Electronic Journal of Statistics
Vol. 12 (2018) 461–529
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1395

Stochastic heavy ball*

Sébastien Gadat, Fabien Panloup and Sofiane Saadane

Toulouse School of Economics, UMR 5604
Université de Toulouse, France
e-mail: sebastien.gadat@math.univ-toulouse.fr

Laboratoire Angevin de Recherche en Mathématiques, UMR 6093
Université d'Angers, France
e-mail: panloup@math.univ-angers.fr

Institut de Mathématiques de Toulouse, UMR 5219
Université de Toulouse, France
e-mail: sofiane.saadane@math.univ-toulouse.fr

Abstract: This paper deals with a natural stochastic optimization procedure derived from the so-called Heavy-ball method differential equation, which was introduced by Polyak in the 1960s with his seminal contribution [Pol64]. The Heavy-ball method is a second-order dynamics that was investigated to minimize convex functions f. The family of second-order methods recently received a large amount of attention, until the famous contribution of Nesterov [Nes83], leading to the explosion of large-scale optimization problems. This work provides an in-depth description of the stochastic heavy-ball method, which is an adaptation of the deterministic one when only unbiased evalutions of the gradient are available and used throughout the iterations of the algorithm. We first describe some almost sure convergence results in the case of general non-convex coercive functions f. We then examine the situation of convex and strongly convex potentials and derive some non-asymptotic results about the stochastic heavy-ball method. We end our study with limit theorems on several rescaled algorithms.

MSC 2010 subject classifications: Primary 60J70, 35H10, 60G15, 35P15.
Keywords and phrases: Stochastic optimization algorithms, second-order methods, random dynamical systems.

Received September 2016.

Contents

*The authors gratefully acknowledge the reviewers and the associate editor for their constructive remarks and warm encouragements.
† This research article benefited from the support of the FMJH Program PGMO (grant COSAL) and from the support of EDF, Thales, Orange.

1. Introduction

Minimization problems with deterministic methods. Finding the minimum of a function f over a set Ω with an iterative procedure is very popular among numerous scientific communities and has many applications in optimization, image processing, economics and statistics, to name a few. We refer to [NY83] for a general survey on optimization algorithms and discussions related to complexity theory, and to [Nes04, BV04] for a more focused presentation on convex optimization problems and solutions. The most widespread approaches rely on some first-order strategies, with a sequence $(X_k)_{k\geqslant 0}$ that evolves over Ω with a first-order recursive formula $X_{k+1} = \Psi[X_k, f(X_k), \nabla f(X_k)]$ that uses a local approximation of f at point X_k, where this approximation is built with the knowledge of $f(X_k)$ and $\nabla f(X_k)$ alone. Among them, we refer to the steepest descent strategy in the convex unconstrained case, and to the Frank-Wolfe

[FW56] algorithm in the compact convex constrained case. A lot is known about first-order methods concerning their rates of convergence and their complexity. In comparison to second-order methods, first-order methods are generally slower and are significantly degraded on ill-conditioned optimization problems. However, the complexity of each update involved in first-order methods is relatively limited and therefore useful when dealing with a large-scale optimization problem, which is generally expensive in the case of Interior Point and Newton-like methods. A second-order "optimal" method was proposed in [Nes83] in the 1980s' (also see [BT09] for an extension of this method with proximal operators). The so-called *Nesterov Accelerated Gradient Descent* (NAGD) has particularly raised considerable interest due to its numerical simplicity, to its low complexity and to its mysterious behavior, making this method very attractive for large-scale machine learning problems. Among the available interpretations of NAGD, some recent advances have been proposed concerning the second-order dynamical system by [WSC16], being a particular case of the generalized Heavy Ball with Friction method (referred to as HBF in the text), as previously pointed out in [CEG09a, CEG09b]. In particular, as highlighted in [CEG09a], NAGD may be seen as a specific case of HBF after a time rescaling $t = \sqrt{s}$, thus making the acceleration explicit through this change of variable, as well as being closely linked to the modified Bessel functions when f is quadratic.

Stochastic optimization methods. In problems where the effective computation of the gradient is too costly, an idea initiated by the seminal contributions of [RM51] and [KW52] is to *randomize* the gradient and to consider a so-called stochastic gradient descent (S.G.D. for short). This situation typically appears when the function to minimize is an integral, an expectation of a given random variable or in discrete minimization problems with a very large number of points. Even if this field of investigation has been initiated in the fifties, the study of S.G.D. algorithms has met a great regain of interest in the large scale machine learning community (see, *e.g.*, [GY07, Bot10]), owing in particular to its ability of being parallelized. In this setting, stochastic versions of deterministic accelerated algorithms recently received a growth of interest (see *e.g.* [JKK+17, Nit15]) and led to many open questions (as mentioned in the communication `http://praneethnetrapalli.org/ASGD-long.pdf`). In the sequel, we are going to focus on some of them for the HBF model.

Objectives and motivations. The HBF ordinary differential equation, whose equation is given by (2.1) is a second order system which can be viewed as a gradient descent with memory (see (2.2)). The aim of this paper, which is mainly theoretical, is then to study stochastic optimization algorithms derived from these deterministic dynamical systems. Before going further in the presentation of this procedure, let us go deeper in the general objectives and motivations of the paper.

From a theoretical point of view, we could formulate the general motivation as follows: what are the consequences of the memory on the convergence of the HBF-optimization procedure ? To this (too) general question, our first general

objective is to exhibit some conditions on the memory which guarantee the $a.s.$-convergence towards local of global minima. This part of our work can be viewed at the middle of two topics: the study of the long-time behavior of HBF ordinary differential equations (see [CEG09a, CEG09b]) and, on the other hand of, HBF stochastic differential equations (on this topic, see [GP14], [MS17]). In particular, the hazard involved in stochastic algorithms is located between the fully deterministic dynamics of an O.D.E. and the purely randomized dynamics involved in stochastic differential equations and it could be interesting to fill the gap between these two settings.

At a second level, we aim at studying the rate of convergence of the HBF-procedure. In particular, compared to the standard stochastic gradient descent, what are the effects of the memory on the (asymptotic of non-asymptotic) error? We will tackle this question from a theoretical and numerical point of view.

From a dynamical point of view, our original motivation was to take advantage of the exploration abilities of the HBF. Actually, as a second order method, the deterministic HBF ordinary differential equation already possesses the ability to escape some local traps/minimizers of the function (which is not the case for the standard gradient descent). As a complement of the above theoretical questions, it may be of interest to wonder about the relevance of this optimization procedure in a multi-wells setting. This question seems to be very difficult to tackle in full generality but may be of primary importance for nowadays machine learning problems where non-convex multi-modal are commonly encountered, for example in the matrix completion problem (see, *e.g.*, [BM05]). Starting from this multi-modal motivation, some old works investigated the ability of the S.G.D. to escape local traps (see, *e.g.*, [BD96, Pem90] for pioneering almost sure convergence results towards local minimizers). Recently, [LSJR16] establishes the convergence towards a local minimizer with probability 1 when the initialization point is randomly sampled, whereas [JKN16] studies the particular case of the matrix completion problem with the S.G.D. Beyond the natural exploration of the state space ability of the HBF, the recent work [JNJ17] has also investigated the escape properties of another second order stochastic algorithms with inertia and has shown the ease of stochastic accelerated gradient descent to escape from local minimizers faster than the standard S.G.D.

State of art. As a stochastic version of the HBF strategy, our work falls into the field of second order stochastic gradient algorithms with a memory that produces an acceleration of the drift. We detail below some important references relevant with these themes of research.

Standard S.G.D. and Averaging. As mentioned before, the development of efficient methods to minimize functions when only *noisy* gradients are available is an important problem in view of applications.

To this end, let us recall the existing results for the standard S.G.D., generally called Robbins-Monro algorithm. In this setting, it can be shown in a strongly convex setting that the algorithm can attain the rate $O(1/n)$ (see *e.g.* [Duf97]), but is really sensitive to the step sizes used. This remark led [PJ92] to develop an averaging method that makes it possible to use longer step sizes of the Robbins-

Monro algorithm, and to then average these iterates with a Cesaro procedure so that this method produces optimal results in the minimax sense (see [NY83]) for convex and strongly convex minimization problems, as pointed out in [BM11].

HBF-algorithm as a pertubed second order O.D.E. Numerous studies have addressed a dynamical system point of view and studied the close links between stochastic algorithms and their deterministic counterparts for some general function f (*i.e.*, even non convex). These links originate in the famous Kushner-Clark Theorem (see [KY03]) and successful improvements have been obtained using differential geometry by [BH96, Ben06] on the long-time behavior of stochastic algorithms. In particular, a growing field of interest concerns the behavior of self-interacting stochastic algorithms (see, among others, [BLR02] and [GP14]) because these non-Markovian processes produce interesting features from the modeling point of view (an illustration may be found in [GMP15]). Our work is also linked with random dynamical systems $(X_n, Y_n)_{n\geqslant 1}$ where the two coordinates do not evolve at the same speed: this will be the case when we handle a specific polynomial form of memory function (see below). This field of research has been investigated by the pioneering work [FW84] where homogeneization methods are developed for stochastic differential equations. For optimization procedures, this two-scales setting appears in [Bor97] (see also [Bor08]) where under an appropriate control of the noise and some uniqueness conditions, the *a.s.* -convergence is obtained through a pseudo-trajectory approach (on this topic, see [BH96]). We will come back on this connexion in the beginning of Subsection 3.2.1 (see (3.2)).

Accelerated stochastic methods Several theoretical contributions to the study of specific second-order stochastic optimization algorithms exist. [Lan12] explores some adaptations of the NAGD in the stochastic case for composite (strongly or not) convex functions. Other authors [GL13, GL16] obtained convergence results for the stochastic version of a variant of NAGD for non-convex optimization for gradient Lipschitz functions but these methods cannot be used for the analysis of the Heavy-ball algorithm. Finally, a recent work [YLL16] proposes a unified study of some stochastic momentum algorithms while assuming restrictive conditions on the noise of each gradient evaluation and on the constant step size used. It should be noted that [YLL16] provides a preliminary result on the behavior of the stochastic momentum algorithms in the non-convex case with possible multi-well situations. Our work aims to study the properties of a stochastic optimization algorithm naturally derived from the generalized heavy ball with friction method.

Organisation

Our paper is organized as follows: Section 2 introduces the stochastic algorithm as well as the main assumptions needed to obtain some results on this optimization algorithm. For the sake of readability, these results are then provided in Section 2.4 without too many technicalities. Sections 3, 4 and 5 are devoted to the proof of these results (some technical details are postponed in the appendix

sections). More precisely, Section 3 is dedicated to the almost sure convergence result we can obtain in the case of a non-convex function f with several local minima. Section 4 establishes the convergence rates of the stochastic heavy ball in the strongly convex case. Section 5 provides a central limit theorem in a particular case of the algorithm. Finally, in Section 6, we focus on a series of numerical experiments.

2. Stochastic heavy ball

We begin with a brief description of what is known about the underlying ordinary differential equation (referred to as a dynamical system below).

2.1. Deterministic heavy ball

This method introduced by Polyak in [Pol64] is inspired from the physical idea of producing some inertia on the trajectory to speed up the evolution of the underlying dynamical system: a ball evolves over the graph of a function f and is submitted to both damping (due to a friction on the graph of f) and acceleration. More precisely, this method is a second-order dynamical system described by the following O.D.E.:

$$\ddot{x}_t + \gamma_t \dot{x}_t + \nabla f(x_t) = 0, \tag{2.1}$$

where $(\gamma_t)_{t\geqslant 0}$ corresponds to the damping coefficient, which is a key parameter of the method. In particular, it is shown in [CEG09a] that the trajectory converges only under some restrictive conditions on the function $(\gamma_t)_{t\geqslant 0}$, namely:

- if $\int_0^{+\infty} \gamma_s ds = \infty$, then $(f(x_t))_{t\geqslant 0}$ converges,
- if $\int_0^{\infty} e^{-\int_0^t \gamma_s ds} dt < \infty$, then $(x_t)_{t\geqslant 0}$ converges towards one of the minima of any convex function f.

Intuitively, these conditions translate the oscillating nature of the solutions of (2.1) into a quantitative setting for the convergence of the trajectories: if the convergence $\gamma_t \to 0$ is sufficiently fast, then the trajectory cannot converge (the limiting case being $\ddot{x} + \nabla f(x) = 0$). These properties lead us to consider two natural families of functions $(\gamma_t)_{t\geqslant 0}$: $\gamma_t = r/t$ with $r > 1$ and $\gamma_t = \gamma > 0$. To convert (2.1) into a tractable iterative algorithm, it is necessary to rewrite this O.D.E. using some coupled equations on position/speed, such equations are commonly referred to as momentum equations (see, *e.g.* [Nes83] for an example). Consistent with [CEG09b], (2.1) is *equivalent* to the following integro-differential equation:

$$\dot{x}_t = -\frac{1}{k(t)} \int_0^t h(s) \nabla f(x_s) ds, \tag{2.2}$$

where h and k are two increasing functions related to γ. This *equivalent* feature of the integro-differential formulation given by Equation (2.2) should be understood as a differential equation that produces the same integral curve, up to a suitable change of time, than the one produced by Equation (2.1).

Even though any couple of increasing functions may be chosen for h and k, it is natural to consider only the situation where $h = \dot{k}$ to produce an integral over $[0, t]$ that corresponds to a weighted average of $(\nabla f(x_s))_{s\in[0,t]}$. In such a case, h then represents the amount of weight on the past we consider in (2.2). Through the introduction of the auxiliary function $y_t = k(t)^{-1}\int_0^t h(s)\nabla f(x_s)ds$, it can be checked that Equation (2.2) can be rewritten as a first order o.d.e. In the special case $h = \dot{k}$, this leads to the system

$$\begin{cases} \dot{x}_t = -y_t \\ \dot{y}_t = r(t)(\nabla f(x_t) - y_t) \end{cases} \quad \text{with } r(t) = \frac{h(t)}{k(t)} = \frac{\dot{k}(t)}{k(t)}. \tag{2.3}$$

In the spirit of [GP14] (in a stochastic setting), we will mainly consider this weighted averaged setting for two typical situations that correspond to a stable convergent dynamical system in the deterministic case (see [CEG09a] for further details):

- The *exponentially memoried HBF*: $k(t) = e^{\lambda t}$ and $h(t) = \dot{k}(t) = \lambda e^{\lambda t}$ (and to a constant damping function $\gamma_s = \sqrt{\lambda}$). In this case, $r(t) = \lambda$ so that (2.3) is an homogeneous o.d.e.
- The *polynomially memoried HBF*: $k(t) = t^{\alpha+1}$ and $h(t) = (\alpha+1)t^{\alpha}$ so that $r(t) = \frac{\alpha+1}{t}$. Here, the damping parameter satisfies $\gamma_s = \frac{2\alpha+1}{s}$). In this case, we retrieve the o.d.e. of the NAGD when $\alpha = 1$ (see [WSC16] and their "magic" constant $3 = 2\alpha + 1$ in that case).

2.2. Stochastic HBF

We now define the stochastic Heavy Ball algorithm as a noisy gradient discretized system related to (2.3). More precisely, we set $(X_0, Y_0) = (x, y) \in \mathbb{R}^{2d}$ and for all $n \geqslant 0$:

$$\begin{cases} X_{n+1} = X_n - \gamma_{n+1} Y_n \\ Y_{n+1} = Y_n + \gamma_{n+1} r_n (\nabla f(X_n) - Y_n) + \gamma_{n+1} r_n \Delta M_{n+1}, \end{cases} \tag{2.4}$$

where the natural filtration of the sequence $(X_n, Y_n)_{n\geqslant 0}$ is denoted $(\mathcal{F}_n)_{n\geqslant 1}$ and:

- (ΔM_n) is a sequence of $(\mathcal{F}_n)$-martingale increments. For applications, ΔM_{n+1} usually represents the difference between the "true" value of $\nabla f(X_n)$ and the one observed at iteration n denoted $\partial_x F(X_n, \xi_n)$, where $(\xi_n)_n$ is a sequence of i.i.d. random variables and F is an $\mathbb{R}^d$-valued measurable function such that:

$$\forall u \in \mathbb{R}^d \qquad \mathbb{E}\left[\partial_x F(u, \xi)\right] = \nabla f(u)$$

In this case,

$$\Delta M_{n+1} = \nabla f(X_n) - \partial_x F(X_n, \xi_n). \tag{2.5}$$

The randomness appears in the second component of the algorithm (2.4), whereas it was handled in the first component in [GP14]. We will introduce some assumptions on f and on the martingale sequence later.

- $(\gamma_n)_{n\geqslant 1}$ corresponds to the step size used in the stochastic algorithm, associated with the "time" of the algorithm represented by:

$$\Gamma_n = \sum_{k=1}^{n} \gamma_k \qquad \text{such that} \qquad \lim_{n\longrightarrow+\infty} \Gamma_n = +\infty.$$

For the sake of convenience, we also define:

$$\Gamma_n^{(2)} = \sum_{k=1}^{n} \gamma_k^2,$$

which may converge or not according to the choice of the sequence $(\gamma_k)_{k\geqslant 1}$.

- $(r_n)_{n\geqslant 1}$ is a deterministic sequence that mimics the function $t \longmapsto r(t)$ defined as:

$$r_n = \frac{h(\Gamma_n)}{k(\Gamma_n)}. \tag{2.6}$$

In particular, when an exponentially weighted HBF with $k(t) = e^{rt}$ is chosen, we have $r_n = r > 0$, regardless of the value of n. In the other situation where $k(t) = t^r$, we obtain $r_n = r\Gamma_n^{-1}$.

2.3. Baseline assumptions

We introduce some of the general assumptions we will work with below. Some of these conditions are very general, whereas others are more specifically dedicated to the analysis of the strongly convex situation. We will use the notation $\|.\|$ (resp. $\|.\|_F$) below to refer to the Euclidean norm on $\mathbb{R}^d$ (resp. the Frobenius norm on $\mathcal{M}_{d,d}(\mathbb{R})$). Finally, when $A \in \mathcal{M}_{d,d}(\mathbb{R})$, $\|A\|_\infty$ will refer to the maximal size of the modulus of the coefficients of A: $\|A\|_\infty := \sup_{i,j} |A_{i,j}|$. Our theoretical results will obviously not involve all of these hypotheses simultaneously.

Function f. We begin with a brief enumeration of assumptions on the function f.

- Assumption $(\mathbf{H}_s)$: f is a function in $\mathcal{C}^2(\mathbb{R}^d, \mathbb{R})$ such that:

$$\lim_{|x|\longrightarrow+\infty} f(x) = +\infty \text{ and } \|D^2 f\|_\infty := \sup_{x\in\mathbb{R}^d} \|D^2 f(x)\|_F < +\infty \text{ and } \|\nabla f\|^2 \leqslant c_f f.$$

The assumption $(\mathbf{H}_s)$ is weak: it essentially requires that f be smooth, coercive and have, at the most, a quadratic growth on ∞. In particular, no convexity

hypothesis is made when f satisfies $(\mathbf{H}_s)$. It would be possible to extend most of our results to the situation where f is L-smooth (with a L-Lipschitz gradient), but we preferred to work with a slightly more stringent condition to avoid additional technicalities.

• Assumption $(\mathbf{H}_{SC}(\alpha))$: f is a convex function such that $D^2 f$ is Lipschitz and

$$\alpha = \inf_{x \in \mathbb{R}^d} \mathrm{Sp}\left(D^2 f(x)\right) > 0.$$

In particular, $(\mathbf{H}_{SC}(\alpha))$ implies that f is α-strongly convex, meaning that:

$$\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d \qquad f(x) \geqslant f(y) + \langle \nabla f(y), x - y \rangle + \frac{\alpha}{2}\|x - y\|^2.$$

Of course, $(\mathbf{H}_{SC}(\alpha))$ is still standard and is the most favorable case when dealing with convex optimization problems, leading to the best possible achievable rates. $(\mathbf{H}_{SC}(\alpha))$ translates the fact that the spectrum of the Hessian matrix at point x, denoted by $\mathrm{Sp}\left(D^2 f(x)\right)$, is lower bounded by $\alpha > 0$, uniformly over $\mathbb{R}^d$. The fact that $D^2 f$ is assumed to be Lipschitz will be useful to achieve convergence rates in Section 4.2.

Noise sequence $(\Delta M_{n+1})_{n \geqslant 1}$**.** We will essentially use three types of assumptions alternatively on the noise of the stochastic algorithm (2.4). The first and second assumptions are concerned with a concentration-like hypothesis. The first one is very weak and asserts that the noise has a bounded $\mathbb{L}^2$ norm.

• Assumption $(\mathbf{H}_{\boldsymbol{\sigma},\boldsymbol{p}})$: $(p \geqslant 1)$ For any integer n, we have:

$$\mathbb{E}(\|\Delta M_{n+1}\|^p | \mathcal{F}_n) \leqslant \sigma^2 (1 + f(X_n))^{\frac{p}{2}}.$$

The assumption $(\mathbf{H}_{\boldsymbol{\sigma},\mathbf{2}})$ is a standard convergence assumption for general stochastic algorithms. For some non-asymptotic rates of convergence results, we will rely on $(\mathbf{H}_{\boldsymbol{\sigma},\boldsymbol{p}})$ for any $p \geqslant 1$. In this case, we will denote the assumption by $(\mathbf{H}_{\boldsymbol{\sigma},\infty})$. Finally, let us note that the condition could be slightly alleviated by replacing the right-hand member by $\sigma^2(1 + f(X_n) + |Y_n|^2)^p$. However, in view of the standard case (2.5), this improvement has little interest in practice, which explains our choice.

• Assumption $(\mathbf{H}_{\boldsymbol{Gauss},\boldsymbol{\sigma}})$: For any integer n, the Laplace transform of the noise satisfies:

$$\forall t \geqslant 0 \qquad \mathbb{E}\left[\exp(t \Delta M_{n+1}) | \mathcal{F}_n\right] \leqslant e^{\frac{\sigma^2 t^2}{2}}.$$

This hypothesis is much stronger than $(\mathbf{H}_{\boldsymbol{\sigma},\boldsymbol{p}})$ and translates a sub-Gaussian behavior of $(\Delta M_{n+1})_{n \geqslant 1}$. In particular, it can be easily shown that $(\mathbf{H}_{\boldsymbol{Gauss},\boldsymbol{\sigma}})$ implies $(\mathbf{H}_{\boldsymbol{\sigma},\boldsymbol{p}})$. Hence, $(\mathbf{H}_{\boldsymbol{Gauss},\boldsymbol{\sigma}})$ is somewhat restrictive and will be used only to obtain one important result in the non-convex situation for the almost sure limit of the stochastic heavy ball with multiple wells.

• Assumption $(\mathbf{H}_{\mathcal{E}})$: For any iteration n, the noise of the stochastic algorithm satisfies:

$$\forall v \in \mathcal{S}^{d-1} \qquad \mathbb{E}\left(|\langle \Delta M_n, v\rangle| \ | X_n, Y_n\right) \geqslant c_v > 0,$$

where $\mathcal{S}^{d-1}$ stands for the unit Euclidean sphere of $\mathbb{R}^d$.

This assumption will be essential to derive an almost sure convergence result towards minimizers of f. Roughly speaking, this assumption states that the noise is uniformly elliptic given any current position of the algorithm at step n: the projection of the noise has a non-vanishing component over all directions v. We will use this assumption to guarantee the ability of (2.4) to get out of any unstable point.

Step sizes. One important step in the use of stochastic minimization algorithms relies on an efficient choice of the step sizes involved in the recursive formula (*e.g.* in Equation 2.4). We will deal with the following sequences $(\gamma_n)_{n\geqslant 0}$ below.

• Assumption $(\mathbf{H}^{\gamma}_{\beta})$: The sequence $(\gamma_n)_{n\geqslant 0}$ satisfies:

$$\forall n \in \mathbb{N} \qquad \gamma_n = \frac{\gamma}{n^{\beta}} \qquad \text{with} \qquad \beta \in (0,1],$$

leading to:

$$\forall \beta \in (0,1) \qquad \Gamma_n \sim \frac{\gamma}{1-\beta} n^{1-\beta} \qquad \text{whereas} \qquad \Gamma_n \sim \gamma \log n \quad \text{when} \quad \beta = 1.$$

Memory size. We consider the exponentially and polynomially-weighted HBF as a unique stochastic algorithm parameterized by the memory function $(r_n)_{n\geqslant 1}$. From the definition of r_n given in (2.6), we note that in the exponential case, $r_n = r$ remains constant while the inertia brought by the memory term in the polynomial case $(r_n)_{n\in\mathbb{N}}$ is defined by $r_n = \frac{r}{\Gamma_n}$. Under Assumption $(\mathbf{H}^{\gamma}_{\beta})$, we can show that regardless of the memory, we have:

$$\sum_{n\in\mathbb{N}} \gamma_n r_n = +\infty.$$

This is true when $r_n = r$ because $\gamma_n = \gamma n^{-\beta}$ with $\beta \leqslant 1$. It is also true when we deal with a polynomial memory since in that case:

- if $\beta < 1$, then $\gamma_n r_n \sim \gamma n^{-\beta} \times r(1-\beta)\gamma^{-1} n^{-1+\beta} \sim r(1-\beta) n^{-1}$
- if $\beta = 1$, then $\gamma_n r_n \sim \dfrac{r}{n \log n}$ and $\sum_{k\leqslant n} \gamma_k r_k \sim \log(\log n)$.

Similarly, we also have that in the polynomial case, regardless of β:

$$\sum_{n} \gamma_n^2 r_n < +\infty,$$

although this bound holds in the exponential situation when $\beta > 1/2$. Below, we will use these properties on the sequences $(\gamma_n)_{n\geqslant 0}$ and $(r_n)_{n\geqslant 0}$ and define the next set of assumptions:

• Assumption ($\mathbf{H_r}$): The sequence $(r_n)_{n\geqslant 0}$ is a non-increasing sequence such that:

$$\sum_{n\geqslant 1} \gamma_{n+1} r_n = +\infty \qquad \text{and} \qquad \sum_{n\geqslant 1} \gamma_{n+1}^2 r_n < +\infty$$

and

$$\limsup_{n\to+\infty} \frac{1}{2\gamma_{n+1}} \left(\frac{1}{r_n} - \frac{1}{r_{n-1}} \right) =: c_r < 1.$$

In the exponential case, $c_r = 0$, whereas if $r_n = r/\Gamma_n$, it can be shown that $c_r = \frac{1}{2r}$ and the last point is true when $r > 1/2$. In any case, r_∞ will refer to the limiting value of r_n when $n \longrightarrow +\infty$, which is either 0 or $r > 0$.

2.4. Main results

Section 3 is dedicated to the situation of a general coercive function f. We obtain the almost sure convergence of the stochastic HBF towards a critical point of f.

Theorem 2.1. *Assume that f satisfies $(\mathbf{H}_s)$, that $(\mathbf{H_{\sigma,2}})$ holds and that the sequences $(\gamma_n)_{n\geqslant 1}$ and $(r_n)_{n\geqslant 1}$ are chosen such that $(\mathbf{H}_\beta^\gamma)$ and $(\mathbf{H_r})$ are fulfilled. If for any z, $\{x, f(x) = z\} \cap \{x, \nabla f(x) = 0\}$ is locally finite, then (X_n) a.s. converges towards a critical point of f.*

This result obviously implies the convergence when f has a unique critical point. In the next theorem, we focus on the case where this uniqueness assumption fails, under the additional elliptic assumption $(\mathbf{H}_\mathcal{E})$.

Theorem 2.2. *Assume that f satisfies $(\mathbf{H}_s)$, that the noise is elliptic, i.e., $(\mathbf{H}_\mathcal{E})$ holds, and the sequence $(\gamma_n)_{n\geqslant 1}$ is chosen such that $(\mathbf{H}_\beta^\gamma)$ and $(\mathbf{H_r})$ are fulfilled. If for any z, $\{x, f(x) = z\} \cap \{x, \nabla f(x) = 0\}$ is locally finite, we have:*

(*a*) *If $r_n = r$ (exponential memory) and $(\mathbf{H_{\sigma,2}})$ holds, then (X_n) a.s. converges towards a local minimum of f.*
(*b*) *If $r_n = r\Gamma_n^{-1}$ and the noise is sub-Gaussian, i.e., $(\mathbf{H_{Gauss,\sigma}})$ holds, then (X_n) a.s. converges towards a local minimum of f when $\beta < 1/3$.*

Remark 2.1. ▷ *The previous result provides some guarantees when f is a multiwell potential. In (a), we consider the exponentially weighted HBF and show that the convergence towards a local minimum of f always holds under the additional assumption $(\mathbf{H}_\mathcal{E})$. To derive this result, we will essentially use the former results of [BD96] on "homogeneous" stochastic algorithms.*
▷ *Point (b) is concerned by polynomially-weighted HBF and deserves more comment:*

- *First, the result is rather difficult because of the time inhomogeneity of the stochastic algorithm, which can be written as $Z_{n+1} = Z_n + \gamma_{n+1}F_n(Z_n) + \gamma_{n+1}\Delta M_{n+1}$: the drift term F_n depends on Z_n and on the integer n, which will induce technical difficulties in the proof of the result. In particular, the assumption $\beta < 1/3$ will be necessary to obtain a good lower bound of the*

drift term in the unstable manifold direction with the help of the Poincaré Lemma near hyperbolic equilibrium of a differential equation.

- *Second, the sub-Gaussian assumption* $(\mathbf{H_{Gauss,\sigma}})$ *is less general than* $(\mathbf{H_{\sigma,2}})$ *even though it is still a reasonable assumption within the framework of a stochastic algorithm. To prove* (b)*, we will need to control the fluctuations of the stochastic algorithm around its deterministic drift, which will be quantified by the expectation of the random variable* $\sup_{k\geqslant n}\gamma_k^2\|\Delta M_k\|^2$*. The sub-Gaussian assumption will be mainly used to obtain an upper bound of such an expectation, with the help of a coupling argument. Our proof will follow a strategy used in [Pem90] and [Ben06] where this kind of expectation has to be upper bounded. Nevertheless, the novelty of our work is also to generalize the approach to unbounded martingale increments: the arguments of [Pem90, Ben06] are only valid for a bounded martingale increment, which is a somewhat restrictive framework.*

In Section 4, we focus on the consistency rate under stronger assumptions on the convexity of f. In the exponential memory case, we are able to control the quadratic error and to establish a CLT for the stochastic algorithm under the general assumption $(\mathbf{H}_{SC}(\alpha))$. In the polynomial case, the problem is more involved and we propose a result for the quadratic error only when f is a quadratic function (see Remark 2.2 for further comments on this restriction). More precisely, using the notation $\lesssim$ to refer to an inequality, up to a universal multiplicative constant, we establish the following results.

