Convergence and Dynamical Behavior of the Adam Algorithm for Non Convex Stochastic Optimization

Anas Barakat, Pascal Bianchi

CONVERGENCE AND DYNAMICAL BEHAVIOR OF THE ADAM ALGORITHM FOR NON-CONVEX STOCHASTIC OPTIMIZATION

ANAS BARAKAT AND PASCAL BIANCHI *

Abstract. ADAM is a popular variant of stochastic gradient descent for finding a local minimizer of a function. In the constant stepsize regime, assuming that the objective function is differentiable and non-convex, we establish the convergence in the long run of the iterates to a stationary point under a stability condition. The key ingredient is the introduction of a continuous-time version of ADAM, under the form of a non-autonomous ordinary differential equation. This continuous-time system is a relevant approximation of the ADAM iterates, in the sense that the interpolated ADAM process converges weakly towards the solution to the ODE. The existence and the uniqueness of the solution are established. We further show the convergence of the solution towards the critical points of the objective function and quantify its convergence rate under a Łojasiewicz assumption. Then, we introduce a novel decreasing stepsize version of ADAM. Under mild assumptions, it is shown that the iterates are almost surely bounded and converge almost surely to critical points of the objective function. Finally, we analyze the fluctuations of the algorithm by means of a conditional central limit theorem.

Key words. Stochastic approximation, Dynamical systems, Adaptive gradient methods.

AMS subject classifications. 62L20, 65K05, 34A12, 37C60

1. Introduction. Consider the problem of finding a local minimizer of the expectation $F(x) := \mathbb{E}(f(x,\xi))$ w.r.t. $x \in \mathbb{R}^d$, where $f(\,.\,,\xi)$ is a possibly non-convex function depending on some random variable ξ. The distribution of ξ is assumed unknown, but revealed online by the observation of iid copies $(\xi_n : n \geq 1)$ of the r.v. ξ. Stochastic gradient descent (SGD) is the most classical algorithm to search for such a minimizer. Variants of SGD which include an inertial term have also become very popular. In these methods, the update rule depends on a parameter called the *learning rate*, which is generally assumed constant or vanishing. These algorithms, although widely used, have at least two limitations. First, the choice of the learning rate is generally difficult; large learning rates result in large fluctuations of the estimate, whereas small learning rates induce slow convergence. Second, a common learning rate is used for every coordinate despite the possible discrepancies in the values of the gradient vector's coordinates.

To alleviate these limitations, the popular ADAM algorithm [18] adjusts the learning rate coordinate-wise, as a function of the past values of the squared gradient vectors' coordinates. The algorithm thus combines the assets of inertial methods with an adaptive per-coordinate learning rate selection. Finally, the algorithm includes a so-called *bias correction* step. Acting on the current estimate of the gradient vector, this step is especially useful during the early iterations.

Despite the growing popularity of the algorithm, only few works investigate its behavior from a theoretical point of view (see the discussion in Section 6). The present paper studies the convergence of ADAM from a dynamical system viewpoint.

Contributions

- We introduce a continuous-time version of the ADAM algorithm under the form of a non-autonomous ordinary differential equation (ODE). Building on the existence of an explicit Lyapunov function for the ODE, we show the existence of a unique

*LTCI, Télécom Paris, Institut polytechnique de Paris, France. (firstname.name@telecom-paristech.fr)

global solution to the ODE. This first result turns out to be non-trivial due to the irregularity of the vector field. We then establish the convergence of the continuous-time ADAM trajectory to the set of critical points of the objective function F. The proposed continuous-time version of ADAM provides useful insights on the effect of the bias correction step. It is shown that, close to the origin, the objective function F is non-increasing along the ADAM trajectory, suggesting that early iterations of ADAM can only improve the initial guess.

- Under a Łojasiewicz-type condition, we prove that the solution to the ODE converges to a single critical point of the objective function F. We provide convergence rates in this case.
- In discrete time, we first analyze the ADAM iterates in the constant stepsize regime as originally introduced in [18]. In this work, it is shown that the discrete-time ADAM iterates shadow the behavior of the non-autonomous ODE in the asymptotic regime where the stepsize parameter γ of ADAM is small. More precisely, we consider the interpolated process $\mathsf{z}^\gamma(t)$ which consists of a piecewise linear interpolation of the ADAM iterates. The random process z^γ is indexed by the parameter γ, which is assumed constant during the whole run of the algorithm. In the space of continuous functions on $[0,+\infty)$ equipped with the topology of uniform convergence on compact sets, we establish that z^γ converges in probability to the solution to the non-autonomous ODE when γ tends to zero.
- Under a stability condition, we prove the asymptotic ergodic convergence of the probability of the discrete-time ADAM iterates to approach the set of critical points of the objective function in the doubly asymptotic regime where $n \to \infty$ then $\gamma \to 0$.
- Beyond the original constant stepsize ADAM, we propose a decreasing stepsize version of the algorithm. We provide sufficient conditions ensuring the stability and the almost sure convergence of the iterates towards the critical points of the objective function.
- We establish a convergence rate of the stochastic iterates of the decreasing stepsize algorithm under the form of a conditional central limit theorem.

We claim that our analysis can be easily extended to other adaptive algorithms such as e.g. RMSPROP or ADAGRAD [23, 12] and AMSGRAD (see Section 6).

The paper is organized as follows. In Section 2, we present the ADAM algorithm and the main assumptions. Our main results are stated in Sections 3 to 5. We provide a review of related works in Section 6. The rest of the paper addresses the proofs of our results (Sections 7 to 9).

Notations. If x, y are two vectors on $\mathbb{R}^d$ for some $d \geq 1$, we denote by $x \odot y$, $x^{\odot 2}$, x/y, $|x|$, $\sqrt{|x|}$ the vectors on $\mathbb{R}^d$ whose i-th coordinates are respectively given by $x_i y_i$, x_i^2, x_i/y_i, $|x_i|$, $\sqrt{|x_i|}$. Inequalities of the form $x \leq y$ are read componentwise. Denote by $\|\cdot\|$ the standard Euclidean norm. For any vector $v \in (0,+\infty)^d$, write $\|x\|_v^2 = \sum_i v_i x_i^2$. Notation A^T represents the transpose of a matrix A. If $z \in \mathbb{R}^d$ and A is a non-empty subset of $\mathbb{R}^d$, we use the notation $\mathsf{d}(z,A) := \inf\{\|z - z'\| : z' \in A\}$. If A is a set, we denote by $\mathbb{1}_A$ the function equal to one on that set and to zero elsewhere. We denote by $C([0,+\infty),\mathbb{R}^d)$ the space of continuous functions from $[0,+\infty)$ to $\mathbb{R}^d$ endowed with the topology of uniform convergence on compact intervals.

2. The ADAM Algorithm.

2.1. Algorithm and Assumptions. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $(\Xi, \mathfrak{S})$ denote an other measurable space. Consider a measurable map $f : \mathbb{R}^d \times$

Algorithm 2.1 ADAM$(\gamma, \alpha, \beta, \varepsilon)$.

Initialization: $x_0 \in \mathbb{R}^d, m_0 = 0, v_0 = 0$.
for $n = 1$ **to** n_{iter} **do**
$m_n = \alpha m_{n-1} + (1-\alpha)\nabla f(x_{n-1}, \xi_n)$
$v_n = \beta v_{n-1} + (1-\beta)\nabla f(x_{n-1}, \xi_n)^{\odot 2}$
$\hat{m}_n = m_n/(1-\alpha^n)$ {bias correction step}
$\hat{v}_n = v_n/(1-\beta^n)$ {bias correction step}
$x_n = x_{n-1} - \gamma \hat{m}_n/(\varepsilon + \sqrt{\hat{v}_n})$.
end for

$\Xi \to \mathbb{R}$, where d is an integer. For a fixed value of ξ, the mapping $x \mapsto f(x, \xi)$ is supposed to be differentiable, and its gradient w.r.t. x is denoted by $\nabla f(x, \xi)$. Define $\mathcal{Z} := \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$, $\mathcal{Z}_+ := \mathbb{R}^d \times \mathbb{R}^d \times [0, +\infty)^d$ and $\mathcal{Z}_+^* := \mathbb{R}^d \times \mathbb{R}^d \times (0, +\infty)^d$. ADAM generates a sequence $z_n := (x_n, m_n, v_n)$ on $\mathcal{Z}_+$ given by Algorithm 2.1. It satisfies: $z_n = T_{\gamma,\alpha,\beta}(n, z_{n-1}, \xi_n)$, for every $n \geq 1$, where for every $z = (x, m, v)$ in $\mathcal{Z}_+$, $\xi \in \Xi$,

$$T_{\gamma,\alpha,\beta}(n, z, \xi) := \begin{pmatrix} x - \frac{\gamma(1-\alpha^n)^{-1}(\alpha m + (1-\alpha)\nabla f(x,\xi))}{\varepsilon + (1-\beta^n)^{-1/2}(\beta v + (1-\beta)\nabla f(x,\xi)^{\odot 2})^{1/2}} \\ \alpha m + (1-\alpha)\nabla f(x, \xi) \\ \beta v + (1-\beta)\nabla f(x, \xi)^{\odot 2} \end{pmatrix}. \tag{2.1}$$

Remark 2.1. The iterates z_n form a non-homogeneous Markov chain, because $T_{\gamma,\alpha,\beta}(n, z, \xi)$ depends on n. This is due to the so-called debiasing step, which consists of replacing m_n, v_n in Algorithm 2.1 by their "debiased" versions $\hat{m}_n, \hat{v}_n$. The motivation becomes clear when expanding the expression:

$$\hat{m}_n = \frac{m_n}{1-\alpha^n} = \frac{1-\alpha}{1-\alpha^n} \sum_{k=0}^{n-1} \alpha^k \nabla f(x_k, \xi_{k+1}).$$

From this equation, it is observed that, $\hat{m}_n$ forms a convex combination of the past gradients. This is unlike m_n, which may be small during the first iterations.

Assumption 2.2. The mapping $f : \mathbb{R}^d \times \Xi \to \mathbb{R}$ satisfies the following.

i) For every $x \in \mathbb{R}^d$, $f(x, .)$ is $\mathfrak{S}$-measurable.
ii) For almost every ξ, the map $f(., \xi)$ is continuously differentiable.
iii) There exists $x_* \in \mathbb{R}^d$ s.t. $\mathbb{E}(|f(x_*, \xi)|) < \infty$ and $\mathbb{E}(\|\nabla f(x_*, \xi)\|^2) < \infty$.
iv) For every compact subset $K \subset \mathbb{R}^d$, there exists $L_K > 0$ such that for every $(x, y) \in K^2$, $\mathbb{E}(\|\nabla f(x, \xi) - \nabla f(y, \xi)\|^2) \leq L_K^2 \|x - y\|^2$.

Under Assumption 2.2, it is an easy exercise to show that the mappings $F : \mathbb{R}^d \to \mathbb{R}$ and $S : \mathbb{R}^d \to \mathbb{R}^d$, given by:

$$F(x) := \mathbb{E}(f(x, \xi)) \quad \text{and} \quad S(x) := \mathbb{E}(\nabla f(x, \xi)^{\odot 2}) \tag{2.2}$$

are well defined; F is continuously differentiable and by Lebesgue's dominated convergence theorem, $\nabla F(x) = \mathbb{E}(\nabla f(x, \xi))$ for all x. Moreover, ∇F and S are locally Lipschitz continuous.

Assumption 2.3. F is coercive.

Assumption 2.4. For every $x \in \mathbb{R}^d$, $S(x) > 0$.

It follows from our assumptions that the set of critical points of F, denoted by

$$\mathcal{S} := \nabla F^{-1}(\{0\}),$$

is non-empty. Assumption 2.4 means that there is *no* point $x \in \mathbb{R}^d$ satisfying $\nabla f(x,\xi) = 0$ with probability one (w.p.1). This is a mild hypothesis in practice.

2.2. Asymptotic Regime. We address the constant stepsize regime, where γ is fixed along the iterations (the default value recommended in [18] is $\gamma = 0.001$). As opposed to the decreasing stepsize context, the sequence $z_n^\gamma := z_n$ *cannot* in general converge as n tends to infinity, in an almost sure sense. Instead, we investigate the asymptotic behavior of the family of processes $(n \mapsto z_n^\gamma)_{\gamma>0}$ indexed by γ, in the regime where $\gamma \to 0$. We use the so-called ODE method (see e.g. [5]). The interpolated process z^γ is the piecewise linear function defined on $[0,+\infty) \to \mathcal{Z}_+$ for all $t \in [n\gamma,(n+1)\gamma)$ by:

$$\mathsf{z}^\gamma(t) := z_n^\gamma + (z_{n+1}^\gamma - z_n^\gamma)\left(\frac{t-n\gamma}{\gamma}\right). \tag{2.3}$$

We establish the convergence in probability of the family of random processes $(\mathsf{z}^\gamma)_{\gamma>0}$ as γ tends to zero, towards a deterministic continuous-time system defined by an ODE. The latter ODE, which we provide below at Eq. (ODE), will be referred to as the continuous-time version of Adam.

Before describing the ODE, we need to be more specific about our asymptotic regime. As opposed to SGD, Adam depends on two parameters α, β, in addition to the stepsize γ. The paper [18] recommends choosing the constants α and β close to one (the default values $\alpha = 0.9$ and $\beta = 0.999$ are suggested). It is thus legitimate to assume that α and β tend to one, as γ tends to zero. We set $\alpha := \bar{\alpha}(\gamma)$ and $\beta := \bar{\beta}(\gamma)$, where $\bar{\alpha}(\gamma)$ and $\bar{\beta}(\gamma)$ converge to one as $\gamma \to 0$.

Assumption 2.5. The functions $\bar{\alpha} : \mathbb{R}_+ \to [0,1)$ and $\bar{\beta} : \mathbb{R}_+ \to [0,1)$ are s.t. the following limits exist:

$$a := \lim_{\gamma \downarrow 0} \frac{1-\bar{\alpha}(\gamma)}{\gamma}, \quad b := \lim_{\gamma \downarrow 0} \frac{1-\bar{\beta}(\gamma)}{\gamma}. \tag{2.4}$$

Moreover, $a > 0$, $b > 0$, and the following condition holds: $b \leq 4a$.

Note that the condition $b \leq 4a$ is compatible with the default settings recommended by [18]. In our model, we shall now replace the map $T_{\gamma,\alpha,\beta}$ by $T_{\gamma,\bar{\alpha}(\gamma),\bar{\beta}(\gamma)}$. Let $x_0 \in \mathbb{R}^d$ be fixed. For any fixed $\gamma > 0$, we define the sequence (z_n^γ) generated by Adam with a fixed stepsize $\gamma > 0$:

$$z_n^\gamma := T_{\gamma,\bar{\alpha}(\gamma),\bar{\beta}(\gamma)}(n, z_{n-1}^\gamma, \xi_n), \tag{2.5}$$

the initialization being chosen as $z_0^\gamma = (x_0, 0, 0)$.

3. Continuous-Time System.

3.1. Ordinary Differential Equation. In order to gain insight into the behavior of the sequence (z_n^γ) defined by (2.5), it is convenient to rewrite the Adam iterations under the following equivalent form, for every $n \geq 1$:

$$z_n^\gamma = z_{n-1}^\gamma + \gamma h_\gamma(n, z_{n-1}^\gamma) + \gamma \Delta_n^\gamma, \tag{3.1}$$

where we define for every $\gamma > 0$, $z \in \mathcal{Z}_+$,

$$h_\gamma(n,z) := \gamma^{-1}\mathbb{E}(T_{\gamma,\bar{\alpha}(\gamma),\bar{\beta}(\gamma)}(n,z,\xi) - z), \tag{3.2}$$

and where $\Delta_n^\gamma := \gamma^{-1}(z_n^\gamma - z_{n-1}^\gamma) - h_\gamma(n, z_{n-1}^\gamma)$. Note that (Δ_n^γ) is a martingale increment noise sequence in the sense that $\mathbb{E}(\Delta_n^\gamma | \mathcal{F}_{n-1}) = 0$ for all $n \geq 1$, where $\mathcal{F}_n$ stands for the σ-algebra generated by the r.v. $\xi_1, \dots, \xi_n$. Define the map $h : (0, +\infty) \times \mathcal{Z}_+ \to \mathcal{Z}$ for all $t > 0$, all $z = (x, m, v)$ in $\mathcal{Z}_+$ by:

$$h(t,z) = \begin{pmatrix} -\frac{(1-e^{-at})^{-1} m}{\varepsilon + \sqrt{(1-e^{-bt})^{-1} v}} \\ a(\nabla F(x) - m) \\ b(S(x) - v) \end{pmatrix}, \tag{3.3}$$

where a, b are the constants defined in Assumption 2.5. We prove that, for any fixed (t, z), the quantity $h(t, z)$ coincides with the limit of $h_\gamma(\lfloor t/\gamma \rfloor, z)$ as $\gamma \downarrow 0$. This remark along with Eq. (3.1) suggests that, as $\gamma \downarrow 0$, the interpolated process z^γ shadows the non-autonomous differential equation

$$\dot{z}(t) = h(t, z(t)) \,. \tag{ODE}$$

3.2. Existence, Uniqueness, Convergence. Since $h(\,.\,, z)$ is non-continuous at point zero for a fixed $z \in \mathcal{Z}_+$, and since $h(t, \,.\,)$ is not locally Lipschitz continuous for a fixed $t > 0$, the existence and uniqueness of the solution to (ODE) do not stem directly from off-the-shelf theorems.

Let x_0 be fixed. A continuous map $z : [0, +\infty) \to \mathcal{Z}_+$ is said to be a global solution to (ODE) with initial condition $(x_0, 0, 0)$ if z is continuously differentiable on $(0, +\infty)$, if Eq. (ODE) holds for all $t > 0$, and if $z(0) = (x_0, 0, 0)$.

THEOREM 3.1 (Existence and uniqueness). *Let Assumptions 2.2 to 2.5 hold true. There exists a unique global solution $z : [0, +\infty) \to \mathcal{Z}_+$ to (ODE) with initial condition $(x_0, 0, 0)$. Moreover, $z([0, +\infty))$ is a bounded subset of $\mathcal{Z}_+$.*

On the other hand, we note that a solution may not exist for an initial point(x_0, m_0, v_0) with arbitrary (non-zero) values of m_0, v_0.

THEOREM 3.2 (Convergence). *Let Assumptions 2.2 to 2.5 hold true. Assume that $F(\mathcal{S})$ has an empty interior. Let $z : t \mapsto (x(t), m(t), v(t))$ be the global solution to (ODE) with the initial condition $(x_0, 0, 0)$. Then, the set $\mathcal{S}$ is non-empty and $\lim_{t\to\infty} \mathsf{d}(x(t), \mathcal{S}) = 0$, $\lim_{t\to\infty} m(t) = 0$, $\lim_{t\to\infty} S(x(t)) - v(t) = 0$.*

Lyapunov function. The proof of Th. 3.1 relies on the existence of a Lyapunov function for the non-autonomous equation (ODE). Define $V : (0, +\infty) \times \mathcal{Z}_+ \to \mathbb{R}$ by

$$V(t,z) := F(x) + \frac{1}{2} \|m\|^2_{U(t,v)^{-1}} \,, \tag{3.4}$$

for every $t > 0$ and every $z = (x, m, v)$ in $\mathcal{Z}_+$, where $U : (0, +\infty) \times [0, +\infty)^d \to \mathbb{R}^d$ is the map given by:

$$U(t,v) := a(1 - e^{-at}) \left(\varepsilon + \sqrt{\frac{v}{1 - e^{-bt}}} \right) . \tag{3.5}$$

Then, $t \mapsto V(t, z(t))$ is decreasing if $z(\cdot)$ is the global solution to (ODE).

Cost decrease at the origin. As F itself is not a Lyapunov function for (ODE), there is no guarantee that $F(x(t))$ is decreasing w.r.t. t. Nevertheless, the statement holds at the origin. Indeed, it can be shown that $\lim_{t\downarrow 0} V(t, z(t)) = F(x_0)$ (see Prop. 7.6). As a consequence,

$$\forall t \geq 0, \;\; F(x(t)) \leq F(x_0) \,. \tag{3.6}$$

In other words, the (continuous-time) ADAM procedure *can only improve* the initial guess x_0. This is the consequence of the so-called bias correction steps in ADAM (see Algorithm 2.1). If these debiasing steps were deleted in the ADAM iterations, the early stages of the algorithm could degrade the initial estimate x_0.

