Stochastic optimization with momentum: Convergence, fluctuations, and traps avoidance[*]

Anas Barakat[1], Pascal Bianchi[1], Walid Hachem[2], and Sholom Schechtman[2]

[1]*LTCI, Télécom Paris, IP Paris, France*
e-mail: anas.barakat@telecom-paris.fr; pascal.bianchi@telecom-paris.fr

[2]*LIGM, CNRS, Univ Gustave Eiffel, ESIEE Paris, F-77454 Marne-la-Vallée, France*
e-mail: walid.hachem@univ-eiffel.fr; sholom.schechtman@univ-eiffel.fr

Abstract: In this paper, a general stochastic optimization procedure is studied, unifying several variants of the stochastic gradient descent such as, among others, the stochastic heavy ball method, the Stochastic Nesterov Accelerated Gradient algorithm (S-NAG), and the widely used ADAM algorithm. The algorithm is seen as a noisy Euler discretization of a non-autonomous ordinary differential equation, recently introduced by Belotto da Silva and Gazeau, which is analyzed in depth. Assuming that the objective function is non-convex and differentiable, the stability and the almost sure convergence of the iterates to the set of critical points are established. A noteworthy special case is the convergence proof of S-NAG in a non-convex setting. Under some assumptions, the convergence rate is provided under the form of a Central Limit Theorem. Finally, the non-convergence of the algorithm to undesired critical points, such as local maxima or saddle points, is established. Here, the main ingredient is a new avoidance of traps result for non-autonomous settings, which is of independent interest.

MSC2020 subject classifications: Primary 62L20; secondary 34A12, 68T99, 60F99.
Keywords and phrases: Stochastic approximation, dynamical systems, adaptive gradient methods with momentum, Nesterov accelerated gradient, ADAM, avoidance of traps.

Received December 2020.

Contents

---

1. Introduction

Given a probability space Ξ, an integer $d > 0$, and a function $f : \mathbb{R}^d \times \Xi \to \mathbb{R}$, consider the problem of finding a local minimum of the function $F(x) \triangleq \mathbb{E}_\xi[f(x,\xi)]$ w.r.t. $x \in \mathbb{R}^d$, where $\mathbb{E}_\xi$ represents the expectation w.r.t. the random variable ξ on Ξ. The paper focuses on the case where F is possibly non-convex. It is assumed that the function F is unknown to the observer, either because the distribution of ξ is unknown, or because the expectaction cannot be evaluated. Instead, a sequence $(\xi_n : n \geq 1)$ of i.i.d. copies of the random variable ξ is revealed online.

While the Stochastic Gradient Descent is the most classical algorithm that is used to solve such a problem, recently, several other algorithms became very popular. These include the Stochastic Heavy Ball (SHB), the stochastic version

A. Barakat et al.

of Nesterov's Accelerated Gradient method (S-NAG) and the large class of the so-called *adaptive* gradient algorithms, among which ADAM [29] is perhaps the most used in practice. As opposed to the vanilla Stochastic Gradient Descent, the study of such algorithms is more elaborate, for three reasons. First, the update of the iterates involves a so-called *momentum* term, or inertia, which has the effect of "smoothing" the increment between two consecutive iterates. Second, the update equation at the time index n is likely to depend on n, making these systems inherently *non-autonomous*. Third, as far as adaptive algorithms are concerned, the update also depends on some additional variable (*a.k.a.* the learning rate) computed online as a function of the history of the computed gradients.

In this work, we study in a unified way the asymptotic behavior of these algorithms in the situation where F is a differentiable function which is not necessarily convex, and where the stepsize of the algorithm is decreasing.

Our starting point is a generic non-autonomous Ordinary Differential Equation (ODE) introduced by Belotto da Silva and Gazeau [9] (see also [8] for ADAM), depicting the continuous-time versions of the aforementioned florilegium of algorithms. The solutions to the ODE are shown to converge to the set of critical points of F. This suggests that a general provably convergent algorithm can be obtained by means of an Euler discretization of the ODE, including possible stochastic perturbations. Special cases of our general algorithm include SHB, ADAM and S-NAG. We establish the almost sure boundedness and the convergence to critical points. Under additional assumptions, we obtain convergence rates, under the form of a central limit theorem. These results are new. They extend the works of [23, 8] to a general setting. In particular, we highlight the almost sure convergence result of S-NAG in a non-convex setting, which is new to the best of our knowledge.

Next, we address the question of the avoidance of "traps". In a non-convex setting, the set of critical points of a function F is generally larger than the set of local minimizers. A "trap" stands for a critical point at which the Hessian matrix of F has negative eigenvalues, namely, it is a local maximum or saddle point. We establish that the iterates cannot converge to such a point, if the noise is exciting in some directions. The result extends previous works of [23] obtained in the context of SHB. This result not only allows to study a broader class of algorithms but also significantly weakens the assumptions. In particular, [23] uses a sub-Gaussian assumption on the noise and a rather stringent assumption on the stepsizes. The main difficulty in the approach of [23] lies in the use of the classical autonomous version of Poincaré's invariant manifold theorem. The key ingredient of our proof is a general avoidance of traps result, adapted to non-autonomous settings, which we believe to be of independent interest. It extends usual avoidance of traps results to a non-autonomous setting, by making use of a non-autonomous version of Poincaré's theorem [16, 30].

Paper organization In Section 2, we introduce and study the ODE's governing our general stochastic algorithm. We establish the existence and uniqueness of the solutions, as well as the convergence to the set of critical points. In Sec-

tion 3, we introduce the main algorithm. We provide sufficient conditions under which the iterates are bounded and converge to the set of critical points. A central limit theorem is stated. Section 4 introduces a general avoidance of traps result for non-autonomous settings. Next, this result is applied to the proposed algorithm. Sections 5, 6 and 7 are devoted to the proofs of the results of Sections 2, 3 and 4, respectively.

Notations Given an integer $d \geq 1$, two vectors $x, y \in \mathbb{R}^d$, and a real α, we denote by $x \odot y$, $x^{\odot \alpha}$, x/y, $|x|$, and $\sqrt{|x|}$ the vectors in $\mathbb{R}^d$ whose i-th coordinates are respectively given by $x_i y_i$, x_i^α, x_i/y_i, $|x_i|$, $\sqrt{|x_i|}$. Inequalities of the form $x \leq y$ are to be read componentwise. The standard Euclidean norm is denoted $\|\cdot\|$. Notation M^{T} represents the transpose of a matrix M. For $x \in \mathbb{R}^d$ and $\rho > 0$, the notation $B(x, \rho)$ stands for the open ball of $\mathbb{R}^d$ with center x and radius ρ. We also write $\mathbb{R}_+ = [0, \infty)$. If $z \in \mathbb{R}^d$ and $A \subset \mathbb{R}^d$, we write $\mathrm{dist}(z, A) \triangleq \inf\{\|z - z'\| : z' \in A\}$. By $\mathbb{1}_A(x)$, we refer to the function that is equal to one if $x \in A$ and to zero elsewhere. The set of zeros of a function $h : \mathbb{R}^d \to \mathbb{R}^{d'}$ is $\mathrm{zer}\, h = \{x : h(x) = 0\}$. Let D be a domain in $\mathbb{R}^d$. Given an integer $k \geq 0$, the class $\mathcal{C}^k(D, \mathbb{R})$ is the class of $D \to \mathbb{R}$ maps such that all their partial derivatives up to the order k exist and are continuous. For a function $h \in \mathcal{C}^k(D, \mathbb{R})$ and for every $i \in \{1, \ldots, d\}$, we denote as $\partial_i^k h(x_1, \ldots, x_d)$ the kth partial derivative of the function h with respect to x_i. When $k = 1$, we just write $\partial_i h(x_1, \ldots, x_d)$. The gradient of a function $F : \mathbb{R}^d \to \mathbb{R}$ at a point $x \in \mathbb{R}^d$ is denoted as $\nabla F(x)$, and its Hessian matrix at x is $\nabla^2 F(x)$ as usual. For a function $S : \mathbb{R}^d \to \mathbb{R}^d$, the notation $\nabla S(x)$ stands for the jacobian matrix of S at point x.

2. Ordinary differential equations

2.1. A general ODE

Our starting point will be a non-autonomous ODE which is almost identical to the one introduced in [9] and close to the one in [8]. Let F be a function in $\mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$, let S be a continuous $\mathbb{R}^d \to \mathbb{R}_+^d$ function, let $\mathsf{h}, \mathsf{r}, \mathsf{p}, \mathsf{q} : (0, \infty) \to \mathbb{R}_+$ be four continuous functions, and let $\varepsilon > 0$. Let $v_0 \in \mathbb{R}_+^d$ and $x_0, m_0 \in \mathbb{R}^d$. Starting at $\mathsf{v}(0) = v_0$, $\mathsf{m}(0) = m_0$, and $\mathsf{x}(0) = x_0$, our ODE on $\mathbb{R}_+$ with trajectories in $\mathcal{Z}_+ \triangleq \mathbb{R}_+^d \times \mathbb{R}^d \times \mathbb{R}^d$ reads

$$\begin{cases} \dot{\mathsf{v}}(t) & = \mathsf{p}(t)S(\mathsf{x}(t)) - \mathsf{q}(t)\mathsf{v}(t) \\ \dot{\mathsf{m}}(t) & = \mathsf{h}(t)\nabla F(\mathsf{x}(t)) - \mathsf{r}(t)\mathsf{m}(t) \\ \dot{\mathsf{x}}(t) & = -\mathsf{m}(t)/\sqrt{\mathsf{v}(t) + \varepsilon} \end{cases} \qquad \text{(ODE-1)}$$

This ODE can be rewritten compactly as follows. Write $z_0 = (v_0, m_0, x_0)$, and let $\mathsf{z}(t) = (\mathsf{v}(t), \mathsf{m}(t), \mathsf{x}(t)) \in \mathcal{Z}_+$ for $t \in \mathbb{R}_+$. Let $\mathcal{Z} \triangleq \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$, and define

the map $g : \mathcal{Z}_+ \times (0, \infty) \to \mathcal{Z}$ as

$$
g(z, t) = \begin{bmatrix} \mathsf{p}(t)S(x) - \mathsf{q}(t)v \\ \mathsf{h}(t)\nabla F(x) - \mathsf{r}(t)m \\ -m/\sqrt{v + \varepsilon} \end{bmatrix} \tag{1}
$$

for $z = (v, m, x) \in \mathcal{Z}_+$. With these notations, we can rewrite (ODE-1) as

$$
\mathsf{z}(0) = z_0, \quad \dot{\mathsf{z}}(t) = g(\mathsf{z}(t), t) \text{ for } t > 0.
$$

By setting $S(x) = \nabla F(x)^{\odot 2}$ when necessary and by properly choosing the functions h, r, p, and q, a large number of iterative algorithms used in Machine Learning can be obtained by an Euler's discretization of this ODE. For instance, choosing $\mathsf{h}(t) = \mathsf{r}(t) = a(t, \lambda, \alpha_1)$ and $\mathsf{p}(t) = \mathsf{q}(t) = a(t, \lambda, \alpha_2)$ with $a(t, \lambda, \alpha) = \lambda^{-1}(1 - \exp(-\lambda\alpha))/(1 - \exp(-\alpha t))$ and $\lambda, \alpha_1, \alpha_2 > 0$, one obtains a version of the ADAM algorithm [29] (see [9, Sections 2.4-4.2] for details). To give another less specific example, if we set $\mathsf{p} = \mathsf{q} \equiv 0$, then the resulting ODE covers a family of algorithms to which the well-known HEAVY BALL with friction algorithm [5] belongs. For a comprehensive and more precise view of the deterministic algorithms that can be deduced from (ODE-1) by an Euler's discretization, the reader is referred to [9, Table 1].

In this paper, since we are rather interested in stochastic versions of these algorithms, Eq. (ODE-1) will be the basic building block of the classical "ODE method" which is widely used in the field of stochastic approximation [10]. In order to analyze the behavior of this equation in preparation of the stochastic analysis, we need the following assumptions.

Assumption 2.1. The function F belongs to $\mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ and ∇F is locally Lipschitz continuous.

Assumption 2.2. F is coercive, *i.e.*, $F(x) \to +\infty$ as $\|x\| \to +\infty$.

Note that this assumption implies that the infimum $F_\star$ of F is finite, and the set zer ∇F of zeros of ∇F is nonempty.

Assumption 2.3. The map $S : \mathbb{R}^d \to \mathbb{R}^d_+$ is locally Lipschitz continuous.

Assumption 2.4. The continuous functions $\mathsf{h}, \mathsf{r}, \mathsf{p}, \mathsf{q} : (0, +\infty) \to \mathbb{R}_+$ satisfy:

i) $\mathsf{h} \in \mathcal{C}^1((0, +\infty), \mathbb{R}_+)$, $\dot{\mathsf{h}}(t) \leq 0$ on $(0, +\infty)$ and the limit $h_\infty \triangleq \lim_{t \to \infty} \mathsf{h}(t)$ is positive.

ii) r and q are non-increasing and $r_\infty \triangleq \lim_{t \to \infty} \mathsf{r}(t)$, $q_\infty \triangleq \lim_{t \to \infty} \mathsf{q}(t)$ are positive.

iii) p converges towards p_∞ as $t \to \infty$.

iv) For all $t \in (0, +\infty)$, $\mathsf{r}(t) \geq \mathsf{q}(t)/4$ and $r_\infty > q_\infty/4$.

These assumptions are sufficient to prove the existence and the uniqueness of the solution to (ODE-1) starting at a time $t_0 > 0$. The following additional assumption extends the solution to $t_0 = 0$.

Assumption 2.5. Either $\mathsf{h}, \mathsf{r}, \mathsf{p}, \mathsf{q} \in \mathcal{C}^1([0, +\infty), \mathbb{R}_+)$, or the following holds:

i) For every $x \in \mathbb{R}^d$, we have $S(x) \geq \nabla F(x)^{\odot 2}$.

ii) The functions $\frac{\mathsf{h}}{\mathsf{p}}$, $\frac{\mathsf{h}}{\mathsf{q}-2\mathsf{r}}$, $t \mapsto t\mathsf{h}(t)$, $t \mapsto t\mathsf{r}(t)$, $t \mapsto t\mathsf{p}(t)$, $t \mapsto t\mathsf{q}(t)$ are bounded near zero.

iii) There exists $t_0 > 0$ such that for all $t < t_0$, $2\mathsf{r}(t) - \mathsf{q}(t) > 0$.

iv) There exists $\delta > 0$ such that $\frac{\mathsf{h}}{\mathsf{r}}$, $\frac{\mathsf{p}}{\mathsf{q}} \in \mathcal{C}^1([0,\delta), \mathbb{R}_+)$.

v) The initial condition $z_0 = (v_0, m_0, x_0) \in \mathcal{Z}_+$ satisfies

$$m_0 = \nabla F(x_0) \lim_{t \downarrow 0} \frac{\mathsf{h}(t)}{\mathsf{r}(t)} \quad \text{and} \quad v_0 = S(x_0) \lim_{t \downarrow 0} \frac{\mathsf{p}(t)}{\mathsf{q}(t)}.$$

Remark 1. The functions $\mathsf{h}, \mathsf{r}, \mathsf{p}, \mathsf{q}$ corresponding to ADAM satisfy these conditions. We leave the straightforward verifications to the reader. We just observe here that the function S that will correspond to our stochastic algorithm in Section 3 below will satisfy Assumption 2.5–*i)* by an immediate application of Jensen's inequality.

The following theorem slightly generalizes the results of [9, Th. 3 and Th. 5].

Theorem 2.1. Let Assumptions 2.1 to 2.4 hold true. Consider $z_0 \in \mathcal{Z}_+$ and $t_0 > 0$. Then, there exists a unique global solution $\mathsf{z} : [t_0, +\infty) \to \mathcal{Z}_+$ to (ODE-1) with initial condition $\mathsf{z}(t_0) = z_0$. Moreover, $\mathsf{z}([t_0, +\infty))$ is a bounded subset of $\mathcal{Z}_+$. As $t \to +\infty$, $\mathsf{z}(t)$ converges towards the set

$$\Upsilon \triangleq \{z_\star = (p_\infty S(x_\star)/q_\infty, 0, x_\star) \ : \ x_\star \in \operatorname{zer} \nabla F\}. \tag{2}$$

If, additionally, Assumption 2.5 holds, then we can take $t_0 = 0$.

Remark 2. Th. 2.1 only shows the convergence of the trajectory $\mathsf{z}(t)$ towards a set. Convergence of the trajectory towards a single point is not guaranteed when the set Υ is not countable.

Remark 3. A simpler version of (ODE-1) is obtained when omitting the momentum term. It reads:

$$\begin{cases} \dot{\mathsf{v}}(t) &= \mathsf{p}(t)S(\mathsf{x}(t)) - \mathsf{q}(t)\mathsf{v}(t) \\ \dot{\mathsf{x}}(t) &= -\nabla F(\mathsf{x}(t))/\sqrt{\mathsf{v}(t) + \varepsilon}. \end{cases} \tag{ODE-1$'$}$$

This ODE encompasses the algorithms of the family of RMSPROP [42], as shown in [8, 9]. The approach for proving the previous theorem can be adapted to (ODE-1$'$) with only minor modifications. In the proofs below, we will point out the particularities of (ODE-1$'$) when necessary.

The following paragraph is devoted to a particular case of (ODE-1), which does not satisfy Assumption 2.4, and which requires a more involved treatment than (ODE-1$'$).

2.2. The Nesterov case

The authors of [13], [41] and others studied the ODE

$$\ddot{\mathsf{x}}(t) + \frac{\alpha}{t}\dot{\mathsf{x}}(t) + \nabla F(\mathsf{x}(t)) = 0, \quad \alpha > 0, \quad F \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}),$$

which Euler's discretization generates the well-known Nesterov's accelerated gradient algorithm, see also [6, 7]. This ODE can be rewritten as

$$\begin{cases} \dot{\mathsf{m}}(t) & = \nabla F(\mathsf{x}(t)) - \frac{\alpha}{t}\mathsf{m}(t) \\ \dot{\mathsf{x}}(t) & = -\mathsf{m}(t), \end{cases} \tag{ODE-N}$$

which is formally the particular case of (ODE-1) that is taken for $\mathsf{p}(t) = \mathsf{q}(t) = 0$, $\mathsf{h}(t) = 1$, and $\mathsf{r}(t) = \alpha/t$. Obviously, this case is not covered by Assumption 2.4. Moreover, it turns out that, contrary to the situation described in Remark 3 above, this case cannot be dealt with by a straightforward adaptation of the proof of Th. 2.1. The reason for this is as follows. Heuristically, the proof of Th. 2.1 is built around the fact that the solution of (ODE-1) "shadows" for large t the solution of the autonomous ODE

$$\begin{cases} \dot{\mathsf{v}}(t) & = p_\infty S(\mathsf{x}(t)) - q_\infty \mathsf{v}(t) \\ \dot{\mathsf{m}}(t) & = h_\infty \nabla F(\mathsf{x}(t)) - r_\infty \mathsf{m}(t) \\ \dot{\mathsf{x}}(t) & = -\frac{\mathsf{m}(t)}{\sqrt{\mathsf{v}(t)+\varepsilon}}, \end{cases}$$

and the latter can be shown to converge to the set Υ defined in Eq. (2), either under Assumption 2.4 or for the algorithms covered by Remark 3. This idea does not work anymore for (ODE-N), for its large–t autonomous counterpart

$$\begin{cases} \dot{\mathsf{m}}(t) & = \nabla F(\mathsf{x}(t)) \\ \dot{\mathsf{x}}(t) & = -\mathsf{m}(t). \end{cases}$$

can have solutions that do not converge to the critical points of F. As an example of such solutions, take $d = 1$ and $F(x) = x^2/2$. Then, $t \mapsto (\cos(t), \sin(t))$ is an oscillating solution of the latter ODE.

Yet, we have the following result. Up to our knowledge, the proof of the convergence below as $t \to +\infty$ is new.

Theorem 2.2. Let Assumptions 2.1 and 2.2 hold true. Then, for each $x_0 \in \mathbb{R}^d$, there exists a unique bounded global solution $(\mathsf{m}, \mathsf{x}) : \mathbb{R}_+ \to \mathbb{R}^d \times \mathbb{R}^d$ to (ODE-N) with the initial condition $(\mathsf{m}(0), \mathsf{x}(0)) = (0, x_0)$. As $t \to +\infty$, $(\mathsf{m}(t), \mathsf{x}(t))$ converges towards the set

$$\bar{\Upsilon} \triangleq \{(0, x_\star) : x_\star \in \operatorname{zer} \nabla F\}. \tag{3}$$

2.3. Related works

The continuous-time dynamical system (ODE-1) we consider was first introduced in [9, Eq. (2.1)] with $S = \nabla F^{\odot 2}$. Th. 2.1 above is roughly the same as [9, Ths. 3 and 5], with some slight differences regarding the assumptions on the function F, or Assumption 2.4-*iv)*. We point out that the main focus of [9] is to study the properties of the *deterministic continous-time* dynamical system (ODE-1). In the present work, we highlight that the purpose of Th. 2.1 is

to pave the way to our analysis of the corresponding *stochastic algorithms* in Section 3.

Concerning Th. 2.2, the existence and the uniqueness of a global solution to (ODE-N) has been previously shown in the literature, for instance in [13, Prop. 2.1] or in [41, Th. 1]. The convergence statement in Th. 2.2 is new to the best of our knowledge. In particular, we stress that we do not make any convexity assumption on F. The closest result we are aware of is the one of Cabot-Engler-Gadat [13]. In [13, Prop. 2.5], it is shown that if $\mathsf{x}(t)$ converges towards some point $\bar{x}$, then necessarily $\bar{x}$ is a critical point of F. Our result in Th. 2.2 strengthens this statement, by establishing that $\mathsf{x}(t)$ actually converges to the set of critical points.

3. Stochastic algorithms

In this section, we discuss the asymptotic behavior of stochastic algorithms that consist in noisy Euler's discretizations of (ODE-1) and (ODE-N) studied in the previous section.

We first set the stage. Let $(\Xi, \mathscr{T}, \mu)$ be a probability space. Denoting as $\mathscr{B}(\mathbb{R}^d)$ the Borel σ-algebra on $\mathbb{R}^d$, consider a $\mathscr{B}(\mathbb{R}^d) \otimes \mathscr{T}$–measurable function $f : \mathbb{R}^d \times \Xi \to \mathbb{R}$ that satisfies the following assumption.

Assumption 3.1. The following conditions hold:

  i) For every $x \in \mathbb{R}^d$, $f(x, \cdot)$ is μ–integrable.
  ii) For every $s \in \Xi$, the map $f(\cdot, s)$ is differentiable. Denoting as $\nabla f(x, s)$ its gradient w.r.t. x, the function $\nabla f(x, \cdot)$ is integrable.
  iii) There exists a measurable map $\kappa : \mathbb{R}^d \times \Xi \to \mathbb{R}_+$ s.t. for every $x \in \mathbb{R}^d$:

   a) The map $\kappa(x, \cdot)$ is μ–integrable,
   b) There exists $\varepsilon > 0$ s.t. for every $s \in \Xi$,

   $$\forall\, u, v \in B(x, \varepsilon),\ \|\nabla f(u, s) - \nabla f(v, s)\| \le \kappa(x, s)\|u - v\|\,.$$

Under Assumption 3.1, we can define the mapping $F : \mathbb{R}^d \to \mathbb{R}$ as

$$F(x) = \mathbb{E}_\xi[f(x, \xi)] \qquad (4)$$

for all $x \in \mathbb{R}^d$, where we write $\mathbb{E}_\xi \varphi(\xi) = \int \varphi(\xi)\mu(d\xi)$. It is easy to see that the mapping F is differentiable,

$$\nabla F(x) = \mathbb{E}_\xi[\nabla f(x, \xi)]$$

for all $x \in \mathbb{R}^d$, and ∇F is locally Lipschitz.

Let $(\gamma_n)_{n \ge 1}$ be a sequence of positive real numbers satisfying

Assumption 3.2. $\gamma_{n+1}/\gamma_n \to 1$ and $\sum_n \gamma_n = +\infty$.

Define for every integer $n \ge 1$

$$\tau_n = \sum_{k=1}^n \gamma_k\,.$$

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, and let $(\xi_n : n \geq 1)$ be a sequence of iid random variables defined from $(\Omega, \mathscr{F}, \mathbb{P})$ into $(\Xi, \mathscr{T}, \mu)$ with the distribution μ.

3.1. General algorithm

Our first algorithm is a discrete and noisy version of (ODE-1).

Let $z_0 = (v_0, m_0, x_0) \in \mathcal{Z}_+$ and $h_0, r_0, p_0, q_0 \in (0, \infty)$. Define for every $n \geq 1$

$$h_n = \mathsf{h}(\tau_n), \ r_n = \mathsf{r}(\tau_n), \ p_n = \mathsf{p}(\tau_n), \text{ and } q_n = \mathsf{q}(\tau_n). \tag{5}$$

The algorithm is written as follows.

Algorithm 1 (general algorithm)

Initialization: $z_0 \in \mathcal{Z}_+$.
for $n = 1$ **to** n_{iter} **do**
  $v_{n+1} = (1 - \gamma_{n+1} q_n) v_n + \gamma_{n+1} p_n \nabla f(x_n, \xi_{n+1})^{\odot 2}$
  $m_{n+1} = (1 - \gamma_{n+1} r_n) m_n + \gamma_{n+1} h_n \nabla f(x_n, \xi_{n+1})$
  $x_{n+1} = x_n - \gamma_{n+1} m_{n+1} / \sqrt{v_{n+1} + \varepsilon}$.
end for

We suppose throughout the paper that $1 - \gamma_{n+1} q_n \geq 0$ for all $n \in \mathbb{N}$. This will guarantee that the quantity $\sqrt{v_n + \varepsilon}$ is always well-defined (see Algorithm 1). This mild assumption is satisfied as soon as $q_0 \leq \frac{1}{\gamma_1}$ since the sequence (q_n) is non-increasing and the sequence of stepsizes (γ_n) can also be supposed to be non-increasing.

Since this algorithm makes use of the function $\nabla f(x, \xi)^{\odot 2}$, a strengthening of Assumption 3.1 is required:

Assumption 3.3. In Assumption 3.1, Conditions *ii)* and *iii)* are respectively replaced with the stronger conditions

ii') For each $x \in \mathbb{R}^d$, the function $\nabla f(x, \cdot)^{\odot 2}$ is μ -integrable.
iii') There exists a measurable map $\kappa : \mathbb{R}^d \times \Xi \to \mathbb{R}_+$ s.t. for every $x \in \mathbb{R}^d$:

  a) The map $\kappa(x, \cdot)$ is μ–integrable.

  b) There exists $\varepsilon > 0$ s.t. for every $u, v \in B(x, \varepsilon)$,

$$\|\nabla f(u, s) - \nabla f(v, s)\| \vee \|\nabla f(u, s)^{\odot 2} - \nabla f(v, s)^{\odot 2}\| \leq \kappa(x, s)\|u - v\|.$$

Under Assumption 3.3, we can also define the mapping $S : \mathbb{R}^d \to \mathbb{R}^d$ as:

$$S(x) = \mathbb{E}_\xi[\nabla f(x, \xi)^{\odot 2}]$$

for all $x \in \mathbb{R}^d$. Notice that Assumptions 2.1 and 2.3 are satisfied for F and S.