Theorem 2.3. *Denote by* $x^\star$ *the unique minimizer of* f *and assume that* $(\mathbf{H}_\beta^\gamma)$, $(\mathbf{H}_s)$, $(\mathbf{H}_{SC}(\alpha))$ *and* $(\mathbf{H_{\sigma,2}})$ *hold, we have:*

(a) *When* $r_n = r$ *(exponential memory) and* $\beta < 1$*, we have:*

$$\mathbb{E}\left[\|X_n - x^\star\|^2 + \|Y_n\|^2\right] \lesssim \gamma_n$$

If $(\mathbf{H_{\sigma,\infty}})$ *holds and* $\beta = 1$*, set* $\alpha_r = r\left(1 - \sqrt{1 - \frac{(4\underline{\lambda})\wedge r}{r}}\right)$ *where* $\underline{\lambda}$ *denotes the smallest eigenvalue of* $D^2 f(x^\star)$*. We have, for any* $\varepsilon > 0$*:*

$$\mathbb{E}\left[\|X_n - x^\star\|^2 + \|Y_n\|^2\right] \lesssim \begin{cases} n^{-1} & \text{if } \gamma\alpha_r > 1 \\ n^{-\alpha_r+\varepsilon} & \text{if } \gamma\alpha_r \leqslant 1. \end{cases}$$

(b) *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be a quadratic function. Assume that* $r_n = r\Gamma_n^{-1}$ *(polynomial memory) with* $\beta < 1$*. Then, if* $r > \frac{1+\beta}{2(1-\beta)}$*, we have:*

$$\mathbb{E}\left[\|X_n - x^\star\|^2 + \Gamma_n\|Y_n\|^2\right] \lesssim \gamma_n$$

When $r_n = r\Gamma_n^{-1}$ *(polynomial memory) and* $\beta = 1$*, we have:*

$$\mathbb{E}\left[\|X_n - x^\star\|^2 + \log n\|Y_n\|^2\right] \lesssim \frac{1}{\log n}.$$

For (a), the case $\beta < 1$ is a consequence of Proposition 4.3 (or Proposition 4.1 in the quadratic case), whereas the (more involved) case $\beta = 1$ is dealt with Propositions 4.1 and 4.4 for the quadratic and the non-quadratic cases, respectively. We first stress that that when $\beta < 1$, the noise only needs to satisfy $(\mathbf{H}_{\boldsymbol{\sigma},\boldsymbol{p}})$ to obtain our upper bound. When we deal with $\beta = 1$, we could prove a positive result in the quadratic case when we only assume $(\mathbf{H}_{\boldsymbol{\sigma},\boldsymbol{p}})$. Nevertheless, the stronger assumption $(\mathbf{H}_{\boldsymbol{\sigma},\infty})$ is necessary to produce a result in the general strongly convex situation. Finally, (b) is a consequence of Proposition 4.2.

Remark 2.2. ▷ *It is worth noting that in* (a) *(*$\beta = 1$*), the dependency of the parameter* α_r *in* $D^2 f$ *only appears through the smallest eigenvalue of* $D^2 f(x^\star)$*. In particular, it does not depend on* $\inf_{x\in\mathbb{R}^d} \lambda_{D^2 f(x)}$ *as it could be expected in this type of result. In other words, we are almost able to retrieve the conditions that appear when* f *is quadratic. This optimization of the constraint is achieved with a "power increase" argument, but this involves a stronger assumption* $(\mathbf{H}_{\boldsymbol{\sigma},\infty})$ *on the noise.*

▷ *The restriction to quadratic functions in the polynomial case may appear surprising. In fact, the "power increase" argument does not work in this non-homogeneous case. However, when* $\beta < 1$*, it would be possible to extend to non-quadratic functions through a Lyapunov argument (on this topic, see Remark 4.3), but under some quite involved conditions on* r*,* β *and the Hessian of* f*. Hence, we chose to only focus on the quadratic case and to try to obtain some potentially optimal conditions on* r *and* β *only (in particular, there is no dependence to the spectrum of* $D^2 f$*). The interesting point is that it is possible to preserve the standard rate order when* $\beta < 1$ *but under the constraint* $r > \frac{1+\beta}{2(1-\beta)}$*, which increases with* β*. In particular, the rate* $\mathcal{O}(n^{-1})$ *cannot be attained in this case (see Remark 4.2 for more details).*

Finally, we conclude by a central limit theorem related to the stochastic algorithm the *exponential memory case*.

Theorem 2.4. *Assume* $(\mathbf{H}_s)$ *and* $(\mathbf{H}_{SC}(\alpha))$ *are true. Suppose that* $r_n = r$ *and that* $(\mathbf{H}_\beta^\gamma)$ *holds with* $\beta \in (0,1)$ *or,* $\beta = 1$ *and* $\gamma\alpha_r > 1$*. Assume that* $(\mathbf{H}_{\boldsymbol{\sigma},\boldsymbol{p}})$ *holds with* $p > 2$ *when* $\beta < 1$ *and* $p = \infty$ *when* $\beta = 1$*. Finally, suppose that the following condition is fulfilled:*

$$\mathbb{E}\left[(\Delta M_{n+1})(\Delta M_{n+1})^t | \mathcal{F}_{n-1}\right] \xrightarrow{n\to+\infty} \mathcal{V} \quad \text{in probability} \tag{2.7}$$

where $\mathcal{V}$ *is a symmetric positive* $d \times d$*-matrix. Let* σ *be a* $d \times d$*-matrix such that* $\sigma\sigma^t = \mathcal{V}$*. Then,*

(i) *The normalized algorithm* $\left(\frac{X_n}{\sqrt{\gamma_n}}, \frac{Y_n}{\sqrt{\gamma_n}}\right)_n$ *converges in law to a centered Gaussian distribution* $\mu_\infty^{(\beta)}$*, which is the invariant distribution of the (linear) diffusion with infinitesimal generator* $\mathcal{L}$ *defined on* $\mathcal{C}^2$*-functions by:*

$$\mathcal{L}g(z) = \left\langle \nabla g(z), \left(\frac{1}{2\gamma} 1_{\{\beta=1\}} I_{2d} + H\right) z \right\rangle + \frac{1}{2}\mathrm{Tr}(\Sigma^T D^2 g(z)\Sigma)$$

with

$$H = \begin{pmatrix} 0 & -I_d \\ rD^2 f(x^\star) & -rI_d \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} 0 & 0 \\ 0 & \sigma \end{pmatrix}.$$

(*ii*) *In the simple situation where* $\mathcal{V} = \sigma_0^2 I_d$ *(*$\sigma_0 > 0$*) and* $\beta < 1$*. In this case, the covariance of* $\mu_\infty^{(\beta)}$ *is given by*

$$\frac{\sigma_0^2}{2} \begin{pmatrix} \{D^2 f(x^\star)\}^{-1} & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times d} & rI_d \end{pmatrix}$$

In particular,

$$\frac{X_n}{\sqrt{\gamma_n}} \Longrightarrow \mathcal{N}\left(0, \frac{\sigma_0^2}{2}\{D^2 f(x^\star)\}^{-1}\right). \tag{2.8}$$

Remark 2.3. ▷ *As a first comment of the above theorem, let us note that in the fundamental example where:*

$$\Delta M_{n+1} = \nabla f(X_n) - \partial_x F(X_n, \xi_n), \quad n \geqslant 1,$$

the additional assumption (2.7) *is a continuity assumption. Actually, in this case:*

$$\mathbb{E}[\Delta M_n \Delta M_n^t | \mathcal{F}_{n-1}] = \bar{\mathcal{V}}(X_n), \quad \text{with } \bar{\mathcal{V}}(x) = \mathrm{Cov}(F(x, \xi_1)).$$

Thus, since $X_n \to x^\star$ *a.s., Assumption* (2.7) *is equivalent to the continuity of* $\bar{\mathcal{V}}$ *in* $x^\star$ *so that:*

$$\mathcal{V} = \bar{\mathcal{V}}(x^\star).$$

▷ *Point* (*ii*) *of Theorem 2.4 reveals the behavior of the asymptotic variance of* Y *increases with* r*. This translates the fact that the instantaneous speed coordinate* Y *is proportional to* r *in Equation* (2.4)*, which then implies a large variance of the* Y *coordinate when we use an important value of* r*.*

▷ *When* $\beta = 1$*, it is also possible (but rather technical) to make the limit variance explicit. The expression obtained with the classical stochastic gradient descent with step-size* γn^{-1} *and Hessian* λ*, the asymptotic variance is* $\gamma/(2\lambda\gamma - 1)$*, whose optimal value is attained when* $\gamma = \lambda^{-1}$ *(it attains the Cramer-Rao lower bound). Concerning now the stochastic HBF, for example, when* $d = 1$ *and* $r \geqslant 4\lambda$ *(the result is still valid in higher dimensions, see Section 5), we can show that:*

$$\lim_{n\longrightarrow+\infty} \gamma_n^{-1}\mathbb{E}[X_n^2] = \sigma_0^2 \frac{2\lambda r\gamma^3}{(\gamma r - 1)(2\lambda\gamma - \check{\alpha}_-)(2\lambda\gamma - \check{\alpha}_+)},$$

where $\check{\alpha}_+ = 1 + \sqrt{1 - \frac{4\lambda}{r}}$ *and* $\check{\alpha}_- = 1 - \sqrt{1 - \frac{4\lambda}{r}}$*. Similar expressions may be obtained when* $r < 4\lambda$*. Note also that we assumed that* $\gamma\alpha_r > 1$*, and it is easy to check that this condition implies that* $\gamma r > 1$ *because* $\alpha_r \leqslant r$*, regardless of* r*. In the meantime, this condition also implies that* $2\lambda\gamma > \check{\alpha}_+ \geqslant \check{\alpha}_-$*.*

Finally, this explicit value could be used to find the optimal calibration of the parameters to obtain the best asymptotic variance. Unfortunately, the expressions are rather technical and we can see that such calibrations are far from being independent of λ*, the a priori unknown Hessian of* f *on* $x^\star$*.*

3. Almost sure convergence of the stochastic heavy ball

In this section, the baseline assumption on the function f is $(\mathbf{H}_s)$, and we are thus interested in the almost sure convergence of the stochastic HBF. In particular, *we do not make any convexity assumption on* f.

Below, we will sometimes use standard and sometimes more intricate normalizations for the coupled process $Z_n = (X_n, Y_n)$. These normalizations will be of a different nature and, to be as clear as possible, we will always use the same notation $\check{Z}_n$ and $\check{Z}_n$ to refer to a rotation of the initial vector Z_n, whereas $\widetilde{Z}_n$ will introduce a scaling in the Y_n component of Z_n by a factor $\sqrt{r_n}$.

3.1. Preliminary result

We first state a useful upper bound that makes it possible to derive a Lyapunov-type control for the mean evolution of the stochastic algorithm $(X_n, Y_n)_{n\geqslant 1}$ described by (2.4). This result is based on the important function $(x, y) \longmapsto V_n(x, y)$ that depends on two parameters $(a, b) \in \mathbb{R}_+^2$ defined by:

$$V_n(x,y) = (a + br_{n-1})f(x) + \frac{a}{2r_{n-1}}\|y\|^2 - b\langle \nabla f(x), y\rangle. \tag{3.1}$$

We will show that V_n plays the role of a (potentially time-dependent) Lyapunov function for the sequence $(X_n, Y_n)_{n\geqslant 1}$. The construction of V_n shares a lot of similarity with other Lyapunov functions built to control second-order systems. If the two first terms are classical and generate a $-\|y\|^2$ term, the last one is more specific to hypo-coercive dynamics and was already used in [Har91]. Recent works fruitfully exploit this kind of Lyapunov function (see, among others, the kinetic Fokker-Planck equations in [Vil09] and the memory gradient diffusion in [GP14]). This function is obtained by the introduction of some Lie brackets of differential operators, leading to the presence of $\langle \nabla f(x), y\rangle$ that generates a mean reverting effect on the variable x.

With the help of V_n, we derive the first important result on $(X_n, Y_n)_{n\geqslant 1}$. The proof is deferred to the appendix paragraph in Section A.1.

Proposition 3.1. *If* $(\mathbf{H}_{\boldsymbol{\sigma},\mathbf{2}})$ *and* $(\mathbf{H}_s)$ *hold and* $(r_n)_{n\geqslant 1}$ *satisfies* $(\mathbf{H}_\mathbf{r})$*, then we have:*

(i)
$$\sup_{n\geqslant 1}\left(\mathbb{E}[f(X_n)] + \frac{1}{r_n}\mathbb{E}[\|Y_n\|^2]\right) < +\infty$$

(ii) $(V_n(X_n, Y_n))_{n\geqslant 1}$ *is a.s.-convergent to* $V_\infty \in \mathbb{R}_+$*. In particular,* $(X_n)_{n\geqslant 1}$ *and* $(Y_n/\sqrt{r_n})_{n\geqslant 1}$ *are a.s.-bounded.*

(iii) $\displaystyle\sum_{n\geqslant 1} \gamma_{n+1} r_n \left(\frac{\|Y_n\|^2}{r_n} + \|\nabla f(X_n)\|^2\right) < +\infty$ *a.s.*

(iv) $(Y_n/\sqrt{r_n})_{n\geqslant 0}$ *tends to* 0 *since* $n \to +\infty$ *and every limit point of* $(X_n)_{n\geqslant 0}$ *belong to* $\{x, \nabla f(x) = 0\}$. *Furthermore, if for any* z, $\{x, f(x) = z\} \cap \{x, \nabla f(x) = 0\}$ *is locally finite,* $(X_n)_{n\geqslant 0}$ *converges towards a critical point of* f *a.s.*

Note that if $(\mathbf{H_r})$ holds, then *(iii)* provides a strong repelling effect on the system (x, y) because in that case, $\sum \gamma_{n+1} r_n = +\infty$. This makes it possible to obtain a more precise a.s. convergence result, which is stated in *(iv)*.

3.2. Convergence to a local minimum

3.2.1. Nature of the result and theoretical difficulties

To motivate the next theoretical study, we address the result of Proposition 3.1. We have shown in this corollary the almost sure convergence of (2.4) towards a point of the form $(x_\infty, 0)$ in both exponential and polynomial cases where x_∞ is a critical point of f. This result is obtained under very weak assumptions on f and on the noise $(\Delta M_{n+1})_{n\geqslant 1}$ and is rather close to Theorems 3-4 of [YLL16] (obtained within a different framework). Unfortunately, it only provides a very partial answer to the problem of minimizing f because nothing is said about the stability of the limit of the sequence $(X_n)_{n\geqslant 0}$ by Proposition 3.1: the attained critical point may be a local maximum, a saddle point or a local minimum. This result is made more precise below and we establish some sufficient guarantees for the a.s. convergence of (X_n) towards a *minimum* of f, even if f possesses some local traps. To derive this important and stronger key result, we need to introduce the additional assumption $(\mathbf{H}_\mathcal{E})$, which translates an elliptic behavior of the martingale noise $(\Delta M_{n+1})_{n\geqslant 1}$ and we have to overcome several difficulties.

- The proof follows the approach described in [BD96] and [Ben06] but requires some careful adaptations because of the hypo-elliptic noise of the algorithm (there is no noise on the x-component) for both the exponentially and polynomially-weighted memory. Therefore, even though the global probabilistic argument relies on the approach of [Ben06], the estimations of the exit times of the neighborhoods of unstable equilibria (local maxima or saddle points) deserve a particular study because of the hypo-ellipticity.
- Moreover, the linearization of the inhomogeneous drift around a critical point of f in the polynomial memory case is a supplementary difficulty we need to bypass because in this situation, the algorithm $(X_n, Y_n)_{n\geqslant 1}$ does not evolve at the same time-scale on the two coordinates. We should emphasize that one should think of the use of the recent contributions of [Bor97, Bor08] on dynamical systems with two different time scales. Let us briefly discuss on the approach developed in these works: [Bor08] investigates the behaviour of

$$\begin{aligned}(x_{n+1}, y_{n+1}) &= (x_n, y_n) \\ &\quad + \left(a_n[h(x_n, y_n) + \Delta M^1_{n+1}], b_n[g(x_n, y_n) + \Delta M^2_{n+1}]\right), \end{aligned} \tag{3.2}$$

where $b_n = o(a_n)$. This is exactly our setting in the polynomial memory case since $\gamma_n r_n = o(\gamma_n)$. Unfortunately, [Bor08] assumes that the differential equation $\dot{x} = h(x, y)$ has a globally asymptotically stable equilibrium for any given and fixed $y \in \mathbb{R}^d$, which is false in our case since $\dot{x} = -y$ is solved by $x_t = x_0 - ty$ and has no stable equilibrium except when $y = 0$. Therefore, it is not possible to use the former works of [Bor97, Bor08] in our polynomial memory case.

Note that some recent works on stochastic algorithms (see, *e.g.*, [LSJR16]) deal with the convergence to minimizers of f of *deterministic* gradient descent with a randomized initialization. In our case, we will obtain a rather different result because of the randomization of the algorithm at each iteration. Note, however that the main ingredient of the proofs below will be the stable manifold theorem (the Poincaré Lemma on stable/unstable hyperbolic points of [Poi86]) and its consequence around hyperbolic points. This geometrical result is also used in [LSJR16].

3.2.2. Exponential memory $r_n = r > 0$

The exponential memory case may be (almost) seen as an application of Theorem 1 of [BD96]. More precisely, if $Z_n = (X_n, Y_n)$ and $h(x, y) = (-y, r\nabla f(x) - ry)$, then the underlying stochastic algorithm may be written as:

$$Z_{n+1} = Z_n + \gamma_n h(Z_n) + \gamma_n \Delta M_n,$$

When $r_n = r > 0$ (exponential memory), Proposition 3.1 applies and $Z_n \xrightarrow{a.s.} Z_\infty = (X_\infty, 0)$ where X_∞ is a critical point of f. For the analysis of the dynamics around a critical point of the drift, the critical poinf of f is denoted x_0 and we can linearize the drift around $(x_0, 0) \in \mathbb{R}^d \times \mathbb{R}^d$ as:

$$h(x, y) = \begin{pmatrix} 0 & -I_d \\ rD^2(f)(x_0) & -rI_d \end{pmatrix} \begin{pmatrix} x - x_0 \\ y \end{pmatrix} + O(\|x - x_0\|^2),$$

where I_d is the $d \times d$ identity-squared matrix and $D^2(f)(x_0)$ is the Hessian matrix of f at point x_0. When x_0 is not a local minimum of f, the spectral decomposition of $D^2(f)(x_0)$ leads to the spectral decomposition:

$$\exists P \in \mathcal{O}_d(\mathbb{R}) \qquad D^2(f)(x_0) = P^{-1}\Lambda P,$$

where Λ is a diagonal matrix with at least one negative eigenvalue $\lambda < 0$. Considering now $\breve{Z}_n = (\breve{X}_n, \breve{Y}_n)$ where $\breve{X}_n = PX_n$ and $\breve{Y}_n = PY_n$, we have:

$$\breve{Z}_{n+1} = \breve{Z}_n + \gamma_n \breve{h}(\breve{Z}_n) + \gamma_n P \Delta M_n,$$

where $\breve{h}$ may be linearized as:

$$\breve{h}(\breve{x}, \breve{y}) = \begin{pmatrix} 0 & -I_d \\ r\Lambda & -rI_d \end{pmatrix} \begin{pmatrix} \breve{x} - \breve{x}_0 \\ \breve{y} \end{pmatrix} + O(\|\breve{x} - \breve{x}_0\|^2) \qquad \text{where} \qquad \breve{x}_0 = Px_0.$$

In particular, if e_λ is an eigenvector associated with the eigenvalue $\lambda < 0$ of $D^2 f(x_0)$, we can see that the linearization of $\tilde{h}$ on the space $Span(e_\lambda) \otimes (1, 0, \ldots, 0)$ acts as:

$$A_{\lambda,r} = \begin{pmatrix} 0 & -1 \\ r\lambda & -r \end{pmatrix}.$$

Its spectrum is $Sp(A_{\lambda,r}) = -\frac{r}{2} \pm \sqrt{\frac{r^2}{4} - r\lambda}$. The important fact is that when $\lambda < 0$, the eigenvalue $-\frac{r}{2} + \sqrt{\frac{r^2}{4} - r\lambda}$ is positive and whose corresponding eigenspace is $E_\lambda^+ = \left(1, \frac{1}{2} - \sqrt{\frac{1}{4} - \lambda/r}\right)$. In the initial space $\mathbb{R}^d \times \mathbb{R}^d$ (without applying the change of basis through $P \otimes P$), the corresponding eigenvector is:

$$e_\lambda^+ = e_\lambda \otimes \left(\frac{1}{2} - \sqrt{\frac{1}{4} - \lambda/r}\right) e_\lambda$$

Consequently, when x_0 is not a local minimum of f, it generates a hyperbolic equilibrium of h and we can apply the "general" local trap Theorem 1 of [BD96]. If $\Pi_{E_\lambda^+}$ denotes the projection on the eigenspace $Span(e_\lambda^+)$, then the noise in the direction E_λ^+ is:

$$\xi_n^+ = \Pi_{E_\lambda^+}(0, \Delta M_n) = \frac{\langle \Delta M_n, e_\lambda \rangle}{\|e_\lambda\|^2} e_\lambda.$$

Now, Assumption $(\mathbf{H}_{\boldsymbol{\mathcal{E}}})$ implies that:

$$\liminf_{n \longrightarrow +\infty} \mathbb{E} \left\| \Pi_{E_\lambda^+}(0, \Delta M_n) \right\| \geqslant c_{e_\lambda} > 0.$$

We can then apply Theorem 1 of [BD96] and conclude the following result.

Theorem 3.1. *If* $(\mathbf{H}_{\boldsymbol{\sigma,2}})$, $(\mathbf{H}_s)$ *and* $(\mathbf{H}_{\mathcal{E}})$ *hold and* $r_n = r$, *then* X_n *a.s. converges towards a local minimum of* f.

3.2.3. Polynomial memory $r_n = r\Gamma_n^{-1} \longrightarrow 0$

We introduce a key normalization of the speed coordinate and define the rescaled process:

$$\widetilde{X}_n = X_n \qquad \text{and} \qquad \widetilde{Y}_n = \sqrt{\Gamma_n} Y_n.$$

We can note that $\widetilde{Y}_n = \sqrt{r} Y_n r_n^{-1/2}$ and the important conclusion brought by (iv) of Proposition 3.1 is that $(\widetilde{X}_n, \widetilde{Y}_n) \xrightarrow{a.s.} (X_\infty, 0)$ still holds (under the assumptions of Proposition 3.1) We can write the recursive upgrade of the couple $(\widetilde{X}_n, \widetilde{Y}_n)$. The evolution of $(\widetilde{X}_n)_{n \geqslant 0}$ is easy to write: $\widetilde{X}_{n+1} = \widetilde{X}_n - \frac{\gamma_{n+1}}{\sqrt{\Gamma_n}} \widetilde{Y}_n$. The recursive formula satisfied by $(\widetilde{Y}_n)_{n \geqslant 0}$ is:

$$\begin{aligned}\widetilde{Y}_{n+1} &= \sqrt{\Gamma_{n+1}}\,[Y_n + \gamma_{n+1} r_{n+1}\,(\nabla f(X_n) - Y_n + \Delta M_{n+1})] \\ &= \frac{\sqrt{\Gamma_{n+1}}}{\sqrt{\Gamma_n}}\widetilde{Y}_n + r\frac{\gamma_{n+1}}{\sqrt{\Gamma_n}} \times \frac{\sqrt{\Gamma_{n+1}}}{\sqrt{\Gamma_n}}\nabla f(\widetilde{X}_n) - r\frac{\gamma_{n+1}}{\sqrt{\Gamma_n}} \times \frac{\sqrt{\Gamma_{n+1}}}{\Gamma_n}\widetilde{Y}_n \\ &\quad + r\frac{\gamma_{n+1}}{\sqrt{\Gamma_n}} \times \frac{\sqrt{\Gamma_{n+1}}}{\sqrt{\Gamma_n}}\Delta M_{n+1}\end{aligned}$$

Hence, the couple $(\widetilde{X}_n, \widetilde{Y}_n)$ evolves as an almost standard stochastic algorithm, whose step size is $\widetilde{\gamma}_{n+1} = \gamma_{n+1}\Gamma_n^{-1/2}$:

$$\begin{cases} \widetilde{X}_{n+1} = \widetilde{X}_n - \widetilde{\gamma}_{n+1}\widetilde{Y}_n \\ \widetilde{Y}_{n+1} = \widetilde{Y}_n + r\widetilde{\gamma}_{n+1}\nabla f(\widetilde{X}_n) + \widetilde{\gamma}_{n+1} q_{n+1}\Delta M_{n+1} + \widetilde{\gamma}_{n+1} U_{n+1}, \end{cases} \tag{3.3}$$

where $q_{n+1} = \sqrt{\Gamma_{n+1}/\Gamma_n} = 1 + o(n^{-1})$ as $n \longrightarrow +\infty$ and $(U_{n+1})_{n\geqslant 1}$ is defined by:

$$U_{n+1} = \frac{1/2 - r q_{n+1} + o(n^{-1})}{\sqrt{\Gamma_n}}\widetilde{Y}_n + r(q_{n+1} - 1)\nabla f(\widetilde{X}_n).$$

This dynamical system is related to the deterministic one $\begin{cases} \dot{x}_t = -y_t \\ \dot{y}_t = r\nabla f(x_t) \end{cases}$ or equivalently:

$$\dot{z}_t = F(z_t) \quad \text{with} \quad F(z) = F(x,y) = (-y, r\nabla f(x)). \tag{3.4}$$

It is easy to see that when x_∞ is a local maximum of f, then the above drift is unstable near $z_\infty = (x_\infty, 0)$. Unfortunately, Theorem 1 of [BD96] cannot be applied because of the size of the remainder terms involved in (3.3) and the a.s. convergence of $(X_n, Y_n)_{n\geqslant 0}$ requires further investigation. From [Ben06], we borrow a tractable construction of a "Lyapunov" function η in the neighborhood of each hyperbolic point, which translates a mean repelling effect of the unstable points. This construction still relies on the Poincaré Lemma (see [Poi86] and [Har82] for a recent reference). Again, in the neighborhood of any hyperbolic point, we will treat the projection Π_+ as a projection on the unstable manifold.

Proposition 3.2 ([Ben06])**.** *For any local maximum point x_∞ of f, a compact neighborhood $\mathcal{N}$ of $z_\infty = (x_\infty, 0)$ and a positive function $\eta \in \mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}_+^\star)$ exist such that:*

(*i*) *$\forall z = (x,y) \in \mathcal{N}, D\eta(z) : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}^d \times \mathbb{R}^d$ is Lipschitz, convex and positively homogeneous.*

(*ii*) *Two constants $k > 0$ and $c_1 > 0$ and a neighborhood U of $(0,0)$ exist such that:*

$$\forall z \in \mathcal{N} \quad \forall u \in U \qquad \eta(z+u) \geqslant \eta(z) + \langle D\eta(z), u\rangle - k\|u\|^2,$$

and if $\lfloor\,\rfloor_+$ denotes the positive part:

$$\forall z \in \mathcal{N} \quad \forall u \in U \qquad \lfloor D\eta(z)(u)\rfloor_+ \geqslant c_1\|\Pi_+(u)\|.$$

(iii) A positive constant κ exists such that:

$$\forall z \in \mathcal{N} \qquad \langle D\eta(z), F(z)\rangle \geqslant \kappa\eta(z)$$

When $d = 1$, it is possible to check that if λ is a negative eigenvalue of the Hessian of f around a local maximum x_∞, then the drift may be linearized in $(-y, \lambda(x - x_\infty))$ and a reasonable approximation of η is given by $\eta(x, y) = \frac{1}{2}\|y - \sqrt{-\lambda}x\|^2$. Nevertheless, the situation is more involved in higher dimensions and the construction of the function η relies on the Poincare stable manifold theorem. We are now able to state the next important result.

Theorem 3.2. *Assume that the noise satisfies* $(\mathbf{H}_{\boldsymbol{Gauss,\sigma}})$ *and* $(\mathbf{H}_{\mathcal{E}})$*, that the function satisfies* $(\mathbf{H}_s)$*, and that* $\gamma_n = \gamma n^{-\beta}$ *with* $\beta < 1/3$*, then* $(X_n)_{n\geqslant 0}$ *a.s. converges towards a local minimum of* f*.*

The proof relies on an argument of [Pem90, Ben06] even though it requires major modifications to deal with the time inhomogeneity of the process and the unbounded noise, which are assumed in these previous works. We denote $\mathcal{N}$ as any neighborhood of z_∞ and consider any integer $n_0 \in \mathbb{N}$. We then introduce $\widetilde{Z}_n = (\widetilde{X}_n, \widetilde{Y}_n)$ and the stopping time:

$$T := \inf\left\{n \geqslant n_0 \,:\, \widetilde{Z}_n \notin \mathcal{N}\right\}.$$

We will show that $\mathbb{P}(T < +\infty) = 1$, which implies the conclusion. We introduce two sequences $(\Omega_n)_{n\geqslant n_0}$ and $(S_n)_{n\geqslant n_0}$:

$$\Omega_{n+1} = [\eta(\widetilde{Z}_{n+1}) - \eta(\widetilde{Z}_n)]\mathbf{1}_{n<T} + \tilde{\gamma}_{n+1}\mathbf{1}_{n\geqslant T} \quad \text{and} \quad S_n = \eta(\tilde{Z}_{n_0}) + \sum_{k=n_0+1}^{n} \Omega_k. \tag{3.5}$$

Note that the construction of η implies that $z \longmapsto D\eta(z)$ is Lipschitz, so that the following inequality holds:

$$\eta(z + u) - \eta(z) \geqslant \langle D\eta(z), u\rangle - \frac{\|D\eta\|_{Lip}\|u\|^2}{2}.$$

This inequality provides some information when u is small. In the meantime, η is positive so that:

$$\begin{aligned} &\forall \alpha \in (0, 1] \quad \exists k_\alpha > 0 \quad \forall (z, u) \in \mathcal{N} \times \mathbb{R}^d \\ &\eta(z + u) - \eta(z) \geqslant \langle D\eta(z), u\rangle - k_\alpha\|u\|^{1+\alpha} \end{aligned} \tag{3.6}$$

The family of inequalities described in (3.6) will be used with an appropriate value of α in the next result.