Derivatives at the origin. The proof of Th. 3.1 reveals that the initial derivative is given by $\dot{x}(0) = -\nabla F(x_0)/(\varepsilon + \sqrt{S(x_0)})$ (see Lemma 7.3). In the absence of debiasing steps, the initial derivative $\dot{x}(0)$ would be a function of the initial parameters m_0, v_0, and the user would be required to tune these hyperparameters. No such tuning is required thanks to the debiasing step. When ε is small and when the variance of $\nabla f(x_0, \xi)$ is small (*i.e.*, $S(x_0) \simeq \nabla F(x_0)^{\odot 2}$), the initial derivative $\dot{x}(0)$ is approximately equal to $-\nabla F(x_0)/|\nabla F(x_0)|$. This suggests that in the early stages of the algorithm, the ADAM iterations are comparable to the *sign* variant of the gradient descent, the properties of which were discussed in previous works, see [3].

3.3. Convergence rates. In this paragraph, we establish the convergence to a single critical point of F and quantify the convergence rate, using the following assumption [19].

Assumption 3.3 (Łojasiewicz property). For any $x^* \in \mathcal{S}$, there exist $c > 0\,, \sigma > 0$ and $\theta \in (0, \frac{1}{2}]$ s.t.

$$\forall x \in \mathbb{R}^d \text{ s.t } \|x - x^*\| \leq \sigma\,, \quad \|\nabla F(x)\| \geq c|F(x) - F(x^*)|^{1-\theta}\,. \tag{3.7}$$

Assumption 3.3 holds for real-analytic functions and semialgebraic functions. We refer to [16, 1, 7] for a discussion and a review of applications. We will call any θ satisfying (3.7) for some $c, \sigma > 0$, as a Łojasiewicz exponent of f at x^*. The next result establishes the convergence of the function $x(t)$ generated by the ODE to a single critical point of f, and provides the convergence rate as a function of the Łojasiewicz exponent of f at this critical point. The proof is provided in subsection 7.4.

THEOREM 3.4. *Let Assumptions 2.2 to 2.5 and 3.3 hold true. Assume that $F(\mathcal{S})$ has an empty interior. Let $x_0 \in \mathbb{R}^d$ and let $z : t \mapsto (x(t), m(t), v(t))$ be the global solution to (ODE) with initial condition $(x_0, 0, 0)$. Then, there exists $x^* \in \mathcal{S}$ such that $x(t)$ converges to x^* as $t \to +\infty$.*

Moreover, if $\theta \in (0, \frac{1}{2}]$ is a Łojasiewicz exponent of f at x^*, there exists a constant $C > 0$ s.t. for all $t \geq 0$,

$$\|x(t) - x^*\| \leq C t^{-\frac{\theta}{1-2\theta}}\,, \quad \text{if } 0 < \theta < \frac{1}{2}\,,$$
$$\|x(t) - x^*\| \leq C e^{-\delta t}\,, \quad \text{for some } \delta > 0 \text{ if } \theta = \frac{1}{2}\,.$$

4. Discrete-Time System: Convergence of ADAM.

Assumption 4.1. The sequence $(\xi_n : n \geq 1)$ is iid, with the same distribution as ξ.

Assumption 4.2. Let $p > 0$. Assume either one of the following conditions.

i) For every compact set $K \subset \mathbb{R}^d$, $\sup_{x \in K} \mathbb{E}(\|\nabla f(x, \xi)\|^p) < \infty$.
ii) For every compact set $K \subset \mathbb{R}^d$, $\exists p_K > p$, $\sup_{x \in K} \mathbb{E}(\|\nabla f(x, \xi)\|^{p_K}) < \infty$.

The value of p will be specified in the sequel, in the statement of the results. Clearly, Assumption 4.2 ii) is stronger than Assumption 4.2 i). We shall use either the latter or the former in our statements.

THEOREM 4.3. *Let Assumptions 2.2 to 2.5 and 4.1 hold true. Let Assumption 4.2 ii) hold with $p = 2$. Consider $x_0 \in \mathbb{R}^d$. For every $\gamma > 0$, let $(z_n^\gamma : n \in \mathbb{N})$ be*

the random sequence defined by the ADAM *iterations (2.5) and* $z_0^\gamma = (x_0, 0, 0)$. *Let* z^γ *be the corresponding interpolated process defined by Eq. (2.3). Finally, let* z *denote the unique global solution to (ODE) issued from* $(x_0, 0, 0)$. *Then,*

$$\forall\, T > 0,\ \forall\, \delta > 0,\quad \lim_{\gamma \downarrow 0} \mathbb{P}\left(\sup_{t \in [0,T]} \|\mathsf{z}^\gamma(t) - z(t)\| > \delta \right) = 0\,.$$

Recall that a family of r.v. $(X_\alpha)_{\alpha \in I}$ is called *bounded in probability*, or *tight*, if for every $\delta > 0$, there exists a compact set K s.t. $\mathbb{P}(X_\alpha \in K) \geq 1 - \delta$ for every $\alpha \in I$.

Assumption 4.4. There exists $\bar{\gamma}_0 > 0$ s.t. the family of r.v. $(z_n^\gamma : n \in \mathbb{N}, 0 < \gamma < \bar{\gamma}_0)$ is bounded in probability.

THEOREM 4.5. *Consider* $x_0 \in \mathbb{R}^d$. *For every* $\gamma > 0$, *let* $(z_n^\gamma : n \in \mathbb{N})$ *be the random sequence defined by the* ADAM *iterations (2.5) and* $z_0^\gamma = (x_0, 0, 0)$. *Let Assumptions 2.2 to 2.5, 4.1 and 4.4 hold. Let Assumption 4.2 ii) hold with* $p = 2$. *Then, for every* $\delta > 0$,

$$\lim_{\gamma \downarrow 0} \limsup_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{P}(\mathsf{d}(x_k^\gamma, \mathcal{S}) > \delta) = 0\,. \tag{4.1}$$

Convergence in the long run. When the stepsize γ is constant, the sequence (x_n^γ) cannot converge in the almost sure sense as $n \to \infty$. Convergence may only hold in the doubly asymptotic regime where $n \to \infty$ then $\gamma \to 0$.

Randomization. For every n, consider a r.v. N_n uniformly distributed on $\{1, \ldots, n\}$. Define $\tilde{x}_n^\gamma = x_{N_n}^\gamma$. We obtain from Th. 4.5 that for every $\delta > 0$,

$$\limsup_{n \to \infty}\ \mathbb{P}(\mathsf{d}(\tilde{x}_n^\gamma, \mathcal{S}) > \delta) \underset{\gamma \downarrow 0}{\longrightarrow} 0\,.$$

Relationship between discrete and continuous time ADAM. Th. 4.3 means that the family of random processes $(\mathsf{z}^\gamma :\ \gamma > 0)$ converges in probability as $\gamma \downarrow 0$ towards the unique solution to (ODE) issued from $(x_0, 0, 0)$. This motivates the fact that the non-autonomous system (ODE) is a relevant approximation to the behavior of the iterates $(z_n^\gamma : n \in \mathbb{N})$ for a small value of the stepsize γ.

Stability. Assumption 4.4 ensures that the iterates z_n^γ do not explode in the long run. A sufficient condition is for instance that $\sup_{n,\gamma} \mathbb{E}\|z_n^\gamma\| < \infty$. In theory, this assumption can be difficult to verify. Nevertheless, in practice, a projection step on a compact set can be introduced to ensure the boundedness of the estimates.

5. A Decreasing Stepsize ADAM Algorithm.

5.1. Algorithm. ADAM inherently uses constant stepsizes. Consequently, the iterates (2.5) do not converge in the almost sure sense. In order to achieve convergence, we introduce in this section a decreasing stepsize version of ADAM. The iterations are given in Algorithm 5.1. The algorithm generates a sequence $z_n = (x_n, m_n, v_n)$ with initial point $z_0 = (x_0, 0, 0)$, where $x_0 \in \mathbb{R}^d$. Apart from the fact that the hyperparameters $(\gamma_n, \alpha_n, \beta_n)$ now depend on n, the main difference w.r.t Algorithm 2.1 lies in the expression of the debiasing step. As noted in Remark 2.1, the aim is to rescale m_n (resp. v_n) in such a way that the rescaled version $\hat{m}_n$ (resp. $\hat{v}_n$) is a convex combination of past stochastic gradients (resp. squared gradients). While in the constant step case the rescaling coefficient is $(1 - \alpha^n)^{-1}$ (resp. $(1 - \beta^n)^{-1}$), the

Algorithm 5.1 ADAM- decreasing stepsize $(((\gamma_n, \alpha_n, \beta_n) : n \in \mathbb{N}^*), \varepsilon)$**.**

Initialization: $x_0 \in \mathbb{R}^d, m_0 = 0$, $v_0 = 0$, $r_0 = \bar{r}_0 = 0$.
for $n = 1$ **to** n_{iter} **do**
$m_n = \alpha_n m_{n-1} + (1 - \alpha_n)\nabla f(x_{n-1}, \xi_n)$
$v_n = \beta_n v_{n-1} + (1 - \beta_n)\nabla f(x_{n-1}, \xi_n)^{\odot 2}$
$r_n = \alpha_n r_{n-1} + (1 - \alpha_n)$
$\bar{r}_n = \beta_n \bar{r}_{n-1} + (1 - \beta_n)$
$\hat{m}_n = m_n / r_n$ {bias correction step}
$\hat{v}_n = v_n / \bar{r}_n$ {bias correction step}
$x_n = x_{n-1} - \gamma_n \hat{m}_n / (\varepsilon + \sqrt{\hat{v}_n})$.
end for

decreasing step case requires dividing m_n by the coefficient $r_n = 1 - \prod_{i=1}^n \alpha_i$ (resp. v_n by $\bar{r}_n = 1 - \prod_{i=1}^n \beta_i$), which keeps track of the previous weights:

$$\hat{m}_n = \frac{m_n}{r_n} = \frac{\sum_{k=1}^n \rho_{n,k} \nabla f(x_{k-1}, \xi_k)}{\sum_{k=1}^n \rho_{n,k}},$$

where for every n, k, $\rho_{n,k} = \alpha_n \cdots \alpha_{k+1}(1 - \alpha_k)$. A similar equation holds for $\hat{v}_n$.

5.2. Almost sure convergence.

Assumption 5.1 (Stepsizes). The following holds.
i) For all $n \in \mathbb{N}$, $\gamma_n > 0$ and $\gamma_{n+1}/\gamma_n \to 1$,
ii) $\sum_n \gamma_n = +\infty$ and $\sum_n \gamma_n^2 < +\infty$,
iii) For all $n \in \mathbb{N}$, $0 \le \alpha_n \le 1$ and $0 \le \beta_n \le 1$,
iv) There exist a, b s.t. $0 < b < 4a$, $\gamma_n^{-1}(1 - \alpha_n) \to a$ and $\gamma_n^{-1}(1 - \beta_n) \to b$.

THEOREM 5.2. *Let Assumptions 2.2 to 2.4, 4.1 and 5.1 hold. Let Assumption 4.2 i) hold with $p = 4$. Assume that $F(\mathcal{S})$ has an empty interior and that the random sequence $((x_n, m_n, v_n) : n \in \mathbb{N})$ given by Algorithm 5.1 is bounded, with probability one. Then, w.p.1, $\lim_{n\to\infty} \mathsf{d}(x_n, \mathcal{S}) = 0$, $lim_{n\to\infty} m_n = 0$ and $lim_{n\to\infty}(S(x_n) - v_n) = 0$. If moreover $\mathcal{S}$ is finite or countable, then w.p.1, there exists $x^* \in \mathcal{S}$ s.t. $\lim_{n\to\infty}(x_n, m_n, v_n) = (x^*, 0, S(x^*))$.*

Th. 5.2 establishes the almost sure convergence of x_n to the set of critical points of F, under the assumption that the sequence $((x_n, m_n, v_n))$ is a.s. bounded. The next result provides a sufficient condition under which almost sure boundedness holds.

Assumption 5.3. The following holds.
i) ∇F is Lipschitz continuous.
ii) There exists $C > 0$ s.t. for all $x \in \mathbb{R}^d$, $\mathbb{E}[\|\nabla f(x, \xi)\|^2] \le C(1 + F(x))$.
iii) We assume the condition: $\limsup_{n\to\infty} \left(\frac{1}{\gamma_n} - \left(\frac{1-\alpha_{n+2}}{1-\alpha_{n+1}}\right)\frac{1}{\gamma_{n+1}}\right) < 2(a - \frac{b}{4})$, which is satisfied for instance if $b < 4a$ and $1 - \alpha_{n+1} = a\gamma_n$.

THEOREM 5.4. *Let Assumptions 2.2, 2.3, 4.1, 5.1 and 5.3 hold. Let Assumption 4.2 i) hold with $p = 4$. Then, the sequence $((x_n, m_n, v_n) : n \in \mathbb{N})$ given by Algorithm 5.1 is bounded with probability one.*

5.3. Central Limit Theorem.

Assumption 5.5. Let $x^* \in \mathcal{S}$. There exists a neighborhood $\mathcal{V}$ of x^* s.t.
i) F is twice continuously differentiable on $\mathcal{V}$, and the Hessian $\nabla^2 F(x^*)$ of F at x^* is positive definite.

ii) S is continuously differentiable on $\mathcal{V}$.

Define $D := \operatorname{diag}\left((\varepsilon + \sqrt{S_1(x^*)})^{-1}, \cdots, (\varepsilon + \sqrt{S_d(x^*)})^{-1}\right)$. Let P be an orthogonal matrix s.t. the following spectral decomposition holds:

$$D^{1/2}\nabla^2 F(x^*)D^{1/2} = P\operatorname{diag}(\lambda_1, \cdots, \lambda_d)P^{-1},$$

where $\lambda_1, \cdots, \lambda_d$ are the (positive) eigenvalues of $D^{1/2}\nabla^2 F(x^*)D^{1/2}$. Define

$$H := \begin{pmatrix} 0 & -D & 0 \\ a\nabla^2 F(x^*) & -aI_d & 0 \\ b\nabla S(x^*) & 0 & -bI_d \end{pmatrix} \tag{5.1}$$

where I_d represents the $d \times d$ identity matrix and $\nabla S(x^*)$ is the Jacobian matrix of S at x^*. The largest real part of the eigenvalues of H coincides with $-L$, where

$$L := b \wedge \frac{a}{2}\left(1 - \sqrt{\left(1 - \frac{4\lambda_1}{a}\right) \vee 0}\right) > 0\,. \tag{5.2}$$

Finally, define the $3d \times 3d$ matrix

$$Q := \begin{pmatrix} 0 & 0 & 0 \\ 0 & \mathbb{E}\left[\begin{pmatrix} a\nabla f(x^*,\xi) \\ b(\nabla f(x^*,\xi)^{\odot 2} - S(x^*)) \end{pmatrix}\begin{pmatrix} a\nabla f(x^*,\xi) \\ b(\nabla f(x^*,\xi)^{\odot 2} - S(x^*)) \end{pmatrix}^T\right] \\ 0 \end{pmatrix}. \tag{5.3}$$

Assumption 5.6. The following holds.

i) There exist $\kappa \in (0,1]$, $\gamma_0 > 0$, s.t. the sequence (γ_n) satisfies $\gamma_n = \gamma_0/(n+1)^\kappa$ for all n. If $\kappa = 1$, we assume moreover that $\gamma_0 > \frac{1}{2L}$.

ii) The sequences $\left(\frac{1}{\gamma_n}(\frac{1-\alpha_n}{\gamma_n} - a)\right)$ and $\left(\frac{1}{\gamma_n}(\frac{1-\beta_n}{\gamma_n} - b)\right)$ are bounded.

For an arbitrary sequence (X_n) of random variables on some Euclidean space, a probability measure μ on that space and an event Γ s.t. $\mathbb{P}(\Gamma) > 0$, we say that X_n converges in distribution to μ *given* Γ if the measures $\mathbb{P}(X_n \in \cdot\,|\Gamma)$ converge weakly to μ.

THEOREM 5.7. *Let Assumptions 2.2, 2.4, 4.1, 5.5 and 5.6 hold true. Let Assumption 4.2 ii) hold with $p = 4$. Consider the iterates $z_n = (x_n, m_n, v_n)$ given by Algorithm 5.1. Set $z^* = (x^*, 0, S(x^*))$. Set $\zeta := 0$ if $0 < \kappa < 1$ and $\zeta := \frac{1}{2\gamma_0}$ if $\kappa = 1$. Assume $\mathbb{P}(z_n \to z^*) > 0$. Then, given the event $\{z_n \to z^*\}$, the rescaled vector $\sqrt{\gamma_n}^{-1}(z_n - z^*)$ converges in distribution to a zero mean Gaussian distribution on $\mathbb{R}^{3d}$ with a covariance matrix Σ which is solution to the Lyapunov equation: $(H + \zeta I_{3d})\,\Sigma + \Sigma\left(H^T + \zeta I_{3d}\right) = -Q$. In particular, given $\{z_n \to z^*\}$, the vector $\sqrt{\gamma_n}^{-1}(x_n - x^*)$ converges in distribution to a zero mean Gaussian distribution with a covariance matrix Σ_1 given by:*

$$\Sigma_1 = D^{1/2}P\left(\frac{C_{k,\ell}}{(1-\frac{2\zeta}{a})(\lambda_k + \lambda_\ell - 2\zeta + \frac{2}{a}\zeta^2) + \frac{1}{2(a-2\zeta)}(\lambda_k - \lambda_\ell)^2}\right)_{k,\ell=1\ldots d} P^{-1}D^{1/2} \tag{5.4}$$

where $C := P^{-1}D^{1/2}\mathbb{E}\left(\nabla f(x^*,\xi)\nabla f(x^*,\xi)^T\right)D^{1/2}P$.

The following remarks are useful.

- The variable v_n has an impact on the limiting covariance Σ_1 through its limit $S(x^*)$ (used to define D), but the fluctuations of v_n and the parameter b have no effect on Σ_1. As a matter of fact, Σ_1 coincides with the limiting covariance matrix that would have been obtained by considering iterates of the form

$$\begin{cases} x_{n+1} &= x_n - \gamma_{n+1} p_{n+1} \\ p_{n+1} &= p_n + a\gamma_{n+1}(D\nabla f(x_n, \xi_{n+1}) - p_n)\,, \end{cases}$$

  which can be interpreted as a preconditioned version of the stochastic heavy ball algorithm [14]. Of course, the above iterates are not implementable because the preconditioning matrix D is unknown.
- When a is large, Σ_1 is close to the matrix $\Sigma_1^{(0)}$ obtained when letting $a \to +\infty$ in Eq. (5.4). The matrix $\Sigma_1^{(0)}$ is the solution to the Lyapunov equation

$$(D\nabla^2 F(x^*) - \zeta I_d)\Sigma_1^{(0)} + \Sigma_1^{(0)}(\nabla^2 F(x^*)D - \zeta I_d) = D\mathbb{E}\left(\nabla f(x^*,\xi)\nabla f(x^*,\xi)^T\right) D\,.$$

  The matrix $\Sigma_1^{(0)}$ can be interpreted as the asymptotic covariance matrix of the x-variable in the absence of the inertial term (that is, when one considers RmsProp instead of Adam). The matrix $\Sigma_1^{(0)}$ approximates Σ_1 in the sense that $\Sigma_1 = \Sigma_1^{(0)} + \frac{1}{a}\Delta + O(\frac{1}{a^2})$ for some symmetric matrix Δ which can be explicited. The matrix Δ is neither positive nor negative definite in general. This suggests that the question of the potential benefit of the presence of an inertial term is in general problem dependent.
- In the statement of Th. 5.7, the conditioning event $\{z_n \to z^*\}$ can be replaced by the event $\{x_n \to x^*\}$ under the additional assumption that $\sum_n \gamma_n^2 < +\infty$.

6. Related Works. Although the idea of adapting the (per-coordinate) learning rates as a function of past gradient values is not new (see *e.g.* variable metric methods such as the BFGS algorithms), AdaGrad [12] led the way to a new class of algorithms that are sometimes referred to as adaptive gradient methods. AdaGrad consists of dividing the learning rate by the square root of the sum of previous gradients squared componentwise. The idea was to give larger learning rates to highly informative but infrequent features instead of using a fixed predetermined schedule. However, in practice, the division by the cumulative sum of squared gradients may generate small learning rates, thus freezing the iterates too early. Several works proposed heuristical ways to set the learning rates using a less aggressive policy. The work [23] introduced an unpublished, yet popular, algorithm referred to as RmsProp where the cumulative sum used in AdaGrad is replaced by a moving average of squared gradients. Adam combines the advantages of both AdaGrad, RmsProp and inertial methods.