Assumption 3.4. Assume either of the following conditions.

i) There exists $q \geq 2$ s.t. for every compact set $\mathcal{K} \subset \mathbb{R}^d$,

$$\sup_{x \in \mathcal{K}} \mathbb{E}_\xi \|\nabla f(x, \xi)\|^{2q} < \infty \quad \text{and} \quad \sum_n \gamma_n^{1+q/2} < \infty.$$

ii) For every compact set $\mathcal{K} \subset \mathbb{R}^d$, there exists a real $\sigma_{\mathcal{K}} \neq 0$ s.t.

$$\mathbb{E}_\xi \exp\langle u, \nabla f(x,\xi) - \nabla F(x)\rangle \mathbb{1}_{x \in \mathcal{K}} \leq \exp\left(\sigma_{\mathcal{K}}^2 \|u\|^2/2\right) \text{ and}$$

$$\mathbb{E}_\xi \exp\langle u, \nabla f(x,\xi)^{\odot 2} - S(x)\rangle \mathbb{1}_{x \in \mathcal{K}} \leq \exp\left(\sigma_{\mathcal{K}}^2 \|u\|^2/2\right) ,$$

for every $x, u \in \mathbb{R}^d$. Moreover, for every $\alpha > 0$, $\sum_n \exp(-\alpha/\gamma_n) < \infty$.

Remark 4. We make the following comments regarding Assumption 3.4.

- Assumption 3.4-i) allows to use larger stepsizes in comparison to the classical condition $\sum_n \gamma_n^2 < \infty$ which corresponds to the particular case $q = 2$.
- Recall that a random vector X is said to be subgaussian if there exists a real $\sigma \neq 0$ s.t. $\mathbb{E}e^{\langle u, X\rangle} \leq e^{\sigma^2 \|u\|^2/2}$ for every constant vector $u \in \mathbb{R}^d$. In Assumption 3.4-ii), the subgaussian noise offers the possibility to use a sequence of stepsizes with an even slower decay rate than in Assumption 3.4–i).

Assumption 3.5. The set $F(\{x : \nabla F(x) = 0\})$ has an empty interior.

Remark 5. Assumption 3.5 excludes a pathological behavior of the objective function F at critical points. It is satisfied when $F \in \mathcal{C}^k(\mathbb{R}^d, \mathbb{R})$ for $k \geq d$. Indeed, in this case, Sard's theorem stipulates that the Lebesgue measure of $F(\{x : \nabla F(x) = 0\})$ is zero in $\mathbb{R}$.

Theorem 3.1. Let Assumptions 2.2, 2.4, and 3.2–3.5 hold true. Assume that the random sequence $(z_n = (v_n, m_n, x_n) : n \in \mathbb{N})$ given by Algorithm 1 is bounded with probability one. Then, w.p.1, the sequence (z_n) converges towards the set Υ defined in Eq. (2). If, in addition, the set of critical points of the objective function F is finite or countable, then w.p.1, the sequence (z_n) converges to a single point of Υ.

We now deal with the boundedness problem of the sequence (z_n). We introduce an additional assumption for this purpose.

Assumption 3.6. The following conditions hold.

i) ∇F is (globally) Lipschitz continuous.
ii) There exists $C > 0$ s.t. for all $x \in \mathbb{R}^d$, $\mathbb{E}_\xi[\|\nabla f(x,\xi)\|^2] \leq C(1 + F(x))$,
iii) $\sum_n \gamma_n^2 < \infty$.

Theorem 3.2. Let Assumptions 2.2, 2.4, 3.2, 3.3, 3.4-i) (with $q = 2$) and 3.6 hold. Then, the sequence (v_n, m_n, x_n) given by Algorithm 1 is bounded with probability one.

Remark 6. The above stability result requires square summable step sizes. Showing the same boundedness result under the Assumption 3.4 that allows for larger step sizes is a challenging problem in the general case. In these situations, the boundedness of the iterates can be sometimes ensured by *ad hoc* means.

Remark 7. We can also consider the noisy discretization of (ODE-1′) intro-
duced in Remark 3 above. This algorithm reads

$$
\begin{cases}
v_{n+1} & = (1 - \gamma_{n+1}q_n)v_n + \gamma_{n+1}p_n\nabla f(x_n, \xi_{n+1})^{\odot 2} & \text{(6a)} \\
x_{n+1} & = x_n - \gamma_{n+1}\nabla f(x_n, \xi_{n+1})/\sqrt{v_{n+1} + \varepsilon} & \text{(6b)}
\end{cases}
$$

for $(v_0, x_0) \in \mathbb{R}_+^d \times \mathbb{R}^d$. With only minor adaptations, Th. 3.1 and Th. 3.2 can
be shown to hold as well for this algorithm. We refer to the concomitant paper
[22, Sec. 2.2] for the link between this algorithm and the seminal algorithms
ADAGRAD [21] and RMSPROP [42].

3.2. Stochastic Nesterov's Accelerated Gradient (S-NAG)

S-NAG is the noisy Euler's discretization of (ODE-N). Given $\alpha > 0$, it generates
the sequence (m_n, x_n) on $\mathbb{R}^d \times \mathbb{R}^d$ given by Algorithm 2.

Algorithm 2 (S-NAG with decreasing steps)

Initialization: $m_0 = 0, x_0 \in \mathbb{R}^d$.
for $n = 1$ **to** n_{iter} **do**
    $m_{n+1} = (1 - \alpha\gamma_{n+1}/\tau_n)m_n + \gamma_{n+1}\nabla f(x_n, \xi_{n+1})$
    $x_{n+1} = x_n - \gamma_{n+1}m_{n+1}$.
end for

Assumption 3.7. Assume either of the following conditions.

i) There exists $q \geq 2$ s.t. for every compact set $\mathcal{K} \subset \mathbb{R}^d$,

$$
\sup_{x \in \mathcal{K}} \mathbb{E}_\xi \|\nabla f(x, \xi)\|^q < \infty \quad \text{and} \quad \sum_n \gamma_n^{1+q/2} < \infty \,.
$$

ii) For every compact set $\mathcal{K} \subset \mathbb{R}^d$, there exists a real $\sigma_\mathcal{K} \neq 0$ s.t.

$$
\mathbb{E}_\xi \exp\langle u, \nabla f(x, \xi) - \nabla F(x)\rangle \mathbb{1}_{x \in \mathcal{K}} \leq \exp\left(\sigma_\mathcal{K}^2 \|u\|^2/2\right) \,,
$$

for every $x, u \in \mathbb{R}^d$. Moreover, for every $\alpha > 0$, $\sum_n \exp(-\alpha/\gamma_n) < \infty$.

Theorem 3.3. Let Assumptions 2.2, 3.1, 3.2, 3.5 and 3.7 hold true. Assume that
the random sequence $(y_n = (m_n, x_n) : n \in \mathbb{N})$ given by Algorithm 2 is bounded
with probability one. Then, w.p.1, the sequence (y_n) converges towards the
set $\bar{\Upsilon}$ defined in Eq. (3). If, in addition, the set of critical points of the objective
function F is finite or countable, then w.p.1, the sequence (y_n) converges to a
single point of $\bar{\Upsilon}$.

The almost sure boundedness of the sequence (y_n) is handled in what follows.

Theorem 3.4. Let Assumptions 2.2, 3.1, 3.2 and 3.6 hold. Then, the sequence
$(y_n = (m_n, x_n) : n \in \mathbb{N})$ given by Algorithm 2 is bounded with probability one.

Remark 8. Assumption 3.4-i) in Th. 3.2 is not needed for Th. 3.4.

3.3. Central limit theorem

In this section, we establish a conditional central limit theorem for Algorithm 1.

Assumption 3.8. Let $x_\star \in \operatorname{zer} \nabla F$. The following holds.

   i) F is twice continuously differentiable on a neighborhood of $x_\star$ and the Hessian $\nabla^2 F(x_\star)$ is positive definite.

   ii) S is continuously differentiable on a neighborhood of $x_\star$.

   iii) There exists $M > 0$ and $b_M > 4$ s.t.

$$\sup_{x \in B(x_\star, M)} \mathbb{E}_\xi[\|\nabla f(x, \xi)\|^{b_M}] < \infty . \tag{7}$$

Under Assumptions 2.4-*i)* to *iii)*, it follows from Eq. (5) that the sequences $(h_n), (r_n), (p_n)$ and (q_n) of nonnegative reals converge respectively to $h_\infty, r_\infty, p_\infty$ and q_∞ where h_∞, r_∞ and q_∞ are supposed positive. Define $v_\star \triangleq q_\infty^{-1} p_\infty S(x_\star)$. Consider the matrix

$$V \triangleq \operatorname{diag}\left((\varepsilon + v_\star)^{\odot - \frac{1}{2}}\right) . \tag{8}$$

Let P be an orthogonal matrix s.t. the following spectral decomposition holds:

$$V^{\frac{1}{2}} \nabla^2 F(x_\star) V^{\frac{1}{2}} = P \operatorname{diag}(\pi_1, \cdots, \pi_d) P^{-1} ,$$

where $\pi_1 \leq \cdots \leq \pi_d$ are the (positive) eigenvalues of $V^{\frac{1}{2}} \nabla^2 F(x_\star) V^{\frac{1}{2}}$. Define

$$\mathcal{H} \triangleq \begin{bmatrix} -r_\infty I_d & h_\infty \nabla^2 F(x_\star) \\ -V & 0 \end{bmatrix}$$

where I_d is the $d \times d$ identity matrix. Then the matrix $\mathcal{H}$ is Hurwitz. Indeed, it can be shown that the largest real part of the eigenvalues of $\mathcal{H}$ coincides with $-L$, where

$$L \triangleq \frac{r_\infty}{2} \left(1 - \sqrt{\left(1 - \frac{4h_\infty \pi_1}{r_\infty^2}\right) \vee 0}\right) > 0 . \tag{9}$$

Assumption 3.9. The sequence (γ_n) is given by $\gamma_n = \frac{\gamma_0}{n^\alpha}$ for some $\alpha \in (0, 1]$, $\gamma_0 > 0$. Moreover, if $\alpha = 1$, we assume that $\gamma_0 > \frac{1}{2(L \wedge q_\infty)}$.

Theorem 3.5. Let Assumptions 2.4-*i)* to *iii)*, 3.3, 3.8 and 3.9 hold. Consider the iterates $z_n = (v_n, m_n, x_n)$ given by Algorithm 1. Set $\theta \triangleq 0$ if $\alpha < 1$ and $\theta \triangleq 1/(2\gamma_0)$ if $\alpha = 1$. Assume that the event $\{z_n \to z_\star\}$, where $z_\star = (v_\star, 0, x_\star)$, has a positive probability. Then, given that event,

$$\frac{1}{\sqrt{\gamma_n}} \begin{bmatrix} m_n \\ x_n - x_\star \end{bmatrix} \Rightarrow \mathcal{N}(0, \Gamma) ,$$

where $\Rightarrow$ stands for the convergence in distribution and $\mathcal{N}(0, \Gamma)$ is a centered Gaussian distribution on $\mathbb{R}^{2d}$ with a covariance matrix Γ given by the unique solution to the Lyapunov equation

$$(\mathcal{H} + \theta I_{2d})\Gamma + \Gamma(\mathcal{H} + \theta I_{2d})^{\mathrm{T}} = - \begin{bmatrix} \operatorname{Cov}(h_\infty \nabla f(x_\star, \xi)) & 0 \\ 0 & 0 \end{bmatrix} .$$

In particular, given $\{z_n \to z_\star\}$, the vector $\sqrt{\gamma_n}^{-1}(x_n - x_\star)$ converges in distribution to a centered Gaussian distribution with a covariance matrix given by:

$$\Gamma_2 = V^{\frac{1}{2}} P \left[\frac{C_{k,\ell}}{\frac{r_\infty - 2\theta}{h_\infty}(\pi_k + \pi_\ell + \frac{2\theta(\theta - r_\infty)}{h_\infty}) + \frac{(\pi_k - \pi_\ell)^2}{2(r_\infty - 2\theta)}} \right]_{k,\ell = 1\ldots d} P^{-1} V^{\frac{1}{2}} \quad (10)$$

where $C \triangleq P^{-1} V^{\frac{1}{2}} \mathbb{E}_\xi \left[\nabla f(x_\star, \xi) \nabla f(x_\star, \xi)^{\mathrm{T}} \right] V^{\frac{1}{2}} P$.

A few remarks are in order.

- The matrix Γ_2 coincides with the limiting covariance matrix associated to the iterates

$$\begin{cases} m_{n+1} &= m_n + \gamma_{n+1}(h_\infty V \nabla f(x_n, \xi_{n+1}) - r_\infty m_n) \\ x_{n+1} &= x_n - \gamma_{n+1} m_{n+1}\,. \end{cases}$$

  This procedure can be seen as a preconditioned version of the stochastic heavy ball algorithm [23] although the iterates are not implementable because of the unknown matrix V. Notice also that the limiting covariance Γ_2 depends on $v_\star$ but does not depend on the fluctuations of the sequence (v_n).
- When $h_\infty = r_\infty$ (which is the case for ADAM), we recover the expression of the asymptotic covariance matrix previously provided in [8, Section 5.3] and the remarks formulated therein.
- The assumption $r_\infty > 0$ is crucial to establish Th. 3.5. For this reason, Th. 3.5 does not generalize immediately to Algorithm 2. The study of the fluctuations of Algorithm 2 is left for future works.

3.4. Related works

In [23], Gadat, Panloup and Saadane study the SHB algorithm, which is a noisy Euler's discretization of (ODE-1) in the situation where $\mathsf{h} = \mathsf{r}$ and $\mathsf{p} = \mathsf{q} \equiv 0$ (i.e., there is no v variable). In this framework, if we set $\mathsf{h} = \mathsf{r} \equiv r > 0$ in Algorithm 1 above, then Th. 3.1 above recovers the analogous case in [23, Th. 2.1], which is termed as the exponential memory case. The other important case treated in [23] is the case where $\mathsf{h}(t) = \mathsf{r}(t) = r/t$ for some $r > 0$, referred to as the polynomially memory case. Actually, it is known that the ODE obtained for $\mathsf{h}(t) = \mathsf{r}(t) = r/t$ and $\mathsf{p} = \mathsf{q} \equiv 0$ boils down to (ODE-N) after a time variable change (see, e.g., Lem. 5.3 below). Nevertheless, we highlight that the stochastic algorithm that stems from this ODE and that is studied in [23] is different from the S-NAG algorithm introduced above which stems from a different ODE (ODE-N). Hence, the convergence result of Th. 3.3 for the S-NAG algorithm we consider is not covered by the analysis of [23].

The specific case of the ADAM algorithm is analyzed in [8] in both the constant and vanishing stepsize settings (see [8, Ths. 5.2-5.4] which are the analogues of our Ths. 3.1-3.2). Note that we deal with a more general algorithm in the present paper. Indeed, Algorithm 1 offers some freedom in the choice of the functions h, r, p, q satisfying Assumption 2.4 beyond the specific case of

the ADAM algorithm studied in [8]. Apart from this generalization, we also emphasize some small improvements. Regarding Theorem 3.1, we provide noise conditions allowing to choose larger stepsizes (see Assumption 3.4 compared to [8, Assumption 4.2]). Concerning the stability result (Th. 3.2), we relax [8, Assumption 5.3-(iii)] which is no more needed in the present paper (see Assumption 3.6) thanks to a modification of the discretized Lyapunov function used in the proof (see Section 6.4 compared to [8, Section 9.2]).

In most generality, the almost sure convergence result of the iterates of Algorithm 1 using vanishing stepsizes (Ths. 3.1-3.2) is new to the best of our knowledge. Moreover, while some recent results exist for S-NAG in the constant stepsize and for convex objective functions (see for e.g. [4]), Ths. 3.3 and 3.4 which tackle the possibly non-convex setting are also new to the best of our knowledge.

In the work [22] that was posted on the arXiv repository a few days after our submission, Gadat and Gavra study the specific case of the algorithm described in Eq. (6) encompassing both ADAGRAD and RMSPROP, with the possibility to use mini-batches. For this specific algorithm, the authors establish a similar almost sure convergence result to ours [22, Th. 1] for decreasing stepsizes and derive some quantitative results bounding in expectation the gradient of the objective function along the iterations for constant stepsizes [22, Th. 2]. We highlight though that they do not consider the presence of momentum in the algorithm. Therefore, their analysis does not cover neither Algorithm 1 nor Algorithm 2.

In contrast to our analysis, some works in the literature explore the constant stepsize regime for some stochastic momentum methods either for smooth [43] or weakly convex objective functions [32]. Furthermore, concerning ADAM-like algorithms, several recent works control the minimum of the norms of the gradients of the objective function evaluated at the iterates of the algorithm over N iterations in expectation or with high probability [17, 45, 14, 46, 15, 44, 1, 18, 2] and establish regret bounds in the convex setting [2].

Similar central limit theorems to Th. 3.5 are established in the cases of the stochastic heavy ball algorithm with exponential memory [23, Th. 2.4] and ADAM [8, Th. 5.7]. In comparison to [23], we precise that our theorem recovers their result and provides a closed formula for the asymptotic covariance matrix Γ_2. Our proof of Th. 3.5 differs from the strategies adopted in [23] and [8].

4. Avoidance of traps

In Th. 3.1 and Th. 3.3 above, we established the almost sure convergence of the iterates x_n towards the set of critical points of the objective function F for both Algorithms 1 and 2. However, the landscape of F can contain what is known as "traps" for the algorithm, namely, critical points where the Hessian matrix of F has negative eigenvalues, making these critical points local maxima or saddle points. In this section, we show that the convergence of the iterates to these traps does not take place if the noise is exciting in some directions.

Starting with the contributions of Pemantle [38] and Brandière and Duflo [12], the numerous so-called avoidance of traps results that can be found in the literature deal with the case where the ODE that underlies the stochastic algorithm is an autonomous ODE. Obviously, this is neither the case of (ODE-1), nor of (ODE-N). To deal with this issue, we first state a general avoidance of traps result that extends [38, 12] to a non-autonomous setting, and that has an interest of its own. We then apply this result to Algorithms 1 and 2.

4.1. A general avoidance-of-traps result in a non-autonomous setting

The notations in this subsection and in Sections 7.1–7.2 are independent from the rest of the paper. We recall that for a function $h : \mathbb{R}^d \to \mathbb{R}^{d'}$, we denote by $\partial_i^k h(x_1, \ldots, x_d)$ the kth partial derivative of the function h with respect to x_i.

The setting of our problem is as follows. Given an integer $d > 0$ and a continuous function $b : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^d$, we consider a stochastic algorithm built around the non-autonomous ODE $\dot{\mathsf{z}}(t) = b(\mathsf{z}(t), t)$. Let $z_\star \in \mathbb{R}^d$, and assume that on $\mathcal{V} \times \mathbb{R}_+$ where $\mathcal{V}$ is a certain neighborhood of $z_\star$, the function b can be developed as

$$b(z, t) = D(z - z_\star) + e(z, t), \tag{11}$$

where $e(z_\star, \cdot) \equiv 0$, and where the matrix $D \in \mathbb{R}^{d \times d}$ is assumed to admit the following spectral factorization: Given $0 \leq d^- < d$ and $0 < d^+ \leq d$ with $d^- + d^+ = d$, we can write

$$D = Q\Lambda Q^{-1}, \quad \Lambda = \begin{bmatrix} \Lambda^- & \\ & \Lambda^+ \end{bmatrix}, \tag{12}$$

where the Jordan blocks that constitute $\Lambda^- \in \mathbb{R}^{d^- \times d^-}$ (respectively $\Lambda^+ \in \mathbb{R}^{d^+ \times d^+}$) are those that contain the eigenvalues λ_i of D for which $\Re\lambda_i \leq 0$ (respectively $\Re\lambda_i > 0$). Since $d^+ > 0$, the point $z_\star$ is an unstable equilibrium point of the ODE $\dot{\mathsf{z}}(t) = b(\mathsf{z}(t), t)$, in the sense that the ODE solution will only be able to converge to $z_\star$ along a specific so-called invariant manifold which precise characterization will be given in Section 7.1 below.

We now consider a stochastic algorithm that is built around this ODE. The condition $d^+ > 0$ makes that $z_\star$ is a trap that the algorithm should desirably avoid. The following theorem states that this will be the case if the noise term of the algorithm is omnidirectional enough. The idea is to show that the case being, the algorithm trajectories will move away from the invariant manifold mentioned above.

Theorem 4.1. Given a sequence (γ_n) of nonnegative deterministic stepsizes such that $\sum_n \gamma_n = +\infty$, $\sum_n \gamma_n^2 < +\infty$, and a filtration $(\mathscr{F}_n)$, consider the stochastic approximation algorithm in $\mathbb{R}^d$

$$z_{n+1} = z_n + \gamma_{n+1} b(z_n, \tau_n) + \gamma_{n+1}\eta_{n+1} + \gamma_{n+1}\rho_{n+1}$$

where $\tau_n = \sum_{k=1}^n \gamma_k$. Assume that the sequences (η_n) and (ρ_n) are adapted to $\mathscr{F}_n$, and that z_0 is $\mathscr{F}_0$–measurable. Assume that there exists $z_\star \in \mathbb{R}^d$ such that Eq. (11) holds true on $\mathcal{V} \times \mathbb{R}_+$, where $\mathcal{V}$ is a neighborhood of $z_\star$. Consider the spectral factorization (12), and assume that $d^+ > 0$. Assume moreover that the function e at the right hand side of Eq. (11) satisfies the conditions:

  i) $e(z_\star, \cdot) \equiv 0$.
  ii) On $\mathcal{V} \times \mathbb{R}_+$, the functions $\partial_2^n \partial_1^k e(z, t)$ exist and are continuous for $0 \leq n < 2$ and $0 \leq k + n \leq 2$.
  iii) The following convergence holds:

$$\lim_{(z,t) \to (z_\star, \infty)} \|\partial_1 e(z, t)\| = 0. \tag{13}$$

  iv) There exist $t_0 > 0$ and a neighborhood $\mathcal{W} \subset \mathbb{R}^d$ of $z_\star$ s.t.

$$\sup_{z \in \mathcal{W}, t \geq t_0} \|\partial_2 e(z, t)\| < +\infty \quad \text{and} \quad \sup_{z \in \mathcal{W}, t \geq t_0} \|\partial_1^2 e(z, t)\| < +\infty.$$

Moreover, suppose that:

  v) $\sum_n \|\rho_{n+1}\|^2 \mathbb{1}_{z_n \in \mathcal{W}} < \infty$ almost surely.
  vi) $\limsup \mathbb{E}[\|\eta_{n+1}\|^4 \,|\, \mathscr{F}_n] \mathbb{1}_{z_n \in \mathcal{W}} < \infty$, and $\mathbb{E}[\eta_{n+1} \,|\, \mathscr{F}_n] \mathbb{1}_{z_n \in \mathcal{W}} = 0$.
  vii) Writing $\tilde{\eta}_n = Q^{-1}\eta_n = (\tilde{\eta}_n^-, \tilde{\eta}_n^+)$ with $\tilde{\eta}_n^\pm \in \mathbb{R}^{d^\pm}$, for some $c^2 > 0$, it holds that

$$\liminf \mathbb{E}[\|\tilde{\eta}_{n+1}^+\|^2 \,|\, \mathscr{F}_n] \mathbb{1}_{z_n \in \mathcal{W}} \geq c^2 \mathbb{1}_{z_n \in \mathcal{W}}.$$

Then, $\mathbb{P}([z_n \to z_\star]) = 0$.

Remark 9. Assumptions *i)* to *iv)* of Th. 4.1 are related to the function e defined in Eq. (11), which can be seen as a non-autonomous perturbation of the autonomous linear ODE $\dot{z}(t) = D(z(t) - z_\star)$. These assumptions guarantee the existence of a local (around the unstable equilibrium $z_\star$) non-autonomous invariant manifold of the non-autonomous ODE $\dot{z}(t) = b(z(t), t)$ with enough regularity properties, as provided by Prop. 7.1 and Prop. 7.3 below.

4.2. Application to the stochastic algorithms

4.2.1. Trap avoidance of the general algorithm 1

In Th. 3.1 above, we showed that the sequence (z_n) generated by Algorithm 1 converges almost surely towards the set Υ defined in Eq. (2). Our purpose now is to show that the traps in Υ (to be characterized below) are avoided by the stochastic algorithm 1 under a proper omnidirectionality assumption on the noise.

Our first task is to write Algorithm 1 in a manner compatible with the statement of Th. 4.1. The following decomposition holds for the sequence $(z_n = (v_n, m_n, x_n), n \in \mathbb{N})$ generated by this algorithm:

$$z_{n+1} = z_n + \gamma_{n+1} g(z_n, \tau_n) + \gamma_{n+1} \eta_{n+1} + \gamma_{n+1} \tilde{\rho}_{n+1},$$

where $\tilde{\rho}_{n+1} = \left(0, 0, \frac{m_n}{\sqrt{v_n+\varepsilon}} - \frac{m_{n+1}}{\sqrt{v_{n+1}+\varepsilon}}\right)$, and where η_{n+1} is the martingale increment with respect to the filtration $(\mathscr{F}_n)$ which is defined by Eq. (28).

Observe from Eq. (2) that each $z_\star \in \Upsilon$ is written as $z_\star = (v_\star, 0, x_\star)$ where $x_\star \in \mathrm{zer}\,\nabla F$, and $v_\star = q_\infty^{-1} p_\infty S(x_\star)$ (in particular, $x_\star$ and $z_\star$ are in a one-to-one correspondence). We need to linearize the function $g(\cdot, t)$ around $z_\star$. The following assumptions will be required.

Assumption 4.1. The functions F and S belong respectively to $\mathcal{C}^3(\mathbb{R}^d, \mathbb{R})$ and $\mathcal{C}^2(\mathbb{R}^d, \mathbb{R}_+^d)$.

Assumption 4.2. The functions $\mathsf{h}, \mathsf{r}, \mathsf{p}, \mathsf{q}$ belong to $\mathcal{C}^1((0,\infty), \mathbb{R}_+)$ and have bounded derivatives on $[t_0, +\infty)$ for some $t_0 > 0$.

Lemma 4.2. Let Assumptions 2.4-*i)* to *iii)*, 4.1 and 4.2 hold. Let $z_\star = (v_\star, 0, x_\star)$ $\in \Upsilon$. Then, for every $z \in \mathcal{Z}_+$ and every $t > 0$, the following decomposition holds true:

$$g(z, t) = D(z - z_\star) + e(z, t) + c(t),$$

where $D = \begin{bmatrix} -q_\infty I_d & 0 & p_\infty \nabla S(x_\star) \\ 0 & -r_\infty I_d & h_\infty \nabla^2 F(x_\star) \\ 0 & -V & 0 \end{bmatrix}$, $\quad c(t) = \begin{bmatrix} \mathsf{p}(t)S(x_\star) - \mathsf{q}(t)v_\star \\ 0 \\ 0 \end{bmatrix}$,

and the function $e(z, t)$ (defined in Section 7.3.1 below for conciseness) has the same properties as its analogue in the statement of Th. 4.1.

Using this lemma, the algorithm iterate z_{n+1} can be rewritten as an instance of the algorithm in the statement of Th. 4.1, namely,

$$z_{n+1} = z_n + \gamma_{n+1} b(z_n, \tau_n) + \gamma_{n+1} \eta_{n+1} + \gamma_{n+1} \rho_{n+1}, \tag{14}$$

where in our present setting, $b(z, t) = g(z, t) - c(t) = D(z - z_\star) + e(z, t)$ and $\rho_n = c(\tau_{n-1}) + \tilde{\rho}_n$. In the following assumption, we use the well-known fact that a symmetric matrix H has the same inertia as AHA^T for an arbitrary invertible matrix A.

