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 the stochastic gradient descent for finding a local minimizer of a function. The objective function is unknown but a random estimate of the current gradient vector is observed at each round of the algorithm. 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. The key ingredient is the introduction of a continuous-time version of ADAM, under the form of a non-autonomous ordinary differential equation. The existence and the uniqueness of the solution are established, as well as the convergence of the solution towards the stationary points of the objective function. The continuous-time system is a relevant approximation of the ADAM iterates, in the sense that the interpolated ADAM process converges weakly to the solution to the ODE.

Key words. Stochastic approximation with constant step, Dynamical systems, Weak convergence of stochastic processes.

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. The stochastic gradient descent (SGD) is the most classical algorithm to search for such a minimizer [33]. Variants of SGD which include a momentum term have also become very popular [31, 29]. In these methods, the update equation depends on a parameter called the *learning rate*, which is generally assumed constant or vanishing. These algorithms 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.

In ADAM [25], the learning rate is adjusted coordinate-wise, as a function of the past values of the squared gradient vectors' coordinates. The algorithm thus combines the assets of momentum methods with an adaptive per-coordinate learning rate selection. Last but not least, the algorithm includes a so-called *bias correction* step acting on the current estimate of the gradient vector, which is revealed useful especially during the early iterations. However, despite its growing popularity, only few works investigate the behavior of the algorithm from a theoretical point of view (see the discussion in Section 2). The present paper studies the convergence of ADAM from a dynamical system viewpoint.

*LTCI, Télécom ParisTech, IP Paris, 75013, Paris, France. (firstname.name@telecom-paristech.fr)

Contributions

- We introduce a continuous-time version of ADAM, under the form of a non autonomous ordinary differential equation (ODE). Both the existence and the uniqueness of a global solution to the ODE turn out to be non trivial problems due to the irregularity of the vector field. The proof relies on the existence of a Lyapunov function for the ODE. We establish the convergence of the continuous-time ADAM trajectory to the set of stationary 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.
- We show that the discrete-time ADAM iterates shadow the behavior of the non-autonomous ODE in the asymptotic regime where the step size parameter γ of ADAM is small. More precisely, we consider the interpolated process $\mathsf{z}^\gamma(t)$ associated with the discrete-time version of ADAM, which consists in a piecewise linear interpolation of the iterates. The random process z^γ is indexed by the parameter γ, which is assumed constant during the whole run of the algorithm. We establish that when γ tends to zero, the interpolated process z^γ converges weakly[1] to the solution to the non-autonomous ODE.
- Under a stability condition, we prove the convergence of the discrete-time ADAM iterates in the doubly asymptotic regime where $n \to \infty$ then $\gamma \to 0$.

We claim that our analysis can be easily extended to other adaptive algorithms such as e.g. RMSPROP or ADAGRAD [38, 18] and AMSGRAD (see Section 2).

The paper is organized as follows. In Section 2, we provide a review of related works. In Section 3, we introduce the ADAM algorithm and the main assumptions. In Section 4, we introduce the continuous-time version of ADAM. In Section 5, our main results are stated. Section 6 is devoted to the proofs of existence and uniqueness of the solution to the ODE. Section 7 establishes the convergence of the continuous-time solution to the equilibrium points of the ODE. Section 8 establishes the weak convergence of the ADAM interpolated process towards the solution to the ODE. Section 9 proves the convergence in the long run of the iterates of ADAM. Finally, Section 10 contains numerical experiments sustaining our claims.

2. 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 [19]), ADAGRAD [18] led the way into a new class of algorithms sometimes refered to as adaptive gradient methods. ADAGRAD consists in 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 cumulated 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, see *e.g.* [35]. The work [38] introduced an unpublished but yet popular algorithm refered to as RMSPROP where the cumulated sum used in ADAGRAD is replaced by a moving average of squared gradients. Variants SC-ADAGRAD and SC-RMSPROP were proposed for strongly convex objectives with logarithmic regret bounds [28]. ADAM combines the

[1] in the space of continuous functions on $[0, +\infty)$ equipped with the topology of uniform convergence on compact sets.

advantages of both AdaGrad, RmsProp and momentum methods [31].

As opposed to AdaGrad, for which theoretical convergence guarantees exist [18, 16, 43, 39], Adam is comparatively less studied. The initial paper [25] suggests a $\mathcal{O}(\frac{1}{\sqrt{T}})$ average regret bound in the convex setting, but [32] 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 [32] to propose a variant called AmsGrad with the aim to recover, at least in the convex case, the sought guarantees. The work [6] 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 (the function f is non-random) of Adam is considered in [8]. Under quite specific values of the Adam-hyperparameters, it is shown that for every $\delta > 0$, there exists some time instant (non explicit, but with an explicit upper bound) for which the norm of the gradient of the objective at the current iterate is no larger than δ. The recent paper [16] provides a similar result for AmsGrad and AdaGrad, but the generalization to Adam is subject to conditions which are not easily verifiable. The paper [42] provides a convergence result for RmsProp. To that end, the objective F is used as a Lyapunov function, however our work suggests that unlike RmsProp, Adam does not admit F as a Lyapunov function, which makes the approach of [42] hardly generalizable to Adam. Moreover, [42] 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 weak limit of the (interpolated) Adam iterates as the step size tends to zero. The idea of approximating a discrete time stochastic system by a deterministic continuous one, often refered to as the ODE method, traces back to the works of [27] (see also [26]). The method can be summarized as follows. Given a certain stochastic algorithm parametrized by a step size γ, the interpolated process is the continuous piecewise linear function defined on $[0, +\infty)$ whose value coincides with the n-th iterate at time $n\gamma$. The interpolated process is a random variable on the space of continuous functions (equipped with the topology of uniform convergence on compact sets). As γ tends to zero, the aim of the ODE method is to establish the weak convergence of the interpolated process to a deterministic continuous function, generally defined as the unique solution to an ODE. This property is the crux to establish further convergence properties of the algorithm in the long run [11, 34, 13].

Recently, several works have raised a new interest in the analysis of deterministic continuous-time systems, as a way to understand the dynamics of numerical optimization algorithms [40, 41, 36]. A recent example is given by [37] which introduces a second-order continuous-time ODE to analyze Nesterov's accelerated gradient method [29] (see also [2, 5]). A generalization including an additional perturbation is provided by [3], where the rate of convergence of the continuous-time solutions is as well studied. This also generalizes earlier works of [4], where the so-called *heavy ball with friction* (HBF) dynamical system is introduced. It is shown that the continuous-time HBF solution converges towards a critical point of the objective function. The works [14, 15, 21] explore the asymptotic properties of a generalized HBF system with a vanishing time-dependent damping coefficient. Existence of global solutions is established and a Lyapunov function is introduced (see also [30]). The convergence towards the critical points of the objective function is shown under some hypotheses. The paper [22] studies a stochastic version of the celebrated heavy ball algorithm.

Algorithm 3.1 ADAM$(\gamma, \alpha, \beta, \varepsilon)$**.**

Input: data x_i, number of iterations n_{iter}.
Parameters: $\gamma > 0, \varepsilon > 0$, $(\alpha, \beta) \in [0,1)^2$.
Initialization: $x_0 \in \mathbb{R}^d, m_0 = 0$, $v_0 = 0$.
for $n = 1$ **to** n_{iter} **do**
$m_n = \alpha m_{n-1} + (1-\alpha)\nabla f(x_{n-1}, \xi_n)$
$v_n = \beta v_{n-1} + (1-\beta)\nabla f(x_{n-1}, \xi_n)^2$
$\hat{m}_n = m_n/(1-\alpha^n)$ {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

Almost sure convergence is established in a decreasing step size regime. The analysis again relies on the study of the deterministic continuous-time version of the algorithm.

We also point out [17] which is concomittant to the present paper ([17] was posted only four weeks after the first version of the present work [7]) and studies the asymptotic behavior of a similar dynamical system as the one introduced here. The work [17] establishes 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 present 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, [17] studies a slightly modified version of ADAM allowing to recover an ODE with a locally Lipschitz continuous vector field, whereas the original ADAM algorithm [25] leads on the otherhand to an ODE with an irregular vector field. This technical issue is tackled in the present paper.

3. The ADAM Algorithm.

Notations. If x, y are two vectors on $\mathbb{R}^d$, we denote by xy, x/y, x^α, $|x|$ the vectors on $\mathbb{R}^d$ whose k-th coordinates are respectively given by $x_k y_k$, x_k/y_k, x_k^α, $|x_k|$. Inequalities of the form $x \leq y$ are read componentwise. For any vector $v \in (0,+\infty)^d$, write $\|x\|_v^2 = \sum_k v_k x_k^2$. If (E, d) is a metric space, $z \in E$ and A is a non-empty subset of E, we use the notation $\mathsf{d}(z, A) := \inf\{\mathsf{d}(z, z') : z' \in A\}$.

3.1. 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 $f : \mathbb{R}^d \times \Xi \to \mathbb{R}$, 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 3.1. 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 $\mathcal{Z}_+$, $\xi \in \Xi$,

$$T_{\gamma,\alpha,\beta}(n, z, \xi) := \begin{pmatrix} x - \frac{\gamma(1-\alpha^n)^{-1}(\alpha m + (1-\alpha)\nabla f(x,\xi))}{\varepsilon + (1-\beta^n)^{-1/2}(\beta v + (1-\beta)\nabla f(x,\xi)^2)^{1/2}} \\ \alpha m + (1-\alpha)\nabla f(x,\xi) \\ \beta v + (1-\beta)\nabla f(x,\xi)^2 \end{pmatrix} . \tag{3.1}$$

Assumption 3.1. 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 3.1, 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:

$$F(x) := \mathbb{E}(f(x,\xi)) \tag{3.2}$$

$$S(x) := \mathbb{E}(\nabla f(x,\xi)^2) \tag{3.3}$$

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 3.2. F is coercive.

Assumption 3.3. 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 3.3 means that there is *no* point $x \in \mathbb{R}^d$ satisfying $\nabla f(x,\xi) = 0$ with probability one. This is a mild hypothesis in practice.