As opposed to AdaGrad, for which theoretical convergence guarantees exist [12, 9, 26, 24], Adam is comparatively less studied. The initial paper [18] suggests a $\mathcal{O}(\frac{1}{\sqrt{T}})$ average regret bound in the convex setting, but [21] exhibits a counterexample in contradiction with this statement. The latter counterexample implies that the average regret bound of Adam does not converge to zero. A first way to overcome the problem is to modify the Adam iterations themselves in order to obtain a vanishing average regret. This led [21] to propose a variant called AmsGrad with the aim to recover, at least in the convex case, the sought guarantees. The work [3] interprets Adam as a variance-adapted sign descent combining an update direction given by the sign and a magnitude controlled by a variance adaptation principle. A "noiseless" version of Adam is considered in [4]. Under quite specific values of the Adam-hyperparameters,

it is shown that for every $\delta > 0$, there exists some time instant for which the norm of the gradient of the objective at the current iterate is no larger than δ. The recent paper [9] provides a similar result for AMSGRAD and ADAGRAD, but the generalization to ADAM is subject to conditions which are not easily verifiable. The paper [25] provides a convergence result for RMSPROP using the objective function F as a Lyapunov function. However, our work suggests that unlike RMSPROP, ADAM does not admit F as a Lyapunov function. This makes the approach of [25] hardly generalizable to ADAM. Moreover, [25] considers biased gradient estimates instead of the debiased estimates used in ADAM.

In the present work, we study the behavior of an ODE, interpreted as the limit in probability of the (interpolated) ADAM iterates as the stepsize tends to zero. Closely related continuous-time dynamical systems are also studied in [2, 8]. We leverage the idea of approximating a discrete time stochastic system by a deterministic continuous one, often referred to as the ODE method. A recent work [14] fruitfully exploits this method to study a stochastic version of the celebrated heavy ball algorithm. We refer to [11] for the reader interested in the non-differentiable setting with an analysis of the stochastic subgradient algorithm for non-smooth non-convex objective functions.

Concomitant to the present paper, Da Silva and Gazeau [10] (posted only four weeks after the first version of the present work) study the asymptotic behavior of a similar dynamical system as the one introduced here. They establish several results in continuous time, such as avoidance of traps as well as convergence rates in the convex case; such aspects are out of the scope of this paper. However, the question of the convergence of the (discrete-time) iterates is left open. In the current paper, we also exhibit a Lyapunov function which allows, amongst others, to draw useful conclusions on the effect of the debiasing step of ADAM. Finally, [10] studies a slightly modified version of ADAM allowing to recover an ODE with a locally Lipschitz continuous vector field, whereas the original ADAM algorithm [18] leads to an ODE with an irregular vector field. This technical issue is tackled in the present paper.

7. Proofs of Section 3.

7.1. Preliminaries.

The results in this section are not specific to the case where F and S are defined as in Eq. (2.2): they are stated for *any* mappings F, S satisfying the following hypotheses.

Assumption 7.1. The function $F : \mathbb{R}^d \to \mathbb{R}$ is s.t.: F is continuously differentiable and ∇F is locally Lipschitz continuous.

Assumption 7.2. The map $S : \mathbb{R}^d \to [0, +\infty)^d$ is locally Lipschitz continuous.

In the sequel, we consider the following generalization of Eq. (ODE) for any $\eta > 0$:

$$\dot{z}(t) = h(t + \eta, z(t))\,. \tag{ODE$_\eta$}$$

When $\eta = 0$, Eq. (ODE$_\eta$) boils down to the equation of interest (ODE). The choice $\eta \in (0, +\infty)$ will be revealed useful to prove Th. 3.1. Indeed, for $\eta > 0$, a solution to Eq. (ODE$_\eta$) can be shown to exist (on some interval) due to the continuity of the map $h(\,.\,+\eta,\,.\,)$. Considering a family of such solutions indexed by $\eta \in (0, 1]$, the idea is to prove the existence of a solution to (ODE) as a cluster point of the latter family when $\eta \downarrow 0$. Indeed, as the family is shown to be equicontinuous, such a cluster point does exist thanks to the Arzelà-Ascoli theorem. When $\eta = +\infty$, Eq. (ODE$_\eta$) rewrites

$$\dot{z}(t) = h_\infty(z(t))\,, \tag{ODE$_\infty$}$$

where $h_\infty(z) := \lim_{t\to\infty} h(t,z)$. It is useful to note that for $(x,m,v) \in \mathcal{Z}_+$,

$$h_\infty((x,m,v)) = \left(-m/(\varepsilon+\sqrt{v})\,,\, a(\nabla F(x)-m)\,,\, b(S(x)-v)\right)\,. \tag{7.1}$$

Contrary to Eq. (ODE), Eq. (ODE$_\infty$) defines an autonomous ODE. The latter admits a unique global solution for any initial condition in $\mathcal{Z}_+$, and defines a dynamical system $\mathcal{D}$. We shall exhibit a strict Lyapunov function for this dynamical system $\mathcal{D}$, and deduce that any solution to (ODE$_\infty$) converges to the set of equilibria of $\mathcal{D}$ as $t\to\infty$. On the otherhand, we will prove that the solution to (ODE) with a proper initial condition is a so-called asymptotic pseudotrajectory (APT) of $\mathcal{D}$. Due to the existence of a strict Lyapunov function, the APT shall inherit the convergence behavior of the autonomous system as $t\to\infty$, which will prove Th. 3.2.

It is convenient to extend the map $h : (0,+\infty)\times\mathcal{Z}_+ \to \mathcal{Z}$ on $(0,+\infty)\times\mathcal{Z}\to\mathcal{Z}$ by setting $h(t,(x,m,v)) := h(t,(x,m,|v|))$ for every $t>0$, $(x,m,v)\in\mathcal{Z}$. Similarly, we extend h_∞ as $h_\infty((x,m,v)) := h_\infty((x,m,|v|))$. For any $T\in(0,+\infty]$ and any $\eta\in[0,+\infty]$, we say that a map $z:[0,T)\to\mathcal{Z}$ is a solution to (ODE$_\eta$) on $[0,T)$ with initial condition $z_0\in\mathcal{Z}_+$, if z is continuous on $[0,T)$, continuously differentiable on $(0,T)$, and if (ODE$_\eta$) holds for all $t\in(0,T)$. When $T=+\infty$, we say that the solution is global. We denote by $Z_T^\eta(z_0)$ the subset of $C([0,T),\mathcal{Z})$ formed by the solutions to (ODE$_\eta$) on $[0,T)$ with initial condition z_0. For any $K\subset\mathcal{Z}_+$, we define $Z_T^\eta(K) := \bigcup_{z\in K} Z_T^\eta(z)$.

Lemma 7.3. *Let Assumptions 7.1 and 7.2 hold. Consider $x_0\in\mathbb{R}^d$, $T\in(0,+\infty]$ and let $z\in Z_T^0((x_0,0,0))$, which we write $z(t)=(x(t),m(t),v(t))$. Then, z is continuously differentiable on $[0,T)$, $\dot m(0)=a\nabla F(x_0)$, $\dot v(0)=bS(x_0)$ and $\dot x(0)=\frac{-\nabla F(x_0)}{\varepsilon+\sqrt{S(x_0)}}$.*

Proof. By definition of $z(\,.\,)$, $m(t)=\int_0^t a(\nabla F(x(s))-m(s))ds$ for all $t\in[0,T)$ (and a similar relation holds for $v(t)$). The integrand being continuous, it holds that m and v are differentiable at zero and $\dot m(0)=a\nabla F(x_0)$, $\dot v(0)=bS(x_0)$. Similarly, $x(t)=x_0+\int_0^t h_x(s,z(s))ds$, where $h_x(s,z(s)) := -(1-e^{-as})^{-1}m(s)/(\varepsilon+\sqrt{(1-e^{-bs})^{-1}v(s)})$. Note that $m(s)/s\to\dot m(0)=a\nabla F(x_0)$ as $s\downarrow 0$. Thus, $(1-e^{-as})^{-1}m(s)\to\nabla F(x_0)$ as $s\to 0$. Similarly, $(1-e^{-bs})^{-1}v(s)\to S(x_0)$. It follows that $h_x(s,z(s))\to-(\varepsilon+\sqrt{S(x_0)})^{-1}\nabla F(x_0)$. Thus, $s\mapsto h_x(s,z(s))$ can be extended to a continuous map on $[0,T)\to\mathbb{R}^d$ and the differentiability of x at zero follows. ∎

Lemma 7.4. *Let Assumptions 2.4, 7.1 and 7.2 hold. For every $\eta\in[0,+\infty]$, $T\in(0,+\infty]$, $z_0\in\mathcal{Z}_+$, $z\in Z_T^\eta(z_0)$, it holds that $z((0,T))\subset\mathcal{Z}_+^*$.*

Proof. Set $z(t)=(x(t),m(t),v(t))$ for all t. Consider $i\in\{1,\dots,d\}$. Assume by contradiction that there exists $t_0\in(0,T)$ s.t. $v_i(t_0)<0$. Set $\tau := \sup\{t\in[0,t_0] : v_i(t)\geq 0\}$. Clearly, $\tau<t_0$ and $v_i(\tau)=0$ by the continuity of v_i. Since $v_i(t)\leq 0$ for all $t\in(\tau,t_0]$, it holds that $\dot v_i(t)=b(S_i(x(t))-v_i(t))$ is nonnegative for all $t\in(\tau,t_0]$. This contradicts the fact that $v_i(\tau)>v_i(t_0)$. Thus, $v_i(t)\geq 0$ for all $t\in[0,T)$. Now assume by contradiction that there exists $t\in(0,T)$ s.t. $v_i(t)=0$. Then, $\dot v_i(t)=bS_i(x(t))>0$. Thus, $\lim_{\delta\downarrow 0}\frac{v_i(t-\delta)}{-\delta}=bS_i(x(t))$. In particular, there exists $\delta>0$ s.t. $v_i(t-\delta)\leq-\frac{\delta b}{2}S_i(x(t))$. This contradicts the first point. ∎

Recall the definitions of V and U from Eqs. (3.4) and (3.5). Clearly, $U_\infty(v) := \lim_{t\to\infty}U(t,v)=a(\varepsilon+\sqrt{v})$ is well defined for every $v\in[0,+\infty)^d$. Hence, we can also define $V_\infty(z) := \lim_{t\to\infty}V(t,z)$ for every $z\in\mathcal{Z}_+$.

Lemma 7.5. *Let Assumptions 7.1 and 7.2 hold. Assume that $0<b\leq 4a$. Consider $(t,z)\in(0,+\infty)\times\mathcal{Z}_+^*$ and set $z=(x,m,v)$. Then, V and V_∞ are differentiable*

at points (t,z) *and* z *respectively. Moreover,* $\langle \nabla V_\infty(z), h_\infty(z)\rangle \leq -\varepsilon \left\| \frac{am}{U_\infty(v)} \right\|^2$ *and*

$$\langle \nabla V(t,z), (1, h(t,z))\rangle \leq -\frac{\varepsilon}{2} \left\| \frac{a\, m}{U(t,v)} \right\|^2 .$$

Proof. We only prove the second point, the proof of the first point follows the same line. Consider $(t,z) \in (0,+\infty)\times \mathcal{Z}_+^*$. We decompose $\langle \nabla V(t,z), (1,h(t,z))\rangle = \partial_t V(t,z) + \langle \nabla_z V(t,z), h(t,z)\rangle$. After tedious but straightforward derivations, we get:

(7.2)
$$\partial_t V(t,z) = -\sum_{i=1}^d \frac{a^2 m_i^2}{U(t,v_i)^2} \left(\frac{e^{-at}\varepsilon}{2} + \left(\frac{e^{-at}}{2} - \frac{b e^{-bt}(1-e^{-at})}{4a(1-e^{-bt})} \right) \sqrt{\frac{v_i}{1-e^{-bt}}} \right) ,$$

where $U(t,v_i) = a(1-e^{-at})\left(\varepsilon + \sqrt{\frac{v_i}{1-e^{-bt}}}\right)$ and $\langle \nabla_z V(t,z), h(t,z)\rangle$ is equal to:

$$\sum_{i=1}^d \frac{-a^2 m_i^2 (1-e^{-at})}{U(t,v_i)^2} \left(\varepsilon + (1-\frac{b}{4a})\sqrt{\frac{v_i}{1-e^{-bt}}} + \frac{b S_i(x)}{4a\sqrt{v_i(1-e^{-bt})}} \right) .$$

Using that $S_i(x) \geq 0$, we obtain:

$$\langle \nabla V(t,z), (1,h(t,z))\rangle \leq -\sum_{i=1}^d \frac{a^2 m_i^2}{U(t,v_i)^2} \left((1-\frac{e^{-at}}{2})\varepsilon + c_{a,b}(t)\sqrt{\frac{v_i}{1-e^{-bt}}} \right) , \tag{7.3}$$

where $c_{a,b}(t) := 1 - \frac{e^{-at}}{2} - \frac{b}{4a}\frac{1-e^{-at}}{1-e^{-bt}}$. Using inequality $1-e^{-at}/2 \geq 1/2$ in (7.3), the inequality (7.3) proves the Lemma, provided that one is able to show that $c_{a,b}(t) \geq 0$, for all $t>0$ and all a,b satisfying $0<b\leq 4a$. We prove this last statement. It can be shown that the function $b \mapsto c_{a,b}(t)$ is decreasing on $[0,+\infty)$. Hence, $c_{a,b}(t) \geq c_{a,4a}(t)$. Now, $c_{a,4a}(t) = q(e^{-at})$ where $q:[0,1)\to\mathbb{R}$ is the function defined for all $y\in[0,1)$ by $q(y) = y\left(y^4-2y^3+1\right)/(2(1-y^4))$. Hence $q\geq 0$. Thus, $c_{a,b}(t) \geq q(e^{-at}) \geq 0$. ∎

7.2. Proof of Th. 3.1.

7.2.1. Boundedness. Define $\mathcal{Z}_0 := \{(x,0,0) : x\in\mathbb{R}^d\}$. Let $\bar{e} : (0,+\infty)\times\mathcal{Z}_+ \to \mathcal{Z}_+$ be defined by $\bar{e}(t,z) := (x, m/(1-e^{-at}), v/(1-e^{-bt}))$ for every $t>0$ and every $z=(x,m,v)$ in $\mathcal{Z}_+$.

Proposition 7.6. *Let Assumptions 2.3, 2.4, 7.1 and 7.2 hold. Assume that* $0<b\leq 4a$. *For every* $z_0\in\mathcal{Z}_0$, *there exists a compact set* $K\subset\mathcal{Z}_+$ *s.t. for all* $\eta\in[0,+\infty)$, *all* $T\in(0,+\infty]$ *and all* $z\in Z_T^\eta(z_0)$, $\{\bar{e}(t+\eta, z(t)) : t\in(0,T)\}\subset K$. *Moreover, choosing* z_0 *of the form* $z_0=(x_0,0,0)$ *and* $z(t)=(x(t),m(t),v(t))$, *it holds that* $F(x(t))\leq F(x_0)$ *for all* $t\in[0,T)$.

Proof. Let $\eta\in[0,+\infty)$. Consider a solution $z_\eta(t) = (x_\eta(t), m_\eta(t), v_\eta(t))$ as in the statement, defined on some interval $[0,T)$. Define $\hat{m}_\eta(t) = m_\eta(t)/(1-e^{-a(t+\eta)})$, $\hat{v}_\eta(t) = v_\eta(t)/(1-e^{-b(t+\eta)})$. By Lemma 7.4, $t\mapsto V(t+\eta, z(t))$ is continuous on $[0,T)$, and continuously differentiable on $(0,T)$. By Lemma 7.5, $\dot{V}(t+\eta, z_\eta(t)) \leq 0$ for all $t>0$. As a consequence, $t\mapsto V(t+\eta, z_\eta(t))$ is non-increasing on $[0,T)$. Thus, for all $t\geq 0$, $F(x_\eta(t)) \leq \lim_{t'\downarrow 0} V(t'+\eta, z_\eta(t'))$. Note that $V(t+\eta, z_\eta(t)) \leq F(x_\eta(t)) + \frac{1}{2}\sum_{i=1}^d \frac{m_{\eta,i}(t)^2}{a(1-e^{-a(t+\eta)})\varepsilon}$. If $\eta>0$, every term in the sum in the righthand side tends to zero, upon noting that $m_\eta(t)\to 0$ as $t\to 0$. The statement still holds if $\eta=0$. Indeed, by Lemma 7.3, for a given $i\in\{1,\dots,d\}$, there exists $\delta>0$ s.t. for all $0<t<\delta$, $m_{\eta,i}(t)^2 \leq 2a^2(\partial_i F(x_0))^2 t^2$ and $1-e^{-at}\geq (at)/2$. As a

consequence, each term of the sum is no larger than $4(\partial_i F(x_0))^2 t/\varepsilon$, which tends to zero as $t \to 0$. We conclude that for all $t \geq 0$, $F(x_\eta(t)) \leq F(x_0)$. In particular, $\{x_\eta(t) : t \in [0,T)\} \subset \{F \leq F(x_0)\}$, the latter set being bounded by Assumption 2.3.

We prove that $v_{i,\eta}(t)$ is (upper)bounded. Define $R_i := \sup S_i(\{F \leq F(x_0)\})$, which is finite by continuity of S. Assume by contradiction that the set $\{t \in [0,T) : v_{\eta,i}(t) \geq R_i + 1\}$ is non-empty, and denote its infimum by τ. By continuity of $v_{\eta,i}$, one has $v_{\eta,i}(\tau) = R_i + 1$. This by the way implies that $\tau > 0$. Hence, $\dot{v}_{\eta,i}(\tau) = b(S_i(x_\eta(\tau)) - v_{\eta,i}(\tau)) \leq -b$. This means that there exists $\tau' < \tau$ s.t. $v_{\eta,i}(\tau') > v_{\eta,i}(\tau)$, which contradicts the definition of τ. We have shown that $v_{\eta,i}(t) \leq R_i + 1$ for all $t \in (0,T)$. In particular, when $t \geq 1$, $\hat{v}_{\eta,i}(t) = v_{\eta,i}(t)/(1-e^{-bt}) \leq (R_i+1)/(1-e^{-b})$. Consider $t \in (0, 1 \wedge T)$. By the mean value theorem, there exists $\tilde{t}_\eta \in [0,t]$ s.t. $v_{\eta,i}(t) = \dot{v}_{\eta,i}(\tilde{t}_\eta)t$. Thus, $v_{\eta,i}(t) \leq bS_i(x(\tilde{t}_\eta))t \leq bR_i t$. Using that the map $y \mapsto y/(1-e^{-y})$ is increasing on $(0,+\infty)$, it holds that for all $t \in (0, 1 \wedge T)$, $\hat{v}_{\eta,i}(t) \leq bR_i/(1-e^{-b})$. We have shown that, for all $t \in (0,T)$ and all $i \in \{1,\dots,d\}$, $0 \leq \hat{v}_{\eta,i}(t) \leq M$, where $M := (1-e^{-b})^{-1}(1+b)(1+\max\{R_\ell : \ell \in \{1,\dots,d\}\})$.

As $V(t+\eta, z_\eta(t)) \leq F(x_0)$, we obtain: $F(x_0) \geq F(x_\eta(t)) + \frac{1}{2}\|m_\eta(t)\|^2_{U(t+\eta,v_\eta(t))^{-1}}$. Thus, $F(x_0) \geq \inf F + \frac{1}{2a(\varepsilon+\sqrt{M})}\|m_\eta(t)\|^2$. Therefore, $m_\eta(\,.\,)$ is bounded on $[0,T)$, uniformly in η. The same holds for $\hat{m}_\eta$ by using the mean value theorem in the same way as for $\hat{v}_\eta$. The proof is complete. ∎

PROPOSITION 7.7. *Let Assumptions 2.3, 2.4, 7.1 and 7.2 hold. Assume that $0 < b \leq 4a$. Let K be a compact subset of $\mathcal{Z}_+$. Then, there exists an other compact set $K' \subset \mathcal{Z}_+$ s.t. for every $T \in (0,+\infty]$ and every $z \in Z_T^\infty(K)$, $z([0,T)) \subset K'$.*

Proof. The proof follows the same line as Prop. 7.6 and is omitted. ∎

For any $K \subset \mathcal{Z}_+$, define $v_{\min}(K) := \inf\{v_k : (x,m,v) \in K, i \in \{1,\dots,d\}\}$.