Proposition 3.3. *The random variables* $(\Omega_n)_{n\geqslant 0}$ *satisfy the following conditions:*

(i) A constant c exists such that:

$$\mathbb{E}[\Omega_{n+1}^2|\mathcal{F}_n] \leqslant c\tilde{\gamma}_{n+1}^2$$

(*ii*) *A sequence* $(\epsilon_n)_{n\geqslant 0}$ *exists such that:*

$$\mathbf{1}_{S_n\geqslant\epsilon_n}\mathbb{E}[\Omega_{n+1}|\mathcal{F}_n]\geqslant 0,$$

with $\epsilon_n\sim cn^{-(1-\alpha)/2}$ *for a large enough* c *and* $\alpha=(1-\beta)/(1+\beta)$.

(*iii*) *Assume that* $\beta<\frac{1}{3}$, *then* $(S_n^2)_{n\geqslant 0}$ *has a submartingale increment:*

$$\mathbb{E}[S_{n+1}^2-S_n^2|\mathcal{F}_n]\geqslant a\tilde{\gamma}_{n+1}^2$$

for a small enough constant a.

The proof of this technical proposition is deferred to the appendix paragraph A.2 We use now the key estimations derived from Proposition 3.3 to obtain the proof of Theorem 3.2.

Proof of Theorem 3.2: The proof is split into three parts. We consider:

$$S_n=S_0+\sum_{k=1}^{n}\Omega_k\qquad\text{and define}\qquad\delta_n=\sum_{i\geqslant n}\tilde{\gamma}_i^2.$$

In our case, we have chosen $\beta\in(0,1/3)$ and we can check that:

$$\tilde{\gamma}_n\sim n^{-(1+\beta)/2}\quad\text{so that}\quad\delta_n\sim n^{-\beta}. \tag{3.7}$$

We consider the sequence ϵ_n defined in Proposition 3.3:

$$\epsilon_n\sim\Gamma_n^{-1/2}\sim\tilde{\gamma}_{n+1}^{\alpha}\quad\text{with}\quad\alpha=\frac{1-\beta}{1+\beta}>1/2.$$

In this case, we have:

$$\epsilon_n=n^{-(1-\beta)/2}=o(n^{-\beta/2})=o(\sqrt{\delta_n})\quad\text{because}\quad\beta<1/3<1/2.$$

The proof now proceeds by considering the sequential crossings $S_n\leqslant c\sqrt{\delta_n}$ and $S_n\geqslant c\sqrt{\delta_n}$ for a suitable value of c.

Step 1: S_n *becomes greater than* $\sqrt{b\delta_n}$ *with a positive probability.*

For a given constant b and a positive $n\in\mathbb{N}$, we introduce the stopping time:

$$\mathcal{T}=\inf\left\{i\geqslant n\,:\,S_i\geqslant\sqrt{b\delta_i}\right\},$$

and we show that an $\epsilon>0$ exists such that $\mathbb{P}\left(\mathcal{T}<\infty\right)\geqslant 1-\epsilon$. For a given by (*iii*) of Proposition 3.3, we consider:

$$\mathcal{M}_k=S_k^2-a\sum_{i=0}^{k}\tilde{\gamma}_i^2.$$

$(\mathcal{M}_k)_{k\geqslant n}$ is a submartingale, so that $(\mathcal{M}_{k\wedge\mathcal{T}})_{k\geqslant n}$ is also a stopped submartingale. This yields:

$$\mathbb{E}\left[S_{m\wedge\mathcal{T}}^2-S_n^2|\mathcal{F}_n\right]\geqslant a\mathbb{E}\left[\sum_{n+1}^{m\wedge\mathcal{T}}\tilde{\gamma}_i^2|\mathcal{F}_n\right]\geqslant a\left(\sum_{n+1}^{m}\tilde{\gamma}_i^2\right)\mathbb{P}\left(\mathcal{T}>m|\mathcal{F}_n\right). \tag{3.8}$$

In the meantime, we can decompose $S^2_{m\wedge\mathcal{T}} - S^2_n$ into:

$$\begin{aligned} S^2_{m\wedge\mathcal{T}} - S^2_n &= S^2_{m\wedge\mathcal{T}} - S^2_{m\wedge\mathcal{T}-1} + S^2_{m\wedge\mathcal{T}-1} - S^2_n \\ &\leqslant 2S_{m\wedge\mathcal{T}-1}\Omega_{m\wedge\mathcal{T}} + \Omega^2_{m\wedge\mathcal{T}} + S^2_{m\wedge\mathcal{T}-1} \\ &\leqslant 2S^2_{m\wedge\mathcal{T}-1} + 2\Omega^2_{m\wedge\mathcal{T}} \\ &\leqslant 2b\delta_{m\wedge\mathcal{T}-1} + 2\Omega^2_{m\wedge\mathcal{T}}. \end{aligned}$$

Since $(\delta_k)_{k\geqslant n}$ is decreasing, we then have $\delta_{m\wedge\mathcal{T}-1} \leqslant \delta_n$. We then study the remaining term. We can use Equation (3.3) and the Lipschitz continuity of η over the neighborhood $\mathcal{N}$ (before time T) to obtain a large enough C such that:

$$\begin{aligned} \Omega^2_{m\wedge\mathcal{T}} &= \Omega^2_{m\wedge\mathcal{T}}\left[\mathbf{1}_{m\wedge\mathcal{T}-1<T} + \mathbf{1}_{m\wedge\mathcal{T}-1\geqslant T}\right] \\ &= \left[\eta(\tilde{Z}_{m\wedge\mathcal{T}}) - \eta(\tilde{Z}_{m\wedge\mathcal{T}-1})\right]^2 \mathbf{1}_{m\wedge\mathcal{T}-1<T} + \tilde{\gamma}^2_{m\wedge\mathcal{T}}\mathbf{1}_{m\wedge\mathcal{T}-1\geqslant T} \\ &\leqslant C[\tilde{\gamma}^2_{m\wedge\mathcal{T}} + \tilde{\gamma}^2_{m\wedge\mathcal{T}}\|\Delta M_{m\wedge\mathcal{T}}\|^2]. \end{aligned}$$

However, nothing more is known about the stopped process $\|\Delta M_{m\wedge\mathcal{T}}\|^2$ and we are forced to use:

$$\mathbb{E}\left[S^2_{m\wedge\mathcal{T}} - S^2_n|\mathcal{F}_n\right] \leqslant 2b\delta_n + 2C\left[\tilde{\gamma}^2_n + \mathbb{E}\left[\sup_{k\geqslant n}\tilde{\gamma}^2_k\|\Delta M_k\|^2\right]\right].$$

Given that all ΔM_k are independent sub-Gaussian random variables that satisfy Inequality (A.6), we can use Theorem A.1 and obtain that a constant C large enough exists such that for any $\epsilon > 0$:

$$\mathbb{E}\left[S^2_{m\wedge\mathcal{T}} - S^2_n|\mathcal{F}_n\right] \leqslant 2b\delta_n + 2C\tilde{\gamma}^2_n\log(\tilde{\gamma}_n^{-2}). \tag{3.9}$$

We can plug the estimate (3.9) into Inequality (3.8) to obtain:

$$\mathbb{P}\left(\mathcal{T} > m|\mathcal{F}_n\right) \leqslant \frac{2b\delta_n + 2C\tilde{\gamma}^2_n\log(\tilde{\gamma}_n^{-2})}{a\sum_{i=n+1}^m \tilde{\gamma}^2_i}.$$

Letting $m \longrightarrow +\infty$, we deduce that:

$$\mathbb{P}\left(\mathcal{T} = \infty|\mathcal{F}_n\right) \leqslant \frac{2b}{a} + \frac{2C\tilde{\gamma}^2_n\log(\tilde{\gamma}_n^{-2})}{a\delta_n}.$$

According to the calibration (3.7), we have $\tilde{\gamma}^2_n\log(\tilde{\gamma}_n^{-2}) = o(\delta_n)$. Consequently, we can choose n large enough such that:

$$\mathbb{P}\left(\mathcal{T} < \infty|\mathcal{F}_n\right) \geqslant 1 - \frac{3b}{a}. \qquad \diamond$$

Step 2: The sequence $(S_k)_{k\geqslant n}$ may remain larger than $\sqrt{b/2\delta_n}$ with a positive probability.

We introduce the stopping time $\mathcal{S}$ and the event $E_n \in \mathcal{F}_n$:

$$\mathcal{S} = \inf\{i \geqslant n \,:\, S_i < \frac{\sqrt{b}}{2}\sqrt{\delta_n}\} \qquad \text{and} \qquad E_n = \left\{S_n \geqslant \sqrt{b}\sqrt{\delta_n}\right\}.$$

Since the sequence $(\delta_i)_{i\geqslant n}$ is non-increasing, (ii) of Proposition 3.3 yields:

$$\begin{aligned}\mathbb{E}\left[S_{(i+1)\wedge\mathcal{S}} - S_{i\wedge\mathcal{S}}|\mathcal{F}_i\right] &= \mathbf{1}_{\mathcal{S}>i}\mathbb{E}\left[S_{i+1}-S_i|\mathcal{F}_i\right]\\ &= \mathbf{1}_{\mathcal{S}>i}\mathbf{1}_{S_i\geqslant\sqrt{b/2\delta_n}}\mathbb{E}\left[S_{i+1}-S_i|\mathcal{F}_i\right]\\ &\geqslant \mathbf{1}_{\mathcal{S}\geqslant i}\mathbf{1}_{S_i\geqslant\sqrt{b/2\delta_i}}\mathbb{E}\left[X_{i+1}|\mathcal{F}_i\right]\\ &\geqslant \mathbf{1}_{\mathcal{S}\geqslant i}\mathbf{1}_{S_i\geqslant\epsilon_i}\mathbb{E}\left[X_{i+1}|\mathcal{F}_i\right]\geqslant 0.\end{aligned}$$

Hence, $(S_{i\wedge\mathcal{S}})_{i\geqslant n}$ is a submartingale and the Doob decomposition reads $S_{i\wedge\mathcal{S}} = M_i + I_i$ where $(M_i)_{i\geqslant n}$ is a Martingale and (I_i) is a predictable increasing process such that $I_n = 0$. Hence,

$$\mathbb{P}(\mathcal{S}=\infty|\mathcal{F}_n) = \mathbb{P}_{|\mathcal{F}_n}\left(\forall i\geqslant n\,:\,S_i\geqslant\frac{\sqrt{b}}{2}\sqrt{\delta_n}\right)\geqslant\mathbb{P}_{|\mathcal{F}_n}\left(\forall i\geqslant n\,:\,M_i\geqslant\frac{\sqrt{b}}{2}\sqrt{\delta_n}\right)$$

On the event E_n, $S_n = M_n \geqslant \sqrt{b}\sqrt{\delta_n}$ so that $M_i - M_n \leqslant M_i - \sqrt{b}\sqrt{\delta_n}$. Therefore:

$$\mathbb{P}\left(\forall i\geqslant n\,:\,M_i\geqslant\frac{\sqrt{b}}{2}\sqrt{\delta_n}\,|\,\mathcal{F}_n\right)\mathbf{1}_{E_n}\geqslant\mathbb{P}\left(\forall i\geqslant n\,:\,M_i-M_n\geqslant-\frac{\sqrt{b}}{2}\sqrt{\delta_n}\,|\,\mathcal{F}_n\right)\mathbf{1}_{E_n}.$$

The rest of the proof follows a standard martingale argument:

$$\begin{aligned}\mathbb{E}\left((M_i-M_n)^2|\mathcal{F}_n\right) &= \sum_{j=n}^{i-1}\mathbb{E}\left((M_{j+1}-M_j)^2|\mathcal{F}_n\right)\\ &= \sum_{j=n}^{i-1}\mathbb{E}\left(\mathbb{E}\left((M_{j+1}-M_j)^2|\mathcal{F}_j\right)|\mathcal{F}_n\right)\\ &= \sum_{j=n}^{i-1}\mathbb{E}\left(\mathbb{E}\left((S_{j+1}-S_j)^2|\mathcal{F}_j\right)-(I_{j+1}-I_j)^2|\mathcal{F}_n\right)\\ &\leqslant \sum_{j=n}^{i-1}\mathbb{E}\left((S_{j+1}-S_j)^2|\mathcal{F}_n\right)\leqslant\sum_{j=n}^{i-1}\mathbb{E}\left(\Omega_{j+1}^2|\mathcal{F}_n\right)\\ &\leqslant c\sum_{j=n}^{i}\tilde{\gamma}_{j+1}^2\leqslant c\delta_n.\end{aligned}$$

where we used the upper bound given by (i) of Proposition 3.3 in the last line. Now, the Doob inequality implies that:

$$\begin{aligned}\mathbb{P}(\inf_{n\leqslant i\leqslant m}(M_i-M_n)\leqslant -s|\mathcal{F}_n) &= \mathbb{P}(\inf_{n\leqslant i\leqslant m}(M_i-M_n-t)\leqslant -s-t|\mathcal{F}_n)\\ &\leqslant \mathbb{P}(\sup_{n\leqslant i\leqslant m}|M_i-M_n-t|\leqslant s+t|\mathcal{F}_n)\\ &\leqslant \frac{E\left((M_m-M_n-t)^2|\mathcal{F}_n\right)}{(s+t)^2}\\ &= \frac{E\left((M_m-M_n)^2|\mathcal{F}_n\right)+t^2}{(s+t)^2}=\frac{c\delta_n+t^2}{(s+t)^2}.\end{aligned}$$

We apply this inequality with $s = \frac{\sqrt{b}}{2}\sqrt{\delta_n}$ and use $(s+t)^2 \leqslant (1+\vartheta)s^2+(1+\vartheta^{-1})t^2$ for any $\vartheta > 0$. It leads to:

$$\mathbb{P}\left(\inf_{n\leqslant i\leqslant m}(M_i - M_n) \leqslant -\frac{\sqrt{b}}{2}\sqrt{\delta_n}|\mathcal{F}_n\right) \leqslant \frac{c\delta_n + t^2}{(1+\vartheta)b\delta_n/4 + (1+\vartheta^{-1})t^2}.$$

We now choose $\vartheta = 4c/b$, $t = \sqrt{\delta_n}$ and deduce that:

$$\mathbb{P}\left(\inf_{n\leqslant i\leqslant m}(M_i - M_n) \leqslant -\frac{\sqrt{b}}{2}\sqrt{\delta_n}|\mathcal{F}_n\right) \leqslant \frac{c+1}{c+1+b/4c}.$$

Consequently, we deduce that:

$$\begin{aligned}\mathbb{P}(\mathcal{S}=\infty|\mathcal{F}_n)\mathbf{1}_{E_n} &\geqslant \mathbb{P}_{|\mathcal{F}_n}\left(\forall i \geqslant n : M_i \geqslant \frac{\sqrt{b}}{2}\sqrt{\delta_n}\right)\mathbf{1}_{E_n} \\ &\geqslant \left(1 - \frac{c+1}{c+1+b/4c}\right)\mathbf{1}_{E_n} = \frac{b}{b+4c+4c^2}\mathbf{1}_{E_n} \qquad \diamond\end{aligned}$$

Step 3: $(S_n)_{n\geqslant 0}$ does not converge to 0 with probability 1.

We denote $\mathcal{G}$ as the event that $(S_n)_{n\geqslant 0}$ does not converge to 0. For any integer n, we have the inclusion:

$$\{\mathcal{S} = +\infty\} = \left\{\forall i \geqslant n \,:\, S_i \geqslant \sqrt{b/4}\sqrt{\delta_n}\right\} \subset \mathcal{G},$$

which implies:

$$\mathbb{E}[\mathbf{1}_{\mathcal{G}}|\mathcal{F}_i]\mathbf{1}_{\mathcal{T}=i} = \mathbb{E}[\mathbf{1}_{\mathcal{G}}|\mathcal{F}_i]\mathbf{1}_{\mathcal{T}=i}\mathbf{1}_{E_i} \geqslant \frac{b}{b+4c+4c^2}\mathbf{1}_{\mathcal{T}=i}\mathbf{1}_{E_i} = \frac{b}{b+4c+4c^2}\mathbf{1}_{\mathcal{T}=i}$$

Hence,

$$\begin{aligned}\mathbb{E}[\mathbf{1}_{\mathcal{G}}|\mathcal{F}_n] &= \sum_{i\geqslant n}\mathbb{E}[\mathbf{1}_{\mathcal{G}}\mathbf{1}_{\mathcal{T}=i}|\mathcal{F}_n] = \mathbb{E}\left[\mathbb{E}[\mathbf{1}_{\mathcal{G}}|\mathcal{F}_i]\mathbf{1}_{\mathcal{T}=i}\,|\mathcal{F}_n\right] \\ &\geqslant \frac{b}{b+4c+4c^2}\sum_{i\geqslant n}\mathbb{E}\left[\mathbf{1}_{\mathcal{T}=i}\,|\mathcal{F}_n\right] \\ &\geqslant \frac{b}{b+4c+4c^2}\mathbb{P}\left(\mathcal{T} < +\infty|\mathcal{F}_n\right) \geqslant \frac{b}{b+4c+4c^2}\left(1 - \frac{3b}{a}\right) > 0.\end{aligned}$$

Since $\mathbf{1}_{\mathcal{G}} \in \mathcal{F}_\infty$, we have $\lim_{n\longrightarrow+\infty}\mathbb{E}[\mathbf{1}_{\mathcal{G}}|\mathcal{F}_n] = \mathbf{1}_{\mathcal{G}}$. The previous lower bound implies that $\mathcal{G}$ almost surely holds. $\diamond$

Conclusion of the proof: *The stochastic algorithm does not converge to a local trap.*

Consider $\mathcal{N}$ a neighborhood of a local maximum of f, and its associated function η given by Proposition 3.2. We then consider the random variables $(\Omega_n)_{n\geqslant 0}$

and $(S_n)_{n\geqslant 0}$. We have seen that S_n does not converge to 0 with probability 1. We define:

$$\mathcal{T}_{\mathcal{N}} := \inf\left\{n \geqslant 0 \,:\, \widetilde{Z}_n \notin \mathcal{N}\right\}.$$

and assume that $\mathcal{T}_{\mathcal{N}} = +\infty$. In that case, we always have:

$$\Omega_{n+1} = \eta(\widetilde{Z}_{n+1}) - \eta(\widetilde{Z}_n) \qquad \text{and} \qquad S_n = \eta(\widetilde{Z}_n).$$

The limit set of $(\widetilde{Z}_n)_{n\geqslant 0}$ is a non empty compact subset of $\mathcal{N}$, which is left invariant by the flow $(\Phi_t)_{t\geqslant 0}$ of the O.D.E. whose drift is F. Now, consider z in $\overline{(\widetilde{Z}_n)_{n\geqslant 0}}$ and apply (iii) of Proposition 3.2. We then have $\eta(\Phi_t(z)) \geqslant e^{\kappa t}\eta(y)$. Since $\eta(\Phi_t(z)) \leqslant \sup_{\mathcal{N}} \eta$, we therefore deduce that $\eta(z) = 0$. Hence, the unique limiting value for $(S_n)_{n\geqslant 0}$ is zero, meaning that $S_n \longrightarrow 0$ as $n \longrightarrow +\infty$. However, we have seen in Step 3 that S_n does not converge to 0 with probability 1. Therefore, $\mathbb{P}(\mathcal{T}_{\mathcal{N}} = +\infty) = 0$ and the process does not converge towards a local maximum of f with probability 1. □

4. Convergence rates for strongly convex functions

This section focuses on the convergence rates of algorithm (2.4) according to the step-size $\gamma_n = \gamma n^{-\beta}$ for λ-strongly convex function f with a L-Lipschitz gradient, corresponding to the assumptions $(\mathbf{H}_{SC}(\lambda))$ and $(\mathbf{H}_s)$.

4.1. Quadratic case

We first study the benchmark case of a purely quadratic function f, meaning that ∇f is linear. In this case, $f(x) = \frac{1}{2}\|Ax\|^2$ and $\nabla f(x) = Sx$, leading to the following form of the algorithm:

$$\begin{cases} X_{n+1} = X_n - \gamma_{n+1} Y_n \\ Y_{n+1} = Y_n + \gamma_{n+1} r_n (S X_n - Y_n) + \gamma_{n+1} r_n \Delta M_{n+1}, \end{cases} \tag{4.1}$$

where S is a $d \times d$ squared matrix defined by $S = A'A$. The matrix S is assumed to be positive definite with lower bounded eigenvalues, *e.g.*, $Sp(S) \subset [\lambda, +\infty[$ when f is $(\mathbf{H}_{SC}(\lambda))$ with $\lambda > 0$.

4.1.1. Reduction to a two dimensional system

Equation (4.1) may be parameterized in a simpler form using the spectral decomposition of $S = P^{-1}\Lambda P$, where P is orthogonal, and Λ is a diagonal matrix:

$$\forall (i,j) \in \{1 \dots d\}^2 \qquad \Lambda_{i,j} = \lambda_i \delta_{i,j} \geqslant \lambda > 0.$$

Keeping the notation $(\breve{X}_n, \breve{Y}_n)_{n\geqslant 1}$ for the change of basis induced by P, we define $\breve{X}_n = PX_n$ and $\breve{Y}_n = PY_n$ and obtain:

$$\begin{cases} \breve{X}_{n+1} = \breve{X}_n - \gamma_{n+1} \breve{Y}_n \\ \breve{Y}_{n+1} = \breve{Y}_n + \gamma_{n+1} r_n (\Lambda \breve{X}_n - \breve{Y}_n) + \gamma_{n+1} r_n P \Delta M_{n+1}, \end{cases}$$

Since Λ is diagonal, we are now led to study the evolution of d couples of stochastic algorithms:

$$\forall i \in \{1 \dots d\} \qquad \begin{cases} \breve{x}_{n+1}^{(i)} = \breve{x}_n^{(i)} - \gamma_{n+1}\breve{y}_n^{(i)} \\ \breve{y}_{n+1}^{(i)} = \breve{y}_n^{(i)} + \gamma_{n+1} r_n(\lambda_i \breve{x}_n^{(i)} - \breve{y}_n^{(i)}) + \gamma_{n+1} r_n \Delta \widetilde{M}_{n+1}^{(i)}, \end{cases}$$

where we used the notations $\breve{X}_n = (\breve{x}_n^{(i)})_{1 \leqslant i \leqslant d}$ and $\breve{Y}_n = (\breve{y}_n^{(i)})_{1 \leqslant i \leqslant d}$. Consequently, in the quadratic case, the stochastic HBF may be reduced to d couples of 2-dimensional random dynamical systems:

$$\forall i \in \{1, \dots, d\}^2 \qquad \breve{Z}_{n+1}^{(i)} = (I_2 + \gamma_{n+1} C_n^{(i)})\breve{Z}_n^{(i)} + \gamma_{n+1} r_n \Sigma_2 \Delta N_{n+1}^{(i)}, \tag{4.2}$$

where

$$\breve{Z}_n^{(i)} := (\breve{x}_n^{(i)}, \breve{y}_n^{(i)}) \quad \text{and} \quad C_n^{(i)} = \begin{pmatrix} 0 & -1 \\ \lambda^{(i)} r_n & -r_n \end{pmatrix} \qquad \text{and} \qquad \Sigma_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

$\lambda^{(i)} = \Lambda_{i,i} \geqslant \lambda > 0$ and $(\Delta N_n^{(i)})_{n \geqslant 1}$ is a sequence of martingale increments.

It is worth noting that due to the multiplication by the matrix P, the martingale increment $\Delta N_{n+1}^{(i)}$ potentially depends on the whole coordinate $(\breve{Z}_n^{(j)})_{1 \leqslant j \leqslant d}$. In a completely general case, this involves technicalities mainly due to the fact that the system (4.2) is not completely autonomous (in general, the components $\breve{Z}_n^{(i)}$ and $\breve{Z}_n^{(j)}$ do not evolve independently). To overcome this difficulty, the idea is to obtain some general controls for a system solution to (4.2) and to then bring the controls of each coordinate together. For the sake of simplicity, we propose in the sequel to state the results in the general case but to only make the proof for (4.2) with the assumption that:

$$\mathbb{E}[|\Delta N_{n+1}^{(j)}|^2 | \mathcal{F}_n] \leqslant C(1 + \|\breve{X}_n^{(j)}\|^2). \tag{4.3}$$

From now on, we will omit the indexation by j to alleviate the notations. An easy computation shows that the characteristic polynomial of C_n is given by:

$$\chi_{C_n}(t) = \left(t + \frac{r_n}{2}\right)^2 + \frac{r_n(4\lambda - r_n)}{4}.$$

We now consider the two different cases:

- For all $n \geqslant 1$, C_n has two real or complex eigenvalues whose values do not change from n to n, which corresponds to $r_n = r$. This case necessarily corresponds to an exponentially-weighted memory and r_n is thus kept fixed constant: $r_n = r \geqslant 4\lambda$ or $r_n = r < 4\lambda$.
- For a large enough n, C_n has two complex conjugate and vanishing eigenvalues. This situation may occur if we use a polynomially-weighted memory because, in that case, $r_n \longrightarrow 0$ as $n \longrightarrow +\infty$.

4.1.2. Exponential memory $r_n = r$

We first study the situation when $r_n = r$, which is easier to deal with from a technical point of view.

Proposition 4.1. *Let $\sigma > 0$. Assume that a.s. $\forall n \geqslant 1$, $\mathbb{E}(\|\Delta M_{n+1}\|^2|\mathcal{F}_n) \leqslant \sigma^2(1 + f(X_n))$. Let $(Z_n)_{n\geqslant 0}$ be defined by (4.1) with $Sp(S) \subset [\lambda, +\infty[$ and $r_n = r$. Set:*

$$\alpha_r = \begin{cases} r\left(1 - \sqrt{1 - \frac{4\lambda}{r}}\right), & \text{if} \quad r \geqslant 4\lambda \\ r & \text{if} \quad r < 4\lambda, \end{cases}.$$

Assume that $\gamma_n = \gamma n^{-\beta}$, we then have:

(i) If $\beta < 1$, then a constant $c_{r,\lambda,\gamma}$ exists such that:

$$\forall n \geqslant 1 \qquad \mathbb{E}\left[\|X_n\|^2 + \|Y_n\|^2\right] \leqslant c_{r,\lambda,\gamma}\gamma_n.$$

(ii) If $\beta = 1$, then a constant $c_{r,\lambda,\gamma}$ exists such that:

$$\forall n \geqslant 1 \qquad \mathbb{E}\left[\|X_n\|^2 + \|Y_n\|^2\right] \leqslant c_{r,\lambda,\gamma} n^{-(1\wedge\gamma\alpha_r)} \log(n)^{\mathbf{1}_{\{\gamma\alpha_r=1\}}}.$$

Proof of Proposition 4.1: According to Subsection 4.1.1, we only make the proof for a system solution to (4.2) with the assumption that (4.3) holds. We begin with the simplest case where $r \geqslant 4\lambda$. The above computations show that:

$$Sp(C_n) = \left\{\mu_+ = \frac{-r + \sqrt{(r-4\lambda)r}}{2}; \mu_- = \frac{-r - \sqrt{(r-4\lambda)r}}{2}\right\}, \tag{4.4}$$

while the associated eigenvectors are given by $e_+ = \begin{pmatrix} 1 \\ -\mu_+ \end{pmatrix}$ and $e_- = \begin{pmatrix} 1 \\ -\mu_- \end{pmatrix}$ and are kept fixed throughout the iterations of the algorithm. Consequently, (4.2) may be rewritten in an even simpler way:

$$\breve{Z}_{n+1} = \begin{pmatrix} 1 + \gamma_{n+1}\mu_+ & 0 \\ 0 & 1 + \gamma_{n+1}\mu_- \end{pmatrix} \breve{Z}_n + r\gamma_{n+1}\breve{\xi}_{n+1}, \tag{4.5}$$

where $\breve{Z}_n = QZ_n$ ((Z_n) being defined by (4.2)) where Q is an invertible matrix such that $C_n = Q^{-1}\begin{pmatrix} \mu_+ & 0 \\ 0 & \mu_- \end{pmatrix} Q$ and $\breve{\xi}_{n+1} = Q\Sigma_2\Delta N_{n+1}$. The squared norm of $(\breve{Z}_n)_{n\geqslant 1}$ is now controlled using a standard martingale argument and Assumption $(\mathbf{H}_{\sigma,\mathbf{2}})$:

$$\mathbb{E}\left[\|\breve{Z}_{n+1}\|^2|\mathcal{F}_n\right] \leqslant [(1 + \mu_+\gamma_{n+1})^2 + C\gamma_{n+1}^2]\|\breve{Z}_n\|^2 + C\gamma_{n+1}^2,$$

so that by setting $u_n = \mathbb{E}[\|\breve{Z}_n\|^2]$, this yields:

$$u_{n+1} \leqslant (1 + 2\mu_+\gamma_{n+1} + C_1\gamma_{n+1}^2) + C_2\gamma_{n+1}^2. \tag{4.6}$$

The result then follows from Propositions B.1 (*iii*) and B.2 (*iii*) (see Appendix B).