Assumption 4.3. Let $x_\star \in \mathrm{zer}\,\nabla F$, let $v_\star = q_\infty^{-1} p_\infty S(x_\star)$, and define the diagonal matrix $V = \mathrm{diag}((v_\star + \varepsilon)^{\odot - \frac{1}{2}})$ as in (8). Assume the following conditions:

i) $\sum_n (q_\infty p_n - p_\infty q_n)^2 < \infty$,
ii) The Hessian matrix $\nabla^2 F(x_\star)$ has a negative eigenvalue.
iii) There exists $\delta > 0$ such that $\sup_{x \in B(x_\star, \delta)} \mathbb{E}_\xi[\|\nabla f(x, \xi)\|^8] < \infty$.
iv) Defining Π_u as the orthogonal projector on the eigenspace of $V^{\frac{1}{2}} \nabla^2 F(x_\star) V^{\frac{1}{2}}$ that is associated with the negative eigenvalues of this matrix, it holds that

$$\Pi_\mathrm{u} V^{\frac{1}{2}} \mathbb{E}_\xi (\nabla f(x_\star, \xi) - \nabla F(x_\star))(\nabla f(x_\star, \xi) - \nabla F(x_\star))^\mathrm{T} V^{\frac{1}{2}} \Pi_\mathrm{u} \neq 0.$$

Theorem 4.3. Let Assumptions 2.4, 3.3, and 4.1, 4.2 hold true. Let $z_\star \in \Upsilon$ be such that Assumption 4.3 holds true for this $z_\star$. Then, the eigenspace associated with the eigenvalues of D with positive real parts has the same dimension

as the eigenspace of $\nabla^2 F(x_\star)$ associated with the negative eigenvalues of this matrix. Let $(z_n = (v_n, m_n, x_n) : n \in \mathbb{N})$ be the random sequence generated by Algorithm 1 with stepsizes satisfying $\sum_n \gamma_n = +\infty$ and $\sum_n \gamma_n^2 < +\infty$. Then, $\mathbb{P}([z_n \to z_\star]) = 0$.

The assumptions and the result call for some comments.

Remark 10. The definition of a trap as regards the general algorithm in the statement of Th. 4.1 is that the matrix D in Eq. (11) has eigenvalues with positive real parts. Th. 4.3 states that this condition is equivalent to $\nabla^2 F(x_\star)$ having negative eigenvalues. What's more, the dimension of the invariant subspace of D corresponding to the eigenvalues with positive real parts is equal to the dimension of the negative eigenvalue subspace of $\nabla^2 F(x_\star)$. Thus, Assumption 4.3–*iv)* provides the "largest" subspace where the noise energy must be non zero for the purpose of avoiding the trap.

Remark 11. Assumptions 4.2 and 4.3-*i)* are satisfied by many widely studied algorithms, among which RMSPROP and ADAM.

Remark 12. The results of Th. 4.3 can be straightforwardly adapted to the case of (ODE-1′). Assumption 4.3-*iv)* on the noise is unchanged.

In the case of the S-NAG algorithm, the assumptions become particularly simple. We state the afferent result separately.

4.2.2. Trap avoidance for S-NAG

Assumption 4.4. Let $x_\star \in \text{zer} \nabla F$ and let the following conditions hold.

- *i)* The Hessian matrix $\nabla^2 F(x_\star)$ has a negative eigenvalue.
- *ii)* There exists $\delta > 0$ such that $\sup_{x \in B(x_\star, \delta)} \mathbb{E}_\xi[\|\nabla f(x, \xi)\|^4] < \infty$.
- *iii)* $\tilde{\Pi}_u \mathbb{E}_\xi(\nabla f(x_\star, \xi) - \nabla F(x_\star))(\nabla f(x_\star, \xi) - \nabla F(x_\star))^{\mathrm{T}} \tilde{\Pi}_u \neq 0$, where $\tilde{\Pi}_u$ is the orthogonal projector on the eigenspace of $\nabla^2 F(x_\star)$ associated with its negative eigenvalues.

Theorem 4.4. Let Assumptions 2.4, 3.1, 4.1 and 4.4 hold. Define $y_\star = (0, x_\star)$. Let $(y_n = (m_n, x_n) : n \in \mathbb{N})$ be the random sequence given by Algorithm 2 with stepsizes satisfying $\sum_n \gamma_n = +\infty$ and $\sum_n \gamma_n^2 < +\infty$. Then, $\mathbb{P}([y_n \to y_\star]) = 0$.

4.3. Related works

Up to our knowledge, all the avoidance of traps results that can be found in the literature, starting from [38, 12], refer to stochastic algorithms that are discretizations of autonomous ODE's (see for e.g., [10, Sec. 9] for general Robbins Monro algorithms and [33, Sec. 4.3] for SGD). In this line of research, a powerful class of techniques relies on Poincaré's invariant manifold theorem for an autonomous ODE in a neighborhood of some unstable equilibrium point. In our work, we extend the avoidance of traps results to a non-autonomous setting,

by borrowing a non-autonomous version of Poincaré's theorem from the rich literature that exists on the subject [16, 30].

In [23], the authors succeeded in establishing an avoidance of traps result for their non-autonomous stochastic algorithm which is close to our S-NAG algorithm (see the discussion at the end of Section 3.4 above), at the expense of a sub-Gaussian assumption on the noise and a rather stringent assumption on the stepsizes. The main difficulty in the approach of [23] lies in the use of the classical autonomous version of Poincaré's theorem (see [23, Remark 2.1]). This kind of difficulty is avoided by our approach, which allows to obtain avoidance of traps results with close to minimal assumptions. More recently, in the contribution of [22] discussed in Sec. 3.4, the authors establish an avoidance of traps result ([22, Th. 3]) for the algorithm described in Eq. (6) using techniques inspired from [38, 10]. As previously mentioned, this recent work does not handle momentum and hence neither Algorithm 1 nor Algorithm 2. Moreover, as indicated in our discussion of [23], our strategy of proof is different.

Taking another point of view as concerns the trap avoidance, some recent works [31, 20, 27, 35, 36] address the problem of escaping saddle points when the algorithm is deterministic but when the initialization point is random. In contrast to this line of research, our work considers a stochastic algorithm for which randomness enters into play at each iteration of the algorithm via noisy gradients.

5. Proofs for Section 2

5.1. Proof of Th. 2.1

The arguments of the proof of this theorem that we provide here follow the approach of [9] with some small differences. Close arguments can be found in [8]. We provide the proof here for completeness and in preparation of the proofs that will be related with the stochastic algorithms.

5.1.1. Existence and uniqueness

The following lemma guarantees that the term $\sqrt{\mathsf{v}(t) + \varepsilon}$ in (ODE-1) is well-defined.

Lemma 5.1. Let $t_0 \in \mathbb{R}_+$ and $T \in (t_0, \infty]$. Assume that there exists a solution $\mathsf{z}(t) = (\mathsf{v}(t), \mathsf{m}(t), \mathsf{x}(t))$ to (ODE-1) on $[t_0, T)$ for which $\mathsf{v}(t_0) \geq 0$. Then, for all $t \in [t_0, T)$, $\mathsf{v}(t) \geq 0$.

Proof. Assume that $\nu \triangleq \inf\{t \in [t_0, T), \mathsf{v}(t) < 0\}$ satisfies $\nu < T$. If $\mathsf{v}(t_0) > 0$, Gronwall's lemma implies that $\mathsf{v}(t) \geq \mathsf{v}(t_0) \exp(-\int_{t_0}^{t} q(t))$ on $[t_0, \nu]$ which is in contradiction with the fact that $\mathsf{v}(\nu) = 0$. If $\mathsf{v}(t_0) = 0$, since $\nu < T$, there exists $t_1 \in (t_0, \nu)$ s.t. $\dot{\mathsf{v}}(t_1) < 0$. Hence, using the first equation from (ODE-1), we obtain $\mathsf{v}(t_1) > 0$. This brings us back to the first case, replacing t_0 by t_1. $\quad\square$

Recall that $F_\star = \inf F$ is finite by Assumption 2.2. Of prime importance in the proof will be the energy (Lyapunov) function $\mathcal{E} : \mathbb{R}_+ \times \mathcal{Z}_+ \to \mathbb{R}$, defined as

$$\mathcal{E}(h, z) = h(F(x) - F_\star) + \frac{1}{2} \left\| \frac{m}{(v + \varepsilon)^{\odot \frac{1}{4}}} \right\|^2 , \tag{15}$$

for every $h \geq 0$ and every $z = (v, m, x) \in \mathcal{Z}_+$. This function is slightly different from its analogues that were used in [3, 8, 9].

Consider $(t, z) \in (0, +\infty) \times \mathcal{Z}_+$ and set $z = (v, m, x)$. Then, using Assumption 2.1, we can write

$$\partial_t \mathcal{E}(h(t), z) + \langle \nabla_z \mathcal{E}(h(t), z), g(z, t) \rangle$$

$$= \dot{h}(t)(F(x) - F_\star) - \frac{1}{4} \langle \frac{m^{\odot 2}}{(v + \varepsilon)^{\odot \frac{3}{2}}}, p(t)S(x) - q(t)v \rangle$$

$$+ \langle \frac{m}{(v + \varepsilon)^{\odot \frac{1}{2}}}, h(t)\nabla F(x) - r(t)m \rangle - \langle \frac{m}{(v + \varepsilon)^{\odot \frac{1}{2}}}, h(t)\nabla F(x) \rangle$$

$$\leq - \left(r(t) - \frac{q(t)}{4} \right) \left\| \frac{m}{(v + \varepsilon)^{\odot \frac{1}{4}}} \right\|^2 + \dot{h}(t)(F(x) - F_\star) - \frac{p(t)}{4} \langle S(x), \frac{m^{\odot 2}}{(v + \varepsilon)^{\odot \frac{3}{2}}} \rangle. \tag{16}$$

With the help of this function, we can now establish the existence, the uniqueness and the boundedness of the solution of (ODE-1) on $[t_0, \infty)$ for an arbitrary $t_0 > 0$.

Lemma 5.2. For each $t_0 > 0$ and $z_0 \in \mathcal{Z}_+$, (ODE-1) has a unique solution on $[t_0, \infty)$ starting at $z(t_0) = z_0$. Moreover, the orbit $\{z(t) : t \geq t_0\}$ is bounded.

Proof. Let $t_0 > 0$, and fix $z_0 \in \mathcal{Z}_+$. On each set of the type $[t_0, t_0 + A] \times \bar{B}(z_0, R)$ where $A, R > 0$ and $\bar{B}(z_0, R) \subset (-\varepsilon, \infty)^d \times \mathbb{R}^d \times \mathbb{R}^d$, we easily obtain from our assumptions that the function g defined in (1) is continuous, and that $g(\cdot, t)$ is uniformly Lipschitz on $t \in [t_0, t_0 + A]$. In these conditions, Picard's theorem asserts that (ODE-1) starting from $z(t_0) = z_0$ has a unique solution on a certain maximal interval $[t_0, T)$. Lem. 5.1 shows that $v(t) \geq 0$ on this interval.

Let us show that $T = \infty$. Applying Ineq. (16) with $(v, m, x) = (v(t), m(t), x(t))$ and using Assumption 2.4, we obtain that the function $t \mapsto \mathcal{E}(h(t), z(t))$ is decreasing on $[t_0, T)$. By the coercivity of F (Assumption 2.2) and Assumption 2.4–*i*), we get that the trajectory $\{x(t)\}$ is bounded. Recall the equation $\dot{m}(t) = h(t)\nabla F(x(t)) - r(t)m(t)$. Using the continuity of the functions ∇F, h and r along with Gronwall's lemma, we get that $\{m(t)\}$ is bounded if $T < \infty$. We can show a similar result for $\{v(t)\}$. Thus, $\{z(t)\}$ is bounded on $[t_0, T)$ if $T < \infty$ which is a contradiction, see, *e.g.*, [25, Cor.3.2].

It remains to show that the trajectory $\{z(t)\}$ is bounded. To that end, let us apply the variation of constants method to the equation $\dot{m}(t) = h(t)\nabla F(x(t)) - r(t)m(t)$. Writing $R(t) = \int_{t_0}^t r(u)\,du$, we get that

$$\frac{d}{dt} \left(e^{R(t)} m(t) \right) = e^{R(t)} h(t) \nabla F(x(t)).$$

Therefore, for every $t \geq t_0$,

$$\mathsf{m}(t) = e^{-R(t)}\mathsf{m}(t_0) + \int_{t_0}^{t} e^{R(u)-R(t)}\mathsf{h}(u)\nabla F(\mathsf{x}(u))du\,.$$

Using the continuity of ∇F together with the boundedness of x, Assumption 2.4 and the triangle inequality, we obtain the existence of a constant $C > 0$ independent of t s.t.

$$\|\mathsf{m}(t) - \mathsf{m}(t_0)\| - \|\mathsf{m}(t_0)\| \leq C\mathsf{h}(t_0) \int_{t_0}^{t} e^{-\int_{u}^{t}\mathsf{r}(s)\,ds}du$$

$$\leq C\mathsf{h}(t_0) \int_{t_0}^{t} e^{-r_\infty(t-u)}du \leq \frac{C\mathsf{h}(t_0)}{r_\infty}\,.$$

The same reasoning applies to $\mathsf{v}(t)$ using the continuity of S and Assumption 2.4. This completes the proof. $\qquad\square$

We can now extend this solution to $t_0 = 0$ along the approach of [9], where the detailed derivations can be found. The idea is to replace $\mathsf{h}(t)$ with $\mathsf{h}(\max(\eta, t))$ for some $\eta > 0$ and to do the same for p, q, and r. It is then easy to see that the ODE that is obtained by doing these replacements has a unique global solution on $\mathbb{R}_+$. By making $\eta \to 0$ and by using the Arzelà-Ascoli theorem along with Assumption 2.5, we obtain that (ODE-1) has a unique solution on $\mathbb{R}_+$.

5.1.2. *Convergence*

The first step in this part consists in transforming (ODE-1) into an autonomous ODE by including the time variable into the state vector. More specifically, we start with the following ODE:

$$\begin{bmatrix} \dot{\mathsf{z}}(t) \\ \dot{u}(t) \end{bmatrix} = \begin{bmatrix} g(\mathsf{z}(t), u(t)) \\ 1 \end{bmatrix} \quad \text{with} \quad \begin{bmatrix} \mathsf{z}(0) \\ u(0) \end{bmatrix} = \begin{bmatrix} z_0 \\ t_0 \end{bmatrix},$$

then, we perform the following change of variable in time

$$\begin{bmatrix} z \\ u \end{bmatrix} \mapsto \begin{bmatrix} z \\ s = 1/u \end{bmatrix}$$

allowing the solution to lie in a compact set.

We initialize the above ODE at a time instant $t_0 > 0$. Define the functions $\mathsf{H}, \mathsf{R}, \mathsf{P}, \mathsf{Q} : \mathbb{R}_+ \to \mathbb{R}_+$ by setting $\mathsf{H}(s) = \mathsf{h}(1/s)$, $\mathsf{R}(s) = \mathsf{r}(1/s)$, $\mathsf{P}(s) = \mathsf{p}(1/s)$; $\mathsf{Q}(s) = \mathsf{q}(1/s)$ for $s > 0$; $\mathsf{H}(0) = h_\infty$, $\mathsf{R}(0) = r_\infty$, $\mathsf{P}(0) = p_\infty$ and $\mathsf{Q}(0) = q_\infty$. Our autonomous dynamical system can then be described by the following system of equations:

$$\begin{cases} \dot{\mathsf{v}}(t) &= \mathsf{P}(\mathsf{s}(t))S(\mathsf{x}(t)) - \mathsf{Q}(\mathsf{s}(t))\mathsf{v}(t) \\ \dot{\mathsf{m}}(t) &= \mathsf{H}(\mathsf{s}(t))\nabla F(\mathsf{x}(t)) - \mathsf{R}(\mathsf{s}(t))\mathsf{m}(t) \\ \dot{\mathsf{x}}(t) &= -\dfrac{\mathsf{m}(t)}{\sqrt{\mathsf{v}(t)+\varepsilon}} \\ \dot{\mathsf{s}}(t) &= -\mathsf{s}(t)^2 \end{cases} \qquad (17)$$

Since the solution of the ODE $\dot{\mathsf{s}}(t) = -\mathsf{s}(t)^2$ for which $\mathsf{s}(t_0) = 1/t_0$ is $\mathsf{s}(t) = 1/t$, the trajectory $\{\mathsf{s}(t)\}$ is bounded. The three remaining equations are a reformulation of (ODE-1) for which the trajectories have already been shown to exist and to be bounded in Lem. 5.2. In the sequel, we denote by $\Phi : \mathcal{Z}_+ \times \mathbb{R}_+ \to \mathcal{Z}_+ \times \mathbb{R}_+$ the semiflow induced by the autonomous ODE (17), *i.e.*, for every $u = (z, s) \in \mathcal{Z}_+ \times \mathbb{R}_+$, $\Phi(u, \cdot)$ is the unique global solution to the autonomous ODE (17) initialized at u. Observe that the orbits of this semiflow are precompact. Moreover, the function $\Phi((z, 0), \cdot)$ is perfectly defined for each $z \in \mathcal{Z}_+$ since the associated solution satisfies the ODE (19) defined below, which three first equations satisfy the hypotheses of Lem. 5.2.

Consider now a continuous function $V : \mathcal{Z}_+ \times \mathbb{R}_+ \to \mathbb{R}$ defined by:

$$V(u) = \mathcal{E}\left(\mathsf{H}(s), z\right), \quad u = (z, s) \in \mathcal{Z}_+ \times (0, \infty).$$

As for Ineq. (16) above, we have here that

$$\frac{d}{dt} V\left(\Phi(u, t)\right) \leq -\left(\mathsf{r}(t) - \frac{\mathsf{q}(t)}{4}\right) \left\|\frac{\mathsf{m}(t)}{(\mathsf{v}(t) + \varepsilon)^{\odot \frac{1}{4}}}\right\|^2$$
$$+ \dot{\mathsf{h}}(t)(F(\mathsf{x}(t)) - F_\star) - \frac{\mathsf{p}(t)}{4} \langle S(\mathsf{x}(t)), \frac{\mathsf{m}(t)^{\odot 2}}{(\mathsf{v}(t) + \varepsilon)^{\odot \frac{3}{2}}} \rangle$$

if $s > 0$, and the same inequality with $(\dot{\mathsf{h}}(t), \mathsf{p}(t), \mathsf{r}(t), \mathsf{q}(t))$ being replaced with $(0, p_\infty, r_\infty, q_\infty)$ otherwise.

Since $V \circ \Phi(u, \cdot)$ is non-increasing and nonnegative, we can define $V_\infty \triangleq \lim_{t \to \infty} V(\Phi(u, t))$. Let $\omega(u) \triangleq \bigcap_{s > 0} \overline{\bigcup_{t \geq s} \Phi(u, t)}$ be the ω-limit set of the semiflow Φ issued from u. Recall that $\omega(u)$ is an invariant set for the flow $\Phi(u, \cdot)$, and that

$$\mathrm{dist}(\Phi(u, t), \omega(u)) \xrightarrow[t \to \infty]{} 0,$$

see, *e.g.*, [24, Th. 1.1.8]). In order to finish the proof of Th. 2.1, we need to make explicit the structure of $\omega(u)$.

We know from La Salle's invariance principle that $\omega(u) \subset V^{-1}(V_\infty)$. In particular,

$$\forall y \in \omega(u), \ \forall t \geq 0, \ V(\Phi(y, t)) = V(y) = V_\infty \tag{18}$$

by the invariance of $\omega(u)$.

From ODE (17), we have that any $y \in \omega(u)$ is of the form $y = (z, 0)$ since $s(t) \to 0$. As a consequence, $\Phi(y, \cdot)$ is a solution to the autonomous ODE

$$\begin{cases} \dot{\mathsf{v}}(t) & = p_\infty S(\mathsf{x}(t)) - q_\infty \mathsf{v}(t) \\ \dot{\mathsf{m}}(t) & = h_\infty \nabla F(\mathsf{x}(t)) - r_\infty \mathsf{m}(t) \\ \dot{\mathsf{x}}(t) & = -\frac{\mathsf{m}(t)}{\sqrt{\mathsf{v}(t) + \varepsilon}} \\ \dot{\mathsf{s}}(t) & = 0 \,. \end{cases} \tag{19}$$

The three first equations can be written in a more compact form:

$$\dot{\mathsf{z}}(t) = g_\infty(\mathsf{z}(t)) \tag{20}$$

where $\mathsf{z}(t) = (\mathsf{v}(t), \mathsf{m}(t), \mathsf{x}(t))$, and

$$g_\infty(z) = \lim_{t \to \infty} g(z, t) = \begin{bmatrix} p_\infty S(x) - q_\infty v \\ h_\infty \nabla F(x) - r_\infty m \\ -m/\sqrt{v + \varepsilon} \end{bmatrix}$$

for each $z \in \mathcal{Z}_+$. Consider $y = (v, m, x, 0) \in \omega(u)$. Using Eq. (18), we obtain that $dV(\Phi(y, t))/dt = 0$, which implies that

$$\left(r_\infty - \frac{q_\infty}{4} \right) \left\| \frac{\mathsf{m}(t)}{(\mathsf{v}(t) + \varepsilon)^{\odot \frac{1}{4}}} \right\|^2 + \frac{p_\infty}{4} \langle S(\mathsf{x}(t)), \frac{\mathsf{m}(t)^{\odot 2}}{(\mathsf{v}(t) + \varepsilon)^{\odot \frac{3}{2}}} \rangle = 0$$

for all $(\mathsf{v}(t), \mathsf{m}(t), \mathsf{x}(t), 0) = \Phi(y, t)$. As a consequence, Assumption 2.4-*iv)* gives $\mathsf{m}(t) = m = 0$, and then, $\mathsf{x}(t) = x$ for some x s.t. $\nabla F(x) = 0$ using ODE (19). We now turn to showing that $\mathsf{v}(t) = v = p_\infty S(x)/q_\infty$. We have proved so far that any element $y \in \omega(u)$ is written $y = (v, 0, x, 0)$ where $\nabla F(x) = 0$. The component $\mathsf{v}(\cdot)$ of $\Phi(y, \cdot)$ is a solution to the ODE $\dot{\mathsf{v}}(t) = p_\infty S(x) - q_\infty \mathsf{v}(t)$ and is thus written

$$\mathsf{v}(t) = \frac{p_\infty S(x)}{q_\infty} + e^{-q_\infty t} \left(v - \frac{p_\infty S(x)}{q_\infty} \right). \tag{21}$$

Fixing x, let $\mathcal{S}_x$ be the section of $\omega(u)$ defined by:

$$\mathcal{S}_x \omega(u) = \left\{ y \in \omega(u) \,:\, y = (\tilde{v}, 0, x, 0) \,,\, \tilde{v} \in \mathbb{R}_+^d \right\}.$$

As $\omega(u)$ is invariant, we have $\mathcal{S}_x \omega(u) = \mathcal{S}_x \Phi(\omega(u), t)$ for all $t \geq 0$. Since the set $\{\tilde{v} \in \mathbb{R}_+^d \text{ s.t. } (\tilde{v}, 0, x, 0) \in \mathcal{S}_x \omega(u)\}$ lies in a compact, we deduce from Eq. (21) that this set is reduced to the singleton $\{p_\infty S(x)/q_\infty\}$ and in particular $v = p_\infty S(x)/q_\infty$. Therefore, the union of ω-limit sets of the semiflow Φ induced by ODE (17) coincides with the set of equilibrium points of this semiflow. The latter set itself corresponds to the set of points $(z, 0)$ s.t. $z \in \operatorname{zer} g_\infty$. It remains to notice that $\Upsilon = \operatorname{zer} g_\infty$ to finish the proof.

Remark 13. Commenting on Remark 3, the same proof works for (ODE-1$'$) by using the function $F - F_\star$ as a Lyapunov function. The corresponding limit set (as $t \to +\infty$) is then of the form

$$\{\tilde{z}_\infty = (\tilde{v}_\infty, \tilde{x}_\infty) \in \mathbb{R}_+^d \times \mathbb{R}^d \,:\, \nabla F(\tilde{x}_\infty) = 0 \,,\, \tilde{v}_\infty = p_\infty S(\tilde{x}_\infty)/q_\infty\}.$$

Similarly, if we set $\mathsf{p} = \mathsf{q} \equiv 0$ in (ODE-1) and we keep what remains in Assumption 2.4, the function $\mathsf{h}(t)(F(x) - F_\star) + \frac{1}{2}\|m\|^2$ works as a Lyapunov function, and the limit set has the form $\{(0, x) : \nabla F(x) = 0\}$.

5.2. *Proof of Th.* 2.2

The existence and the uniqueness of the solution to (ODE-N) have been shown in the literature. We refer to [13, Prop. 2.1-2.2.c)] for an identical statement of this result and [41, Th. 1, Appendix A] for a complete proof. The boundedness

of the solution follows immediately from the coercivity of F together with the fact that the function $t \mapsto F(\mathsf{x}(t)) + \frac{1}{2}\|\mathsf{m}(t)\|^2$ is nonincreasing.

Concerning the convergence statement, our proof is based on comparing the solutions of (ODE-N) to the solutions of the ODE in [23, Eq. (2.3)]. We first note that under a change of variable, a solution to (ODE-N) gives a solution to [23, Eq. (2.3)].

Lemma 5.3. Let (m, x) be a solution to (ODE-N). Define $\mathsf{y}(t) = \frac{\kappa \mathsf{m}(\kappa\sqrt{t})}{2\sqrt{t}}$, $\mathsf{u}(t) = \mathsf{x}(\kappa\sqrt{t})$, with $\kappa = \sqrt{2\alpha + 2}$ and $\beta = \frac{\kappa^2}{4}$. Then, (y, u) verifies

$$\begin{cases} \dot{\mathsf{y}}(t) &= \frac{\beta}{t}(\nabla F(\mathsf{u}(t))) - \mathsf{y}(t)) \\ \dot{\mathsf{u}}(t) &= -\mathsf{y}(t) \,. \end{cases} \tag{22}$$

Proof. By simple differentiation, we get:

$$\dot{\mathsf{y}}(t) = \frac{\beta}{t}\left[\nabla F\left(\mathsf{x}(\kappa\sqrt{t})\right) - \frac{\alpha}{\kappa\sqrt{t}}\mathsf{m}\left(\kappa\sqrt{t}\right)\right] - \frac{\kappa}{4t^{\frac{3}{2}}}\mathsf{m}\left(\kappa\sqrt{t}\right) = \frac{\beta}{t}\left(\nabla F(\mathsf{u}(t)) - \mathsf{y}(t)\right),$$
$$\dot{\mathsf{u}}(t) = -\frac{\kappa}{2\sqrt{t}}\mathsf{m}\left(\kappa\sqrt{t}\right) = -\mathsf{y}(t) \,. \qquad \square$$

Consider a solution (m, x) of (ODE-N) starting at $(m_0, x_0) \in \mathbb{R}^d \times \mathbb{R}^d$. As in Section 5.1.2, for every $t_0 > 0$, on $[t_0, +\infty)$, we have that $(\mathsf{m}, \mathsf{x}, \mathsf{s})$ is a solution to the autonomous ODE

$$\begin{cases} \dot{\mathsf{m}}(t) &= \nabla F(\mathsf{x}(t)) - \alpha\mathsf{s}(t)\mathsf{m}(t) \\ \dot{\mathsf{x}}(t) &= -\mathsf{m}(t) \\ \dot{\mathsf{s}}(t) &= -\mathsf{s}(t)^2 \,, \end{cases} \tag{23}$$

starting at $(m_0, x_0, 1/t_0)$. Denote by $\Phi_N = (\Phi_N^m, \Phi_N^x, \Phi_N^s)$ the semiflow induced by ODE (23) and $\omega_N((m_0, x_0, 1/t_0))$ its limit set.