LEMMA 7.8. *Under Assumptions 2.3, 2.4, 7.1 and 7.2, the following holds true.*

i) For every compact set $K \subset \mathcal{Z}_+$, there exists $c > 0$, s.t. for every $z \in Z_\infty^\infty(K)$, of the form $z(t) = (x(t), m(t), v(t))$, $v_i(t) \geq c \min\left(1, \frac{v_{\min}(K)}{2c} + t\right)$ $(\forall t \geq 0, \forall i \in \{1,\dots,d\})$.

ii) For every $z_0 \in \mathcal{Z}_0$, there exists $c > 0$ s.t. for every $\eta \in [0,+\infty)$ and every $z \in Z_\infty^\eta(z_0)$, $v_i(t) \geq c \min(1,t)$ $(\forall t \geq 0, \forall i \in \{1,\dots,d\})$.

Proof. We prove the first point. Consider a compact set $K \subset \mathcal{Z}_+$. By Prop. 7.7, one can find a compact set $K' \subset \mathcal{Z}_+$ s.t. for every $z \in Z_\infty^\infty(K)$, it holds that $\{z(t) : t \geq 0\} \subset K'$. Denote by L_S the Lipschitz constant of S on the compact set $\{x : (x,m,v) \in K'\}$. Introduce the constants $M_1 := \sup\{\|m/(\varepsilon+\sqrt{v})\|_\infty : (x,m,v) \in K'\}$, $M_2 := \sup\{\|S(x)\|_\infty : (x,m,v) \in K'\}$. The constants L_S, M_1, M_2 are finite. Now consider a global solution $z(t) = (x(t), m(t), v(t))$ in $Z_\infty^\infty(K)$. Choose $i \in \{1,\dots,d\}$ and consider $t \geq 0$. By the mean value theorem, there exists $t' \in [0,t]$ s.t. $v_i(t) = v_i(0) + \dot{v}_i(t')t$. Thus, $v_i(t) = v_i(0) + \dot{v}_i(0)t + b(S_i(x(t')) - v_i(t') - S_i(x(0)) + v_i(0))t$, which in turn implies $v_i(t) \geq v_i(0) + \dot{v}_i(0)t - bL_S\|x(t') - x(0)\|t - b|v_i(t') - v_i(0)|t$. Using again the mean value theorem, for every $\ell \in \{1,\dots,d\}$, there exists $t'' \in [0,t']$ s.t. $|x_\ell(t') - x_\ell(0)| = t'|\dot{x}_\ell(t'')| \leq tM_1$. Therefore, $\|x(t') - x(0)\| \leq \sqrt{d}M_1 t$. Similarly, there exists $\tilde{t}$ s.t.: $|v_i(t') - v_i(0)| = t'|\dot{v}_i(\tilde{t})| \leq t'bS_i(x(\tilde{t})) \leq tbM_2$. Putting together the above inequalities, $v_i(t) \geq v_i(0)(1-bt) + bS_i(x(0))t - bCt^2$, where $C := (M_2 + L_S\sqrt{d}M_1)$. For every $t \leq 1/(2b)$, $v_i(t) \geq \frac{v_{\min}}{2} + tbC\left(\frac{S_{\min}}{C} - t\right)$, where we defined $S_{\min} := \inf\{S_i(x) : i \in \{1,\dots,d\}, (x,m,v) \in K\}$. Setting $\tau := 0.5\min(1/b, S_{\min}/C)$,

$$\forall t \in [0,\tau],\ v_i(t) \geq \frac{v_{\min}}{2} + \frac{bS_{\min}t}{2}. \tag{7.4}$$

Set $\kappa_1 := 0.5(v_{\min} + bS_{\min}\tau)$. Note that $v_i(\tau) \geq \kappa_1$. Define $S'_{\min} := \inf\{S_i(x) : i \in \{1,\dots,d\}, (x,m,v) \in K'\}$. Note that $S'_{\min} > 0$ by Assumptions 7.2 and 2.4. Finally, define $\kappa = 0.5\min(\kappa_1, S'_{\min})$. By contradiction, assume that the set $\{t \geq \tau : v_i(t) < \kappa\}$ is non-empty, and denote by τ' its infimum. It is clear that $\tau' > \tau$ and $v_i(\tau') = \kappa$. Thus, $b^{-1}\dot{v}_i(\tau') = S_i(x(\tau')) - \kappa$. We obtain that $b^{-1}\dot{v}_i(\tau') \geq 0.5S'_{\min} > 0$. As a consequence, there exists $t \in (\tau, \tau')$ s.t. $v_i(t) < v_i(\tau')$. This contradicts the definition of τ'. We have shown that for all $t \geq \tau$, $v_i(t) \geq \kappa$. Putting this together with Eq. (7.4) and using that $\kappa \leq v_{\min} + bS_{\min}\tau$, we conclude that: $\forall t \geq 0$, $v_i(t) \geq \min\left(\kappa, \frac{v_{\min}}{2} + \frac{bS_{\min}t}{2}\right)$. Setting $c := \min(\kappa, bS_{\min}/2)$, the result follows.

We prove the second point. By Prop. 7.6, there exists a compact set $K \subset \mathcal{Z}_+$ s.t. for every $\eta \geq 0$, every $z \in Z^\eta_\infty(x_0)$ of the form $z(t) = (x(t), m(t), v(t))$ satisfies $\{(x(t), \hat{m}(t), \hat{v}(t)) : t \geq 0\} \subset K$, where $\hat{m}(t) = m(t)/(1 - e^{-a(t+h)})$ and $\hat{v}(t) = v(t)/(1-e^{-b(t+h)})$. Denote by L_S the Lipschitz constant of S on the set $\{x : (x,m,v) \in K\}$. Introduce the constants $M_1 := \sup\{\|m/(\varepsilon + \sqrt{v})\|_\infty : (x,m,v) \in K\}$, $M_2 := \sup\{\|S(x)\|_\infty : (x,m,v) \in K'\}$. These constants being introduced, the rest of the proof follows the same line as the proof of the first point. ∎

7.2.2. Existence.

Corollary 7.9. *Let Assumptions 2.3, 2.4, 7.1 and 7.2 hold. Assume that* $0 < b \leq 4a$. *For every* $z_0 \in \mathcal{Z}_+$, $Z^\infty_\infty(z_0) \neq \emptyset$. *For every* $(z_0, \eta) \in \mathcal{Z}_0 \times (0, +\infty)$, $Z^\eta_\infty(z_0) \neq \emptyset$.

Proof. We prove the first point (the proof of the second point follows the same line). Under Assumptions 7.1 and 7.2, h_∞ is continuous. Therefore, Cauchy-Peano's theorem guarantees the existence of a solution to the (ODE) issued from z_0, which we can extend over a maximal interval of existence $[0, T_{\max})$.We conclude that the solution is global ($T_{\max} = +\infty$) using the boundedness of the solution given by Prop. 7.7. ∎

Lemma 7.10. *Let Assumptions 2.3, 2.4, 7.1 and 7.2 hold. Assume that* $0 < b \leq 4a$. *Consider* $z_0 \in \mathcal{Z}_0$. *Denote by* $(z_\eta : \eta \in (0, +\infty))$ *a family of functions on* $[0, +\infty) \to \mathcal{Z}_+$ *s.t. for every* $\eta > 0$, $z_\eta \in Z^\eta_\infty(z_0)$. *Then,* $(z_\eta)_{\eta>0}$ *is equicontinuous.*

Proof. For every such solution z_η, we set $z_\eta(t) = (x_\eta(t), m_\eta(t), v_\eta(t))$ for all $t \geq 0$, and define $\hat{m}_\eta$ and $\hat{v}_\eta$ as in Prop. 7.6. By Prop. 7.6, there exists a constant M_1 s.t. for all $\eta > 0$ and all $t \geq 0$, $\max(\|x_\eta(t)\|, \|\hat{m}_\eta(t)\|_\infty, \|\hat{v}_\eta(t)\|) \leq M_1$. Using the continuity of ∇F and S, there exists an other finite constant M_2 s.t. $M_2 \geq \sup\{\|\nabla F(x)\|_\infty : x \in \mathbb{R}^d, \|x\| \leq M_1\}$ and $M_2 \geq \sup\{\|S(x)\|_\infty : x \in \mathbb{R}^d, \|x\| \leq M_1\}$. For every $(s,t) \in [0,+\infty)^2$, we have for all $i \in \{1,\dots,d\}$, $|x_{\eta,i}(t) - x_{\eta,i}(s)| \leq \int_s^t \left|\frac{\hat{m}_{\eta,i}(u)}{\varepsilon + \sqrt{\hat{v}_{\eta,i}(u)}}\right| du \leq \frac{M_1}{\varepsilon}|t-s|$, and similarly $|m_{\eta,i}(t) - m_{\eta,i}(s)| \leq a(M_1+M_2)|t-s|$, $|v_{\eta,i}(t) - v_{\eta,i}(s)| \leq b(M_1+M_2)|t-s|$. Therefore, there exists a constant M_3, independent from η, s.t. for all $\eta > 0$ and all $(s,t) \in [0,+\infty)^2$, $\|z_\eta(t) - z_\eta(s)\| \leq M_3|t-s|$.

Proposition 7.11. *Let Assumptions 2.3, 2.4, 7.1 and 7.2 hold. Assume that* $0 < b \leq 4a$. *For every* $z_0 \in \mathcal{Z}_0$, $Z^0_\infty(z_0) \neq \emptyset$ i.e., *(ODE) admits a global solution issued from* z_0.

Proof. By Cor. 7.9, there exists a family $(z_\eta)_{\eta>0}$ of functions on $[0,+\infty) \to \mathcal{Z}$ s.t. for every $\eta > 0$, $z_\eta \in Z^\eta_\infty(z_0)$. We set as usual $z_\eta(t) = (x_\eta(t), m_\eta(t), v_\eta(t))$. By Lemma 7.10, and the Arzelà-Ascoli theorem, there exists a map $z : [0,+\infty) \to \mathcal{Z}$ and a sequence $\eta_n \downarrow 0$ s.t. z_{η_n} converges to z uniformly on compact sets, as $n \to \infty$. Considering some fixed scalars $t > s > 0$, $z(t) = z(s) + \lim_{n\to\infty} \int_s^t h(u+\eta_n, z_{\eta_n}(u))du$. By Prop. 7.6, there exists a compact set $K \subset \mathcal{Z}_+$ s.t. $\{z_{\eta_n}(t) : t \geq 0\} \subset K$ for all n. Moreover, by Lemma 7.8, there exists a constant $c > 0$ s.t. for all n and all $u \geq 0$,

$v_{\eta_n,k}(u) \geq c\min(1,u)$. Denote by $\bar{K} := K \cap (\mathbb{R}^d \times \mathbb{R}^d \times [c\min(1,s),+\infty)^d)$. It is clear that $\bar{K}$ is a compact subset of $\mathcal{Z}_+^*$. Since h is continuously differentiable on the set $[s,t]\times\bar{K}$, it is Lipschitz continuous on that set. Denote by L_h the corresponding Lipschitz constant. We obtain:

$$\int_s^t \|h(u+\eta_n, z_{\eta_n}(u)) - h(u,z(u))\| du \leq L_h \left(\eta_n + \sup_{u\in[s,t]} \|z_{\eta_n}(u) - z(u)\|\right)(t-s)\,,$$

and the righthand side converges to zero. As a consequence, for all $t > s$, $z(t) = z(s) + \int_s^t h(u,z(u))du$. Moreover, $z(0) = z_0$. This proves that $z \in Z_\infty^0(z_0)$. ∎

7.2.3. Uniqueness.

PROPOSITION 7.12. *Let Assumptions 2.3, 2.4, 7.1 and 7.2 hold. Assume $b \leq 4a$. For every $z_0 \in \mathcal{Z}_0$, $Z_\infty^0(z_0)$ is a singleton* i.e.*, there exists a unique global solution to (ODE) with initial condition z_0.*

Proof. Consider solutions z and z' in $Z_\infty^0(z_0)$. We denote by $(x(t),m(t),v(t))$ the blocks of $z(t)$, and we define $(x'(t),m'(t),v'(t))$ similarly. For all $t > 0$, we define $\hat{m}(t) := m(t)/(1-e^{-at})$, $\hat{v}(t) := v(t)/(1-e^{-bt})$, and we define $\hat{m}'(t)$ and $\hat{v}'(t)$ similarly. By Prop. 7.6, there exists a compact set $K \subset \mathcal{Z}_+$ s.t. $(x(t),\hat{m}(t),\hat{v}(t))$ and $(x'(t),\hat{m}'(t),\hat{v}'(t))$ are both in K for all $t > 0$. We denote by L_S and $L_{\nabla F}$ the Lipschitz constants of S and ∇F on the compact set $\{x : (x,m,v) \in K\}$. These constants are finite by Assumptions 7.1 and 7.2. We define $M := \sup\{\|m\|_\infty : (x,m,v) \in K\}$. Define $u_x(t) := \|x(t) - x'(t)\|^2$, $u_m(t) := \|\hat{m}(t) - \hat{m}'(t)\|^2$ and $u_v(t) := \|\hat{v}(t) - \hat{v}'(t)\|^2$. Let $\delta > 0$. Define: $u^{(\delta)}(t) := u_x(t) + \delta u_m(t) + \delta u_v(t)$. By the chain rule and the Cauchy-Schwarz inequality, $\dot{u}_x(t) \leq 2\|x(t)-x'(t)\|\|\frac{\hat{m}(t)}{\varepsilon+\sqrt{\hat{v}(t)}} - \frac{\hat{m}'(t)}{\varepsilon+\sqrt{\hat{v}'(t)}}\|$. Thus, using Lemma 7.8, there exists $c > 0$ s.t.

$$\dot{u}_x(t) \leq 2\|x(t)-x'(t)\| \left(\varepsilon^{-1}\|\hat{m}(t)-\hat{m}'(t)\| + \frac{M}{2\varepsilon^2\sqrt{c\min(1,t)}}\|\hat{v}(t)-\hat{v}'(t)\|\right).$$

For any $\delta > 0$, $2\|x(t)-x'(t)\|\,\|\hat{m}(t)-\hat{m}'(t)\| \leq \delta^{-1/2}(u_x(t)+\delta u_m(t)) \leq \delta^{-1/2}u^{(\delta)}(t)$. Similarly, $2\|x(t)-x'(t)\|\,\|\hat{v}(t)-\hat{v}'(t)\| \leq \delta^{-1/2}u^{(\delta)}(t)$. Thus, for any $\delta > 0$,

$$\dot{u}_x(t) \leq \left(\frac{1}{\varepsilon\sqrt{\delta}} + \frac{M}{2\varepsilon^2\sqrt{\delta c\min(1,t)}}\right)u^{(\delta)}(t)\,. \tag{7.5}$$

We now study $u_m(t)$. For all $t > 0$, we obtain after some algebra: $\frac{d}{dt}\hat{m}(t) = a(\nabla F(x(t)) - \hat{m}(t))/(1-e^{-at})$. Therefore, $\dot{u}_m(t) \leq \frac{2aL_{\nabla F}}{1-e^{-at}}\|\hat{m}(t)-\hat{m}'(t)\|\,\|x(t)-x'(t)\|$. For any $\theta > 0$, it holds that $2\|\hat{m}(t)-\hat{m}'(t)\|\,\|x(t)-x'(t)\| \leq \theta u_x(t) + \theta^{-1}u_m(t)$. In particular, letting $\theta := 2L_{\nabla F}$, we obtain that for all $\delta > 0$,

$$\delta\dot{u}_m(t) \leq \tfrac{a}{2(1-e^{-at})}\left(4\delta L_{\nabla F}^2 u_x(t) + \delta u_m(t)\right) \leq \left(\tfrac{a}{2} + \tfrac{1}{2t}\right)\left(4\delta L_{\nabla F}^2 u_x(t) + \delta u_m(t)\right), \tag{7.6}$$

where the last inequality is due to the fact that $y/(1-e^{-y}) \leq 1 + y$ for all $y > 0$. Using the exact same arguments, we also obtain that

$$\delta\dot{u}_v(t) \leq \left(\frac{b}{2} + \frac{1}{2t}\right)\left(4\delta L_S^2 u_x(t) + \delta u_m(t)\right). \tag{7.7}$$

We now choose any δ s.t. $4\delta \leq 1/\max(L_S^2, L_{\nabla F}^2)$. Then, Eq. (7.6) and (7.7) respectively imply that $\delta \dot{u}_m(t) \leq 0.5(a + t^{-1})u^{(\delta)}(t)$ and $\delta \dot{u}_v(t) \leq 0.5(b + t^{-1})u^{(\delta)}(t)$. Summing these inequalities along with Eq. (7.5), we obtain that for every $t > 0$, $\dot{u}^{(\delta)}(t) \leq \psi(t)u^{(\delta)}(t)$, where: $\psi(t) := \frac{a+b}{2} + \frac{1}{\varepsilon\sqrt{\delta}} + \frac{M}{2\varepsilon^2\sqrt{\delta c \min(1,t)}} + \frac{1}{t}$. From Grönwall's inequality, it holds that for every $t > s > 0$, $u^{(\delta)}(t) \leq u^{(\delta)}(s) \exp\left(\int_s^t \psi(s')ds'\right)$. We first consider the case where $t \leq 1$. We set $c_1 := (a + b)/2 + (\varepsilon\sqrt{\delta})^{-1}$ and $c_2 := M/(\varepsilon^2\sqrt{\delta c})$. With these notations, $\int_s^t \psi(s')ds' \leq c_1 t + c_2\sqrt{t} + \ln \frac{t}{s}$. Therefore, $u^{(\delta)}(t) \leq \frac{u^{(\delta)}(s)}{s} \exp\left(c_1 t + c_2\sqrt{t} + \ln t\right)$. By Lemma 7.3, recall that $\dot{x}(0)$ and $\dot{x}'(0)$ are both well defined (and coincide). Thus,

$$u_x(s) = \|x(s) - x'(s)\|^2 \leq 2\|x(s) - x(0) - \dot{x}(0)s\|^2 + 2\|x'(s) - x'(0) - \dot{x}'(0)s\|^2 .$$

It follows that $u_x(s)/s^2$ converges to zero as $s \downarrow 0$. We now show the same kind of result for $u_m(s)$ and $u_v(s)$. Consider $i \in \{1, \ldots, d\}$. By the mean value theorem, there exists $\tilde{s}$ (resp. $\tilde{s}'$) in $[0, t]$ s.t. $m_i(s) = \dot{m}_i(\tilde{s})s$ (resp. $m_i'(s) = \dot{m}_i'(\tilde{s}')s$). Thus, $\hat{m}_i(s) = \frac{as}{1-e^{-as}}(\partial_i F(x(\tilde{s})) - m_i(\tilde{s}))$, and a similar equality holds for $\hat{m}_i'(s)$. Then, given that $\|x(\tilde{s}) - x'(\tilde{s}')\| \vee \|m(\tilde{s}) - m'(\tilde{s}')\| \leq \|z(\tilde{s}) - z'(\tilde{s}')\|$, $\tilde{s} \leq s$ and $\tilde{s}' \leq s$,

$$\frac{|\hat{m}_i(s) - \hat{m}_i'(s)|}{s} \leq \frac{2a(L_{\nabla F} \vee 1)s}{1 - e^{-as}} \left(\frac{\|z(\tilde{s}) - z(0)\|}{\tilde{s}} + \frac{\|z'(\tilde{s}') - z'(0)\|}{\tilde{s}'} \right) .$$

By Lemma 7.3, z and z' are differentiable at point zero. Then, the above inequality gives $\limsup_{s\downarrow 0} \frac{|\hat{m}_i(s) - \hat{m}_i'(s)|}{s} \leq 4(L_{\nabla F} \vee 1)\|\dot{z}(0)\|$ and $\limsup_{s\downarrow 0} \frac{u_m(s)}{s^2} \leq 16d(L_{\nabla F}^2 \vee 1)\|\dot{z}(0)\|^2$ .Therefore, $u_m(s)/s$ converges to zero as $s \downarrow 0$. By similar arguments, it can be shown that $\limsup_{s\downarrow 0} u_v(s)/s^2 \leq 16d(L_S^2 \vee 1)\|\dot{z}(0)\|^2$, thus $\lim u_v(s)/s = 0$. Finally, we obtain that $u^{(\delta)}(s)/s$ converges to zero as $s \downarrow 0$. Letting s tend to zero, we obtain that for every $t \leq 1$, $u^{(\delta)}(t) = 0$. Setting $s = 1$ and $t > 1$, and noting that ψ is integrable on $[1, t]$, it follows that $u^{(\delta)}(t) = 0$ for all $t > 1$. This proves that $z = z'$. ∎

We recall that a semiflow Φ on the space (E, d) is a continuous map Φ from $[0, +\infty) \times E$ to E defined by $(t, x) \mapsto \Phi(t, x) = \Phi_t(x)$ such that Φ_0 is the identity and $\Phi_{t+s} = \Phi_t \circ \Phi_s$ for all $(t, s) \in [0, +\infty)^2$.

