) is not restrictive and allows a polynomial dependency in \(f(\theta_n)\) of the moments of \(\Delta M_n\), which is much weaker than the bounded increments used in [5]. 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 [12] that only provides an almost sure convergence of \((\theta_n)_{n \geq 1}\) towards \(\theta^*\) without any rate. In this setting, we can state the following proposition, whose proof is left to the reader and up to some minor modifications, is contained in the more general result stated in Theorem 11.

**Proposition 1.** Assume that \(f\) is \(SC(\alpha)\), \(x \mapsto D^2 f(x)\) is Lipschitz bounded and \((\Delta M_n)_{n \geq 1}\) satisfies \((H_{\Sigma^p}^{SC})\). Then \((\theta_n)_{n \geq 1}\) is \((L^p, \sqrt{\gamma_n})\)-consistent, i.e.

\[
\forall p \geq 1 \quad \exists c_p > 0 \quad \mathbb{E}[|\theta_n - \theta^*|^p] \leq c_p (\gamma_n)^{p/2}.
\]

A direct consequence of Proposition 1 and Theorem 3 is the next corollary.

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

\[
\forall n \in \mathbb{N}^* \quad \mathbb{E} \left[ \left| \hat{\theta}_n - \theta^* \right|^2 \right] \leq \frac{\text{Tr}(\Sigma^*)}{n} + C_\beta(c_4, f, S) \left( \frac{\sqrt{d}}{\mu} \right)^3 n^{-r_\beta}
\]

where \(r_\beta\) is defined in Theorem 3 and \(C_\beta(c_4, f, S)\) in Proposition 7.
2.3.2 \((L^p, \sqrt{n})\)-consistency without strong convexity

In some interesting cases, the latter \(SC(\alpha)\) is not suitable because the repelling effect towards \(\theta^*\) 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 \(\nabla f\) is asymptotically flat for large values of \(|x|\). Motivated by these examples, we generalize the class of functions \(f\) for which the \((L^p, \sqrt{n})\)-consistency property holds. For this purpose, we define Assumption \((H_\phi)\) by:

**Assumption \((H_\phi)\)** \(D^2f\) is bounded and Lipschitz, \(D^2f(\theta^*)\) invertible and:

- i) \(\phi\) is \(C^2(\mathbb{R}_+, \mathbb{R}_+)\) non-decreasing and \(\exists x_0 \geq 0 : \forall x \geq x_0, \phi''(x) \leq 0\).
- ii) Two positive numbers \(m\) and \(M\) exist such that \(\forall x \in \mathbb{R}^d\setminus\{\theta^*\}:
  \begin{align*}
  0 < m \leq \phi'(f(x))|\nabla f(x)|^2 + \frac{|\nabla f(x)|^2}{f(x)} \leq M.
  \end{align*}

Roughly speaking, \(\phi\) quantifies the deficit of convexity far from \(\theta^*\).

When \(\phi \equiv 1\), we recover the previous case: \(SC(\alpha) \implies (H_\phi)\) with \(\phi \equiv 1\). Actually, in this case, \(\alpha_1 > 0\) and \(\alpha_2 > 0\) exist such that

\[
\frac{\alpha_1}{2}|x - \theta^*|^2 \leq f(x) \leq \frac{\alpha_2}{2}|x - \theta^*|^2, \quad \text{and} \quad \alpha_1|x - \theta^*| \leq |\nabla f(x)| \leq \alpha_2|x - \theta^*|.
\]

But \((H_\phi)\) is more general since it can be true even when \(D^2f\) vanishes. The opposite case is \(\phi(x) = x\). In this setting, \((H_\phi)\) is satisfied when \(m \leq |\nabla f(x)|^2 \leq M\) with some positive \(m\) and \(M\). Note that this framework includes the logistic regression and the recursive quantile (see Subsection 2.4).

For practical purposes, we introduced in Section 1.2 a parametric version of Assumption \((H_\phi)\) denoted by \((H_{KL}^\phi)\), which may be seen as a global Kurdyka-Lojasiewicz gradient inequality (see, e.g., [18, 19] and Subsection 2.4 for details). \((H_\phi)\) and \((H_{KL}^\phi)\) are linked by the following proposition.

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

\[
\lim\inf_{|x| \to +\infty} \frac{f(x)|x|^{1-2r}}{|x|} > 0.
\]

The implication is easy to prove: near \(\theta^*\), \(f(x)\) is \(|\nabla f(x)|^2\) and \(|x - \theta^*| \leq |\nabla f(x)|\) since \(\nabla f(\theta^*) = 0\) and \(D^2f(\theta^*) > 0\) (in the sense of symmetric matrices) so that \(x \mapsto \frac{|\nabla f(x)|^2}{|x|}\) is lower-bounded by a positive constant near \(\theta^*\) (and on any compact set of \(\mathbb{R}^d\)). 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 \(|D^2f| < +\infty\) and from (7) again. The proof of (10) is postponed to Appendix 4.2. Note that this property will be important to derive the \((L^p, \sqrt{n})\)-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 \((H_\phi)\) and \((H_{KL}^\phi)\).

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

**Assumption \((H_{\Sigma_p}^\phi)\)** A constant \(\bar{\Sigma}_p\) exists such that:

\[
\forall n \geq 0, \quad \mathbb{E}[|\Delta M_{n+1}|^{2p+2}e^{\phi(|\Delta M_{n+1}|^2)}|F_n] \leq \bar{\Sigma}_p \quad \text{a.s.}
\]

**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 \((\Delta M_n)_{n \geq 1}\) is bounded, \((H_{\Sigma_p}^\phi)\) is automatically satisfied (this is the case for the recursive quantile and for the logistic regression of bounded variables; see Subsection 2.5). Finally, the constraint on the noise increases with \(\phi\), i.e., with the deficit of convexity.
We state the main result for a potentially non-convex function \( f \).

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

i) A constant \( c_p \) exists such that:
\[
\mathbb{E}\left[f^p(\theta_n)e^{\phi(f(\theta_n)))}\right] \leq c_p\{\gamma_n\}^p.
\]

ii) If \( \liminf_{|x| \to +\infty} |x|^{-2p}f^p(x)e^{\phi(f(x))} > 0 \), then \( (\theta_n)_{n \geq 1} \) is \((L^{2p}, \sqrt{n})\)-consistent:
\[
\mathbb{E}|\theta_n - \theta^*|^2 \leq c_p\{\gamma_n\}^p.
\]

iii) If \((\mathbf{H}_{KL}^\phi)\) holds, \((\theta_n)_{n \geq 1} \) is \((L^{2p}, \sqrt{n})\)-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 \( \inf_{x \in \mathbb{R}^d \setminus \{0\}} \tau(x)|x - \theta^*|^{-2p} > 0 \). Near \( \theta^* \), \( D^2f(\theta^*) \) is positive-definite and \( x \rightarrow \tau(x)|x - \theta^*|^{-2p} \) is lower-bounded by a positive constant. Since \( \tau \) 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 \((\mathbf{H}_{KL}^\phi)\). It is a consequence of (10) and \( \phi(x) = (1 + |x|^2)^{\frac{1}{2p}} \). \( \square \)

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

**Corollary 6.** Assume \((\mathbf{H}_S), (\mathbf{H}_\phi)\) and \((\mathbf{H}_{\Sigma_p}^\phi)\) with \( p = 2 \), then:
\[
\forall n \in \mathbb{N}^* \quad \mathbb{E}\left[|\tilde{\theta}_n - \theta^*|^2\right] \leq \frac{\text{Tr}(\Sigma^\star)}{n} + C_{\beta}(c_4, f, S) \left(\frac{\sqrt{d}}{\mu}\right)^3 n^{-r_\beta},
\]
where \( c_4 \) is given in Theorem 11 and \( r_\beta \) 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 \( \theta^* \) without strong convexity, including, for example, some situations where \( f(x) \sim |x| \) as \( |x| \rightarrow +\infty \). 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 [20]). 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 \( C^2(\mathbb{R}^d, \mathbb{R}) \) and have a positive curvature around \( \theta^* \).

In particular, the worst case is attained in [2] through linear combinations of shifted piecewise affine functions \( x \rightarrow |x + 1/2| \) and \( x \rightarrow |x - 1/2| \), functions for which Assumption \((\mathbf{H}_\phi)\) is obviously not satisfied. According to the results in Appendix H of [4], the local curvature near \( \theta^* \) 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 \( (\theta_n)_{n \geq 0} \). We will use a Lyapunov analysis that will involve \( f^p e^{\phi(t)} \) instead of \( f \) itself for the sequence \( (\theta_n)_{n \geq 0} \). The descent property will then be derived from Equation (9) in ii) of \( (H_\phi) \). Thereafter, we will be able to exploit a spectral analysis of the dynamical system that governs \( (\tilde{\theta}_n)_{n \geq 0} \). 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. [12]). As will be shown later on, \( (H_\phi) \) will be sufficient to derive efficient convergence rates for the averaged sequence \( (\tilde{\theta}_n)_{n \geq 0} \) without any strong convexity.

**f is necessarily sub-quadratic and \( L \)-smooth**  \( (H_\phi) \) entails an a priori upper bound for \( f \) that cannot increase faster than a quadratic form. We have:

\[
\forall x \in \mathbb{R}^d \quad \frac{|\nabla f(x)|^2}{f(x)} \leq M \implies |\nabla \sqrt{f}| \leq \frac{\sqrt{M}}{2} \\
\implies f(x) \leq \frac{M}{4} |x - \theta^*|^2.
\]

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 [21] 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. [6]).

**About the Kurdyka-Lojasiewicz inequality**  \( (H_\phi) \) should be related to the KL inequalities. The Lojasiewicz gradient inequality [19] with exponent \( r \) is:

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

while a generalization (see, e.g., [18]) is governed by the existence of a concave increasing "desingularizing" function \( \psi \) such that: \( |\nabla (\psi \circ f)| \geq 1 \). The Lojasiewicz gradient inequality is then just a particular case of the previous inequality while choosing \( \psi(t) = ct^{1-r} \). We refer to [7] 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 \( (H_{\text{KL}}) \) with \( r \in [0, 1/2] \), which implies \( (H_\phi) \) (see Proposition 2). However, it should be noticed that \( (H_{\text{KL}}) \) 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 \( D^2 f(\theta^*) > 0 \), it is easy to prove that \( (H_{\text{KL}}) \) implies (12) everywhere: around \( \theta^* \), \( D^2 f(\theta^*) \) is positive definite so that we could choose \( r = 1/2 \) and then satisfy the Lojasiewicz gradient inequality (12) near \( \theta^* \) so that the link between \( (H_{\text{KL}}) \) given in (7) and (12) has to be understood for large values of \( |x| \).