3.2. Asymptotic Regime. In this paper, we focus on the constant step size regime, where γ is fixed along the iterations (the default value recommended in [25] is $\gamma = 0.001$). As opposed to the decreasing step size context (*i.e.*, when γ vanishes along the iteration index n), here 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 [27, 26, 11]. 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:

$$\mathsf{z}^\gamma(t) := z_n^\gamma + (z_{n+1}^\gamma - z_n^\gamma)\left(\frac{t-n\gamma}{\gamma}\right). \tag{3.4}$$

We establish the weak convergence 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 refered 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 step size γ. The paper [25] recommends to choose 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. This boils down to $\alpha := \bar{\alpha}(\gamma)$ and $\beta := \bar{\beta}(\gamma)$, where $\bar{\alpha}$ and $\bar{\beta}$ are some mappings on $\mathbb{R}_+ \to [0,1)$ s.t. $\bar{\alpha}(\gamma)$ and $\bar{\beta}(\gamma)$ converge to one as $\gamma \to 0$.

Assumption 3.4. The functions $\bar{\alpha} : \mathbb{R}_+ \to [0,1)$ and $\bar{\beta} : \mathbb{R}_+ \to [0,1)$ are s.t. the following limits exist:

$$a := \lim_{\gamma \downarrow 0} \frac{1-\bar{\alpha}(\gamma)}{\gamma}, \quad b := \lim_{\gamma \downarrow 0} \frac{1-\bar{\beta}(\gamma)}{\gamma}. \tag{3.5}$$

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

Note that the condition $b \leq 4a$ is compatible with the default settings recommended by [25]. 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 step size $\gamma > 0$:

$$z_n^\gamma := T_{\gamma,\bar{\alpha}(\gamma),\bar{\beta}(\gamma)}(n, z_{n-1}^\gamma, \xi_n)\,, \tag{3.6}$$

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

4. Continuous-Time System. In order to have insights about the behavior of the sequence (z_n^γ) defined by (3.6), it is convenient to rewrite the ADAM iterations under the following equivalent form, for every $n \geq 1$:

$$z_n^\gamma = z_{n-1}^\gamma + \gamma h_\gamma(n, z_{n-1}^\gamma) + \gamma \Delta_n^\gamma\,, \tag{4.1}$$

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

$$h_\gamma(n, z) := \gamma^{-1}\mathbb{E}(T_{\gamma,\bar{\alpha}(\gamma),\bar{\beta}(\gamma)}(n, z, \xi) - z)\,, \tag{4.2}$$

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. $\xi_1, \ldots, \xi_n$. 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:

$$h(t, z) = \begin{pmatrix} -\frac{(1-e^{-at})^{-1} m}{\varepsilon + \sqrt{(1-e^{-bt})^{-1} v}} \\ a(\nabla F(x) - m) \\ b(S(x) - v) \end{pmatrix}\,, \tag{4.3}$$

where a, b are the constants defined in Assumption 3.4. 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. (4.1) suggests that, as $\gamma \downarrow 0$, the interpolated process z^γ shadows the non-autonomous differential equation

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

More formally, we shall demonstrate below that the family $(\mathsf{z}^\gamma : \gamma \in (0, \gamma_0])$ (where $\gamma_0 > 0$ is any fixed constant), interpreted as a family of r.v. on $C([0, +\infty), \mathcal{Z}_+)$ equipped with the topology of uniform convergence on compact sets, converges weakly as $\gamma \to 0$ to a solution to (ODE), under technical hypotheses. This legitimates the fact that (ODE) is a relevant approximation of the behavior of z^γ when γ is small.

Remark 4.1. Since $h(\,.\,, z)$ is non continuous at point zero for a fixed $z \in \mathcal{Z}_+$, and since moreover $h(t, \,.\,)$ is not locally Lipschitz continuous for a fixed $t > 0$, the existence and uniqueness of the solution to (ODE) cannot be directly solved using off-the-shelf theorems.

5. Main Results.

5.1. Continuous Time: Analysis of the ODE. Let $x_0 \in \mathbb{R}^d$. 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 5.1 (Existence and uniqueness). *Let Assumptions 3.1 to 3.4 hold true. Let $x_0 \in \mathbb{R}^d$. 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 $\mathcal{Z}_+$.*

THEOREM 5.2 (Convergence). *Let Assumptions 3.1 to 3.4 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, the set $\mathcal{S}$ is non-empty and $\lim_{t\to\infty} \mathsf{d}(x(t), \mathcal{S}) = 0$. Moreover, $\lim_{t\to\infty} m(t) = 0$, $\lim_{t\to\infty} S(x(t)) - v(t) = 0$.*

Denote by $z(t) = (x(t), m(t), v(t))$ the global solution to (ODE) issued from $(x_0, 0, 0)$.

Lyapunov function. The proof of Th. 5.1 (see section 6) relies on the existence of a Lyapunov function for the non-autonomous equation (ODE). By Lyapunov function, we mean a continuous function $V : (0, +\infty) \times \mathcal{Z}_+ \to \mathbb{R}$ s.t. $t \mapsto V(t, z(t))$ is decreasing on $(0, +\infty)$. Such a function V is given by:

$$V(t, z) := F(x) + \frac{1}{2} \|m\|^2_{U(t,v)^{-1}} \,, \tag{5.1}$$

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:

$$U(t, v) := a(1 - e^{-at}) \left(\varepsilon + \sqrt{\frac{v}{1 - e^{-bt}}} \right) . \tag{5.2}$$

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. 6.6). As a consequence,

$$\forall t \geq 0, \;\; F(x(t)) \leq F(x_0) \,. \tag{5.3}$$

This is an important feature of the algorithm. The (continuous-time) ADAM procedure *can only improve* the initial guess x_0. This is the consequence of the so-called bias correction step in ADAM *i.e.*, the fact that m_n and v_n are respectively divided by $(1 - \alpha^n)$ and $(1 - \beta^n)$ before being injected in the update of the iterate x_n. If the 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. 5.1 reveals that the initial derivative is given by $\dot{x}(0) = -\nabla F(x_0)/(\varepsilon + \sqrt{S(x_0)})$ (see Lemma 6.3). In the absence of debiasing step, 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.*, $S(x_0) \simeq \nabla F(x_0)^2$), 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, whose properties were discussed in previous works, see [12, 6].

ADAM as a Heavy Ball with Friction (HBF). It follows from our proof that the estimate $x(t)$ is twice differentiable and satisfies for every $t > 0$,

$$c_1(t)\, \ddot{x}(t) + c_2(t) \dot{x}(t) + \nabla F(x(t)) = 0 \,, \tag{5.4}$$

where $c_1(t) := a^{-2} U(t, v(t))$ and $c_2(t)$ is a term which can be explicited (the expression is omitted) and satisfies $c_2(t) > \frac{\dot{U}(t,v(t)))}{2a^2}$ for all $t > 0$. In the sense of (5.4), $x(t)$ can be interpreted as the solution to a generalized HBF problem, where both the mass of the particle and the viscosity depend on time [1, 4, 14, 22, 21].

5.2. Discrete Time: Convergence of ADAM.

Assumption 5.3. For every compact set $K \subset \mathbb{R}^d$, there exists $r_K > 0$ s.t.

$$\sup_{x \in K} \mathbb{E}(\|\nabla f(x, \xi)\|^{2+r_K}) < \infty \,.$$

Assumption 5.4. The sequence $(\xi_n : n \geq 1)$ is iid, with the same distribution as ξ.

THEOREM 5.5. *Let Assumptions 3.1 to 3.4, 5.3, and 5.4 hold true. 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 (3.6) and $z_0^\gamma = (x_0, 0, 0)$. Let z^γ be the corresponding interpolated process defined by Eq. (3.4). 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 5.6. There exists $\gamma_0 > 0$ s.t. the family of r.v. $(z_n^\gamma : n \in \mathbb{N}, 0 < \gamma < \gamma_0)$ is bounded in probability.

THEOREM 5.7. *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 (3.6) and $z_0^\gamma = (x_0, 0, 0)$. Under Assumptions 3.1 to 3.4, 5.3, 5.4, and 5.6, it holds that for every $\delta > 0$,*

$$\lim_{\gamma \downarrow 0} \limsup_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \mathbb{P}(\mathsf{d}(x_n^\gamma, \mathcal{S}) > \delta) = 0 \,. \tag{5.5}$$

Convergence in the long run. When the step size 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 $\gamma \to 0$. This doubly asymptotic regime is refered to as the convergence in the long run following the terminology of [34]. Theorem 5.7 establishes the convergence in the long run of the iterates of ADAM, in an ergodic sense.

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 Theorem 5.7 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. Theorem 5.5 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 $(x_0, 0, 0)$. Convergence in probability is understood here in the space $C([0, +\infty), \mathcal{Z}_+)$ of continuous functions on $[0, +\infty)$ endowed with the topology of uniform convergence on compact sets. 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 step size γ.

Stability. Assumption 5.6 is a stability condition ensuring that the iterates z_n^γ do not explode in the long run. A sufficient condition is for instance that $\sup_{n,\gamma} \mathbb{E}\|z_n^\gamma\| < \infty$. Checking this assumption is not easy, and left for future works. Note that in practice, a projection step on a compact set is often introduced in order to avoid numerical issues, in which case Assumption 5.6 is automatically satisfied.

6. Boundedness, Existence and Uniqueness.

6.1. Preliminaries. The results in this section are not specific to the case where F and S are defined as in Eq. (3.2)–(3.3): they are stated for *any* mappings F, S satisfying the following hypotheses.

Assumption 6.1. The function $F : \mathbb{R}^d \to \mathbb{R}$ is s.t.: F is continuously differentiable and ∇F is locally Lipschitz-continuous.

Assumption 6.2. 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. 5.1. 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 $(x, m, v) \in \mathcal{Z}_+$,

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

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. We shall exhibit a strict Lyapunov function for this dynamical system, and deduce that any solution to (ODE$_\infty$) converges to the set of equilibria of the dynamical system 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 the dynamical system. 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. 5.2.