PROPOSITION 7.13. *Let Assumptions 2.3, 2.4, 7.1 and 7.2 hold. Assume that $0 < b \leq 4a$. The map Z_∞^∞ is single-valued on $\mathcal{Z}_+ \to C([0, +\infty), \mathcal{Z}_+)$* i.e.*, there exists a unique global solution to (ODE$_\infty$) starting from any given point in $\mathcal{Z}_+$. Moreover, the following map is a semiflow:*

$$\begin{array}{rccc} \Phi : [0, +\infty) \times \mathcal{Z}_+ & \to & \mathcal{Z}_+ \\ (t, z) & \mapsto & Z_\infty^\infty(z)(t) \end{array} \tag{7.8}$$

Proof. The result is a direct consequence of Lemma 7.12. ∎

7.3. Proof of Th. 3.2.

7.3.1. Convergence of the semiflow. We first recall some useful definitions and results. Let Ψ represent any semiflow on an arbitrary metric space (E, d). A point $z \in E$ is called an *equilibrium point* of the semiflow Ψ if $\Psi_t(z) = z$ 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 $\mathsf{V}(\Psi_t(z)) \leq \mathsf{V}(z)$ for all $z \in E$ and all $t \geq 0$. It is called a *strict Lyapunov function* if, moreover, $\{z \in E : \forall t \geq 0, \mathsf{V}(\Psi_t(z)) = \mathsf{V}(z)\} = \Lambda_\Psi$. If V is a strict Lyapunov function

for Ψ and if $z \in E$ is a point s.t. $\{\Psi_t(z) : t \geq 0\}$ is relatively compact, then it holds that $\Lambda_\Psi \neq \emptyset$ and $\mathsf{d}(\Psi_t(z), \Lambda_\Psi) \to 0$, see [15, Th. 2.1.7]. A continuous function $z : [0, +\infty) \to E$ is said to be an asymptotic pseudotrajectory (APT) for the semiflow Ψ if for every $T \in (0, +\infty)$, $\lim_{t\to+\infty} \sup_{s\in[0,T]} \mathsf{d}(z(t+s), \Psi_s(z(t))) = 0$. The following result follows from [5, Th. 5.7] and [5, Prop. 6.4].

PROPOSITION 7.14 ([5]).
Consider a semiflow Ψ on (E, d) and a map $z : [0, +\infty) \to E$. Assume the following:
i) Ψ admits a strict Lyapunov function V.
ii) The set Λ_Ψ of equilibrium points of Ψ is compact.
iii) $\mathsf{V}(\Lambda_\Psi)$ has an empty interior.
iv) z is an APT of Ψ.
v) $z([0, \infty))$ is relatively compact.
Then, $\bigcap_{t\geq 0} \overline{z([t, \infty))}$ is a compact connected subset of Λ_Ψ .

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

$$W_\delta(x, m, v) := V_\infty(x, m, v) - \delta\langle \nabla F(x), m\rangle + \delta\|S(x) - v\|^2 , \tag{7.9}$$

where we recall that $V_\infty(z) := \lim_{t\to\infty} V(t, z)$ for every $z \in \mathcal{Z}_+$ and V is defined by Eq.(3.4). Consider the set $\mathcal{E} := h_\infty^{-1}(\{0\})$ of all equilibrium points of (ODE$_\infty$), namely: $\mathcal{E} = \{(x, m, v) \in \mathcal{Z}_+ : \nabla F(x) = 0, m = 0, v = S(x)\}$. The set $\mathcal{E}$ is non-empty by Assumption 2.3.

PROPOSITION 7.15. *Let Assumptions 2.3, 2.4, 7.1 and 7.2 hold. Assume that $0 < b \leq 4a$. Let $K \subset \mathcal{Z}_+$ be a compact set. Define $K' := \overline{\{\Phi(t, z) : t \geq 0, z \in K\}}$. Let $\overline{\Phi} : [0, +\infty) \times K' \to K'$ be the restriction of the semiflow Φ to K'* i.e., *$\overline{\Phi}(t, z) = \Phi(t, z)$ for all $t \geq 0, z \in K'$. Then,*
i) K' is compact.
ii) $\overline{\Phi}$ is well defined and is a semiflow on K'.
iii) The set of equilibrium points of $\overline{\Phi}$ is equal to $\mathcal{E} \cap K'$.
iv) There exists $\delta > 0$ s.t. W_δ is a strict Lyapunov function for the semiflow $\overline{\Phi}$.

Proof. The first point is a consequence of Prop. 7.7. The second point stems from Prop. 7.13. The third point is immediate from the definition of $\mathcal{E}$ and the fact that $\overline{\Phi}$ is valued in K'. We now prove the last point. Consider $z \in K'$ and write $\overline{\Phi}_t(z)$ under the form $\overline{\Phi}_t(z) = (x(t), m(t), v(t))$. For *any* map $\mathsf{W} : \mathcal{Z}_+ \to \mathbb{R}$, define for all $t > 0$, $\mathcal{L}_\mathsf{W}(t) := \limsup_{s\to 0} s^{-1}(\mathsf{W}(\overline{\Phi}_{t+s}(z)) - \mathsf{W}(\overline{\Phi}_t(z)))$. Introduce $G(z) := -\langle \nabla F(x), m\rangle$ and $H(z) := \|S(x) - v\|^2$ for every $z = (x, m, v)$. Consider $\delta > 0$ (to be specified later on). We study $\mathcal{L}_{W_\delta} = \mathcal{L}_V + \delta\mathcal{L}_G + \delta\mathcal{L}_H$. Note that $\overline{\Phi}_t(z) \in K' \cap \mathcal{Z}_+^*$ for all $t > 0$ by Lemma 7.4. Thus, $t \mapsto V_\infty(\overline{\Phi}_t(z))$ is differentiable at any point $t > 0$ and the derivative coincides with $\mathcal{L}_V(t) = \dot{V}_\infty(\overline{\Phi}_t(z))$. Define $C_1 := \sup\{\|v\|_\infty : (x, m, v) \in K'\}$. Then, by Lemma 7.5, $\mathcal{L}_V(t) \leq -\varepsilon(\varepsilon + \sqrt{C_1})^{-2} \|m(t)\|^2$. Let $L_{\nabla F}$ be the Lipschitz constant of ∇F on $\{x : (x, m, v) \in K'\}$. For every $t > 0$,

$$\begin{aligned}
\mathcal{L}_G(t) &\leq \limsup_{s\to 0} s^{-1}\|\nabla F(x(t)) - \nabla F(x(t + s))\|\|m(t + s)\| - \langle \nabla F(x(t)), \dot{m}(t)\rangle \\
&\leq L_{\nabla F}\varepsilon^{-1}\|m(t)\|^2 - a\|\nabla F(x(t))\|^2 + a\langle \nabla F(x(t)), m(t)\rangle \\
&\leq -\frac{a}{2}\|\nabla F(x(t))\|^2 + \left(\frac{a}{2} + \frac{L_{\nabla F}}{\varepsilon}\right)\|m(t)\|^2 .
\end{aligned}$$

Denote by L_S the Lipschitz constant of S on $\{x : (x,m,v) \in K'\}$. For every $t > 0$,

$$\begin{aligned}\mathcal{L}_H(t) &= \limsup_{s\to 0} s^{-1}(\|S(x(t+s)) - S(x(t)) + S(x(t)) - v(t+s)\|^2 - \|S(x(t)) - v(t)\|^2) \\ &= -2\langle S(x(t)) - v(t), \dot v(t)\rangle + \limsup_{s\to 0} 2s^{-1}\langle S(x(t+s)) - S(x(t)), S(x(t)) - v(t+s)\rangle \\ &\le -2b\|S(x(t)) - v(t)\|^2 + 2L_S\varepsilon^{-1}\|m(t)\|\|S(x(t)) - v(t)\| \,.\end{aligned}$$

Using that $2\|m(t)\|\|S(x(t)) - v(t)\| \le \frac{L_S}{b\varepsilon}\|m(t)\|^2 + \frac{b\varepsilon}{L_S}\|S(x(t)) - v(t)\|^2$, we obtain $\mathcal{L}_H(t) \le -b\|S(x(t)) - v(t)\|^2 + \frac{L_S^2}{b\varepsilon^2}\|m(t)\|^2$. Hence, for every $t > 0$,

$$\mathcal{L}_{W_\delta}(t) \le -M(\delta)\|m(t)\|^2 - \frac{a\delta}{2}\|\nabla F(x(t))\|^2 - \delta b\|S(x(t)) - v(t)\|^2 \,.$$

where $M(\delta) := \varepsilon(\varepsilon + \sqrt{C_1})^{-2} - \frac{\delta L_S^2}{b\varepsilon^2} - \delta\left(\frac{a}{2} + \frac{L_{\nabla F}}{\varepsilon}\right)$. Choosing δ s.t. $M(\delta) > 0$,

$$\forall t > 0, \quad \mathcal{L}_{W_\delta}(t) \le -c\left(\|m(t)\|^2 + \|\nabla F(x(t))\|^2 + \|S(x(t)) - v(t)\|^2\right) , \tag{7.10}$$

where $c := \min\{M(\delta), \frac{a\delta}{2}, \delta b\}$. It can easily be seen that for every $z \in K'$, $t \mapsto W_\delta(\overline{\Phi}_t(z))$ is Lipschitz continuous, hence absolutely continuous. Its derivative almost everywhere coincides with $\mathcal{L}_{W_\delta}$, which is non-positive. Thus, W_δ is a Lyapunov function for $\overline{\Phi}$. We prove that the Lyapunov function is strict. Consider $z \in K'$ s.t. $W_\delta(\overline{\Phi}_t(z)) = W_\delta(z)$ for all $t > 0$. The derivative almost everywhere of $t \mapsto W_\delta(\overline{\Phi}_t(z))$ is identically zero, and by Eq. (7.10), this implies that $-c\left(\|m_t\|^2 + \|\nabla F(x_t)\|^2 + \|S(x_t) - v_t\|^2\right)$ is equal to zero for every t a.e. (hence, for every t, by continuity of $\overline{\Phi}$). In particular for $t = 0$, $m = \nabla F(x) = 0$ and $S(x) - v = 0$. Hence, $z \in h_\infty^{-1}(\{0\})$. ∎

Corollary 7.16. *Let Assumptions 2.3, 2.4, 7.1 and 7.2 hold. Assume that $0 < b \le 4a$. For every $z \in \mathcal{Z}_+$, $\lim_{t\to\infty} \mathsf{d}(\Phi(z,t), \mathcal{E}) = 0$.*

Proof. Use Prop. 7.15 with $K := \{z\}$. and [15, Th. 2.1.7]. ∎

7.3.2. Asymptotic Behavior of the Solution to (ODE).

Proposition 7.17 (APT). *Let Assumptions 2.3, 2.4, 7.1 and 7.2 hold true. Assume that $0 < b \le 4a$. Then, for every $z_0 \in \mathcal{Z}_0$, $Z_\infty^0(z_0)$ is an asymptotic pseudotrajectory of the semiflow Φ given by (7.8).*

Proof. Consider $z_0 \in \mathcal{Z}_0$, $T \in (0, +\infty)$ and define $z := Z_\infty^0(z_0)$. Consider $t \ge 1$. For every $s \ge 0$, define $\Delta_t(s) := \|z(t+s) - \Phi(z(t))(s)\|$. The aim is to prove that $\sup_{s\in[0,T]} \Delta_t(s)$ tends to zero as $t \to \infty$. Putting together Prop. 7.6 and Lemma 7.8, the set $K := \overline{\{z(t) : t \ge 1\}}$ is a compact subset of $\mathcal{Z}_+^*$. Define $C(t) := \sup_{s\ge 0}\sup_{z'\in K}\|h(t+s,z') - h_\infty(z')\|$. It can be shown that $\lim_{t\to\infty} C(t) = 0$. We obtain that for every $s \in [0,T]$, $\Delta_t(s) \le TC(t) + \int_0^s \|h_\infty(z(t+s')) - h_\infty(\Phi(z(t))(s'))\|ds'$. By Lemma 7.8, $K' := \overline{\bigcup_{z'\in\Phi(K)} z'([0,\infty))}$ is a compact subset of $\mathcal{Z}_+^*$. It is immediately seen from the definition that h_∞ is Lipschitz continuous on every compact subset of $\mathcal{Z}_+^*$, hence on $K \cup K'$. Therefore, there exists a constant L, independent from t, s, s.t. $\Delta_t(s) \le TC(t) + \int_0^s L\Delta_t(s')ds' \qquad (\forall t \ge 1, \forall s \in [0,T])$. Using Grönwall's lemma, it holds that for all $s \in [0,T]$, $\Delta_t(s) \le TC(t)e^{Ls}$. As a consequence, $\sup_{s\in[0,T]} \Delta_t(s) \le TC(t)e^{LT}$ and the righthand side converges to zero as $t \to \infty$. ∎

End of the Proof of Th. 3.2. By Prop. 7.6, the set $K := \overline{Z^0_\infty(z_0)([0,\infty))}$ is a compact subset of $\mathcal{Z}_+$. Define $K' := \overline{\{\Phi(t,z) : t \geq 0, z \in K\}}$, and let $\overline{\Phi} : [0,+\infty) \times K' \to K'$ be the restriction Φ to K'. By Prop. 7.15, there exists $\delta > 0$ s.t. W_δ is a strict Lyapunov function for the semiflow $\overline{\Phi}$. Moreover, the set of equilibrium points coincides with $\mathcal{E} \cap K'$. In particular, the equilibrium points of $\overline{\Phi}$ form a compact set. By Prop. 7.17, $Z^0_\infty(z_0)$ is an APT of $\overline{\Phi}$. Note that every $z \in \mathcal{E}$ can be written under the form $z = (x, 0, S(x))$ for some $x \in \mathcal{S}$. From the definition of W_δ in (7.9), $W_\delta(z) = W_\delta(x,0,S(x)) = V_\infty(x,0,S(x)) = F(x)$. Since $F(\mathcal{S})$ is assumed to have an empty interior, the same holds for $W_\delta(\mathcal{E} \cap K')$. By Prop. 7.14, $\bigcap_{t\geq 0} \overline{Z^0_\infty(z_0)([t,\infty))} \subset \mathcal{E} \cap K'$. The set in the righthand side coincides with the set of limits of convergent sequences of the form $Z^0_\infty(z_0)(t_n)$ for $t_n \to \infty$. As $Z^0_\infty(z_0)([0,\infty))$ is a bounded set, $\mathsf{d}(Z^0_\infty(z_0)(t), \mathcal{E})$ tends to zero.

7.4. Proof of Th. 3.4. The proof follows the path of [16, Th. 10.1.6, Th. 10.2.3], but requires specific adaptations to deal with the dynamical system at hand. Define for all $\delta > 0$, $t > 0$, and $z = (x, m, v)$,

$$\tilde{W}_\delta(t,(x,m,v)) := V(t,(x,m,v)) - \delta\langle \nabla F(x), m\rangle + \delta\|S(x) - v\|^2 . \tag{7.11}$$

The function $\tilde{W}_\delta$ is the non-autonomous version of the function (7.9). Consider a fixed $x_0 \in \mathbb{R}^d$, and define $w_\delta(t) := \tilde{W}_\delta(t, z(t))$ where $z(t) = (x(t), m(t), v(t))$ is the solution to (ODE) with initial condition $(x_0, 0, 0)$. The proof uses the following steps.

i) Upper-bound on $w_\delta(t)$. From Eq. (3.4), we obtain that for every $t \geq 1$, $V(t, z(t)) \leq |F(x(t))| + \frac{\|m(t)\|^2}{2a\varepsilon(1-e^{-a})}$. Using $\langle \nabla F(x), m\rangle \leq (\|\nabla F(x)\|^2 + \|m\|^2)/2$, we obtain that there exists a constant c_1 (depending on δ) s.t. for every $t \geq 1$,

$$w_\delta(t) \leq c_1 \left(|F(x(t))| + \|m(t)\|^2 + \|\nabla F(x(t))\|^2 + \|S(x(t)) - v(t)\|^2\right) . \tag{7.12}$$

ii) Upper-bound on $\frac{d}{dt} w_\delta(t)$. The function w_δ is absolutely continuous on $[1,+\infty)$. Moreover, there exist $\delta > 0$, $c_2 > 0$ (both depending on x_0) s.t. for every $t \geq 1$ a.e.,

$$\frac{d}{dt} w_\delta(t) \leq -c_2 \left(\|m(t)\|^2 + \|\nabla F(x(t))\|^2 + \|S(x(t)) - v(t)\|^2\right) . \tag{7.13}$$

The proof of Eq. (7.13) uses arguments that are similar to the ones used in the proof of Prop. 7.15 (just use Lemma 7.5 to bound the derivative of the first term in Eq. (7.11)). For this reason, it is omitted.

iii) Positivity of $w_\delta(t)$. By Lemma 7.5, the function $t \mapsto V(t, z(t))$ is decreasing. As it is lower bounded, $\ell := \lim_{t\to\infty} V(t, z(t))$ exists. By Th. 3.2, $m(t)$ tends to zero, hence this limit coincides with $\lim_{t\to\infty} F(x(t))$. Replacing F with $F - \ell$, one can assume without loss of generality that $\ell = 0$. By Eq. (7.13), w_δ is non-increasing on $[1,+\infty)$, hence converging to some limit. Using again Th. 3.2, $\langle \nabla F(x(t)), m(t)\rangle \to 0$ and $S(x(t)) - v(t) \to 0$. Thus, $\lim_{t\to\infty} w_\delta(t) = \ell = 0$. Assume that there exists $t_0 \geq 1$ s.t. $w_\delta(t_0) = 0$. Then, w_δ is constant on $[t_0, +\infty)$. By Eq. (7.13), this implies that $m(t) = 0$ on this interval. Hence, $dx(t)/dt = 0$. This means that $x(t) = x(t_0)$ for all $t \geq t_0$. By Th. 3.2, it follows that $x(t_0) \in \mathcal{S}$. In that case, the final result is shown. Therefore, one can assume that $w_\delta(t) > 0$ for all $t \geq 1$.

iv) Putting together (7.12) and (7.13) using the Łojasiewicz condition. By Prop. 7.14 and 7.17, the set $L := \overline{\bigcup_{s\geq 0}\{z(t) : t \geq s\}}$ is a compact connected subset of $\mathcal{E} = \{(x, 0, S(x)) : \nabla F(x) = 0\}$. The set $\mathcal{U} := \{x : (x, 0, S(x)) \in L\}$ is a compact and connected subset of $\mathcal{S}$. Using Assumption 3.3 and [16, Lemma 2.1.6], there exist constants $\sigma, c > 0$ and $\theta \in (0, \frac{1}{2}]$, s.t. $\|\nabla F(x)\| \geq c|F(x)|^{1-\theta}$ for all x

s.t. $\mathsf{d}(x,\mathcal{U}) < \sigma$. As $\mathsf{d}(x(t),\mathcal{U}) \to 0$, there exists $T \geq 1$ s.t. for all $t \geq T$, $\|\nabla F(x(t))\| \geq c|F(x(t))|^{1-\theta}$. Thus, we may replace the term $\|\nabla F(x(t))\|^2$ in the righthand side of Eq. (7.13) using $\|\nabla F(x(t))\|^2 \geq \frac{1}{2}\|\nabla F(x(t))\|^2 + \frac{1}{2}|F(x(t))|^{2(1-\theta)}$. Upon noting that $2(1-\theta) \geq 1$, we thus obtain that there exists a constant c_3 and some $T' \geq 1$ s.t. for $t \geq T'$ a.e.,

$$\frac{d}{dt} w_\delta(t) \leq -c_3 \left(\|m(t)\|^2 + \|\nabla F(x(t))\|^2 + |F(x(t))| + \|S(x(t)) - v(t)\|^2 \right)^{2(1-\theta)} .$$

Putting this inequality together with Eq. (7.12), we obtain that for some constant $c_4 > 0$ and for all $t \geq T'$ a.e., $\frac{d}{dt} w_\delta(t) \leq -c_4 w_\delta(t)^{2(1-\theta)}$.

v) *End of the proof.* Following the arguments of [16, Th. 10.1.6], by integrating the preceding inequality, over $[T', t]$, we obtain $w_\delta(t) \leq c_5 t^{-\frac{1}{1-2\theta}}$ for $t \geq T'$ in the case where $\theta < \frac{1}{2}$, whereas $w_\delta(t)$ decays exponentially if $\theta = \frac{1}{2}$. From now on, we focus on the case $\theta < \frac{1}{2}$. By definition of (ODE), $\|\dot{x}(t)\|^2 \leq \|m(t)\|^2/((1-e^{-aT'})^2\varepsilon^2)$ for all $t \geq T'$. Since Eq. (7.13) implies $\|m(t)\|^2 \leq -\dot{w}_\delta(t)/c_2$, we deduce that there exists $c, c' > 0$ s.t. for all $t \geq T'$, $\int_t^{2t} \|\dot{x}(s)\|^2 ds \leq c w_\delta(t) \leq c' t^{-\frac{1}{1-2\theta}}$ .Applying [16, Lemma 2.1.5], it follows that $\int_t^\infty \|\dot{x}(s)\|^2 ds \leq c t^{-\frac{\theta}{1-2\theta}}$ for some other constant c. Therefore $x^* := \lim_{t\to+\infty} x(t)$ exists by Cauchy's criterion and for all $t \geq T'$, $\|x(t) - x^*\| \leq c t^{-\frac{\theta}{1-2\theta}}$. Finally, since $x(t) \to a$, we remark that, using the same arguments, the global Łojasiewicz exponent θ can be replaced by any Łojasiewicz exponent of f at x^*. When $\theta = \frac{1}{2}$, the proof follows the same line.