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

Finally, the interest of \( (H_\phi) \) in the stochastic framework is related to the behavior of the algorithm when \( (\theta_n)_{n \geq 1} \) is far from \( \theta^* \), whereas in the deterministic framework, the main interest of the desingularizing function \( \psi \) is used around \( \theta^* \) to derive fast linear rates even in non strongly convex situations (see e.g. [8]). 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 \( \theta^* \) 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 \( \nabla f \) around \( \theta^* \) (see, e.g. [8]).
Dissipative condition  We also observe from Proposition 2 that (Hₙ) 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

\[ f(x) \leq \alpha |x|^2 - \beta. \]

Consider \( f(x) = |x|^\rho \), we verify easily that dissipativity holds for \( \rho \geq 2 \) whereas \( (H_\rho) \) is verified when \( \rho \leq 2 \).

Counter-examples of the global KL inequality  Finally, we should have in mind what kind of functions do not satisfy the global Lojasiewicz inequality (12). Since we assumed \( f \) to have a unique minimizer \( \theta^* \) with \( D^2 f(\theta^*) \) invertible, \( f^{-1}[\nabla f] \geq m > 0 \) should only fail asymptotically. From Equation (10) of Proposition 2, we know that \( |x| \leq \omega f(x) \) 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 \( f' \geq 0 \) with longer and longer areas near \( f' = 0 \) when \( |x| \) is increasing. We refer to [8] for the following function that does not satisfy KL for any \( r \geq 2 \): \( f(x) = x^{2r}[2 + \cos(x^{-1})] \) 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., [12]), \( \theta \rightarrow f(\theta) \) is quadratic. In that case, the problem is strongly convex, and the noise increment satisfies:

\[ \mathbb{E}[\Delta M_n|^{2p}|F_n] \leq \Sigma_p (1 + (f(\theta_n))^p \quad \text{a.s.} \]

Then Proposition 1 yields the \( (L^p, \sqrt{n}) \) consistency rate of \( (\hat{\theta}_n)_{n \geq 1} \), which implies a first-order optimal excess risk for \( (\hat{\theta}_n)_{n \geq 1} \) with a \( O(n^{-5/4}) \) second-order term. We stress that [5] 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 \( \Delta M_n \). Hence, Corollary 4 yields a stronger result in that case.

Assumptions \( (H_\rho) \) and \( (H_{\Delta \rho}^\infty) \) hold in many situations

- Semi-algebraic case  Before explicit examples, an argument relies on the statement of Theorem 2 of [7]: a coercive convex proper and semi-algebraic continuous function \( f \) (see [7] for some details), satisfies the KL inequality.

- On-line logistic regression  The logistic regression corresponds to:

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

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

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

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

\[ \theta_{n+1} = \theta_n + \gamma_{n+1} \frac{Y_nX_n}{1 + e^{<\theta_n,\theta_n>}} = \theta_n - \gamma_{n+1} \nabla f(\theta_n) + \gamma_{n+1} \Delta M_{n+1}. \tag{15} \]

We state the following result:

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

i) \( f \) defined in (13) is convex with \( D^2 f \) bounded and Lipschitz continuous, \( D^2 f(\theta^*) \) is invertible and \( f \) satisfies \( (H_{KL}^r) \) with \( r = 0 \).
Theorem 3 for the averaging procedure \((\hat{\theta}_n)_{n \geq 1}\), 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., [12] for details). For a given cumulative distribution $G$ defined over $\mathbb{R}$, the problem is to find $q_\alpha$ such that $G(q_\alpha) = 1 - \alpha$. We assume that we observe a sequence of i.i.d. $(X_i)_{i \geq 1}$ distributed with a cumulative distribution $G$. The recursive quantile is then:

$$
\theta_{n+1} = \theta_n - \gamma_{n+1} [1_{X_n \leq \theta_n} - (1 - \alpha)] = \theta_n - \gamma_{n+1} [G(\theta_n) - (1 - \alpha)] + \gamma_{n+1} \Delta M_{n+1},
$$

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

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

where $p$ is the density with respect to the Lebesgue measure such that $G(g) = \int_{-\infty}^{g} p$. 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:

$$
f(\theta) := \int_{q_\alpha}^{\theta} \sum_{u} p(s)dsdu,
$$

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 $f'(\theta) \to 1 - \alpha$ when $\theta \to +\infty$ while $f'(\theta) \to -\alpha$ when $\theta \to -\infty$. Therefore, $f$ satisfies $(H_\phi)$ since $(H_{KL})$ and Equation (12) hold with $r = 0$ and $\phi(t) = t$. Again, regarding Assumption $(H_\phi^s)$, we can observe that the martingale increments are bounded. Therefore, Inequality (11) is obviously satisfied since $\phi$ is a monotone increasing function. Corollary 6 implies that $(\hat{\theta}_n)_{n \geq 1}$ satisfies:

$$
\forall n \geq 1 \quad \mathbb{E}[\hat{\theta}_n - q_\alpha]^2 \leq \frac{\alpha(1 - \alpha)}{p(q_\alpha)} n + \mathcal{O} \left( \frac{n^{-5/4}}{p(q_\alpha)^2} \right)
$$

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 $(\hat{\theta}_n)_{n \geq 1}$ 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 $(L^p, \sqrt{n})$-consistency under $(H_\phi)$ (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 $\theta^* = 0$ and that $f(\theta^*) = 0$. Our proof relies on a spectral strategy developed by [14] 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 $n$:

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

using the Taylor formula and the fact that $\nabla f(\theta^*) = 0$. 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:

$$Z_{n+1} = \left( I_d - \gamma_{n+1}A_n + \frac{1}{n+1}(I_d - \gamma_{n+1}A_n) \right) Z_n + \gamma_{n+1} \left( \frac{\Delta M_{n+1}}{\Delta \varphi_{n+1}} \right). \tag{17}$$

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

$$\forall n \in \mathbb{N} \quad \left\{ \begin{array}{l}
\theta_{n+1} = \theta_n - \gamma_{n+1} \nabla f(\theta_n) + \gamma_{n+1} \Delta M_{n+1} \\
\hat{\theta}_{n+1} = \hat{\theta}_n (1 - \frac{1}{n+1}) + \frac{1}{n+1}(\theta_n - \gamma_{n+1} \nabla f(\theta_n) + \gamma_{n+1} \Delta M_{n+1})
\end{array} \right..$$

The result then follows from (16).

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

Proposition 5. $Q \in \mathcal{O}_d(\mathbb{R})$ exists such that $\dot{Z}_n = \left( \begin{array}{cc} Q & 0 \\ 0 & Q \end{array} \right) Z_n$ satisfies:

$$\dot{Z}_{n+1} = A_n \dot{Z}_n + \gamma_{n+1} \left( \frac{Q \Delta M_{n+1}}{Q \varphi_{n+1}} \right) + \gamma_{n+1} \left( \frac{Q(\Lambda^* - \Lambda_n) \theta_n}{Q(\Lambda_n - \Lambda^*) \theta_n} \right), \tag{18}$$

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

$$A_n := \left( I_d - \gamma_{n+1}D^* + \frac{1}{n+1}(I_d - \gamma_{n+1}D^*) \right) \left( 1 - \frac{1}{n+1} \right) I_d \tag{19}.\quad \frac{\theta_n}{\Lambda^* - \Lambda_n}$$

Proof: We write $\Lambda_n = D^2 f(\theta^*) + (\Lambda_n - D^2 f(\theta^*))$ and use the spectrum of $\Lambda^*$.

$$Z_{n+1} = \left( I_d - \gamma_{n+1}A^* + \frac{1}{n+1}(I_d - \gamma_{n+1}A^*) \right) Z_n + \gamma_{n+1} \left( \frac{\Delta M_{n+1}}{\Delta \varphi_{n+1}} \right) + v_n, \tag{20}$$

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

$$v_n := \gamma_{n+1} \left( \frac{\theta_n}{\Lambda_n - \Lambda^*} \right).$$

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

$$\dot{Z}_n = \left( \begin{array}{cc} Q & 0 \\ 0 & Q \end{array} \right) Z_n = \left( \begin{array}{c} Q \theta_n \\ Q \hat{\theta}_n \end{array} \right). \tag{21}$$

Using $Q\Lambda^* = D^* Q$, we obtain the equality described in Equation (18). An important feature about $(Z_n)_{n \geq 1}$ is the blockwise structure of $A_n$:

$$A_n = \begin{pmatrix}
1 - \gamma_{n+1} \mu_1 & 0 & \cdots & 0 \\
0 & 1 - \gamma_{n+1} \mu_2 & \cdots & \vdots \\
\vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & 1 - \gamma_{n+1} \mu_d \\
\frac{1 - \gamma_{n+1} \mu_1}{n+1} & 0 & \cdots & 0 \\
0 & \frac{1 - \gamma_{n+1} \mu_2}{n+1} & \cdots & \vdots \\
\vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & \frac{1 - \gamma_{n+1} \mu_d}{n+1}
\end{pmatrix} \left( 1 - \frac{1}{n+1} \right) I_d$$

\[13\]
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 \(n \geq 1\), 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:
  \[
  \epsilon_{\mu,n+1} := \frac{1 - \mu \gamma_{n+1}}{1 - \mu \gamma_{n+1}(n + 1)},
  \]
  \(\text{The eigenvalues of } E_{\mu,n} \text{ are then given by } \text{Sp}(E_{\mu,n}) = \left\{ 1 - \mu \gamma_{n+1}, 1 - \frac{1}{n+1} \right\}, \text{ whereas the associated eigenvectors are:} \]
  \[
  u_{\mu,n} = \begin{pmatrix} 1 \\ \epsilon_{\mu,n+1} \end{pmatrix} \text{ and } v = \begin{pmatrix} 0 \\ 1 \end{pmatrix};
  \]
- 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 \((\hat{\theta}_n)_{n \geq 1}\) (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 \((\hat{Z}_n)_{n \geq 1}\).
- 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 [25, 13, 9, 14]).
- The first eigenvalue of \(E_{\mu,n}\) is \(1 - \mu \gamma_{n+1}\), and essentially acts on the component \(\theta_n\) 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 \(\mu\). 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 \((\hat{\theta}_n)_{n \geq 1}\).