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 6.3. *Let Assumptions 6.1 and 6.2 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 $[0, T)$, and it holds that $\dot{m}(0) = a\nabla F(x_0)$, $\dot{v}(0) = bS(x_0)$ 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 follows

from the fundamental theorem of calculus 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 6.4. *Let Assumptions 3.3, 6.1, and 6.2 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 $k \in \{1, \dots, d\}$. Assume by contradiction that there exists $t_0 \in (0, T)$ s.t. $v_k(t_0) < 0$. Set $\tau := \sup\{t \in [0, t_0] : v_k(t) \geq 0\}$. Clearly, $\tau < t_0$ and $v_k(\tau) = 0$ by the continuity of v_k. Since $v_k(t) \leq 0$ for all $t \in (\tau, t_0]$, it holds that $\dot{v}_k(t) = b(S_k(x(t)) - v_k(t))$ is non negative for all $t \in (\tau, t_0]$. This contradicts the fact that $v_k(\tau) > v_k(t_0)$. Thus, $v_k(t) \geq 0$ for all $t \in [0, T)$. Now assume by contradiction that there exists $t \in (0, T)$ s.t. $v_k(t) = 0$. Then, $\dot{v}_k(t) = bS_k(x(t)) > 0$. Thus, $\lim_{\delta \downarrow 0} \frac{v_k(t-\delta)}{-\delta} = bS_k(x(t))$. In particular, there exists $\delta > 0$ s.t. $v_k(t - \delta) \leq -\frac{\delta b}{2} S_k(x(t))$. This contradicts the first point. ∎

We define $V_\infty(z) := \lim_{t \to \infty} V(t, z)$ for every $z \in \mathcal{Z}_+$, and $U_\infty(v) := \lim_{t \to \infty} U(t, v) = a(\varepsilon + \sqrt{v})$ for every $v \in [0, +\infty)^d$.

LEMMA 6.5. *Let Assumptions 6.1 and 6.2 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 lines and can be found in [7, Lemma 5.3]. 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 obtain:

(6.1)
$$\partial_t V(t, z) = -\sum_{k=1}^d \frac{a^2 m_k^2}{U(t, v_k)^2} \left(\frac{e^{-at}\varepsilon}{2} + \left(\frac{e^{-at}}{2} - \frac{be^{-bt}(1 - e^{-at})}{4a(1 - e^{-bt})} \right) \sqrt{\frac{v_k}{1 - e^{-bt}}} \right) ,$$

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

$$\sum_{k=1}^d \frac{-a^2 m_k^2 (1 - e^{-at})}{U(t, v_k)^2} \left(\varepsilon + (1 - \frac{b}{4a}) \sqrt{\frac{v_k}{1 - e^{-bt}}} + \frac{bS_k(x)}{4a\sqrt{v_k(1 - e^{-bt})}} \right) .$$

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

$$\langle \nabla V(t, z), (1, h(t, z)) \rangle \leq -\sum_{k=1}^d \frac{a^2 m_k^2}{U(t, v_k)^2} \left((1 - \frac{e^{-at}}{2})\varepsilon + c_{a,b}(t) \sqrt{\frac{v_k}{1 - e^{-bt}}} \right) , \tag{6.2}$$

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 (6.2), the inequality (6.2) 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, $c_{a,b}(t) \geq Q(e^{-at}) \geq 0$.∎

6.2. 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 for every $t > 0$ and every $z = (x, m, v)$ in $\mathcal{Z}_+$ by:

$$\bar{e}(t, z) := (x, m/(1 - e^{-at}), v/(1 - e^{-bt})). \tag{6.3}$$

PROPOSITION 6.6. *Let Assumptions 3.2, 6.1, and 6.2 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. Consider $\eta \in [0, +\infty)$. 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 6.4, $t \mapsto V(t + \eta, z(t))$ is continuous on $[0, T)$, and continuously differentiable on $(0, T)$. By Lemma 6.5, $\dot{V}(t + \eta, z_\eta(t)) = \langle \nabla V(t + \eta, z_\eta(t)), (1, h(t + \eta, z_\eta(t)))\rangle \leq 0$ 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 $V(t + \eta, z_\eta(t)) \leq F(x_\eta(t)) + \frac{1}{2}\sum_{k=1}^d \frac{m_{\eta,k}(t)^2}{a(1-e^{-a(t+\eta)})\varepsilon}$. If $\eta > 0$, every term in the sum in the righthand side tends to zero, upon noting that $m_{\eta,k}(t) \to 0$ as $t \to 0$, for every $k \in \{1, \ldots, d\}$. The statement still holds if $\eta = 0$. Indeed, by Lemma 6.3, for a given $k \in \{1, \ldots, d\}$, there exists $\delta > 0$ s.t. for all $0 < t < \delta$, $m_{\eta,k}(t)^2 \leq 2a^2(\partial_k F(x_0))^2 t^2$ and $1 - e^{-at} \geq (at)/2$. As a consequence, each term of the sum in the righthand side of (4) is no larger than $4(\partial_k F(x_0))^2 t/\varepsilon$, 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 3.2.

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

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 6.7. *Let Assumptions 3.2, 6.1, and 6.2 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. 6.6 and is omitted. ∎

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

LEMMA 6.8. *Under Assumptions 3.2, 6.1, and 6.2, the following statements hold.*

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 $z(t) = (x(t), m(t), v(t))$, $v_k(t) \geq c \min\left(1, \frac{v_{\min}(K)}{2c} + t\right) \qquad (\forall t \geq 0, \forall k \in \{1,\dots,d\})$.

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

Proof. We prove the first point. Consider a compact set $K \subset \mathcal{Z}_+$. By Prop. 6.7, 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 $k \in \{1,\dots,d\}$ and consider $t \geq 0$. By the mean value theorem, there exists $t' \in [0,t]$ s.t. $v_k(t) = v_k(0) + \dot{v}_k(t')t$. Thus, $v_k(t) = v_k(0) + \dot{v}_k(0)t + b(S_k(x(t')) - v_k(t') - S_k(x(0)) + v_k(0))t$, which in turn implies $v_k(t) \geq v_k(0) + \dot{v}_k(0)t - bL_S\|x(t') - x(0)\|t - b|v_k(t') - v_k(0)|t$. 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.: $|v_k(t') - v_k(0)| = t'|\dot{v}_k(\tilde{t})| \leq t'bS_k(x(\tilde{t})) \leq tbM_2$. Putting together the above inequalities, $v_k(t) \geq v_k(0)(1 - bt) + bS_k(x(0))t - bCt^2$, where $C := (M_2 + L_S\sqrt{d}M_1)$. For every $t \leq 1/(2b)$, $v_k(t) \geq \frac{v_{\min}}{2} + tbC\left(\frac{S_{\min}}{C} - t\right)$, where we defined $S_{\min} := \inf\{S_k(x) : k \in \{1,\dots,d\}, (x,m,v) \in K\}$. Setting $\tau := 0.5\min(1/b, S_{\min}/C)$,

$$\forall t \in [0,\tau],\ v_k(t) \geq \frac{v_{\min}}{2} + \frac{bS_{\min}t}{2}\,. \tag{6.4}$$

Set $\kappa_1 := 0.5(v_{\min} + bS_{\min}\tau)$. Note that $v_k(\tau) \geq \kappa_1$. Define $S'_{\min} := \inf\{S_k(x) : k \in \{1,\dots,d\}, (x,m,v) \in K'\}$. Note that $S'_{\min} > 0$ by Assumptions 6.2 and 3.3. Finally, define $\kappa = 0.5\min(\kappa_1, S'_{\min})$. By contradiction, assume that the set $\{t \geq \tau : v_k(t) < \kappa\}$ is non empty, and denote by τ' its infimum. It is clear that $\tau' > \tau$ and $v_k(\tau') = \kappa$. Thus, $b^{-1}\dot{v}_k(\tau') = S_k(x(\tau')) - \kappa$. We obtain that $b^{-1}\dot{v}_k(\tau') \geq 0.5S'_{\min} > 0$. As a consequence, there exists $t \in (\tau, \tau')$ s.t. $v_k(t) < v_k(\tau')$. This contradicts the definition of τ'. We have shown that for all $t \geq \tau$, $v_k(t) \geq \kappa$. Putting this together with Eq. (6.4) and using that $\kappa \leq v_{\min} + bS_{\min}\tau$, we conclude that: $\forall t \geq 0$, $v_k(t) \geq \min\left(\kappa, \frac{v_{\min}}{2} + \frac{bS_{\min}t}{2}\right)$. Setting $c := \min(\kappa, bS_{\min}/2)$, the result follows.

We prove the second point. By Prop. 6.6, there exists a compact set $K \subset \mathcal{Z}_+$ s.t. for every $\eta \geq 0$, every $z \in Z_\infty^\eta(x_0)$ 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. ∎

6.3. Existence.

COROLLARY 6.9. *Let Assumptions 3.2, 6.1, and 6.2 hold. Assume that $0 < b \leq 4a$. For every $z_0 \in \mathcal{Z}_+$, $Z_\infty^\infty(z_0) \neq \emptyset$. For every $(z_0, \eta) \in \mathcal{Z}_0 \times (0, +\infty)$, $Z_\infty^\eta(z_0) \neq \emptyset$.*

Proof. We prove the first point (the proof of the second point follows the same lines). Under assumptions 3.2, 6.1 and 6.2, 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 $[0, T_{max})$ [24, Th. 2.1, Th. 3.1]. We conclude that the solution is global ($T_{max} = +\infty$) using the boundedness of the solution given by Prop. 6.7 and [24, Cor. 3.2]. ∎

LEMMA 6.10. *Let Assumptions 3.2, 6.1, and 6.2 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 $\eta > 0$, $z_\eta \in Z_\infty^\eta(z_0)$. 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. 6.6. By Prop. 6.6, 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 $k \in \{1, \dots, d\}$,

$$
\begin{aligned}
|x_{\eta,k}(t) - x_{\eta,k}(s)| &\leq \int_s^t \left| \frac{\hat{m}_{\eta,k}(u)}{\varepsilon + \sqrt{\hat{v}_{\eta,k}(u)}} \right| du \leq \frac{M_1}{\varepsilon} |t - s| \\
|m_{\eta,k}(t) - m_{\eta,k}(s)| &\leq \int_s^t a \, |\partial_k F(x_\eta(u)) - m_{\eta,k}(u)| \, du \leq a(M_1 + M_2)|t - s| \\
|v_{\eta,k}(t) - v_{\eta,k}(s)| &\leq \int_s^t b \, |S_k(x_\eta(u)) - v_{\eta,k}(u)| \, du \leq b(M_1 + M_2)|t - s| \, .
\end{aligned}
$$

Therefore, there exists a constant M_3, independent from η, s.t. for all $\eta > 0$ and all $(s,t) \in [0, +\infty)^2$, $\|z_\eta(t) - z_\eta(s)\| \leq M_3|t - s|$, which concludes the proof. ∎