We now study the situation $r < 4\lambda$. In this case, C_n possesses two conjugate complex eigenvalues:

$$Sp(C_n) = \left\{ \mu_+ = \frac{-r + \mathrm{i}\sqrt{r(4\lambda - r)}}{2}; \mu_- = \frac{-r - \mathrm{i}\sqrt{r(4\lambda - r)}}{2}. \right\}, \tag{4.7}$$

Once again, we use the notation $(\check{Z}_n)_{n\geqslant 1}$ defined as $\check{Z}_n = QZ_n$ with Q an invertible (complex) matrix such that $S_n = Q^{-1}\begin{pmatrix} \mu_+ & 0 \\ 0 & \mu_- \end{pmatrix} Q$ and $\check{\xi}_{n+1} = Q\Sigma_2 \Delta N_{n+1}$. The squared norm of $(\check{Z}_n)_{n\geqslant 1}$ may be controlled while paying attention to the modulus of complex numbers, and we obtain an inequality similar to (4.6).

$$\begin{aligned} \mathbb{E}\left[\|\check{Z}_{n+1}\|^2 | \mathcal{F}_n\right] &\leqslant \max\left(|1+\mu_+\gamma_{n+1}|^2 ; |1+\mu_-\gamma_{n+1}|^2\right) \|\check{Z}_n\|^2 + C_2\gamma_{n+1}^2, \\ &\leqslant \left(\left(1 - \frac{\gamma_{n+1} r}{2}\right)^2 + C_1\gamma_{n+1}^2\right) \|\check{Z}_n\|^2 + C_2\gamma_{n+1}^2, \\ &\leqslant \left(1 - \gamma_{n+1} r + C_1\gamma_{n+1}^2\right) \|\check{Z}_n\|^2 + C_2\gamma_{n+1}^2. \end{aligned}$$

Once again, we can apply (*iii*) of Propositions B.1(*iii*) and B.2(*iii*) to obtain the desired conclusion. □

Remark 4.1. *In the above proposition, the constants $c_{r,\lambda,\gamma}$ are not made explicit. However, it is possible to obtain an estimation if we assume that*

$$\mathbb{E}[\|\Delta M_{n+1}\|^2] \leqslant \sigma^2 \quad \textit{and} \quad r \geqslant 4\lambda.$$

In this particular case, with the notations of (4.6), we have:

$$u_{n+1} \leqslant (1 - \alpha_r\gamma_n)\, u_n + r^2\sigma^2 \|Q_r\|^2 \gamma_{n+1}^2,$$

where $u_n = \mathbb{E}\|\check{Z}_n\|^2$. The Propositions B.1 (iii) and B.2 (iii) now imply that:

$$\mathbb{E}\left[\|\check{Z}_n\|^2\right] \leqslant \mathbb{E}\left[\|\check{Z}_0\|^2\right] e^{-\alpha_r\Gamma_n} + C_\gamma \frac{2r^2\|Q_r\|^2}{\alpha_r}\sigma^2\gamma_n,$$

which, in the end, provide an explicit upper bound of $\mathbb{E}\|Z_n\|^2$ since $Z_n = Q_r^{-1}\check{Z}_n$.

A more important issue concerns the rate obtained when $\beta = 1$ and we can remark in the statement of Proposition 4.1 that this rate depends on the size of γ and of α_r. In particular, the best rate (of order $\mathcal{O}(n^{-1})$) is obtained when $\gamma\alpha_r > 1$, meaning that α_r must be as large as possible to optimize the performance of the algorithm and we therefore obtain a non-adaptive rate. It is easy to see that $r \longmapsto \alpha_r$ increases on $[0, 4\lambda]$ and decreases on $[4\lambda, +\infty)$. It attains its maximal value ($\max_r \alpha_r = 4\lambda$) when $r = 4\lambda$. This maximal value is twice the size of the eigenvalue of the (standard) stochastic gradient descent (SGD).

Finally, $\lim_{r\longrightarrow+\infty}\alpha_r = 2\lambda$. *This limiting value* 2λ *corresponds to the size of the eigenvalue of the SGD. In other words, the limit* $r = +\infty$ *in HBF may be seen as an almost identical situation to SGD.*

If we compare the rate of convergence of HBF to the one of SGD using the same step size $\gamma_n = \gamma n^{-1}$, *we see that choosing a reasonably large* r *makes it possible to obtain a less stringent condition on* γ *to recover the (optimal) rate* $\mathcal{O}(n^{-1})$. *In particular, the rate of the HBF is better when* $r \geqslant 2\lambda$ *than the one attained by the SGD. Unfortunately, it seems impossible to obtain an adaptive procedure on the choice of* (γ, r) *that guarantees the rate* $\mathcal{O}(n^{-1})$, *unlike the Polyak-Ruppert averaging procedure.*

4.1.3. Polynomial memory $r_n = r\Gamma_n^{-1} \longrightarrow 0$

This case is more intricate because of the variations with n of the eigenvectors of the matrix C_n defined in (4.2).

Proposition 4.2. *Let* $\sigma > 0$. *Assume that a.s.* $\forall n \geqslant 1$, $\mathbb{E}(\|\Delta M_{n+1}\|^2|\mathcal{F}_n) \leqslant \sigma^2(1 + f(X_n))$. *Let* $(Z_n)_{n\geqslant 0}$ *be defined by* (4.1) *with* $Sp(S) \subset [\lambda, +\infty[$ *and* $r_n = \frac{r}{\Gamma_n}$.

(*i*) *If* $\beta < 1$ *and* $r > \frac{1+\beta}{2(1-\beta)}$, *a constant* $c_{\beta,\lambda,r}$ *exists such that:*

$$\forall n \geqslant 1 \qquad \mathbb{E}\|X_n\|^2 \leqslant c_{\beta,\lambda,r}\gamma_n,$$

and

$$\forall n \geqslant 1 \qquad \mathbb{E}\|Y_n\|^2 \leqslant c_{\beta,\lambda}\gamma_n r_n.$$

(*ii*) *If* $\beta = 1$, *a constant* C *exists such that:*

$$\forall n \geqslant 1 \qquad \mathbb{E}\|X_n\|^2 \leqslant \frac{C}{\log n}$$

and

$$\forall n \geqslant 1 \qquad \mathbb{E}\|Y_n\|^2 \leqslant \frac{C}{n\log n}$$

Remark 4.2. *We can observe that when* $\beta < 1$, *the rates of the exponential case are preserved under a constraint on* r *which becomes harder and harder when* β *is close to* 1*:* r *needs to be greater than* $\frac{1+\beta}{2(1-\beta)}$. *Carefully following the proof of this result, we could in fact show that when* $1/2 < r < \frac{1+\beta}{2(1-\beta)}$, *then* $\mathbb{E}\|X_n\|^2 \leqslant Cn^{-(r-\frac{1}{2})(1-\beta)}$. *Since* $(r-\frac{1}{2})(1-\beta) \longrightarrow 0$ *as* $\beta \longrightarrow 1$, *our upper bound in* $(\log n)^{-1}$ *related to the case* $\beta = 1$ *becomes reasonable. Another possible interpretation of the poor convergence rate in that case is that the size of the negative real part of the eigenvalues of* C_n *is on the order* $\frac{1}{n\log n}$, *which leads to a contraction of the bias equivalent to* $\mathcal{O}\left(e^{-c\sum_1^n \frac{1}{k\log k}}\right)$. *Regardless of* c, *we cannot obtain a polynomial rate of convergence in that case since* $\sum_1^n \frac{1}{k\log k} \sim \log\log n$.

Proof of Proposition 4.2:
Proof of (i): We study the case $\beta < 1$ here. According to the arguments used in the proof of Proposition 4.1 and Subsection 4.1.1, the dynamical system may be reduced to d couples of systems in the form $(x_n^{(i)}, y_n^{(i)})_{n\geqslant 1}$ so that we only make the proof for a system solution to (4.2) under assumption (4.3). Another key feature of the polynomial case has been observed in the proof of the a.s. convergence of the algorithm (Theorem 3.2): the study of the rate in the polynomial case involves a normalization of the algorithm with a $\sqrt{r_n}$-scaling of the Y coordinate. Therefore, we set $\tilde{Z}_n = (\tilde{X}_n, \tilde{Y}_n)$ with $\tilde{X}_n = X_n$ and $\tilde{Y}_n = Y_n/\sqrt{r_n}$. With these notations, we obtain (similar to Lemma A.2):

$$\tilde{Z}_{n+1} = (I_2 + \tilde{\gamma}_{n+1}\tilde{C}_n)\tilde{Z}_n + \tilde{\gamma}_{n+1}\sqrt{\frac{r_n}{r_{n+1}}}\Sigma_2 \Delta N_{n+1}, \tag{4.8}$$

with $\tilde{\gamma}_{n+1} = \gamma_{n+1}\sqrt{r_n}$ and:

$$\tilde{C}_n = \begin{pmatrix} 0 & -1 \\ \lambda\sqrt{\frac{r_n}{r_{n+1}}} & \rho_n \end{pmatrix}$$

with

$$\rho_n := \frac{1}{\tilde{\gamma}_{n+1}}\left(\sqrt{\frac{r_n}{r_{n+1}}} - 1\right) - \frac{r_n}{\sqrt{r_{n+1}}}.$$

Since $r_n = r\Gamma_n^{-1}$, the following expansion holds:

$$\rho_n = \frac{1}{\sqrt{\Gamma_n}}\left(\frac{1}{2\sqrt{r}} - \sqrt{r}\right) + O\left(\frac{\gamma_n}{\Gamma_n^{\frac{3}{2}}}\right). \tag{4.9}$$

In particular, for a large enough n, $\rho_n < 0$ if and only if $r > 1/2$. Furthermore, an integer $n_0 \in \mathbb{N}$ exists such that for any $n \geqslant n_0$, $\tilde{C}_n$ has complex eigenvalues given by:

$$\mu_{\pm}^{(n)} = \frac{1}{2}\left(\rho_n \pm i\sqrt{4\lambda\sqrt{\frac{r_n}{r_{n+1}}} - \rho_n^2}\right) \xrightarrow{n\to+\infty} \pm i\sqrt{\lambda}.$$

We define the diagonal matrix:

$$\Lambda_n := \begin{pmatrix} \mu_+^{(n)} & 0 \\ 0 & \mu_-^{(n)} \end{pmatrix}$$

and let Q_n be the matrix that satisfies $Q_n^{-1}\Lambda_n Q_n = \tilde{C}_n$. We have:

$$Q_n^{-1} = \begin{pmatrix} 1 & 1 \\ -\mu_+^{(n)} & -\mu_-^{(n)} \end{pmatrix} \quad \text{and} \quad Q_n = \frac{1}{\mu_+^{(n)} - \mu_-^{(n)}}\begin{pmatrix} -\mu_-^{(n)} & -1 \\ \mu_+^{(n)} & 1 \end{pmatrix}.$$

We can now introduce the change of basis brought by Q_n and the new coordinates $\breve{Z}_n := Q_n \tilde{Z}_n$. We have:

$$\begin{aligned}\breve{Z}_{n+1} &= Q_{n+1}(I_2 + \tilde{\gamma}_{n+1}\tilde{C}_n)Q_n^{-1}\breve{Z}_n + \tilde{\gamma}_{n+1}\sqrt{\frac{r_n}{r_{n+1}}}Q_{n+1}\Sigma_2\Delta N_{n+1} \\ &= Q_{n+1}Q_n^{-1}(I_2 + \tilde{\gamma}_{n+1}\Lambda_n)\breve{Z}_n + \tilde{\gamma}_{n+1}\sqrt{\frac{r_n}{r_{n+1}}}Q_{n+1}\Sigma_2\Delta N_{n+1}. \qquad (4.10)\end{aligned}$$

We now observe that:

$$Q_{n+1}Q_n^{-1} = I_2 + \Upsilon_n \quad \text{with} \quad \Upsilon_n = (Q_{n+1} - Q_n)Q_n^{-1}$$

and that for n large enough:

$$\|\Upsilon_n\|_\infty \leqslant C\|Q_{n+1} - Q_n\|_\infty = O(|\mu_+^{(n+1)} - \mu_+^{(n)}|) = O\big(|\rho_{n+1} - \rho_n| + |\mathfrak{Im}(\mu_+^{(n+1)} - \mu_+^{(n)})|\big).$$

Expansion (4.9), the fact that $\sqrt{\frac{r_n}{r_{n+1}}} = 1 + \frac{1}{2}\frac{\gamma_{n+1}}{\Gamma_n} + O\left(\frac{\gamma_{n+1}^2}{\Gamma_n^2}\right)$ and the Lipschitz continuity of $x \mapsto \sqrt{1+x}$ on $[-1/2, +\infty)$ yield:

$$\|\Upsilon_n\|_\infty = O\left(\frac{\gamma_n}{\Gamma_n^{\frac{3}{2}}} + \frac{\gamma_n - \gamma_{n-1}}{\Gamma_n}\right) = O\left(\frac{\gamma_n}{\Gamma_n^{\frac{3}{2}}}\right) = O\left(n^{-\frac{\beta+3}{2}}\right).$$

From the above, we obtain, for any $z \in \mathbb{R}^2$,

$$\begin{aligned}\|Q_{n+1}Q_n^{-1}(I_2 + \tilde{\gamma}_{n+1}\Lambda_n)z\|^2 \quad \leqslant \quad & \left(1 + \tilde{\gamma}_{n+1}\frac{\rho_n}{2} + O\left(\frac{\gamma_n}{\Gamma_n^{\frac{3}{2}}}\right)\right)^2 \|z\|^2 \\ & + \left(\tilde{\gamma}_{n+1}\mathfrak{Im}(\mu_+^{(n)}) + O\left(\frac{\gamma_n}{\Gamma_n^{\frac{3}{2}}}\right)\right)^2 \|z\|^2\end{aligned}$$

which after several computations yields:

$$\|Q_{n+1}Q_n^{-1}(I_2 + \tilde{\gamma}_{n+1}\Lambda_n)z\|^2 \leqslant \left(1 + \frac{\gamma_{n+1}}{\Gamma_n}\left(\frac{1}{2} - r + o(1)\right)\right)\|z\|^2.$$

Note that a universal constant C (independent of n) exists such that $\|Q_{n+1}\|_\infty \leqslant C$ and the upper bounds above can be used into (4.10) to deduce that:

$$\begin{aligned}\|\breve{Z}_{n+1}\|^2 \quad \leqslant \quad & \left(1 + \frac{\gamma_{n+1}}{\Gamma_n}\left(\frac{1}{2} - r\right) + b\left(\frac{\gamma_{n+1}}{\Gamma_n}\right)^2\right)\|\breve{Z}_n\|^2 \\ & + \quad \tilde{\gamma}_{n+1}\Delta\widetilde{M}_n + C\frac{\gamma_{n+1}^2}{\Gamma_n}\|\Delta N_{n+1}\|^2, \qquad (4.11)\end{aligned}$$

where $(\Delta\widetilde{M}_n)_{n\geqslant 1}$ is a sequence of martingale increments and b a large enough constant.

When $\gamma_n = \gamma n^{-\beta}$ with $\beta < 1$, the fact that $\Gamma_n = \frac{n^{1-\beta}}{1-\beta} + O(1)$ combined with the upper bound of the variance of the martingale (4.3) imply that:

$$\mathbb{E}[\|\check{Z}_{n+1}\|^2] \leqslant \left(1 - \frac{\alpha}{n} + \frac{b}{n^2}\right) \mathbb{E}[\|\check{Z}_n\|^2] + Cn^{-1-\beta} \tag{4.12}$$

where $\alpha := (r - \frac{1}{2})(1 - \beta)$. Under the condition $r > \frac{1+\beta}{2(1-\beta)}$, we observe that:

$$\alpha > \beta.$$

An induction based on Inequality (4.12) yields:

$$\begin{aligned}
\mathbb{E}[\|\check{Z}_{n+1}\|^2] &\leqslant \mathbb{E}[\|\check{Z}_{n_\varepsilon}\|^2] \prod_{\ell=n_\varepsilon}^{n} \left(1 - \frac{\alpha}{\ell} + \frac{b}{\ell^2}\right) \\
&+ C \sum_{k=n_\varepsilon+1}^{n} k^{-1-\beta} \prod_{\ell=k+1}^{n} \left(1 - \frac{\alpha}{\ell} + \frac{b}{\ell^2}\right) \\
&\leqslant Cn^{-\beta}
\end{aligned}$$

where in the second line, we repeated an argument used in the proof of Propositions B.2 and made use of the property $\alpha > \beta$. To conclude the proof, it remains to observe that $\|Q_{n+1}^{-1}\|_\infty \leqslant C$ regardless of n. ◇

(*ii*) When $\beta = 1$, Inequality 4.11 leads to:

$$\mathbb{E}[\|\check{Z}_{n+1}\|^2] \leqslant \left(1 - \frac{\alpha}{n \log n} + \frac{b}{n^2 \log n}\right) \mathbb{E}[\|\check{Z}_n\|^2] + \frac{C}{n^2 \log n}$$

and a procedure similar to the one used above (given that $\sum_{k=1}^{n} (k \log k)^{-1} \sim \log(\log n)$) leads to the desired result. ◇□

4.2. The non-quadratic case under exponential memory

The objective of this subsection is to extend the results of the quadratic case to strongly convex functions satisfying ($\mathbf{H}_{SC}(\alpha)$) for a given positive α. As pointed out in Remark 2.2, we are not able to obtain neat and somewhat intrinsic results in the polynomial memory case, so we therefore preferred to only consider the exponential memory one.

With the help of Subsection 4.1.1, we can restrain the study to the situation where $d = 1$ and f has a unique minimum in $x^\star$ and we denote $\lambda = f''(x^\star)$, which is assumed to be positive. We also assume that $\underline{f''} = \inf_{x \in \mathbb{R}} f''(x) > 0$. It is worth noting that in this setting, we are able to obtain some non-asymptotic bounds with some assumptions on λ only. This means that our results do not involve the quantity $\underline{f''}$. To only involve the value of the second derivative in $x^\star$, the main argument is a *power increase* stated in the next lemma.

Lemma 4.1. *Let $(u_n^{(k)})_{n\geqslant 0,k\geqslant 1}$ be a sequence of non-negative numbers satisfying for every integers $n \geqslant 0$ and $k \geqslant 1$,*

$$u_{n+1}^{(k)} \leqslant (1 - a_k\gamma_{n+1} + b_k\gamma_{n+1}^2)u_n^{(k)} + C_k(\gamma_{n+1}^2 + \gamma_{n+1}u_n^{(k+1)}) \tag{4.13}$$

where $(a_k)_{k\geqslant 1}$ and $(b_k)_{k\geqslant 1}$ are sequences of positive numbers. Furthermore, assume that $K \geqslant 2$ exists and a constant $C > 0$ exists such that:

$$\forall n \geqslant 1, \quad u_n^{(K)} \leqslant C\gamma_n. \tag{4.14}$$

Then, suppose that $\gamma_n = \gamma n^{-\beta}$ ($\gamma > 0$, $\beta \in (0, 1]$) and that $\underline{a} := \min_{k\leqslant K} a_k > 0$ and $\bar{b} := \max_{k\leqslant K} b_k < +\infty$.

(i) If $\beta \in (0, 1)$, a constant $C > 0$ exists such that for every $k \in \{1, \dots, K\}$,

$$\forall n \geqslant 1, \quad u_n^{(k)} \leqslant C\gamma_n.$$

(ii) If $\beta = 1$ and $\underline{a}\gamma > 1$, a constant $C > 0$ exists such that for every $k \in \{1, \dots, K\}$,

$$\forall n \geqslant 2, \quad u_n^{(k)} \leqslant Cn^{-1}. \tag{4.15}$$

Proof of Lemma 4.1:
Let $K \geqslant 2$. We proceed by a decreasing induction on $k \in \{1, \dots, K\}$. The initialization is given by (4.14). Then, let $k \in \{1, \dots, K-1\}$ and assume that $u_n^{(k+1)} \leqslant C_{k+1}\gamma_n$ (where C_k is a positive constant that does not depend on n). We can use this upper bound in the second term of the right hand side of (4.13) and obtain:

$$u_{n+1}^{(k)} \leqslant (1 - \underline{a}\gamma_{n+1} + \bar{b}\gamma_{n+1}^2)u_n^{(k)} + C\gamma_{n+1}^2$$

where C is a constant that does not depend on n.

When $\beta < 1$, it follows from Proposition B.1(*iii*) that:

$$\forall n \geqslant 1, \quad u_n^{(k)} \lesssim \gamma_n. \qquad \diamond$$

If $\beta = 1$ and $\underline{a}\gamma > 1$ now, the above control is a consequence of Proposition B.2(*iii*). This concludes the proof. $\diamond\square$

We will apply this lemma to $u_n^{(k)} = \mathbb{E}[|\breve{Z}_n|^{2k}]$ where $\breve{Z}_n$ is an appropriate linear transformation of Z_n. Therefore, we will mainly have to check that Conditions (4.13) and (4.14) hold.

Proposition 4.3. *Assume $(\mathbf{H}_s)$, $(\mathbf{H}_{SC}(\alpha))$ and $(\mathbf{H}_{\sigma,\infty})$ with $p \geqslant 1$. Let a and b be some positive numbers such that (A.1) holds. Then, an integer $K \geqslant 1$ exists such that for any $p \geqslant K$:*

$$\mathbb{E}[V_n^p(X_n, Y_n)] \leqslant C_p\gamma_n. \tag{4.16}$$

Furthermore, if $r_n = r$ and $\gamma_n = \gamma n^{-\beta}$ with $\beta \in (0, 1)$, then (4.16) holds for $p = K = 1$ under $(\mathbf{H}_{\sigma,\mathbf{2}})$ instead of $(\mathbf{H}_{\boldsymbol{\sigma},\infty})$. As a consequence,

$$\mathbb{E}[\|X_n - x^\star\|^{2K} + \|Y_n\|^{2K}] \leqslant C\gamma_n. \tag{4.17}$$

Remark 4.3. *Note that the second assertion* (4.17) *easily follows from Equations* (A.2) *and* (4.16) *and from the fact that under* $(\mathbf{H}_{SC}(\alpha))$*, a constant* c *exists such that for all* x*,* $f(x) \geqslant c\|x\|^2$*.*

Moreover, note that this proposition is not restricted to the exponential memory case. In particular, as suggested in Remark 2.2, this Lyapunov approach could lead to some (rough) controls of the quadratic error in the polynomial case when the function is not quadratic.

Proof of Proposition 4.3:
We begin by the first assertion under Assumption $(\mathbf{H}_{\sigma,\infty})$. Going back to the proof of Lemma A.1 (and to the associated notations), we obtain the existence of some positive a and b such that

$$V_{n+1}(X_{n+1},Y_{n+1}) \leqslant V_n(X_n,Y_n) + \gamma_{n+1}\Delta_{n+1} \quad \text{with}$$
$$\Delta_{n+1} = -c_{a,b}\|Y_n\|^2 - r_n b\|\nabla f(X_n)\|^2 - br_n\langle \nabla f(X_n), \Delta M_{n+1}\rangle + \Delta R_{n+1} \quad (c_{a,b}>0).$$

Denoting the smallest (positive) eigenvalue of $D^2 f(x^\star)$ by $\underline{\lambda}$, we have:

$$\|\nabla f(x)\|^2 \geqslant \underline{\lambda}\|x\|^2 \geqslant C\ \underline{\lambda} f(x).$$

Following the arguments of the proof of Lemma A.1 once again, we can easily deduce the existence of some positive ε and C such that:

$$\mathbb{E}[\Delta_{n+1}|\mathcal{F}_n] \leqslant (-\varepsilon + C\gamma_{n+1})r_n V_n(X_n,Y_n) + C\gamma_{n+1}r_n.$$

Using $(\mathbf{H}_{\sigma,\infty})$, we also obtain for every $r \geqslant 1$:

$$\mathbb{E}[\|\Delta_{n+1}\|^r|\mathcal{F}_n] \leqslant C_r(1+V_n^r(X_n,Y_n)).$$

As a consequence, a binomial expansion of $(V_n(X_n,Y_n)+\gamma_{n+1}\Delta_{n+1})^K$ yields:

$$\mathbb{E}[V_{n+1}^K(X_{n+1},Y_{n+1})|\mathcal{F}_n] \leqslant (1 - K\varepsilon\gamma_{n+1}r_n + C\gamma_{n+1}^2 r_n)V_n^K(X_n,Y_n) + C\gamma_{n+1}^2 r_n.$$

Setting $u_n = \mathbb{E}[V_{n+1}^K(X_{n+1},Y_{n+1})]$, we obtain:

$$u_{n+1} \leqslant (1 - K\varepsilon\gamma_{n+1}r_n + C\gamma_{n+1}^2 r_n)u_n + C\gamma_{n+1}^2 r_n.$$

Now, assume that $\gamma_n = \gamma n^{-\beta}$ with $\beta \in (0,1]$ and successively consider exponential and polynomial cases:

- If $r_n = r$ and $\beta < 1$, the result holds with $K = 1$ by Proposition B.1(*iii*). ◇
- If $r_n = r$ and $\beta = 1$, we have to choose K large enough in order that $K\varepsilon\gamma > 1$. In this case, Proposition B.2(*iii*) yields the result. ◇
- If $r_n = r/\Gamma_n$ and $\beta < 1$ now, then the above inequality yields the existence of a $\rho > \beta$ and a $n_0 \geqslant 1$ for K large enough such that:

$$\forall n \geqslant n_0, \quad u_{n+1} \leqslant \left(1 - \frac{\rho}{n}\right)u_n + Cn^{-\beta-1}.$$

We have:

$$u_n \leqslant u_{n_0} \prod_{k=n_0}^{n} \left(1 - \frac{\rho}{k}\right) + C \sum_{k=n_0+1}^{n} k^{-\beta-1} \prod_{\ell=k+1}^{n} \left(1 - \frac{\rho}{k}\right).$$

Given that $1 - x \leqslant \exp(-x)$ and that $\sum_{k=1}^{n} \frac{1}{k} = \log n + O(1)$, we obtain:

$$u_n \leqslant Cn^{-\rho}(1 + \sum_{k=n_0+1}^{n} k^{-\beta-1+\rho}) \leqslant Cn^{-\beta}$$

where in the last inequality, we deduced that $-\beta - 1 + \rho > -1$ since $\rho < \beta$. ◇□

Proposition 4.4. *Assume* $(\mathbf{H}_s)$, $(\mathbf{H}_{SC}(\alpha))$*and* $(\mathbf{H}_{\boldsymbol{\sigma},\infty})$ *and* $r_n = r$ *for all* $n \geqslant 1$. *Set* $\lambda = f''(x^\star)$. *Then, assume that* $\gamma_n = \gamma n^{-\beta}$ *with* $\beta \in (0, 1]$.

- *If* $\beta < 1$, *then:*
$$\mathbb{E}[\|X_n - x^\star\|^2] + \mathbb{E}[\|Y_n\|^2] \leqslant C\gamma_n.$$
- *If* $\beta = 1$, *then for every* $\varepsilon > 0$, *a constant* C_ε *exists such that*
$$\mathbb{E}[\|X_n - x^\star\|^2] \leqslant C_\varepsilon n^{-((r+\varepsilon-\sqrt{r^2-4\lambda r}1_{r\geqslant 4\lambda})\gamma)\wedge 1}.$$

Proof of Proposition 4.4:
The starting point is to linearize the gradient:

$$f'(X_n) = \lambda(X_n - x^\star) + \phi_n \quad \text{where} \quad \phi_n = (f''(\xi_n) - f''(x^\star))(X_n - x^\star).$$

Since f'' is Lipschitz continuous, then:

$$|\phi_n| \leqslant C(X_n - x^\star)^2. \tag{4.18}$$

Let us begin with the case where the matrix C_n defined in (4.2) has real eigenvalues μ_+ and μ_- (given by (4.4)). With the notations introduced in (4.5),

$$\breve{Z}_{n+1} = \begin{pmatrix} 1 + \gamma_{n+1}\mu_+ & 0 \\ 0 & 1 + \gamma_{n+1}\mu_- \end{pmatrix} \breve{Z}_n + r\gamma_{n+1} Q \begin{pmatrix} 0 \\ \phi_n \end{pmatrix} + r\gamma_{n+1}\breve{\xi}_{n+1}. \tag{4.19}$$

As a consequence,

$$\|\breve{Z}_{n+1}\|^2 \leqslant (1 + \mu_+\gamma_{n+1})^2\|\breve{Z}_n\|^2 + C\gamma_{n+1}\|\breve{Z}_n\|^3 + \gamma_{n+1}^2(\|\breve{Z}_n\|^4 + \|\Delta N_{n+1}\|^2) + \Delta\mathcal{M}_{n+1}$$

where $(\Delta\mathcal{M}_n)$ is a sequence of martingale increments. Using the elementary inequality $|x| \leqslant \varepsilon + C_\varepsilon|x|^2$, $x \in \mathbb{R}$ (available for any $\varepsilon > 0$),

$$\begin{aligned} \|\breve{Z}_{n+1}\|^2 &\leqslant [(1 + (2\mu_+ + \varepsilon)\gamma_{n+1} + C\gamma_{n+1}^2)]\|\breve{Z}_n\|^2 \\ &+ C_\varepsilon\gamma_{n+1}\|\breve{Z}_n\|^4 + C\gamma_{n+1}^2\|\Delta N_{n+1}\|^2 + \Delta N_{n+1}. \end{aligned}$$

Then, by Assumption $(\mathbf{H}_{\boldsymbol{\sigma},\infty})$ and the fact $\sup_n \mathbb{E}[|\breve{Z}_n|^r] < +\infty$ for any $r > 1$ (by Proposition 4.3 for example), we obtain, for any $k \geqslant 1$,

$$\mathbb{E}\left[\|\check{Z}_{n+1}\|^{2k}\right] \leqslant (1 + k(2\mu_+ + \varepsilon)\gamma_{n+1} + C_k\gamma_{n+1}^2)\mathbb{E}[\|\check{Z}_n\|^{2k}] + C_{k,\varepsilon}(\gamma_{n+1}\mathbb{E}[\|\check{Z}_n\|^{2k+2}] + \gamma_{n+1}^2).$$

At this stage, we observe that Assumption (4.13) is satisfied with $u_n^{(k)} = \mathbb{E}[\|\check{Z}_n\|^{2k}]$ and $a_k = k(2\mu_+ + \varepsilon)$. Using Proposition 4.3 and Lemma A.1(*i*), we check that the second assumption of Lemma 4.1 also holds. Thus, the result follows in this case from this lemma. □

5. Limit of the rescaled algorithm

In this paragraph, we establish a (functional) Central Limit Theorem when the memory is exponential, *i.e.*, when $r_n = r$ and when $(\mathbf{H}_{SC}(\alpha))$ holds. In particular, f admits a unique minimum $x^\star$. Without loss of generality, we assume that $x^\star = 0$.