Define (y, u) as in Lem. 5.3. Starting at $(\mathsf{y}(t_0), \mathsf{u}(t_0), 1/t_0)$, we also have that $(\mathsf{y}, \mathsf{u}, \mathsf{s})$ is a solution on $[t_0, +\infty)$ to the "autonomized" Heavy-Ball ODE

$$\begin{cases} \dot{\mathsf{y}}(t) &= \beta\mathsf{s}(t)(\nabla F(\mathsf{u}(t))) - \mathsf{y}(t) \\ \dot{\mathsf{u}}(t) &= -\mathsf{y}(t) \\ \dot{\mathsf{s}}(t) &= -\mathsf{s}(t)^2 \,. \end{cases} \tag{24}$$

Denote by $\Phi_H = (\Phi_H^y, \Phi_H^u, \Phi_H^s)$ the semiflow induced by ODE (24) and $\omega_H((\mathsf{y}(t_0), \mathsf{u}(t_0), 1/t_0))$ its limit set.

Lemma 5.4. For any compact set $K \subset \mathbb{R}^{2d+1}$ and any $T > 0$, the family of functions $\left\{\Phi(z, \cdot) : [0, T] \to \mathbb{R}^{2d+1}\right\}_{z \in K}$, where Φ is either Φ_H or Φ_N, is relatively compact in $(\mathcal{C}^0([0, T], \mathbb{R}^{2d+1}), \|\cdot\|_\infty)$.

Proof. The map $\Phi : \mathbb{R}^{2d+1} \times \mathbb{R}_+ \to \mathbb{R}^{2d+1}$ is continuous, hence uniformly continuous on $K \times [0, T]$. The result follows from the application of the Arzelà-Ascoli theorem to the family $\left\{\Phi(z, \cdot) : [0, T] \to \mathbb{R}^{2d+1}\right\}_{z \in K}$. $\square$

Let $(m, x, 0) \in \omega_N((m_0, x_0, 1/t_0))$. There exists a sequence (t_k) of nonnegative reals such that $(m, x, 0) = \lim_{k \to \infty}(\mathsf{m}(t_k), \mathsf{x}(t_k), 1/t_k)$. For any $T > 0$, using Lem. 5.4, up to an extraction, we can say that the sequence of functions $\{\Phi_N((\mathsf{m}(t_k), \mathsf{x}(t_k), 1/t_k), \cdot)\}_k$ converges towards $(\tilde{\mathsf{m}}, \tilde{\mathsf{x}}, 0)$ in $\mathcal{C}^0([0, T], \mathbb{R}^d)$, where $(\tilde{\mathsf{m}}, \tilde{\mathsf{x}})$ is a solution to

$$\begin{cases} \dot{\tilde{\mathsf{m}}}(t) & = \nabla F(\tilde{\mathsf{x}}(t)) \\ \dot{\tilde{\mathsf{x}}}(t) & = -\tilde{\mathsf{m}}(t), \end{cases} \tag{25}$$

with $(\tilde{\mathsf{m}}(0), \tilde{\mathsf{x}}(0)) = (m, x)$. Moreover, by Lem. 5.3, we also have that:

$$\sup_{h \in [0, T^2/\kappa^2]} \left\| \tilde{\mathsf{x}}(\kappa\sqrt{h}) - \Phi_N^x((\mathsf{m}(t_k), \mathsf{x}(t_k), 1/t_k), \kappa\sqrt{h}) \right\|$$

$$= \sup_{h \in [0, T^2/\kappa^2]} \left\| \tilde{\mathsf{x}}(\kappa\sqrt{h}) - \Phi_H^u((\mathsf{m}(t_k), \mathsf{x}(t_k), 1/t_k), h) \right\| \xrightarrow[k \to +\infty]{} 0. \tag{26}$$

Using Lem. 5.4, up to an additional extraction, we get on $\mathcal{C}^0([0, T^2/\kappa^2], \mathbb{R}^{2d+1})$ that $\{\Phi_H((\mathsf{x}(t_k), \mathsf{m}(t_k), 1/t_k), \cdot)\}_k$ converges to $(\mathsf{u}, \mathsf{y}, 0)$, where (u, y) is a solution to

$$\begin{cases} \dot{\mathsf{y}}(t) & = 0 \\ \dot{\mathsf{u}}(t) & = -\mathsf{y}(t). \end{cases} \tag{27}$$

Therefore, $\mathsf{u}(t) = A + Bt$ for some A and B in $\mathbb{R}^d$. Imagine that $B \neq 0$. We previously proved that x (and therefore u) is bounded by some constant $C > 0$. Let $T' > \frac{C + \|A\|}{\|B\|}$. Up to an extraction, we obtain that $\{\Phi_H((\mathsf{x}(t_k), \mathsf{m}(t_k), 1/t_k), \cdot)\}_k$ converges to u' on $\mathcal{C}^0([0, T'], \mathbb{R}^{2d+1})$, with $\mathsf{u}'(t) = A' + B't$ for some A' and B' in $\mathbb{R}^d$. We then have by uniqueness of the limit that $A' = A$ and $B' = B$. As a consequence, $\|\mathsf{u}'(T')\| = \|A + BT'\| > C$ and we obtain a contradiction. Hence $B = 0$.

This implies that u is constant. Then, if we go back to Eqs. (26) and (25), we get that $\tilde{\mathsf{x}}$ is constant, hence $\tilde{\mathsf{m}} \equiv 0$ and then $\nabla F(\tilde{\mathsf{x}}) \equiv 0$. In particular, this means that $m = \tilde{\mathsf{m}}(0) = 0$ and $\nabla F(x) = \nabla F(\tilde{\mathsf{x}}(0)) = 0$.

6. Proofs for Section 3

6.1. Preliminaries

We first recall some useful definitions and results. Let Ψ represent any semiflow on an arbitrary metric space (E, d). As in the previous section, a point $z \in E$ is called an equilibrium point of the semiflow Ψ if $\Psi(z, t) = z$ for all $t \geq 0$. We denote by Λ_Ψ the set of equilibrium points of Ψ. A continuous function $\mathsf{V} : E \to \mathbb{R}$ is called a Lyapunov function for the semiflow Ψ if $\mathsf{V}(\Psi(z, t)) \leq \mathsf{V}(z)$ for all $z \in E$ and all $t \geq 0$. It is called a *strict* Lyapunov function if, moreover, $\{z \in E : \forall t \geq 0, \mathsf{V}(\Psi(z, t)) = \mathsf{V}(z)\} = \Lambda_\Psi$. If V is a strict Lyapunov function for Ψ and if $z \in E$ is a point s.t. $\{\Psi(z, t) : t \geq 0\}$ is relatively compact, then it holds that $\Lambda_\Psi \neq \varnothing$ and $\mathsf{d}(\Psi(z, t), \Lambda_\Psi) \to 0$, see [24, Th. 2.1.7]. A continuous

function $z : [0, +\infty) \to E$ is said to be an asymptotic pseudotrajectory (APT, [11]) for the semiflow Ψ if $\lim_{t \to +\infty} \sup_{s \in [0,T]} \mathsf{d}(z(t+s), \Psi(z(t), s)) = 0$ for every $T \in (0, +\infty)$.

6.2. Proof of Th. 3.1

Recall that Φ is the semiflow induced by the autonomous ODE (17) which is an "autonomized" version of our initial (ODE-1). In the remainder of this section, the proof will be divided into two main steps: (a) we show that a certain continuous-time linearly interpolated process constructed from the iterates of our algorithm 1 is an APT of Φ; (b) we exhibit a strict Lyapunov function for a restriction to a carefully chosen compact set of a well chosen semiflow related to Φ. Then, we characterize the limit set of the APT using [10, Th. 5.7] and [10, Prop. 6.4]. The sequence (z_n) converges almost surely to this same limit set.

(a) APT For every $n \geq 1$, define $\bar{z}_n = (v_n, m_n, x_{n-1})$ (note the shift in the index of the variable x). We have the decomposition

$$\bar{z}_{n+1} = \bar{z}_n + \gamma_{n+1} g(\bar{z}_n, \tau_n) + \gamma_{n+1} \eta_{n+1} + \gamma_{n+1} \varsigma_{n+1},$$

where g is defined in Eq. (1),

$$\eta_{n+1} = \big(p_n(\nabla f(x_n, \xi_{n+1})^{\odot 2} - S(x_n)), \ h_n(\nabla f(x_n, \xi_{n+1}) - \nabla F(x_n)), \ 0 \big), \quad (28)$$

is a martingale increment and where we set $\varsigma_{n+1} = (\varsigma_{n+1}^v, \varsigma_{n+1}^m, \varsigma_{n+1}^x)$ with the components defined by:

$$\begin{cases} \varsigma_{n+1}^v &= p_n(S(x_n) - S(x_{n-1})) \\ \varsigma_{n+1}^m &= h_n(\nabla F(x_n) - \nabla F(x_{n-1})) \\ \varsigma_{n+1}^x &= (\frac{\gamma_n}{\gamma_{n+1}} - 1) \frac{m_n}{\sqrt{v_n + \varepsilon}}. \end{cases}$$

We first prove that $\varsigma_n \to 0$ a.s. by considering the components separately. The components ς_{n+1}^m and ς_{n+1}^v converge a.s. to zero by using Assumptions 2.1, 2.3, together with the boundedness of the sequences (p_n) and (h_n) (which are both convergent). Indeed, since ∇F is locally Lipschitz continuous and the sequence (z_n) is supposed to be almost surely bounded, there exists a constant C s.t. $\|\nabla F(x_n) - \nabla F(x_{n-1})\| \leq C\|x_n - x_{n-1}\| \leq \frac{C}{\varepsilon}\gamma_n\|m_n\|$. The same inequality holds when replacing ∇F by S which is also locally Lipschitz continuous. The component ς_{n+1}^x also converges a.s. to zero by observing that $\|\varsigma_{n+1}^x\| \leq |1 - \frac{\gamma_n}{\gamma_{n+1}}|.\|m_n\|/\sqrt{\varepsilon}$ and using Assumption 3.2 together with the a.s. boundedness of (z_n). Now consider the martingale increment sequence (η_n), adapted to $\mathcal{F}_n$. Take $\delta > 0$. Since (z_n) is a.s bounded, there is a constant $C' > 0$ such that $\mathbb{P}(\sup \|x_n\| > C') \leq \delta$. Denoting $\tilde{\eta}_n \triangleq \eta_n \mathbb{1}_{\|x_n\| \leq C'}$ and combining Assumptions 2.4 with 3.4-i) we can show using convexity inequalities that

$$\sup_n \mathbb{E}\|\tilde{\eta}_{n+1}\|^q < \infty.$$

Then, we deduce from this result together with the corresponding stepsize assumption from 3.4-i) and [10, Prop. 4.2] (see also [34, Prop. 8]) the key property:

$$\forall T > 0, \quad \max\left\{\left\|\sum_{k=n}^{L-1} \gamma_{k+1}\tilde{\eta}_{k+1}\right\| \;:\; L = n+1, \ldots, J(\tau_n + T)\right\} \xrightarrow[n\to\infty]{\text{a.s.}} 0 \quad (29)$$

where $J(t) = \max\{n \geq 0 \;:\; \tau_n \leq t\}$. Hence, for all $T > 0$, with probability at least $1 - \delta$:

$$\max\left\{\left\|\sum_{k=n}^{L-1} \gamma_{k+1}\eta_{k+1}\right\| \;:\; L = n+1, \ldots, J(\tau_n + T)\right\} \xrightarrow[n\to\infty]{} 0. \quad (30)$$

Since δ can be chosen arbitrary small, Eq. (30) remains true with probability 1. This result also holds under Assumption 3.4-ii) (instead of 3.4-i)) by applying [10, Prop. 4.4].

Let $\boldsymbol{z} : [0, +\infty) \to \mathcal{Z}_+$ be the continous-time linearly interpolated process given by

$$\boldsymbol{z}(t) = \bar{z}_n + (t - \tau_n)\frac{\bar{z}_{n+1} - \bar{z}_n}{\gamma_{n+1}} \qquad (\forall n \in \mathbb{N}, \, \forall t \in [\tau_n, \tau_{n+1}))$$

(where $\tau_n = \sum_{k=1}^{n} \gamma_k$). Let $t_0 > 0$. Define $\boldsymbol{u} : [t_0, \infty) \to \mathcal{Z} \times (0, 1/t_0]$ by

$$\boldsymbol{u}(t) = \begin{bmatrix} \boldsymbol{z}(t) \\ 1/t \end{bmatrix}, \quad \text{for} \quad t \geq t_0 > 0.$$

Using Eq. (30) and the almost sure boundedness of the sequence (z_n) along with the fact that ς_n converges a.s. to zero, it follows from [10, Prop. 4.1, Remark 4.5] that $\boldsymbol{u}(t)$ is an APT of the already defined semiflow Φ induced by (17). Remark that it also holds that $\boldsymbol{z}(t)$ is an APT of the semiflow Φ^∞ induced by (20). As the trajectory of $\boldsymbol{u}(t)$ is precompact, the limit set

$$\boldsymbol{L}(\boldsymbol{u}) = \bigcap_{t \geq t_0} \overline{\boldsymbol{u}([t, \infty))}$$

is compact. Moreover, it has the form

$$\boldsymbol{L}(\boldsymbol{u}) = \begin{bmatrix} \boldsymbol{S} \\ 0 \end{bmatrix}, \quad \text{where} \quad \boldsymbol{S} \triangleq \bigcap_{t \geq t_0} \overline{\boldsymbol{z}([t, \infty))}. \quad (31)$$

Our objective now is to prove that

$$\boldsymbol{S} \subset \Lambda_{\Phi^\infty}. \quad (32)$$

In order to establish this inclusion, we study the behavior of the restriction $\Phi|\boldsymbol{L}$ of the semiflow Φ to the set $\boldsymbol{L}$ (which is well-defined since $\boldsymbol{L}$ is Φ-invariant). Remark that

$$\Phi|\boldsymbol{L} = \begin{bmatrix} \Phi^\infty|\boldsymbol{S} \\ 0 \end{bmatrix},$$

where Φ^∞ is the semiflow associated to (20). In the second part of the proof, we establish Eq. (32) combining item (a) we just proved with [10, Th. 5.7] and [10, Prop. 6.4]. In order to use the latter proposition, we prove a useful proposition in item (b).

(b) Strict Lyapunov function and convergence For every $\delta > 0$ and every $z = (v, m, x) \in \mathcal{Z}_+$, define:

$$W_\delta(v, m, x) \triangleq \mathcal{E}_\infty(z) - \delta\langle\nabla F(x), m\rangle + \delta\|q_\infty v - p_\infty S(x)\|^2, \qquad (33)$$

where, under Assumption 2.4-*i)*, the function $\mathcal{E}_\infty$ is defined by

$$\mathcal{E}_\infty(z) \triangleq \lim_{t\to+\infty}\mathcal{E}(t, z) = h_\infty(F(x) - F_\star) + \frac{1}{2}\left\|\frac{m}{(v+\varepsilon)^{\odot\frac{1}{4}}}\right\|^2. \qquad (34)$$

Proposition 6.1. Let $t_0 > 0$ and let Assumptions 2.1 to 2.4 and 3.5 hold true. Let $\boldsymbol{S}$ be the limit set defined in Eq. (31). Let $\overline{\Phi}^\infty : \boldsymbol{S} \times [t_0, +\infty) \to \boldsymbol{S}$ be the restriction of the semiflow Φ^∞ to $\boldsymbol{S}$ *i.e.*, $\overline{\Phi}^\infty(z, t) = \Phi^\infty(z, t)$ for all $z \in \boldsymbol{S}, t \geq t_0$. Then,

- *i)* $\boldsymbol{S}$ is compact.
- *ii)* $\overline{\Phi}^\infty$ is a well-defined semiflow on $\boldsymbol{S}$.
- *iii)* The set of equilibrium points of $\overline{\Phi}^\infty$ is equal to $\Lambda_{\Phi^\infty} \cap \boldsymbol{S}$.
- *iv)* There exists $\delta > 0$ s.t. W_δ is a strict Lyapunov function for the semiflow $\overline{\Phi}^\infty$.

Proof. The first point is a consequence of the definition of $\boldsymbol{S}$ and the boundedness of $\boldsymbol{z}$. The second point stems from the definition of Φ^∞. Observing that $\overline{\Phi}^\infty$ is valued in $\boldsymbol{S}$, the third point is immediate from the definition of Λ_{Φ^∞}. We now prove the last point. Consider $z \in \boldsymbol{S}$ and write $\overline{\Phi}^\infty(z, t)$ under the form $\overline{\Phi}^\infty(z, t) = (\mathsf{v}(t), \mathsf{m}(t), \mathsf{x}(t))$. Notice that this quantity is bounded as a function of the variable t. For *any* map $\mathsf{W} : \mathcal{Z}_+ \to \mathbb{R}$, define for all $t \geq t_0$, $\mathcal{L}_\mathsf{W}(t) \triangleq \limsup_{s\to 0} s^{-1}(\mathsf{W}(\overline{\Phi}^\infty(z, t+s)) - \mathsf{W}(\overline{\Phi}^\infty(z, t)))$. Introduce $G(z) \triangleq -\langle\nabla F(x), m\rangle$ and $H(z) \triangleq \|q_\infty v - p_\infty S(x)\|^2$ for every $z = (v, m, x) \in \mathcal{Z}_+$. Consider $\delta > 0$ (to be specified later on). We study $\mathcal{L}_{W_\delta} = \mathcal{L}_{\mathcal{E}_\infty} + \delta\mathcal{L}_G + \delta\mathcal{L}_H$. Note that $\overline{\Phi}^\infty(z, t) \in \boldsymbol{S} \cap \mathcal{Z}_+$ for all $t \geq t_0$ by an analogous result to Lem. 5.1 for Φ^∞. Thus, $t \mapsto \mathcal{E}_\infty(\overline{\Phi}^\infty(z, t))$ is differentiable at any point $t \geq t_0$ and $\mathcal{L}_{\mathcal{E}_\infty}(t) = \frac{d}{dt}\mathcal{E}_\infty(\overline{\Phi}^\infty(z, t))$. Using similar derivations to Ineq. (16), we obtain that

$$\mathcal{L}_{\mathcal{E}_\infty}(t) \leq -\left(r_\infty - \frac{q_\infty}{4}\right)\left\|\frac{\mathsf{m}(t)}{(\mathsf{v}(t)+\varepsilon)^{\odot\frac{1}{4}}}\right\|^2. \qquad (35)$$

We now study $\mathcal{L}_G$. For every $t \geq t_0$,

$$\mathcal{L}_G(t) = \limsup_{s\to 0} s^{-1}(-\langle\nabla F(\mathsf{x}(t+s)), \mathsf{m}(t+s)\rangle + \langle\nabla F(\mathsf{x}(t)), \mathsf{m}(t)\rangle)$$

$$\leq \limsup_{s\to 0} s^{-1}\|\nabla F(\mathsf{x}(t)) - \nabla F(\mathsf{x}(t+s))\|\|\mathsf{m}(t+s)\| - \langle\nabla F(\mathsf{x}(t)), \dot{\mathsf{m}}(t)\rangle.$$

Let $L_{\nabla F}$ be the Lipschitz constant of ∇F on the bounded set $\{x : (v, m, x) \in \boldsymbol{S}\}$. Define $C_1 \triangleq \sup_t \|\sqrt{v(t) + \varepsilon}\|$. Then,

$$
\begin{aligned}
\mathcal{L}_G(t) &\leq L_{\nabla F} \limsup_{s \to 0} s^{-1} \|x(t) - x(t+s)\| \|m(t+s)\| - \langle \nabla F(x(t)), \dot{m}(t) \rangle \\
&\leq L_{\nabla F} \|\dot{x}(t)\| \|m(t)\| - \langle \nabla F(x(t)), \dot{m}(t) \rangle \\
&\leq L_{\nabla F} \|\dot{x}(t)\| \|m(t)\| - h_\infty \|\nabla F(x(t))\|^2 + r_\infty \langle \nabla F(x(t)), m(t) \rangle \\
&\leq \left(\frac{L_{\nabla F} C_1^{\frac{1}{2}}}{\varepsilon^{\frac{1}{4}}} + \frac{r_\infty C_1}{2u_1^2} \right) \left\| \frac{m(t)}{(v(t) + \varepsilon)^{\odot \frac{1}{4}}} \right\|^2 - \left(h_\infty - \frac{r_\infty u_1^2}{2} \right) \|\nabla F(x(t))\|^2
\end{aligned}
\tag{36}
$$

where we used the classical inequality $|\langle a, b \rangle| \leq \|a\|^2/(2u^2) + u^2 \|b\|^2/2$ for any non-zero real u to derive the last above inequality. We now study $\mathcal{L}_H$. For every $t \geq t_0$,

$$
\begin{aligned}
\mathcal{L}_H(t) &= \limsup_{s \to 0} s^{-1} (\|q_\infty v(t+s) - p_\infty S(x(t+s))\|^2 - \|q_\infty v(t) - p_\infty S(x(t))\|^2) \\
&= \limsup_{s \to 0} s^{-1} (p_\infty^2 \|S(x(t)) - S(x(t+s))\|^2 \\
&\quad + 2p_\infty \langle S(x(t)) - S(x(t+s)), q_\infty v(t+s) - p_\infty S(x(t)) \rangle \\
&\quad + \lim_{s \to 0} s^{-1} (\|q_\infty v(t+s) - p_\infty S(x(t))\|^2 - \|q_\infty v(t) - p_\infty S(x(t))\|^2).
\end{aligned}
$$

The second term in the righthand side coincides with

$$
-2q_\infty \langle p_\infty S(x(t)) - q_\infty v(t), \dot{v}(t) \rangle = -2q_\infty \|p_\infty S(x(t)) - q_\infty v(t)\|^2.
$$

Denote by L_S the Lipschitz constant of S on the set $\{x : (v, m, x) \in \boldsymbol{S}\}$. Note that $s^{-1}(\|S(x(t+s)) - S(x(t))\|^2) \leq L_S^2 s \|s^{-1}(x(t+s) - x(t))\|^2$ which converges to zero as $s \to 0$. Thus,

$$
\begin{aligned}
\mathcal{L}_H(t) &= -2q_\infty \|p_\infty S(x(t)) - q_\infty v(t)\|^2 \\
&\quad + \limsup_{s \to 0} 2p_\infty s^{-1} \langle S(x(t)) - S(x(t+s)), q_\infty v(t+s) - p_\infty S(x(t)) \rangle \\
&\leq -2q_\infty \|p_\infty S(x(t)) - q_\infty v(t)\|^2 + 2p_\infty \|\dot{x}(t)\| L_S \|q_\infty v(t) - p_\infty S(x(t))\| \\
&\leq \frac{p_\infty}{\varepsilon^{\frac{1}{2}} u_2^2} \left\| \frac{m(t)}{(v(t) + \varepsilon)^{\odot \frac{1}{4}}} \right\|^2 - (2q_\infty - p_\infty u_2^2 L_S^2) \|p_\infty S(x(t)) - q_\infty v(t)\|^2.
\end{aligned}
\tag{37}
$$

Recalling that $\mathcal{L}_{W_\delta} = \mathcal{L}_{\mathcal{E}_\infty} + \delta \mathcal{L}_G + \delta \mathcal{L}_H$ and combining Eqs. (35), (36) and (37), we obtain for every $t \geq t_0$,

$$
\begin{aligned}
\mathcal{L}_{W_\delta}(t) &\leq -M(\delta) \left\| \frac{m(t)}{(v(t) + \varepsilon)^{\odot \frac{1}{4}}} \right\|^2 - \delta \left(h_\infty - \frac{r_\infty u_1^2}{2} \right) \|\nabla F(x(t))\|^2 \\
&\quad - \delta \left(2q_\infty - p_\infty u_2^2 L_S^2 \right) \|p_\infty S(x(t)) - q_\infty v(t)\|^2. \quad (38)
\end{aligned}
$$

where $M(\delta) \triangleq r_\infty - \frac{q_\infty}{4} - \delta \left(\frac{r_\infty C_1}{2u_1^2} + \frac{L_{\nabla F} C_1^{\frac{1}{2}}}{\varepsilon^{\frac{1}{4}}} + \frac{p_\infty}{\varepsilon^{\frac{1}{2}} u_2^2} \right)$. Now select u_1, u_2 small enough s.t. $h_\infty - r_\infty u_1^2/2 > 0$ and $2q_\infty - p_\infty u_2^2 L_S^2 > 0$. Then, choose δ in such a way that $M(\delta) > 0$. Thus, there exists a constant c depending on δ s.t. for every $t \geq t_0$,

$$\mathcal{L}_{W_\delta}(t) \leq -c \left(\left\| \frac{\mathsf{m}(t)}{(\mathsf{v}(t) + \varepsilon)^{\odot \frac{1}{4}}} \right\|^2 + \|\nabla F(\mathsf{x}(t))\|^2 + \|p_\infty S(\mathsf{x}(t)) - q_\infty \mathsf{v}(t)\|^2 \right). \tag{39}$$

It can easily be seen that for every $z \in \boldsymbol{S}$, $t \mapsto W_\delta(\overline{\Phi}^\infty(z,t))$ is Lipschitz continuous, hence absolutely continuous. Its derivative almost everywhere coincides with $\mathcal{L}_{W_\delta}$, which is nonpositive. Thus, W_δ is a Lyapunov function for $\overline{\Phi}^\infty$. We prove that the Lyapunov function is strict. Consider $z = (v, m, x) \in \boldsymbol{S}$ s.t. $W_\delta(\overline{\Phi}^\infty(z,t)) = W_\delta(z)$ for all $t \geq t_0$. The derivative almost everywhere of $t \mapsto W_\delta(\overline{\Phi}^\infty(z,t))$ is identically zero, and by Eq. (39), this implies that

$$-c \left(\left\| \frac{\mathsf{m}(t)}{(\mathsf{v}(t) + \varepsilon)^{\odot \frac{1}{4}}} \right\|^2 + \|\nabla F(\mathsf{x}(t))\|^2 + \|p_\infty S(\mathsf{x}(t)) - q_\infty \mathsf{v}(t)\|^2 \right)$$

is equal to zero for every $t \geq t_0$ a.e. (hence, for every $t \geq t_0$, by continuity of $\overline{\Phi}^\infty$). In particular for $t = t_0$, $m = \nabla F(x) = 0$ and $p_\infty S(x) - q_\infty v = 0$. Hence, $z \in \operatorname{zer} g_\infty \cap \boldsymbol{S}$. This concludes the proof since $\Lambda_{\Phi^\infty} = \operatorname{zer} g_\infty$. $\qquad\square$

End of the Proof of Th. 3.1 Finally, Assumption 3.5 implies that the set $W_\delta(\Lambda_{\Phi^\infty} \cap \boldsymbol{S})$ is of empty interior. Recall that Assumptions 2.1 and 2.3 both follow from Assumption 3.3 made in Th. 3.1. Given Prop. 6.1, the proof is concluded by applying [10, Prop. 6.4] to the restricted semiflow $\bar{\Phi}^\infty$ (with $(M, \Lambda) = (\boldsymbol{S}, \Lambda_{\bar{\Phi}^\infty})$). Note that a Lyapunov function for $\Lambda_{\bar{\Phi}^\infty}$ is what is called a strict Lyapunov function. Such a function is provided by Prop. 6.1. We obtain as a conclusion of [10, Prop. 6.4] that $\boldsymbol{S} \subset \Lambda_{\bar{\Phi}^\infty}$. This gives the desired result (Eq. (32)) given Prop. 6.1-*iii)*.