PROPOSITION 6.11. *Let Assumptions 6.1 and 6.2 hold. Assume that $0 < b \leq 4a$. For every $z_0 \in \mathcal{Z}_0$, $Z_\infty^0(z_0) \neq \emptyset$* i.e., *(ODE) admits a global solution issued from z_0.*

Proof. By Corollary 6.9, there exists a family $(z_\eta)_{\eta>0}$ of functions on $[0, +\infty) \to \mathcal{Z}$ s.t. for every $\eta > 0$, $z_\eta \in Z_\infty^\eta(z_0)$. We set as usual $z_\eta(t) = (x_\eta(t), m_\eta(t), v_\eta(t))$. By Lemma 6.10, 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. 6.6, 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 6.8, 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 $z \in Z_\infty^0(z_0)$. ∎

6.4. Uniqueness.

PROPOSITION 6.12. *Let Assumptions 6.1 and 6.2 hold. Assume that $0 < b \leq 4a$.*

i) 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.*

ii) For every compact subset K of $\mathcal{Z}_+$, there exist non negative constants c_1, c_2 s.t. for every $(z, z') \in Z_\infty^\infty(K)^2$,

$$\forall t \geq 0,\ \|z(t) - z'(t)\|^2 \leq \|z(0) - z'(0)\|^2 \exp(c_1 + c_2 t)\,.$$

Proof. i) 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. 6.6, 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 6.1 and 6.2. 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

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

For every $k \in \{1, \dots, d\}$, $\left| \sqrt{\hat{v}_k(t)} - \sqrt{\hat{v}'_k(t)} \right| = \frac{|\hat{v}_k(t) - \hat{v}'_k(t)|}{|\sqrt{\hat{v}_k(t)} + \sqrt{\hat{v}'_k(t)}|}$. By Lemma 6.8, there exists $c > 0$ s.t. for all $t \geq 0$, $\hat{v}_k(t) \wedge \hat{v}'_k(t) \geq c \min(1, t)$. Thus,

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

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 $\delta > 0$,

$$\dot{u}_x(t) \leq \left(\frac{1}{\varepsilon\sqrt{\delta}} + \frac{M}{2\varepsilon^2 \sqrt{\delta c \min(1, t)}} \right) u^{(\delta)}(t)\,. \tag{6.5}$$

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,

$$\begin{aligned} \dot{u}_m(t) &= \frac{2a}{1 - e^{-at}} \langle \hat{m}(t) - \hat{m}'(t), \nabla F(x(t)) - \hat{m}(t) - \nabla F(x'(t)) + \hat{m}'(t) \rangle \\ &\leq \frac{2aL_{\nabla F}}{1 - e^{-at}} \|\hat{m}(t) - \hat{m}'(t)\| \|x(t) - x'(t)\|\,. \end{aligned}$$

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 $\delta > 0$,

$$\begin{aligned} \delta \dot{u}_m(t) &\leq \frac{a}{2(1 - e^{-at})} \left(4\delta L_{\nabla F}^2 u_x(t) + \delta u_m(t) \right) \\ &\leq \left(\frac{a}{2} + \frac{1}{2t} \right) \left(4\delta L_{\nabla F}^2 u_x(t) + \delta u_m(t) \right)\,, \end{aligned} \tag{6.6}$$

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

$$\delta \dot{u}_v(t) \leq \left(\frac{b}{2} + \frac{1}{2t} \right) \left(4\delta L_S^2 u_x(t) + \delta u_m(t) \right)\,. \tag{6.7}$$

We now choose any δ s.t. $4\delta \leq 1/\max(L_S^2, L_{\nabla F}^2)$. Then, Eq. (6.6) and (6.7) 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. (6.5), 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 6.3, 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 $k \in \{1, \ldots, d\}$. By the mean value theorem, there exists $\tilde{s}$ (resp. $\tilde{s}'$) in the interval $[0, t]$ s.t. $m_k(s) = \dot{m}_k(\tilde{s})s$ (resp. $m'_k(s) = \dot{m}'_k(\tilde{s}')s$). Thus, $\hat{m}_k(s) = \frac{as}{1-e^{-as}} \left(\partial_k F(x(\tilde{s})) - m_k(\tilde{s})\right)$, and a similar equality holds for $\hat{m}'_k(s)$. As a consequence,

$$\begin{aligned}
|\hat{m}_k(s) - \hat{m}'_k(s)| &\leq \frac{as}{1 - e^{-as}} \left(|\partial_k F(x(\tilde{s})) - \partial_k F(x'(\tilde{s}'))| + |m_k(\tilde{s}) - m'_k(\tilde{s}')|\right) \\
&\leq \frac{as}{1 - e^{-as}} \left(L_{\nabla F}\|x(\tilde{s}) - x'(\tilde{s}')\| + |m_k(\tilde{s}) - m'_k(\tilde{s}')|\right) \\
&\leq \frac{2a(L_{\nabla F} \vee 1)s}{1 - e^{-as}} \|z(\tilde{s}) - z'(\tilde{s}')\| ,
\end{aligned}$$

where we used $\|x(\tilde{s}) - x'(\tilde{s}')\| \leq \|z(\tilde{s}) - z'(\tilde{s}')\|$ and $|m_k(\tilde{s}) - m'_k(\tilde{s}')| \leq \|z(\tilde{s}) - z'(\tilde{s}')\|$ to obtain the last inequality. Using that $\tilde{s} \leq s$ and $\tilde{s}' \leq s$, it follows that:

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

By Lemma 6.3, z and z' are differentiable at point zero. Thus, the righthand side of the above inequality has a limit as $s \downarrow 0$: $\limsup_{s\downarrow 0} \frac{|\hat{m}_k(s) - \hat{m}'_k(s)|}{s} \leq 4(L_{\nabla F} \vee 1)\|\dot{z}(0)\|$. Thus,

$$\limsup_{s\downarrow 0} \frac{u_m(s)}{s^2} \leq 16d(L_{\nabla F}^2 \vee 1)\|\dot{z}(0)\|^2 .$$

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 $z = z'$.

ii) Consider the compact set K, and introduce the compact set $K' \subset \mathcal{Z}_+$ as in Prop. 6.7, and the constant $c > 0$ defined in Lemma 6.8. Define $K'_x = \{x : (x, m, v) \in K'\}$. The set is compact in $\mathbb{R}^d$. Respectively denote by L_S and $L_{\nabla F}$ the Lipschitz constants of S and ∇F on K'_x. Introduce the constant $M := \sup\{\|m\|_\infty : (x, m, v) \in K'\}$. Consider $(z_0, z'_0) \in K^2$ and two global solutions $z(\,.\,)$ and $z'(\,.\,)$ starting at z_0 and z'_0 respectively. We denote by $(x(t), m(t), v(t))$ the blocks of $z(t)$, and we define $(x'(t), m'(t), v'(t))$ similarly. Set $u(t) := \|z(t) - z'(t)\|^2$. Set also $u_x(t) := \|x(t) - x'(t)\|^2$ and define $u_m(t)$ and $u_v(t)$ similarly, hence, $u(t) = u_x(t) +$

$u_m(t) + u_v(t)$. Using the same derivations as above, we establish for all $t \geq 0$ that: $\dot{u}_m(t) \leq aL_{\nabla F} u_x(t) + a(L_{\nabla F} + 2)u_m(t)$. Similarly, $\dot{u}_v(t) \leq bL_S u_x(t) + b(L_S + 2)u_v(t)$. Moreover, $\dot{u}_x(t) \leq (\varepsilon^{-1} + \varepsilon^{-2} MC(t))u_x(t) + \varepsilon^{-1} u_m(t) + \varepsilon^{-2} MC(t) u_v(t)$ where we set $C(t) := \|(\sqrt{v(t)} + \sqrt{v'(t)})^{-1}\|_\infty$. Putting all pieces together, we obtain that there exists non negative constants c_1 and c_2, depending on K, s.t. $\dot{u}(t) \leq (c_1 + c_2 C(t))u(t)$. By Lemma 6.8, there exist two other non negative constants c_1', c_2' depending on K, s.t. for all $t > 0$, $\dot{u}(t) \leq (c_1' + c_2' \max(1, t^{-1/2}))u(t)$. Using Grönwall's Lemma, we obtain that for all $t \geq 0$, $u(t) \leq u(0) \exp\left(\int_0^t (c_1' + c_2' \max(1, s^{-1/2}))ds\right)$. It is easy to show that the integral in the exponential is no larger than $2c_2' + (c_1' + c_2')t$. ∎

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 6.13. *Let Assumptions 6.1 and 6.2 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_∞) starting from any given point in $\mathcal{Z}_+$. Moreover, the map*

$$\begin{array}{rccc} \Phi : [0, +\infty) \times \mathcal{Z}_+ & \to & \mathcal{Z}_+ \\ (t, z) & \mapsto & Z_\infty^\infty(z)(t) \end{array} \tag{6.8}$$

is a semiflow.

Proof. The result is a direct consequence of Lemma 6.12. ∎

7. Convergence of the Trajectories.

7.1. Convergence of the semiflow. In this paragraph, Ψ represents 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 [23, Th. 2.1.7]. A continuous function $z : [0, +\infty) \to E$ is said to be an asymptotic pseudotrajectory (APT, [10]) 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 [9, Th. 5.7] and [9, Prop. 6.4].

PROPOSITION 7.1 ([9]).
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, $\bigcap_{t \geq 0} \overline{z([t, \infty))} \subset \Lambda_\Psi$.

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

$$W_\delta(x, m, v) := V_\infty(x, m, v) - \delta \langle \nabla F(x), m \rangle + \delta \|S(x) - v\|^2, \tag{7.1}$$

where V_∞ is defined by Eq.(5.1). Consider the set $\mathcal{E} := h_\infty^{-1}(\{0\})$ of all equilibrium points of (ODE_∞), 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 3.2.