8. Proofs of Section 4.

8.1. Proof of Th. 4.3.

Given an initial point $x_0 \in \mathbb{R}^d$ and a stepsize $\gamma > 0$, we consider the iterates z_n^γ given by (2.5) and $z_0^\gamma := (x_0, 0, 0)$. For every $n \in \mathbb{N}^*$ and every $z \in \mathcal{Z}_+$, we define

$$H_\gamma(n, z, \xi) := \gamma^{-1}(T_{\gamma,\bar{\alpha}(\gamma),\bar{\beta}(\gamma)}(n, z, \xi) - z) .$$

Thus, $z_n^\gamma = z_{n-1}^\gamma + \gamma H_\gamma(n, z_{n-1}^\gamma, \xi_n)$ for every $n \in \mathbb{N}^*$. For every $n \in \mathbb{N}^*$ and every $z \in \mathcal{Z}$ of the form $z = (x, m, v)$, we define $e_\gamma(n, z) := (x, (1-\bar{\alpha}(\gamma)^n)^{-1} m, (1-\bar{\beta}(\gamma)^n)^{-1} v)$, and set $e_\gamma(0, z) := z$.

LEMMA 8.1. *Let Assumptions 2.2, 2.5 and 4.2 hold true. There exists $\bar{\gamma}_0 > 0$ s.t. for every $R > 0$, there exists $s > 0$,*

(8.1)
$$\sup \left\{ \mathbb{E}\left(\|H_\gamma(n+1, z, \xi)\|^{1+s} \right) : \gamma \in (0, \bar{\gamma}_0], n \in \mathbb{N}, z \in \mathcal{Z}_+ \text{ s.t. } \|e_\gamma(n, z)\| \leq R \right\} < \infty .$$

Proof. Let $R > 0$. We denote by $(H_{\gamma,\mathsf{x}}, H_{\gamma,\mathsf{m}}, H_{\gamma,\mathsf{v}})$ the block components of H_γ. There exists a constant C_s depending only on s s.t. $\|H_\gamma\|^{1+s} \leq C_s(\|H_{\gamma,\mathsf{x}}\|^{1+s} + \|H_{\gamma,\mathsf{m}}\|^{1+s} + \|H_{\gamma,\mathsf{v}}\|^{1+s})$. Hence, it is sufficient to prove that Eq. (8.1) holds respectively when replacing H_γ with each of $H_{\gamma,\mathsf{x}}, H_{\gamma,\mathsf{m}}, H_{\gamma,\mathsf{v}}$. Consider $z = (x, m, v)$ in $\mathcal{Z}_+$. We write: $\|H_{\gamma,\mathsf{x}}(n+1, z, \xi)\| \leq \varepsilon^{-1}(\|\frac{m}{1-\bar{\alpha}(\gamma)^n}\| + \|\nabla f(x, \xi)\|)$. Thus, for every z s.t. $\|e_\gamma(n, z)\| \leq R$, there exists a constant C depending only on ε, R and s s.t. $\|H_{\gamma,\mathsf{x}}(n+1, z, \xi)\|^{1+s} \leq C(1 + \|\nabla f(x, \xi)\|^{1+s})$. By Assumption 4.2, (8.1) holds for $H_{\gamma,\mathsf{x}}$ instead of H_γ. Similar arguments hold for $H_{\gamma,\mathsf{m}}$ and $H_{\gamma,\mathsf{v}}$ upon noting that the functions $\gamma \mapsto (1-\bar{\alpha}(\gamma))/\gamma$ and $\gamma \mapsto (1-\bar{\beta}(\gamma))/\gamma$ are bounded under Assumption 2.5. ∎

For every $R > 0$, and every arbitrary sequence $z = (z_n : n \in \mathbb{N})$ on $\mathcal{Z}_+$, we define $\tau_R(z) := \inf\{n \in \mathbb{N} : \|e_\gamma(n, z_n)\| > R\}$ with the convention that $\tau_R(z) = +\infty$

when the set is empty. We define the map $B_R : \mathcal{Z}_+^{\mathbb{N}} \to \mathcal{Z}_+^{\mathbb{N}}$ given for any arbitrary sequence $z = (z_n : n \in \mathbb{N})$ on $\mathcal{Z}_+$ by $B_R(z)(n) = z_n \mathbb{1}_{n<\tau_R(z)} + z_{\tau_R(z)} \mathbb{1}_{n\geq\tau_R(z)}$. We define the random sequence $z^{\gamma,R} := B_R(z^\gamma)$. Recall that a family $(X_i : i \in I)$ of random variables on some Euclidean space is called *uniformly integrable* if $\lim_{A\to+\infty} \sup_{i\in I} \mathbb{E}(\|X_i\| \mathbb{1}_{\|X_i\|>A}) = 0$.

LEMMA 8.2. *Let Assumptions 2.2, 2.5, 4.2 and 4.1 hold true. There exists $\bar\gamma_0 > 0$ s.t. for every $R > 0$, the family of r.v. $(\gamma^{-1}(z_{n+1}^{\gamma,R} - z_n^{\gamma,R}) : n \in \mathbb{N}, \gamma \in (0, \bar\gamma_0])$ is uniformly integrable.*

Proof. Let $R > 0$. As the event $\{n < \tau_R(z^\gamma)\}$ coincides with $\bigcap_{k=0}^n \{\|e_\gamma(k, z_k^\gamma)\| \leq R\}$, it holds that for every $n \in \mathbb{N}$,

$$\frac{z_{n+1}^{\gamma,R} - z_n^{\gamma,R}}{\gamma} = \frac{z_{n+1}^{\gamma} - z_n^{\gamma}}{\gamma} \mathbb{1}_{n<\tau_R(z^\gamma)} = H_\gamma(n+1, z_n^\gamma, \xi_{n+1}) \prod_{k=0}^n \mathbb{1}_{\|e_\gamma(k,z_k^\gamma)\|\leq R} \,.$$

Choose $\bar\gamma_0 > 0$ and $s > 0$ as in Lemma 8.1. For every $\gamma \leq \bar\gamma_0$,

$$\mathbb{E}\left(\left\|\gamma^{-1}(z_{n+1}^{\gamma,R} - z_n^{\gamma,R})\right\|^{1+s}\right) \leq \sup\left\{\mathbb{E}\left(\|H_{\gamma'}(\ell+1, z, \xi)\|^{1+s}\right) : \gamma' \in (0, \bar\gamma_0], \ell \in \mathbb{N}, z \in \mathcal{Z}_+, \|e_\gamma(\ell, z)\| \leq R\right\}.$$

By Lemma 8.1, the righthand side is finite and does not depend on (n, γ). ∎

For a fixed $\gamma > 0$, we define the interpolation map $\mathsf{X}_\gamma : \mathcal{Z}^{\mathbb{N}} \to C([0, +\infty), \mathcal{Z})$ as follows for every sequence $z = (z_n : n \in \mathbb{N})$ on $\mathcal{Z}$:

$$\mathsf{X}_\gamma(z) : t \mapsto z_{\lfloor \frac{t}{\gamma} \rfloor} + (t/\gamma - \lfloor t/\gamma \rfloor)(z_{\lfloor \frac{t}{\gamma} \rfloor + 1} - z_{\lfloor \frac{t}{\gamma} \rfloor}) \,.$$

For every $\gamma, R > 0$, we define $\mathsf{z}^{\gamma,R} := \mathsf{X}_\gamma(z^{\gamma,R}) = \mathsf{X}_\gamma \circ B_R(z^\gamma)$. Namely, $\mathsf{z}^{\gamma,R}$ is the interpolated process associated with the sequence $(z_n^{\gamma,R})$. It is a random variable on $C([0, +\infty), \mathcal{Z})$. We recall that $\mathcal{F}_n$ is the σ-algebra generated by the r.v. $(\xi_k : 1 \leq k \leq n)$. For every γ, n, R, we use the notation: $\Delta_0^{\gamma,R} := 0$ and

$$\Delta_{n+1}^{\gamma,R} := \gamma^{-1}(z_{n+1}^{\gamma,R} - z_n^{\gamma,R}) - \mathbb{E}(\gamma^{-1}(z_{n+1}^{\gamma,R} - z_n^{\gamma,R}) | \mathcal{F}_n) \,.$$

LEMMA 8.3. *Let Assumptions 2.2, 2.5, 4.2 and 4.1 hold true. There exists $\bar\gamma_0 > 0$ s.t. for every $R > 0$, the family of r.v. $(\mathsf{z}^{\gamma,R} : \gamma \in (0, \bar\gamma_0])$ is tight. Moreover, for every $\delta > 0$, $\mathbb{P}\left(\max_{0\leq n\leq \lfloor \frac{T}{\gamma} \rfloor} \gamma \left\|\sum_{k=0}^n \Delta_{k+1}^{\gamma,R}\right\| > \delta\right) \xrightarrow{\gamma\to 0} 0$.*

Proof. It is an immediate consequence of Lemma 8.2 and [6, Lemma 6.2] ∎

The proof of the following lemma is omitted.

LEMMA 8.4. *Let Assumptions 2.2 and 2.5 hold true. Consider $t > 0$ and $z \in \mathcal{Z}_+$. Let (φ_n, z_n) be a sequence on $\mathbb{N}^* \times \mathcal{Z}_+$ s.t. $\lim_{n\to\infty} \gamma_n \varphi_n = t$ and $\lim_{n\to\infty} z_n = z$. Then, $\lim_{n\to\infty} h_{\gamma_n}(\varphi_n, z_n) = h(t, z)$ and $\lim_{n\to\infty} e_{\gamma_n}(\varphi_n, z_n) = \bar e(t, z)$.*

End of the Proof of Th. 4.3 Consider $x_0 \in \mathbb{R}^d$ and set $z_0 = (x_0, 0, 0)$. Define $R_0 := \sup\left\{\|\bar e(t, Z_\infty^0(z_0)(t))\| : t > 0\right\}$. By Prop. 7.6, $R_0 < +\infty$. We select an arbitrary R s.t. $R \geq R_0 + 1$. For every $n \geq 0$, $z \in \mathcal{Z}_+$,

$$z_{n+1}^{\gamma,R} = z_n^{\gamma,R} + \gamma H_\gamma(n+1, z_n^{\gamma,R}, \xi_{n+1}) \mathbb{1}_{\|e_\gamma(n, z_n^{\gamma,R})\|\leq R} \,.$$

Define for every $n \geq 1$, $z \in \mathcal{Z}_+$, $h_{\gamma,R}(n, z) := h_\gamma(n, z) \mathbb{1}_{\|e_\gamma(n-1,z)\|\leq R}$. Then, $\Delta_{n+1}^{\gamma,R} = \gamma^{-1}(z_{n+1}^{\gamma,R} - z_n^{\gamma,R}) - h_{\gamma,R}(n+1, z_n^{\gamma,R})$. Define also for every $n \geq 0$,

$M_n^{\gamma,R} := \sum_{k=1}^n \Delta_k^{\gamma,R} = \gamma^{-1}(z_n^{\gamma,R} - z_0) - \sum_{k=0}^{n-1} h_{\gamma,R}(k+1, z_k^{\gamma,R})$. Consider $t \geq 0$ and set $n := \lfloor t/\gamma \rfloor$. For any $T > 0$, it holds that :

$$\sup_{t\in[0,T]} \left\| \mathsf{z}^{\gamma,R}(t) - z_0 - \int_0^t h_{\gamma,R}(\lfloor s/\gamma \rfloor + 1, \mathsf{z}^{\gamma,R}(\gamma\lfloor s/\gamma \rfloor))ds \right\| \leq \max_{0\leq n\leq \lfloor T/\gamma \rfloor + 1} \gamma \|M_n^{\gamma,R}\| .$$

By Lemma 8.3,

$$\mathbb{P}\left(\sup_{t\in[0,T]} \left\| \mathsf{z}^{\gamma,R}(t) - z_0 - \int_0^t h_{\gamma,R}\left(\lfloor s/\gamma \rfloor + 1, \mathsf{z}^{\gamma,R}(\gamma\lfloor s/\gamma \rfloor)\right) ds \right\| > \delta \right) \xrightarrow{\gamma\to 0} 0 . \tag{8.2}$$

As a second consequence of Lemma 8.3, the family of r.v. $(\mathsf{z}^{\gamma,R} : 0 < \gamma \leq \bar{\gamma}_0)$ is tight, where $\bar{\gamma}_0$ is chosen as in Lemma 8.3 (it does not depend on R). By Prokhorov's theorem, there exists a sequence $(\gamma_k : k \in \mathbb{N})$ s.t. $\gamma_k \to 0$ and s.t. $(\mathsf{z}^{\gamma_k,R} : k \in \mathbb{N})$ converges in distribution to some probability measure ν on $C([0,+\infty), \mathcal{Z}_+)$. By Skorohod's representation theorem, there exists a r.v. z on some probability space $(\Omega', \mathcal{F}', \mathbb{P}')$, with distribution ν, and a sequence of r.v. $(\mathsf{z}_{(k)} : k \in \mathbb{N})$ on that same probability space where for each $k \in \mathbb{N}$, the r.v. $\mathsf{z}_{(k)}$ has the same distribution as the r.v. $\mathsf{z}^{\gamma_k,R}$, and s.t. for every $\omega \in \Omega'$, $\mathsf{z}_{(k)}(\omega)$ converges to $\mathsf{z}(\omega)$ uniformly on compact sets. Now select a fixed $T > 0$. According to Eq. (8.2), the sequence

$$\sup_{t\in[0,T]} \left\| \mathsf{z}_{(k)}(t) - z_0 - \int_0^t h_{\gamma_k,R}\left(\lfloor s/\gamma_k \rfloor + 1, \mathsf{z}_{(k)}(\gamma_k\lfloor s/\gamma_k \rfloor)\right) ds \right\| ,$$

indexed by $k \in \mathbb{N}$, converges in probability to zero as $k \to \infty$. One can therefore extract a further subsequence $\mathsf{z}_{(\varphi_k)}$, s.t. the above sequence converges to zero almost surely. In particular, since $\mathsf{z}_{(k)}(t) \to \mathsf{z}(t)$ for every t, we obtain that

$$\mathsf{z}(t) = z_0 + \lim_{k\to\infty} \int_0^t h_{\gamma_{\varphi_k},R}\left(\lfloor s/\gamma_{\varphi_k} \rfloor + 1, \mathsf{z}_{(\varphi_k)}(\gamma_{\varphi_k}\lfloor s/\gamma_{\varphi_k} \rfloor)\right) ds \quad (\forall t \in [0,T]) . \tag{8.3}$$

Consider $\omega \in \Omega'$ s.t. the r.v. z satisfies (8.3) at point ω. From now on, we consider that ω is fixed, and we handle z as an element of $C([0,+\infty), \mathcal{Z}_+)$, and no longer as a random variable. Define $\tau := \inf\{t \in [0,T] : \|\bar{e}(t, \mathsf{z}(t))\| > R_0 + \frac{1}{2}\}$ if the latter set is non-empty, and $\tau := T$ otherwise. Since $\mathsf{z}(0) = z_0$ and $\|z_0\| < R_0$, it holds that $\tau > 0$ using the continuity of z. Choose any (s,t) s.t. $0 < s < t < \tau$. Note that $\mathsf{z}_{(k)}(\gamma_k\lfloor s/\gamma_k \rfloor) \to \mathsf{z}(s)$ and $\gamma_k(\lfloor s/\gamma_k \rfloor + 1) \to s$. Thus, by Lemma 8.4, $h_{\gamma_k}\left(\lfloor s/\gamma_k \rfloor + 1, \mathsf{z}_{(k)}(\gamma_k\lfloor s/\gamma_k \rfloor)\right)$ converges to $h(s, \mathsf{z}(s))$ and $e_{\gamma_k}(\lfloor s/\gamma_k \rfloor, \mathsf{z}_{(k)}(\gamma_k\lfloor s/\gamma_k \rfloor))$ converges to $\bar{e}(s, \mathsf{z}(z))$. Since $s < \tau$, $\bar{e}(s, \mathsf{z}(z)) \leq R_0 + \frac{1}{2}$. As $R \geq R_0 + 1$, there exists a certain $K(s)$ s.t. for every $k \geq K(s)$, $\mathbb{1}_{\|e_{\gamma_k}(\lfloor s/\gamma_k \rfloor, \mathsf{z}_{(k)}(\gamma_k\lfloor s/\gamma_k \rfloor))\| \leq R} = 1$. As a consequence, $h_{\gamma_k,R}(\lfloor s/\gamma_k \rfloor + 1, \mathsf{z}_{(k)}(\gamma_k\lfloor s/\gamma_k \rfloor))$ converges to $h(s, \mathsf{z}(s))$ as $k \to \infty$. Using Lebesgue's dominated convergence theorem, we obtain, for all $t \in [0, \tau]$: $\mathsf{z}(t) = z_0 + \int_0^t h(s, \mathsf{z}(s)) ds$. Therefore $\mathsf{z}(t) = Z_\infty^0(x_0)(t)$ for every $t \in [0, \tau]$. In particular, $\|\mathsf{z}(\tau)\| \leq R_0$ and this means that $\tau = T$. Thus, $\mathsf{z}(t) = Z_\infty^0(x_0)(t)$ for every $t \in [0,T]$ (and consequently for every $t \geq 0$). We have shown that for every $R \geq R_0 + 1$, the sequence of r.v. $(\mathsf{z}^{\gamma,R} : \gamma \in (0, \bar{\gamma}_0])$ is tight and converges in probability to $Z_\infty^0(z_0)$ as $\gamma \to 0$. Therefore, for every $T > 0$,

$$\forall \delta > 0, \ \lim_{\gamma\to 0} \mathbb{P}\left(\sup_{t\in[0,T]} \left\| \mathsf{z}^{\gamma,R}(t) - Z_\infty^0(x_0)(t) \right\| > \delta \right) = 0 . \tag{8.4}$$

In order to complete the proof, we show that $\mathbb{P}\left(\sup_{t\in[0,T]}\left\|\mathsf{z}^{\gamma,R}(t)-\mathsf{z}^{\gamma}(t)\right\|>\delta\right)\to 0$ as $\gamma\to 0$, for all $\delta>0$. where we recall that $\mathsf{z}^{\gamma}=\mathsf{X}_{\gamma}(z^{\gamma})$. Note that $\left\|\mathsf{z}^{\gamma,R}(t)\right\|\leq \left\|\mathsf{z}^{\gamma,R}(t)-Z^0_{\infty}(z_0)(t)\right\|+R_0$ by the triangular inequality. Therefore, for every $T,\delta>0$,

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in[0,T]}\left\|\mathsf{z}^{\gamma,R}(t)-\mathsf{z}^{\gamma}(t)\right\|>\delta\right)&\leq\mathbb{P}\left(\sup_{t\in[0,T]}\left\|\mathsf{z}^{\gamma,R}(t)\right\|\geq R\right)\\&\leq\mathbb{P}\left(\sup_{t\in[0,T]}\left\|\mathsf{z}^{\gamma,R}(t)-Z^0_{\infty}(z_0)(t)\right\|\geq R-R_0\right).\end{aligned}$$

By Eq. (8.4), the RHS of the above inequality tends to zero as $\gamma\to 0$. The proof is complete.