The last assertion of Th. 3.1 is a consequence of [10, Cor. 6.6].

6.3. Proof of Th. 3.3

We can rewrite the iterates from Algorithm 2 as follows:

$$\begin{cases} m_{n+1} & = m_n + \gamma_{n+1}(\nabla F(x_n) - \frac{\alpha}{\tau_n} m_n) + \gamma_{n+1}(\nabla f(x_n, \xi_{n+1}) - \nabla F(x_n)) \\ x_{n+1} & = x_n - \gamma_{n+1} m_{n+1} \,. \end{cases} \tag{40}$$

We prove that the sequence $(y_n = (m_n, x_n) : n \in \mathbb{N})$ of iterates of this algorithm converges almost surely towards the set $\bar{\Upsilon}$ defined in Eq. (3) if it is supposed to be bounded with probability one. The proof follows a similar path to the proof in Section 5.2.

Indeed, denote by X and M the linearly interpolated processes constructed respectively from the sequences (x_n) and (m_n) and let $\mathsf{s}(t) = 1/t$. Recall that $\Phi_N = (\Phi_N^m, \Phi_N^x, \Phi_N^s)$ is the semiflow induced by (23). As in Section 6.2, we have that $\mathsf{Z} \triangleq (\mathsf{M}, \mathsf{X}, \mathsf{s})$ is an APT of (23). In particular, this means that

$$\forall T > 0, \quad \sup_{h \in [0,T]} \|\mathsf{X}(t+h) - \Phi_N^x(\mathsf{Z}(t), h)\| \xrightarrow[t \to \infty]{} 0. \tag{41}$$

By Lem. 5.3, we also have that

$$\sup_{h \in [0, T^2/\kappa^2]} \left\| \mathsf{X}(t + \kappa\sqrt{h}) - \Phi_N^x(\mathsf{Z}(t), \kappa\sqrt{h}) \right\|$$

$$= \sup_{h \in [0, T^2/\kappa^2]} \left\| \mathsf{X}(t + \kappa\sqrt{h}) - \Phi_H^u(\mathsf{Z}(t), h) \right\| \xrightarrow[t \to \infty]{} 0. \tag{42}$$

Let (m, x) be a limit point of the sequence (y_n) and let $T > 0$. Using Lem. 5.4, we can proceed in the same manner as in Section 5.2 and get a sequence (t_k) such that

$$(\mathsf{M}(t_k + \cdot), \mathsf{X}(t_k + \cdot)) \to (\mathsf{m}, \mathsf{x}) \text{ and } (\Phi_H^y(\mathsf{Z}(t_k), \cdot), \Phi_H^u(\mathsf{Z}(t_k), \cdot)) \to (\mathsf{y}, \mathsf{u}),$$

where $(\mathsf{m}(0), \mathsf{x}(0)) = (m, x)$, and (m, x) and (x, u) are respectively solutions to (25) and (27). As in the end of Section 5.2, we obtain that u and x are constant, therefore $\mathsf{m} \equiv 0$ and $\nabla F(\mathsf{x}) \equiv 0$, which finishes the proof.

6.4. Proof of Th. 3.2

The idea of the proof is to apply Robbins-Siegmund's theorem [40] to

$$V_n = h_{n-1} F(x_n) + \frac{1}{2} \langle m_n^{\odot 2}, \frac{1}{\sqrt{v_n + \varepsilon}} \rangle$$

(note the similarity of V_n with the energy function (15)). Since $\inf F > -\infty$, we assume without loss of generality that $F \geq 0$. In this subsection, we use the notation ∇f_{n+1} as a shorthand notation for $\nabla f(x_n, \xi_{n+1})$ and C denotes some positive constant which may change from line to line. We write $\mathbb{E}_n = \mathbb{E}[\cdot \,|\, \mathscr{F}_n]$ for the conditional expectation w.r.t the σ-algebra $\mathscr{F}_n$. Define $P_n \triangleq \frac{1}{2} \langle D_n, m_n^{\odot 2} \rangle$, with $D_n \triangleq \frac{1}{\sqrt{v_n + \varepsilon}}$. We have the decomposition:

$$P_{n+1} - P_n = \frac{1}{2} \langle D_{n+1} - D_n, m_{n+1}^{\odot 2} \rangle + \frac{1}{2} \langle D_n, m_{n+1}^{\odot 2} - m_n^{\odot 2} \rangle. \tag{43}$$

We estimate the vector

$$D_{n+1} - D_n = \frac{\sqrt{v_n + \varepsilon} - \sqrt{v_{n+1} + \varepsilon}}{\sqrt{v_{n+1} + \varepsilon} \odot \sqrt{v_n + \varepsilon}}.$$

Remarking that $v_{n+1} \geq (1 - \gamma_{n+1} q_n) v_n$ and using the update rule of v_n, we obtain for a sufficiently large n that

$$\sqrt{v_n + \varepsilon} - \sqrt{v_{n+1} + \varepsilon} = \gamma_{n+1} \frac{q_n v_n - p_n \nabla f_{n+1}^{\odot 2}}{\sqrt{v_n + \varepsilon} + \sqrt{v_{n+1} + \varepsilon}}$$

$$\leq \gamma_{n+1} q_n \frac{v_n}{(1 + \sqrt{1 - \gamma_{n+1} q_n}) \sqrt{v_n + \varepsilon}}$$

$$= \frac{\gamma_{n+1} q_n}{1 + \sqrt{1 - \gamma_{n+1} q_n}} \sqrt{v_n} \odot \frac{\sqrt{v_n}}{\sqrt{v_n + \varepsilon}}$$

$$\leq c_{n+1} \sqrt{v_{n+1}}, \tag{44}$$

where $c_{n+1} \triangleq \frac{\gamma_{n+1} q_n}{\sqrt{1 - \gamma_{n+1} q_n}(1 + \sqrt{1 - \gamma_{n+1} q_n})}$. It is easy to see that $c_{n+1}/\gamma_n \to q_\infty/2$. Thus, for any $\delta > 0$, $c_{n+1} \leq (q_\infty + 2\delta)\gamma_n/2$ for all n large enough. Using also that $\sqrt{v_{n+1}}/\sqrt{v_{n+1} + \varepsilon} \leq 1$, we obtain

$$D_{n+1} - D_n \leq \frac{q_\infty + 2\delta}{2} \gamma_n D_n. \tag{45}$$

Substituting the above inequality in Eq. (43), we obtain

$$P_{n+1} - P_n \leq \left(\frac{q_\infty + 2\delta}{2} \right) \frac{\gamma_n}{2} \langle D_n, m_{n+1}^{\odot 2} \rangle + \frac{1}{2} \langle D_n, m_{n+1}^{\odot 2} - m_n^{\odot 2} \rangle$$

$$\leq \frac{q_\infty + 2\delta}{2} \gamma_n P_n + \left(1 + \frac{q_\infty + 2\delta}{2} \gamma_n \right) \frac{1}{2} \langle D_n, m_{n+1}^{\odot 2} - m_n^{\odot 2} \rangle.$$

Using $m_{n+1}^{\odot 2} - m_n^{\odot 2} = 2m_n \odot (m_{n+1} - m_n) + (m_{n+1} - m_n)^{\odot 2}$, and noting that $\mathbb{E}_n(m_{n+1} - m_n) = \gamma_{n+1} h_n \nabla F(x_n) - \gamma_{n+1} r_n m_n$,

$$\mathbb{E}_n \frac{1}{2} \langle D_n, m_{n+1}^{\odot 2} - m_n^{\odot 2} \rangle = \gamma_{n+1} h_n \langle \nabla F(x_n), \frac{m_n}{\sqrt{v_n + \varepsilon}} \rangle - 2\gamma_{n+1} r_n P_n$$

$$+ \frac{1}{2} \langle D_n, \mathbb{E}_n[(m_{n+1} - m_n)^{\odot 2}] \rangle.$$

There exists $\delta > 0$ such that $r_\infty - \frac{q_\infty}{4} - \frac{\delta}{2} > 0$ by Assumption 2.4-*iv)*. As $\frac{\gamma_{n+1}}{\gamma_n} r_n - \frac{q_\infty}{4} \to r_\infty - \frac{q_\infty}{4}$, for all n large enough, $\frac{\gamma_{n+1}}{\gamma_n} r_n - \frac{q_\infty}{4} > r_\infty - \frac{q_\infty}{4} - \frac{\delta}{2} > 0$. Hence, for all n large enough,

$$\mathbb{E}_n P_{n+1} - P_n \leq -2 \left(r_\infty - \frac{q_\infty}{4} - \frac{\delta}{2} \right) \gamma_n P_n + \gamma_{n+1} h_n \langle \nabla F(x_n), \frac{m_n}{\sqrt{v_n + \varepsilon}} \rangle$$

$$+ C \gamma_n^2 \langle \nabla F(x_n), \frac{m_n}{\sqrt{v_n + \varepsilon}} \rangle + C \langle D_n, \mathbb{E}_n[(m_{n+1} - m_n)^{\odot 2}] \rangle. \tag{46}$$

Using the inequality $\langle u, v \rangle \leq (\|u\|^2 + \|v\|^2)/2$ and Assumption 3.6-*ii)*, it is easy to show the inequality $\langle \nabla F(x_n), \frac{m_n}{\sqrt{v_n + \varepsilon}} \rangle \leq C(1 + F(x_n) + P_n)$. Moreover, using the componentwise inequality $(h_n \nabla f_{n+1} - r_n m_n)^{\odot 2} \leq 2h_n^2 \nabla f_{n+1}^{\odot 2} + 2r_n^2 m_n^{\odot 2}$

A. Barakat et al.

along with Assumption 3.6-*ii)* and the boundedness of the sequences $(h_n), (r_n)$ and (γ_{n+1}/γ_n), we obtain

$$\langle D_n, \mathbb{E}_n[(m_{n+1} - m_n)^{\odot 2}]\rangle \le C\gamma_n^2(1 + F(x_n) + P_n). \qquad (47)$$

Combining Eq. (46) and Eq. (47), we get

$$\mathbb{E}_n(P_{n+1} - P_n) \le \gamma_{n+1}h_n\langle\nabla F(x_n), m_n \odot D_n\rangle + C\gamma_n^2(1 + F(x_n) + P_n). \qquad (48)$$

Denoting by M the Lipschitz coefficient of ∇F, we also have

$$F(x_{n+1}) \le F(x_n) - \gamma_{n+1}\langle\nabla F(x_n), m_{n+1} \odot D_{n+1}\rangle + \frac{\gamma_{n+1}^2 M}{2}\|m_{n+1} \odot D_{n+1}\|^2. \qquad (49)$$

Using (45) and the update rule of m_n, we have $\|m_{n+1} \odot D_{n+1} - m_n \odot D_n\|^2$

$$\begin{aligned}
&\le C\|(m_{n+1} - m_n) \odot D_n\|^2 + C\|m_{n+1} \odot (D_{n+1} - D_n)\|^2 \\
&\le C\gamma_{n+1}^2(\|\nabla f_{n+1}\|^2 + \|m_n \odot D_n\|^2) + C\gamma_{n+1}^2\|m_{n+1} \odot D_n\|^2 \qquad (50) \\
&\le C\gamma_{n+1}^2(\|m_n \odot D_n\|^2 + \|\nabla f_{n+1}\|^2).
\end{aligned}$$

Finally, recalling that $V_n = h_{n-1}F(x_n) + P_n$, (h_n) is decreasing, combining Eq. (48), (49), (50), and using Assumption 3.6, we have

$$\begin{aligned}
\mathbb{E}_n[V_{n+1}] &\le V_n + \gamma_{n+1}h_n\langle\nabla F(x_n), \mathbb{E}_n[m_n \odot D_n - m_{n+1} \odot D_{n+1}]\rangle \\
&\quad + C\gamma_{n+1}^2\left(1 + F(x_n) + P_n + \|m_n \odot D_n\|^2\right) \\
&\quad + C\gamma_{n+1}^2\mathbb{E}_n[\|m_n \odot D_n - m_{n+1} \odot D_{n+1}\|^2] \\
&\le V_n + C\gamma_n^2\left(1 + F(x_n) + P_n + \|m_n \odot D_n\|^2 + \mathbb{E}_n\left[\|\nabla f_{n+1}\|^2\right]\right) \\
&\le V_n + C\gamma_n^2(1 + F(x_n) + P_n) \\
&\le (1 + C\gamma_n^2)V_n + C\gamma_n^2,
\end{aligned}$$

where we used Cauchy-Schwarz's inequality and the fact that $\|m_n \odot D_n\|^2 \le CP_n$. By the Robbins-Siegmund's theorem [40], the sequence (V_n) converges almost surely to a finite random variable $V_\infty \in \mathbb{R}^+$. Then, the coercivity of F implies that (x_n) is almost surely bounded.

We now establish the almost sure boundedness of (m_n). Assume in the sequel that n is large enough to have $(1 - \gamma_{n+1}r_n) \ge 0$. Consider the martingale difference sequence $\Delta_{n+1} \triangleq \nabla f_{n+1} - \nabla F(x_n)$. We decompose $m_n = \bar{m}_n + \tilde{m}_n$ where $\bar{m}_{n+1} = (1 - \gamma_{n+1}r_n)\bar{m}_n + \gamma_{n+1}h_n\nabla F(x_n)$ and $\tilde{m}_{n+1} = (1 - \gamma_{n+1}r_n)\tilde{m}_n + \gamma_{n+1}h_n\Delta_{n+1}$, setting $\bar{m}_0 = 0$ and $\tilde{m}_0 = m_0$. We prove that both terms $\bar{m}_n$ and $\tilde{m}_n$ are bounded. Consider the first term: $\|\bar{m}_{n+1}\| \le (1 - \gamma_{n+1}r_n)\|\bar{m}_n\| + \gamma_{n+1}\sup_k\|h_k\nabla F(x_k)\|$, where the supremum in the above inequality is almost surely finite by continuity of ∇F. We immediately get that if $\|\bar{m}_n\| \ge \frac{\sup_k\|h_k\nabla F(x_k)\|}{r_\infty}$, then $\|\bar{m}_{n+1}\| \le \|\bar{m}_n\|$. Thus

$$\|\bar{m}_{n+1}\| \leq \frac{\sup_k \|h_k \nabla F(x_k)\|}{r_\infty} + \sup_k \gamma_{k+1} \|h_k \nabla F(x_k)\|,$$

which implies that $\bar{m}_n$ is bounded. Consider now the term $\tilde{m}_n$:

$$\mathbb{E}_n[\|\tilde{m}_{n+1}\|^2] = (1 - \gamma_{n+1} r_n)^2 \|\tilde{m}_n\|^2 + \gamma_{n+1}^2 h_n^2 \mathbb{E}_n[\|\Delta_{n+1}\|^2]$$
$$\leq \|\tilde{m}_n\|^2 + \gamma_{n+1}^2 h_n^2 \mathbb{E}_n[\|\Delta_{n+1}\|^2].$$

Then, the inequality $\mathbb{E}_n[\|\Delta_{n+1}\|^2] \leq \mathbb{E}_n[\|\nabla f_{n+1}\|^2]$ combined with Assumption 3.4-i) and the a.s. boundedness of the sequence (x_n) imply that there exists a finite random variable $C_{\mathcal{K}}$ (independent of n) s.t. $\mathbb{E}_n[\|\nabla f_{n+1}\|^2] \leq C_{\mathcal{K}}$. As a consequence, since $\sum_n \gamma_{n+1}^2 < \infty$ and the sequence (h_n) is bounded, we obtain that a.s.:

$$\sum_{n \geq 0} \gamma_{n+1}^2 h_n^2 \mathbb{E}_n[\|\Delta_{n+1}\|^2] \leq C C_{\mathcal{K}} \sum_{n \geq 0} \gamma_{n+1}^2 < +\infty.$$

Hence, we can apply the Robbins-Siegmund theorem to obtain that $\sup_n \|\tilde{m}_n\|^2 < \infty$ w.p.1. Finally, it can be shown that (v_n) is almost surely bounded using the same arguments, decomposing v_n into $\bar{v}_n + \tilde{v}_n$ as above. Indeed, first, we have:

$$\mathbb{E}_n[\|\tilde{v}_{n+1}\|^2] \leq \|\tilde{v}_n\|^2 + \gamma_{n+1}^2 p_n^2 \mathbb{E}_n[\|\nabla f_{n+1}^{\odot 2} - S(x_n)\|^2].$$

Second, it also holds that:

$$\mathbb{E}_n[\|\nabla f_{n+1}^{\odot 2} - S(x_n)\|^2] \leq \mathbb{E}_n[\|\nabla f_{n+1}^{\odot 2}\|^2] \leq \mathbb{E}_n[\|\nabla f_{n+1}\|^4].$$

Then, using Assumption 3.4-i) and the a.s. boundedness of the sequence (x_n), there exists a finite random variable $C'_{\mathcal{K}}$ (independent of n) s.t. $\mathbb{E}_n[\|\nabla f_{n+1}\|^4] \leq C'_{\mathcal{K}}$. Moreover, the sequence (p_n) is bounded and $\sum_n \gamma_{n+1}^2 < \infty$. As a consequence, it holds that a.s:

$$\sum_{n \geq 0} \gamma_{n+1}^2 p_n^2 \mathbb{E}_n[\|\nabla f_{n+1}^{\odot 2} - S(x_n)\|^2] \leq C C'_{\mathcal{K}} \sum_{n \geq 0} \gamma_{n+1}^2 < +\infty.$$

It follows that the Robbins-Siegmund theorem can be applied to the sequence $\|\tilde{v}_n\|^2$ as for the sequence $\|\tilde{m}_n\|^2$ to obtain that $\sup_n \|\tilde{v}_n\|^2 < \infty$ w.p.1.

6.5. *Proof of Th.* 3.4

The proof of Th. 3.2 easily adapts to Algorithm 2 by replacing V_n by

$$\tilde{V}_n \triangleq F(x_n) + \frac{1}{2} \|m_n\|^2.$$

The boundedness of (m_n) is an immediate consequence of the convergence of $\tilde{V}_n$.

6.6. Proof of Th. 3.5

We shall use the following result.

Theorem 6.2 (adapted from [37], Th. 7). Let $k \geq 1$. On some probability space equipped with a filtration $\mathscr{F} = (\mathscr{F}_n)_{n \in \mathbb{N}}$, consider a sequence of r.v. on $\mathbb{R}^k$ given by

$$Z_{n+1} = (I + \gamma_{n+1}\bar{H})Z_n + \gamma_{n+1}b_{n+1} + \sqrt{\gamma_{n+1}}\eta_{n+1}$$

and $\mathbb{E}[\|Z_0\|^2] < \infty$, where $\bar{H}$ is a $k \times k$ Hurwitz matrix, (b_n) and (η_n) are random sequences, and $\gamma_n = \gamma_0 n^{-\alpha}$ for some $\gamma_0 > 0$ and $\alpha \in (0,1]$. Let $\Omega_0 \in \mathscr{F}_\infty$ have a positive probability. Assume that the following holds almost surely on Ω_0:

  i) $\mathbb{E}[\eta_{n+1}|\mathscr{F}_n] = 0$.
 ii) There exists a constant $\bar{b} > 2$ s.t. $\sup_{n \geq 0} \mathbb{E}[\|\eta_{n+1}\|^{\bar{b}}|\mathscr{F}_n] < \infty$.
iii) $\mathbb{E}[\eta_{n+1}\eta_{n+1}^{\mathrm{T}}|\mathscr{F}_n] = \Sigma + \Delta_n$ where $\mathbb{E}[\|\Delta_n\|\mathbb{1}_{\Omega_0}] \to 0$ and Σ is a positive semidefinite matrix.
 iv) The sequence (b_n) is the sum of two sequences $(b_{n,1})$ and $(b_{n,2})$, adapted to $\mathscr{F}$, s.t. $\sup_{n \geq 0} \mathbb{E}[\|b_{n,1}\|^2] < \infty$, $\mathbb{E}[\|b_{n,1}\|\mathbb{1}_{\Omega_0}] \to 0$ and $b_{n,2} \to 0$ a.s. on Ω_0.

Then, given Ω_0, (Z_n) converges in distribution to the unique stationary distribution $\mu_\star$ of the generalized Ornstein-Uhlenbeck process

$$dX_t = \bar{H}X_t dt + \sqrt{\Sigma}dB_t$$

where (B_t) is the standard Brownian motion and $\sqrt{\Sigma}$ is the unique positive semidefinite square root of Σ. The distribution $\mu_\star$ is the zero mean Gaussian distribution with covariance matrix Γ given as the solution to $(\bar{H} + \frac{\mathbb{1}_{\alpha=1}}{2\gamma_0}I_k)\Gamma + \Gamma(\bar{H} + \frac{\mathbb{1}_{\alpha=1}}{2\gamma_0}I_k)^{\mathrm{T}} = -\Sigma$.

Proof. The proof is identical to the proof of [37, Th. 7], only substituting the inverse of the square root of Σ by the Moore-Penrose inverse. Finally, the uniqueness of the stationary distribution $\mu_\star$ and its expression follow from [28, Th. 6.7, p. 357]. $\square$

We define $v_n = \bar{v}_n + \delta_n$ where $\delta_0 = 0$, $\bar{v}_0 = v_0$ and

$$\delta_{n+1} = (1 - \gamma_{n+1}q_n)\delta_n + \gamma_{n+1}(p_n - q_n q_\infty^{-1}p_\infty)S(x_n),$$

$$\begin{aligned}\bar{v}_{n+1} &= (1 - \gamma_{n+1}q_n)\bar{v}_n + \gamma_{n+1}q_n q_\infty^{-1}p_\infty S(x_n) \\ &\quad + \gamma_{n+1}p_n(\nabla f(x_n, \xi_{n+1})^{\odot 2} - S(x_n)).\end{aligned}$$

For every $z = (v, m, x) \in \mathcal{Z}_+$ and $\delta \geq 0$, we define

$$r_n(z, \delta) \triangleq \begin{bmatrix} q_n q_\infty^{-1} p_\infty (S(x - \gamma_n \frac{m}{\sqrt{v+\delta+\varepsilon}}) - S(x)) \\ h_n(\nabla F(x - \gamma_n \frac{m}{\sqrt{v+\delta+\varepsilon}}) - \nabla F(x)) \\ \frac{\gamma_n}{\gamma_{n+1}}(\frac{1}{\sqrt{v+\varepsilon}} - \frac{1}{\sqrt{v+\delta+\varepsilon}}) \odot m \end{bmatrix}.$$

Moreover, for every $z = (v, m, x) \in \mathcal{Z}_+$ and every $n \in \mathbb{N}$, we set

$$g_n(z) = \begin{bmatrix} q_n q_\infty^{-1} p_\infty S(x) - q_n v \\ h_n \nabla F(x) - r_n m \\ -\frac{\gamma_n}{\gamma_{n+1}} \frac{m}{\sqrt{v+\varepsilon}} \end{bmatrix} .$$

Defining $\zeta_n = (\bar{v}_n, m_n, x_{n-1})$ and recalling the definition of (η_n) from Eq. (28), we have the decomposition

$$\zeta_{n+1} = \zeta_n + \gamma_{n+1} g_n(\zeta_n) + \gamma_{n+1} \eta_{n+1} + \gamma_{n+1} r_n(\zeta_n, \delta_n) .$$

Define $z_\star \triangleq (x_\star, 0, v_\star)$. Note that $g_n(z_\star) = 0$. Evaluating the Jacobian matrix G_n of g_n at $z_\star$, we obtain that there exist constants $C > 0$, $\bar{M} > 0$ and $n_0 \in \mathbb{N}$ s.t. for all $n \geq n_0$,

$$\|g_n(z) - G_n(z - z_\star)\| \leq C \|z - z_\star\|^2 \quad (\forall z \in B(z_\star, \bar{M})), \tag{51}$$

where G_n is given by

$$G_n \triangleq \begin{bmatrix} -q_n I_d & 0 & q_n q_\infty^{-1} p_\infty \nabla S(x_\star) \\ 0 & -r_n I_d & h_n \nabla^2 F(x_\star) \\ 0 & -\frac{\gamma_n}{\gamma_{n+1}} V & 0 \end{bmatrix} ,$$

where ∇S is the Jacobian of S and the matrix V is defined in Eq. (8). We define

$$G_\infty \triangleq \lim_n G_n = \begin{bmatrix} -q_\infty I_d & 0 & p_\infty \nabla S(x_\star) \\ 0 & -r_\infty I_d & h_\infty \nabla^2 F(x_\star) \\ 0 & -V & 0 \end{bmatrix} .$$

One can verify that G_∞ is Hurwitz, and that the largest real part of its eigenvalues is $-L'$, where $L' \triangleq L \wedge q_\infty$ and L is defined in Eq. (9).

We define $\Omega^{(0)} \triangleq \{z_n \to z_\star\}$. We assume $\mathbb{P}(\Omega^{(0)}) > 0$. Using for instance [19, Lem. 4 and Lem. 5], it holds that $\delta_n(\omega) \to 0$ for every $\omega \in \Omega^{(0)}$, and since $x_n(\omega) - x_{n-1}(\omega) \to 0$ on that set, we obtain that $\Omega^{(0)} = \{\zeta_n \to z_\star\}$. Let $M \in (0, \bar{M})$ be a constant, whose value will be specified later on. For every $N_0 \in \mathbb{N}$, define $\Omega_{N_0}^{(0)} \triangleq \{\zeta_n \to z_\star$ and $\sup_{n \geq N_0} \|\zeta_n - z_\star\| \leq M\}$. We seek to show that $\sqrt{\gamma_n}^{-1}(\zeta_n - z_\star) \Rightarrow \nu$ given $\Omega^{(0)}$, for some Gaussian measure ν, using Th. 6.2. As $\Omega_{N_0}^{(0)} \uparrow \Omega^{(0)}$, it is sufficient to show that the latter convergence holds given $\Omega_{N_0}^{(0)}$, for every N_0 large enough. From now on, we consider that N_0 is fixed. We define the sequence $(\tilde{\zeta}_n)_{n \geq N_0}$ as $\tilde{\zeta}_{N_0} = \zeta_{N_0}$ and for every $n \geq N_0$,

$$\tilde{\zeta}_{n+1} = \tilde{\zeta}_n + \gamma_{n+1} \tilde{g}_n(\tilde{\zeta}_n) + \gamma_{n+1}(\eta_{n+1} + r_n(\tilde{\zeta}_n, \delta_n)) \mathbb{1}_{\mathcal{A}_n}$$

where $\mathcal{A}_n$ is the event defined by

$$\mathcal{A}_n \triangleq \bigcap_{k=N_0}^{n} \{\|x_k - x_\star\| \leq M\} \cap \{\|\tilde{\zeta}_n - z_\star\| \leq M\}$$

and

$$\tilde{g}_n(z) \triangleq g_n(z)\mathbb{1}_{\|z-z_\star\|\leq M} - K(z - z_\star)\mathbb{1}_{\|z-z_\star\|>M},$$

where $K > 0$ is a large constant which will be specified later on. The sequences $(\tilde{\zeta}_n)_{n\geq N_0}$ and $(\zeta_n)_{n\geq N_0}$ coincide on $\Omega_{N_0}^{(0)}$. Thus, it is sufficient to study the weak convergence of $(\tilde{\zeta}_n)_{n\geq N_0}$.