PROPOSITION 7.2. *Let Assumptions 6.1 and 6.2 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. 6.7. The second point is a consequence of Prop. 6.13. 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 6.4. Thus, $t \mapsto V_\infty(\overline{\Phi}_t(z))$ is differentiable at any point $t > 0$ and the derivative coincides with $\mathcal{L}_V(t) = \dot{V}_\infty(\overline{\Phi}_t(z))$. By Lemma 6.5,

$$\mathcal{L}_V(t) = \langle \nabla V_\infty(\overline{\Phi}_t(z)), h_\infty(\overline{\Phi}_t(z))\rangle \leq -\frac{\varepsilon}{(\varepsilon + \sqrt{\|v(t)\|_\infty})^2}\|m(t)\|^2 .$$

Define $C_1 := \sup\{\|v\|_\infty : (x,m,v) \in K'\}$. Then, $\mathcal{L}_V(t) \leq -\varepsilon(\varepsilon + \sqrt{C_1})^{-2}\|m(t)\|^2$. Let $L_{\nabla F}$ be the Lipschitz constant of ∇F on $\{x : (x,m,v) \in K'\}$. For every $t > 0$,

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

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

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

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

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

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. $M(\delta) > 0$,

$$\forall t > 0, \quad \mathcal{L}_{W_\delta}(t) \leq -c\left(\|m(t)\|^2 + \|\nabla F(x(t))\|^2 + \|S(x(t)) - v(t)\|^2\right) , \tag{7.2}$$

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. (7.2), 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 7.3. *Let Assumptions 6.1 and 6.2 hold. Assume that $0 < b \leq 4a$. For every $z \in \mathcal{Z}_+$, $\lim_{t\to\infty} \mathsf{d}(\Phi(z,t), \mathcal{E}) = 0$.*

Proof. Use Prop. 7.2 with $K := \{z\}$. and [23, Th. 2.1.7]. ∎

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

Proposition 7.4 (APT). *Let Assumptions 6.1 and 6.2 hold true. Assume that $0 < b \leq 4a$. Then, for every $z_0 \in \mathcal{Z}_0$, $Z_\infty^0(z_0)$ is an asymptotic pseudotrajectory of the semiflow Φ given by (6.8).*

Proof. Consider $z_0 \in \mathcal{Z}_0$, $T \in (0, +\infty)$ and define $z := Z_\infty^0(z_0)$. Consider $t \geq 1$. For every $s \geq 0$, 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. 6.6 and Lemma 6.8, the set $K := \overline{\{z(t) : t \geq 1\}}$ is a compact subset of $\mathcal{Z}_+^*$. Define $C(t) := \sup_{s\geq 0} \sup_{z'\in K} \|h(t+s, z') - h_\infty(z')\|$. It can be shown that $\lim_{t\to\infty} C(t) = 0$. We obtain that for every $s \in [0,T]$, $\Delta_t(s) \leq TC(t) + \int_0^s \|h_\infty(z(t+s')) - h_\infty(\Phi(z(t))(s'))\| ds'$. By Lemma 6.8, $K' := \overline{\bigcup_{z'\in\Phi(K)} z'([0,\infty))}$ is a compact subset of $\mathcal{Z}_+^*$. It is immediately seen from the definition that h_∞ is Lispchitz 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. $\Delta_t(s) \leq TC(t) + \int_0^s L\Delta_t(s')ds' \qquad (\forall t \geq 1, \forall s \in [0,T])$. Using Grönwall's lemma, it holds that for all $s \in [0,T]$, $\Delta_t(s) \leq TC(t)e^{Ls}$. As a consequence, $\sup_{s\in[0,T]} \Delta_t(s) \leq TC(t)e^{LT}$ and the righthand side converges to zero as $t \to \infty$. ∎

End of the Proof of Th. 5.2. By Prop. 6.6, the set $K := \overline{Z_\infty^0(z_0)([0,\infty))}$ 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. 7.2, 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. 7.4, $Z_\infty^0(z_0)$ 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 (7.1), $W_\delta(z) = W_\delta(x, 0, S(x)) = V_\infty(x, 0, S(x)) = F(x)$. Since $F(\mathcal{S})$ is assumed to have an empty interior, the same holds for $W_\delta(\mathcal{E} \cap K')$. By Prop. 7.1, $\bigcap_{t\geq 0} \overline{Z_\infty^0(z_0)([t,\infty))} \subset \mathcal{E} \cap K'$. The set in the righthand side coincides with the set of limits of convergent sequences of the form $Z_\infty^0(z_0)(t_n)$ for $t_n \to \infty$. As $Z_\infty^0(z_0)([0,\infty))$ is bounded set, $\mathsf{d}(Z_\infty^0(z_0)(t), \mathcal{E})$ tends to zero.

8. Proof of Theorem 5.5. Given an initial point $x_0 \in \mathbb{R}^d$ and a step size $\gamma > 0$, we consider the iterates z_n^γ given by (3.6) 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 3.1, 3.4, and 5.3 hold true. There exists $\gamma_0 > 0$ s.t. for every $R > 0$, there exists $r > 0$,*

(8.1)
$$\sup \left\{ \mathbb{E}\left(\|H_\gamma(n+1, z, \xi)\|^{1+r} \right) : \gamma \in (0, \gamma_0], n \in \mathbb{N}, z \in \mathcal{Z}_+ \text{ s.t. } \|e_\gamma(n, z)\| \le R \right\} < \infty \,.$$

Proof. By Assumption 3.4, the functions $\gamma \mapsto (1-\bar\alpha(\gamma))/\gamma$ and $\gamma \mapsto (1-\bar\beta(\gamma))/\gamma$ converge as $\gamma \downarrow 0$. Thus, there exists $\gamma_0 > 0$ and a constant $A > 0$ s.t. both functions are upper bounded by A on the interval $(0, \gamma_0]$. Let $R > 0$. By Assumption 5.3, there exists $r > 0$ and a finite constant $C > 0$ s.t. $\mathbb{E}(\|\nabla f(x, \xi)\|^{2+2r}) \le C$ for every x s.t. $\|x\| \le R$. We denote the block components of H_γ by $(H_{\gamma,\mathsf{x}}, H_{\gamma,\mathsf{m}}, H_{\gamma,\mathsf{v}}) := H_\gamma$. There exists a constant C_r depending only on r s.t. $\|H_\gamma\|^{1+r} \le C_r(\|H_{\gamma,\mathsf{x}}\|^{1+r} + \|H_{\gamma,\mathsf{m}}\|^{1+r} + \|H_{\gamma,\mathsf{v}}\|^{1+r})$. As a consequence, it is sufficient to prove that Eq. (8.1) holds respectively when replacing H_γ with each of its three components $H_{\gamma,\mathsf{x}}, H_{\gamma,\mathsf{m}}, H_{\gamma,\mathsf{v}}$. In the sequel, we write $\alpha := \bar\alpha(\gamma)$ and $\beta = \bar\beta(\gamma)$. Consider $z = (x, m, v)$ in $\mathcal{Z}_+$. We write: $\|H_{\gamma,\mathsf{x}}(n+1, z, \xi)\| \le \varepsilon^{-1}(\|\frac{m}{1-\alpha^n}\| + \|\nabla f(x, \xi)\|)$. Thus, for every z s.t. $\|e_\gamma(n, z)\| \le R$, there exists a constant C depending only on ε, R and r s.t. $\|H_{\gamma,\mathsf{x}}(n+1, z, \xi)\|^{1+r} \le C(1 + \|\nabla f(x, \xi)\|^{1+r})$. By Assumption 5.3, (8.1) holds for $H_{\gamma,\mathsf{x}}$ instead of H_γ. Consider $H_{\gamma,\mathsf{m}}$. For every $\gamma < \gamma_0$, it holds that: $\|H_{\gamma,\mathsf{m}}(n+1, z, \xi)\| = \frac{1-\alpha}{\gamma} \|\nabla f(x, \xi) - m\|$. For every z s.t. $\|e_\gamma(n, z)\| \le R$, $\|H_{\gamma,\mathsf{m}}(n+1, z, \xi)\| \le A(\|\nabla f(x, \xi)\| + R)$. Just as above, we deduce that $\mathbb{E}(\|H_{\gamma,\mathsf{x}}(n+1, z, \xi)\|^{1+r})$ is uniformly bounded on the set $\{(\gamma, n, z) : \gamma \in (0, \gamma_0], \|e_\gamma(n, z)\| \le R\}$. Finally, $H_{\gamma,\mathsf{v}}$ satisfies the same kind of inequality for every z s.t. $\|e_\gamma(n, z)\| \le R$, $\mathbb{E}(\|H_{\gamma,\mathsf{v}}(n+1, z, \xi)\|^{1+r} \le C'(1 + \mathbb{E}(\|\nabla f(x, \xi)\|^{2(1+r)}))$, which is again bounded uniformly in (γ, n, z) s.t. $\gamma \in (0, \gamma_0]$ and $\|e_\gamma(n, z)\| \le R$ by Assumption 5.3. ∎

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 $B_R(z)(n) = z_n \mathbb{1}_{n < \tau_R(z)} + z_{\tau_R(z)} \mathbb{1}_{n \ge \tau_R(z)}$. 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 3.1, 3.4, 5.3, and 5.4 hold true. There exists $\gamma_0 > 0$ s.t. for every $R > 0$, the family of r.v. $(\gamma^{-1}(z_{n+1}^{\gamma,R} - z_n^{\gamma,R}) : n \in \mathbb{N}, \gamma \in (0, \gamma_0])$ is uniformly integrable.*

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

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

Choose $\gamma_0 > 0$ and $r > 0$ as in Lemma 8.1. For every $\gamma \le \gamma_0$,

$$\begin{aligned} \mathbb{E}\left(\left\| \gamma^{-1}(z_{n+1}^{\gamma,R} - z_n^{\gamma,R}) \right\|^{1+r} \right) &\le \mathbb{E}\left(\|H_\gamma(n+1, z_n^\gamma, \xi_{n+1})\|^{1+r} \, \mathbb{1}_{\|e_\gamma(n, z_n^\gamma)\| \le R} \right) \\ &\le \sup \left\{ \mathbb{E}\left(\|H_{\gamma'}(\ell+1, z, \xi)\|^{1+r} \right) : \gamma' \in (0, \gamma_0], \ell \in \mathbb{N}, z \in \mathcal{Z}_+, \|e_\gamma(\ell, z)\| \le R \right\} . \end{aligned}$$

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

We endow the space $C([0,+\infty), \mathcal{Z})$ of continuous functions on $[0,+\infty) \to \mathcal{Z}$ with the topology of uniform convergence on compact sets. 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. $(\xi_k : 1 \le k \le n)$. For every γ, n, R, we use the notation:

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

and $\Delta_0^{\gamma,R} := 0$.