From the factorization \(E_{\mu,n} = \left( \begin{array}{cc} 1 & 0 \\ \epsilon_{\mu,n+1} & 1 \end{array} \right) \left( \begin{array}{cc} 1 - \mu \gamma_{n+1} & 0 \\ 0 & 1 - \frac{1}{n+1} \end{array} \right) \left( -\epsilon_{\mu,n+1} & 1 \\ 1 & 0 \end{array} \right)\), we define the diagonal matrix \(\mathcal{E}_{n,D^*}\) by:
\[
\mathcal{E}_{n,D^*} = \text{Diag}(\epsilon_{\mu^1,n+1}, \ldots, \epsilon_{\mu^m,n+1});
\]
we then deduce the spectral decomposition of \(A_n\):
\[
A_n = \left( I_d \begin{array}{cc} 0 & 0 \\ 0 & I_d \end{array} \right) \left( I_d - \gamma_{n+1} D^* \begin{array}{cc} 0 & 0 \\ 0 & (1 - \frac{1}{n+1})I_d \end{array} \right) \left( -I_d \begin{array}{cc} 0 & 0 \\ 0 & I_d \end{array} \right).
\]
We introduce the last change of basis as:
\[
\bar{Z}_n := \begin{pmatrix} I_d \\ -\mathcal{E}_{n,D^*} \end{pmatrix} Z_n =: \begin{pmatrix} \bar{Z}_n^{(1)} \\ \bar{Z}_n^{(2)} \end{pmatrix};
\]
For the sequence \((\bar{Z}_n)_{n \geq 1}\), the following proposition holds:

**Proposition 7.** Assume \((H_S)\) and \((\hat{\theta}_n)_{n \geq 1}\) is a \((L^4, \sqrt{n})\)-consistent:

- i) A constant \(c_4\) exists such that: \(\forall n \geq 1\)
  \[
  \mathbb{E} \left| \bar{Z}_n^{(1)} \right|^2 \leq c_4 \gamma_n^2.
  \]
\( \bullet \) ii) Let \( \beta \in [1/2, 1] \), \( r_\beta = ((\beta + 1/2) \land (2 - \beta)) > 1 \) and \( n_1 \) be the positive integer defined in Lemma 13. For any \( n \geq n_1 \):

\[
\mathbb{E} \left| \tilde{Z}^{(2)}_n \right|^2 \leq \frac{\text{Tr}(\Sigma^*)}{n} + C_\beta(c_4, f, S) \left( \frac{\sqrt{d}}{\mu} \right)^3 \mathcal{O}_{\text{id}}(n^{-r_\beta}).
\]

where \( C_\beta(c_4, f, S) = c_4(1 - \beta)^{-1}C_{f, S} \) with \( C_{f, S} = \|D^2f\|_{\text{Lip}} + \|S\|_{\text{Lip}} \).

We are driven to the “optimal” choice \( \beta = 3/4 \), which in turns implies that:

\[
\forall n \in \mathbb{N}^* \quad \mathbb{E} \left| \tilde{Z}^{(2)}_n \right|^2 \leq \frac{\text{Tr}(\Sigma^*)}{n} + c_4 S_{f, S} \left( \frac{\sqrt{d}}{\mu} \right)^3 \mathcal{O}_{\text{id}}(n^{-5/4}).
\]

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

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

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

\[
\tilde{Z}_{n+1} = \begin{pmatrix}
I_d & 0 \\
-E_{n+1, D^*} & I_d
\end{pmatrix} \tilde{Z}_n
\]

\[
= \begin{pmatrix}
I_d & 0 \\
-E_{n+1, D^*} & I_d
\end{pmatrix} \begin{pmatrix}
A_n \tilde{Z}_n + \gamma_{n+1} \left( \frac{Q\Delta M_{n+1}}{\gamma_{n+1}} \right) + \tilde{v}_n \\
0
\end{pmatrix}
\]

\[
= \begin{pmatrix}
I_d & 0 \\
-E_{n+1, D^*} & I_d
\end{pmatrix} \begin{pmatrix}
I_d - \gamma_{n+1}D^* & 0 \\
0 & I_d - \frac{1}{n+1}I_d
\end{pmatrix} \tilde{Z}_n
\]

\[
+ \gamma_{n+1} \left( \frac{Q\Delta M_{n+1}}{\gamma_{n+1}} + \frac{1}{n+1}Q\Delta \right) + \left( (E_{n+1, D^*} - \frac{1}{n+1}I_d) \tilde{Z}_n + \gamma_{n+1} \right)
\]

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

\[
\|\Lambda^* - \Lambda_n\| \leq \frac{1}{n} \int_0^1 \|D^2f(t\theta_n) - D^2f(0)\| \, dt \leq \frac{1}{2} \|D^2f\|_{\text{Lip}} |\theta_n|. 
\]

Then, we deduce that:

\[
\begin{align*}
\tilde{Z}^{(1)}_{n+1} &= (I_d - \gamma_{n+1}D^*) \tilde{Z}^{(1)}_n + \gamma_{n+1} \left( \frac{Q\Delta M_{n+1}}{\gamma_{n+1}} + \mathcal{O}_{\text{id}}(\|D^2f\|_{\text{Lip}} |\theta_n|^2) \right) \\
\tilde{Z}^{(2)}_{n+1} &= (I_d - \gamma_{n+1}D^*) \tilde{Z}^{(2)}_n + \Omega_n \tilde{Z}^{(1)}_n + \gamma_{n+1} Y_n \left( \frac{Q\Delta M_{n+1}}{\gamma_{n+1}} + \mathcal{O}_{\text{id}}(\|D^2f\|_{\text{Lip}} |\theta_n|^2) \right),
\end{align*}
\]

with \( \Omega_n = (E_{n, D^*} - E_{n+1, D^*})(I_d - \gamma_{n+1}D^*) \) and \( Y_n = E_{n+1, D^*} - \frac{I_d}{n+1} \).

**Step 2:** \( \mathbb{E}[|\tilde{Z}^{(2)}_n|^2] = \mathcal{O}_{\text{id}}(n^{-1}) \) We introduce the covariation:

\[
\forall i \in \{1, \ldots, d\} \quad \omega_n(i) = \mathbb{E}[\tilde{Z}_n(i)\tilde{Z}_n(i+1)] = \mathbb{E}[(\tilde{Z}^{(1)}_n)_i(\tilde{Z}^{(2)}_n)_{i+1}],
\]

and the useful coefficient:

\[
\forall i \in \{1, \ldots, d\} \quad \alpha_{n}^i = 2 \left( 1 - \frac{1}{n+1} \right) \{\Omega_n\}_{i,i}.
\]

We use the Young inequality \( ab \leq \frac{\alpha}{2}a^2 + \frac{\beta}{2}b^2 \) with \( \epsilon = n^{\beta - \frac{1}{2}} \).

\[
\mathbb{E}[|\tilde{Z}^{(2)}_n|^2] \leq \frac{n^{\beta - \frac{1}{2}}}{2n} \mathbb{E}[|\theta_n|^4] + \frac{n^{\beta + \frac{1}{2}}}{2n} \mathbb{E}[|\tilde{Z}^{(2)}_n|^2] \leq C_4 \gamma_n \left( n^{-\frac{1}{2}} + n^{-\frac{1}{2}} \mathbb{E}[|\tilde{Z}^{(2)}_n|^2] \right).
\]
Second, Lemma 13 implies that $i \in \{1, \ldots, d\}$: $|\alpha_n| \leq \frac{1}{\mu n} |\tilde{Z}_n(1)|^2 + \frac{1}{\mu n} |\tilde{Z}_n(1)|^2 \leq \frac{c_n^2 n^{\frac{1}{2}}}{\mu^2} + n^{\frac{1}{2}} \mathbb{E} |\tilde{Z}_n(1)|^2$. Hence
\[
\sum_{i=1}^d |\alpha_n^i \omega_n(i)| \leq \frac{1}{\mu n} |\tilde{Z}_n(1)|^2 \leq \frac{c_n^2 n^{\frac{1}{2}}}{\mu^2} + n^{\frac{1}{2}} \mathbb{E} |\tilde{Z}_n(1)|^2.
\]
We use this inequality into Lemma 13 ii), an integer $n_1$ exists (see Lemma 13):
\[
\forall n \geq n_1 \quad \mathbb{E}[|\tilde{Z}_{n+1}^{(2)}|^2] \leq \left( 1 - \frac{1}{n+1} \right)^2 + \left[ n^{-\frac{1}{2}} + n^\frac{1}{2} \right] \mathbb{E}[|\tilde{Z}_n^{(2)}|^2] + \frac{\operatorname{Tr} \left( \Sigma^* \right)}{n+1} + \mathcal{O}_{\text{id}} \left( \frac{c_n^2 n^{\frac{1}{2}}}{\mu^2} + \frac{c_n^2 n^{\frac{1}{2}}}{\mu^2} |\tilde{Z}_n^{(2)}|^2 \right).
\]
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 $C_2 = \mathcal{O}_{\text{id}} \left( \frac{c_n^2 C_S}{n} \right)$. Setting $N = n_1$ and $u_n = \mathbb{E}[|\tilde{Z}_n^{(2)}|^2]$, we apply Lemma 16 and deduce that:
\[
\forall n \geq N \quad \mathbb{E}[|\tilde{Z}_{n+1}^{(2)}|^2] \leq \frac{\operatorname{Tr} \left( \Sigma^* \right)}{n} + \mathcal{O}_{\text{id}} \left( \frac{u_n n_1^2}{n^2} + \operatorname{Tr} \left( \Sigma^* \right) n^{-r} + C_2 n^{-q} \right).
\]
Using the arguments of (50), $\operatorname{Tr} \left( \Sigma^* \right) \leq d \mu^{-2} \| \Sigma \|$. The definition of $C_S$ yields:
\[
\forall n \geq n_1, \quad \mathbb{E}[|\tilde{Z}_n^{(2)}|^2] \leq \frac{c_n d}{\mu^3} C_S n^{-1} + \frac{u_n n_1}{n}.
\]
Remark that Lemma 15 entails $u_n \leq \frac{c_n^2 n^{\frac{1}{2}}}{\mu^2}$ so that:
\[
\frac{u_n n_1}{n} \leq \frac{c_n^2 n^{\frac{1}{2}}}{\mu^2} \leq \frac{c_n^2 n^{\frac{1}{2}}}{\mu^2} \left( n_0^{1-\beta} + \frac{c_2}{c_0} \right) \leq \frac{c_n^2 n^{\frac{1}{2}}}{\mu^2} \leq \frac{c_n d}{\mu^3} C_S n^{-1}.
\]
using in particular that $c_2 \leq \sqrt{c_4}$. As a conclusion, we finally get:
\[
\forall n \geq n_1, \quad \mathbb{E}[|\tilde{Z}_n^{(2)}|^2] \leq \frac{c_n d}{\mu^3} C_S n^{-1}.
\]
\textbf{Step 3: Control of the covariance} Inequality (30) yields for $n \geq n_1$:
\[
\mathbb{E}[|\tilde{Z}_n^{(2)}|^2] \leq \sqrt{\mathbb{E}[|\tilde{Z}_n^{(2)}]^2] \mathbb{E}[|\tilde{Z}_n^{(2)}|^2] = c_4 \sqrt{\left( \frac{d \mu^3}{\mu - \beta} \right) \mathbb{E}[\tilde{Z}_n^{(2)}]} \mathcal{O}_{\text{id}} \left( \frac{\gamma_n}{\sqrt{n}} \right).
\]
Plugging this control into Lemma 13 i), we obtain that for all $i \in \{1, \ldots, d\}$:
\[
|\omega_{n+1}(i) - (1 - \gamma_{n+1} \mu_0^*) \frac{n}{n+1} \omega_n(i)| \leq \frac{c_n d}{\mu^3} \mathcal{O}_{\text{id}} \left( \frac{\gamma_n}{\sqrt{n}} \right).
\]
Now, remark that $\gamma_n \leq \mu n^{-1/2}$ so that we conclude that $\mathbb{E}[|\theta_n|^2 | \hat{Z}_n^{(2)}]$ shall be neglected in $(\omega_n(i))_{n \geq 1}$. Now, set