An estimate of $\|r_n(\tilde{\zeta}_n, \delta_n)\|\mathbb{1}_{\mathcal{A}_n}$ We start by studying the sequence $(\|\delta_n\|\mathbb{1}_{\mathcal{A}_n})$. Unfolding the update rule defining δ_n and using the fact that (q_n) is a sequence of positive reals converging to $q_\infty > 0$, we obtain that

$$\|\delta_n\|\mathbb{1}_{\mathcal{A}_n} \leq \sum_{k=1}^n \left[\prod_{j=k+1}^n |1 - \gamma_j q_{j-1}|\right]\gamma_k|p_{k-1} - q_{k-1}q_\infty^{-1}p_\infty|\|S(x_{k-1})\|\mathbb{1}_{\mathcal{A}_n}$$

$$\leq C\sum_{k=1}^n \exp\left(-\beta\sum_{j=k+1}^n \gamma_j\right)\gamma_k|p_{k-1} - q_{k-1}q_\infty^{-1}p_\infty| \triangleq w_n,$$

for some $\beta > 0$. The sequence (w_n) is deterministic and converges to zero by [19, Lem. 4]. There exists $n_1 \geq n_0$ s.t. $w_n \leq M$. As $v \mapsto \frac{1}{\sqrt{v+\varepsilon}}$ is Lipschitz and ∇F and S are locally Lipschitz, for every $z = (v, m, x)$ and δ s.t. $\|z - z_\star\| \leq M$ and $\|\delta\| \leq M$, we have

$$\|r_n(z, \delta)\| \leq C\gamma_{n+1}\|(v + \delta + \varepsilon)^{\odot-\frac{1}{2}}\|\|m\|$$

$$+ C\|(v + \delta + \varepsilon)^{\odot-\frac{1}{2}} - (v + \varepsilon)^{\odot-\frac{1}{2}}\|\|m\|$$

$$\leq C\gamma_{n+1}\|z - z_\star\| + C\|\delta\|\|z - z_\star\|.$$

This implies that for every $n \geq n_1$,

$$\|r_n(\tilde{\zeta}_n, \delta_n)\|\mathbb{1}_{\mathcal{A}_n} \leq C(\gamma_{n+1} + w_n)\|\tilde{\zeta}_n - z_\star\|. \tag{52}$$

Tightness of $\sqrt{\gamma_n}^{-1}(\tilde{\zeta}_n - z_\star)$ We decompose

$$\tilde{\zeta}_{n+1} - z_\star = (I_{3d} + \gamma_{n+1}G_n)(\tilde{\zeta}_n - z_\star) + \gamma_{n+1}\left(g_n(\tilde{\zeta}_n) - G_n(\tilde{\zeta}_n - z_\star)\right)\mathbb{1}_{\|\tilde{\zeta}_n - z_\star\|\leq M}$$

$$- \gamma_{n+1}(K + G_n)(\tilde{\zeta}_n - z_\star)\mathbb{1}_{\|\tilde{\zeta}_n - z_\star\|>M} + \gamma_{n+1}(\eta_{n+1} + r_n(\tilde{\zeta}_n, \delta_n))\mathbb{1}_{\mathcal{A}_n}. \tag{53}$$

For a given $t > 0$, we write $G_\infty = B_t^{-1}G_tB_t$ the Jordan-like decomposition of G_∞, where the ones of the second diagonal of the usual Jordan decomposition are replaced by t, and where B_t is some invertible matrix. We define $S_n \triangleq B_t(\tilde{\zeta}_n - z_\star)$. Setting $G_n^{(t)} \triangleq B_tG_nB_t^{-1}$, we obtain

$$S_{n+1} = (I_{3d} + \gamma_{n+1}G_n^{(t)})S_n + \gamma_{n+1}B_t\left(g_n(\tilde{\zeta}_n) - G_n(\tilde{\zeta}_n - z_\star)\right)\mathbb{1}_{\|\tilde{\zeta}_n - z_\star\|\leq M}$$

$$- \gamma_{n+1}(K + G_n^{(t)})S_n\mathbb{1}_{\|\tilde{\zeta}_n - z_\star\|>M} + \gamma_{n+1}B_t(\eta_{n+1} + r_n(\tilde{\zeta}_n, \delta_n))\mathbb{1}_{\mathcal{A}_n}.$$

Choose $A \in (0, 2L')$ and $A' \in (A, 2L')$. There exists $\bar{\gamma}$ and $t > 0$ s.t. for every $\gamma < \bar{\gamma}$, $\|I + \gamma G_t\|_2 \leq 1 - \gamma(A' + 2L')/2$, where $\|\cdot\|_2$ is the spectral norm. As $G_n^{(t)} \to G^t$, there exists $n_2 \geq n_1$, such that for all $n \geq n_2$, $\|I + \gamma G_n^{(t)}\|_2 \leq 1 - \gamma A'$. Recall the notation $\mathbb{E}_n = \mathbb{E}[\cdot \,|\, \mathscr{F}_n]$. We expand $\|S_{n+1}\|^2$ and use the inequality $\left\|g_n(\tilde{\zeta}_n) - G_n(\tilde{\zeta}_n - z_\star)\right\|^2 \mathbb{1}_{\|\tilde{\zeta}_n - z_\star\| \leq M} \leq C\|S_n\|^2$ to obtain after straightforward algebra

$$
\begin{aligned}
\mathbb{E}_n \|S_{n+1}\|^2 &\leq (1 - \gamma_{n+1} A')\|S_n\|^2 + C\gamma_{n+1}^2\|S_n\|^2 \\
&\quad + C\gamma_{n+1}^2(\mathbb{E}_n\|\eta_{n+1}\|^2 + \|r_n(\tilde{\zeta}_n, \delta_n)\|^2)\mathbb{1}_{\mathcal{A}_n} \\
&\quad + 2\gamma_{n+1}\mathrm{Re}\left(S_n^* B_t \left(g_n(\tilde{\zeta}_n) - G_n(\tilde{\zeta}_n - z_\star)\right)\right)\mathbb{1}_{\|\tilde{\zeta}_n - z_\star\| \leq M}
\end{aligned}
$$
$$
-2\gamma_{n+1}\mathrm{Re}\left(S_n^*(K + G_n^{(t)})S_n\right)\mathbb{1}_{\|\tilde{\zeta}_n - z_\star\| > M} + 2\gamma_{n+1}\mathrm{Re}\left(S_n^* B_t r_n(\tilde{\zeta}_n, \delta_n)\right)\mathbb{1}_{\mathcal{A}_n}.
$$

Choose $c \triangleq (A' - A)/2$. If M is chosen small enough,

$$
\|g_n(\tilde{\zeta}_n) - G_n(\tilde{\zeta}_n - z_\star)\|\mathbb{1}_{\|\tilde{\zeta}_n - z_\star\| \leq M} \leq \frac{c}{2}\|B_t\|^{-1}\|B_t^{-1}\|\|\tilde{\zeta}_n - z_\star\|.
$$

Moreover, choosing $K > \sup_n \|G_n^{(t)}\|_2$, it holds that $\mathrm{Re}\left(S_n^*(K + G_n^{(t)})S_n\right) \geq 0$. Then,

$$
\begin{aligned}
\mathbb{E}_n \|S_{n+1}\|^2 &\leq (1 - \gamma_{n+1}(A' - c))\|S_n\|^2 + C\gamma_{n+1}^2\|S_n\|^2 \\
&+ C\gamma_{n+1}^2(\mathbb{E}_n\|\eta_{n+1}\|^2 + \|r_n(\tilde{\zeta}_n, \delta_n)\|^2)\mathbb{1}_{\mathcal{A}_n} + 2\gamma_{n+1}\|B_t\|\|S_n\|\|r_n(\tilde{\zeta}_n, \delta_n)\|\mathbb{1}_{\mathcal{A}_n}.
\end{aligned}
$$

Using Eq. (52),

$$
\begin{aligned}
\mathbb{E}_n \|S_{n+1}\|^2 &\leq (1 - \gamma_{n+1}(A' - c - w_n))\|S_n\|^2 + C\gamma_{n+1}^2(1 + w_n^2)\|S_n\|^2 \\
&\quad\quad\quad\quad\quad\quad + C\gamma_{n+1}^2\mathbb{E}_n\|\eta_{n+1}\|^2 \mathbb{1}_{\mathcal{A}_n}.
\end{aligned}
$$

Therefore, there exists $n_3 \geq n_2$ s.t. for all $n \geq n_3$,

$$
\mathbb{E}\|S_{n+1}\|^2 \leq (1 - \gamma_{n+1}A)\mathbb{E}\|S_n\|^2 + C\gamma_{n+1}^2\mathbb{E}(\|\eta_{n+1}\|^2 \mathbb{1}_{\|x_n - x_\star\| \leq M}).
$$

The second expectation in the righthand side is bounded uniformly in n by the condition (7). Using [19, Lem. 4 and Lem. 5], we conclude that $\sup_n \gamma_n^{-1}\mathbb{E}\|S_n\|^2 < \infty$. Therefore, $\sup_n \gamma_n^{-1}\mathbb{E}\|\tilde{\zeta}_n - z_\star\|^2 < \infty$, which in turn implies $\sup_n \gamma_n^{-1}\mathbb{E}(\|\zeta_n - z_\star\|^2 \mathbb{1}_{\Omega_{N_0}^{(0)}}) < \infty$.

Strongly perturbed iterations We define $\tilde{y}_n = \sqrt{\gamma_n}^{-1}(\tilde{\zeta}_n - z_\star)$. Define

$$
\bar{G}_n \triangleq \gamma_{n+1}^{-1}\left(\sqrt{\frac{\gamma_n}{\gamma_{n+1}}} - 1\right)I_{3d} + \sqrt{\frac{\gamma_n}{\gamma_{n+1}}}G_n.
$$

The sequence $\bar{G}_n$ converges to $\bar{G}_\infty \triangleq G_\infty + \frac{\mathbb{1}_{\alpha=1}}{2\gamma_0}I_{3d}$. Recalling Eq. (53), we can write

$$
\tilde{y}_{n+1} = (I_{3d} + \gamma_{n+1}\bar{G}_\infty)\tilde{y}_n + \gamma_{n+1}\bar{r}_n + \sqrt{\gamma_{n+1}}\bar{\eta}_{n+1}
$$

where $\bar\eta_{n+1} = \eta_{n+1}\mathbb{1}_{\mathcal{A}_n}$ and $\bar r_n = \bar r_{n,1} + \bar r_{n,2} + \bar r_{n,3}$, where

$$\bar r_{n,1} \triangleq \sqrt{\gamma_{n+1}}^{-1} r_n(\tilde\zeta_n, \delta_n)\mathbb{1}_{\mathcal{A}_n} + (\bar G_n - \bar G_\infty)\tilde y_n$$

$$\bar r_{n,2} \triangleq \sqrt{\gamma_{n+1}}^{-1}\left(g_n(\tilde\zeta_n) - G_n(\tilde\zeta_n - z_\star)\right)\mathbb{1}_{\|\tilde\zeta_n - z_\star\| \le M}$$

$$\bar r_{n,3} \triangleq -\sqrt{\gamma_{n+1}}^{-1}(K + G_n)(\tilde\zeta_n - z_\star)\mathbb{1}_{\|\tilde\zeta_n - z_\star\| > M}\,.$$

We now check that the assumptions of Th. 6.2 are fulfilled. On the event $\Omega_{N_0}^{(0)}$, we recall that $\tilde\zeta_n = \zeta_n$, hence $\bar r_{n,3}$ is identically zero. Moreover, using Eq. (52), it holds that for all n large enough,

$$\|\bar r_{n,1}\| \le C\left(\sqrt{\frac{\gamma_n}{\gamma_{n+1}}}(\gamma_{n+1} + w_n) + \|\bar G_n - \bar G_\infty\|\right)\|\tilde y_n\|$$

and therefore, $\mathbb{E}[\|\bar r_{n,1}\|^2] \to 0$. Now consider the term $\bar r_{n,2}$. By Eq. (51),

$$\|\bar r_{n,2}\| \le C\sqrt{\gamma_{n+1}}^{-1}\|\tilde\zeta_n - z_\star\|^2\mathbb{1}_{\|\tilde\zeta_n - z_\star\| \le M}\,.$$

Thus, $\|\bar r_{n,2}\|^2 \le C\|\tilde y_n\|^2$ which implies that $\sup_{n \ge N_0}\mathbb{E}[\|r_{n,2}\|^2] < \infty$. Moreover, $\mathbb{E}[\|\bar r_{n,2}\|] \le C\sqrt{\gamma_{n+1}}\mathbb{E}\|\tilde y_n\|^2$ tends to zero. Finally, consider $\bar\eta_{n+1}$. Using condition (7), there exist $M > 0$ and $b_M > 4$ s.t.

$$\mathbb{E}_n[\|\bar\eta_{n+1}\|^{b_M/2}] \le \mathbb{E}_n[\|\eta_{n+1}\|^{b_M/2}]\mathbb{1}_{\|x_n - x_\star\| \le M}$$
$$\le C\mathbb{E}_n[\|\nabla f(x_n, \xi_{n+1})\|^{b_M}]\mathbb{1}_{\|x_n - x_\star\| \le M} \le C\,.$$

Moreover, $\mathbb{E}_n[\bar\eta_{n+1}] = 0$ and finally, almost surely on $\Omega_N^{(0)}$, $\mathbb{E}_n[\bar\eta_{n+1}\bar\eta_{n+1}^{\mathrm{T}}]$ converges to

$$\Sigma \triangleq \left[\begin{array}{cc} \mathbb{E}_\xi\left[\begin{bmatrix} p_\infty(\nabla f(x_\star, \xi)^{\odot 2} - S(x_\star)) \\ h_\infty\nabla f(x_\star, \xi) \\ 0 \end{bmatrix}\begin{bmatrix} p_\infty(\nabla f(x_\star, \xi)^{\odot 2} - S(x_\star)) \\ h_\infty\nabla f(x_\star, \xi) \\ 0 \end{bmatrix}^{\mathrm{T}}\right] & \begin{array}{c} 0 \\ 0 \\ 0 \end{array} \\ & 0 \end{array}\right].$$
$$\tag{54}$$

Therefore, the assumptions of Th. 6.2 are fulfilled for the sequence $\tilde y_n$. We obtain the desired result for the sequence (m_n, x_{n-1}). We now show that the same result also holds for the sequence (m_n, x_n). For this purpose, observe that

$$\frac{1}{\sqrt{\gamma_n}}\begin{bmatrix} m_n \\ x_n - x_\star \end{bmatrix} = \frac{1}{\sqrt{\gamma_n}}\begin{bmatrix} m_n \\ x_{n-1} - x_\star \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{\sqrt{\gamma_n}}(x_n - x_{n-1}) \end{bmatrix}\,.$$

Then, notice that $\|\frac{x_n - x_{n-1}}{\sqrt{\gamma_n}}\| = \sqrt{\gamma_n}\|\frac{m_n}{\sqrt{v_n + \varepsilon}}\| \le \sqrt{\frac{\gamma_n}{\varepsilon}}\|m_n\| \to 0$ as $n \to \infty$ since it is assumed that $z_n \to z_\star$ (which implies in particular that $m_n \to 0$). Hence, it holds that $\sqrt{\gamma_n}^{-1}(x_n - x_{n-1})$ converges a.s. to 0. We conclude by invoking Slutsky's lemma.

Proof of Eq. (10). We have the subsystem:

$$\tilde H\Gamma + \Gamma\tilde H^{\mathrm{T}} = \begin{bmatrix} -h_\infty^2\mathcal{Q} & 0 \\ 0 & 0 \end{bmatrix} \qquad \text{where } \tilde H \triangleq \begin{bmatrix} (\theta - r_\infty)I_d & h_\infty\nabla^2 F(x_\star) \\ -V & \theta I_d \end{bmatrix} \tag{55}$$

and where $\mathcal{Q} \triangleq \mathrm{Cov}\left(\nabla f(x_\star, \xi)\right)$. The next step is to triangularize the matrix $\tilde{H}$ in order to decouple the blocks of Γ. For every $k = 1, \ldots, d$, set $\nu_k^\pm \triangleq -\frac{r_\infty}{2} \pm \sqrt{r_\infty^2/4 - h_\infty \pi_k}$ with the convention that $\sqrt{-1} = \imath$ (inspecting the characteristic polynomial of H, these are the eigenvalues of H). Set $M^\pm \triangleq \mathrm{diag}\left(\nu_1^\pm, \cdots, \nu_d^\pm\right)$ and $R^\pm \triangleq V^{-\frac{1}{2}} P M^\pm P^{\mathrm{T}} V^{-\frac{1}{2}}$. Using the identities $M^+ + M^- = -r_\infty I_d$ and $M^+ M^- = h_\infty \mathrm{diag}\left(\pi_1, \cdots, \pi_d\right)$, it can be checked that

$$\mathcal{R}\tilde{H} = \begin{bmatrix} R^- V + \theta I_d & 0 \\ -V & V R^+ + \theta I_d \end{bmatrix} \mathcal{R}, \text{ where } \mathcal{R} \triangleq \begin{bmatrix} I_d & R^+ \\ 0 & I_d \end{bmatrix}.$$

Set $\tilde{\Gamma} \triangleq \mathcal{R}\Gamma\mathcal{R}^{\mathrm{T}}$. Denote by $(\tilde{\Gamma}_{i,j})_{i,j=1,2}$ the blocks of $\tilde{\Gamma}$. Note that $\tilde{\Gamma}_{2,2} = \Gamma_{2,2}$. By left/right multiplication of Eq. (55) respectively by $\mathcal{R}$ and $\mathcal{R}^{\mathrm{T}}$, we obtain

$$(R^- V + \theta I_d)\tilde{\Gamma}_{1,1} + \tilde{\Gamma}_{1,1}(V R^- + \theta I_d) = -h_\infty^2 \mathcal{Q} \tag{56}$$

$$(R^- V + \theta I_d)\tilde{\Gamma}_{1,2} + \tilde{\Gamma}_{1,2}(R^+ V + \theta I_d) = \tilde{\Gamma}_{1,1} V \tag{57}$$

$$(V R^+ + \theta I_d)\tilde{\Gamma}_{2,2} + \tilde{\Gamma}_{2,2}(R^+ V + \theta I_d) = V\tilde{\Gamma}_{1,2} + \tilde{\Gamma}_{1,2}^{\mathrm{T}} V. \tag{58}$$

Set $\bar{\Gamma}_{1,1} = P^{-1} V^{\frac{1}{2}} \tilde{\Gamma}_{1,1} V^{\frac{1}{2}} P$. Define $C \triangleq P^{-1} V^{\frac{1}{2}} \mathcal{Q} V^{\frac{1}{2}} P$. Eq. (56) yields

$$(M^- + \theta I_d)\bar{\Gamma}_{1,1} + \bar{\Gamma}_{1,1}(M^- + \theta I_d) = -h_\infty^2 C.$$

Set $\bar{\Gamma}_{1,2} = P^{-1} V^{\frac{1}{2}} \tilde{\Gamma}_{1,2} V^{-\frac{1}{2}} P$. Eq. (57) is rewritten $(M^- + \theta I_d)\bar{\Gamma}_{1,2} + \bar{\Gamma}_{1,2}(M^+ + \theta I_d) = \bar{\Gamma}_{1,1}$. The component (k, ℓ) is given by

$$\bar{\Gamma}_{1,2}^{k,\ell} = (\nu_k^- + \nu_\ell^+ + 2\theta)^{-1}\bar{\Gamma}_{1,1}^{k,\ell} = \frac{-h_\infty^2 C_{k,\ell}}{(\nu_k^- + \nu_\ell^+ + 2\theta)(\nu_k^- + \nu_\ell^- + 2\theta)}.$$

Set finally $\bar{\Gamma}_{2,2} = P^{-1} V^{-\frac{1}{2}} \Gamma_{2,2} V^{-\frac{1}{2}} P$. Eq. (58) becomes

$$(M^+ + \theta I_d)\bar{\Gamma}_{2,2} + \bar{\Gamma}_{2,2}(M^+ + \theta I_d) = \bar{\Gamma}_{1,2} + \bar{\Gamma}_{1,2}^{\mathrm{T}}.$$

Thus,

$$\begin{aligned}
\bar{\Gamma}_{2,2}^{k,\ell} &= \frac{\bar{\Gamma}_{1,2}^{k,\ell} + \bar{\Gamma}_{1,2}^{\ell,k}}{\nu_k^+ + \nu_\ell^+ + 2\theta} \\
&= \frac{-h_\infty^2 C_{k,\ell}}{(\nu_k^+ + \nu_\ell^+ + 2\theta)(\nu_k^- + \nu_\ell^- + 2\theta)} \left(\frac{1}{\nu_k^- + \nu_\ell^+ + 2\theta} + \frac{1}{\nu_k^+ + \nu_\ell^- + 2\theta}\right).
\end{aligned}$$

After tedious but straightforward computations, we obtain

$$\bar{\Gamma}_{2,2}^{k,\ell} = \frac{h_\infty^2 C_{k,\ell}}{(r_\infty - 2\theta)(h_\infty(\pi_k + \pi_\ell) + 2\theta(\theta - r_\infty)) + \frac{h_\infty^2 (\pi_k - \pi_\ell)^2}{2(r_\infty - 2\theta)}},$$

and the result is proved. $\qquad\square$

7. Proofs for Section 4

7.1. Preliminaries

Most of the avoidance of traps results in the stochastic approximation literature deal with the case where the ODE that underlies the stochastic algorithm under study is an autonomous ODE $\dot{\mathsf{z}} = h(\mathsf{z})$. In this setting, a point $z_\star \in \operatorname{zer} h$ is called a trap if $h(z)$ admits an expansion around $z_\star$ of the type $h(z) = D(z - z_\star) + o(\|z - z_\star\|)$, where the matrix D has at least one eigenvalue which real part is (strictly) positive. Initiated by Pemantle [38] and by Brandière and Duflo [12], the most powerful class of techniques for establishing avoidance of traps results makes use of Poincaré's invariant manifold theorem for the ODE $\dot{\mathsf{z}} = h(\mathsf{z})$ in a neighborhood of some point $z_\star \in \operatorname{zer} h$. The idea is to show that with probability 1, the stochastic algorithm strays away from the maximal invariant manifold of the ODE where the convergence to $z_\star$ of the ODE flow can take place. As previously mentioned, since we are dealing with algorithms derived from non-autonomous ODEs, we extend the results of [38, 12] to this setting. The proof of Th. 4.1 relies on a non-autonomous version of Poincaré's theorem. We borrow this result from the rich literature that exists on the subject [16, 30].

Let us start by setting the context for the non-autonomous version that we shall need for the invariant manifold theorem. Given an integer $d > 0$ and a matrix $D \in \mathbb{R}^{d \times d}$, consider the linear autonomous differential equation

$$\dot{\mathsf{z}}(t) = D\mathsf{z}(t), \tag{59}$$

which solution is obviously $\mathsf{z}(t) = e^{Dt}\mathsf{z}(0)$ for $t \in \mathbb{R}$. Let us factorize D as in (12), and write $D = Q\Lambda Q^{-1}$ with $\Lambda = \begin{bmatrix} \Lambda^- & \\ & \Lambda^+ \end{bmatrix}$ where we recall that the Jordan blocks that constitute $\Lambda^- \in \mathbb{R}^{d^- \times d^-}$ (respectively $\Lambda^+ \in \mathbb{R}^{d^+ \times d^+}$) are those that contain the eigenvalues λ_i of D such that $\Re \lambda_i \leq 0$ (respectively $\Re \lambda_i > 0$). Let us assume here that both d^- and d^+ are positive. It will be convenient to work in the basis of the columns of Q by making the variable change

$$z \mapsto y = \begin{bmatrix} y^- \\ y^+ \end{bmatrix} = Q^{-1}z,$$

where $y^\pm \in \mathbb{R}^{d^\pm}$. In this new basis, the ODE (59) is written as

$$\begin{bmatrix} \dot{\mathsf{y}}^- \\ \dot{\mathsf{y}}^+ \end{bmatrix} = \begin{bmatrix} \Lambda^- & \\ & \Lambda^+ \end{bmatrix} \begin{bmatrix} \mathsf{y}^- \\ \mathsf{y}^+ \end{bmatrix}, \tag{60}$$

which solution is $\mathsf{y}^\pm(t) = \exp(t\Lambda^\pm)\mathsf{y}^\pm(0)$. One can readily check that for each couple of real numbers α^+ and α^- that satisfy

$$0 < \alpha^- < \alpha^+ < \min\{\Re \lambda_i \ : \ \Re \lambda_i > 0\}, \tag{61}$$

there exists a so-called exponential dichotomy of the ODE solutions, which amounts in our case to the existence of two constants $K^-, K^+ \geq 1$ such that

$$\| \exp(t\Lambda^-) \| \leq K^- e^{\alpha^- t} \quad \text{for } t \geq 0,$$
$$\| \exp(t\Lambda^+) \| \leq K^+ e^{\alpha^+ t} \quad \text{for } t \leq 0,$$

see, *e.g.*, [26].

We now consider a non-autonomous perturbation of this ODE, which is represented in the basis of the columns of Q as

$$\dot{\mathsf{y}}(t) = h(\mathsf{y}(t), t) \quad \text{with} \quad h(y, t) = \begin{bmatrix} \Lambda^- & \\ & \Lambda^+ \end{bmatrix} y + \varepsilon(y, t), \tag{62}$$

and $\varepsilon : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ is a continuous function. In the sequel, we shall be interested in the asymptotic behavior of this equation for the large values of t, and therefore, restrict our study to the interval $\mathbb{I} = [t_0, \infty)$ for some given $t_0 \geq 0$ that we shall fix later. We assume that $\varepsilon(0, \cdot) = 0$ on $\mathbb{I}$. We denote as $\phi : \mathbb{I} \times \mathbb{I} \times \mathbb{R}^d \to \mathbb{R}^d$ the so-called general solution of (62), which is defined by the fact that $\phi(\cdot, t, x)$ is the unique noncontinuable solution of (62) such that $\phi(t, t, x) = x$ for $t \in \mathbb{I}$ and $x \in \mathbb{R}^d$, assuming this solution exists and is unique for each $(x, t) \in \mathbb{R}^d \times \mathbb{I}$.

In the linear autonomous case provided by the ODE (60), the subspace

$$\mathcal{G} = \left\{ \left(t, \begin{bmatrix} y^- \\ 0 \end{bmatrix} \right) \in \mathbb{R} \times \mathbb{R}^d \ : \ y^- \in \mathbb{R}^{d^-} \right\}$$

is trivially invariant in the sense that if $(t, y) \in \mathcal{G}$, then, $(s, \phi(s, t, y)) \in \mathcal{G}$ for each $s \in \mathbb{R}$. This concept can be generalized to the non-linear and non-autonomous case. We say that the $\mathcal{C}^1$ function $w : \mathbb{R}^{d^-} \times \mathbb{I} \to \mathbb{R}^{d^+}$ defines a global non-autonomous invariant manifold for the ODE (62) if $w(0, t) = 0$ for all $t \in \mathbb{I}$, and, furthermore, if for each $t \in \mathbb{I}$ and each $y^- \in \mathbb{R}^{d^-}$, writing $y = (y^-, w(y^-, t))$, the general solution $\phi(s, t, y) = (\phi^-(s, t, y), \phi^+(s, t, y))$ with $\phi^\pm(s, t, y) \in \mathbb{R}^{d^\pm}$ verifies $\phi^+(s, t, y) = w(\phi^-(s, t, y), s)$ for each $s \in \mathbb{I}$. The non-autonomous invariant manifold is the set

$$\mathcal{G} = \left\{ \left(t, \begin{bmatrix} y^- \\ w(y^-, t) \end{bmatrix} \right) \in \mathbb{I} \times \mathbb{R}^d \ : \ y^- \in \mathbb{R}^{d^-} \right\},$$

which obviously satisfies $(t, y) \in \mathcal{G} \Rightarrow (s, \phi(s, t, y)) \in \mathcal{G}$ for each $s \in \mathbb{I}$.