5.1. Rescaling stochastic HBF

We start with an appropriate rescaling by a factor $\sqrt{\gamma_n}$. More precisely, we define a sequence $(\bar{Z}_n)_{n\geqslant 1}$:

$$\bar{Z}_n = \frac{Z_n}{\sqrt{\gamma_n}} = \left(\frac{X_n}{\sqrt{\gamma_n}}, \frac{Y_n}{\sqrt{\gamma_n}}\right).$$

Given that f is $\mathcal{C}^2$ (and that $x^\star = 0$), we "linearize" ∇f around 0 with a Taylor formula and obtain that $\xi_n \in [0, X_n]$ exists such that:

$$\nabla f(X_n) = D^2 f(\xi_n) X_n.$$

Therefore, we can compute that:

$$\bar{Z}_{n+1} = \bar{Z}_n + \gamma_{n+1} b_n(\bar{Z}_n) + \sqrt{\gamma_{n+1}} \begin{pmatrix} 0 \\ \Delta M_{n+1} \end{pmatrix}$$

where b_n is defined by:

$$b_n(z) = \frac{1}{\gamma_{n+1}}\left(\sqrt{\frac{\gamma_n}{\gamma_{n+1}}} - 1\right) z + \bar{C}_n z, \quad z \in \mathbb{R}^{2d}, \tag{5.1}$$

where:

$$\bar{C}_n := \sqrt{\frac{\gamma_n}{\gamma_{n+1}}} \begin{pmatrix} 0 & -I_d \\ rD^2 f(\xi_n) & -rI_d \end{pmatrix}. \tag{5.2}$$

It is important to observe that:

$$\frac{1}{\gamma_{n+1}}\left(\sqrt{\frac{\gamma_n}{\gamma_{n+1}}} - 1\right) = \gamma^{-1}(n+1)^\beta \left[1 + \frac{\beta}{2n} + o(n^{-1}) - 1\right] = \begin{cases} o(n^{\beta-1}) & \text{if } \beta < 1 \\ \frac{1}{2\gamma} + o(1) & \text{if } \beta = 1 \end{cases} \tag{5.3}$$

We associate to the sequence $(\bar{Z}_n)_{n\geqslant 1}$ a sequence $(\bar{Z}^{(n)})_{n\geqslant 1}$ of continuous-time processes defined by:

$$\bar{Z}_t^{(n)} = \bar{Z}_n + B_t^{(n)} + M_t^{(n)}, \quad t \geqslant 0, \tag{5.4}$$

where:

$$B_t^{(n)} = \sum_{k=n+1}^{\tilde{N}(n,t)} \gamma_k b_{k-1}(\bar{Z}_{k-1}) + (t - \underline{t}_n) b_{\tilde{N}(n,t)}(\bar{Z}_{\tilde{N}(n,t)}),$$

$$M_t^{(n)} = \sum_{k=n+1}^{\tilde{N}(n,t)} \sqrt{\gamma_k} \begin{pmatrix} 0 \\ \Delta M_k \end{pmatrix} + \sqrt{t - \underline{t}_n} \begin{pmatrix} 0 \\ \Delta M_{\tilde{N}(n,t)+1} \end{pmatrix}.$$

We used the standard notations $\underline{t}_n = \Gamma_{\tilde{N}(n,t)} - \Gamma_n$ above where $N(n,t) = \min\left\{ m \geqslant n, \sum_{k=n+1}^{m} \gamma_k > t \right\}$.

To obtain a CLT, we show that $(\bar{Z}^{(n)})_{n\geqslant 1}$ converges in distribution to a stationary diffusion, following a classical roadmap based on a tightness result and on an identification of the limit as a solution to a martingale problem.

5.2. Tightness

The next lemma holds for any sequence of processes that satisfy (5.4).

Lemma 5.1. *Assume that $D^2 f$ is bounded, that $\sup_{k\geqslant 1} \mathbb{E}[\|\bar{Z}_k\|^2] < +\infty$ and that a $p > 2$ exists such that $\sup_{k\geqslant 1} \mathbb{E}[\|\Delta M_k\|^p] < +\infty$, then $(\bar{Z}^{(n)})_{n\geqslant 1}$ is tight for the weak topology induced by the weak convergence on compact intervals.*

Proof of Lemme 5.1:
First, note that $\bar{Z}_0^{(n)} = \bar{Z}_n$, the assumption $\sup_{k\geqslant 1} \mathbb{E}[\|\bar{Z}_k\|^2] < +\infty$ implies the tightness of $(\bar{Z}_0^{(n)})_{n\geqslant 1}$ (on $\mathbb{R}^{2d}$). Then, by a classical criterion (see, *e.g.*, [Bil95, Theorem 8.3]), we deduce that a sufficient condition for the tightness of $(\bar{Z}^{(n)})_{n\geqslant 1}$ (for the weak topology induced by the uniform convergence on compacts intervals) is the following property: for any $T > 0$, for any positive ε and η, a $\delta > 0$ exist and an integer n_0 such that for any $t \in [0, T]$ and $n \geqslant n_0$,

$$\mathbb{P}\Big(\sup_{s\in[t,t+\delta]} \|\bar{Z}_s^{(n)} - \bar{Z}_t^{(n)}\| \geqslant \varepsilon \Big) \leqslant \eta\delta.$$

We consider $B^{(n)}$ and $M^{(n)}$ separately and begin by the drift term $B^{(n)}$. On the one hand,

$$\mathbb{P}\left(\sup_{s\in[t,t+\delta]} \|B_s^{(n)} - B_t^{(n)}\| \geqslant \varepsilon \right) \leqslant \mathbb{P}\left(\sum_{k=N(n,t)}^{N(n,t+\delta)+1} \gamma_k \|b_{k-1}(\bar{Z}_{k-1})\| \geqslant \varepsilon \right).$$

The Chebyschev inequality and the fact that $\|b_k(z)\| \leqslant C(1+\|z\|)$ (where C does not depend on k) yield:

$$\mathbb{P}\left(\sup_{s\in[t,t+\delta]}\|B_s^{(n)}-B_t^{(n)}\|\geqslant\varepsilon\right)\leqslant\varepsilon^{-2}\mathbb{E}\left[\left(\sum_{k=N(n,t)}^{N(n,t+\delta)+1}\gamma_k(1+\|(\bar{Z}_{k-1})\|)\right)^2\right]$$

The Jensen inequality and the fact that $\sum_{k=N(n,t)}^{N(n,t+\delta)+1}\gamma_k \leqslant 2\delta$ when n is large enough imply that a constant C exists such that for large enough n and for a small enough δ:

$$\mathbb{P}\left(\sup_{s\in[t,t+\delta]}\|B_s^{(n)}-B_t^{(n)}\|\geqslant\varepsilon\right)\leqslant\varepsilon^{-2}\times C\delta^2(1+\sup_{k\geqslant 1}\mathbb{E}[\|\bar{Z}_k\|^2])\leqslant\eta\delta \qquad \diamond$$

We now consider the martingale component $M^{(n)}$: if we denote $\alpha=\sqrt{\frac{t-t_n}{\gamma_{N(n,t)+1}}}$, we have for any $t\geqslant 0$,

$$M_s^{(n)}=(1-\alpha)M_{N(n,s)}^{(n)}+\alpha M_{N(n,s)+1}^{(n)}$$

so that $\|M_s^{(n)}-M_t^{(n)}\|\leqslant\max\{\|M_{N(n,s)}^{(n)}-M_t^{(n)}\|,\|M_{N(n,s)+1}^{(n)}-M_t^{(n)}\|\}$. As a consequence,

$$\mathbb{P}\left(\sup_{s\in[t,t+\delta]}\|M_s^{(n)}-M_t^{(n)}\|\geqslant\varepsilon\right)\leqslant\mathbb{P}\left(\sup_{N(n,t)+1\leqslant k\leqslant N(n,t+\delta)+1}\|M_{\Gamma_k}^{(n)}-M_t^{(n)}\|\geqslant\varepsilon\right)$$

Let $p>2$ and applying the Doob inequality, the assumption of the lemma leads to:

$$\mathbb{P}\left(\sup_{s\in[t,t+\delta]}\|M_s^{(n)}-M_t^{(n)}\|\geqslant\varepsilon\right)\leqslant\varepsilon^{-p}\mathbb{E}\left[\|M_{\tilde{N}(n,t+\delta)+1}^{(n)}-M_t^{(n)}\|^p\right]$$

and the Minkowski inequality yields:

$$\mathbb{P}\left(\sup_{s\in[t,t+\delta]}\|M_s^{(n)}-M_t^{(n)}\|\geqslant\varepsilon\right)\leqslant\varepsilon^{-p}\sum_{k=N(n,t)+1}^{N(n,t+\delta)+1}\gamma_k^{\frac{p}{2}}\mathbb{E}\left[\|\Delta M_k\|^p\right].$$

Under the assumptions of the lemma, $\mathbb{E}[[\|\Delta M_k\|^p]\leqslant C$. Furthermore, we can use the rough upper bound:

$$\sum_{k=N(n,t)+1}^{N(n,t+\delta)+1}\gamma_k^{\frac{p}{2}}\leqslant\gamma_n^{\frac{p}{2}-1}\sum_{k=N(n,t)+1}^{N(n,t+\delta)+1}\gamma_k\leqslant\eta\delta$$

for large enough n. This concludes the proof. $\diamond\square$

Corollary 5.1. *Let the assumptions of Theorem 2.4 hold, then $(\bar{Z}^{(n)})_{n\geqslant 1}$ is tight.*

Proof of Corollary 5.1:
To prove this result, it is enough to check that the assumptions of Lemma 5.1 are satisfied. First, one remarks that the assumptions of Theorem 2.4 imply the ones of Theorem 2.3(*a*) so that $\mathbb{E}[\|Z_n - z^\star\|^2] \leqslant C\gamma_n$ (this also holds when $\beta = 1$ since we assume that $\gamma\alpha_r > 1$). As a consequence, $\sup_{k\geqslant 1} \mathbb{E}[\|\bar{Z}_k\|^2] < +\infty$.

On the other hand, since $(\mathbf{H}_{\sigma,\mathbf{p}})$ holds for a given $p > 2$, we can derive by following the lines of the proof of Proposition 4.3 that $\sup_{n\geqslant 1} \mathbb{E}[V^p(X_n, Y_n)] < +\infty$. As a consequence, $\sup_n \mathbb{E}[f^p(X_n)] < +\infty$ and $(\mathbf{H}_{\sigma,\mathbf{p}})$ leads to:

$$\sup_{n\geqslant 1} \mathbb{E}[\|\Delta M_n\|^p] \lesssim \sup_n \mathbb{E}[f^p(X_n)] < +\infty. \qquad \square$$

5.3. Identification of the limit

Starting from our compactness result above, we now characterize the potential weak limits of $(\bar{Z}^{(n)})_{n\geqslant 1}$. This step is strongly based on the following lemma.

Lemma 5.2. *Suppose that the assumptions of Lemma 5.1 hold and that:*

$$\mathbb{E}[\Delta M_n (\Delta M_n)^t | \mathcal{F}_{n-1}] \xrightarrow{n\to+\infty} \mathcal{V} \quad \text{in probability,}$$

where σ^2 is a positive symmetric $d \times d$-matrix. Then, for every C^2-function $g : \mathbb{R}^{2d} \to \mathbb{R}$, compactly supported with Lipschitz continuous second derivatives, we have:

$$\mathbb{E}(g(\bar{Z}_{n+1}) - g(\bar{Z}_n)|\mathcal{F}_n) = \gamma_{n+1}\mathcal{L}g(\bar{Z}_n) + R_n^g$$

where $\gamma_{n+1}^{-1} R_n^g \to 0$ in L^1 and $\mathcal{L}$ is the infinitesimal generator defined in Theorem 2.4.

Remark 5.1. *We recall that $\mathcal{L}$ is the infinitesimal generator of the following stochastic differential equation:*

$$d\bar{Z}_t = \bar{H}\bar{Z}_t dt + \Sigma dB_t$$

where: $\bar{H} = \frac{1}{2\gamma} 1_{\{\beta=1\}} I_{2d} + H$ and Σ is defined in Theorem 2.4. $(\bar{Z}_t)_{t\geqslant 0}$ lies in the family of Ornstein-Uhlenbeck processes: on the one hand, the drift and diffusion coefficients being respectively linear and constant, $(\bar{Z}_t)_{t\geqslant 0}$ is a Gaussian diffusion; on the other hand, since $\bar{H}$ has negative eigenvalues, $(\bar{Z}_t)_{t\geqslant 0}$ is ergodic.

Proof of Lemma 5.2:
C will denote an absolute constant whose value may change from line to line, for the sake of convenience. We use a Taylor expansion between $\bar{Z}_n$ and $\bar{Z}_{n+1}$ and obtain that θ_n exists in $[0, 1]$ such that:

$$g(\bar{Z}_{n+1}) - g(\bar{Z}_n) = \langle \nabla g(\bar{Z}_n), (\bar{Z}_{n+1} - \bar{Z}_n)\rangle + \frac{1}{2}(\bar{Z}_{n+1} - \bar{Z}_n)^T D^2 g(\bar{Z}_n)(\bar{Z}_{n+1} - \bar{Z}_n) + \underbrace{\frac{1}{2}(\bar{Z}_{n+1} - \bar{Z}_n)^T (D^2 g(\theta \bar{Z}_n + (1-\theta)\bar{Z}_{n+1}) - D^2 g(\bar{Z}_n))(\bar{Z}_{n+1} - \bar{Z}_n)}_{R^{(1)}_{n+1}}. \tag{5.5}$$

We first deal with the remainder term $R^{(1)}_{n+1}$ and observe that $(\bar{C}_n)$ introduced in (5.2) is uniformly bounded so that a constant C exists such that $\|b_n(z)\| \leqslant C\|z\|$. We thus conclude that:

$$\|\bar{Z}_{n+1} - \bar{Z}_n\| \leqslant C\left(\gamma_{n+1}\|\bar{Z}_n\| + \sqrt{\gamma_{n+1}}\|\Delta M_{n+1}\|\right).$$

Using $(\mathbf{H}_{\boldsymbol{\sigma},\boldsymbol{p}})$, we deduce that for any $\bar{p} \leqslant p$,

$$\mathbb{E}\left[\|\bar{Z}_{n+1} - \bar{Z}_n\|^{\bar{p}}\right] \leqslant C\gamma_{n+1}^{\frac{\bar{p}}{2}}. \tag{5.6}$$

Since $D^2 g$ is Lipschitz continuous and compactly supported, $D^2 g$ is also ε-Hölder for all $\varepsilon \in (0,1]$. We choose ε such that $2 + \varepsilon \leqslant p$ and obtain:

$$\mathbb{E}\left[|R_{n+1}|\right] \leqslant C\mathbb{E}\left[\|\bar{Z}_{n+1} - \bar{Z}_n\|^{2+\varepsilon}\right] \leqslant C\gamma_{n+1}^{1+\frac{\varepsilon}{2}}.$$

We deduce that $\gamma_{n+1}^{-1} R^{(1)}_{n+1} \to 0$ in L^1. ◇

Second, we can express (5.3) when $\gamma_n = \gamma n^{-\beta}$ with $\beta \in (0,1]$ in the following form:

$$\epsilon_n := \frac{1}{\gamma_{n+1}}\left(\sqrt{\frac{\gamma_n}{\gamma_{n+1}}} - 1\right) - \frac{1}{2\gamma}1_{\{\beta=1\}} = o(1).$$

Then, given that $D^2 f$ is Lipschitz (and that $x^\star = 0$), it follows that:

$$\forall z \in \mathbb{R}^d \times \mathbb{R}^d \qquad \left\| b_n(z) - \left(\frac{1}{2\gamma}1_{\{\beta=1\}} I_{2d} + H\right) z \right\| \leqslant (\varepsilon_n + \|\bar{X}_n\|)\|z\|$$

where $(\varepsilon_n)_{n\geqslant 1}$ is a deterministic sequence such that $\lim_{n\to+\infty} \varepsilon_n = 0$.

Under the conditions of Theorem 2.4, we may apply the convergence rates obtained in Theorem 2.3 and observe that $\sup_n \mathbb{E}[\|X_n\|^2] \lesssim \gamma_n$, meaning that $\sup_n \mathbb{E}[\|\bar{Z}_n\|^2] < +\infty$. Since $\|\bar{X}_n\| \leqslant \|\bar{Z}_n\|$, we deduce that:

$$\mathbb{E}[\langle \nabla g(\bar{Z}_n), (\bar{Z}_{n+1} - \bar{Z}_n)\rangle | \mathcal{F}_n] = \gamma_{n+1}\langle \nabla g(\bar{Z}_n), (\frac{1}{4\gamma\sqrt{r}}1_{\{\beta=1\}} I_{2d} + H)\bar{Z}_n\rangle + R^{(2)}_n$$

where $\gamma_{n+1}^{-1} R^{(2)}_n \to 0$ in L^1 as $n \to +\infty$. Let us now consider the second term of the right-hand side of (5.5). We have:

$$\mathbb{E}[(\bar{Z}_{n+1} - \bar{Z}_n)^T D^2 g(\bar{Z}_n)(\bar{Z}_{n+1} - \bar{Z}_n)|\mathcal{F}_n] = \gamma_{n+1}\sum_{i,j} D^2_{y_i y_j} g(\bar{Z}_n)\mathbb{E}[\Delta M^i_{n+1}\Delta M^j_{n+1}|\mathcal{F}_n] + R^{(3)}_n$$

where

$$|\gamma_{n+1}^{-1} R_n^{(3)}| \leqslant C\gamma_{n+1}\|\bar{Z}_n\|^2 \xrightarrow{n\to+\infty} 0 \quad \text{in } L^1$$

under the assumptions of the lemma. To conclude the proof, it remains to note that under the assumptions of the lemma for any i and j, $(\mathbb{E}[\Delta M_{n+1}^i \Delta M_{n+1}^j | \mathcal{F}_n])_{n\geqslant 1}$ is a uniformly integrable sequence that satisfies:

$$\mathbb{E}[\Delta M_{n+1}^i \Delta M_{n+1}^j | \mathcal{F}_n] = \mathcal{V}_{i,j} \quad \text{in probability.}$$

Thus, the convergence also holds in L^1. The conclusion of the lemma easily follows from the boundedness of $D^2 g$. ◇□

We are now able to prove Theorem 2.4:

Proof of Theorem 2.4, (i): Note that under the assumptions of Theorem 2.4, we can apply Lemma 5.1 and Lemma 5.2 and obtain that the sequence of processes $(\bar{Z}^{(n)})_{n\geqslant 1}$ is tight. The rest of the proof is then divided into two steps. In the first one, we prove that every weak limit of $(\bar{Z}^{(n)})_{n\geqslant 1}$ is a solution of the martingale problem $(\mathcal{L}, \mathcal{C})$ where $\mathcal{C}$ denotes the class of $\mathcal{C}^2$-functions with compact support and Lipschitz-continuous second derivatives. Before going further, let us recall that, owing to the Lipschitz continuity of the coefficients, this martingale problem is well-posed, *i.e.*, tha,t existence and uniqueness hold for the weak solution starting from a given initial distribution μ (see, *e.g.*, [EK86] or [SV06]).

In a second step, we prove the uniqueness of the invariant distribution related to the operator $\mathcal{L}$ and the convergence in distribution to this invariant measure. We end this proof by showing that $(\bar{Z}^{(n)})$ converges to this invariant distribution, so that the sequence $(\bar{Z}^{(n)})_{n\geqslant 1}$ converges to a stationary solution of the previously introduced martingale problem. We will characterize this invariant (Gaussian) distribution in the next paragraph.

Step 1: Let g belong to $\mathcal{C}$ and let $(\mathcal{F}_t^{(n)})_{t\geqslant 0}$ be the natural filtration of $\bar{Z}^{(n)}$. To prove that any weak limit of $(\bar{Z}^{(n)})_{n\geqslant 1}$ solves the martingale problem $(\mathcal{L}, \mathcal{C})$, it is enough to show that:

$$\forall t \geqslant 0, \quad g(\bar{Z}_t^{(n)}) - g(\bar{Z}_0^{(n)}) - \int_0^t \mathcal{L} g(\bar{Z}_s^{(n)}) ds = \mathcal{M}_t^{(n,g)} + \mathcal{R}_t^{(n,g)}$$

where $(\mathcal{M}_t^{(n,g)})_{t\geqslant 0}$ is an $(\mathcal{F}_t^{(n)})$-adapted martingale and $\mathcal{R}_t^{(n,g)} \to 0$ in probability for any $t \geqslant 0$. We set:

$$\mathcal{M}_t^{(n,g)} = \sum_{k=n+1}^{N(n,t)} g(\bar{Z}_{k+1}) - g(\bar{Z}_k) - \mathbb{E}[g(\bar{Z}_{k+1}) - g(\bar{Z}_k) | \mathcal{F}_{k-1}].$$

By construction, $(\mathcal{M}_t^{(n,g)})_{t\geqslant 0}$ is an $(\mathcal{F}_t^{(n)})$-adapted martingale (given that $\mathcal{F}_s^{(n)} = \mathcal{F}_{\underline{s}_n}^{(n)}$) and:

$$\mathcal{R}_t^{(n,g)} = g(\bar{Z}_t^{(n)}) - g(\bar{Z}_{\underline{t}_n}^{(n)}) - \int_{\underline{t}_n}^{t} \mathcal{L}g(\bar{Z}_s^{(n)})ds + \int_0^{\underline{t}_n} \left(\mathcal{L}g(\bar{Z}_{\underline{s}_n}^{(n)}) - \mathcal{L}g(\bar{Z}_s^{(n)})\right) ds + \sum_{k=n}^{N(n,t)-1} R_k^g$$

where $(R_k^g)_{k\geqslant 1}$ has been defined in Lemma 5.2. Using an argument similar to (5.6), we can check that for any $t \geqslant 0$:

$$\sup_{s\leqslant t} \mathbb{E}[\|\bar{Z}_s^{(n)} - \bar{Z}_{\underline{s}_n}^{(n)}\|^2] \leqslant C\sqrt{\gamma}_n.$$

This inequality combined with the Lipschitz continuity of g and its derivatives implies that the first three terms tend to 0 when $n \to +\infty$. Now, concerning the last one, the previous lemma yields:

$$\mathbb{E}\left[\left|\sum_{k=n}^{N(n,t)-1} R_k^g\right|\right] \leqslant Ct \sup_{k\geqslant n} \mathbb{E}\left[\left|\gamma_k^{-1} R_k^g\right|\right] \xrightarrow{n\to+\infty} 0. \qquad \diamond$$

Step 2: First, let us prove that uniqueness holds for the invariant distribution related to $\mathcal{L}$. We denote it by $\mu_\infty^{(\beta)}$ below. In this simple setting where the coefficients are linear, we could use the fact that the process, which is solution to the martingale problem, is Gaussian so that any invariant distribution is so. Uniqueness could then be deduced through the characterization of the mean and the variance through the relationship $\int \mathcal{L}f(x)\mu_\infty^{(\beta)}(dx) = 0$ (see next subsection for such an approach). However, at this stage, we prefer to use a more general strategy related to the hypoellipticity of $\mathcal{L}$ (see, *e.g.*, [GP14] for a similar approach). More precisely, set $L_D := -\langle y, \partial_x\rangle + r\langle D^2 f(x^\star)x - y], \partial_y\rangle$ and $\sigma_i := \sum_{j=1}^d \sigma_i^j \partial_{y_j}$, where σ satisfies $\sigma\sigma^t = \mathcal{V}$ (where $\mathcal{V}$ is defined by (2.7)). We have assumed that σ is invertible, so that:

$$\text{span}(\sigma_1, \ldots, \sigma_d) = \text{span}(\partial_{y_1}, \ldots, \partial_{y_d}).$$

Therefore,

$$\text{Lie}\,(L_D, \sigma_1, \ldots, \sigma_d) = \text{Lie}\,(L_D, \partial_{y_1}, \ldots, \partial_{y_d})$$

Now, it is straightforward to check that:

$$\forall i \in \{1, \ldots, d\} \qquad [L_D, \partial_{y_i}](f) = -\partial_{x_i}(f),$$

and we deduce that $\text{Lie}\,(L_D, \sigma_1, \ldots, \sigma_d) = \text{Lie}\,(\partial_{x_1}, \ldots, \partial_{x_d}, \partial_{y_1}, \ldots, \partial_{y_d})$. This means that the Hormandër bracket condition holds at any point z of $\mathbb{R}^{2d}$, which implies that the process admits a density $(p_t(z,.))_{t\geqslant 0}$ such that for any $t > 0$, $(z, z') \mapsto p_t(z, z')$, which is smooth on $\mathbb{R}^{2d} \times \mathbb{R}^{2d}$. It is moreover possible to show that these densities are positive, for any $t > 0$, given that the linear vector field is approximately controllable: for any time $T > 0$, any $\eta > 0$ and any couple

of initial points (x_0, y_0) and ending points (x_T, y_T), we can build a function φ such that $\dot{\varphi} \in \mathbb{L}^2$ and such that the controlled trajectory:

$$\left\{ \begin{array}{rcl} \dot{x}(t) & = & -y(t) \\ \dot{y}(t) & = & r(t)(\nabla U(x(t)) - y(t)) + \sigma\dot{\varphi}, \end{array} \right. \tag{5.7}$$

satisfies: $z_0 = (x_0, y_0)$ and $\|z_T - (x_T, y_T)\| \leqslant \eta$. This implies the irreducibility of the diffusion and, therefore, the uniqueness of the invariant distribution. We refer to [GP14] for more details on this controllability problem.

Then, checking that $\mathcal{L}\|x\|^2 \leqslant \beta - \alpha\|x\|^2$ for positive α and β, it can be classically deduced from the Meyn-Tweedie-type arguments (see [MT93]) that the process converges locally uniformly, exponentially fast in total variation to $\mu_\infty^{(\beta)}$. For more details, we refer to [MSH02, Theorem 4.4]. Below, we will only use the following corollary: for any bounded Lipschitz-continuous function f, for any compact set K of $\mathbb{R}^{2d}$,

$$\sup_{z \in K} |P_t f(z) - \mu_\infty^{(\beta)}(f)| \xrightarrow{t \to +\infty} 0 \tag{5.8}$$

where $(P_t)_{t \geqslant 0}$ denotes the semi-group related to the (well-posed) martingale problem $(\mathcal{L}, \mathcal{C})$. ◇

Step 3: Let $(\bar{Z}_{n_k})_{k \geqslant 1}$ be a (weakly) convergent subsequence of $(\bar{Z}_n)_{n \geqslant 1}$ to a probability ν. We have to prove that $\nu = \mu_\infty^{(\beta)}$. To do this, we take advantage of the "shifted" construction of the sequence $(\bar{Z}^{(n)})_{n \in \mathbb{N}}$. More precisely, as a result of construction, for any positive T, a sequence $(\psi(n_k, T))_{k \geqslant 1}$ exists such that:

$$N(T, \psi(n_k, T)) = n_k.$$

In other words,

$$\bar{Z}^{(\psi(n_k,T))}_{\underline{T}_{\psi(n_k,T)}} = \bar{Z}_{n_k}.$$

At the price of a potential extraction, $(\bar{Z}^{(\psi(n_k,T))})_{k \geqslant 1}$ is convergent to a continuous process, which is denoted by $Z^{\infty,T}$ below. Given that $\bar{Z}^{(n)}_T - \bar{Z}^{(n)}_{\underline{T}_n}$ tends to 0 as $n \to +\infty$ in probability, it follows that $Z^{\infty,T}_T$ has distribution ν. However, according to Step 1, $Z^{\infty,T}$ is also a solution to the martingale problem $(\mathcal{L}, \mathcal{C})$ so that for any Lipschitz continuous function f,

$$\mathbb{E}[f(Z^{\infty,T}_T)] - \mu_\infty^{(\beta)}(f) = \int_{\mathbb{R}^{2d}} \Big(P_T f(z) - \mu_\infty^{(\beta)}(f) \Big) \, \mathbb{P}_{Z^{\infty,T}_0}(dz).$$

Denote by $\mathcal{P}$, the set of weak limits of $(\bar{Z}_n)_{n \geqslant 1}$. $\mathcal{P}$ is tight and as a result of construction, $Z^{\infty,T}_0$ belongs to $\mathcal{P}$. Thus, for any $\varepsilon > 0$, a compact set K_ε exists such that for any $T > 0$,

$$\left| \int_{K_\varepsilon^c} \Big(P_T f(z) - \mu_\infty^{(\beta)}(f) \Big) \, \mathbb{P}_{Z^{\infty,T}_0}(dz) \right| \leqslant 2\|f\|_\infty \sup_{\mu \in \mathcal{P}} \mu(K_\varepsilon^c) \leqslant 2\|f\|_\infty \varepsilon.$$

On the other hand,

$$\left|\int_{K_\varepsilon}\left(P_T f(z)-\mu_\infty^{(\beta)}(f)\right)\mathbb{P}_{Z_0^{\infty,T}}(dz)\right| \leqslant \sup_{z\in K_\varepsilon}|P_T f(z)-\mu_\infty^{(\beta)}(f)|$$

and it follows from Step 2 that the right-hand member tends to 0 as $T\to+\infty$. From this, we can therefore conclude that for any bounded Lipschitz-continuous function f, a large enough T exists such that:

$$\left|\mathbb{E}[f(Z_T^{\infty,T})]-\mu_\infty^{(\beta)}(f)\right| \leqslant C_f\varepsilon.$$

Since $\mathbb{E}[f(Z_T^{\infty,T})]=\nu(f)$, it follows that $\nu(f)=\mu_\infty^{(\beta)}(f)$. Finally, the set $\mathcal{P}$ is reduced to a single element $\mathcal{P}=\{\mu_\infty^{(\beta)}\}$, and the whole sequence $(\bar{Z}_n)_{n\geqslant 1}$ converges to $\mu_\infty^{(\beta)}$.