Lemma 8.3. *Let Assumptions 3.1, 3.4, 5.3, and 5.4 hold true. There exists $\gamma_0 > 0$ s.t. for every $R > 0$, the family of r.v. $(\mathsf{z}^{\gamma,R} : \gamma \in (0, \gamma_0])$ is tight. Moreover, for every $\delta > 0$,*

$$\mathbb{P}\left(\max_{0 \le n \le \lfloor \frac{T}{\gamma} \rfloor} \gamma \left\| \sum_{k=0}^{n} \Delta_{k+1}^{\gamma,R} \right\| > \delta \right) \xrightarrow{\gamma \to 0} 0 \, . \tag{8.2}$$

Proof. It is an immediate consequence of Lemma 8.2 and [13, Lemma 6.2] ∎

The proof of the following lemma is omitted and can be found in [7, Lemma 7.4].

Lemma 8.4. *Let Assumptions 3.1 and 3.4 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 Theorem 5.5 Consider $x_0 \in \mathbb{R}^d$ and set $z_0 = (x_0, 0, 0)$. Define $R_0 := \sup \left\{ \|\bar{e}(t, Z_\infty^0(x_0)(t))\| : t > 0 \right\}$. By Prop. 6.6, $R_0 < +\infty$. We select an arbitrary R s.t. $R \ge R_0 + 1$. For every $n \ge 0$, $z \in \mathcal{Z}_+$,

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

Thus, $\Delta_{n+1}^{\gamma,R} = \gamma^{-1}(z_{n+1}^{\gamma,R} - z_n^{\gamma,R}) - \mathbb{E}(H_\gamma(n+1, z_n^{\gamma,R}, \xi_n) \mathbb{1}_{\|e_\gamma(n, z_n^{\gamma,R})\| \le R} | \mathcal{F}_n)$, Define for every $n \ge 1$, $z \in \mathcal{Z}_+$, $h_{\gamma,R}(n,z) := h_\gamma(n,z) \mathbb{1}_{\|e_\gamma(n-1,z)\| \le R}$. 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 $n \ge 0$:

$$M_n^{\gamma,R} := \sum_{k=1}^{n} \Delta_k^{\gamma,R} = \gamma^{-1}(z_n^{\gamma,R} - z_0) - \sum_{k=0}^{n-1} h_{\gamma,R}(k+1, z_k^{\gamma,R}) \, .$$

Consider $t \ge 0$ and set $n := \lfloor t/\gamma \rfloor$. It holds that:

$$\mathsf{z}^{\gamma,R}(t) = z_0 + \int_0^t h_{\gamma,R}(\lfloor s/\gamma \rfloor + 1, \mathsf{z}^{\gamma,R}(\gamma \lfloor s/\gamma \rfloor)) ds + \gamma M_n^{\gamma,R} + (t - n\gamma) \Delta_{n+1}^{\gamma,R} \, .$$

As a consequence,

$$\left\| \mathsf{z}^{\gamma,R}(t) - z_0 - \int_0^t h_{\gamma,R}(\lfloor s/\gamma \rfloor + 1, \mathsf{z}^{\gamma,R}(\gamma \lfloor s/\gamma \rfloor)) ds \right\| \le \| \gamma M_n^{\gamma,R} + (t - n\gamma) \Delta_{n+1}^{\gamma,R} \| \, .$$

Therefore, for any $T > 0$,

$$\sup_{t\in[0,T]} \left\| \mathsf{z}^{\gamma,R}(t) - z_0 - \int_0^t h_{\gamma,R}(\lfloor s/\gamma \rfloor + 1, \mathsf{z}^{\gamma,R}(\gamma\lfloor s/\gamma\rfloor))ds \right\| \leq \sqrt{2} \max_{0\leq n\leq \lfloor T/\gamma\rfloor+1} \gamma \|M_n^{\gamma,R}\| .$$

By Lemma 8.3,

$$\mathbb{P}\left(\sup_{t\in[0,T]} \left\| \mathsf{z}^{\gamma,R}(t) - z_0 - \int_0^t h_{\gamma,R}\left(\lfloor s/\gamma \rfloor + 1, \mathsf{z}^{\gamma,R}(\gamma\lfloor s/\gamma\rfloor)\right) ds \right\| > \delta \right) \xrightarrow{\gamma\to 0} 0 . \tag{8.3}$$

As a second consequence of Lemma 8.3, the family of r.v. $(\mathsf{z}^{\gamma,R} : 0 < \gamma \leq \gamma_0)$ is tight, where γ_0 is chosen as in Lemma 8.3 (it does not depend on R). By Prokorov'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. $(\bar{\mathsf{z}}_{(k)} : k \in \mathbb{N})$ on that same probability space, 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. (8.3), 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 z_{φ_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

$$\mathsf{z}(t) = z_0 + \lim_{k\to\infty} \int_0^t h_{\gamma_{\varphi_k},R}\left(\lfloor s/\gamma_{\varphi_k} \rfloor + 1, \mathsf{z}_{(\varphi_k)}(\gamma_{\varphi_k}\lfloor s/\gamma_{\varphi_k}\rfloor)\right) ds \quad (\forall t \in [0,T]) . \tag{8.4}$$

Consider $\omega \in \Omega'$ s.t. the r.v. z satisfies (8.4) 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, $\|\mathsf{z}(\tau)\| \leq R_0$. Recalling the definition of τ, 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. $(\mathsf{z}^{\gamma,R} : \gamma \in (0,\gamma_0])$ is tight and converges weakly to $Z_\infty^0(x_0)$ as $\gamma \to 0$. Therefore, for every $T > 0$,

$$\forall \delta > 0, \ \lim_{\gamma\to 0} \mathbb{P}\left(\sup_{t\in[0,T]} \left\| \mathsf{z}^{\gamma,R}(t) - Z_\infty^0(x_0)(t) \right\| > \delta \right) = 0 . \tag{8.5}$$

In order to complete the proof, it is now sufficient to establish that:

$$\forall \delta > 0,\ \lim_{\gamma \to 0} \mathbb{P}\left(\sup_{t \in [0,T]} \left\| \mathsf{z}^{\gamma,R}(t) - \mathsf{z}^{\gamma}(t) \right\| > \delta \right) = 0\,, \tag{8.6}$$

where we recall that $\mathsf{z}^\gamma = sX^\gamma(z^\gamma)$. Note that for every $T, \delta > 0$,

$$\mathbb{P}\left(\sup_{t \in [0,T]} \left\| \mathsf{z}^{\gamma,R}(t) - \mathsf{z}^{\gamma}(t) \right\| > \delta \right) \leq \mathbb{P}\left(\sup_{t \in [0,T]} \left\| \mathsf{z}^{\gamma,R}(t) \right\| \geq R \right)\,.$$

By the triangular inequality, $\left\| \mathsf{z}^{\gamma,R}(t) \right\| \leq \left\| \mathsf{z}^{\gamma,R}(t) - Z_\infty(x_0)(t) \right\| + R_0$. Therefore,

$$\mathbb{P}\left(\sup_{t \in [0,T]} \left\| \mathsf{z}^{\gamma,R}(t) - \mathsf{z}^{\gamma}(t) \right\| > \delta \right) \leq \mathbb{P}\left(\sup_{t \in [0,T]} \left\| \mathsf{z}^{\gamma,R}(t) - Z_\infty(x_0)(t) \right\| \geq R - R_0 \right)\,.$$

By Eq. (8.5), the righthand side of the above inequality tends to zero as $\gamma \to 0$. This shows that Eq. (8.6) holds true. The proof is complete.

9. Proof of Theorem 5.7. We start by stating a general result. Consider an Euclidean space X equipped with its Borel σ-field $\mathcal{X}$. Let $\gamma_0 > 0$, and consider two families $(P_{\gamma,n} : 0 < \gamma < \gamma_0, n \in \mathbb{N}^*)$ and $(\bar{P}_\gamma : 0 < \gamma < \gamma_0)$ 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 $(\mathbb{P}^{\gamma,\nu} : 0 < \gamma < \gamma_0, \nu \in \mathcal{P}(\mathsf{X}))$ and $(\bar{\mathbb{P}}^{\gamma,\nu} : 0 < \gamma < \gamma_0, \nu \in \mathcal{P}(\mathsf{X}))$ 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 9.1. Let $\nu \in \mathcal{P}(\mathsf{X})$.

i) For every γ, $\bar{P}_\gamma$ is Feller.

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

iii) For every $\gamma \in (0, \gamma_0)$ 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 $t > 0$,

$$\lim_{\gamma \to 0} \sup_{x \in K} \bar{P}^{\gamma,x}\left(\|X_{\lfloor t/\gamma \rfloor} - \Psi_t(x)\| > \delta \right) = 0\,.$$

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

THEOREM 9.2. *Consider $\nu \in \mathcal{P}(\mathsf{X})$ s.t. Assumption 9.1 holds true. Then, for every $\delta > 0$, $\lim_{\gamma \to 0} \limsup_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^n \mathbb{P}^{\gamma,\nu}\left(d(X_k, BC_\Psi) > \delta \right) = 0$.*

Proof. For every γ, n, define $\mu_{\gamma,n} := \nu P_{\gamma,1} \cdots P_{\gamma,n}$ with the convention that $\mu_{\gamma,0} = \nu$. Otherwise stated, $\mu_{\gamma,n} = \mathbb{P}_{\gamma,\nu} X_n^{-1}$. Define $\Pi_{\gamma,n} := \frac{1}{n+1} \sum_{k=0}^n \mu_{\gamma,k}$ for every $n \in \mathbb{N}$. Assumption 9.1 implies that for any fixed γ, $(\Pi_{\gamma,n} : n \in \mathbb{N})$ is tight. By Prokhorov's theorem, it admits a cluster point π_γ. For such a cluster point, consider a subsequence

φ_n s.t. $\Pi_{\gamma,\varphi_n} \Rightarrow \pi_\gamma$, where $\Rightarrow$ stands for the weak convergence of probability measures. Consider a bounded Lipschitz-continuous function $f : \mathsf{X} \to \mathbb{R}$. It holds that $\Pi_{\gamma,n}(f)$ and $\Pi_{\gamma,n}(\bar{P}_\gamma f)$ respectively converge to $\pi_\gamma(f)$ and $\pi_\gamma(\bar{P}_\gamma f)$ along the subsequence, because $\bar{P}_\gamma$ is Feller. We observe that

$$\left|\Pi_{\gamma,n}\bar{P}_\gamma f - \Pi_{\gamma,n} f\right| \leq \frac{1}{n+1}\sum_{k=0}^{n} |\mu_{\gamma,k}(\bar{P}_\gamma f - P_{\gamma,k+1} f)| + \frac{2\|f\|_\infty}{n+1}\,.$$

Choose $\delta > 0$ and a compact set $K \subset \mathsf{X}$ s.t. $\sup_k \mu_{\gamma,k}(K^c) < \delta$. For every k, $|\mu_{\gamma,k}(\bar{P}_\gamma f - P_{\gamma,k+1} f)| \leq \sup_{x\in K} |\bar{P}_\gamma f(x) - P_{\gamma,k+1} f(x)| + 2\|f\|_\infty\delta$. By Assumption 9.1*iii)*, it holds that $\limsup_n \left|\Pi_{\gamma,n}\bar{P}_\gamma f - \Pi_{\gamma,n} f\right| \leq 2\|f\|_\infty\delta$. As δ is arbitrary, $\Pi_{\gamma,n}\bar{P}_\gamma f - \Pi_{\gamma,n} f \to 0$, which shows that $\pi_\gamma \bar{P}_\gamma f - \pi_\gamma f = 0$. We have shown that every cluster point of $(\Pi_{\gamma,n} : n \in \mathbb{N})$ is an invariant measure of $\bar{P}_\gamma$.