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

We have

$$\forall n \geq 1 \quad \left| \omega_{n+1}(i) - (1 - \gamma_{n+1} \mu^*) \frac{n}{n+1} \omega_n(i) \right| \leq O_{\text{id}} \left( \frac{c_4 \sqrt{d} C_{f,S} \gamma_n}{\mu^{3/2}/n} \right).$$

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

$$\forall i \in \{1, \ldots, d\} \quad |\omega_n(i)| \leq O_{\text{id}} \left( \frac{c_4 \sqrt{d} C_{f,S} \gamma_n}{\mu^{3/2}/n} \right).$$  \hfill (32)

**Step 4: Conclusion of the proof**  

From (32) and (31), we have:

$$\sum_{i=1}^d \alpha^i_n \omega_n(i) = \frac{c_4 d^{3/2} C_{f,S}}{\mu^{3/2}} O_{\text{id}} \left( \frac{1}{\mu^3 \gamma_n} \right) \text{ and }\n
$$

$$\mathbb{E}[|\theta_n|^2 | \hat{Z}_n^{(2)}] = \frac{c_4 \sqrt{d}}{\mu^{3/2} C_{f,S}} O_{\text{id}} \left( \frac{\gamma_n}{\mu^{3/2}} \right).$$

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

$$\mathbb{E}[|\hat{Z}_{n+1}^{(2)}|^2] \leq \left(1 - \frac{1}{n+1}\right)^2 \mathbb{E}[|\hat{Z}_n^{(2)}|^2] + \frac{\text{Tr}(\Sigma^*)}{(n+1)^2} \left( n^{-3+\beta} \vee n^{-3/2-\beta} \right),$$

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. \hfill \Box

### 3.2 End of the proof of Theorem 3

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

$$|\hat{\theta}_n|^2 \leq 2 \left( |\hat{Z}_n^{(2)}|^2 + \rho \langle \mathcal{E}_n, D \nu \rangle^2 |\hat{Z}_n^{(1)}|^2 \right),$$

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

$$\mathbb{E}[|\hat{\theta}_n|^2] \leq \frac{c_2 n^{-\beta}}{1 - \beta} \leq \frac{c_2}{1 - \beta} R_n^{\beta-\beta} n^{-r_\beta}.$$  

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

$$n_1 \leq \mu^{-\frac{1}{1-\beta}} + \frac{c_2}{\mu^{1/2}} + \|D^2f\|_{\mathcal{X}},$$

we get

$$c_2 n_1^{r_\beta-\beta} \leq \mu^{-\frac{1}{1-\beta}} + \frac{c_2}{\mu^{1/2}} + \|D^2f\|_{\mathcal{X}}.$$

Up to a universal constant, this last upper bound is smaller than $\frac{c_4 d^{3/2} C_{f,S}}{(1-\beta)\mu^3}$. \hfill \Box
3.3 Further remarks on the second order term

When \( x \mapsto D^2 f(x) \) is constant (or also when the function \( f \) to minimize is \( C^3 \) with third partial derivatives Lipschitz and null at \( \theta^* \)), we remark that \( \Lambda_n - \Lambda^* = O(n^{-1}|\theta|^2) \). Following the proof of Lemma 13, the error term is replaced by \( n^{-1}O_{id}(E[|\theta| |Z_n(2)|]) \leq n^{-1}O(n^{-1}|\theta|^2) \) if the \((L^p, \sqrt{n})\)-consistency holds. Hence, we obtain:

\[
E[|Z_{n+1}(2)|^2] \leq \left(1 - \frac{1}{n+1}\right)^2 E[|Z_n(2)|^2] + O(n^{-3}|\theta|^2) + O\left(\frac{\sqrt{n}}{n^2}\right),
\]

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 \( \beta = \frac{2}{3} \) and \( r_\beta = \frac{1}{4} \).

The previous remark shows that we may obtain a different size of the second order terms when \( f \) is locally symmetric around \( \theta^* \) (which occurs when \( D^3 f(\theta^*) = 0 \)) 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) \( n \rightarrow n^p \left(E[|\hat{\theta}_n - \theta^*|^2] - \frac{Tr(\Sigma^*)}{n}\right) \) with \( \rho = \frac{5}{4} \) and \( \beta = \frac{3}{4} \) for a locally non-symmetric \( f_1 \) around \( \theta^* \) and \( n \rightarrow n^p \left(E[|\hat{\theta}_n - \theta^*|^2] - \frac{Tr(\Sigma^*)}{n}\right) \) 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 \rightarrow n^p \left(E[|\hat{\theta}_n - \theta^*|^2] - \frac{Tr(\Sigma^*)}{n}\right)](image)

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 \((L^p, \sqrt{n})\)-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 \( V_p : \mathbb{R}^d \rightarrow \mathbb{R} \) defined as:

\[
\forall p \geq 1 \quad V_p(x) = f^p(x) \exp(\phi(f(x))).
\]

4.1 Taylor’s expansion on \( V_p \)

To prove our main result (Theorem 11), we need to establish some technical results related to \( \phi \) and \( V_p \). The first result is a simple sub-additive property on \( \phi \) that essentially relies on the concavity property on \([x_0, +\infty)\).
Lemma 7. If $\phi$ satisfies (H$_\phi$)(i), then a constant $c_\phi$ exists such that:

$$\forall (x, y) \in \mathbb{R}^d \quad \phi(x + y) \leq \phi(x) + \phi(y) + c_\phi.$$ 

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

$$\forall x \geq x_0 \quad \phi(x + y) \leq \phi(x) + \phi(x_0 + y) - \phi(x_0).$$

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

$$\forall x \geq x_0, \forall y \geq 0 \quad \phi(x + y) \leq \phi(x) + \phi(y) + Cx_0 - \phi(x_0).$$

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

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

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 $p \in \mathbb{N}^*$ and any $x \in \mathbb{R}^d \setminus \{\theta^*\}$, 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 $\psi_1$ and $\psi_2$ are given by:

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

and $\psi_2(x) := \frac{p}{f(x)} + \phi'(f(x))$.

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 (H$_\phi$): $0 < m \leq \phi'(f)|\nabla f|^2 + \frac{|
abla f|^2}{f} \leq M$, then

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

ii) $$\forall \xi \in \mathbb{R}^d \quad \rho(D^2 V_p(\xi)) \leq (1 + M^2 + \rho_\infty(f)) \left( V_{p-1}(\xi) + \frac{V_p(\xi)}{1 + |
abla f(\xi)|^2} \right).$$

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

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

The result then follows from Assumption (H$_\phi$) ii).

ii) We apply Lemma 8 ii). We have that $\forall y \in \mathbb{R}^d.$
\[
\begin{align*}
\langle y, D^2V_p(\xi)y \rangle \|y\|^2 & = \frac{V_p(\xi)}{\|y\|^2} \left[ \psi_1(\xi)\langle y, \nabla f(\xi) \otimes \nabla f(\xi)y \rangle + \psi_2(\xi)\langle y, D^2 f(\xi)y \rangle \right] \\
& \leq V_p(\xi) \left( \frac{2p^2}{f^2(\xi)} + 2\phi'(f(\xi)) + \phi''(f(\xi)) \right) \|\nabla f(\xi)\|^2 + \left[ \frac{p}{f(\xi)} + \phi'(f(\xi)) \right] \rho_\xi(f) \right).
\end{align*}
\]

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

\[
\frac{\|\nabla f(\xi)\|^2}{f^2(\xi)} \leq \frac{M}{f(\xi)} \quad \text{and} \quad \phi'(f(\xi))^2 \|\nabla f(\xi)\|^2 \leq M\phi'(f(\xi)).
\]

Using the definition of \(M\), we deduce that:

\[
\begin{align*}
\frac{\langle y, D^2V_p(\xi)y \rangle}{\|y\|^2} & \leq_{id} M V_p(\xi) \left[ \frac{1}{f(\xi)} + \phi'(f(\xi)) + \phi''(f(\xi)) + M \rho_\xi(f) \right] V_p(\xi) \\
& \leq_{id} (M + \rho_\xi(f)) V_{p-1}(\xi) + M V_p(\xi) \left( \phi'(f(\xi)) + \phi''(f(\xi)) + M \phi'(f(\xi)) \right).
\end{align*}
\]

\(H\(_\phi\)\) implies that \(\phi''\) is negative for \(u \geq x_0\) so that \(\phi'\) is bounded (it is a non-negative function and non-increasing on \([x_0, +\infty)\)). Now, \(H\(_\phi\)\) yields

\[
\sup_{x \in \mathbb{R}^d} \left( \phi'(f(\xi)) + \phi''(f(\xi)) \right) (1 + \|f(\xi)\|^2) \\
\leq (\|\phi'\|_\infty + M + \sup_{x \in [0, x_0]} \phi''(x)(x_0(1 + Mx_0)) \\
\leq_{id} 1 + M.
\]

We then deduce that

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

The second assertion follows. \(\square\)

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

i) Assume that \(\delta > 0\) is such that \(\rho_\xi(f)\delta \leq 1/2\), then:

\[
f(x_{\delta, \varepsilon, x, t}) \leq f(x) + \delta |\varepsilon|^2.
\]

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

\[
D^2V_p(\xi)_{\delta, \varepsilon, x, t}(-\nabla f(x) + \varepsilon)^{\otimes 2} \leq q_p(\phi) V_{p-1}(x + \delta^{-1}),
\]

with \(q_p(\phi) = O_{id}(1 + M^3 + \rho_\xi(f))\).

iii) When \(\phi = 0\), if \(\rho_\xi(f)\delta \leq 1/2\), then a \(q_p(\phi)\) exists such that for all \(x \in \mathbb{R}^d:\)

\[
D^2f^p(\xi)_{\delta, \varepsilon, x, t}(-\nabla f(x) + \varepsilon)^{\otimes 2} \leq q_p(\phi)(1 + \Sigma_p)^{-1} \times V_{p-1}(x + \delta^{-1}),
\]

Furthermore, \(q_p(\phi) = O_{id}(1 + M^2 + \rho_\xi(f))(1 + \Sigma_p)\).
We consider Lemma 7 and Inequality \( T \phi \).

The elementary inequality

\[
\langle u, v \rangle \leq \frac{1}{2} \left( \| u \|^2 + \| v \|^2 \right).
\]

where in the last line we use that \( |\varepsilon| \leq 1 \) and the condition \( \delta \rho_x(f) \leq 1/2 \).