8.2. Proof of Th. 4.5. We start by stating a general result. Consider a Euclidean space X equipped with its Borel σ-field $\mathcal{X}$. Let $\bar\gamma_0>0$, and consider two families $(P_{\gamma,n}:0<\gamma<\bar\gamma_0, n\in\mathbb{N}^*)$ and $(\bar P_\gamma:0<\gamma<\bar\gamma_0)$ of Markov transition kernels on X. Denote by $\mathcal{P}(\mathsf{X})$ the set of probability measures on X. Let $X=(X_n:n\in\mathbb{N})$ be the canonical process on X. Let $(\mathbb{P}^{\gamma,\nu}:0<\gamma<\bar\gamma_0,\nu\in\mathcal{P}(\mathsf{X}))$ and $(\bar{\mathbb{P}}^{\gamma,\nu}:0<\gamma<\bar\gamma_0,\nu\in\mathcal{P}(\mathsf{X}))$ be two families of measures on the canonical space $(X^{\mathbb{N}},\mathcal{X}^{\otimes\mathbb{N}})$ such that the following holds:

- Under $\mathbb{P}^{\gamma,\nu}$, X is a non-homogeneous Markov chain with transition kernels $(P_{\gamma,n}:n\in\mathbb{N}^*)$ and initial distribution ν, that is, for each $n\in\mathbb{N}^*$, $\mathbb{P}^{\gamma,\nu}(X_n\in dx|X_{n-1})=P_{\gamma,n}(X_{n-1},dx)$.
- Under $\bar{\mathbb{P}}^{\gamma,\nu}$, X is an homogeneous Markov chain with transition kernel $\bar P_\gamma$ and initial distribution ν.

In the sequel, we will use the notation $\bar P^{\gamma,x}$ as a shorthand notation for $\bar P^{\gamma,\delta_x}$ where δ_x is the Dirac measure at some point $x\in\mathsf{X}$. Finally, let Ψ be a semiflow on X. A Markov kernel P is *Feller* if Pf is continuous for every bounded continuous f.

Assumption 8.5. Let $\nu\in\mathcal{P}(\mathsf{X})$.

i) For every γ, $\bar P_\gamma$ is Feller.

ii) $(\mathbb{P}^{\gamma,\nu}X_n^{-1}:n\in\mathbb{N},0<\gamma<\bar\gamma_0)$ is a tight family of measures.

iii) For every $\gamma\in(0,\bar\gamma_0)$ and every bounded Lipschitz continuous function $f:\mathsf{X}\to\mathbb{R}$, $P_{\gamma,n}f$ converges to $\bar P_\gamma f$ as $n\to\infty$, uniformly on compact sets.

iv) For every $\delta>0$, for every compact set $K\subset\mathsf{X}$, for every $t>0$, $\lim_{\gamma\to0}\sup_{x\in K}\bar P^{\gamma,x}\left(\|X_{\lfloor t/\gamma\rfloor}-\Psi_t(x)\|>\delta\right)=0$.

Let BC_Ψ be the Birkhof center of Ψ *i.e.*, the closure of the set of recurrent points.

THEOREM 8.6. *Consider $\nu\in\mathcal{P}(\mathsf{X})$ s.t. Assumption 8.5 holds true. Then, for every $\delta>0$,* $\lim_{\gamma\to0}\limsup_{n\to\infty}\frac{1}{n+1}\sum_{k=0}^n\mathbb{P}^{\gamma,\nu}\left(d(X_k,BC_\Psi)>\delta\right)=0$.

We omit the proof of this result which follows a similar reasoning to [6, Th. 5.5 and Proof in section 8.4] and makes use of results from [13].

End of the Proof of Th. 4.5. We apply Th. 8.6 in the case where $P_{\gamma,n}$ is the kernel of the non-homogeneous Markov chain (z_n^γ) defined by (2.5) and $\bar P_\gamma$ is the kernel of the homogeneous Markov chain $(\bar z_n^\gamma)$ given by $\bar z_n^\gamma=\bar z_{n-1}^\gamma+\gamma H_\gamma(\infty,\bar z_{n-1}^\gamma,\xi_n)$ for every $n\in\mathbb{N}^*$ and $\bar z_0\in\mathcal{Z}_+$ where $H_\gamma(\infty,\bar z_{n-1}^\gamma,\xi_n):=\lim_{k\to\infty}H_\gamma(k,\bar z_{n-1}^\gamma,\xi_n)$. The task is merely to verify Assumption 8.5*iii)*, the other assumptions being easily verifiable using Th. 4.3, Consider $\gamma\in(0,\bar\gamma_0)$. Let $f:\mathcal{Z}\to\mathbb{R}$ be a bounded M-Lipschitz continuous

function and K a compact. For all $z=(x,m,v)\in K$:

$$|P_{\gamma,n}(f)(z)-\bar{P}_\gamma(f)(z)| \le M\gamma\mathbb{E}\left\|\frac{(1-\alpha^n)^{-1}\tilde{m}_\xi}{\varepsilon+(1-\beta^n)^{-\frac{1}{2}}\tilde{v}_\xi^{1/2}}-\frac{\tilde{m}_\xi}{\varepsilon+\tilde{v}_\xi^{1/2}}\right\|$$
$$\le \tfrac{M\gamma\alpha^n}{\varepsilon(1-\alpha^n)}\sup_{x,m}\left(\alpha||m||+(1-\alpha)\mathbb{E}||\nabla f(x,\xi)||\right)+\tfrac{M\gamma\mathbb{E}||\tilde{m}_\xi\odot\tilde{v}_\xi^{1/2}||}{\varepsilon^2}\left(1-\tfrac{1}{(1-\beta^n)^{1/2}}\right)$$

where we write $\alpha=\bar{\alpha}(\gamma)$, $\beta=\bar{\beta}(\gamma)$, $\tilde{m}_\xi:=\alpha m+(1-\alpha)\nabla f(x,\xi)$ and $\tilde{v}_\xi:=\beta v+(1-\beta)\nabla f(x,\xi)^{\odot 2}$. Thus, condition 8.5*iii)* follows. Finally, Cor. 7.16 implies $BC_\Phi=\mathcal{E}$.

9. Proofs of Section 5. In this section, we denote by $\mathbb{E}_n=\mathbb{E}(\cdot|\mathcal{F}_n)$ the conditional expectation w.r.t. $\mathcal{F}_n$. We also use the notation $\nabla f_{n+1}:=\nabla f(x_n,\xi_{n+1})$.

The following lemma will be useful in the proofs.

LEMMA 9.1. *Let the sequence (r_n) be defined as in Algorithm 5.1. Assume that $0\le\alpha_n\le 1$ for all n and that $(1-\alpha_n)/\gamma_n\to a>0$ as $n\to+\infty$. Then,*

i) $\forall n\in\mathbb{N}, r_n=1-\prod_{i=1}^n\alpha_i$,

ii) The sequence (r_n) is nondecreasing and converges to 1.

iii) Under Assumption 5.6 i), for every $\epsilon>0$, for sufficiently large n, we have $r_n-1\le e^{-\frac{a\gamma_0}{2(1-\kappa)}n^{1-\kappa}}$ *if* $\kappa\in(0,1)$ *and* $r_n-1\le n^{-a\gamma_0/(1+\epsilon)}$ *if* $\kappa=1$.

A similar lemma holds for the sequence $(\bar{r}_n)$.

Proof. i) stems from observing that $r_{n+1}-1=\alpha_{n+1}(r_n-1)$ for every $n\in\mathbb{N}$ and iterating this relation ($r_0=0$). As a consequence, the sequence (r_n) is nondecreasing. We can write : $0\le 1-r_n\le\exp(-\sum_{i=1}^n(1-\alpha_i))$. iii) As $\sum_{n\ge1}\gamma_n=+\infty$ and $(1-\alpha_n)\sim a\gamma_n$, we deduce that $\sum_{i=1}^n(1-\alpha_i)\sim\sum_{i=1}^n a\gamma_i$. The results follow from the fact that $\sum_{i=1}^n\gamma_i\sim\frac{\gamma_0}{1-\kappa}n^{1-\kappa}$ when $\kappa\in(0,1)$ and $\sum_{i=1}^n\gamma_i\sim\gamma_0\ln n$ for $\kappa=1$. ∎

9.1. Proof of Th. 5.2. We define $\bar{z}_n=(x_{n-1},m_n,v_n)$ (note the shift in the index of the variable x). We have

$$\bar{z}_{n+1}=\bar{z}_n+\gamma_{n+1}h_\infty(\bar{z}_n)+\gamma_{n+1}\chi_{n+1}+\gamma_{n+1}\varsigma_{n+1}\,,$$

where h_∞ is defined in Eq. (7.1) and where we set

$$\chi_{n+1}=\left(0,\gamma_{n+1}^{-1}(1-\alpha_{n+1})(\nabla f_{n+1}-\nabla F(x_n)),\gamma_{n+1}^{-1}(1-\beta_{n+1})(\nabla f_{n+1}^{\odot 2}-S(x_n))\right)$$

and $\varsigma_{n+1}=(\varsigma_{n+1}^x,\varsigma_{n+1}^m,\varsigma_{n+1}^v)$ with the components defined by: $\varsigma_{n+1}^x=\frac{m_n}{\varepsilon+\sqrt{v_n}}-\frac{\gamma_n}{\gamma_{n+1}}\frac{\hat{m}_n}{\varepsilon+\sqrt{\hat{v}_n}}$, $\varsigma_{n+1}^m=\left(\frac{1-\alpha_{n+1}}{\gamma_{n+1}}-a\right)(\nabla F(x_n)-m_n)+a(\nabla F(x_n)-\nabla F(x_{n-1}))$ and $\varsigma_{n+1}^v=\left(\frac{1-\beta_{n+1}}{\gamma_{n+1}}-b\right)(S(x_n)-v_n)+b(S(x_n)-S(x_{n-1}))$. We prove that $\varsigma_n\to 0$ a.s. Using the triangular inequality,

$$\|\varsigma_n^x\|\le\left\|\frac{m_n}{\varepsilon+\sqrt{v_n}}-\frac{m_n}{\bar{r}_n^{1/2}\varepsilon+\sqrt{v_n}}\right\|+\left|1-\frac{\gamma_n r_n^{-1}}{\gamma_{n+1}\bar{r}_n^{-1/2}}\right|\left\|\frac{m_n}{\bar{r}_n^{1/2}\varepsilon+\sqrt{v_n}}\right\|$$
$$\le\varepsilon^{-1}|1-\bar{r}_n^{-1/2}|\|m_n\|+\varepsilon^{-1}\left|\bar{r}_n^{-1/2}-\frac{\gamma_n r_n^{-1}}{\gamma_{n+1}}\right|\|m_n\|\,,$$

which converges a.s. to zero because of the boundedness of (z_n) combined with Assumption 5.1 and Lemma 9.1 for $(\bar{r}_n)$. The components ς_{n+1}^m and ς_{n+1}^v converge a.s. to zero, as products of a bounded term and a term converging to zero. Indeed, note that

∇F and S are locally Lipschitz continuous under Assumption 2.2. Hence, 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. Now consider the martingale increment sequence (χ_n), adapted to $\mathcal{F}_n$. Estimating the second order moments, it is easy to show using Assumption 4.2 i) that there exists a constant C' s.t. $\mathbb{E}_n(\|\chi_{n+1}\|^2) \leq C'$. Using that $\sum_k \gamma_k^2 < \infty$, it follows that $\sum_n \mathbb{E}_n(\|\gamma_{n+1}\chi_{n+1}\|^2) < \infty$ a.s. By Doob's convergence theorem, $\lim_{n\to\infty} \sum_{k\leq n} \gamma_k\chi_k$ exists almost surely. Using this result along with the fact that ς_n converges a.s. to zero, it follows from usual stochastic approximation arguments [5] that the interpolated process $\bar{\mathsf{z}} : [0, +\infty) \to \mathcal{Z}_+$ given by

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

(where $\tau_n = \sum_{k=0}^{n} \gamma_k$), is almost surely a bounded APT of the semiflow $\bar{\Phi}$ defined by (ODE_∞). The proof is concluded by applying Prop. 7.14 and Prop. 7.15.

9.2. Proof of Prop. 5.4. As $\inf F > -\infty$, one can assume without loss of generality that $F \geq 0$. In the sequel, C denotes some positive constant which may change from line to line. We define $a_n := (1-\alpha_{n+1})/\gamma_n$ and $P_n := \frac{1}{2a_n r_n}\langle m_n^{\odot 2}, \frac{1}{\varepsilon+\sqrt{\hat{v}_n}}\rangle$. We have $a_n \to a$ and $r_n \to 1$. By Assumption 5.3-*i)*,

$$F(x_n) \leq F(x_{n-1}) - \gamma_n\langle \nabla F(x_n), \frac{\hat{m}_n}{\varepsilon + \sqrt{\hat{v}_n}}\rangle + C\gamma_n^2 P_n. \tag{9.1}$$

We set $u_n := 1 - \frac{a_{n+1}}{a_n}$ and $D_n := \frac{r_n^{-1}}{\varepsilon+\sqrt{\hat{v}_n}}$, so that $P_n = \frac{1}{2a_n}\langle D_n, m_n^{\odot 2}\rangle$. We can write:

$$P_{n+1} - P_n = u_n P_{n+1} + \langle \frac{D_{n+1} - D_n}{2a_n}, m_{n+1}^{\odot 2}\rangle + \langle \frac{D_n}{2a_n}, m_{n+1}^{\odot 2} - m_n^{\odot 2}\rangle. \tag{9.2}$$

We estimate the vector $D_{n+1} - D_n$. Using that (r_n^{-1}) is non-increasing,

$$D_{n+1} - D_n \leq r_n^{-1}\frac{\sqrt{\hat{v}_n} - \sqrt{\hat{v}_{n+1}}}{(\varepsilon + \sqrt{\hat{v}_{n+1}}) \odot (\varepsilon + \sqrt{\hat{v}_n})}\,.$$

Remarking that $v_{n+1} \geq \beta_{n+1}v_n$, recalling that $(\bar{r}_n)$ is nondecreasing and using the update rules of v_n and $\bar{r}_n$, we obtain after some algebra

$$\begin{aligned}\sqrt{\hat{v}_n} - \sqrt{\hat{v}_{n+1}} &= \bar{r}_{n+1}^{-\frac{1}{2}}(1 - \beta_{n+1})\frac{v_n - \nabla f_{n+1}^{\odot 2}}{\sqrt{v_n} + \sqrt{v_{n+1}}} + \frac{\bar{r}_{n+1} - \bar{r}_n}{\sqrt{\bar{r}_n}(\sqrt{\bar{r}_n} + \sqrt{\bar{r}_{n+1}})}\sqrt{\frac{v_n}{\bar{r}_{n+1}}} \\ &\leq c_{n+1}\sqrt{\hat{v}_{n+1}} \text{ where } c_{n+1} := \frac{1 - \beta_{n+1}}{\sqrt{\beta_{n+1}}}\left(\frac{1}{1 + \sqrt{\beta_{n+1}}} + \frac{1 - \bar{r}_n}{2\bar{r}_n}\right)\,.\end{aligned} \tag{9.3}$$

It is easy to see that $c_n/\gamma_n \to b/2$. Thus, for any $\delta > 0$, $c_{n+1} \leq (b + 2\delta)\gamma_n/2$ for all n large enough. Using also that $\sqrt{\hat{v}_{n+1}}/(\varepsilon + \sqrt{\hat{v}_{n+1}}) \leq 1$, we obtain that $D_{n+1} - D_n \leq \frac{b+2\delta}{2}\gamma_n D_n$. Substituting this inequality in Eq. (9.2), we get

$$P_{n+1} - P_n \leq u_n P_{n+1} + \gamma_n\langle \frac{b + 2\delta}{4a_n} D_n, m_{n+1}^{\odot 2}\rangle + \langle \frac{D_n}{2a_n}, m_{n+1}^{\odot 2} - m_n^{\odot 2}\rangle\,.$$

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 $\mathbb{E}_n(m_{n+1} - m_n) = a_n\gamma_n(\nabla F(x_n) - m_n)$,

$$\mathbb{E}_n\langle \frac{D_n}{2a_n}, m_{n+1}^{\odot 2} - m_n^{\odot 2}\rangle = \gamma_n\langle \nabla F(x_n), \frac{\hat{m}_n}{\varepsilon + \sqrt{\hat{v}_n}}\rangle - 2a_n\gamma_n P_n + \langle \frac{D_n}{2a_n}, \mathbb{E}_n[(m_{n+1} - m_n)^{\odot 2}]\rangle$$

As $a_n \to a$, we have $a_n - \frac{b+2\delta}{4} \geq a - \frac{b+\delta}{4}$ for all n large enough. Hence,

$$\mathbb{E}_n P_{n+1} - P_n \leq u_n P_{n+1} - 2(a - \frac{b+\delta}{4})\gamma_n P_n + \gamma_n \langle \nabla F(x_n), \frac{\hat{m}_n}{\varepsilon + \sqrt{\hat{v}_n}} \rangle + \gamma_n^2 \frac{b+2\delta}{2} \langle \nabla F(x_n), \frac{\hat{m}_n}{\varepsilon + \sqrt{\hat{v}_n}} \rangle + C \langle \frac{D_n}{2a_n}, \mathbb{E}_n[(m_{n+1} - m_n)^{\odot 2}] \rangle \,.$$

Using the Cauchy-Schwartz inequality and Assumption 5.3 *ii)*, it is easy to show the inequality $\langle \nabla F(x_n), \frac{\hat{m}_n}{\varepsilon + \sqrt{\hat{v}_n}} \rangle \leq C(1 + F(x_n) + P_n)$. Moreover, using the componentwise inequality $(\nabla f_{n+1} - m_n)^{\odot 2} \leq 2\nabla f_{n+1}^{\odot 2} + 2m_n^{\odot 2}$ along with Assumption 5.3 *ii)*, we obtain

$$\langle \tfrac{D_n}{2a_n}, \mathbb{E}_n[(m_{n+1} - m_n)^{\odot 2}] \rangle \leq 2(1-\alpha_{n+1})^2 \langle \tfrac{D_n}{2a_n}, \mathbb{E}_n[\nabla f_{n+1}^{\odot 2}] + m_n^{\odot 2} \rangle \leq C\gamma_n^2 (1 + F(x_n) + P_n) \,.$$

Putting all pieces together with Eq. (9.1),

(9.4)
$$\mathbb{E}_n(F(x_n) + P_{n+1}) \leq F(x_{n-1}) + P_n + u_n P_{n+1} - 2(a - \tfrac{b+\delta}{4})\gamma_n P_n + C\gamma_n^2 (1 + F(x_n) + P_n) \,.$$

Define $V_n := (1 - C\gamma_{n-1}^2)F(x_{n-1}) + (1 - u_{n-1})P_n$ where the constant C is fixed so that Eq. (9.4) holds. Then,

$$\mathbb{E}_n(V_{n+1}) \leq V_n - \left(2a - \frac{b+\delta}{2} - \frac{u_{n-1}}{\gamma_n}\right) \gamma_n P_n + C\gamma_n^2 (1 + P_n) + C\gamma_{n-1}^2 F(x_{n-1}) \,.$$

By Assumption 5.3, $\limsup_n u_{n-1}/\gamma_n < 2a - b/2$ and for δ small enough, we obtain

$$\mathbb{E}_n(V_{n+1}) \leq V_n + C\gamma_n^2 (1 + P_n) + C\gamma_{n-1}^2 F(x_{n-1}) \leq (1 + C'\gamma_n^2) V_n + C\gamma_n^2 \,.$$

By the Robbins-Siegmund's theorem [22], the sequence (V_n) converges almost surely to a finite random variable $V_\infty \in \mathbb{R}^+$. In turn, the coercivity of F implies that (x_n) is almost surely bounded. We now establish the almost sure boundedness of (m_n). Consider the martingale difference sequence $\Delta_{n+1} := \nabla f_{n+1} - \nabla F(x_n)$. We decompose $m_n = \bar{m}_n + \tilde{m}_n$ where $\bar{m}_{n+1} = \alpha_{n+1}\bar{m}_n + (1-\alpha_{n+1})\nabla F(x_n)$ and $\tilde{m}_{n+1} = \alpha_{n+1}\tilde{m}_n + (1-\alpha_{n+1})\Delta_{n+1}$, setting $\bar{m}_0 = \tilde{m}_0 = 0$. We prove that both terms $\bar{m}_n$ and $\tilde{m}_n$ are bounded. Consider the first term: $\|\bar{m}_{n+1}\| \leq \alpha_{n+1}\|\bar{m}_n\| + (1-\alpha_{n+1}) \sup_k \|\nabla F(x_k)\|$. By continuity of ∇F, the supremum in the above inequality is almost surely finite. Thus, for every n, the ratio $\|\bar{m}_n\| / \sup_k \|\nabla F(x_k)\|$ is upperbounded by the bounded sequence r_n. Hence, $(\bar{m}_n)$ is bounded w.p.1. Consider now the term $\tilde{m}_n$:

$$\mathbb{E}_n(\|\tilde{m}_{n+1}\|^2) = \alpha_{n+1}^2 \|\tilde{m}_n\|^2 + (1-\alpha_{n+1})^2 \mathbb{E}_n(\|\Delta_{n+1}\|^2) \leq (1 + (1-\alpha_{n+1})^2)\|\tilde{m}_n\|^2 + (1-\alpha_{n+1})^2 C \,,$$

where C is a constant s.t. $\mathbb{E}_n(\|\nabla f_{n+1}\|^2) \leq C$ by Assumption 4.2 i). Here, we used $\alpha_{n+1}^2 \leq (1 + (1-\alpha_{n+1})^2)$ and the inequality $\mathbb{E}_n(\|\Delta_{n+1}\|^2) \leq \mathbb{E}_n(\|\nabla f_{n+1}\|^2)$. By Assumption 5.1, $\sum_n (1-\alpha_{n+1})^2 < \infty$. By the Robbins-Siegmund theorem, it follows 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.