Consider an arbitrary sequence $\gamma_j \downarrow 0$ as $j \to \infty$, and let π_j be an invariant measure of $\bar{P}_{\gamma_j}$ for every j. It is not difficult to show that the sequence (π_j) is also tight, hence converging to some π^* as $j \to \infty$, along some subsequence. We now prove that such a cluster point π^* is an invariant measure for the semiflow Ψ *i.e.*, $\pi^*\Psi_t^{-1} = \pi^*$ for every $t > 0$. Such a proof can be found for instance in [20], we reproduce it here for completeness. Denote by $\bar{\mathbb{E}}^{\gamma,\nu}$ the expectation associated with $\bar{\mathbb{P}}^{\gamma,\nu}$ and by L the Lipschitz constant of f. For an arbitrary $\delta > 0$, consider a compact set K s.t. $\sup_j \pi_j(K^c) < \delta$. For every j and every $t > 0$, using that $\pi_j = \pi_j \bar{P}_{\gamma_j}^{\lfloor \frac{t}{\gamma_j} \rfloor}$, we obtain, by following the same approach as [20],

$$\left|\int f\circ\Psi_t d\pi_j - \int f d\pi_j\right| = \left|\mathbb{E}^{\gamma_j,\pi_j}(f(\Psi_t(X_0)) - f(X_{\lfloor \frac{t}{\gamma_j} \rfloor}))\right|$$

$$\begin{aligned}
&\leq \mathbb{E}^{\gamma_j,\pi_j}\left(|f(\Psi_t(X_0)) - f(X_{\lfloor \frac{t}{\gamma_j} \rfloor})|\mathbb{1}_K(X_0)\right) + 2\|f\|_\infty\delta \\
&\leq \mathbb{E}^{\gamma_j,\pi_j}\left(\left(2\|f\|_\infty \wedge L\|\Psi_t(X_0) - X_{\lfloor \frac{t}{\gamma_j} \rfloor}\|\right)\mathbb{1}_K(X_0)\right) + 2\|f\|_\infty\delta \\
&\leq \mathbb{E}^{\gamma_j,\pi_j}\left(2\|f\|_\infty \mathbb{1}_K(X_0)\mathbb{1}_{\|\Psi_t(X_0) - X_{\lfloor \frac{t}{\gamma_j} \rfloor}\|>\delta}\right) + L\delta + 2\|f\|_\infty\delta \\
&\leq 2\|f\|_\infty \sup_{x\in K} \mathbb{P}^{\gamma_j,x}\left(\|\Psi_t(x) - X_{\lfloor \frac{t}{\gamma_j} \rfloor}\| > \delta\right) + L\delta + 2\|f\|_\infty\delta\,.
\end{aligned}$$

Thus, $\limsup_j \left|\int f\circ\Psi_t d\pi_j - \int f d\pi_j\right| \leq (L + 2\|f\|_\infty)\delta$, and since δ is arbitrary, the lim sup is equal to zero. Considering the limit along the converging subsequence, it follows that $\int f\circ\Psi_t d\pi^* - \int f d\pi^* = 0$. Hence, π^* is invariant for Ψ. By Poincaré's recurrence theorem, $\pi^*(BC_\Psi) = 1$.

We now conclude the proof of Theorem 9.1. For every $\delta > 0$, set $A_\delta := \{x : d(x, BC_\Psi) \geq \delta\}$ By contradiction, assume that there exists $\delta > 0$, a sequence $\gamma_j \downarrow 0$, and, for every j, a sequence $(\varphi_n^j : n \in \mathbb{N})$ s.t. for every n, $\Pi_{\gamma_j,\varphi_n^j}(A_\delta) > \delta$. For every j, as $(\Pi_{\gamma_j,\varphi_n^j} : n \in \mathbb{N})$ is tight, one can extract a subsequence $(\Pi_{\gamma_j,\bar{\varphi}_n^j} : n \in \mathbb{N})$ converging weakly to some measure π_j which is invariant for $\bar{P}_{\gamma_j}$. By Portmanteau's theorem, $\pi_j(A_\delta) > \delta$. As (π_j) is tight, it converges weakly along some subsequence to some π^* satisfying $\pi^*(BC_\Psi) = 1$. As $\pi^*(A_\delta) > \delta$, this leads to a contradiction. □

End of the Proof of Theorem 5.7. We apply Theorem 5.7 in the case where $P_{\gamma,n}$ is the kernel of the non-homogeneous Markov chain (z_n^γ) defined by (3.6) 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 $\bar{z}_0 \in \mathcal{Z}_+$. The task is merely to verify Assumption 9.1*iii)*, the other assumptions being easily verifiable using Theorem 5.5, Lemma 8.2, Lemma 8.3 and [13, Lemma 6.2]. Consider $\gamma \in (0, \gamma_0)$. Let $f : \mathcal{Z} \to \mathbb{R}$ be a bounded L-Lipschitz-continuous function and K a compact. For all $z = (x, m, v) \in K$:

$$|P_{\gamma,n}(f)(z) - \bar{P}_\gamma(f)(z)| \leq L\gamma\mathbb{E}\left\| \frac{(1-\alpha^n)^{-1}\tilde{m}_\xi}{\varepsilon + (1-\beta^n)^{-\frac{1}{2}}\tilde{v}_\xi^{1/2}} - \frac{\tilde{m}_\xi}{\varepsilon + \tilde{v}_\xi^{1/2}} \right\|$$

$$\leq \frac{L\gamma\alpha^n}{\varepsilon(1-\alpha^n)} \sup_{x,m} (\alpha||m|| + (1-\alpha)\mathbb{E}||\nabla f(x,\xi)||) + \frac{L\gamma\mathbb{E}||\tilde{m}_\xi \tilde{v}_\xi^{1/2}||}{\varepsilon^2} \left(1 - \frac{1}{(1-\beta^n)^{1/2}}\right)$$

where we write $\alpha = \bar{\alpha}(\gamma)$, $\beta = \bar{\beta}(\gamma)$, $\tilde{m}_\xi := \alpha m + (1-\alpha)\nabla f(x,\xi)$ and $\tilde{v}_\xi := \beta v + (1-\beta)\nabla f(x,\xi)^2$. Thus, condition 9.1*iii)* follows. Finally, the fact that $BC_\Phi = \mathcal{E}$ follows from Corollary 7.3.

10. Numerical Examples. In this section, we illustrate our results on two different synthetic problems.

Convergence toward the ODE solution. In the following, we consider a synthetic 2D linear regression problem. Let X be a Bernoulli random variable with parameter $p \in (0,1)$ (i.e $X \in \{0,1\}$ and $\mathbb{P}(X = 1) = p$). Consider a real valued gaussian noise ϵ of zero mean and variance $\sigma^2 > 0$ (ie. $\epsilon \sim \mathcal{N}(0, \sigma^2)$) independent from X. Define $Y = Xx_1^* + (1-X)x_2^* + \epsilon$ where $(x_1^*, x_2^*) = (3, 1)$. Define $\xi = (X, Y)$. Consider now the problem of finding a local minimizer of the expectation $F(x) := \mathbb{E}(f(x,\xi))$ w.r.t. $x \in \mathbb{R}^2$, where $f(\,.\,,\xi) := \frac{1}{2}\left(\left\langle \begin{pmatrix} X \\ 1-X \end{pmatrix}, \cdot \right\rangle - Y\right)^2$. We determine the (ODE) solution using an explicit Euler discretization method. We compute the interpolated process which consists of a linear interpolation of the Adam iterates. Then we plot the solution and the interpolated process on a contour plot of the objective function F, we obtain Figure 1. SGD iterates are also represented for comparison. Figure 1 illustrates the convergence of the (ODE) solution toward the set of critical points of F (see Th. 5.2). We also observe that the interpolated process derived from Adam shadows the (ODE) solution (see Th. 5.5).

In Figure 2, we plot both coordinates of the Adam interpolated process and the (ODE) solution. As expected by Th. 5.5, Figure 2 shows that the interpolated process from the Adam iterates shadows the solution to the non-autonomous differential equation (ODE) in the asymptotic regime where the step size parameter γ of Adam is small. The gradient flow curve represents the continuous-time version of gradient descent which is the solution to the ODE $\dot{x}(t) = -\nabla F(x(t))$.