Before ending this section, let us note that $\mu_\infty^{(\beta)}$ is a Gaussian centered distribution is a simple consequence of Remark 5.1. We therefore leave this point to the reader. ◇□

5.4. Limit variance

We end this section on the analysis of the rescaled algorithm with some considerations on the invariant measure $\mu_\infty^{(\beta)}$ involved in Theorem 2.4 for the exponential memoried stochastic HBF, *i.e.* when $r_n=r$. As shown in the above paragraph, this invariant measure describes the exact asymptotic variance of the initial algorithm. We now focus on its characterization *i.e.*, on the proof of Theorem 2.4(*ii*). In particular, to ease the presentation, we assume that the covariance matrix $\mathcal{V}$ related to $(\Delta M_{n+1})_{n\geqslant 1}$ is proportional to the identity matrix:

$$\lim_{n\longrightarrow+\infty}\mathbb{E}\left[\Delta M_{n+1}\left(\Delta M_{n+1}\right)^t|\mathcal{F}_n\right]=\sigma_0^2 I_d \quad \text{in probability.} \tag{5.9}$$

We also assume that $\gamma_n=\gamma n^{-\beta}$ with $\beta<1$. Then, (i) of Theorem 2.4 states that $(\bar{Z}_n)_{n\geqslant 1}$ weakly converges toward a diffusion process, whose generator $\mathcal{L}$ is the one of an Ornstein-Uhlenbeck process. Assumption (5.9) leads to a simpler expression:

$$\mathcal{L}(\phi)(x,y)=-\langle y,\nabla_x\phi\rangle+r\langle D^2(f)(x^\star)x-y,\nabla_y\phi\rangle+r^2\frac{{\sigma_0}^2}{2}\Delta_y\phi. \tag{5.10}$$

A particular feature of Equation (5.10) when $\gamma_n=\gamma n^{-\beta}$ is that $\mathcal{L}$ does not depend on β nor γ. The invariant measure $\mu_\infty^{(\beta)}$ is a multivariate Gaussian distribution that may be well described in the basis given by the eigenvectors of the Hessian $D^2(f)(x^\star)$. The reduction to d couples of two-dimensional system used in Section 4.1.1 makes it possible to use the spectral decomposition of $D^2(f)(x^\star)=P^{-1}\Lambda P$ where P is an orthonormal matrix and Λ a diagonal matrix with positive eigenvalues. The process $(\check{X}_n,\check{Y}_n)=(P\bar{X}_n,P\bar{Y}_n)$ is therefore

centered and Gaussianly distributed asymptotically. This process is associated with d 2×2 blockwise independent Ornstein-Uhlenbeck processes, whose generator is now

$$\check{\mathcal{L}}(\phi)(\check{x},\check{y}) = -\langle \check{y}, \nabla_{\check{x}}\phi\rangle + r\langle \Lambda\check{x}-\check{y}, \nabla_{\check{y}}\phi\rangle + r^2\frac{{\sigma_0}^2}{2}\Delta_{\check{y}}\phi,$$

where we used $\mathrm{Tr}\left(P^tD^2_{\check{y}}P\right) = \mathrm{Tr}\left(D^2_{\check{y}}PP^t\right) = \mathrm{Tr}\left(D^2_{\check{y}}\right)$ in the last line because $P^tP = I_d$. If we denote $\check{\mu}^{(\beta)}_\infty$ the associated invariant gaussian measure, the tensor structure of $\check{\mathcal{L}}$ leads to

$$\forall i\neq j \qquad \mathbb{E}_{(\check{x},\check{y})\sim\check{\gamma}^\beta_\infty}[\check{x}^{(i)}\check{x}^{(j)}] = \mathbb{E}_{(\check{x},\check{y})\sim\check{\gamma}^\beta_\infty}[\check{x}^{(i)}\check{y}^{(j)}] = \mathbb{E}_{(\check{x},\check{y})\sim\check{\gamma}^\beta_\infty}[\check{y}^{(i)}\check{y}^{(j)}] = 0. \tag{5.11}$$

Now, using the relationship $\int \check{\mathcal{L}}(\phi)d\check{\mu}^{(\beta)}_\infty = 0$ for some well chosen functions ϕ, we can identify the rest of the covariance matrix. Denote i any integer in $\{1,\dots,d\}$. We chose $\phi(\check{x},\check{y}) = \frac{\left\{\check{x}^{(i)}\right\}^2}{2}$ and obtain that $\check{\mathcal{L}}\left(\frac{\left\{\check{x}^{(i)}\right\}^2}{2}\right)(\check{x},\check{y}) = -\check{x}^{(i)}\check{y}^{(i)}$. It then implies that

$$\mathbb{E}_{(\check{x},\check{y})\sim\check{\mu}^{(\beta)}_\infty}[\check{x}^{(i)}\check{y}^{(i)}] = 0. \tag{5.12}$$

Picking now $\phi(\check{x},\check{y}) = \frac{\left\{\check{y}^{(i)}\right\}^2}{2}$, we obtain $\check{\mathcal{L}}\left(\frac{\left\{\check{y}^{(i)}\right\}^2}{2}\right)(\check{x},\check{y}) = r\lambda_i\check{x}^{(i)}\check{y}^{(i)} - r\left\{\check{y}^{(i)}\right\}^2 + \frac{r^2{\sigma_0}^2}{2}$ so that

$$\mathbb{E}_{(\check{x},\check{y})\sim\check{\mu}^{(\beta)}_\infty}[\{\check{y}^{(i)}\}^2] = \frac{r{\sigma_0}^2}{2}. \tag{5.13}$$

Finally, we chose $\phi(\check{x},\check{y}) = \check{x}^{(i)}\check{y}^{(i)}$ and obtain $\check{\mathcal{L}}\left(\check{x}^{(i)}\check{y}^{(i)}\right)(\check{x},\check{y}) = -\left\{\check{y}^{(i)}\right\}^2 + r\lambda_i\{\check{x}^{(i)}\}^2 - r\check{x}^{(i)}\check{y}^{(i)}$. Therefore, we deduce that:

$$\mathbb{E}_{(\check{x},\check{y})\sim\check{\mu}^{(\beta)}_\infty}[\{\check{x}^{(i)}\}^2] = \frac{{\sigma_0}^2}{2\lambda_i}. \tag{5.14}$$

We can sum-up formulae (5.11)-(5.14) in $\check{\mu}^{(\beta)}_\infty = \mathcal{N}(0, D_{r,\sigma_0})$ with $D_{r,\sigma_0} = \frac{{\sigma_0}^2}{2}\begin{pmatrix}\Lambda^{-1} & \mathbf{0}_{d\times d}\\ \mathbf{0}_{d\times d} & rI_d\end{pmatrix}$. Since $(\bar{X}_n,\bar{Y}_n) = (P^{-1}\check{X}_n, P^{-1}\check{Y}_n)$, we deduce that:

$$\mu^{(\beta)}_\infty = \mathcal{N}\left(0, \frac{{\sigma_0}^2}{2}\begin{pmatrix}\{D^2f(x^\star)\}^{-1} & \mathbf{0}_{d\times d}\\ \mathbf{0}_{d\times d} & rI_d\end{pmatrix}\right). \qquad \diamond$$

Proof of Theorem 2.4- Step size $\gamma_n = \gamma n^{-1}$
This situation is more involved since we can observe that the drift of the limit diffusion is modified according to the size of γ. In particular, the generator $\mathcal{L}$ in that case is shifted from the one above by $\frac{1}{2\gamma}I$ so that:

$$\mathcal{L}(\phi)(x,y) = \frac{1}{2\gamma}\left[\langle\nabla_x\phi,x\rangle + \langle\nabla_y\phi,y\rangle\right] - \langle y,\nabla_x\phi\rangle + r\langle D^2f(x^\star)x - y, \nabla_y\phi\rangle + r^2\frac{{\sigma_0}^2}{2}\Delta_y\phi.$$

Again, we can use the decomposition $D^2 f(x^\star) = P^{-1}\Lambda P$ where P is an orthonormal matrix, and the generator of the rotated process $(\check{X}_n, \check{Y}_n) = (P\bar{X}_n, P\bar{Y}_n)$ is:

$$\check{\mathcal{A}}(\phi)(x,y) = \left\langle \frac{\check{x}}{2\gamma} - \check{y}, \nabla_{\check{x}}\phi \right\rangle + \left\langle r\Lambda\check{x} + \left(\frac{1}{2\gamma} - r\right)\check{y}, \nabla_{\check{y}}\phi \right\rangle + r^2 \frac{{\sigma_0}^2}{2}\Delta_{\check{y}}\phi.$$

The associated Ornstein-Uhlenbeck process has a unique Gaussian invariant measure $\check{\mu}_\infty^{(1)}$ if and only if $\gamma\alpha_r > 1$ where α_r is the constant defined in the statement of Proposition 4.1. The following equations still hold:

$$\forall i \neq j \qquad \mathbb{E}_{(\check{x},\check{y})\sim\check{\mu}_\infty^{(1)}}[\check{x}^{(i)}\check{x}^{(j)}] = \mathbb{E}_{(\check{x},\check{y})\sim\check{\mu}_\infty^{(1)}}[\check{x}^{(i)}\check{y}^{(j)}] = \mathbb{E}_{(\check{x},\check{y})\sim\check{\mu}_\infty^{(1)}}[\check{y}^{(i)}\check{y}^{(j)}] = 0. \tag{5.15}$$

To determine the rest of the covariance matrix, we follow the same strategy and only address the case $d = 1$ for the sake of convenience. We define: $\sigma_x^2 := \mathbb{E}_{(\check{x},\check{y})\sim\check{\mu}_\infty^{(1)}}\left[\check{x}^2\right]$, $\sigma_y^2 := \mathbb{E}_{(\check{x},\check{y})\sim\check{\mu}_\infty^{(1)}}\left[\check{y}^2\right]$ and $\sigma_{x,y} := \mathbb{E}_{(\check{x},\check{y})\sim\check{\mu}_\infty^{(1)}}\left[\check{x}\check{y}\right]$.

We start by chosing $\phi(\check{x},\check{y}) = \frac{\check{x}^2}{2}$ and obtain $\check{\mathcal{A}}(\phi)(x,y) = \frac{\check{x}^2}{2\gamma} - \check{x}\check{y}$. Therefore, we deduce that:

$$2\gamma\,\sigma_{x,y} = \sigma_x^2 \tag{5.16}$$

Now we pick $\phi(\check{x},\check{y}) = \frac{\check{y}^2}{2}$ and obtain $\check{\mathcal{A}}(\phi)(x,y) = r\lambda\check{x}\check{y} + \left(\frac{1}{2\gamma} - r\right)\check{y}^2 + \frac{r^2{\sigma_0}^2}{2}$ so that:

$$\left(r - \frac{1}{2\gamma}\right)\sigma_y^2 = r\lambda\sigma_{x,y} + \frac{r^2{\sigma_0}^2}{2}. \tag{5.17}$$

Finally, the function $\phi(\check{x},\check{y}) = \check{x}\check{y}$ yields $\check{\mathcal{A}}(\phi)(x,y) = \check{x}\check{y}\left(\frac{1}{\gamma} - r\right) - \check{y}^2 + r\lambda\check{x}^2$, which implies:

$$\sigma_y^2 = r\lambda\sigma_x^2 + \left(\frac{1}{\gamma} - r\right)\sigma_{x,y} \tag{5.18}$$

We are led to the introduction of:

$$\check{\alpha}_- = 1 - \sqrt{1 - \frac{4\lambda}{r}} \quad \text{and} \quad \check{\alpha}_+ = 1 + \sqrt{1 - \frac{4\lambda}{r}},$$

which leads to:

$$\sigma_x^2 = {\sigma_0}^2 \frac{2\lambda r\gamma^3}{(\gamma r - 1)(2\lambda\gamma - \check{\alpha}_-)(2\lambda\gamma - \check{\alpha}_+)}, \quad \sigma_y^2 = {\sigma_0}^2 \frac{\lambda r\gamma(2\lambda r\gamma^2 - r\gamma + 1)}{(\gamma r - 1)(2\lambda\gamma - \check{\alpha}_-)(2\lambda\gamma - \check{\alpha}_+)},$$

and

$$\sigma_{x,y} = {\sigma_0}^2 \frac{\lambda r\gamma^2}{(\gamma r - 1)(2\lambda\gamma - \check{\alpha}_-)(2\lambda\gamma - \check{\alpha}_+)}. \qquad \diamond\ \square$$

6. Numerical experiments

In this paragraph, the aim is to provide numerical tests related to the HBF-algorithm. In a first part, we mainly provide illustrations of some of the theoretical results established previously. Then, we focus on some numerical comparisons with other algorithms widely used in the stochastic approximation field. In particular, we are interested in the convergence rates of each algorithm, as well as their behavior in the setting of non-convex potential f with multiple wells, which is covered by Theorems 2.1 and 2.2.

6.1. About L^2-convergence rates

In Theorem 2.3, we proved that under appropriate conditions, the Mean-Squared Error (MSE) related to the algorithm is $O(\gamma_n)$ when $\gamma_n = \gamma n^{-\beta}$ with $\beta < 1$ and, in the exponential case, the optimal order $O(1/n)$ can be attained when $\beta = 1$ but under conditions on r, γ and the Hessian matrix at the minimum. Note that such types of conditions also appear when $\beta = 1$ in the classical stochastic gradient descent (and can be classically overcame by a Ruppert-Polyak averaging). Figure 1 illustrates some of these properties in the exponential and polynomial cases with the toy-example $f(x) = \frac{x^2}{2}$ in the one-dimensional case. In Figure 1, we focus on the behavior of the (Monte-Carlo estimated) MSE for different values of r by computing $M = 100$ paths until $n = 10^5$ starting from $(X_0, Y_0) = (10, 0)$ with $\sigma = 1$ and $\Delta M_n = Z_n$ where $(Z_n)_{n\geqslant 1}$ is a sequence of *i.i.d.* $\mathcal{N}(0,1)$-distributed random variables. In order to get a more readable illustration of the rate of convergence, we represent the behavior in a logarithmic scale, *i.e.*, we draw the graph of the Monte-Carlo estimation of $\log p \mapsto \log(\mathbb{E}[|X_p - x^\star|^2])$ (here, $|X_k - x^\star|^2 = 2f(X_k)$) and look at the influence of r. Note that in order to avoid some numerical instability of the algorithm at the beginning of the iterations for large values of r (especially when $r = 50$), we introduced an additional truncation trick for the step-size sequence, *i.e*, we computed the algorithm with the sequence $(\gamma_n^{(r)})_{n\geqslant 1}$ instead of $(\gamma_n)_{n\geqslant 1}$, defined by

$$\gamma_n^{(r)} = \min\left(\gamma_n, \frac{1}{r}\right). \tag{6.1}$$

Such modification is classical in numerical investigations related to discretizations of continuous dynamics (see *e.g* [Lem07] for instance in a diffusion setting).

In the exponential case (on the left of Figure 1), the algorithm is computed with $\gamma_k = k^{-1}$. For $r = 2, 5, 10, 50$, the behavior seems to be robust and mainly reproduces the theoretical decrease in $\frac{1}{n}$ obtained in Theorem 2.3 (a) as indicated by the estimated slope of the evolution on the log-log scale, which holds as soon as $\gamma\alpha_r > 1$. For our considered function, it is straigthforward to check that

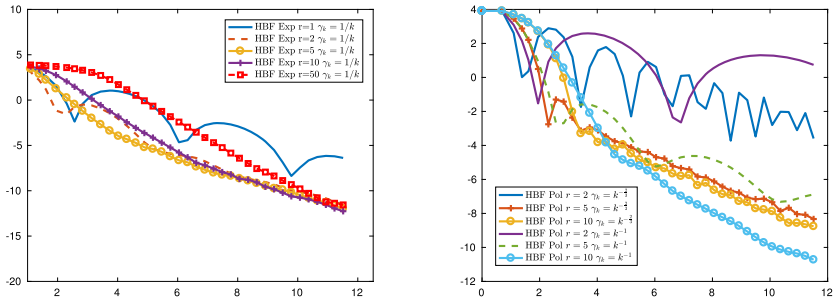

FIG 1. *Evolution of* $\log(\mathbb{E}[|X_k - x^*|^2])$ *with respect to* $\log(k)$. *Left: Exponential memory. Right: Polynomial memory.*

$$\gamma\alpha_r = r\left(1 - \sqrt{\frac{4 \wedge r}{r}}\right),$$

and the condition $\gamma\alpha_r > 1$ holds only when $r = 2, 5, 10, 50$. Oppositely, when $r = 1$, we then obtain $\gamma\alpha_r = 1$, which is illustrated by the worse performances obtained in this case. The algorithm also possesses a lengthy oscillating behavior which is coherent with the complex eigenvalues of the second-order underlying linear differential system (see *e.g.* (4.7)). In the polynomial case (on the right of Figure 1), we focus our attention on the dependency in r for two choices of steps: $\gamma_k = k^{-\beta}$ for $\beta = \frac{2}{3}$ and $\beta = 1$. Once again, we observe in the two cases some oscillations for small values of r (which follow from the same arguments). When $\beta = 2/3$, Theorem 2.3 (b) states that the MSE is $O(\gamma_n)$ if $r > 5/2$, which explains the bad performances when $r = 2$. When $\beta = 1$, our theoretical bound is $O(\frac{1}{\log n})$. However, from a numerical point of view, it seems that, when r is large enough, the behavior is rather close to $O(1/n)$, as it is the case for the exponential case.

6.2. About the central limit theorem

In Theorem 2.4, we stated a CLT for exponential memory under conditions on the parameters of the algorithm that are similar to those of Theorem 2.3. In the context of the previous paragraph, we illustrate Theorem 2.4 in Figure 2. More precisely, on the left side, we compare the estimated density of $X_n/\sqrt{\gamma_n}$ (built by convolution with a Gaussian kernel) with the theoretical one given in Equation (2.8). Then, on the right side, we consider the polynomial case for which no theoretical result has been proved. In fact, by drawing the evolution of $n \mapsto \gamma_n^{-1}\mathbb{E}[|X_n - x^\star|^2]$, we remark that the second moment has an oscillating behavior, which suggests that the convergence in distribution of $(X_n - x^\star)/\sqrt{\gamma_n}$ does not hold in this case, otherwise we should observe a convergence when n increases.

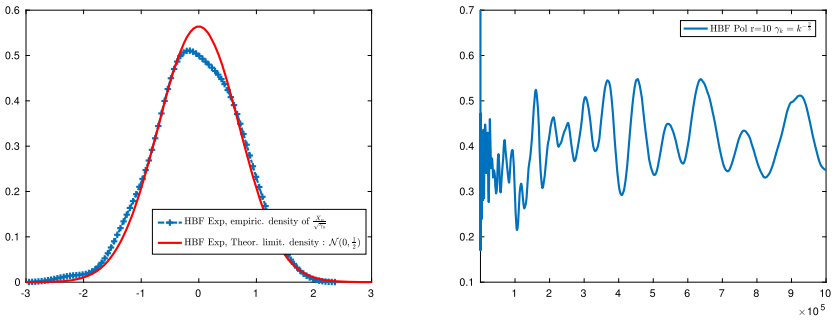

FIG 2. *Left (Exponential) Estimation of the density of* $\frac{X_n}{\sqrt{\gamma_n}}$. *Right (Polynomial): Behavior of* $n \mapsto \gamma_n^{-1}\mathbb{E}[|X_n - x^\star|^2]$.

6.3. Comparisons with other algorithms

Figure 3 compares the HBF-algorithm with several stochastic optimization algorithms: the standard Robbins-Monro stochastic gradient descent (SGD) introduced in [RM51] and several second order algorithms: the "optimal" Ruppert-Polyak averaging algorithm (see [PJ92, Rup88]), the Nesterov accelerated gradient descent [Nes83] adapted in the stochastic framework in a straightforward way using an unbiased evaluation of the gradient in each iteration, and the recent SAGE method introduced in [HPK09]. Note that the Ruppert-Polyak averaging algorithm is used according to the recommendation of [Bac14] with $\gamma_k = k^{-1/2}$. As in the previous parts, $\sigma = 1$ and (ΔM_n) is a sequence of *i.i.d.* $\mathcal{N}(0,1)$ random variables. Finally, the function to optimize is $f(x) = |x|^p/p$ with two values of p: $p = 2$ (strongly convex situation) and $p = 4$ (convex situation, the Hessian being degenerated at 0).

The first elementary remark is that the rate is of course deteriorated by the loss of strong convexity (left side, Figure 3). In this case, the Ruppert-Polyak averaging outperforms other methods and attains the $O(1/\sqrt{n})$ minimax rate (see [NY83]). When f is strongly convex, the second-order algorithms then all share an equivalent efficiency with, apparently a $\mathcal{O}(1/n)$ convergence rate. This corresponds to (ii) of Theorem 2.3 when the Hessian at the critical point is sufficiently large to make this minimax optimal rate possible. Nevertheless, the ability of the stochastic heavy ball in a more general situation may deserve further numerical investigation, which is beyond the scope of this paper. The SGD seems to be a little bit less effective in the strongly convex case. Finally, the Nesterov adaptation to the stochastic case does not lead to an efficient algorithm (in comparison to the other methods tested). However, this remark should be balanced by the fact that we did not use the Lan adaptation of the Nesterov accelerated gradient descent introduced in [Lan12]. It appears that this modification that consists in an addition of an intermediary point in the NAGD seems important to optimize the behavior of the algorithm in the stochastic case.

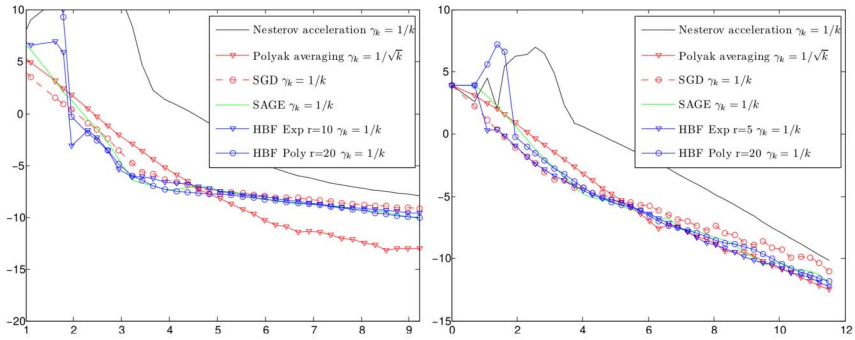

FIG 3. *Evolution of* $\log(\mathbb{E}[f(X_k)])$ *with respect to* $\log(k)$ *with* $f(x) = |x|^p/p$. *Left: Convex case* $p = 4$. *Right: Strongly convex case* $p = 2$.

6.4. Non-convex case

In this paragraph, we investigate the ability of the stochastic algorithm to avoid local traps and, in particular, we focus on the behavior of second order algorithms that may be an intermediary step towards global optimization methods such as simulated annealing. For this purpose, we defined f as:

$$\forall x \in \mathbb{R} \qquad f(x) = ax^4 + b(x-1)^2.$$

with $a = 1/40$ and $b = -1/5$. These values have been fixed to guarantee the numerical stability of the stochastic procedures, but the results we obtained may be replicated for other values. The values of a and b above yield a double-well potential with a global minimizer of f of around $x^\star \simeq -4.9$, although f has a local trap on the positive part at around $x_+ \simeq 4$. The function f is represented on the top left of Figure 4.

We used $\gamma_k = k^{-1}$ for all of the methods and we varied the initialization point of each algorithm from -10 to 10 with 100 Monte-Carlo replications. For each simulation, we arbitrarily stopped the evolution of the algorithm after $T = 10^4$ iterations, and considered that optimization was successful when $|x_T - x^\star| \leqslant 1$. This criterion may be replaced by a more stringent inequality, at the price of an increase of T, without really changing the main conclusions below.

Performances are reported in Figure 4. We observe that both SGD and Ruppert-Polyak algorithms have the same behavior. This fact is absolutely clear because Polyak averaging is built with a Cesaro average of SGD. The target convergence point of SGD and of Polyak averaging are thus the same. We can also note that in the almost no noise setting, the basin of attraction of $x^\star$ for SGD may be roughly approximated by $]-\infty, 1]$. Nevertheless, both SAGE and HBF seem to behave better behaviour with a somewhat larger basin of attraction: in particular, it is possible to start from an initialization point $x_1 = 8$ and still obtain convergence of SAGE or HBF towards $x^\star$. This last point is clearly impossible with SGD. The same conclusions hold for different values of σ (see Figure

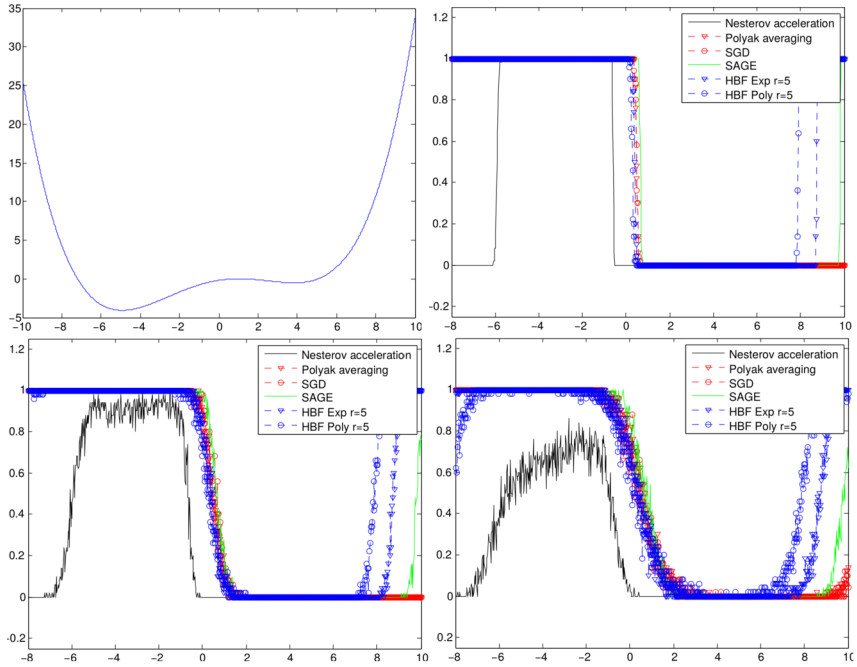

FIG 4. *Top left: function f to be minimized. Top right: probability of success of the stochastic algorithms with respect to the initialization point with small variance: $\sigma = 0.1$. Bottom left: $\sigma = 1$. Bottom right: $\sigma = 2$.*

4, bottom left and right). Finally, we observe that NAGD does not present very good behavior: the probability of failure when the algorithm is initialized at -4 is lower than 1 for $\sigma = 1$ or $\sigma = 2$.

We can calculate a more quantitative indicator of this behavior with the computation of the average rate of success of each algorithm when the initialization point is sampled uniformly over $[-10; 10]$. Table 1 seems to indicate that the stochastic heavy ball leads to a better exploration of the state space, in particular, with reasonable values of r (see Table 2). These conclusions should be understood as numerical observations of experimental results on this particular type of synthetic case, but we do not have any theoretical arguments to strengthen these final observations at this time.

σ	SGD	AV SGD	SAGE	NAGD	HBF Poly r=5	HBF Expo r=5
0.1	0.47	0.47	0.49	0.29	0.58	0.52
1	0.47	0.47	0.49	0.27	0.58	0.55
2	0.47	0.47	0.49	0.20	0.58	0.54

TABLE 1

Average rate of success of each stochastic algorithm with a uniformly sampled initialization over $[-10; 10]$ when σ varies.

Exp 1	Exp 2	Exp 5	Exp 10	Poly 1	Poly 2	Poly 5	Poly 10
0.51	0.53	0.55	0.58	0.26	0.43	0.58	0.50

TABLE 2

Average rate of success of heavy ball stochastic algorithm for several values of r, when $\sigma = 1$ and the initialization point is sampled uniformly over $[-10; 10]$.

Appendix A: Almost sure convergence towards a local minimizer

A.1. Preliminary estimations for convergence towards a critical point

We first establish a preliminary lemma that translates a mean-reverting effect on $\mathbb{E}[V_n(X_n, Y_n)]$ from n to $n+1$.