These invariant manifolds are described by the following proposition, which is a straightforward application of [39, Th. A.1] (see also [30, Th. 6.3 p. 106, Rem. 6.6 p. 111]). It is useful to note that under the conditions provided in the statement of this proposition, the existence of the general solution ϕ of the ODE (62) is ensured by Picard's theorem.

Proposition 7.1. Let $\mathbb{I} = [t_0, \infty)$ for some $t_0 \geq 0$. Assume that the function $\varepsilon(y, t)$ is such that $\varepsilon(0, \cdot) \equiv 0$ on $\mathbb{I}$, the function $\varepsilon(\cdot, t)$ is continuously differentiable for each $t \in \mathbb{I}$, and furthermore, the Jacobian matrix $\partial_1 \varepsilon(y, t)$ satisfies

$$|\varepsilon|_1 \triangleq \sup_{(y,t) \in \mathbb{R}^d \times \mathbb{I}} \| \partial_1 \varepsilon(y, t) \| < \frac{\alpha^+ - \alpha^-}{4K} \tag{63}$$

with $K = K^- + K^+ + K^- K^+ (K^- \vee K^+)$ and α^-, α^+ chosen as in Eq. (61). Then, for each $\delta \in (2K|\varepsilon|_1, (\alpha^+ - \alpha^-)/2)$ and each $\gamma \in (\alpha^- + \delta, \alpha^+ - \delta)$, the set

$$\mathcal{G} = \left\{ (t, y) \in \mathbb{I} \times \mathbb{R}^d \ : \ \sup_{s \geq t} \|\phi(s, t, y)\| \exp(\gamma(t - s)) < \infty \right\}$$

is nonempty, and does not depend on γ. Moreover, this set is a global invariant manifold for the ODE (62) that is defined by a continuously differentiable mapping $w : \mathbb{R}^{d^-} \times \mathbb{I} \to \mathbb{R}^{d^+}$. In addition, if the partial derivatives $\partial_1^k \varepsilon : \mathbb{R}^d \times \mathbb{I}$ exist and are continuous for $k \in \{1, \ldots, m\}$ with globally bounded partial derivatives

$$|\varepsilon|_k \triangleq \sup_{(y,t) \in \mathbb{R}^d \times \mathbb{I}} \|\partial_1^k \varepsilon(y, t)\| < \infty, \tag{64}$$

under the gap condition

$$m\alpha^- < \alpha^+, \ m \in \mathbb{N}^*, \tag{65}$$

the partial derivatives $\partial_1^k w : \mathbb{R}^{d^-} \times \mathbb{I}$ exist and are continuous with

$$\sup_{(y^-,t) \in \mathbb{R}^{d^-} \times \mathbb{I}} \|\partial_1^k w(y^-, t)\| < \infty \quad \text{for all } k \in \{1, \ldots, m\}. \tag{66}$$

Finally, if $\partial_2^n \partial_1^k \varepsilon$ exist and are continuous for $0 \leq n < m$ and $0 \leq k + n \leq m$, then w is m-times continuously differentiable.

Let us partition the function $h(y, t)$ as

$$h(y, t) = \begin{bmatrix} h^-(y, t) \\ h^+(y, t) \end{bmatrix} = \begin{bmatrix} \Lambda^- y^- + \varepsilon^-(y, t) \\ \Lambda^+ y^+ + \varepsilon^+(y, t) \end{bmatrix}, \tag{67}$$

where $h^\pm : \mathbb{R}^d \times \mathbb{I} \to \mathbb{R}^{d^\pm}$, $y^\pm \in \mathbb{R}^{d^\pm}$ and $\varepsilon^\pm : \mathbb{R}^d \times \mathbb{I} \to \mathbb{R}^{d^\pm}$. With these notations, the previous proposition leads to the following lemma.

Lemma 7.2. In the setting of Prop. 7.1, for each t in the interior of $\mathbb{I}$ and each vector $y = (y^-, y^+)$ such that $y^\pm \in \mathbb{R}^{d^\pm}$ and $y^+ = w(y^-, t)$, it holds that

$$h^+(y, t) = \partial_1 w(y^-, t) h^-(y, t) + \partial_2 w(y^-, t). \tag{68}$$

Assume that α^- is small enough so that Ineq. (65) and Eq. (64) hold true with $m = 2$. Assume in addition that $\partial_2^n \partial_1^k \varepsilon$ exists and is continuous for $0 \leq n < 2$ and $0 \leq k + n \leq 2$, and furthermore, that there exists a bounded neighborhood $\mathcal{V} \subset \mathbb{R}^d$ of zero such that

$$\sup_{(y,t) \in \mathcal{V} \times \mathbb{I}} \|\partial_2 \varepsilon(y, t)\| < +\infty. \tag{69}$$

Then, there exists a neighborhood $\mathcal{V}^- \subset \mathbb{R}^{d^-}$ of zero such that

$$\sup_{(y^-,t) \in \mathcal{V}^- \times \mathbb{I}} \|\partial_1 \partial_2 w(y^-, t)\| < +\infty, \tag{70}$$

$$\sup_{(y^-,t) \in \mathcal{V}^- \times \mathbb{I}} \|\partial_2^2 w(y^-, t)\| < +\infty. \tag{71}$$

Proof. By Prop. 7.1, the general solution $\phi(s, t, y)$ of the ODE (62) can be written as $\phi(s, t, y) = (\phi^-(s, t, y), \phi^+(s, t, y))$ with $\phi^+(s, t, y) = w(\phi^-(s, t, y), s)$ for each $s \in \mathbb{I}$. Equating the derivatives with respect to s of the two members of this equation and taking $s = t$, we get the first equation.

Writing $g : \mathbb{R}^{d^-} \times \mathbb{I} \to \mathbb{R}^d, (y^-, t) \mapsto (y^-, w(y^-, t))$, Eq. (68) can be rewritten as

$$\partial_2 w(y^-, t) = h^+(g(y^-, t), t) - \partial_1 w(y^-, t) h^-(g(y^-, t), t). \tag{72}$$

By Prop. 7.1, the function w is twice differentiable, and we can write

$$\partial_2^2 w(y^-, t) = \partial_1 h^+ \partial_2 g + \partial_2 h^+ - (\partial_1 \partial_2 w) h^- - (\partial_1 w)(\partial_1 h^- \partial_2 g + \partial_2 h^-), \tag{73}$$

where, *e.g.*, h^+ is a shorthand notation for $h^+(g(y^-, t), t)$. It holds from Eq. (67) and the assumptions of Prop. 7.1 that for each $(y, t) \in \mathbb{R}^d \times \mathbb{I}$,

$$\|\partial_1 h(y, t)\| \le \|\Lambda\| + \|\partial_1 \varepsilon(y, t)\| \le C, \tag{74}$$

where the constant $C > 0$ is independent of (y, t) and can change from an inequality to another in the remainder of the proof. By the mean value inequality and Prop. 7.1, we also get that

$$\|w(y^-, t)\| = \|w(y^-, t) - w(0, t)\| \le \sup_{(u,s)} \|\partial_1 w(u, s)\| \, \|y^-\| \le C \|y^-\|,$$

thus, $\|g(y^-, t)\| \le C \|y^-\|$. By the mean value inequality again,

$$\left\| h(g(y^-, t), t) \right\| = \left\| h(g(y^-, t), t) - h(0, t) \right\| \le \sup_{(u,t)} \|\partial_1 h(u, t)\| \left\| g(y^-, t) \right\|$$
$$\le C \left\| g(y^-, t) \right\| \le C \|y^-\|.$$

By Eq. (72) and Prop. 7.1, this implies that

$$\left\| \partial_2 g(y^-, t) \right\| = \left\| \partial_2 w(y^-, t) \right\| = \left\| h^+ - (\partial_1 w) h^- \right\| \le C \left\| y^- \right\|, \quad \text{and} \tag{75}$$
$$\left\| \partial_1 \partial_2 w(y^-, t) \right\| = \left\| \partial_1 h^+ \partial_1 g - (\partial_1^2 w) h^- - (\partial_1 w)(\partial_1 h^- \partial_1 g) \right\| \le C(\left\| y^- \right\| + 1). \tag{76}$$

Let $\mathcal{V}^- \subset \mathbb{R}^{d^-}$ be a small enough neighborhood of zero so that $g(y^-, t) \in \mathcal{V}$ for each $y^- \in \mathcal{V}^-$, which is possible by the inequality $\|g(y^-, t)\| \le C \|y^-\|$. By the assumption on $\|\partial_2 \varepsilon(y, t)\|$ in the statement of Lem. 7.2, we have

$$\forall y^- \in \mathcal{V}^-, \quad \left\| \partial_2 h(g(y^-, t), t) \right\| = \left\| \partial_2 \varepsilon(g(y^-, t), t) \right\| \le C. \tag{77}$$

The bound (70) is an immediate consequence of Eq. (76). Getting back to Eq. (73), the bound (71) follows from the inequalities (74)–(77). $\qquad \square$

Prop. 7.1 deals with the case where the function ε is globally Lipschitz continuous. In practical cases, such a strong assumption is not necessarily verified. In particular, for the ODEs we consider for our application, it is not satisfied (see the function e defined in Subsec. 7.3.1 below). Nonetheless, recall that we only

need the existence of a *local* non-autonomous invariant manifold, i.e. defined in the vicinity of an arbitrary solution such as the trivial zero solution (since we suppose here $\varepsilon(0, \cdot) = 0$) whereas the aforementioned strong assumption provides a global non-autonomous invariant manifold. Indeed, as for the avoidance of traps result we intend to show, we will only need to look at the behavior of our ODE in the neighborhood of a trap $z_\star$. Therefore, in prevision of the proof of Th. 4.1, we localize the ODE (62) in the neighborhood of zero. This is the purpose of the next proposition.

Proposition 7.3. Let $\mathbb{I} = [t_0, +\infty)$ for some $t_0 \geq 0$ and let $h : \mathbb{R}^d \times \mathbb{I} \to \mathbb{R}^d$ be defined as in Eq. (62). Assume that $\varepsilon(0, \cdot) \equiv 0$ on $\mathbb{I}$, that the function $\varepsilon(\cdot, t)$ is continuously differentiable for every $t \in \mathbb{I}$ and that

$$\lim_{(y,t) \to (0,+\infty)} \|\partial_1 \varepsilon(y, t)\| = 0 . \tag{78}$$

Then, there exist $\sigma > 0, t_1 > 0$, a function $\tilde{\varepsilon} : \mathbb{R}^d \times \mathbb{I}_1 \to \mathbb{R}^d$ where $\mathbb{I}_1 \triangleq [t_1, +\infty)$ and a function $\tilde{h} : \mathbb{R}^d \times \mathbb{I}_1 \to \mathbb{R}^d$ defined for every $y \in \mathbb{R}^d, t \in \mathbb{I}_1$ by $\tilde{h}(y, t) = \Lambda y + \tilde{\varepsilon}(y, t)$ s.t. $\tilde{h}$ and $\tilde{\varepsilon}$ verify the assumptions of Prop. 7.1 and for every $(y, t) \in B(0, \sigma) \times \mathbb{I}_1$, we have that $\tilde{h}(y, t) = h(y, t)$ and $\tilde{\varepsilon}(y, t) = \varepsilon(y, t)$. Moreover, for any $\delta > 0$, we can choose σ, t_1 respectively small and large enough s.t. the mapping $w : \mathbb{R}^{d^-} \times \mathbb{I}_1 \to \mathbb{R}^{d^+}$ obtained from Prop. 7.1 (applied to $\tilde{h}$ and $\tilde{\varepsilon}$) satisfies

$$|w|_1 = \sup_{(y,t) \in \mathbb{R}^{d^-} \times \mathbb{I}_1} \|\partial_1 w(y, t)\| < \delta . \tag{79}$$

Furthermore, Eq. (68) holds for $\tilde{h}$ and w for all $(y, t) \in B(0, \sigma) \times \mathbb{I}_1$. If, additionally, Eq. (69) holds for ε, then there exists $\sigma_1 \leq \sigma$ such that

$$\sup_{(y^-,t) \in B(0,\sigma_1) \times \mathbb{I}_1} \left\| \partial_1 \partial_2 w(y^-, t) \right\| < +\infty , \tag{80}$$

$$\sup_{(y^-,t) \in B(0,\sigma_1) \times \mathbb{I}_1} \left\| \partial_2^2 w(y^-, t) \right\| < +\infty . \tag{81}$$

Proof. The idea of the proof is to *localize* the function $h(y, t)$ to a neighborhood of zero in the variable y for the purpose of applying Prop. 7.1. This cut-off technique is known in the non-autonomous ODE literature, see, *e.g.*, [30, Th. 6.10]. Let $\psi : \mathbb{R}^d \to [0, 1]$ be a smooth function such that $\psi(y) = 1$ if $\|y\| \leq 1$, and $\psi(y) = 0$ if $\|y\| \geq 2$. Let $C = \max_y \|\nabla \psi(y)\|$ where $\nabla \psi$ is the Jacobian matrix of ψ. Thanks to the convergence (78), we can choose $t_1 > 0$ large enough and $\sigma > 0$ small enough so that

$$\sup_{(t,y) \in [t_1,\infty) \times B(0,2\sigma)} \|\partial_1 \varepsilon(y, t)\| < \frac{\alpha^+ - \alpha^-}{4K(1 + 2C)},$$

and we set $\mathbb{I}_1 = [t_1, \infty)$. Writing $\tilde{\varepsilon}(y, t) = \psi(y/\sigma)\varepsilon(y, t)$, it holds that for each $(t, y) \in \mathbb{I}_1 \times \mathbb{R}^d$,

$$\|\partial_1 \tilde{\varepsilon}(y, t)\| \leq \sigma^{-1} C \mathbb{1}_{\|y\| \leq 2\sigma} \|\varepsilon(y, t)\| + \mathbb{1}_{\|y\| \leq 2\sigma} \|\partial_1 \varepsilon(y, t)\|$$

$$\leq \left(\max_{\|y\|\leq 2\sigma} \|\partial_1\varepsilon(y,t)\| \right) \left(\sigma^{-1}C\|y\| + 1 \right) \mathbb{1}_{\|y\|\leq 2\sigma}$$

$$\leq \frac{\alpha^+ - \alpha^-}{4K},$$

where we used the mean value inequality along with $\varepsilon(0,t) = 0$ to obtain the second inequality. Thus, the function $\tilde{h}(y,t) = \Lambda y + \tilde{\varepsilon}(y,t)$ satisfies all the assumptions of Prop. 7.1. In addition, the function $\tilde{\varepsilon}$ coincides with the function ε on $B(0,\sigma_1) \times \mathbb{I}_1$, and so it is for the functions $\tilde{h}$ and h. Finally, it follows from [30, Th. 6.3] that

$$|w|_1 \leq \frac{2K^2}{\alpha_+ - \alpha_- - 4K|\tilde{\varepsilon}|_1} |\tilde{\varepsilon}|_1$$

(note that L in [30, Th. 6.3] corresponds to $|\tilde{\varepsilon}|_1$ with our notations). Using Eq. (78), we can make $|\tilde{\varepsilon}|_1$ as small as needed by choosing σ, t_1 respectively small and large enough, which gives us Eq. (79). The proof of the last two equations follows from the application of Lemma 7.2 to $\tilde{h}$ and w. The result is immediate after noticing that for $(y,t) \in \mathbb{R}^d \times \mathbb{I}_1$, we have $\|\partial_2\tilde{\varepsilon}(y,t)\| \leq \|\partial_2\varepsilon(y,t)\|$. $\qquad\square$

7.2. Proof of Th. 4.1

We shall rely on the following result of Brandière and Duflo. Recall that $(\Omega, \mathscr{F}, \mathbb{P})$ is a probability space equipped with a filtration $(\mathscr{F}_n)_{n\in\mathbb{N}}$.

Proposition 7.4. ([12, Prop. 4]) Given a sequence (γ_n) of deterministic non-negative stepsizes such that $\sum_k \gamma_k = +\infty$ and $\sum_k \gamma_k^2 < +\infty$, consider the $\mathbb{R}^d$–valued stochastic process $(z_n)_{n\in\mathbb{N}}$ given by

$$z_{n+1} = (I + \gamma_{n+1}H_n)z_n + \gamma_{n+1}\eta_{n+1} + \gamma_{n+1}\rho_{n+1}.$$

Assume that z_0 is $\mathscr{F}_0$–measurable and that the sequences (η_n), (ρ_n) together with the sequence of random matrices (H_n) are $(\mathscr{F}_n)$–adapted. Moreover, on a given event $A \in \mathscr{F}$, assume the following facts:

   i) $\sum_n \|\rho_n\|^2 < \infty$.
   ii) $\limsup \mathbb{E}[\|\eta_{n+1}\|^{2+a} \,|\, \mathscr{F}_n] < \infty$ for some $a > 0$, and $\mathbb{E}[\eta_{n+1} \,|\, \mathscr{F}_n] = 0$.
   iii) $\liminf \mathbb{E}[\|\eta_{n+1}\|^2 \,|\, \mathscr{F}_n] > 0$.

Let $H \in \mathbb{R}^{d\times d}$ be a deterministic matrix such that the real parts of its eigenvalues are all positive. Then,

$$\mathbb{P}\left(A \cap [z_n \to 0] \cap [H_n \to H]\right) = 0.$$

We now enter the proof of Th. 4.1. Recall the development (11) of $b(z,t)$ near $z_\star$ and the spectral factorization (12) of the matrix D. To begin with, it will be convenient to make the variable change $y = Q^{-1}(z - z_\star)$, and set

$$h(y,t) = Q^{-1}b(Qy + z_\star, t) = \Lambda y + \tilde{e}(y,t),$$

with $\tilde{e}(y,t) = Q^{-1}e(Qy + z_\star, t)$, in such a way that our stochastic algorithm is rewritten as

$$y_{n+1} = y_n + \gamma_{n+1}h(y_n, \tau_n) + \gamma_{n+1}\tilde{\eta}_{n+1} + \gamma_{n+1}\tilde{\rho}_{n+1}$$

where $\tilde{\eta}_n$ is as in the statement of the theorem and $\tilde{\rho}_n = Q^{-1}\rho_n$. Observe that the assumptions on the function e in the statement of the theorem remain true for $\tilde{e}$ with $z_\star$ replaced by zero.

If the matrix Λ has only eigenvalues with (strictly) positive real parts, *i.e.*, $d^- = 0$, then we can apply Prop. 7.4 to the sequence (z_n). Henceforth, we deal with the more complicated case where $d^- > 0$.

Apply Prop. 7.3 to h to obtain $\tilde{h}$ and σ, t_1 respectively small and large enough and $w : \mathbb{R}^{d^-} \times \mathbb{I}_1 \to \mathbb{R}^{d^+}$ where $\mathbb{I}_1 := [t_1, +\infty)$. By Assumption *iv)* of Th. 4.1 and Prop. 7.3 we can choose $\sigma_1 \leq \sigma$ such that Eq. (80) and Eq. (81) hold. Now, given $p \in \mathbb{N}$, let us define the event

$$E_p = [\forall n \geq p, \; \|y_n\| < \sigma_1, \tau_n \in \mathbb{I}_1].$$

On E_p, it holds that $h(y_n, \tau_n) = \tilde{h}(y_n, \tau_n)$ and

$$\forall n \geq p, \quad y_{n+1} = y_n + \gamma_{n+1}h(y_n, \tau_n) + \gamma_{n+1}\tilde{\eta}_{n+1} + \gamma_{n+1}\tilde{\rho}_{n+1}$$
$$= \begin{bmatrix} y_n^- \\ y_n^+ \end{bmatrix} + \gamma_{n+1} \begin{bmatrix} h^-(y_n, \tau_n) \\ h^+(y_n, \tau_n) \end{bmatrix} + \gamma_{n+1} \begin{bmatrix} \tilde{\eta}_{n+1}^- \\ \tilde{\eta}_{n+1}^+ \end{bmatrix} + \gamma_{n+1} \begin{bmatrix} \tilde{\rho}_{n+1}^- \\ \tilde{\rho}_{n+1}^+ \end{bmatrix}$$
$$(82)$$

where h is partitioned as in (67), and where $\tilde{\eta}_n^\pm, \tilde{\rho}_n^\pm \in \mathbb{R}^{d^\pm}$. Note that, by Prop. 7.3 and Assumptions *vi)* and *vii)* on the sequence (η_n), we can choose σ, t_1 respectively small and large enough s.t.

$$\liminf \mathbb{E}[\|\tilde{\eta}_{n+1}^+\|^2 |\mathscr{F}_n] \mathbb{1}_{E_p}(y_n)$$
$$- 2\limsup \mathbb{E}[\|\partial_1 w(y_n^-, \tau_n)\tilde{\eta}_{n+1}^-\|^2 |\mathscr{F}_n] \mathbb{1}_{E_p}(y_n) > \frac{c^2}{2}. \quad (83)$$

This inequality will be important in the end of our proof. Let t be in the interior of $\mathbb{I}_1$, and let $y = (y^-, y^+)$ be in a neighborhood of 0. Make the variable change $(y^-, y^+) \mapsto (u^-, u^+)$ with

$$u^+ = y^+ - w(y^-, t),$$
$$u^- = y^-,$$

where w is the function defined in the statement of Prop. 7.3, and let

$$W(u^-, u^+, t) = h^+(y, t) - \partial_1 w(y^-, t)h^-(y, t) - \partial_2 w(y^-, t)$$
$$= h^+((u^-, u^+ + w(u^-, t)), t)$$
$$- \partial_1 w(u^-, t)h^-((u^-, u^+ + w(u^-, t)), t) - \partial_2 w(u^-, t).$$

By Prop. 7.3 and Lem. 7.2, it holds that $W(u^-, 0, t) = 0$. Moreover, $W(u^-, \cdot, t) \in \mathcal{C}^1$ by the assumptions on h. Therefore, writing $y(r) = (u^-, ru^+ + w(u^-, t))$ for $r \in [0, 1]$, and using the decomposition (67), we get that

$$
\begin{aligned}
W(u^-, u^+, t) &= \int_0^1 \partial_2 W(u^-, ru^+, t) u^+ \, dr \\
&= \Lambda^+ u^+ \\
&\quad + \int_0^1 \left(\partial_1 \varepsilon^+(y(r), t) \begin{bmatrix} 0 \\ I_{d^+} \end{bmatrix} - \partial_1 w(u^-, t) \partial_1 \varepsilon^-(y(r), t) \begin{bmatrix} 0 \\ I_{d^+} \end{bmatrix} \right) u^+ dr.
\end{aligned}
$$

We can also write $y(r) = (y^-, ry^+ + (1 - r)w(y^-, t))$. Recalling that $w(0, t) = 0$ and that $\|\partial_1 w(y^-, t)\|$ is bounded on $\mathbb{R}^{d^-} \times \mathbb{I}$, we get by the mean value inequality that $\|w(y^-, t)\| \le C \|y^-\|$ where $C > 0$ is a constant. Thus, $\|y(r)\| \le (1 + C) \|y\|$. Moreover, $\varepsilon(y, t) = Q^{-1} e(Qy, t)$ for $\|y\| < \sigma$. Thus, we get by (13) that $\|\partial_1 \varepsilon(y(r), t)\| \to 0$ as $(y, t) \to (0, \infty)$ uniformly in $r \in [0, 1]$. Using again the boundedness of $\|\partial_1 w(\cdot, \cdot)\|$, we eventually obtain that

$$
W(u^-, u^+, t) = \left(\Lambda^+ + \Delta(y, t) \right) u^+, \quad \text{with} \quad \lim_{(y,t) \to (0, \infty)} \Delta(y, t) = 0.
$$

On the event E_p, assume that $n \ge p$, and write

$$
u_n^+ = y_n^+ - w(y_n^-, \tau_n), \quad u_n^- = y_n^-,
$$

(see Eq. (82)). Choosing $\alpha_- > 0$ small enough so that the gap condition (65) is satisfied with $m = 2$, we have by Taylor's expansion

$$
\begin{aligned}
&w(y_{n+1}^-, \tau_{n+1}) - w(y_n^-, \tau_n) \\
&= w(y_{n+1}^-, \tau_{n+1}) - w(y_n^-, \tau_{n+1}) + w(y_n^-, \tau_{n+1}) - w(y_n^-, \tau_n) \\
&= \partial_1 w(y_n^-, \tau_{n+1})(y_{n+1}^- - y_n^-) + \gamma_{n+1} \partial_2 w(y_n^-, \tau_n) + \epsilon_{n+1} + \epsilon_{n+1}^\gamma,
\end{aligned}
$$

$$
\begin{aligned}
\text{with} \quad &\|\epsilon_{n+1}\| \le \sup_{y^- \in [y_n^-, y_{n+1}^-]} \left\| \partial_1^2 w(y^-, \tau_{n+1}) \right\| \left\| y_{n+1}^- - y_n^- \right\|^2, \\
\text{and} \quad &\left\| \epsilon_{n+1}^\gamma \right\| \le \sup_{\tau \in [\tau_n, \tau_{n+1}]} \left\| \partial_2^2 w(y_n^-, \tau) \right\| \gamma_{n+1}^2.
\end{aligned}
$$

Using this equation, we obtain

$$
\begin{aligned}
u_{n+1}^+ - u_n^+ &= \gamma_{n+1} W(u_n^-, u_n^+, \tau_n) + \gamma_{n+1} \left(\tilde{\eta}_{n+1}^+ - \partial_1 w(y_n^-, \tau_{n+1}) \tilde{\eta}_{n+1}^- \right) \\
&\quad + \gamma_{n+1} \left(\tilde{\rho}_{n+1}^+ - \partial_1 w(y_n^-, \tau_{n+1}) \tilde{\rho}_{n+1}^- \right) - \epsilon_{n+1} - \epsilon_{n+1}^\gamma \\
&\quad + \gamma_{n+1} \left(\partial_1 w(y_n^-, \tau_n) - \partial_1 w(y_n^-, \tau_{n+1}) \right) h^-(y_n, \tau_n),
\end{aligned}
$$

which leads to

$$
u_{n+1}^+ = u_n^+ + \gamma_{n+1} \left(\Lambda^+ + \Delta(y_n, \tau_n) \right) u_n^+ + \gamma_{n+1} \bar{\eta}_{n+1} + \gamma_{n+1} \bar{\rho}_{n+1}, \tag{84}
$$

with $\bar{\eta}_{n+1} = \tilde{\eta}_{n+1}^+ - \partial_1 w(y_n^-, \tau_n)\tilde{\eta}_{n+1}^-$ and

$$\bar{\rho}_{n+1} = \tilde{\rho}_{n+1}^+ - \partial_1 w(y_n^-, \tau_n)\tilde{\rho}_{n+1}^- - \mathbb{1}_{\gamma_{n+1}>0}\frac{\epsilon_{n+1} + \epsilon_{n+1}^\gamma}{\gamma_{n+1}}$$
$$+ \left(\partial_1 w(y_n^-, \tau_n) - \partial_1 w(y_n^-, \tau_{n+1})\right) h^-(y_n, \tau_n). \quad (85)$$

To finish the proof, it remains to check that the noise sequence satisfies the assumptions of Prop. 7.4 on the event $A_p = E_p \cap [y_n \to 0]$. In the remainder, C' will indicate some positive constant which can change from an inequality to another one.