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

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

\[
V_r(\xi) \leq (f(x) + \delta \| \varepsilon \|^2)^r \exp \left( \phi(f(x) + \delta \| \varepsilon \|^2) \right).
\]

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

\[
V_r(\xi) \leq 2^{-1} \left( f^r(x) + \delta^r \| \varepsilon \|^2 \right) e^{\phi(f(x) + \delta \| \varepsilon \|^2)} + c_0.
\]

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

\[
\forall r \geq 0 \quad C_r > 0 \quad V_r(\xi) \leq \exp(\phi(\| \varepsilon \|^2)) \left[ V_r(x) + \delta^r \| \varepsilon \|^2 V_0(x) \right]
\]

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

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

\[
\rho(D^2 V_p(\xi),|\nabla f(x)|^2) \leq q_1 \left( T_{\varepsilon,\delta,p-1}[V_p(\xi) + \frac{1}{1 + \| \nabla f(x) \|^2}] \right) \frac{\| \nabla f(x) \|^2}{1 + \| \nabla f(x) \|^2}.
\]

Under Assumption \( (H_\delta) \), \( V_{p-1} \| \nabla f(x) \|^2 \leq M V_p(x) \) and \( \delta^p-1 \| \nabla f(x) \|^2 \leq M \delta^p-1 f(x) \leq M \delta^p-1 (1 + V_p(x)) \). Inequality \( T_{\varepsilon,\delta,p-1} \leq 2 T_{\varepsilon,\delta,p} \) leads to:

\[
\| D^2 V_p(\xi),|\nabla f(x)|^2 \|
\]

\[
\leq q_1 M T_{\varepsilon,\delta,p-1}[V_p(x) + \delta^p-1] + q_1 T_{\varepsilon,\delta,p}[V_p(x) + \delta^p] \frac{\| \nabla f(x) \|^2}{1 + \| \nabla f(x) \|^2}
\]

where

\[
(1 + M^3 + \rho_x(f)) T_{\varepsilon,\delta,p} \left[ V_p(x) + \delta^p-1 + \frac{[V_p(x) + \delta^p] \| \nabla f(x) \|^2}{1 + \| \nabla f(x) \|^2} \right].
\]
To handle Equation (36), we are driven to derive an upper bound of $\frac{|\nabla f(x)|^2}{1+|\nabla f(x)|^2}$. According to the Taylor formula, a $\xi'$ exists in $[x, \xi]$ such that:

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

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

$$|\nabla f(x)| \leq |\nabla f(\xi)| + \rho_{\varepsilon}(f) \delta (|\nabla f(x)| + \|\varepsilon\|).$$

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

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

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

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

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

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

Finally, since $T_{\varepsilon, \delta} \leq 3T_{\varepsilon, \delta, \delta + 1}$, we conclude that:

$$\rho(V_p(\xi)) |\nabla f(x)|^2 \leq_{id} (1 + M^3 + \rho_{\varepsilon}(f)) T_{\varepsilon, \delta, \delta + 1} \delta^{p-1} + V_p(x),$$

(37)

• Step 3: Upper bound of $\|D^2 V_p(\xi_{\varepsilon, \varepsilon, \varepsilon, \varepsilon})\|_{\varepsilon, \varepsilon, \varepsilon, \varepsilon}$, (35) and Lemma 9 ii) yield:

$$\rho(D^2 V_p(\xi_{\varepsilon, \varepsilon, \varepsilon, \varepsilon})) \cdot \varepsilon^2 \leq_{id} T_{\varepsilon, \delta, \delta + 1} (V_{p-1}(x) + V_p(x) + \delta^{p-1}).$$

(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 $(H_\phi)$ in this case reads $m f(x) \leq |\nabla f(x)|^2 \leq M f(x)$, we deduce that:

$$\rho(D^2 f_p(\xi)) \leq (M + \rho_{\varepsilon}(f)) f_p^{p-1}(\xi).$$

(39)

Using Lemma 10 i), we have, if $\delta \rho_{\varepsilon}(f) \leq 1/2$:

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

so that:

$$\rho(D^2 f_p(\xi)) \left(|\nabla f(x)|^2 + |\varepsilon|^2\right) \leq_{id} (M + \rho_{\varepsilon}(f)) \left(M f_p(x) + f_p^{p-1}(x) |\varepsilon|^2 + \delta^{p-1} M f(x) |\varepsilon|^{2(p-1)} + |\varepsilon|^{2p}\right).$$

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

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

**Theorem 11.** Let \(p \geq 1\), if \((H_\phi)\) and \((H_\phi^p)\) when \(\phi \neq 0\) (or \((H_\phi)\) and \((H_\phi^p)\) if \(\phi = 0\)) hold. Consider \(q_p(\phi)\) from Lemma 10 and:

\[
n_0 := \inf \left\{ n : \gamma_{n+1} q_p(\phi) \leq \frac{m}{4} \text{ and } \forall q \leq p : \left( \frac{\gamma_q^2 - \gamma_{n+1}^2}{m\gamma_{n+1}^2} \right) \leq \frac{1}{8} \right\}.
\]

i) For all \(n \geq n_0\),

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

(40)

ii) \((\tilde{C}_p)_{p \geq 1}\) exists such that for all \(n \geq n_0\), \(\mathbb{E}[V_p(\theta_n)] \leq \tilde{C}_p(\gamma_n)^p\) with

\[
\tilde{C}_1 = \mathbb{E}[V_1(\theta_n)] \frac{q_1(\phi)}{m} + \mathcal{O}_{\text{id}} \left( \frac{q_1(\phi)}{m} \right)
\]

and for every integer \(p \geq 2\),

\[
\tilde{C}_p = \mathbb{E}[V_p(\theta_n)] \left( \frac{q_p(\phi)}{m} \right)^p + \mathcal{O}_{\text{id}} \left( (\tilde{C}_{p-1} + 1) \frac{q_p(\phi)}{m} \right).
\]

(41)

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

\[
V_p(\theta_{n+1}) = V_p(\theta_n) - \gamma_{n+1} \langle \nabla V_p(\theta_n), \nabla f(\theta_n) \rangle + \gamma_{n+1} \langle V_p(\theta_n), \Delta M_{n+1} \rangle \\
+ \frac{\gamma_{n+1}^2}{2} D^2 V_p(\xi_{n+1})(-\nabla f(\theta_n) + \Delta M_{n+1})^2;
\]

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

\[
\forall n \in \mathbb{N}^* \quad V_p(\theta_n) - \gamma_{n+1} \langle \nabla V_p(\theta_n), \nabla f(\theta_n) \rangle \leq V_p(\theta_n)(1 - m\gamma_{n+1}).
\]  

(42)

Now, we need to consider separately the cases \(\phi = 0\) and \(\phi \neq 0\).

• Case \(\phi = 0\): Since \(\gamma_n \rho_x(f) \leq \gamma_q q_p(\phi) \leq 1/2\) for all \(n \geq n_0\), \((H_\phi^p)\) yields

\[
\mathbb{E}[\{1 + (\gamma M_{n+1})^2(\gamma M_{n+1})^p \} \exp(\phi(\gamma M_{n+1})^2)] \leq \Sigma_p.
\]

Thus, we deduce from Lemma 10 ii) that for every \(n \geq n_1\)

\[
\mathbb{E}[V_p(\theta_{n+1}) | \mathcal{F}_n] \\
\leq (1 - m\gamma_{n+1}) V_p(\theta_n) + q_p(\phi) \left( \gamma_{n+1}^2 V_p(\theta_n) + V_{p-1}(\theta_n) (\gamma_{n+1})^{p+1} \right).
\]

This yields

\[
\mathbb{E}[V_p(\theta_{n+1}) | \mathcal{F}_n] \\
\leq (1 - m\gamma_{n+1} + q_p(\phi) \gamma_{n+1}^2) V_p(\theta_n) + q_p(\phi)(\gamma_{n+1}^2 V_{p-1}(\theta_n) + (\gamma_{n+1})^{p+1}).
\]  

(43)

The result follows since \(n \geq n_0\) so that \(1 - m\gamma_{n+1} + q_p(\phi) \gamma_{n+1}^2 \leq 1 - \frac{3m}{4} \gamma_{n+1}\).
• Case \( \phi = 0 \): By Lemma 10 iii) and Assumption \((H_{\Sigma p}^{SC})\), we have for all \( n \geq n_0 \),
\[
\mathbb{E} \left[ D^2 f^p(\xi_{n+1}) (-\nabla f(\theta_n) + \Delta M_{n+1})^{\otimes 2} | \mathcal{F}_n \right] \\
\leq q_p(\phi)(1 + \Sigma_p)^{-1} \left( f^p(\theta_n) + f^{p-1}(\theta_n) \Sigma_1 (1 + f(\theta_n)) + \gamma_{n+1}^{p-1} \Sigma_p (1 + f^p(\theta_n)) \right) \\
\leq q_p(\phi) \frac{1 + \Sigma_1 + \Sigma_p}{1 + \Sigma_p} (f^p(\theta_n) + f^{p-1}(\theta_n)) + \gamma_{n+1}^{p+1})
\]

Since \( |x|^2 \leq 1 + |x|^{2p} \) for any \( p \geq 1 \), \( \Sigma_1 \leq 1 + \Sigma_p \). 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
\[
\gamma_{n+1}^2 \mathbb{E} \left[ D^2 f^p(\xi_{n+1}) (-\nabla f(\theta_n) + \Delta M_{n+1})^{\otimes 2} | \mathcal{F}_n \right] \\
\leq q_p(\phi) \left( \gamma_{n+1}^2 (f^p(\theta_n) + f^{p-1}(\theta_n)) + \gamma_{n+1}^{p+1} \right).
\]

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 first consider the case where \( p = 1 \), we use the elementary inequality \( V_0 \leq 1 + V_1 \) and obtain that
\[
\forall n \geq n_0, \quad \mathbb{E}[V_1(\theta_{n+1})] \leq (1 - \frac{3m}{4} \gamma_{n+1} + q_1(\phi) \gamma_{n+1}^2) \mathbb{E}[V_1(\theta_n)] + 2q_1(\phi) \gamma_{n+1}^2.
\]

According to our choice on \( n \), we deduce that
\[
\forall n \geq n_0, \quad \mathbb{E}[V_1(\theta_{n+1})] \leq (1 - \frac{m}{2} \gamma_{n+1}) \mathbb{E}[V_1(\theta_n)] + 2q_1(\phi) \gamma_{n+1}^2
\]

Set \( v_n = \gamma_n^{-1} \mathbb{E}[V_1(\theta_n)] \). We obtain
\[
\forall n \geq n_0, \quad v_{n+1} \leq (1 - \frac{m}{2} \gamma_{n+1}) v_n \left( 1 + \frac{m}{4} \gamma_{n+1} \right) + q_1(\phi) \gamma_{n+1}.
\]

According to the construction of \( n_0 \), we can check that for all \( n \geq n_0 \),
\[
\left( \frac{\gamma_n}{\gamma_{n+1}} \right)^p \leq 1 + \frac{m}{4} \gamma_{n+1}.
\]