Lemma A.1. *Assume that* $(\mathbf{H}_{\boldsymbol{\sigma,2}})$ *and* $(\mathbf{H}_s)$ *hold and suppose that* $c_r < 1$. *Then, for any* $(a, b) \in \mathbb{R}_+^2$ *such that:*

$$\frac{a}{b} > \left(\frac{1}{2} \vee \frac{\|D^2 f\|_\infty}{1 - c_r} \vee r_\infty (c_f - 1) \right), \tag{A.1}$$

we have:

(*i*) *A constant* $C_1 > 0$ *and an integer* $n_0 \in \mathbb{N}$ *exist such that for any* $n \geqslant n_0$,

$$\forall x, y \in \mathbb{R}^d, \quad V_n(x, y) \geqslant C_1 \left(f(x) + \frac{\|y\|^2}{r_{n-1}} \right). \tag{A.2}$$

(*ii*) *Some positive constants* C_2, C_3 *and* $c_{a,b}$ *exist such that:*

$$\begin{aligned} &\mathbb{E}[V_{n+1}(X_{n+1}, Y_{n+1}) | \mathcal{F}_n] \\ &\quad \leqslant V_n(X_n, Y_n)(1 + C_2 \gamma_{n+1}^2 r_n) - c_{a,b} \gamma_{n+1} \|Y_n\|^2 - b \gamma_{n+1} r_n \|\nabla f(X_n)\|^2 \\ &\quad + C_3 \gamma_{n+1}^2 r_n. \end{aligned} \tag{A.3}$$

Proof:

Point (*i*): For any non-negative u, v, the elementary inequality $uv \leqslant \frac{\rho}{2} u^2 + \frac{1}{2\rho} v^2$ holds for any $\rho > 0$. We apply this inequality with $u = \|\nabla f(x)\|$, $v = \|y\|$ and $\rho = 2r_n$ and obtain:

$$|\langle \nabla f(x), y \rangle| \leqslant r_{n-1} \|\nabla f(x)\|^2 + \frac{a}{4 r_{n-1}} \|y\|^2.$$

It follows from Assumption $(\mathbf{H}_s)$ that $\|\nabla f\|^2 \leqslant c_f f$. Using the above inequality, we obtain that for any $x, y \in \mathbb{R}^d$:

$$V_n(x, y) \geqslant (a + b r_{n-1}(1 - c_f)) f(x) + \frac{1}{2 r_{n-1}} \left[a - \frac{b}{2} \right] \|y\|^2.$$

Choosing now a and b such that $a > b/2$ and $a > b r_\infty (c_f - 1)$, we obtain the first assertion follows from (A.1). ◇

Point (*ii*): The Taylor formula ensures the existence of $\xi_{n+1,1}$ and $\xi_{n+1,2}$ in $\overline{[X_n, X_{n+1}]}$ such that:

$$
\begin{aligned}
&V_{n+1}(X_{n+1}, Y_{n+1}) \\
&= (a + br_n)\left(f(X_n) - \gamma_{n+1}\langle \nabla f(X_n), Y_n\rangle + \frac{\gamma_{n+1}^2}{2} Y_n^t D^2 f(\xi_{n+1,1}) Y_n\right) \\
&+ \frac{a}{2r_n}\left(\|Y_n\|^2 + 2\gamma_{n+1} r_n \left(\langle Y_n, \nabla f(X_n)\rangle - \|Y_n\|^2\right.\right. \\
&+ \left.\left.\langle Y_n + \gamma_{n+1} r_n (\nabla f(X_n)\rangle - Y_n), \Delta M_{n+1}\rangle\right)\right) \\
&+ \frac{a}{2r_n}\gamma_{n+1}^2 r_n^2 \|\Delta M_{n+1}\|^2 \\
&- b\left\langle \nabla f(X_n) - \gamma_{n+1} D^2 f(\xi_{n+1,2}) Y_n, Y_n + \gamma_{n+1} r_n \left(\nabla f(X_n) - Y_n + \Delta M_{n+1}\right)\right\rangle.
\end{aligned}
$$

Combining the similar terms leads to:

$$
\begin{aligned}
V_{n+1}(X_{n+1}, Y_{n+1}) = V_n(X_n, Y_n) &- b(r_n - r_{n-1}) f(X_n) \\
&+ \gamma_{n+1}\langle \nabla f(X_n), Y_n\rangle \left(\underbrace{-a - br_n + a + br_n}_{=0}\right) \\
&- \gamma_{n+1} Y_n^t D_{n+1} Y_n - \gamma_{n+1} r_n b \|\nabla f(X_n)\|^2 \\
&+ \gamma_{n+1} r_n \Delta N_{n+1} + \gamma_{n+1} \Delta R_{n+1},
\end{aligned}
$$

where $(\Delta N_n)_{n\geqslant 1}$ is a sequence of martingale increments, D_n is a $d \times d$-matrix defined by:

$$
D_{n+1} = a\left(1 - \frac{1}{2\gamma_{n+1}}\left(\frac{1}{r_n} - \frac{1}{r_{n-1}}\right)\right) I_d - bD^2 f(\xi_{n+1,2}),
$$

and ΔR_{n+1} is a remainder term. Using $(\mathbf{H}_s)$, we know that $D^2 f$ is bounded, and we have the following bound for ΔR_{n+1}:

$$
\|\Delta R_{n+1}\| \leqslant C_2 \gamma_{n+1} r_n \left(\|Y_n\|^2 + \|\Delta M_{n+1}\|^2 + \|\nabla f(X_n)\|.\|Y_n\|\right),
$$

where C_2 is a deterministic positive constant independent of n. The fact that $(r_n)_{n\geqslant 1}$ is a bounded sequence combined with Assumptions $(\mathbf{H}_{\boldsymbol{\sigma},\mathbf{2}})$ and $(\mathbf{H}_s)$ yields $\mathbb{E}[\|\Delta R_{n+1}\||\mathcal{F}_n] \leqslant C_2 \gamma_{n+1} r_n \left(1 + \|Y_n\|^2 + f(X_n)\right)$. It follows that:

$$
\forall n \geqslant n_0 \qquad \mathbb{E}[\|\Delta R_{n+1}\||\mathcal{F}_n] \leqslant C_2 \gamma_{n+1} r_n V_n(X_n, Y_n).
$$

Second, the condition given by (A.1) shows that an integer $n_1 \geqslant n_0$ and a constant $c_{a,b} > 0$ exist such that:

$$
D_{n+1} Y_n^{\otimes 2} \geqslant c_{a,b} \|Y_n\|^2.
$$

Using the previous bounds in $V_{n+1}(X_{n+1}, Y_{n+1})$ and the fact that $(r_n)_{n\in\mathbb{N}}$ is non-increasing shows that:

$$\exists n_2 \geqslant n_1 \quad \forall n \geqslant n_2 : \quad \mathbb{E}[V_{n+1}(X_{n+1}, Y_{n+1})|\mathcal{F}_n] \leqslant V_n(X_n, Y_n)(1 + C\gamma_{n+1}^2 r_n) - c_{a,b}\gamma_{n+1}\|Y_n\|^2 - b\gamma_{n+1} r_n \|\nabla f(X_n)\|^2. \qquad \diamond \, \square$$

Proof of Proposition 3.1 We use Lemma A.1 to prove the results.
Proof of $(i) - (ii) - (iii)$: Under the conditions on (r_n), we can check that some positive a and b exist such that the conclusions of the previous lemma hold true. We then deduce that:

$$\mathbb{E}[V_{n+1}(X_{n+1}, Y_{n+1})|\mathcal{F}_n] \leqslant V_n(X_n, Y_n)(1 + C\alpha_{n+1}) - U_{n+1},$$

with $\alpha_n = \gamma_n^2 r_n$ and $U_{n+1} = c_{a,b}\gamma_{n+1}\|Y_n\|^2 + b\gamma_{n+1} r_n \|\nabla f(X_n)\|^2$. Subsequently, using the Robbins-Siegmund Theorem (see, *e.g.*, Theorem B.1 in Section B, borrowed from [Duf97]), we deduce, on the one hand, that $\sup_{n\geqslant 1} \mathbb{E}[V_n(X_n, Y_n)] < +\infty$ and that $(V_n(X_n, Y_n))_{n\geqslant 1}$ almost surely (and in L^1) converge towards a random variable $V_\infty \in \mathbb{R}_+$. In particular, the coercivity of f implies the $a.s.$-boundedness of $(X_n)_{n\geqslant 0}$. On the other hand, the Robbins-Siegmund Theorem also implies that:

$$\sum_{n\geqslant 1} \gamma_{n+1} r_n \left(\frac{\|Y_n\|^2}{r_n} + \|\nabla f(X_n)\|^2 \right) < +\infty \quad a.s.$$

Hence, the three first statements follow. $\diamond$
Proof of (iv): The proof relies on the so-called *ODE method* (see, *e.g.*, [Ben06]). Set $r_\infty = \lim_{n\to+\infty} r_n$. We deal with cases $r_\infty > 0$ and $r_\infty = 0$ separately.

Case $\mathbf{r_\infty > 0}$ (exponential memory): Set $\Gamma_n = \sum_{k=0}^n \gamma_k$ with the convention $\gamma_0 = 0$. Denote by $(\bar{z}(t))_{t\geqslant 0}$ the interpolated process defined by $\bar{z}(\Gamma_n) = Z_n = (X_n, Y_n)'$, $n \geqslant 0$, with linear interpolations between times Γ_n and Γ_{n+1} and let $\bar{z}^{(n)}$ be the associated *shifted-sequence* defined by:

$$\bar{z}^{(n)}(t) = \bar{z}(t + \Gamma_n) \quad t \geqslant 0.$$

Setting $\varepsilon_n = (0, (r_{n-1} - r_\infty)(\nabla f(X_n) - Y_n) + \Delta M_n)'$ and $h(x, y) = (-y, r_\infty(\nabla f(x) - y))'$, we have:

$$Z_{n+1} = Z_n + \gamma_{n+1}(h(Z_n) + \varepsilon_{n+1}).$$

Set $N(n, t) = \inf\{k \geqslant n, \gamma_{n+1} + \ldots + \gamma_k \geqslant t\}$ (with the convention $\inf \varnothing = n$). Then, since $(Z_n)_{n\geqslant 0}$ is $a.s.$-bounded, it is a classical result on stochastic algorithm theory (see, *e.g.*, [Duf97], Theorem 9.2.8 and the remark below) that if for any $T > 0$,

$$\limsup_{n\to+\infty} \sup_{t\in[0,T]} \left\| \sum_{k=n+1}^{N(n,t)+1} \gamma_k \varepsilon_k \right\| = 0 \quad a.s., \tag{A.4}$$

then $(\bar{z}^{(n)})_{n\geqslant 0}$ is relatively compact (for the topology of uniform convergence on compact sets) and its limit points are solutions to the ODE $\dot{z} = h(z)$. Let us prove (A.4). Let $T > 0$. Using the Cauchy-Schwarz inequality, we have, for every $t \in [0, T]$:

$$\begin{aligned}
&\sum_{k=n+1}^{N(n,t)+1} \gamma_k(\|\nabla f(X_{k-1})\| + \|Y_k\|) \\
&\leqslant \sqrt{2}\left(\sum_{k=n+1}^{N(n,t)+1} \gamma_k\right)^{\frac{1}{2}} \left(\sum_{k=n+1}^{N(n,t)+1} \gamma_k \left(\|\nabla f(X_{k-1})\|^2 + \|Y_{k-1}\|^2\right)\right)^{\frac{1}{2}} \qquad \text{(A.5)} \\
&\leqslant \sqrt{2(T+\gamma_1)}\left(\sum_{k=n+1}^{+\infty} \gamma_k \left(\|\nabla f(X_{k-1})\|^2 + \|Y_{k-1}\|^2\right)\right)^{\frac{1}{2}} \xrightarrow{n\to+\infty} 0,
\end{aligned}$$

where the last convergence follows from (iii). On the basis of Assumption $(\mathbf{H_{\sigma,2}})$ and (iii), we also note that $(\langle \sum_{k=1}^{n} \gamma_k \Delta M_k \rangle)_{n\geqslant 1}$ is $a.s.$-convergent so that $\sum \gamma_n \Delta M_n$. It easily follows that:

$$\limsup_{n\to+\infty} \sup_{t\in[0,T]} \left\| \sum_{k=n+1}^{N(n,t)+1} \gamma_k \Delta M_k \right\| = 0 \quad a.s.$$

and that (A.4) is satisfied. Now, we again deduce from (A.5) that for any $T > 0$,

$$\sup_{t\leqslant T} \|\bar{z}^{(n)}(t) - \bar{z}^{(n)}(0)\| = \sup_{t\leqslant T} \|\bar{z}^{(n)}(t) - Z_n\| \xrightarrow{n\to+\infty} 0$$

so that each limit point is stationary. At this stage, we have thus proven that every limit point of $(\bar{z}^n)_{n\geqslant 0}$ is a stationary solution to $\dot{z} = h(z)$. This implies that any limit point Z_∞ of $(Z_n)_{n\geqslant 0}$ satisfies $h(Z_\infty) = 0$ (and thus $Y_\infty = \nabla f(X_\infty) = 0$). Actually, let $(Z_{n_k})_{k\geqslant 1}$ be a convergent subsequence of the ($a.s.$ bounded) sequence $(Z_n)_{n\geqslant 0}$ and denote its limit by Z_∞. Up to a second extraction, $(\bar{z}^{(n_k)})$ converges to a stationary solution $\bar{z}^\infty$ of $\dot{z} = h(z)$. As a consequence, $h(\bar{z}^\infty(t)) = 0$ for any $t \geqslant 0$. In particular, $h(\bar{z}^\infty(0)) = h(Z_\infty) = 0$. By (ii) and the fact that $(Y_n)_{n\geqslant 0}$ converges to 0, we also deduce that $(f(X_n))_{n\geqslant 0}$ is $a.s.$-convergent. To conclude the proof, it remains to observe that the set of possible limits of subsequences of $(X_n)_{n\geqslant 1}$ is connected. This is true since $X_n - X_{n-1} = -\gamma_n Y_{n-1} \to 0$ as $n \to +\infty$. ◇

Case $\mathbf{r_\infty = 0}$ (polynomial memory): In this case, the proof is somewhat similar but the identification of the asymptotic dynamics requires an appropriate normalization of Y_n[1]. Let us set:

$$\widetilde{\gamma}_n = \gamma_n \sqrt{r_n},\ \widetilde{\Gamma}_n = \sum_{k=0}^{n} \widetilde{\gamma}_k,\ \widetilde{X}_n = X_n,\ \widetilde{Y}_n = \frac{Y_n}{\sqrt{r_n}}.$$

[1] In fact, due to the asymptotic stationarity, the limiting dynamics is not intrinsic.

Also set by $\widetilde{Z}_n = (\widetilde{X}_n, \widetilde{Y}_n)'$. The dynamic of $\widetilde{Z}_n$ is described by Lemma A.2 below. We denote as $(\tilde{z}(t))_{t\geqslant 0}$ the interpolated process, *i.e.* defined by $\tilde{z}(\widetilde{\Gamma}_n) = \widetilde{Z}_n$, $n \geqslant 0$, with linear interpolations between times $\widetilde{\Gamma}_n$ and $\widetilde{\Gamma}_{n+1}$ and let $\tilde{z}^{(n)}$ be the associated *shifted-sequence* defined by

$$\tilde{z}^{(n)}(t) = \tilde{z}(t + \widetilde{\Gamma}_n) \quad t \geqslant 0.$$

With this setting, the idea is to show that the sequence $(\tilde{z}^{(n)}(t))_{t\geqslant 0}$ is tight with limits being stationary solutions of a homogeneous O.D.E. $\dot{z} = \tilde{h}(z)$ ($\tilde{h}$ being the drift to be determined). The sequence $(\widetilde{Z}_n)_{n\geqslant 0}$ satisfies Lemma A.2 that shows that $\widetilde{Z}_{n+1} = \widetilde{Z}_n + \widetilde{\gamma}_{n+1}\left(\tilde{h}(\widetilde{Z}_n) + \tilde{\varepsilon}_{n+1}\right)$ with $\tilde{h}(\tilde{x}, \tilde{y}) := (-\tilde{y}, \nabla f(\tilde{x}))'$ and:

$$\tilde{\varepsilon}_{n+1} = \begin{pmatrix} 0 \\ \upsilon_n^{(1)} \nabla f(\widetilde{X}_n) + \upsilon_n^{(2)} \widetilde{Y}_n + \sqrt{\frac{r_n}{r_{n+1}}} \Delta M_{n+1} \end{pmatrix},$$

where $\upsilon_n^{(1)}$ and $\upsilon_n^{(2)}$ are given in the statement of Lemma A.2.

On the basis of Assumption $(\mathbf{H_r})$, we know that:

$$\limsup_{n\to+\infty} \frac{1}{2\gamma_{n+1}} \left(\frac{1}{r_{n+1}} - \frac{1}{r_n}\right) < 1,$$

so that:

$$\upsilon_n^{(1)} = O\left(\frac{r_n - r_{n+1}}{r_{n+1}}\right) = O(\widetilde{\gamma}_{n+1}\sqrt{r_n}) \qquad \text{and} \qquad \upsilon_n^{(2)} = O\left(\sqrt{r_n}\right).$$

Thus, $(\upsilon_n^{(1)})_{n\geqslant 1}$ and $(\upsilon_n^{(2)})_{n\geqslant 1}$ converge to 0 as $n \to +\infty$. We can now repeat the arguments used in the situation $r_\infty > 0$ and we obtain:

$$\limsup_{n\to+\infty} \sup_{t\in[0,T]} \left\| \sum_{k=n+1}^{\widetilde{N}(n,t)+1} \tilde{\gamma}_k \tilde{\varepsilon}_k \right\| = 0 \quad a.s.,$$

where $\widetilde{N}(n,t) = \inf\{k \geqslant n, \tilde{\gamma}_{n+1} + \ldots + \tilde{\gamma}_k \geqslant t\}$. We can still combine (A.5) and (iii) to obtain $\sup_{t\leqslant T} |\tilde{z}^{(n)}(t) - \tilde{z}^{(n)}(0)| \xrightarrow{n\to+\infty} 0$ for any $T > 0$. We conclude that $(\tilde{z}^{(n)})_{n\geqslant 0}$ is relatively compact and that its limits are stationary solutions of $\dot{z} = \tilde{h}(z)$. The end of the proof is exactly the same as in the case $r_\infty > 0$. ◇□

Lemma A.2. *If the sequence* $(\widetilde{Z}_n)_{n\geqslant 1}$ *is defined by* $\widetilde{Z}_n = (\widetilde{X}_n, \widetilde{Y}_n) := (X_n, \frac{Y_n}{\sqrt{r_n}})$, *then*

$$\widetilde{Z}_{n+1} = \widetilde{Z}_n + \widetilde{\gamma}_{n+1}\left(\tilde{h}(\widetilde{Z}_n) + \tilde{\varepsilon}_{n+1}\right)$$

where $\tilde{h}(\tilde{x}, \tilde{y}) = (-\tilde{y}, \nabla f(\tilde{x}) - \sqrt{r_\infty}\tilde{y})'$ *and*

$$\tilde{\varepsilon}_{n+1} = \begin{pmatrix} 0 \\ \upsilon_n^{(1)} \nabla f(\widetilde{X}_n) + \upsilon_n^{(2)} \widetilde{Y}_n + \sqrt{\frac{r_n}{r_{n+1}}} \Delta M_{n+1} \end{pmatrix},$$

with $\tilde{\gamma}_n := \gamma_n\sqrt{r_n}$ *and*

$$\upsilon_n^{(1)} = \sqrt{\frac{r_n}{r_{n+1}}} - 1 \quad \text{and} \quad \upsilon_n^{(2)} = \frac{1}{\tilde{\gamma}_{n+1}}\upsilon_n^{(1)} + \left(\sqrt{r_\infty} - \frac{r_n}{\sqrt{r_{n+1}}}\right).$$

Proof: First, the fact that $\widetilde{X}_{n+1} = \widetilde{X}_n - \tilde{\gamma}_{n+1}\widetilde{Y}_n$ is obvious. Second,

$$\widetilde{Y}_{n+1} = \widetilde{Y}_n\sqrt{\frac{r_n}{r_{n+1}}} + \tilde{\gamma}_{n+1}\left(\sqrt{\frac{r_n}{r_{n+1}}}\nabla f(\widetilde{X}_n) - \frac{r_n}{\sqrt{r_{n+1}}}\widetilde{Y}_n + \sqrt{\frac{r_n}{r_{n+1}}}\Delta M_{n+1}\right).$$

The lemma follows. □

A.2. Convergence towards a local minimizer

This paragraph gathers the proof of the technical results used in Section 3.2.

Proof of Proposition 3.3

Proof of (i). When $n \geqslant T$, we have $\Omega_{n+1} = \tilde{\gamma}_{n+1}$ by definition and the conclusion follows. In the other situation when $n \leqslant T$, we use the Lipschitz continuity of η: if $m = \sup_{z\in\mathcal{N}} \|D\eta(z)\|$, then Equation (3.3) yields:

$$\|\eta(\widetilde{Z}_{n+1}) - \eta(\widetilde{Z}_n)\|^2 \leqslant 4m^2\tilde{\gamma}_{n+1}^2\left[\|\widetilde{Y}_n\|^2 + r^2\|\nabla f(\widetilde{X}_n)\|^2 + q_{n+1}^2\|\Delta M_{n+1}\|^2 + \|U_{n+1}\|^2\right].$$

The neighborhood $\mathcal{N}$ being compact, we deduce from the previous inequality that a constant $C > 0$ exists such that:

$$\mathbf{E}\left[\|\Omega_{n+1}\|^2\mathbf{1}_{n<T}\,|\mathcal{F}_n\right] \leqslant \mathbf{E}\left[\|\eta(\widetilde{Z}_{n+1}) - \eta(\widetilde{Z}_n)\|^2\mathbf{1}_{n<T}|\mathcal{F}_n\right] \leqslant C\tilde{\gamma}_{n+1}^2,$$

where we used a uniform upper bound on $\mathbb{E}[\|\Delta M_{n+1}\|^2\mathbf{1}_{n<T}|\mathcal{F}_n]$, leading to the proof of (i). ◇

Proof of (ii). Note that $\mathbf{1}_{n<T}$ and $\mathbf{1}_{n\geqslant T}$ are $\mathcal{F}_n$ measurable and we have:

$$\mathbf{1}_{n\geqslant T}\mathbb{E}\left[\Omega_{n+1}|\mathcal{F}_n\right] = \mathbf{1}_{n\geqslant T}\tilde{\gamma}_{n+1} \geqslant 0.$$

On the complementary set, we also have:

$$\mathbf{1}_{n<T}\mathbb{E}\left[\Omega_{n+1}|\mathcal{F}_n\right] \geqslant \mathbf{1}_{n<T}\mathbb{E}\left[[\eta(\widetilde{Z}_{n+1}) - \eta(\widetilde{Z}_n)]|\mathcal{F}_n\right] = \mathbf{1}_{n<T}\mathbb{E}\left[\eta(\widetilde{Z}_{n+1}) - \eta(\widetilde{Z}_n)|\mathcal{F}_n\right]$$

Hence, we can use the lower bound given by (3.6): for any value of $\alpha \in (0, 1]$:

$$\begin{aligned}
&\mathbf{1}_{n<T}\mathbb{E}\left[\Omega_{n+1}|\mathcal{F}_n\right] \\
&\geqslant \mathbf{1}_{n<T}\left[\tilde{\gamma}_{n+1}\langle D\eta(\widetilde{Z}_n), F(\widetilde{Z}_n)\rangle + \tilde{\gamma}_{n+1}\langle D\eta(\widetilde{Z}_n), \mathbb{E}[\Delta M_{n+1}|\mathcal{F}_n] + U_{n+1}\rangle\right] \\
&- \mathbf{1}_{n<T}k_\alpha\tilde{\gamma}_{n+1}^{1+\alpha}\left[\|\widetilde{Y}_n\| + r\|\nabla f(\widetilde{X}_n)\| + q_{n+1}\|\Delta M_{n+1}\| + \|U_{n+1}\|\right]^{1+\alpha}
\end{aligned}$$

where we used the triangle inequality in the last line to derive an upper bound of $\|\widetilde{Z}_{n+1} - \widetilde{Z}_n\|$. When $n < T$, $\widetilde{Z}_n$ is bounded and we have $\mathbb{E}[\|\Delta M_{n+1}\|^2|\mathcal{F}_n] \leqslant \sigma^2 M$ for a large enough M. Hence, the Hölder inequality implies that:

$$\mathbb{E}[\|\Delta M_{n+1}\|^{1+\alpha}|\mathcal{F}_n] \leqslant \sigma^{1+\alpha}M^{\frac{1+\alpha}{2}}.$$

Therefore, we can find a large enough constant $C_1 > 0$ such that:

$$\mathbf{1}_{n<T}\mathbb{E}\left[\Omega_{n+1}|\mathcal{F}_n\right] \geqslant \mathbf{1}_{n<T}\left[\tilde{\gamma}_{n+1}\langle D\eta(\widetilde{Z}_n), F(\widetilde{Z}_n)\rangle - m\tilde{\gamma}_{n+1}\|U_{n+1}\| - C_1\tilde{\gamma}_{n+1}^{1+\alpha}\right].$$

The lower bound (iii) of Proposition 3.2 and the definition of U_{n+1} implies that a constant C_2 exists such that:

$$\mathbf{1}_{n<T}\mathbb{E}\left[\Omega_{n+1}|\mathcal{F}_n\right] \geqslant \mathbf{1}_{n<T}\tilde{\gamma}_{n+1}\left[\kappa\eta(\widetilde{Z}_n) - C_1\tilde{\gamma}_{n+1}^{\alpha} - \frac{C_2}{\sqrt{\Gamma_n}}\right]$$

We now choose α so that $\tilde{\gamma}_{n+1}^{\alpha} \simeq \Gamma_n^{-1/2}$, which corresponds to the choice:

$$\alpha = \frac{1-\beta}{1+\beta}.$$

Defining $\epsilon_n = \kappa^{-1}\left[C_1\tilde{\gamma}_{n+1} + C_2\Gamma_n^{-1/2}\right]$, we then deduce that if $n < T$, then $S_n = \eta(\widetilde{Z}_n)$ so that:

$$\mathbf{1}_{S_n\geqslant\epsilon_n}\mathbb{E}\left[\Omega_{n+1}|\mathcal{F}_n\right] \geqslant 0,$$

which concludes the proof. In particular, ϵ_n must be chosen on the order $\tilde{\gamma}_{n+1}^{\alpha}$ (or on the order $\Gamma_n^{-1/2} \sim n^{-(1-\beta)/2}$). ◇

Proof of (iii). Observe that $S_{n+1}^2 - S_n^2 = \Omega_{n+1}^2 + 2S_n\Omega_{n+1}$. Now, if $S_n \geqslant \epsilon_n$, then we have seen in the proof of (ii) that:

$$\begin{aligned}\mathbf{1}_{S_n\geqslant\epsilon_n}\mathbb{E}[S_{n+1}^2 - S_n^2|\mathcal{F}_n] &= \mathbf{1}_{S_n\geqslant\epsilon_n}\mathbb{E}[\Omega_{n+1}^2|\mathcal{F}_n] + 2S_n\mathbf{1}_{S_n\geqslant\epsilon_n}\mathbb{E}[\Omega_{n+1}|\mathcal{F}_n]\\ &\geqslant \mathbf{1}_{S_n\geqslant\epsilon_n}\mathbb{E}[\Omega_{n+1}^2|\mathcal{F}_n].\end{aligned}$$

In the other situation, we have $S_n \leqslant \epsilon_n$, meaning that $n < T$ and we have seen in the proof of (ii) that:

$$\begin{aligned}\mathbf{1}_{n<T}\mathbb{E}[\Omega_{n+1}|\mathcal{F}_n] \geqslant \mathbf{1}_{n<T}&\left[\tilde{\gamma}_{n+1}\kappa\eta(\widetilde{Z}_n) + \tilde{\gamma}_{n+1}\langle D\eta(\widetilde{Z}_n), U_{n+1}\rangle\right]\\ &- k_2\tilde{\gamma}_{n+1}^2\left[\|\widetilde{Y}_n\| + r\|\nabla f(\widetilde{X}_n)\| + q_{n+1}\|\Delta M_{n+1}\| + \|U_{n+1}\|\right]^2\end{aligned}$$

Consequently, because of the positivity of η, we deduce that:

$$\mathbf{1}_{n<T}\mathbb{E}[\Omega_{n+1}|\mathcal{F}_n] \geqslant -\|D\eta(\widetilde{Z}_n)\| \times O(\tilde{\gamma}_{n+1}\Gamma_n^{-1/2}) - O(\tilde{\gamma}_{n+1}^2).$$

We know that $D\eta$ is locally bounded on $\mathcal{N}$, we then obtain:

$$\begin{aligned}\mathbf{1}_{S_n\leqslant\epsilon_n}\mathbb{E}[\Omega_{n+1}|\mathcal{F}_n] &= \mathbf{1}_{S_n\leqslant\epsilon_n}\mathbf{1}_{n<T}\mathbb{E}[\Omega_{n+1}|\mathcal{F}_n]\\ &= \mathbf{1}_{\eta(\widetilde{Z}_n)\leqslant\epsilon_n}\mathbf{1}_{n<T}\mathbb{E}[\Omega_{n+1}|\mathcal{F}_n]\\ &\geqslant -\mathbf{1}_{\eta(\widetilde{Z}_n)\leqslant\epsilon_n}\mathbf{1}_{n<T}\left[\|D\eta(\widetilde{Z}_n)\|O(\tilde{\gamma}_{n+1}\Gamma_n^{-1/2}) + O(\tilde{\gamma}_{n+1}^2)\right].\\ &\geqslant -C\tilde{\gamma}_{n+1}\left[\Gamma_n^{-1/2} + \tilde{\gamma}_{n+1}\right],\end{aligned}$$

for a large enough constant C. In the two situations, we then have:

$$\mathbb{E}[S_{n+1}^2 - S_n^2|\mathcal{F}_n] \geqslant \mathbb{E}[\Omega_{n+1}^2|\mathcal{F}_n] - 2C\epsilon_n\tilde{\gamma}_{n+1}^2 - 2C\epsilon_n\tilde{\gamma}_{n+1}\Gamma_n^{-1/2}.$$

Finally, Lemma 9.7 of [Ben06] and our hypoelliptic assumption $(\mathbf{H}_{\mathcal{E}})$ implies that for small enough c:

$$\mathbb{E}[\Omega_{n+1}^2|\mathcal{F}_n] \geqslant c\tilde{\gamma}_{n+1}^2$$

The conclusion follows if $\epsilon_n\tilde{\gamma}_{n+1}\Gamma_n^{-1/2} = o\left(\tilde{\gamma}_{n+1}^2\right)$. Since ϵ_n is chosen on the order $\Gamma_n^{-1/2} \sim \tilde{\gamma}_{n+1}^{\alpha}$ with $\alpha = (1-\beta)/(1+\beta)$, this condition is equivalent to:

$$\tilde{\gamma}_{n+1}^{1+2\alpha} = o\left(\tilde{\gamma}_{n+1}^2\right).$$

meaning that $\alpha > 1/2$. It then implies that β should be less than $1/3$. ◇□

A.3. Supremum of the square of sub-Gaussian random variables

We consider a sequence of independent random variables $(\xi_i)_{i\geqslant n}$ of $\mathbb{R}^d$ such that each coordinate satisfies a sub-Gaussian assumption $(\mathbf{H}_{\boldsymbol{Gauss,\sigma}})$:

$$\forall\lambda\in\mathbb{R} \qquad \forall j\in\{1,\dots,d\} \qquad \forall i\geqslant n \qquad \log\mathbb{E}\left[e^{\lambda\xi_i^j}\right] \leqslant \lambda^2\frac{\sigma^2}{2}, \tag{A.6}$$

where σ^2 is a variance factor. If $(\gamma_k)_{k\geqslant n}$ is a decreasing sequence in $\ell^2(\mathbb{N})$, we are looking for an upper bound of:

$$m_n^{\star} = \mathbb{E}\left[\sup_{k\geqslant n}\left\{\gamma_k^2\|\xi_k\|^2\right\}\right]. \tag{A.7}$$

For any $\nu > 0$ and any decreasing sequence $\gamma_n \sim \gamma n^{-\nu}$, we establish the following result (useful for Theorem 3.2).

Theorem A.1. *If each coordinate ξ_i^j is absolutely continuous w.r.t. the Lebesgue measure and satisfies $(\mathbf{H}_{\boldsymbol{Gauss,\sigma}})$, then:*

$$m_n^{\star} \lesssim \sigma^2 d\,\gamma_n^2\log(\gamma_n^{-2}),$$

where $\lesssim$ refers to an inequality up to a universal constant.

We begin with a preliminary lemma.