Biased vs Unbiased Adam. We consider the following Stochastic Quadratic Problem. Define $f(x,\xi) = \frac{1}{2}(x-\xi)^T Q(x-\xi)$ where $Q \in \mathbb{R}^{d\times d}$ is a symmetric positive definite matrix and $\xi \sim \mathcal{N}(\xi^*, \sigma^2 I)$ with $\sigma \in \mathbb{R}_+$ (see [6, section 2.] where the same problem is considered). Notice that : $F(x) = \mathbb{E}(f(x,\xi)) = \frac{1}{2}(x-\xi^*)^T Q(x-\xi^*) + \frac{1}{2}\sigma^2 \operatorname{tr}(Q)$ with $\nabla F(x) = Q(x-\xi^*)$ and $S(x) = \mathbb{E}(\nabla f(x,\xi)^2) = [Q(x-\xi^*)]^2 + \sigma^2 \operatorname{diag}(Q^2)$ where $[Q(x-\xi^*)]^2$ is computed coordinate-wise and $\operatorname{diag}(Q^2)$ is the diagonal of the matrix Q^2. We consider two versions of Adam : the seminal algorithm Adam introduced by [25] and a biased version of Adam corresponding to the same algorithm without the bias correction steps (see Algorithm 3.1). The continuous-time version of Adam is the solution to the non-autonomous (ODE). For the modified Adam algorithm (without the bias correction steps), the continuous-time version is the solution to an autonomous ODE which writes $\dot{z}(t) = h_\infty(z(t))$ where

for $(x, m, v) \in \mathcal{Z}_+$, $h_\infty(x, m, v) = (-m/(\varepsilon + \sqrt{v}), a(\nabla F(x) - m), b(S(x) - v))$ (see subsection 6.1 for more details). For each one of the two ODEs, we compute the solution $x(t)$ using an explicit Euler discretization scheme with a fixed discretization step size $\eta = 10^{-4}$. In Figure 3, we plot the values of the function $t \mapsto F(x(t))$ in both cases. Figure 3 shows that $F(x(t))$ can increase for the biased Adam, deteriorating the initial estimate x_0. We also observe that the solution to the Adam (ODE) improves the initial guess x_0 as expected (see Ineq. (5.3)).

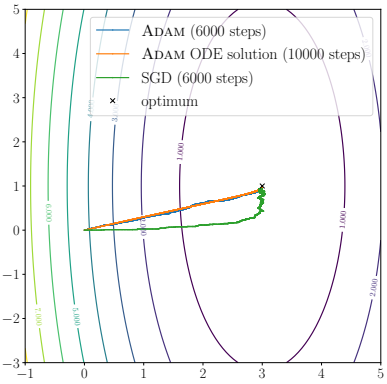

Fig. 1. *Convergence of* Adam *and the corresponding ODE solution to the optimum for a 2D linear regression.*

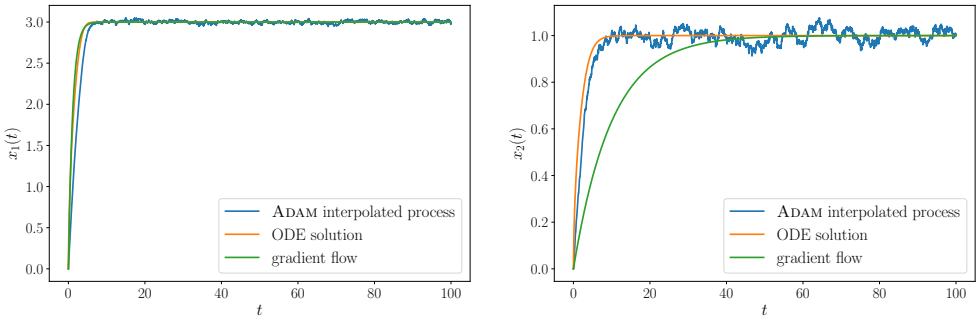

Fig. 2. Adam *: interpolated process and solution to the ODE for a 2D linear regression.*

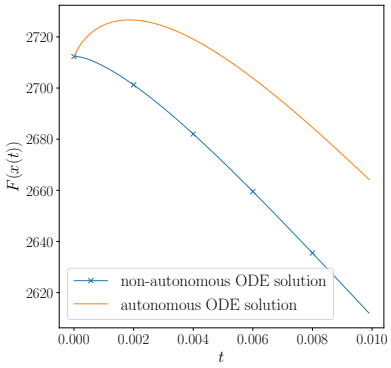

FIG. 3. *Comparison between* ADAM *ODE solution and autonomous* ADAM *ODE solution on a 100-dimensional Stochastic Quadratic Problem.*

11. Conclusion. We introduced a continuous-time version of ADAM relying on the ODE method. This version consists in a non-autonomous ODE. Due to the irregularity of the mean field of the ODE, both the existence and the uniqueness of the global solution turn out to be non-trivial problems. These results are established assuming that the objective function is differentiable but possibly non convex. The convergence of the solution to the set of stationary points of the objective function is obtained. We proved that the linearly interpolated process associated to the discrete-time version of ADAM converges weakly to the solution to the ODE as $\gamma \to 0$. This result is used to establish the convergence in the long-run of the discrete-time ADAM iterates to a stationary point of the objective function.

In future works, it is important to address the question of stability of the Markov chain generated by the ADAM iterations. The case of non-differentiable functions F is worth being studied in order to encompass the case of deep neural networks. Finally, the problem of convergence rates of ADAM is an open question which will be addressed in future works.

REFERENCES

[1] F. ALVAREZ, *On the minimizing property of a second order dissipative system in hilbert spaces*, SIAM Journal on Control and Optimization, 38 (2000), pp. 1102–1119.
[2] V. APIDOPOULOS, J.-F. AUJOL, C. DOSSAL, AND A. RONDEPIERRE, *Convergence rates of an inertial gradient descent algorithm under growth and flatness conditions*, (2018).
[3] H. ATTOUCH, Z. CHBANI, J. PEYPOUQUET, AND P. REDONT, *Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity*, Mathematical Programming, 168 (2018), pp. 123–175.
[4] 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.
[5] J.-F. AUJOL, C. DOSSAL, AND A. RONDEPIERRE, *Optimal convergence rates for nesterov acceleration*, arXiv preprint arXiv:1805.05719, (2018).
[6] 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 (ICML), 2018.
[7] A. BARAKAT AND P. BIANCHI, *Convergence of the adam algorithm from a dynamical system viewpoint*, arXiv preprint https://arxiv.org/abs/1810.02263v1, (4 Oct 2018).
[8] 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).

[9] 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.

[10] M. Benaïm and M. W. Hirsch, *Asymptotic pseudotrajectories and chain recurrent flows, with applications*, J. Dynam. Differential Equations, 8 (1996), pp. 141–176.

[11] M. Benaïm and S. J. Schreiber, *Ergodic properties of weak asymptotic pseudotrajectories for semiflows*, J. Dynam. Differential Equations, 12 (2000), pp. 579–598.

[12] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, *signSGD: Compressed optimisation for non-convex problems*, in Proceedings of the 35th International Conference on Machine Learning, vol. 80 of Proceedings of Machine Learning Research, PMLR, 2018, pp. 560–569.

[13] 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.

[14] 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.

[15] A. Cabot, H. Engler, and S. Gadat, *Second-order differential equations with asymptotically small dissipation and piecewise flat potentials*, Electronic Journal of Differential Equation, 17 (2009), pp. 33–38.

[16] 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.

[17] 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).

[18] 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.

[19] R. Fletcher, *A new approach to variable metric algorithms*, The computer journal, 13 (1970), pp. 317–322.

[20] 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).

[21] S. Gadat and F. Panloup, *Long time behaviour and stationary regime of memory gradient diffusions*, in Annales de l'IHP Probabilités et statistiques, vol. 50, 2014, pp. 564–601.

[22] S. Gadat, F. Panloup, and S. Saadane, *Stochastic heavy ball*, Electronic Journal of Statistics, 12 (2018), pp. 461–529.

[23] A. Haraux, *Systemes dynamiques dissipatifs et applications*, vol. 17, Masson, 1991.

[24] P. Hartman, *Ordinary Differential Equations: Second Edition*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 1982.

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

[26] H. J. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*, vol. 35 of Applications of Mathematics (New York), Springer-Verlag, New York, second ed., 2003. Stochastic Modelling and Applied Probability.

[27] L. Ljung, *Analysis of recursive stochastic algorithms*, IEEE transactions on automatic control, 22 (1977), pp. 551–575.

[28] M. C. Mukkamala and M. Hein, *Variants of RMSProp and Adagrad with logarithmic regret bounds*, in Proceedings of the 34th International Conference on Machine Learning, vol. 70 of Proceedings of Machine Learning Research, PMLR, 2017, pp. 2545–2553.

[29] Y. E. Nesterov, *A method for solving the convex programming problem with convergence rate* $\mathcal{O}(1/k^2)$, in Dokl. Akad. Nauk SSSR, vol. 269, 1983, pp. 543–547.

[30] B. Polyak and P. Shcherbakov, *Lyapunov functions: An optimization theory perspective*, IFAC-PapersOnLine, 50 (2017), pp. 7456–7461.

[31] B. T. Polyak, *Some methods of speeding up the convergence of iteration methods*, USSR Computational Mathematics and Mathematical Physics, 4 (1964), pp. 1–17.

[32] S. J. Reddi, S. Kale, and S. Kumar, *On the convergence of adam and beyond*, in International Conference on Learning Representations, 2018.

[33] H. Robbins and S. Monro, *A stochastic approximation method*, in Herbert Robbins Selected Papers, Springer, 1985, pp. 102–109.

[34] G. Roth and W. H. Sandholm, *Stochastic approximations with constant step size and differential inclusions*, SIAM J. Control Optim., 51 (2013), pp. 525–555.

[35] T. Schaul, S. Zhang, and Y. LeCun, *No more pesky learning rates*, in International Con-

ference on Machine Learning, 2013, pp. 343–351.
[36] B. Shi, S. Du, M. I. Jordan, and W. J. Su, *Understanding the acceleration phenomenon via high-resolution differential equations*, arXiv preprint arXiv:1810.08907, (2018).
[37] W. Su, S. Boyd, and E. J. Candès, *A differential equation for modeling nesterov's accelerated gradient method: Theory and insights*, Journal of Machine Learning Research, 17 (2016), pp. 1–43.
[38] T. Tieleman and G. Hinton, *Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude*, Coursera: Neural networks for machine learning, 4 (2012), pp. 26–31.
[39] R. Ward, X. Wu, and L. Bottou, *Adagrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization*, arXiv preprint arXiv:1806.01811, (2018).
[40] A. Wibisono, A. C. Wilson, and M. I. Jordan, *A variational perspective on accelerated methods in optimization*, proceedings of the National Academy of Sciences, 113 (2016), pp. E7351–E7358.
[41] A. C. Wilson, B. Recht, and M. I. Jordan, *A lyapunov analysis of momentum methods in optimization*, arXiv preprint arXiv:1611.02635, (2016).
[42] 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.
[43] 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).