We then obtain
\[
\forall n \geq n_0, \quad v_{n+1} \leq \left( 1 - \frac{m}{2} \gamma_{n+1} \right) v_n \left( 1 + \frac{m}{4} \gamma_{n+1} \right) + q_1(\phi) \gamma_{n+1}. \quad (44)
\]

Since \( (1 - \frac{m}{2} \gamma_{n+1})(1 + \frac{m}{4} \gamma_{n+1}) \leq 1 - \frac{m}{4} \gamma_{n+1} \), we deduce that for all \( n \geq n_0 \),
\[
v_n \leq v_{n_0} \prod_{k=n_0+1}^n \left( 1 - \frac{m}{4} \gamma_k \right) + q_1(\phi) \sum_{k=n_0+1}^n \gamma_k \prod_{\ell=k}^{n-1} \left( 1 - \frac{m}{4} \gamma_{\ell} \right)
\]

with the convention \( \prod_{\emptyset} = 1 \). By the elementary inequality \( \log(1 + x) \leq x \) for \( x > -1 \), this yields
\[
v_n \leq v_{n_0} e^{-\frac{m}{4} (\Gamma_n - \Gamma_{k-1})} + q_1(\phi) \sum_{k=n_0+1}^n \gamma_k e^{-\frac{m}{4} (\Gamma_{n-1} - \Gamma_{k-1})}.
\]

On the one hand, \( e^{-\frac{m}{4} (\Gamma_n - \Gamma_{k-1})} \leq 1 \), on the other hand, a series/integral comparison yields
\[
\sum_{k=n_0+1}^n \gamma_k e^{-\frac{m}{4} (\Gamma_{n-1} - \Gamma_{k-1})} \leq e^{-\frac{m}{4} \Gamma_{n-1}} \int_0^{\Gamma_n} e^{\frac{m}{4} x} dx \\
\leq \frac{4}{m} e^{-\frac{m}{4} \Gamma_{n-1}} (e^{-\frac{m}{4} \Gamma_n} - 1) \leq \frac{4e^{\frac{m}{4} n_0}}{m} \leq \frac{5}{m}.
\]

24
where we used in the last equality that $m \gamma_n \leq \gamma_n q_1(\phi) \leq 1/2$.

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

$$\forall n \geq n_0, \quad \mathbb{E}[V_p(\theta_{n+1})] \leq \left(1 - \frac{3m}{4} \gamma_{n+1}\right) \mathbb{E}[V_p(\theta_n)] + q_p(\phi) \left(C_{p-1} \gamma_n^{p} + \gamma_n^{p+1}\right).$$

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

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

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

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

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

Acknowledgements The authors gratefully acknowledge Jérôme Bolte and Gersende Fort for stimulating discussions on the Kurdyka-Lojasiewicz 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


Appendix A: Technical lemmas for Theorem 2

In the next lemma, we study some properties of \((\epsilon_{\mu,n})_{n \geq 1}\) involved in the change of basis related to the evolution of \((\hat{\theta}_n)_{n \geq 1}\) (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 \(\underline{\mu} \leq 1\), \(\gamma = 1\) et \(c_2 \geq 1\), \(\|D^2 f\|_{\text{Lip}} \geq 1\).

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

\[
\forall n \geq n_0 := \left[ \left( \frac{2}{\mu} \right)^{1/(1-\beta)} \right], \quad |\epsilon_{\mu,n} - \epsilon_{\mu,n+1}| \leq \frac{1}{\mu} n^{\beta - 2}\text{,} \tag{45}
\]

and \(|\epsilon_{\mu,n}| \leq \frac{1}{\mu} (\mu \gamma_n n)^{-1}\).
Proof: First, remark that for \( n \geq n_0 \), \( \mu \gamma_n n - 1 \geq \frac{1}{2} \mu \gamma_n n \), so that \( \epsilon_{\mu,n} \) is well-defined for any \( n \geq n_0 \) and,

\[
|\epsilon_{\mu,n}| \leq \frac{2}{\mu \gamma_n n} + \frac{2 \mu \gamma_n}{\mu \gamma_n n} \leq \frac{2}{n} \left( \frac{1}{\mu \gamma_n} + 1 \right) \lesssim_{\text{id}} \frac{1}{\mu \gamma_n n},
\]
since \( \mu \gamma_n \leq 1 \) for every \( n \geq 1 \). As concerns (45), we observe that for \( n \geq n_0 \),

\[
|\epsilon_{\mu,n} - \epsilon_{\mu,n+1}| = \left| \frac{1 - \mu \gamma_n}{1 - \mu \gamma_n n} - \frac{1 - \mu \gamma_{n+1}}{1 - \mu \gamma_{n+1} (n + 1)} \right| = \left| \frac{(1 - \mu \gamma_n)(1 - \mu \gamma_{n+1} (n + 1)) - (1 - \mu \gamma_{n+1})(1 - \mu \gamma_n)}{(1 - \mu \gamma_n)(1 - \mu \gamma_{n+1} (n + 1))} \right| \
\leq \mu \frac{\gamma_n - \gamma_{n+1} + |(n + 1)\gamma_{n+1} - n\gamma_n| + \mu \gamma_{n+1} \gamma_{n+1} - 1}{(\mu \gamma_n n - 1)(\mu \gamma_{n+1} (n + 1) - 1)} \
\leq \mu (2 + \mu) \frac{n^{-\beta}}{n \mu \gamma_n (n + 1) \mu \gamma_{n+1}} \lesssim_{\text{id}} \left( 1 + \frac{1}{\mu} \right) n^{2 - \beta},
\]
which yields the result since \( \mu \leq 1 \).

\( \square \)

**Lemma 13.** Set \( n_1 := n_0 \vee \lceil 1 / 2 \rceil \vee \lceil |D^2 f|_{\text{Lip}} \rceil \). Under the assumptions of Proposition 7, we have:

i) For any \( i \in \{1, \ldots, d\} \), \( \omega_n(i) = \mathbb{E}((\tilde{Z}_n^{(1)}\iota(z_n^{(2)}))_i) \) satisfies \( \forall n \geq n_1 \),

\[
\left| \omega_{n+1}(i) - (1 - \gamma_{n+1} \epsilon_{\mu,n}^i) \left( 1 - \frac{1}{n + 1} \right) \omega_n(i) \right| \lesssim_{\text{id}} c \omega_n \frac{\gamma_n}{n} + \gamma_{n+1} \|D^2 f\|_{\text{Lip}} \mathbb{E}[\theta_n^2 \tilde{Z}_n^{(2)}]].
\]

where \( C_\omega = \frac{d(\|S^*\| + |S|_{\text{Lip}}) + c_4 D^2 f_{\text{Lip}}}{\mu \beta} \).

ii) Set

\[
|\Delta Z_{n+1}^{(2)}| := \left| \mathbb{E}[\tilde{Z}_n^{(2)}] - \left( 1 - \frac{1}{n + 1} \right)^2 \mathbb{E}[\tilde{Z}_n^{(2)}]^2 - \sum_{i=1}^d \alpha_n^i \omega_n(i) - \text{Tr}(\Sigma) \right|.
\]

We have:

\[
|\Delta Z_{n+1}^{(2)}| \lesssim_{\text{id}} n^{-\beta/2} \mathbb{E}[|\tilde{Z}_n^{(2)}|^2] + \frac{c_4 d}{\mu \beta} C_S \alpha_n \left( n^{-\beta/2} \vee n^{-3/2} \right)
\]

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

Proof: First, remark that under the definition of \( n_1 \), we have for all \( n \geq n_1 \), \( \mu \gamma_n n - 1 \geq \frac{1}{2} \mu \gamma_n n, c_2 \gamma_n \leq 1 \) and \( |D^2 f|_{\text{Lip}} n^{-1} \leq 1 \). Then, for all \( n \geq n_0 \), \( \gamma_{n+1} \) and \( \Omega_n \) are well-defined deterministic matrices and by Lemma 12, we can verify that

\[
\gamma_{n+1} \|\Upsilon_n\| \lesssim_{\text{id}} \frac{1}{n \mu} \quad \text{and} \quad \gamma_{n+1} \|\Omega_n\| \leq \gamma_{n+1} \|\Upsilon_{n+1,i} - \Upsilon_{n+1,i,j} D^\ast\| \|I_d - \gamma_{n+1} D^\ast\| \lesssim_{\text{id}} \frac{1}{\mu n^2}.
\]

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

\[
\left| \omega_{n+1}(i) - (1 - \gamma_{n+1} \epsilon_{\mu,n}^i) \left( 1 - \frac{1}{n + 1} \right) \omega_n(i) \right| \lesssim_{\text{id}} \gamma_{n+1} \|D^2 f\|_{\text{Lip}} \mathbb{E}[\theta_n^2 \tilde{Z}_n^{(2)}]].
\]

\[
+ \gamma_{n+1}^2 \mathbb{E}[(Q|\Delta M_{n+1}|)_i \{\Upsilon_n Q|\Delta M_{n+1}|_i\} + \gamma_{n+1} r_n^{(1)}],
\]

27
where,
\[
|\tilde{\beta}^{(1)}_n| \lesssim_{id} |\Omega_n| \left( \frac{\mathbb{E}[\tilde{Z}^{(1)}_n^2]}{\gamma_{n+1}} + \mathbb{E}[\|D^2f\|_{\text{Lip}}|\theta_n|^2|\tilde{Z}^{(1)}_n|] \right) + \|Y_n\| \left( \|D^2f\|_{\text{Lip}}\mathbb{E}[|\tilde{Z}^{(1)}_n|]|\theta_n|^2 + \gamma_{n+1}\|D^2f\|_{\text{Lip}}^2\mathbb{E}[|\theta_n|^4] \right).
\]

The Cauchy-Schwarz inequality and $|\tilde{\beta}^{(1)}_n| = |\theta_n|$ yield
\[
\mathbb{E}[|\theta_n|^2|\tilde{Z}^{(1)}_n|] \leq \left\{ \mathbb{E}[|\theta_n|^4] \right\}^{1/2} \left\{ \mathbb{E}[|\tilde{Z}^{(1)}_n|^2] \right\}^{1/2} \leq \sqrt{c_4 \gamma_{n+1}^{-3/2}} \leq c_4 \gamma_{n+1}^{-3/2}.
\]