Lemma A.3. *Assume that X is a real random variable that satisfies $(\mathbf{H}_{\boldsymbol{Gauss,\sigma}})$ with median 0:*

$$\mathbb{P}(X>0) = \mathbb{P}(X<0) = \frac{1}{2}.$$

Then, we can find $Y\sim\mathcal{N}(0,\sigma^2)$ on the same probability space and c large enough s.t.

$$|X| \leqslant c|Y| \qquad a.s.$$

Proof of Lemma A.3: We use a coupling argument. We denote F_X as the cumulative distribution function:

$$F_X(t) = \int_{-\infty}^{t} f_X(u)du = \mathbb{P}[X \leqslant t].$$

Similarly, we also denote Ψ_{σ^2} as the cumulative distribution function of a Gaussian random variable $\mathcal{N}(0, \sigma^2)$:

$$\Psi_{\sigma^2}(t) = \int_{-\infty}^{t} \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi}\sigma} dx = \mathbb{P}[\mathcal{N}(0, \sigma^2) \leqslant t].$$

Our assumption on the distribution on X shows that the generalized inverse of F_X (denoted F_X^{-1}) exists and if $\mathcal{U}$ is a uniform random variable between on $[0, 1]$, then $X \sim F_X^{-1}(\mathcal{U})$. We now consider the random variable $Y \sim F_{\sigma^2}^{-1}(\mathcal{U})$ built with the same realization of $\mathcal{U}$. Of course, Y is distributed according to a Gaussian random variable $\mathcal{N}(0, \sigma^2)$.

We need to show that a sufficiently large $c > 0$ exists such that $|X| \leqslant c|Y|$, that is:

$$\left|F_X^{-1}(u)\right| \leqslant c\left|\Psi_{\sigma^2}^{-1}(u)\right|. \tag{A.8}$$

Using the fact that F_X is an increasing function, and letting $u = \Psi_{\sigma^2}(y)$, it is then equivalent to show that:

$$\forall y \in \mathbb{R} \qquad F_X(-c|y|) \leqslant \Psi_{\sigma^2}(|y|) \leqslant F_X(c|y|) \tag{A.9}$$

We now study two different situations for y. If $y = 0$, then Inequality (A.9) holds since the median of X is 0. If $|y| \leqslant \eta$ is close to 0, the same inequality is satisfied with a first-order Taylor expansion. For example, the right hand side reads:

$$F_X(c|y|) \sim \frac{1}{2} + \int_0^{c|y|} f_X(u)du \geqslant \frac{1}{2} + cf_X(0)|y| + o(|y|),$$

which is greater than $\Psi_{\sigma^2}(|y|)$ for c large enough. Hence, we deduce that Inequality (A.9) holds around 0.

Now, we assume that $|y| > \eta > 0$, the desired upper bound (A.9) is equivalent to:

$$1 - F_X(c|y|) \leqslant 1 - \Psi_{\sigma^2}(|y|).$$

The Chernoff bound associated with the sub-Gaussian assumption ($\mathbf{H}_{\boldsymbol{Gauss,\sigma}}$) on the distribution of X implies that:

$$\mathbb{P}(X > c|y|) \leqslant e^{\inf_{\lambda>0}\left\{\lambda^2\sigma^2/2 - \lambda c|y|\right\}} = e^{-\frac{c^2|y|^2}{2\sigma^2}}.$$

At the same time, the lower bound of the Gaussian tail is given by:

$$1 - \Psi_{\sigma^2}(c|y|) \geqslant \frac{e^{-|y|^2/2\sigma^2}}{\sqrt{2\pi}\sigma}\left[|y|^{-1} - |y|^{-3}\right] \geqslant \kappa(\delta)e^{-|y|^2/2\sigma^2},$$

with $\kappa(\delta)$ a constant independent of $|y| \geqslant \delta$. Hence, the right hand side of (A.9) holds for a large enough c (independent on σ^2). A symmetry argument permits to conclude for the left hand side of (A.9).

Inequality (A.9) being equivalent to (A.8), the conclusion of the proof follows. □

We are now looking at to the proof of Theorem A.1.

Proof of Theorem A.1:

We will shift all of the coordinates of the random variables $(\xi_i)_{i \geqslant n}$ by their corresponding medians. Assuming $(\mathbf{H}_{\boldsymbol{Gauss,\sigma}})$, the coordinates $(\xi_i^j)_{1 \leqslant j \leqslant d}$ are centered and have a second-order moment upper bounded by σ^2 (see [Str94], for example):

$$\forall i \geqslant n \qquad \forall j \in \{1, \dots, d\} \qquad \mathbb{E}[\{\xi_j^i\}^2] \leqslant \sigma^2.$$

The Tchebychev inequality implies that each median m_i^j of the random variables ξ_i^j are bounded by:

$$\forall i \geqslant n \qquad \forall j \in \{1, \dots, d\} \qquad |m_i^j| \leqslant \sqrt{2}\sigma. \tag{A.10}$$

We then consider the centered (w.r.t. their medians) random variables:

$$\tilde{\xi}_i^j = \xi_i^j - m_i^j,$$

and use the inequality $(a+b)^2 \leqslant 2a^2 + 2b^2$ together with the upper bound (A.10) to deduce that:

$$\begin{aligned} m_n^\star &= \mathbb{E} \sup_{k \geqslant n} \gamma_k^2 \|\xi_k\|^2 = \mathbb{E} \sup_{k \geqslant n} \gamma_k^2 \sum_{j=1}^{d} \{\xi_k^j\}^2 \\ &\leqslant \mathbb{E} \sup_{k \geqslant n} \gamma_k^2 \left[2 \sum_{j=1}^{d} \{\xi_k^j - m_j^k\}^2 + 2d\sigma^2 \right] \\ &\leqslant 2d\sigma^2 \gamma_n^2 + 2\mathbb{E} \sup_{k \geqslant n} \gamma_k^2 \|\tilde{\epsilon}_k\|^2. \end{aligned}$$

We can use Lemma A.3 and deduce that up to a multiplicative universal constant:

$$m_n^\star \lesssim 2d\sigma^2 \gamma_n^2 + 2\sigma^2 \mathbb{E} \sup_{k \geqslant n} \gamma_k^2 \|Z_k\|^2,$$

where each $(Z_k)_{k \geqslant n}$ are i.i.d. realizations of Gaussian random variables $\mathcal{N}(0, \sigma^2 I_d)$.

We now aim to apply a chaining argument to control the supremum of the empirical process above. To apply Lemma A.4, we define $\mathcal{T}_n := [\![n; +\infty[\![$ and compute the Laplace transform of the chi-square-like random variables:

$$\log \mathbb{E} e^{\lambda[\gamma_k^2 \|Z_k\|^2 - \gamma_j^2 \|Z_j\|^2]} = \frac{d}{2} \log \left(\frac{1 - 2\lambda\gamma_j^2}{1 - 2\lambda\gamma_k^2} \right)$$

We can check that up to a universal multiplicative constant, we have:

$$\forall \lambda \in \mathbb{R}_+ \quad \forall (a,b) \in \mathbb{R}_+ \times \mathbb{R}_+ : \qquad \log \frac{1-a\lambda}{1-b\lambda} \lesssim \lambda |a-b| + \frac{|a-b|^2 \lambda^2}{1-\lambda |a-b|}.$$

We are naturally driven to define the pseudo-metric on $\mathcal{T}_n$ by:

$$\forall (i,j) \in \mathcal{T}_n^2 \qquad d(i,j) = \left|\gamma_i^2 - \gamma_j^2\right|.$$

It remains to upper bound the covering number of $\mathcal{T}_n$ according to d for any radius $\epsilon > 0$. Indeed, when $2\gamma_n^2 \leqslant \epsilon$, we have $N(\epsilon, \mathcal{T}_n) = 1$ although when $\epsilon \leqslant 2\gamma_n^2$, we use the rough bound:

$$N(\epsilon, \mathcal{T}_n) \leqslant \inf\left\{ j \geqslant n \, : 2\gamma_j^2 \leqslant \epsilon \right\}.$$

In particular, if $\gamma_j = \gamma j^{-\nu}$, we then obtain

$$N(\epsilon, \mathcal{T}_n) \sim \epsilon^{-1/2\nu}.$$

We apply Lemma A.4 and obtain an upper bound for the right hand side of (A.11). The first term is proportionnal to γ_n^2. The other terms lead to the computation of the two integrals (up to some universal multiplicative constants):

$$\int_0^{\gamma_n^2} \sqrt{\log(\epsilon^{-1})} d\epsilon \qquad \text{and} \qquad \int_0^{\gamma_n^2} \log(\epsilon^{-1}) d\epsilon$$

The change of variable $\epsilon = e^{-x}$ and an integration by parts leads to an upper bound whose size is $\log(\gamma_n^{-2})\gamma_n^2$. □

The next Lemma, borrowed from [BLM13] (see Lemma 13.1, Chapter 13), provides a key estimate for the expectation of the suppremum of an empirical process indexed by a pseudo metric space $(\mathcal{T}, d)$. This estimate involves the covering numbers $N(\delta, \mathcal{T})$ associated with the set $\mathcal{T}$ and the pseudo-metric d.

Lemma A.4. *Let $\mathcal{T}$ be a separable metric space and $(X_t)_{t\in\mathcal{T}}$ be a collection of random variables such that for some constants $a, v, c > 0$,*

$$\log \mathbb{E} e^{\lambda[X_i - X_j]} \leqslant a\lambda d(i,j) + \frac{v\lambda^2 d^2(i,j)}{2(1-c\lambda d(i,j)}$$

for all $(i,j) \in \mathcal{T}^2$ and all $0 < \lambda < \{cd(i,j)\}^{-1}$. Then, for any $i_0 \in \mathcal{T}$:

$$\mathbb{E} \sup_{i\in\mathcal{T}} [X_t - X_{i_0}] \leqslant 3a\delta + 12\sqrt{v} \int_0^{\delta/2} \sqrt{H(u,\mathcal{T})} du + 12c \int_0^{\delta/2} H(u,\mathcal{T}) du. \quad \text{(A.11)}$$

Appendix B: Standard tools of stochastic algorithms

We recall below a standard version of the so-called Robbins-Siegmund Theorem (see *e.g.* [Duf97]):

Theorem B.1. *Given a filtration $\mathcal{F}_n$ and four positive, integrable and $\mathcal{F}_n$-adapted sequences $(\alpha_n)_n$,$(\beta_n)_n$, $(U_n)_n$ and $(V_n)_n$ satisfying:*

- *(i) $(\alpha_n)_n$,$(\beta_n)_n$, $(U_n)_n$ are predictible sequences.*
- *(ii) $\sup_\omega \prod_n (1+\alpha_n(\omega)) < \infty$, $\sum_n \mathbb{E}(\beta_n) < \infty$.*
- *(iii) $\forall n \in \mathbb{N}$,*

$$\mathbb{E}(V_{n+1}|\mathcal{F}_n) \leqslant V_n(1+\alpha_{n+1}) + \beta_{n+1} - U_{n+1}$$

Then:

(i) V_n converges to V_∞ in L^1 and $\sup_n \mathbb{E}[V_n] < \infty$.
(ii) $\sum_n \mathbb{E}(U_n) < \infty$, $\sum_n U_n < \infty$ a.s.

B.1. Step sizes $\gamma_n = \gamma\, n^{-\beta}$ with $\beta < 1$

Proposition B.1. *For any positive values $a > 0$ and $b > 0$, for any $\beta \in (0,1)$ and any sequence $(\gamma_n)_{n\geqslant 1}$ defined by $\gamma_n = \gamma n^{-\beta}$, one has:*

$(i) - a$ *If $\beta < 1/2$, then* $\sum_{k=1}^n a\gamma_k - b\gamma_k^2 \geqslant \frac{a\gamma}{1-\beta} n^{1-\beta} - \frac{b\gamma^2}{1-2\beta} n^{1-2\beta}$

$(i) - b$ *If $\beta > 1/2$, then* $\sum_{k=1}^n a\gamma_k - b\gamma_k^2 \geqslant \frac{a\gamma}{1-\beta} n^{1-\beta} - \frac{b\gamma^2}{2\beta-1}$

$(i) - c$ *If $\beta = 1/2$, then* $\sum_{k=1}^n a\gamma_k - b\gamma_k^2 \geqslant \frac{a\gamma}{1-\beta} n^{1-\beta} - b\gamma^2 \log n$

(ii) An integer n_0 exists such that

$$\forall n \geqslant n_0 \quad \sum_{k=1}^n \gamma_k^2 \prod_{l=k+1}^n (1-a\gamma_l)^2 \leqslant \frac{2}{a}\gamma_{n+1}.$$

(iii) An integer n_0 exists such that

$$\forall n \geqslant n_0 \quad \sum_{k=1}^n \gamma_k^2 \prod_{l=k+1}^n (1-a\gamma_l + b\gamma_l^2) \leqslant \frac{2}{a}\gamma_{n+1}.$$

Proof: The upper bounds involved in $(i) - a, b, c$ are straightforward. ◇

Proof of *(ii)*: Using Γ_n introduced in the beginning of Section 2, we write:

$$\begin{aligned}\sum_{k=1}^n \gamma_k^2 \prod_{l=k+1}^n (1-a\gamma_l)^2 &\leqslant \sum_{k=1}^n \gamma_k^2 e^{-a\sum_{k+1}^n \gamma_l} \\ &= \sum_{k=1}^n \gamma_k^2 e^{-a\Gamma_n + a\Gamma_k} \leqslant \gamma^2 e^{-a\Gamma_n} \sum_{k=1}^n k^{-2\beta} e^{\frac{a\gamma}{1-\beta}k^{1-\beta}}\end{aligned}$$

The function $x \longmapsto x^{-2\beta} e^{\frac{a\gamma}{1-\beta}x^{1-\beta}}$ being increasing for $x \geqslant c_{a,\gamma,\beta}$, we then obtain, considering an integer $t > c_{a,\gamma,\beta}$:

$$\sum_{k=1}^n \gamma_k^2 \prod_{l=k+1}^n (1-a\gamma_l)^2 \leqslant \gamma^2 e^{-a\Gamma_n} \left(C_t + \int_t^n x^{-2\beta} e^{\frac{a\gamma}{1-\beta}x^{1-\beta}} dx \right).$$

We can write $x^{-2\beta}e^{Kx^{1-\beta}} = \left(e^{Kx^{1-\beta}}\right)' x^{-\beta}K^{-1}(1-\beta)^{-1}$ and integrating by parts, we obtain for a large enough n:

$$\sum_{k=1}^{n} \gamma_k^2 \prod_{l=k+1}^{n} (1-a\gamma_l)^2 \leqslant \gamma^2 e^{-a\Gamma_n}\left(C_t + \frac{e^{a\Gamma_n}}{a\gamma} n^{-\beta}\right) \leqslant \frac{2}{a}\gamma_n. \qquad \diamond$$

Proof of *(iii)*: We only deal with $\beta < 1/2$, which is the most involved situation. Using Γ_n and $\Gamma_n^{(2)}$ introduced in the beginning of Section 2, we write:

$$\begin{aligned}
\sum_{k=1}^{n} \gamma_k^2 \prod_{l=k+1}^{n} (1-a\gamma_l+b\gamma_l^2) &\leqslant \sum_{k=1}^{n} \gamma_k^2 e^{-a\Gamma_n + a\Gamma_k + b\Gamma_n^{(2)} - b\Gamma_k^{(2)}} \\
&\leqslant e^{-a\Gamma_n + b\Gamma_n^{(2)}} \sum_{k=1}^{n} \gamma_k^2 e^{a\Gamma_k - b\Gamma_k^{(2)}} \\
&\leqslant \gamma^2 e^{-a\Gamma_n + b\Gamma_n^{(2)}} \sum_{k=1}^{n} k^{-2\beta} e^{\frac{a\gamma}{1-\beta}k^{1-\beta} - \frac{b\gamma^2}{1-2\beta}k^{1-2\beta}}.
\end{aligned}$$

The function $x \longmapsto x^{-2\beta} e^{\frac{a\gamma}{1-\beta}x^{1-\beta} - \frac{b\gamma^2}{1-2\beta}x^{1-2\beta}}$ being increasing for $x \geqslant c_{a,b,\gamma,\beta}$, we then obtain considering an integer $t > c_{a,b,\gamma,\beta}$:

$$\begin{aligned}
&\sum_{k=1}^{n} \gamma_k^2 \prod_{l=k+1}^{n} (1-a\gamma_l+b\gamma_l^2) \\
&\leqslant \gamma^2 e^{-a\Gamma_n + b\Gamma_n^{(2)}} \left(\sum_{k=1}^{t} k^{-2\beta} e^{\frac{a\gamma}{1-\beta}k^{1-\beta} - \frac{b\gamma^2}{1-2\beta}k^{1-2\beta}} + \int_t^n x^{-2\beta} e^{\frac{a\gamma}{1-\beta}x^{1-\beta} - \frac{b\gamma^2}{1-2\beta}x^{1-2\beta}} dx\right) \\
&\leqslant \gamma^2 e^{-a\Gamma_n + b\Gamma_n^{(2)}} \left(C_t + \int_t^n x^{-2\beta} e^{\frac{a\gamma}{1-\beta}x^{1-\beta} - \frac{b\gamma^2}{1-2\beta}x^{1-2\beta}} dx\right) \\
&\leqslant \gamma^2 e^{-a\Gamma_n + b\Gamma_n^{(2)}} (C_t \\
&+ \int_t^n x^{-\beta} \left[\frac{3}{2}\frac{a\gamma x^{-\beta} - b\gamma^2 x^{-2\beta}}{a\gamma} + \frac{3b\gamma x^{-2\beta} - a\gamma x^{-\beta}}{2}\right] e^{\frac{a\gamma}{1-\beta}x^{1-\beta} - \frac{b\gamma^2}{1-2\beta}x^{1-2\beta}} dx\Big)
\end{aligned}$$

Now choosing $t \geqslant (3b/a)^{\beta^{-1}}$ yields $3b\gamma x^{-2\beta} \leqslant a\gamma x^{-\beta}$ for any $x \geqslant t$. Integrating by parts, we obtain:

$$\begin{aligned}
\sum_{k=1}^{n} \gamma_k^2 \prod_{l=k+1}^{n} (1-a\gamma_l+b\gamma_l^2) &\leqslant \gamma^2 e^{-a\Gamma_n + b\Gamma_n^{(2)}} \left(C_t + \frac{n^{-\beta}}{a\gamma} e^{-a\Gamma_n + n\Gamma_n^{(2)}}\right) \\
&\leqslant \frac{\gamma n^{-\beta}}{a} + \gamma^2 C_t e^{-a\Gamma_n + b\Gamma_n^{(2)}}.
\end{aligned}$$

Then, choosing n_0 large enough (that depends on a, b, γ and β), we deduce that:

$$\forall n \geqslant n_0 \qquad \sum_{k=1}^{n} \gamma_k^2 \prod_{l=k+1}^{n} (1-a\gamma_l+b\gamma_l^2) \leqslant \frac{2}{a}\gamma_n. \qquad \diamond$$

□

B.2. Step sizes $\gamma_n = \gamma\, n^{-1}$

Proposition B.2. *For any positive values* $a > 0$ *and* $b > 0$ *and any sequence* $(\gamma_n)_{n\geqslant 1}$ *defined by* $\gamma_n = \gamma n^{-1}$*, we have:*

(i) $\displaystyle\sum_{k=1}^{n} a\gamma_k - b\gamma_k^2 \geqslant a\log n - b\pi^2/6$

(ii) $\displaystyle\sum_{k=1}^{n} \gamma_k^2 \prod_{l=k+1}^{n} (1-a\gamma_l)^2 \leqslant C_\gamma \begin{cases} \frac{1}{a\gamma-1} n^{-1} & \text{if} \quad a\gamma > 1 \\ \log n\, n^{-1} & \text{if} \quad a\gamma = 1 \\ \frac{1}{1-a\gamma} n^{-a\gamma} & \text{if} \quad a\gamma < 1 \end{cases}$

(iii) $\displaystyle\sum_{k=1}^{n} \gamma_k^2 \prod_{l=k+1}^{n} (1-a\gamma_l+b\gamma_l^2) \leqslant C_{\gamma,b} \begin{cases} \frac{1}{a\gamma-1} n^{-1} & \text{if} \quad a\gamma > 1 \\ \log n\, n^{-1} & \text{if} \quad a\gamma = 1 \\ \frac{1}{1-a\gamma} n^{-a\gamma} & \text{if} \quad a\gamma < 1 \end{cases}$

(iv) *For any* $\epsilon > 0$*,* $a > 0$ *and* $b > 0$*:* $\displaystyle\sum_{k=1}^{n} \gamma_{k+1} \prod_{l=k+1}^{n} (1-a\gamma_l+b\gamma_l^{1+\epsilon}) \leqslant \frac{2e^{b\Gamma_\infty^{(1+\epsilon)}}}{a}$.

Proof: The upper bounds involved in (i) and (ii) are straightforward. ◇

Proof of (iii): The situation is easier than the one involved in point (ii) of Proposition B.1 because in that case, we have:

$$\forall n \geqslant 1 \qquad \Gamma_n^{(2)} \leqslant \gamma^2\pi^2/6.$$

Therefore, we can repeat the computations above and get:

$$\begin{aligned} \sum_{k=1}^{n} \gamma_k^2 \prod_{l=k+1}^{n} (1-a\gamma_l+b\gamma_l^2) &\leqslant \sum_{k=1}^{n} \gamma_k^2 e^{-a\Gamma_n + a\Gamma_k + b\Gamma_n^{(2)} - b\Gamma_k^{(2)}} \\ &\leqslant e^{-a\Gamma_n + b\gamma^2\pi^2/6} \sum_{k=1}^{n} \gamma_k^2 e^{a\Gamma_k} \\ &\leqslant \gamma^2 e^{b\gamma^2\pi^2/6} n^{-a\gamma} \sum_{k=1}^{n} k^{-2+a\gamma}. \end{aligned}$$

We then deduce that:

$$\sum_{k=1}^{n} \gamma_k^2 \prod_{l=k+1}^{n} (1-a\gamma_l+b\gamma_l^2) = \gamma^2 e^{b\gamma^2\pi^2/6} \begin{cases} \frac{1}{a\gamma-1} n^{-1} & \text{if} \quad a\gamma > 1 \\ \log n\, n^{-1} & \text{if} \quad a\gamma = 1 \\ \frac{1}{1-a\gamma} n^{-a\gamma} & \text{if} \quad a\gamma < 1 \end{cases} \qquad \diamond$$

Proof of (iii): We follow the same guideline: remark that $(\Gamma_n^{(1+\epsilon)})_{n\geqslant 1}$ is a bounded sequence and write

$$\begin{aligned} \sum_{k=1}^{n} \gamma_{k+1} \prod_{l=k+1}^{n} (1-a\gamma_l+b\gamma_l^{1+\epsilon}) &\leqslant \sum_{k=1}^{n} \frac{\gamma}{k+1} e^{-a\gamma\log n + a\gamma\log k + b\Gamma_n^{(1+\epsilon)}} \\ &\leqslant \gamma e^{b\Gamma_\infty^{(1+\epsilon)}} n^{-a\gamma} \int_1^n x^{a\gamma-1} dx \\ &\leqslant \frac{e^{b\Gamma_\infty^{(1+\epsilon)}}}{a}. \end{aligned}$$

□

References

[Bac14] F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15:595–627, 2014.

[BD96] O. Brandière and M. Duflo. Les algorithmes stochastiques contournent-ils les pièges ? *Annales de l'I.H.P. Probabilités et Statistiques*, 32:395–427, 1996.

[Ben06] M. Benaïm. *Dynamics of stochastic approximation algorithms*. Lecture Notes in Mathematics, Séminaire de Probabilités XXXIII. Springer-Verlag, 2006. Characterization and convergence.

[BH96] M. Benaïm and M.W. Hirsh. Asymptotic pseudotrajectories and chain recurrent flows, with applications. *J. Dynam. Differential Equations*, 8:141–176, 1996.

[Bil95] P. Billingsley. *Convergence of Probability Measures*. Wiley series in Probability & Statistics, New York, 1995. MR1700749

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by M. Ledoux.

[BLR02] M. Benaïm, M. Ledoux, and O. Raimond. Self-interacting diffusions. *Probab. Theory Related Fields*, 122:1–41, 2002.

[BM05] S. Burer and R. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Math. Program.*, 103(3, Ser. A):427–444, 2005.

[BM11] F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[Bor97] Vivek S. Borkar. Stochastic approximation with two time scales. *Systems Control Lett.*, 29(5):291–294, 1997.

[Bor08] Vivek S. Borkar. *Stochastic approximation*. Cambridge University Press, Cambridge; Hindustan Book Agency, New Delhi, 2008. A dynamical systems viewpoint. MR2442439

[Bot10] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Physica-Verlag/Springer, Heidelberg, 2010.

[BT09] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

[CEG09a] A. Cabot, H. Engler, and S. Gadat. On the long time behavior of second order differential equations with asymptotically small dissipation. *Trans. Amer. Math. Soc.*, 361(11):5983–6017, 2009. MR2529922

[CEG09b] A. Cabot, H. Engler, and S. Gadat. Second-order differential equations with asymptotically small dissipation and piecewise flat potentials. In *Proceedings of the Seventh Mississippi State–UAB Con-*

ference on Differential Equations and Computational Simulations, volume 17 of *Electron. J. Differ. Equ. Conf.*, pages 33–38. Southwest Texas State Univ., San Marcos, TX, 2009.

[Duf97] M. Duflo. Random iterative models, adaptive algorithms and stochastic approximations,. *Applications of Mathematics (New York). Springer-Verlag, Berlin.*, 22, 1997. MR1485774

[EK86] S. Ethier and T. Kurtz. *Markov Processes*. John Willey and Sons, New York, 1986.

[FW56] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Res. Logist. Quart.*, 3:95–110, 1956. MR0089102

[FW84] M. Freidlin and A. Wentzell. *Random perturbations of dynamical systems*, volume 260 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Heidelberg, third edition, 1984. Translated from the 1979 Russian original by Joseph Szücs. MR0722136

[GL13] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23:2341–2368, 2013.

[GL16] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.*, 156:59–99, 2016.

[GMP15] S. Gadat, L. Miclo, and F. Panloup. A stochastic model for speculative bubbles. *Alea: Latin American journal of probability and mathematical statistics*, 12:491–532, 2015.

[GP14] S. Gadat and F. Panloup. Long time behaviour and stationary regime of memory gradient diffusions. *Annales Institut Henri Poincaré (B)*, 50:564–601, 2014. MR3189085

[GP18] S. Gadat and F. Panloup. Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity *Preprint*, 2018.

[GY07] S. Gadat and L. Younes. A stochastic algorithm for feature selection in pattern recognition. *Journal of Machine Learning Research*, 8:509–547, 2007.

[Har82] P. Hartman. *Ordinary Differential Equations*. Classic in Applied Mathematics. Wiley, 1982.

[Har91] A. Haraux. *Systèmes dynamiques dissipatifs et applications*. R.M.A. Masson, Paris, 1991.

[HPK09] C. Hu, W. Pan, and J. T. Kwok. Accelerated gradient methods for stochastic optimization and online learning. *In Advances in Neural Information Processing Systems*, 2009.

[JKK+17] P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating Stochastic Gradient Descent. *ArXiv e-prints*, April 2017.

[JKN16] C. Jin, S. M. Kakade, and P. Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. *NIPS*, 2016.

[JNJ17] C. Jin, P. Netrapalli, and M. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. *Preprint*, 2017.

[KW52] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.*, 23:462–466, 1952.

[KY03] H. J. Kushner and G. Yin. Stochastic approximation and recursive algorithms and applications. *Springer-Verlag, second edition,*, 2003.

[Lan12] G. Lan. An optimal method for stochastic composite optimization. *Math. Program*, 133(1-2, Ser.A):365–397, 2012.

[Lem07] V. Lemaire. An adaptive scheme for the approximation of dissipative systems. *Stochastic Processes and their Applications*, 117(10):1491–1518, 2007.

[LSJR16] J. Lee, M. Simchowitz, M. Jordan, and B. Recht. Gradient descent converges to minimizers. *Preprint*, 2016.

[MS17] P. Mertikopoulos and M. Staudigl. On the convergence of gradient-like flows with noisy gradient input. *SIAM Journal on Optimization, to appear*, 2017.

[MSH02] J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Process. Appl.*, 101(2):185–232, 2002. MR1931266

[MT93] S. Meyn and R. Tweedie. Stability of Markovian processes. III. Foster-Lyapunov criteria for continuous-time processes. *Adv. in Appl. Probab.*, 25(3):518–548, 1993.

[Nes83] Y. Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[Nes04] Y. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. A basic course.

[Nit15] A. Nitanda. Accelerated Stochastic Gradient Descent for Minimizing Finite Sums. *ArXiv e-prints*, June 2015.

[NY83] A. Nemirovski and D. Yudin. Problem complexity and method efficiency in optimization. *Wiley-Interscience Series in Discrete Mathematics.*, John Wiley, XV, 1983.

[Pem90] R. Pemantle. Non-convergence to unstable points in urn models and stochastic approximations. *Annals of Probability*, 18:698–712, 1990. MR1055428

[PJ92] B. T. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30:838–855, 1992.

[Poi86] H. Poincaré. Mémoire sur les courbes définies par une équation différentielle (iv). *Journal de Mathématiques Pures et Appliquées*, 4:151–217, 1886.

[Pol64] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4:1–17, 1964.

[RM51] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22:400–407, 1951.

[Rup88] D. Ruppert. Efficient estimations from a slowly convergent robbins-

monro process. *Technical Report, 781, Cornell University Operations Research and Industrial Engineering*, 1988.

[Str94] K.R. Stromberg. *Probability for Analysts*. Chapman & Hall, CRC, New York, 1994.

[SV06] D. W. Stroock and S. R. S. Varadhan. *Multidimensional diffusion processes*. Classics in Mathematics. Springer-Verlag, Berlin, 2006. Reprint of the 1997 edition. MR2190038

[Vil09] C. Villani. Hypocoercivity. *Mem. Amer. Math. Soc.*, 202(950), 2009.

[WSC16] S. Boyd W. Su and E. J. Candes. A differential equation for modeling nesterov's accelerated gradient method: theory and insights. *Journal of Machine Learning Research, to appear*, 2016.

[YLL16] T. Yang, Q. Lin, and Z. Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *Preprint*, 2016.