9.3. Proof of Th. 5.7. We use [20, Th. 1]. All the assumptions in the latter can be verified in our case, at the exception of a positive definiteness condition on the limiting covariance matrix, which corresponds, in our case, to the matrix Q given by Eq. (5.3). As Q is not positive definite, it is strictly speaking not possible to just cite and apply [20, Th. 1]. Nevertheless, a detailed inspection of the proofs of [20] shows that only a minor adaptation is needed in order to cover the present case. Therefore,

proving the convergence result of [20] from scratch is worthless. It is sufficient to verify the assumptions of [20, Th. 1] (except the definiteness of Q) and then to point out the specific part of the proof of [20] which requires some adaptation.

Let $z_n = (x_n, m_n, v_n)$ be the output of Algorithm 5.1. Define $z^* = (x^*, 0, S(x^*))$. Define $\eta_{n+1} := (0, a(\nabla f_{n+1} - \nabla F(x_n)), b(\nabla f_{n+1}^{\odot 2} - S(x_n)))$. We have

$$z_{n+1} = z_n + \gamma_{n+1} h_\infty(z_n) + \gamma_{n+1}\eta_{n+1} + \gamma_{n+1}\epsilon_{n+1}\,, \tag{9.5}$$

where $\epsilon_{n+1} := (\epsilon^1_{n+1}, \epsilon^2_{n+1}, \epsilon^3_{n+1})$, whose components are given by

$$\epsilon^1_{n+1} = \tfrac{m_n}{\varepsilon+\sqrt{v_n}} - \tfrac{\hat{m}_{n+1}}{\varepsilon+\sqrt{\hat{v}_{n+1}}};\ \epsilon^2_{n+1} = \left(\tfrac{1-\alpha_{n+1}}{\gamma_{n+1}} - a\right)(\nabla f_{n+1} - m_n)\,;\ \epsilon^3_{n+1} = \left(\tfrac{1-\beta_{n+1}}{\gamma_{n+1}} - b\right)(\nabla f_{n+1}^{\odot 2} - v_n)\,.$$

Here, η_{n+1} is a martingale increment noise and $\epsilon_{n+1} = (\epsilon^1_{n+1}, \epsilon^2_{n+1}, \epsilon^3_{n+1})$ is a remainder term. The aim is to check the assumptions (A1.1) to (A1.3) of [20], where the role of the quantities (h, ε_n, r_n, σ_n, α, ρ, β) in [20] is respectively played by the quantities (h_∞, η_n, ϵ_n, γ_n, κ, 1, 1) of the present paper.

Let us first consider Assumption (A1.1) for h_∞. By construction, $h_\infty(z^*) = 0$. By Assumptions 5.5 and 2.4, h_∞ is continuously differentiable in the neighborhood of z^* and its Jacobian at z^* coincides with the matrix H given by Eq. (5.1). As already discussed, after some algebra, it can be shown that the largest real part of the eigenvalues of H coincides with $-L$ where $L > 0$ is given by Eq. (5.2). Hence, Assumption (A1.1) of [20] is satisfied for h_∞. Assumption (A1.3) is trivially satisfied using Assumption 5.6. The crux is therefore to verify Assumption (A1.2). Clearly, $\mathbb{E}(\eta_{n+1}|\mathcal{F}_n) = 0$. Using Assumption 4.2ii), it follows from straightforward manipulations based on Jensen's inequality that for any $M > 0$, there exists $\delta > 0$ s.t. $\sup_{n\geq 0} \mathbb{E}_n\left(\|\eta_{n+1}\|^{2+\delta}\right) \mathbb{1}_{\{\|z_n - z^*\| \leq M\}} < \infty$. Next, we verify the condition

$$\lim_{n\to\infty} \mathbb{E}\left(\gamma_{n+1}^{-1}\|\epsilon_{n+1}\|^2 \mathbb{1}_{\{\|z_n - z^*\|\leq M\}}\right) = 0\,. \tag{9.6}$$

It is sufficient to verify the latter for ϵ^i_n $(i = 1, 2, 3)$ in place of ϵ_n. The map $(m, v) \mapsto m/(\varepsilon + \sqrt{v})$ is Lipschitz continuous in a neighborhood of $(0, S(x^*))$ by Assumption 2.4. Thus, for M small enough, there exists a constant C s.t. if $\|z_n - z^*\| \leq M$, then $\|\epsilon^1_{n+1}\| \leq C\left\|r_{n+1}^{-1} m_{n+1} - m_n\right\| + C\left\|\bar{r}_{n+1}^{-1} v_{n+1} - v_n\right\|$. Using the triangular inequality and the fact that $r_{n+1}, \bar{r}_{n+1}$ are bounded sequences away from zero, there exists an other constant C s.t.

$$\|\epsilon^1_{n+1}\| \leq C\,\|m_{n+1} - m_n\| + C\,\|v_{n+1} - v_n\| + C|r_{n+1} - 1| + C|\bar{r}_{n+1} - 1|\,.$$

Using Lemma 9.1 under Assumption 5.6 (note that $\gamma_0 > 1/2L \geq 1/a$ when $\kappa = 1$), we obtain that the sequence $|r_n - 1|/\gamma_n$ is bounded, thus $|r_{n+1} - 1| \leq C\gamma_{n+1}$. The sequence $(1 - \alpha_n)/\gamma_n$ being also bounded, it holds that

$$\|m_{n+1} - m_n\|^2 \mathbb{1}_{\{\|z_n - z^*\|\leq M\}} \leq C\gamma^2_{n+1}(1 + \|\nabla f_{n+1}\|^2)\mathbb{1}_{\{\|z_n - z^*\|\leq M\}}\,.$$

By Assumption 4.2 ii), $\mathbb{E}_n(\|\nabla f_{n+1}\|^2|)$ is bounded by a deterministic constant on $\{\|z_n - z^*\| \leq M\}$. Thus, $\mathbb{E}_n(\|m_{n+1} - m_n\|^2 \mathbb{1}_{\{\|z_n - z^*\|\leq M\}}) \leq C\gamma^2_{n+1}$. A similar result holds for $\|v_{n+1} - v_n\|^2$. We have thus shown that $\mathbb{E}_n\left(\|\epsilon^1_{n+1}\|^2 \mathbb{1}_{\{\|z_n - z^*\|\leq M\}}\right) \leq C\gamma^2_{n+1}$. Hence, Eq. (9.6) is proved for ϵ^1_{n+1} in place of ϵ_{n+1}. Under Assumption 5.6, the proof uses the same kind of arguments for ϵ^2_{n+1}, ϵ^3_{n+1} and is omitted. Finally, Eq. (9.6) is proved. Continuing the verification of Assumption (A1.2), we establish that

$$\mathbb{E}_n(\eta_{n+1}\eta_{n+1}^T) \to Q \text{ a.s. on } \{z_n \to z^*\}\,. \tag{9.7}$$

Denote by $\bar{Q}(x)$ the matrix given by the righthand side of Eq. (5.3) when x^* is replaced by an arbitrary $x \in \mathcal{V}$. It is easily checked that $\mathbb{E}_n(\eta_{n+1}\eta_{n+1}^T) = \bar{Q}(x_n)$ and by continuity, $\bar{Q}(x_n) \to Q$ a.s. on $\{z_n \to z^*\}$, which proves (9.7). Therefore, Assumption (A1.2) is fulfilled, except for the point mentioned at the beginning of this section : [20] puts the additional condition that the limit matrix in Eq. (9.7) is positive definite. This condition is not satisfied in our case, but the proof can still be adapted. The specific part of the proof where the positive definiteness comes into play is Th. 7 in [20]. The proof of [20, Th. 1] can therefore be adapted to the case of a positive semidefinite matrix. In the proof of [20, Th. 7], we only substitute the inverse of the square root of Q by the Moore-Penrose inverse. Finally, the uniqueness of the stationary distribution μ and its expression follow from [17, Th. 6.7, p. 357].

Proof of Eq. (5.4). We introduce the $d \times d$ blocks of the $3d \times 3d$ matrix $\Sigma = (\Sigma_{i,j})_{i,j=1,2,3}$ where $\Sigma_{i,j}$ is $d \times d$. We denote by $\tilde{\Sigma}$ the $2d \times 2d$ submatrix $\tilde{\Sigma} := (\Sigma_{i,j})_{i,j=1,2}$. By Th. 5.7, we have the subsystem:

$$\tilde{H}\tilde{\Sigma} + \tilde{\Sigma}\tilde{H}^T = \begin{pmatrix} 0 & 0 \\ 0 & -a^2\tilde{Q} \end{pmatrix} \qquad \text{where } \tilde{H} := \begin{pmatrix} \zeta I_d & -D \\ a\nabla^2 F(x^*) & (\zeta - a)I_d \end{pmatrix} \tag{9.8}$$

and where $\tilde{Q} := \mathrm{Cov}\,(\nabla f(x^*, \xi))$. The next step is to triangularize the matrix $\tilde{H}$ in order to decouple the blocks of $\tilde{\Sigma}$. For every $k = 1, \ldots, d$, set $\nu_k^{\pm} := -\frac{a}{2} \pm \sqrt{a^2/4 - a\lambda_k}$ with the convention that $\sqrt{-1} = \imath$ (inspecting the characteristic polynomial of $\tilde{H}$, these are the eigenvalues of $\tilde{H}$). Set $M^{\pm} := \mathrm{diag}\,(\nu_1^{\pm}, \cdots, \nu_d^{\pm})$ and $R^{\pm} := D^{-1/2} P M^{\pm} P^T D^{-1/2}$. Using the identities $M^+ + M^- = -aI_d$ and $M^+M^- = a\Lambda$ where $\Lambda := \mathrm{diag}\,(\lambda_1, \cdots, \lambda_d)$, it can be checked that

$$\mathcal{R}\tilde{H} = \begin{pmatrix} DR^+ + \zeta I_d & -D \\ 0 & R^-D + \zeta I_d \end{pmatrix} \mathcal{R}, \text{ where } \mathcal{R} := \begin{pmatrix} I_d & 0 \\ R^+ & I_d \end{pmatrix}.$$

Set $\check{\Sigma} := \mathcal{R}\tilde{\Sigma}\mathcal{R}^T$. Denote by $(\check{\Sigma}_{i,j})_{i,j=1,2}$ the blocks of $\check{\Sigma}$. Note that $\check{\Sigma}_{1,1} = \Sigma_{1,1}$. By left/right multiplication of Eq. (9.8) respectively with $\mathcal{R}$ and $\mathcal{R}^T$, we obtain

$$(DR^+ + \zeta I_d)\Sigma_{1,1} + \Sigma_{1,1}(R^+D + \zeta I_d) = \check{\Sigma}_{1,2}D + D\check{\Sigma}_{1,2}^T \tag{9.9}$$

$$(DR^+ + \zeta I_d)\check{\Sigma}_{1,2} + \check{\Sigma}_{1,2}(DR^- + \zeta I_d) = D\check{\Sigma}_{2,2} \tag{9.10}$$

$$(R^-D + \zeta I_d)\check{\Sigma}_{2,2} + \check{\Sigma}_{2,2}(DR^- + \zeta I_d) = -a^2\tilde{Q} \tag{9.11}$$

Set $\bar{\Sigma}_{2,2} = P^{-1}D^{1/2}\check{\Sigma}_{2,2}D^{1/2}P$. Define $C := P^{-1}D^{1/2}\tilde{Q}D^{1/2}P$. Eq. (9.11) yields $(M^- + \zeta I_d)\bar{\Sigma}_{2,2} + \bar{\Sigma}_{2,2}(M^- + \zeta I_d) = -a^2C$. Set $\bar{\Sigma}_{1,2} = P^{-1}D^{-1/2}\check{\Sigma}_{1,2}D^{1/2}P$. Eq. (9.10) rewrites $(M^+ + \zeta I_d)\bar{\Sigma}_{1,2} + \bar{\Sigma}_{1,2}(M^- + \zeta I_d) = \bar{\Sigma}_{2,2}$. We obtain that $\bar{\Sigma}_{1,2}^{k,\ell} = (\nu_k^+ + \nu_\ell^- + 2\zeta)^{-1}\bar{\Sigma}_{2,2}^{k,\ell} = \frac{-a^2C_{k,\ell}}{(\nu_k^+ + \nu_\ell^- + 2\zeta)(\nu_k^- + \nu_\ell^- + 2\zeta)}$. Set $\bar{\Sigma}_{1,1} = P^{-1}D^{-1/2}\Sigma_{1,1}D^{-1/2}P$. Eq. (9.9) becomes $(M^+ + \zeta I_d)\bar{\Sigma}_{1,1} + \bar{\Sigma}_{1,1}(M^+ + \zeta I_d) = \bar{\Sigma}_{1,2} + \bar{\Sigma}_{1,2}^T$. Thus,

$$\begin{aligned}\bar{\Sigma}_{1,1}^{k,\ell} &= \frac{\bar{\Sigma}_{1,2}^{k,\ell} + \bar{\Sigma}_{1,2}^{\ell,k}}{\nu_k^+ + \nu_\ell^+ + 2\zeta} = \frac{-a^2C_{k,\ell}}{(\nu_k^+ + \nu_\ell^+ + 2\zeta)(\nu_k^- + \nu_\ell^- + 2\zeta)}\left(\frac{1}{\nu_k^+ + \nu_\ell^- + 2\zeta} + \frac{1}{\nu_k^- + \nu_\ell^+ + 2\zeta}\right) \\ &= \frac{C_{k,\ell}}{(1 - \frac{2\zeta}{a})(\lambda_k + \lambda_\ell - 2\zeta + \frac{2}{a}\zeta^2) + \frac{1}{2(a-2\zeta)}(\lambda_k - \lambda_\ell)^2},\end{aligned}$$

and the result is proved.

REFERENCES

[1] H. Attouch and J. Bolte, *On the convergence of the proximal algorithm for nonsmooth functions involving analytic features*, Mathematical Programming, 116 (2009), pp. 5–16.
[2] H. Attouch, X. Goudou, and P. Redont, *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 (2000), pp. 1–34.
[3] L. Balles and P. Hennig, *Dissecting adam: The sign, magnitude and variance of stochastic gradients*, in Proceedings of the 35th International Conference on Machine Learning, vol. 80, 2018, pp. 404–413.
[4] A. Basu, S. De, A. Mukherjee, and E. Ullah, *Convergence guarantees for rmsprop and adam in non-convex optimization and their comparison to nesterov acceleration on autoencoders*, arXiv preprint arXiv:1807.06766, (2018).
[5] M. Benaïm, *Dynamics of stochastic approximation algorithms*, in Séminaire de Probabilités, XXXIII, vol. 1709 of Lecture Notes in Math., Springer, Berlin, 1999, pp. 1–68.
[6] P. Bianchi, W. Hachem, and A. Salim, *Constant step stochastic approximations involving differential inclusions: Stability, long-run convergence and applications*, Stochastics, 91 (2019), pp. 288–320.
[7] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming, 146 (2014), pp. 459–494.
[8] A. Cabot, H. Engler, and S. Gadat, *On the long time behavior of second order differential equations with asymptotically small dissipation*, Transactions of the American Mathematical Society, 361 (2009), pp. 5983–6017.
[9] X. Chen, S. Liu, R. Sun, and M. Hong, *On the convergence of a class of adam-type algorithms for non-convex optimization*, in International Conference on Learning Representations, 2019.
[10] A. B. da Silva and M. Gazeau, *A general system of differential equations to model first order adaptive algorithms*, arXiv preprint arXiv:1810.13108, (31 Oct 2018).
[11] D. Davis, D. Drusvyatskiy, S. Kakade, and J. Lee, *Stochastic subgradient method converges on tame functions*, Foundations of Computational Mathematics, 20 (2020), pp. 119–154.
[12] J. Duchi, E. Hazan, and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, Journal of Machine Learning Research, 12 (2011), pp. 2121–2159.
[13] J.-C. Fort and G. Pagès, *Asymptotic behavior of a Markovian stochastic algorithm with constant step*, SIAM J. Control Optim., 37 (1999), pp. 1456–1482 (electronic).
[14] S. Gadat, F. Panloup, and S. Saadane, *Stochastic heavy ball*, Electronic Journal of Statistics, 12 (2018), pp. 461–529.
[15] A. Haraux, *Systemes dynamiques dissipatifs et applications*, vol. 17, Masson, 1991.
[16] A. Haraux and M. Jendoubi, *The convergence problem for dissipative autonomous systems*, SpringerBriefs in Mathematics, Springer International Publishing, 2015.
[17] I. Karatzas and S. Shreve, *Brownian motion and stochastic calculus*, Springer-Verlag, New York, second ed., 1991.
[18] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, in International Conference on Learning Representations, 2015.
[19] S. Łojasiewicz, *Une propriété topologique des sous-ensembles analytiques réels*, Les équations aux dérivées partielles, 117 (1963), pp. 87–89.
[20] M. Pelletier, *Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing*, Annals of Applied Probability, (1998), pp. 10–44.
[21] S. J. Reddi, S. Kale, and S. Kumar, *On the convergence of adam and beyond*, in International Conference on Learning Representations, 2018.
[22] H. Robbins and D. Siegmund, *A convergence theorem for non negative almost supermartingales and some applications*, in Optimizing Methods in Statistics, Academic Press, New York, 1971, pp. 233–257.
[23] T. Tieleman and G. Hinton, *Lecture 6.e-rmsprop: Divide the gradient by a running average of its recent magnitude*, Coursera: Neural networks for machine learning, (2012), pp. 26–31.
[24] R. Ward, X. Wu, and L. Bottou, *AdaGrad stepsizes: Sharp convergence over nonconvex landscapes*, in Proceedings of the 36th International Conference on Machine Learning, vol. 97, 2019, pp. 6677–6686.
[25] M. Zaheer, S. J. Reddi, D. Sachan, S. Kale, and S. Kumar, *Adaptive methods for nonconvex optimization*, in Advances in Neural Information Processing Systems, 2018, pp. 9793–9803.
[26] D. Zhou, Y. Tang, Z. Yang, Y. Cao, and Q. Gu, *On the convergence of adaptive gradient methods for nonconvex optimization*, arXiv preprint arXiv:1808.05671, (2018).