Therefore, using $1 \leq c_2 \leq \sqrt{c_4} \leq c_4$ and $\|D^2f\|_{\text{Lip}}^{-1} \leq 1$, (46) implies:
\[
\gamma_{n+1} \gamma_{n+1}^{(1)} \lesssim_{id} \frac{1}{n^2} \left( c_2 + c_4 \|D^2f\|_{\text{Lip}}^{-1} \right) \left( \frac{\gamma_{n+1}^{3/2}}{n} \right) + \frac{c_4 \|D^2f\|_{\text{Lip}}^{-1}}{n} \left( \frac{\gamma_{n+1}^{3/2}}{n} + \frac{\gamma_n^{3/2}}{n} \right) \lesssim_{id} \frac{c_4 \|D^2f\|_{\text{Lip}}^{-1}}{n^{\frac{1}{2}} \gamma_{n+1}^{3/2}} 
\]

where we used that $\beta \geq 1/2$. In the meantime, under $\textbf{(H}_{\frac{1}{2}}$ and because $Q \in O_d(\mathbb{R})$ and $c_2 \gamma_n < 1$ when $n \geq n_1$, we have $\forall i \in \{1, \ldots, d\}$,
\[
\left| \mathbb{E}[Q\Delta M_{n+1}^i \{ Y_n Q\Delta M_{n+1} \}] \right| \leq \|Y_n\| \mathbb{E}[|\Delta M_{n+1}|^2] \leq \|Y_n\| \mathbb{E}[\text{Tr}(S(\theta_n))] 
\leq \|Y_n\| (\|d\mathbb{E}[|S(\theta_n)|]\|) \leq \|Y_n\| (\|\text{Tr}(S(\theta_n))\| + \|S\|_{\text{Lip}} \mathbb{E}[|\theta_n|]) 
\lesssim_{id} d \|Y_n\| (\|S^*\| + \|S\|_{\text{Lip}}). 
\]

since $c_2 \gamma_n \leq 1$ for $n \geq n_1$. We therefore deduce from (46) and from the previous lines that
\[
\forall i \in \{1, \ldots, d\} \quad \gamma_{n+1}^2 \mathbb{E}[Q\Delta M_{n+1}^i \{ Y_n Q\Delta M_{n+1} \}] \leq \frac{d(\|S^*\| + \|S\|_{\text{Lip}}) \gamma_n}{n^{\frac{1}{2}}}. 
\]

A compilation of the previous bounds (taking into accounts only non-universal constants) leads to
\[
\left| \omega_{n+1}(i) - (1 - \gamma_{n+1} \mu_i) \left( 1 - \frac{1}{n+1} \right) \omega_n(i) \right| \lesssim_{id} \gamma_{n+1} \|D^2f\|_{\text{Lip}} \mathbb{E}[|\theta_n|^2|\tilde{Z}^{(2)}_n|] 
+ \frac{d(\|S^*\| + \|S\|_{\text{Lip}}) + c_4 \|D^2f\|_{\text{Lip}} \gamma_n}{n^{\frac{1}{2}}}. 
\]

\[i) \text{ We set } \Delta N_{n+1} = Y_n Q\Delta M_{n+1} \text{ and recall that } \alpha_n^i \text{ is defined in (29) by } \alpha_n^i = 2(1 - (n + 1)^{-1})(\Omega_n)_{i,i}. 
\]

Starting from (27) and $|\tilde{Z}^{(1)}_n| = |\theta_n|$ with a conditional expectation argument, we use that $\Omega_n$ is diagonal to obtain
\[
\mathbb{E}[\tilde{Z}^{(2)}_n] = \left( 1 - \frac{1}{n+1} \right)^2 \mathbb{E}[\tilde{Z}^{(1)}_n^2] + \sum_{i=1}^{d} \alpha_n^i \omega_n(i) + \gamma_{n+1} \mathbb{E}[\Delta N_{n+1}^1] 
+ \mathbb{E}[|\Omega_n \tilde{Z}^{(1)}_n|^2] + O_{id} \left( \gamma_{n+1} \mathbb{E}[|\theta_n|^2|\tilde{Z}^{(2)}_n|] \right)
+ O_{id}(|\gamma_{n+1} \|D^2f\|_{\text{Lip}} \mathbb{E}[|\tilde{Z}^{(1)}_n|^2]|\theta_n|^2) \|Y_n\| + O_{id}(\gamma_{n+1} \|D^2f\|_{\text{Lip}} \mathbb{E}[|\theta_n|^4]). \tag{47}
\]

First, by (46),
\[
\mathbb{E}[|\Omega_n \tilde{Z}^{(1)}_n|^2] \leq \|\Omega_n\|^2 \mathbb{E}[|\tilde{Z}^{(1)}_n|^2] \leq c_2 \gamma_n |\Omega_n|^2 \leq \frac{c_2 n^{-\beta}}{\gamma_{n+1}^3} \leq \frac{c_4}{n^{\beta}} \mathbb{E}[|\theta_n|^4]. \tag{11}
\]
In the meantime, (46) yields
\[
\gamma_{n+1} |Y_n| \mathbb{E}[|\theta_n|^2 | \tilde{Z}_{n+1}^{(2)}|] = \frac{1}{n\mu} \mathcal{O}_{id} \left( \mathbb{E}[|\theta_n|^2 | \tilde{Z}_{n+1}^{(2)}|] \right) \\
= \frac{1}{n\mu} \mathcal{O}_{id} \left( \mathbb{E}[|\theta_n|^4 v_{n-1}^2 + v_n \mathbb{E}[|\tilde{Z}_{n+1}^{(2)}|^2] / 2 \right).
\]
Choosing \( v_n = \mu n^{1/2-\beta} \), we obtain
\[
O \left( \gamma_{n+1} |Y_n| \mathbb{E}[|\theta_n|^2 | \tilde{Z}_{n+1}^{(2)}|] \right) = \frac{c_4}{\mu^2} \mathcal{O}_{id} \left( n^{-3/2-\beta} \right) + n^{-3/2-\beta} \mathbb{E}[|\tilde{Z}_{n+1}^{(2)}|^2].
\]
The \( (L^p, \sqrt{\gamma_n}) \) consistency property associated to the Cauchy-Schwarz inequality and the fact that \( c_2 \leq \sqrt{c_4} \leq c_4 \) imply that
\[
\gamma_{n+1} |Y_n| \mathbb{E}[|\theta_n|^2 | \tilde{Z}_{n+1}^{(2)}|] = \frac{c_4 |D^2 f|_{\text{Lip}}^2}{\mu^2} \mathcal{O}_{id} \left( n^{-3/2} \right) \quad (13)
\]
Finally, we also obtain that
\[
\gamma_{n+1}^2 |Y_n|^2 \mathbb{E}[|\theta_n|^4] = \frac{c_4 |D^2 f|_{\text{Lip}}^2}{\mu^2} \mathcal{O}_{id} \left( n^{-3/2} \right)
\]
where we used that \( |D^2 f|_{\text{Lip}} n^{-1} \leq 1 \).
To achieve the proof, it remains to study \( \gamma_{n+1}^2 |\Delta N_{n+1}|^2 \). First, set \( B_n = Q^T Y_n^2 Q \). Using that \( Y_n \) is a diagonal matrix, we have
\[
\gamma_{n+1}^2 |\Delta N_{n+1}|^2 = \gamma_{n+1}^2 \text{Tr}(|\Delta N_{n+1}|^2) = \gamma_{n+1}^2 \text{Tr}(\Delta N_{n+1}^T \Delta N_{n+1}) \\
= \gamma_{n+1}^2 \text{Tr}(\Delta \mathcal{M}_{n+1}^T B_n \Delta \mathcal{M}_{n+1}) \\
= \gamma_{n+1}^2 \text{Tr}(B_n \Delta \mathcal{M}_{n+1} \Delta \mathcal{M}_{n+1}^T)
\]
Since the trace is a linear application and \( B_n \) is a deterministic matrix,
\[
\gamma_{n+1}^2 \mathbb{E}[|\Delta N_{n+1}|^2 | \mathcal{F}_n] = \gamma_{n+1}^2 \text{Tr}(B_n \mathbb{E}[\Delta \mathcal{M}_{n+1} \Delta \mathcal{M}_{n+1}^T | \mathcal{F}_n]) = \gamma_{n+1}^2 \text{Tr}(B_n \mathcal{S}(\theta_n))
\]
where we applied Assumption (\( H_3 \)). For \( B_n \), we first remark that
\[-\gamma_{n+1} Y_n = (n+1)^{-1} \{ D^* \}^{-1} + \Delta_{n+1} \]
where \( (\Delta_n)_{n \geq 0} \) is a sequence of matrices defined by:
\[
\Delta_n = \text{Diag} \left\{ \frac{1+(n+1)\mu_i^*}{(n+1)\mu_i^* (n+1)\gamma_{n+1} + \gamma_{n+1}} \cdot \gamma_{n+1}, i = 1, \ldots, d \right\}.
\]
For \( n \geq n_1 \) and every \( i \in \{1, \ldots, d\} \), \( \mu_i^* \gamma_{n+1} - 1 \geq \frac{1}{2} \mu_i^* \gamma_{n+1} \) (by the beginning of the proof of Lemma 12) so that
\[
\frac{1+(n+1)\mu_i^*}{(n+1)\mu_i^* (n+1)\gamma_{n+1} + \gamma_{n+1}} \leq \frac{2}{(n+1)^2 (\mu_i^*)^2 \gamma_{n+1}} + \frac{2\gamma_{n+1}}{n+1}.
\]
Using the diagonal structure of $\Delta_n$, we get for $n \geq n_1$,
\[
\|\Delta_n\| \leq \text{id} \frac{2}{n^2 \gamma_n \mu^2} + \frac{\gamma_{n+1}}{n+1} \leq \text{id} \frac{1}{n^2 \gamma_n \mu^2},
\tag{49}
\]
since $\beta > 1/2$. Then, using that $B_n = Q^T F_n Q$ and that $Q^T (D^*)^{-2} Q = \{\Lambda^*\}^{-1}$, it follows from (48) that
\[
\gamma_{n+1}^2 \mathbb{E}[|\Delta N_{n+1}|^2] = \gamma_{n+1}^2 \mathbb{E}[\text{Tr}(B_n S(\theta_n))]
= \gamma_{n+1}^2 \mathbb{E}\left[\text{Tr}\left(Q^T (D^*)^{-1} + \Delta_{n+1}\right)^2 QS(\theta_n)\right]
= \frac{1}{(n+1)^2} \text{Tr}(\{\Lambda^*\}^{-2} QS(\theta^*)) + \sum_{i=1}^3 R_i^1 + \sum_{i=1}^3 \mathbb{E}[R_i^2]
\]
with,
\[
R_i^1 = \frac{\text{Tr}(\{\Lambda^*\}^{-2} (S(\theta_n) - S(\theta^*)))}{(n+1)^2}, \quad R_i^2 = \frac{2}{n+1} \text{Tr}(Q^T (D^*)^{-1} \Delta_{n+1} QS(\theta_n)),
\]
and $R_3^1 = \text{Tr}(Q^T (\Delta_{n+1})^2 QS(\theta_n))$.