First, we verify that $\sum_n \|\bar{\rho}_n\|^2 < \infty$ on A_p by controlling each one of the terms of $\bar{\rho}_n$. Combining the boundedness of $\partial_1 w(\cdot, \cdot)$ with the summability assumption $\sum_n \|\tilde{\rho}_{n+1}\|^2 \mathbb{1}_{z_n \in \mathcal{W}} < +\infty$ a.s., we immediately obtain on A_p that $\sum_n \|\tilde{\rho}_{n+1}^+ - \partial_1 w(y_n^-, \tau_n)\tilde{\rho}_{n+1}^-\|^2 < +\infty$ given our choice of σ. Moreover, it holds that $\left(\|\epsilon_{n+1}^\gamma\|/\gamma_{n+1}\right)^2 \leq C'\gamma_{n+1}^2$ by invoking Prop. 7.3. In addition, using the boundedness of $\partial_1^2 w(\cdot, \cdot)$, we can write

$$\mathbb{1}_{\gamma_{n+1}>0}\left\|\frac{\epsilon_{n+1}}{\gamma_{n+1}}\right\|^2 \leq \mathbb{1}_{\gamma_{n+1}>0}\frac{C'}{\gamma_{n+1}^2}\|y_{n+1} - y_n\|^4$$
$$\leq C'\gamma_{n+1}^2(\|h(y_n, \tau_n)\|^4 + \|\tilde{\eta}_{n+1}\|^4 + \|\tilde{\rho}_{n+1}\|^4).$$

A coupling argument (see [12, p. 401]) shows that we can simplify the condition $\limsup \mathbb{E}[\|\eta_{n+1}\|^4 \mid \mathscr{F}_n]\mathbb{1}_{z_n \in \mathcal{W}} < \infty$ to $\mathbb{E}[\|\eta_{n+1}\|^4 \mid \mathscr{F}_n]\mathbb{1}_{z_n \in \mathcal{W}} < C'$. The latter condition implies that $\mathbb{E}[\mathbb{1}_{A_p}\sum_n \gamma_{n+1}^2 \|\eta_{n+1}\|^4] \leq \sum_n C'\gamma_{n+1}^2$, and therefore $\sum_n \gamma_{n+1}^2 \|\eta_{n+1}\|^4 \mathbb{1}_{A_p} < +\infty$ a.s. As a consequence, noticing also the boundedness of $(h(y_n, \tau_n))$ and $(\tilde{\rho}_n)$ on A_p, we deduce that $\sum_n \mathbb{1}_{\gamma_{n+1}>0}\left\|\frac{\epsilon_{n+1}}{\gamma_{n+1}}\right\|^2 < +\infty$ on A_p. We now briefly control the last term of $\bar{\rho}_n$. By the mean value inequality, we obtain that

$$\left\|\left(\partial_1 w(y_n^-, \tau_n) - \partial_1 w(y_n^-, \tau_{n+1})\right) h^-(y_n, \tau_n)\right\|$$
$$\leq \gamma_{n+1}\sup_{(y^-,t)}\left\|\partial_2\partial_1 w(y^-, t)\right\| \|h^-(y_n, \tau_n)\| \leq C'\gamma_{n+1},$$

where the last inequality stems from Prop. 7.3-Eq. (80) together with the boundedness of the sequence $(h(y_n, \tau_n))$. In view of Eq. (85) and the above estimates, we deduce that $\sum_n \|\bar{\rho}_{n+1}\|^2 \mathbb{1}_{A_p} < +\infty$ a.s. on A_p.

We verify the remaining conditions on the noise sequence $(\bar{\eta}_n)$. We can easily remark that $\mathbb{E}[\bar{\eta}_{n+1}|\mathscr{F}_n] = 0$ and $\|\bar{\eta}_{n+1}\| \leq C' \|\eta_{n+1}\|$ on A_p. Hence, $\limsup \mathbb{E}[\|\bar{\eta}_{n+1}\|^4 \mid \mathscr{F}_n]\mathbb{1}_{z_n \in \mathcal{W}} < \infty$. The last condition, meaning that the noise is exciting enough, stems from noting that

$$2\liminf \mathbb{E}[\|\bar{\eta}_{n+1}\|^2 \mid \mathscr{F}_n]\mathbb{1}_{A_p} \geq \liminf \mathbb{E}[\|\tilde{\eta}_{n+1}^+\|^2 \mid \mathscr{F}_n]\mathbb{1}_{A_p}$$
$$- 2\limsup \mathbb{E}[\|\partial_1 w(y_n^-, \tau_n)\tilde{\eta}_{n+1}^-\|^2 \mid \mathscr{F}_n]\mathbb{1}_{A_p}$$

$$> \frac{c^2}{2},$$

where we used our choice of σ, t_1 and Eq. (83).

Noticing that $[y_n \to 0] \subset [\Delta(y_n, \tau_n) \to 0]$, we can now apply Prop. 7.4 to the sequence (u_n^+) (see Eq. (84)) with $A = A_p$ to obtain

$$\mathbb{P}\left(A_p \cap [u_n^+ \to 0]\right) = \mathbb{P}\left(A_p \cap [u_n^+ \to 0] \cap [\Delta(y_n, \tau_n) \to 0]\right) = 0.$$

We now show that $[y_n \to 0] \subset [u_n^+ \to 0]$, which amounts to prove that $w(y_n^-, \tau_n) \to 0$ given $y_n \to 0$. To that end, upon noting that $w(0, \cdot) \equiv 0$ and that $\partial_1 w(\cdot, \cdot)$ is bounded, it suffices to apply the mean value inequality, writing:

$$\|w(y_n^-, \tau_n)\| = \|w(y_n^-, \tau_n) - w(0, \tau_n)\| \leq \sup_{(y^-, t)} \|\partial_1 w(y^-, t)\| \, \|y_n^-\| \leq K \|y_n^-\|.$$

We have shown so far that $\mathbb{P}(A_p) = 0$. Since $y_n = Q^{-1} z_n$ and $[y_n \to 0] \subset \bigcup_{p \in \mathbb{N}} E_p$, we finally obtain that

$$\mathbb{P}[z_n \to 0] = \mathbb{P}[y_n \to 0] = \mathbb{P}\left(\bigcup_{p \in \mathbb{N}}([y_n \to 0] \cap E_p)\right) = \mathbb{P}\left(\bigcup_{p \in \mathbb{N}} A_p\right) = 0.$$

Th. 4.1 is proven.

7.3. Proofs for Section 4.2.1

7.3.1. Proof of Lem. 4.2

The matrix D coincides with $\nabla g_\infty(z_\star)$, where the function g_∞ is defined in (20). As such, its expression is immediate. Recalling that $p_\infty S(x_\star) - q_\infty v_\star = 0$, we get

$$
\begin{aligned}
&g(z,t) - D(z - z_\star) \\
&= \begin{bmatrix} \mathsf{p}(t)S(x) - \mathsf{q}(t)v - p_\infty \nabla S(x_\star)(x - x_\star) + q_\infty(v - v_\star) \\ \mathsf{h}(t)\nabla F(x) - \mathsf{r}(t)m - h_\infty \nabla^2 F(x_\star)(x - x_\star) + r_\infty m \\ -m\left((v + \varepsilon)^{-\frac{1}{2}} - (v_\star + \varepsilon)^{-\frac{1}{2}}\right) \end{bmatrix} \\
&= \begin{bmatrix} -\mathsf{q}(t) + q_\infty & 0 & (\mathsf{p}(t) - p_\infty)\nabla S(x_\star) \\ 0 & r_\infty - \mathsf{r}(t) & (\mathsf{h}(t) - h_\infty)\nabla^2 F(x_\star) \\ \frac{m}{2(v_\star + \varepsilon)^{\frac{3}{2}}} & 0 & 0 \end{bmatrix} \begin{bmatrix} v - v_\star \\ m \\ x - x_\star \end{bmatrix} \\
&\quad + \begin{bmatrix} \mathsf{p}(t)(S(x) - S(x_\star) - \nabla S(x_\star)(x - x_\star)) \\ \mathsf{h}(t)(\nabla F(x) - \nabla^2 F(x_\star)(x - x_\star)) \\ -m \odot \left(\frac{1}{\sqrt{v+\varepsilon}} - \frac{1}{\sqrt{v_\star + \varepsilon}} + \frac{v - v_\star}{2(v_\star + \varepsilon)^{\frac{3}{2}}}\right) \end{bmatrix} + \begin{bmatrix} \mathsf{p}(t)S(x_\star) - \mathsf{q}(t)v_\star \\ 0 \\ 0 \end{bmatrix} \\
&\triangleq e(z,t) + c(t).
\end{aligned}
$$

Under the assumptions made, it is easy to see that the function $e(z,t)$ has the properties required in the statement of Th. 4.1.

7.3.2. Proof of Th. 4.3

Consider the matrix D defined in the statement of Lem. 4.2. A spectral analysis of this matrix as regards its eigenvalues with positive real parts is done in the following lemma.

Lemma 7.5. Let D be the matrix provided in the statement of Lem. 4.2. Each eigenvalue ζ of the matrix D such that $\Re\zeta > 0$ is real, and its algebraic and geometric multiplicities are equal. Moreover, there is a one-to-one correspondence φ between these eigenvalues and the negative eigenvalues of $V^{\frac{1}{2}}\nabla^2 F(x_\star)V^{\frac{1}{2}}$. Let d^+ be the dimension of the eigenspace of $V^{\frac{1}{2}}\nabla^2 F(x_\star)V^{\frac{1}{2}}$ that is associated with its negative eigenvalues, let

$$
W = \begin{bmatrix} w_1 \\ \vdots \\ w_{d^+} \end{bmatrix} \in \mathbb{R}^{d^+ \times d}
$$

be a matrix which rows are independent eigenvectors of $V^{\frac{1}{2}}\nabla^2 F(x_\star)V^{\frac{1}{2}}$ that generate this eigenspace, and denote as $\beta_k < 0$ the eigenvalue associated with w_k. Then, the rows of the rank d^+-matrix

$$
A^+ = \begin{bmatrix} 0_{d^+\times d}, & WV^{\frac{1}{2}}, & -\operatorname{diag}(r_\infty + \varphi^{-1}(\beta_k))WV^{-\frac{1}{2}} \end{bmatrix} \in \mathbb{R}^{d^+ \times 3d}
$$

generate the left eigenspace of D, the row k being an eigenvector for the eigenvalue $\varphi^{-1}(\beta_k)$.

Proof. It is obvious that the block lower-triangular matrix D has d eigenvalues equal to $-q_\infty$ and $2d$ eigenvalues which are those of the sub-matrix

$$
\widetilde{D} = \begin{bmatrix} -r_\infty I_d & h_\infty \nabla^2 F(x_\star) \\ -V & 0 \end{bmatrix}.
$$

Given $\lambda \in \mathbb{C}$, we obtain by standard manipulations involving determinants that

$$
\det(\widetilde{D} - \lambda) = \det(\lambda(r_\infty + \lambda) + h_\infty V\nabla^2 F(x_\star))
$$
$$
= \det(\lambda(r_\infty + \lambda) + h_\infty V^{\frac{1}{2}}\nabla^2 F(x_\star)V^{\frac{1}{2}}).
$$

Denoting as $\{\beta_k\}_{k=1}^d$ the eigenvalues of $h_\infty V^{\frac{1}{2}}\nabla^2 F(x_\star)V^{\frac{1}{2}}$ counting the multiplicities, we obtain from the last equation that the eigenvalues of $\widetilde{D}$ are the solutions of the second order equations

$$
\lambda^2 + r_\infty\lambda + \beta_k = 0, \quad k = 1, \ldots, d.
$$

The product of the roots of such an equation is β_k, and their sum is $-r_\infty \leq 0$. Thus, denoting as $\zeta_{k,1}$ and $\zeta_{k,2}$ these roots, it is easy to see that if $\beta_k \geq 0$, then $\Re\zeta_{k,1}, \Re\zeta_{k,2} \leq 0$, while if $\beta_k < 0$, then both $\zeta_{k,i}$ are real, and only one of them is positive. Thus, we have so far shown that the eigenvalues of D which real parts

are positive are themselves real, and there is a one-to-one map φ from the set of positive eigenvalues of D to the set of negative eigenvalues of $V^{\frac{1}{2}}\nabla^2 F(x_\star)V^{\frac{1}{2}}$. Moreover, the algebraic multiplicity of the eigenvalue $\zeta > 0$ of D is equal to the multiplicity of $\varphi(\zeta)$.

Let us now turn to the left (row) eigenvectors of D that correspond to these eigenvalues. To that end, we shall solve the equation

$$uD = \zeta u \quad \text{with } u = [0, u_1, u_2], \quad u_{1,2} \in \mathbb{R}^{1 \times d}, \tag{86}$$

for a given eigenvalue $\zeta > 0$ of D. Developing this equation, we get

$$-r_\infty u_1 - u_2 V = \zeta u_1, \quad h_\infty u_1 \nabla^2 F(x_\star) = \zeta u_2.$$

If we now write $\tilde{u}_1 = u_1 V^{-\frac{1}{2}}$ and $\tilde{u}_2 = u_2 V^{\frac{1}{2}}$, this system becomes

$$-r_\infty \tilde{u}_1 - \tilde{u}_2 = \zeta \tilde{u}_1, \quad h_\infty \tilde{u}_1 V^{\frac{1}{2}}\nabla^2 F(x_\star)V^{\frac{1}{2}} = \zeta \tilde{u}_2,$$

or, equivalently,

$$\tilde{u}_2 = -(r_\infty + \zeta)\tilde{u}_1, \quad \tilde{u}_1\left(\zeta^2 + r_\infty\zeta + h_\infty V^{\frac{1}{2}}\nabla^2 F(x_\star)V^{\frac{1}{2}}\right) = 0,$$

which shows that $\tilde{u}_1$ is a left eigenvector of $V^{\frac{1}{2}}\nabla^2 F(x_\star)V^{\frac{1}{2}}$ associated with the eigenvalue $\varphi(\zeta)$. What's more, assume that r is the multiplicity of $\varphi(\zeta)$, and, without generality loss, that the submatrix $W_{r,\cdot}$ made of the first r rows of W generates the left eigenspace of $\varphi(\zeta)$. Then, the matrix

$$\begin{bmatrix} 0_{r \times d} & W_{r\cdot}V^{\frac{1}{2}} & -(r_\infty + \zeta)W_{r\cdot}V^{-\frac{1}{2}} \end{bmatrix}$$

is a r-rank matrix which rows are independent left eigenvectors that generate the left eigenspace of D for the eigenvalue ζ. In particular, the algebraic and geometric multiplicities of this eigenvalue are equal. The same argument can be applied to the other positive eigenvalues of D. □

We now have all the elements to prove Th. 4.3. Recall Eq. (14):

$$z_{n+1} = z_n + \gamma_{n+1}b(z_n, \tau_n) + \gamma_{n+1}\eta_{n+1} + \gamma_{n+1}\rho_{n+1},$$

where $b(z,t) = g(z,t) - c(t) = D(z - z_\star) + e(z,t)$ and $\rho_n = c(\tau_{n-1}) + \tilde{\rho}_n$. With these same notations, we check that Assumptions *i)–vi)* in the statement of Th. 4.1 are satisfied. The function $e(z,t)$ satisfies Assumptions *i)–iv)* by Lem. 4.2. We now verify that the sequence (ρ_n) fulfills Assumption *v)*. First, observe that $\sum_n \|c(\tau_n)\|^2 < \infty$ under Assumption 4.3-*i)*. Then, we control the second term $(\tilde{\rho}_n)$. After straightforward derivations, one can show the existence of a positive constant C (depending only on ε and a neighborhood $\mathcal{W}$ of $z_\star$) such that

$$\|\tilde{\rho}_{n+1}\|^2 \mathbb{1}_{z_n \in \mathcal{W}} \leq C(\|m_n - m_{n+1}\|^2 + \|v_{n+1} - v_n\|^2)\mathbb{1}_{z_n \in \mathcal{W}}. \tag{87}$$

Using the boundedness of the sequences (h_n) and (r_n) together with the update rule of m_n and Assumption 4.3-*iii)*, there exists a positive constant C' independent of n (which may change from an inequality to another) such that

$$\mathbb{E}\left[\|m_n - m_{n+1}\|^2 \mathbb{1}_{z_n \in \mathcal{W}}\right] \leq \gamma_{n+1}^2 C' \mathbb{E}\left[(1 + \mathbb{E}_\xi\left[\|\nabla f(x_n, \xi)\|^2\right])\mathbb{1}_{z_n \in \mathcal{W}}\right] \leq C' \gamma_{n+1}^2. \tag{88}$$

A similar result holds for $\mathbb{E}\left[\|v_n - v_{n+1}\|^2 \mathbb{1}_{z_n \in \mathcal{W}}\right]$ following the same arguments. In view of Eqs. (87)-(88) and the assumption $\sum_n \gamma_{n+1}^2 < +\infty$, it holds that $\mathbb{E}\left[\sum_n \|\tilde{\rho}_{n+1}\|^2 \mathbb{1}_{z_n \in \mathcal{W}}\right] < +\infty$. Therefore, $\sum_n \|\tilde{\rho}_{n+1}\|^2 \mathbb{1}_{z_n \in \mathcal{W}} < +\infty$ a.s., which completes our verification of condition *v)* of Th. 4.1. Assumption *vi)* follows from condition 4.3-*iii)*. Finally, let us make Assumption *vii)* of Th. 4.1 more explicit. Partitioning the matrix Q^{-1} as $Q^{-1} = \begin{bmatrix} B^- \\ B^+ \end{bmatrix}$ where $B^\pm$ has $d^\pm$ rows, Lem. 7.5 shows that the row spaces of B^+ and A^+ are the same, which implies that Assumption *vii)* can be rewritten equivalently as $\mathbb{E}[\|A^+ \eta_{n+1}\|^2 \mid \mathcal{F}_n]\mathbb{1}_{z_n \in \mathcal{W}} \geq c^2 \mathbb{1}_{z_n \in \mathcal{W}}$. By inspecting the form of η_n provided by Eq. (28) (written as a column vector), one can readily check that Assumption 4.3-*iv)* implies Assumption *vii)* of Th. 4.1 for a small enough neighborhood $\mathcal{W}$, using the continuity of the covariance matrix $V^{\frac{1}{2}} \mathbb{E}_\xi (\nabla f(x, \xi) - \nabla F(x))(\nabla f(x, \xi) - \nabla F(x))^{\mathrm{T}} V^{\frac{1}{2}}$ when x is near $x_\star$.

7.4. Proof of Th. 4.4

As mentioned in Section 4.2.2, the proof of Th. 4.4 is almost identical to the one of Th. 4.3. We point out the main differences here. In Lem. 4.2, replace D by $\tilde{D} = \begin{bmatrix} 0 & h_\infty \nabla^2 F(x_\star) \\ -I_d & 0 \end{bmatrix}$ and set $c(t) = 0$. Then, in Lem. 7.5, replace the matrix $V^{1/2} \nabla^2 F(x_\star) V^{1/2}$ by the Hessian $\nabla^2 F(x_\star)$.

Acknowledgments

References

[1] ALACAOGLU, A., MALITSKY, Y. and CEVHER, V. (2020). Convergence of adaptive algorithms for weakly convex constrained optimization. *arXiv preprint arXiv:2006.06650*.

[2] ALACAOGLU, A., MALITSKY, Y., MERTIKOPOULOS, P. and CEVHER, V. (2020). A new regret analysis for Adam-type algorithms. In *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. SINGH, eds.). *Proceedings of Machine Learning Research* **119** 202–210.

[3] ALVAREZ, F. (2000). On the minimizing property of a second order dissipative system in Hilbert spaces. *SIAM Journal on Control and Optimization* **38** 1102–1119. MR1760062

[4] ASSRAN, M. and RABBAT, M. (2020). On the convergence of Nesterov's accelerated gradient method in stochastic settings. In *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. SINGH, eds.). *Proceedings of Machine Learning Research* **119** 410–420.

[5] ATTOUCH, H., GOUDOU, X. and REDONT, P. (2000). The heavy ball with friction method, I. The continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. *Communications in Contemporary Mathematics* **2** 1–34. MR1753136

[6] ATTOUCH, H., CHBANI, Z., PEYPOUQUET, J. and REDONT, P. (2018). Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming* **168** 123–175. MR3767743

[7] AUJOL, J. F., DOSSAL, C. and RONDEPIERRE, A. (2019). Optimal convergence rates for Nesterov acceleration. *SIAM Journal on Optimization* **29** 3131–3153. MR4043106

[8] BARAKAT, A. and BIANCHI, P. (2021). Convergence and dynamical behavior of the ADAM algorithm for nonconvex stochastic optimization. *SIAM Journal on Optimization* **31** 244–274. MR4199255

[9] BELOTTO DA SILVA, A. and GAZEAU, M. (2020). A general system of differential equations to model first-order adaptive algorithms. *Journal of Machine Learning Research* **21** 1–42. MR4138113

[10] BENAÏM, M. (1999). Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités, XXXIII. Lecture Notes in Math.* **1709** 1–68. Springer, Berlin. MR1767993 (2001d:62081)

[11] BENAÏM, M. and HIRSCH, M. W. (1996). Asymptotic pseudotrajectories and chain recurrent flows, with applications. *J. Dynam. Differential Equations* **8** 141–176. MR1388167 (97d:58165)

[12] BRANDIÈRE, O. and DUFLO, M. (1996). Les algorithmes stochastiques contournent-ils les pièges? *Ann. Inst. H. Poincaré Probab. Statist.* **32** 395–427. MR1387397

[13] CABOT, A., ENGLER, H. and GADAT, S. (2009). On the long time behavior of second order differential equations with asymptotically small dissipation. *Transactions of the American Mathematical Society* **361** 5983–6017. MR2529922

[14] CHEN, J., ZHOU, D., TANG, Y., YANG, Z. and GU, Q. (2018). Closing the generalization gap of adaptive gradient methods in training deep neural networks. *arXiv preprint arXiv:1806.06763*.

[15] CHEN, X., LIU, S., SUN, R. and HONG, M. (2019). On the convergence of a class of Adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*.

[16] DALEC′KIĬ, J. L. and KREĬN, M. G. (1974). *Stability of Solutions of Differential Equations in Banach Space.* American Mathematical Society, Providence, R.I. Translated from the Russian by S. Smith, Translations of Math-

ematical Monographs, Vol. 43. MR0352639

[17] De, S., Mukherjee, A. and Ullah, E. (2018). Convergence guarantees for RMSProp and ADAM in non-convex optimization and their comparison to Nesterov acceleration on autoencoders. *arXiv preprint arXiv:1807.06766*.

[18] Défossez, A., Bottou, L., Bach, F. and Usunier, N. (2020). A simple convergence proof of Adam and Adagrad. *arXiv preprint arXiv:2003.02395*.

[19] Delyon, B., Lavielle, M. and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics* 94–128. MR1701103

[20] Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Singh, A. and Poczos, B. (2017). Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.) 1067–1077. Curran Associates, Inc.

[21] Duchi, J., Hazan, E. and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* **12** 2121–2159. MR2825422

[22] Gadat, S. and Gavra, I. (2020). Asymptotic study of stochastic adaptive algorithm in non-convex landscape. *arXiv preprint arXiv:2012.05640*.

[23] Gadat, S., Panloup, F. and Saadane, S. (2018). Stochastic heavy ball. *Electron. J. Stat.* **12** 461–529. MR3765604

[24] Haraux, A. (1991). *Systèmes Dynamiques Dissipatifs et Applications* **17**. Masson. MR1084372

[25] Hartman, P. (2002). *Ordinary Differential Equations*, Second ed. Society for Industrial and Applied Mathematics. MR1929104

[26] Horn, R. A. and Johnson, C. R. (1994). *Topics in Matrix Analysis*. Cambridge University Press, Cambridge Corrected reprint of the 1991 original. MR1288752

[27] Jin, C., Ge, R., Netrapalli, P., Kakade, S. M. and Jordan, M. I. (2017). How to Escape Saddle Points Efficiently. (D. Precup and Y. W. Teh, eds.). *Proceedings of Machine Learning Research* **70** 1724–1732. PMLR.

[28] Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*, second ed. Springer-Verlag. MR1121940

[29] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

[30] Kloeden, P. E. and Rasmussen, M. (2011). *Nonautonomous Dynamical Systems. Mathematical Surveys and Monographs* **176**. American Mathematical Society, Providence, RI. MR2808288

[31] Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I. and Recht, B. (2019). First-order methods almost always avoid strict saddle points. *Math. Program.* **176** 311–337. MR3960812

[32] Mai, V. V. and Johansson, M. (2020). Convergence of a Stochastic Gradient Method with Momentum for Nonsmooth Nonconvex Optimization.

Proceedings of Machine Learning Research. PMLR.

[33] MERTIKOPOULOS, P., HALLAK, N., KAVIS, A. and CEVHER, V. (2020). On the almost sure convergence of stochastic gradient descent in nonconvex problems. In *Advances in Neural Information Processing Systems* (H. LAROCHELLE, M. RANZATO, R. HADSELL, M. F. BALCAN and H. LIN, eds.) **33** 1117–1128. Curran Associates, Inc.

[34] MÉTIVIER, M. and PRIOURET, P. (1987). Théorèmes de convergence presque sûre pour une classe d'algorithmes stochastiques à pas décroissant. *Probability Theory and Related Fields* **74** 403–428. MR0873887

[35] PANAGEAS, I. and PILIOURAS, G. (2017). Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. In *ITCS*. MR3754926

[36] PANAGEAS, I., PILIOURAS, G. and WANG, X. (2019). First-order methods almost always avoid saddle points: The case of vanishing step-sizes. In *Advances in Neural Information Processing Systems* **32** 6474–6483.

[37] PELLETIER, M. (1998). Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. *Annals of Applied Probability* 10–44. MR1620405

[38] PEMANTLE, R. (1990). Nonconvergence to unstable points in urn models and stochastic approximations. *Ann. Probab.* **18** 698–712. MR1055428

[39] PÖTZSCHE, C. and RASMUSSEN, M. (2006). Taylor approximation of integral manifolds. *J. Dynam. Differential Equations* **18** 427–460. MR2229984 MR2229984

[40] ROBBINS, H. and SIEGMUND, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics* 233–257. Academic Press, New York. MR0343355

[41] SU, W., BOYD, S. and CANDÈS, E. J. (2016). A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *J. Mach. Learn. Res.* **17** Paper No. 153, 43. MR3555044

[42] TIELEMAN, T. and HINTON, G. (2012). Lecture 6.e-rmsprop: Divide the gradient by a running average of its recent magnitude. *Coursera: Neural networks for machine learning* 26–31.

[43] YAN, Y., YANG, T., LI, Z., LIN, Q. and YANG, Y. (2018). A unified analysis of stochastic momentum methods for deep learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* 2955–2961.

[44] ZAHEER, M., REDDI, S., SACHAN, D., KALE, S. and KUMAR, S. (2018). Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems* 9793–9803.

[45] ZHOU, D., TANG, Y., YANG, Z., CAO, Y. and GU, Q. (2018). On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*.

[46] ZOU, F., SHEN, L., JIE, Z., ZHANG, W. and LIU, W. (2019). A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 11127–11135.