Note that for $R_i^2$, we used that $\{D^*\}^{-1}$ and $\Delta_{n+1}$ commute. It remains to bound the remainder terms $R_i^1$, $i = 1, 2, 3$. To this end, let us denote by $\|\cdot\|_F$ the Frobenius norm defined for a square matrix $A$ by $\|A\|_F = \sqrt{\text{Tr}(A^T A)}$. Owing to the sub-multiplicativity of this norm and to the fact that $\|A\|_F \leq \sqrt{d} \|A\|$ (where $\|A\| = \sqrt{\rho(A^T A)}$), we obtain:
\[
\mathbb{E}[R_i^1] \leq \|\{\Lambda^*\}^{-2} F\|_F \|S(\theta_n) - S(\theta^*)\|_F \leq \frac{d \|S\|_{\text{Lip}} \mathbb{E}[|\theta_n|]}{\mu^2 n^2} \leq \frac{d \|S\|_{\text{Lip}} \sqrt{2 \gamma_n}}{\mu^2 n^2}, \tag{50}
\]
Using (49), one can check that
\[
\mathbb{E}[R_i^2] \leq \frac{2}{n+1} \|\{D^*\}^{-1} \Delta_{n+1}\|_F \mathbb{E}[\|S(\theta_n)\|_F] \leq \frac{2d(\|S^*\| + 2 \|S\|_{\text{Lip}})}{n^3 \gamma_n \mu^2}.
\]
Finally,
\[
\mathbb{E}[R_i^2] \leq \|\Delta_{n+1}\|_F \mathbb{E}[\|S(\theta_n)\|_F] \leq \frac{2d(\|S^*\| + 2 \|S\|_{\text{Lip}})}{n^3 \gamma_n \mu^2} \leq \text{id} \frac{d(\|S^*\| + 2 \|S\|_{\text{Lip}})}{n^3 \gamma_n \mu^2},
\]
where in the second inequality, we used that $n^{1 - \beta} \geq 2 \mu^{-1}$ for $n \geq n_0$.

A combination of the above upper bounds of $R_i^1$, $i = 1, 2, 3$ yields:
\[
\gamma_{n+1}^2 \mathbb{E}[|\Delta N_{n+1}|^2] = \frac{\text{Tr}(\Sigma^*)}{(n+1)^2} + c_4 \frac{d(\|S^*\| + 2 \|S\|_{\text{Lip}})}{\gamma_n \mu^3} \text{id} \left(n^{-(2\beta/2)} + n^{-(3-\beta)}\right).
\tag{15}
\]

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
\[
\forall \beta \in (1/2, 1) \quad n^{-3 - \beta/2} \vee n^{-3/2} = O_{id}(n^{-\frac{3}{2} - \beta}).
\]

Considering the worst constant of each term, we obtain that:
\[
|\Delta Z^2_{n+1}| = c_4 d C_{s} S_{id} \left(n^{-\frac{3}{2} - \beta} \vee n^{-3+\beta}\right) + n^{-\frac{3}{2} - \beta} \mathbb{E}|Z^2_{n+1}|.
\]

where $C_{s}$ is defined in the statement of the lemma.
Lemma 14. Assume that \((u_n)_{n \geq 0}\) is a real sequence that satisfies for all \(n \geq n_0\) and for a given \(\mu > 0\):
\[
    u_{n+1} = (1 - \gamma_{n+1}\mu) \frac{n}{n+1} u_n + \beta_{n+1},
\]
with \(\beta_n \leq \frac{\gamma_n}{n}\). Then, a constant \(C\) independent on \(\Box\) exists such that
\[
    u_n \leq \frac{\Box}{n} (n_0 u_{n_0} + \mu^{-1}).
\]

Proof: With the convention \(\prod_{\emptyset} = 1\) and \(\sum_{\emptyset} = 0\), we have for every \(n \geq n_0\):
\[
    u_n = \left( \prod_{k=n_0+1}^{n} (1 - \gamma_k \mu) \frac{k}{k+1} \right) u_{n_0} + \sum_{k=n_0+1}^{n} \beta_k \prod_{\ell=k+1}^{n} (1 - \gamma_\ell \mu) \frac{\ell}{\ell+1}.
\]
Using that for any \(x > -1\), \(\log(1+x) \leq x\), we obtain for every \(n \geq n_0 + 1\)
\[
    \prod_{k=n_0+1}^{n} (1 - \gamma_k \mu) \frac{k}{k+1} \leq \frac{n_0}{n+1} e^{-\mu (\Gamma_n - \Gamma_{n_0})} \leq \frac{n_0}{n+1},
\]
where \(\Gamma_n = \sum_{k=0}^{n} \gamma_k\). Concerning the second term, we have
\[
    \sum_{k=n_0+1}^{n} \beta_k \prod_{\ell=k+1}^{n} (1 - \gamma_\ell \mu) \frac{\ell}{\ell+1} \leq \frac{1}{n+1} \left( e^{-\mu \Gamma_n} \sum_{k=n_0+1}^{n} \beta_k (k+1) e^{\mu \Gamma_k} \right).
\]
Since \(\beta_k (k+1) \leq \Box \gamma_{k+1}\), the monotonicity of \(x \mapsto xe^{\mu x}\) yields:
\[
    \sum_{k=n_0+1}^{n} \beta_k (k+1) e^{\mu \Gamma_k} \leq \sum_{k=n_0+1}^{n} \gamma_{k+1} e^{\mu \Gamma_k} \leq \int_{\Gamma_{n_0+1}}^{\Gamma_{n+1}} e^{\mu x} \, dx.
\]
We deduce that:
\[
    \frac{1}{n+1} \left( e^{-\mu \Gamma_n} \sum_{k=n_0+1}^{n} \beta_k (k+1) e^{\mu \Gamma_k} \right) \leq \frac{\Box}{\mu(n+1)}.
\]

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 \((\theta_n)\) is \((L^2, \sqrt{n})\) consistent. Then, for all \(n \geq 0\),
\[
    \mathbb{E}|\tilde{\theta}_n|^2 \leq_{id} \frac{c_2}{1 - \beta} n^{-\beta}. \tag{51}
\]

Let \(n_1\) be defined in Lemma 13. Under the assumptions of Proposition 7:
\[
    \mathbb{E}|\tilde{Z}_{n_1}^{(2)}|^2 \leq_{id} \frac{c_2}{1 - \beta} n_1^{-\beta}. \tag{52}
\]

Proof: Under the assumption, \(\mathbb{E}|\theta_n|^2 \leq c_2 \gamma_n\). Keeping in mind that \(\theta^* = 0\), we deduce from the Jensen inequality that
\[
    \mathbb{E}\left[|\tilde{\theta}_n|^2\right] \leq \frac{c_2}{n} \sum_{k=1}^{n} \gamma_k \leq_{id} \frac{c_2}{1 - \beta} n^{-\beta}.
\]
For the second part of the proof, we use that:
\[
    \tilde{Z}_n^{(2)} = -\mathcal{E}_{n,D} \tilde{Z}_n^{(1)} + \tilde{Z}_n^{(2)}.
\]
Thus,
\[
|\hat{Z}^{(2)}_n|^2 \leq \rho (\mathcal{E}_{n,D})^2 |\hat{Z}^{(1)}_n|^2 + |\hat{Z}^{(2)}_n|^2
\]
\[
\leq |\epsilon_{\mu,n}|^2 |\theta_n|^2 + |\bar{\theta}_n|^2,
\]
where in the second line, we used that \( Q \) is an orthogonal matrix. By Lemma 12, for every \( n \geq n_1 \geq n_0 \), \( \forall \mu \in \text{Sp}(D^*) \), \( |\epsilon_{\mu,n}| \leq \epsilon_{id} n^{-1-\beta} \), so that
\[
\mathbb{E}|\hat{Z}^{(2)}_n|^2 \leq |\epsilon_{id} n^{-1-\beta}|^2 \frac{c_2}{1-\beta} n^{-\beta} + |\bar{\theta}_n|^2,
\]
where in the second line, we used that \( n_1^{-1+\beta} \leq n_0^{-1+\beta} \leq \mu/2 \).

**Lemma 16.** Let \( N \) be a positive integer and \( (u_n)_{n \geq 0} \) be a sequence which satisfies
\[
\forall n \geq N \quad u_{n+1} \leq u_n \left( 1 - \frac{1}{n+1} \right)^2 + C_1 n^{-r} + \frac{V}{(n+1)^2} + C_2 n^{-q},
\]
with \( r \in (1, 2) \) and \( q \in (2, 3) \). Assume that \((C_1, N)\) satisfies: \( C_1 N^{1-r} \leq 1 \), Then,
\[
\forall n \geq N \quad u_n \leq \frac{V}{n} + O_{id} \left( \|u_N N^2 \|_n^2 + C_1 V n^{-r} + C_2 n^{-(q-1)} \right)
\]
\[
\leq \frac{V}{n} + n^{-r} n^{-(q-1)} O_{id} \left( \|u_N N^r \|_n^2 + C_1 V + C_2 \right).
\]

**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 \( n \geq N \)
\[
u_n \leq u_N \prod_{k=N+1}^n \left( 1 + C_1 \left( \frac{k}{k-1} \right)^2 \right) \]
\[
\leq k^2 \prod_{\ell=k+1}^n \left( 1 + C_1 \left( \frac{\ell}{\ell-1} \right)^2 \right) \leq k^2 \prod_{\ell=k}^n \left( 1 + C_1 \left( \frac{\ell}{\ell-1} \right)^2 \right) \leq k^2 \exp \left( 2C_1 \sum_{\ell=k}^n (\ell-1)^{-r} \right)
\]
where in the last line, we used the inequality \( \log(1+x) \leq x \) for \( x > -1 \) and a comparison between series and integrals. Now, a constant \( c \) exists such that \( \exp(x) \leq 1 + cx \) on \([0, 2]\) and with the condition \( C_1 N^{1-r} \leq 1 \), we get for every \( k \in \{ N+1, \ldots, n-1 \} \),
\[
\prod_{\ell=k+1}^n \left( 1 + C_1 k^{-r} \right) \leq c n^2.
\]
Plugging this inequality into (53) leads to: for all \( n \geq N \),
\[
u_n \leq c \frac{u_N N^2}{n^2} + \frac{V(n-N)}{n^2} + c \frac{C_1 V}{n^2} \sum_{k=N+1}^n k^{-r} + c \frac{C_2}{n^2} \sum_{k=N+1}^n k^{2-q}
\]
\[
\leq \frac{V}{n} + c \left( \frac{u_N N^2}{n^2} + C_1 V n^{-r} + C_2 n^{-(q-1)} \right).
\]
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 $\theta^* = f(\theta^*) = 0$.

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

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

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

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

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

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

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

$$e'(\tau) = \varphi'(f(\chi_x(\tau)))\langle \nabla f(\chi_x(\tau)), \dot{\chi}_x(\tau) \rangle \geq -\varphi'(f(\chi_x(\tau)))\|\nabla f(\chi_x(\tau))\|^2 \leq -m\|\dot{\chi}_x(\tau)\|.$$ 

We deduce that:

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

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

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

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

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

which is the desired conclusion. \qed