Journal of Machine Learning Research 21 (2020) 1-42 Submitted 11/18; Revised 6/20; Published 6/20

A General System of Differential Equations to Model First-Order Adaptive Algorithms

André Belotto da Silva[*] ANDRE-RICARDO.BELOTTO-DA-SILVA@UNIV-AMU.FR
Université Aix-Marseille
Centre de Mathématiques et Informatique,
39, rue F. Joliot Curie,
13013 Marseille, France

Maxime Gazeau* MAXIME.GAZEAU@BOREALISAI.COM
Borealis AI,
MaRS Heritage Building,
101 College St, Suite 350,
Toronto, ON M5G 1L7

Editor: Leon Bottou

Abstract

First-order optimization algorithms play a major role in large scale machine learning. A new class of methods, called *adaptive algorithms*, were recently introduced to adjust iteratively the learning rate for each coordinate. Despite great practical success in deep learning, their behavior and performance on more general loss functions are not well understood. In this paper, we derive a non-autonomous system of differential equations, which is the continuous time limit of adaptive optimization methods. We study the convergence of its trajectories and give conditions under which the differential system, underlying all adaptive algorithms, is suitable for optimization. We discuss convergence to a critical point in the non-convex case and give conditions for the dynamics to avoid saddle points and local maxima. For convex loss function, we introduce a suitable Lyapunov functional which allows us to study its rate of convergence. Several other properties of both the continuous and discrete systems are briefly discussed. The differential system studied in the paper is general enough to encompass many other classical algorithms (such as Heavy Ball and Nesterov's accelerated method) and allow us to recover several known results for these algorithms.

Keywords: Adaptive algorithms, convex and non-convex optimization, first-order methods, differential equation, forward Euler discretization

1. Introduction

Optimization is at the core of many machine learning problems. Estimating the model parameters can often be formulated in terms of an unconstrained optimization problem of the form

$$\min_{\theta \in \mathbb{R}^d} f(\theta) \qquad \text{where } f : \mathbb{R}^d \to \mathbb{R} \text{ is differentiable.} \tag{1.1}$$

[*] Alphabetical order and equal contribution

The emergence of deep learning has spawned the recent popularity of a special class of optimizers to solve Equation (1.1): first order *adaptive* optimization algorithms such as RMSProp (Tieleman and Hinton, 2012), Adagrad (Duchi et al., 2011; Duchi and Singer, 2013), Adadelta (Zeiler, 2013), Adam (Kingma and Ba, 2014) were originally designed to solve unconstrained optimization problem.

Despite its obvious efficiency in deep learning (Gregor et al., 2015; Radford et al., 2015), the reasons of their success are unclear and a large number of fundamental questions are still unanswered. In particular, recent research paper (S. Reddi and Kumar, 2018) shows that Adam may diverge.

Our work started from the intuition that these algorithms are not intrinsically better than gradient descent but rather well suited to the subclass of non-convex functions given by standard deep learning architectures. Studying the convergence of the discrete and stochastic adaptive algorithms for non-convex functional is far too complex and general to get insightful explanation about their efficiency in deep learning. We, therefore, start by studying a deterministic and continuous equation, and we prove that in simple cases (such as convex functional), adaptive algorithm are not converging faster than gradient descent. In particular, the key insights of our analysis are:

1. The convergence rate is nonlinear –in the sense that it depends on the variables– and depends on the history of the dynamics. Initialization is therefore of crucial importance.

2. With the standard choices of hyperparameters, adaptivity degrades the rate of convergence to the global minimum of a convex function compared to gradient descent.

These observations are crucial to unwind the mystery of adaptive algorithms and the next questions to ask are now obvious:

1. Does adaptivity reduces the variance (compared to Stochastic Gradient Descent) and speed up the training for convex functional?

2. Is the fast training observed in deep learning induced by the specificity of the loss surface and common initialization scheme for the weights?

The main contribution of this paper is to provide a theoretical framework to study deterministic adaptive algorithms. Inspired by the history of gradient descent and stochastic gradient descent, we analyse discrete *adaptive* optimization algorithms by introducing their continuous-time counterparts (Equation 2.1), which correspond to the limit of large batch sizes and small learning rates. We focus on Adam given by Equation (4.3). The techniques and analysis are similar for other algorithms and include classical accelerated methods.

This work is intended to serve as a solid foundation for the posterior study in the discrete and stochastic settings, and in this paper, we put an emphasis on the deterministic equation to understand the fundamental properties of adaptive algorithms.

In Section 2 we introduce two general continuous dynamical system (2.1) and its *forward* Euler approximation (2.2). The connection between these equations and optimization algorithms is summarized in table 1 and made precise in Section 4. Section 2.2 contains important properties of the Ordinary Differential Equation (2.1). Section 3 contains the statement of our main results, on the asymptotic behavior of the continuous deterministic trajectories of the ODE (2.1). In the non-convex setting, we prove in Theorem 4 that the gradient converges to zero and in Theorem 5 that the trajectories converge to the critical locus of f. This result is supplemented with the analysis of sufficient conditions in order to avoid convergence to saddle or local maximum points of the loss function f (see Theorem 6). For convex functions, we design a Lyapunov functional (3.2) and obtain, in Theorem 7, a rate of convergence to at least a neighborhood of the critical locus. The rate of convergence crucially depends on the behavior over time of ∇f and on the term v (see Equation 3.3 and the subsequent discussion for more details). In particular, this indicates that the efficiency of adaptive algorithms is highly dependant on the loss function. In Sections 4, we specialize the convergence results to Adam, AdaFom, Heavy Ball and Nesterov. In particular, Corollary 8 provides new results on the convergence of the dynamics of Adam, while Corollary 11 recovers previously known convergence rates of Nesterov's accelerated method. We stress that *Sections 3 and 4 can be read independently.* In Section 5 we provide some empirical observations on adaptive algorithms that are inspired by the continuous analysis. Finally, we collect guidelines for designing new adaptive algorithms in Section 6. Most proofs supporting the paper are postponed to the Appendix.

1.1. Related Work

The study of a continuous dynamical system is often very useful to understand discrete optimization algorithms. For smooth convex or strongly convex functions, Nesterov (2004) introduced an accelerated gradient algorithm which was proven to be optimal (a lower bound matching an upper bound is provided).

However, the key mechanism for acceleration is not well understood and have many interpretations (Bubeck et al., 2015; Hu and Lessard; Lessard et al., 2016). A particular interesting interpretation of acceleration is through the lens of a second order differential equation of the form

$$\ddot{\theta} + a(t)\dot{\theta} + \nabla f(\theta) = 0, \qquad \theta(0) = \theta_0, \quad \dot{\theta}(0) = \psi_0, \tag{1.2}$$

where $t \mapsto a(t)$ is a smooth, positive and decreasing function of time, having possibly a pole at zero. Even if this singularity has important implications for the choice of the initial velocity ψ_0, we are more interested by the long term behavior of the solution to (1.2) and hence at $\lim_{t\to\infty} a(t)$. This system is called dissipative because its energy $E(t) = \frac{1}{2}||\dot{\theta}||^2 + f(\theta)$ decreases over time.

Most accelerated optimization algorithms can be seen as the numerical integration of Equation (1.2). Alvarez et al. (2002); Alvarez (2000) studied the *Heavy Ball method* in which the function a is constant and is called the damping parameter. Gadat and Panloup (2014); Cabot (2009); Cabot et al. (2009) gave conditions on the rate of decay of a and its limit in order for the trajectories of Equation (1.2) to converge to a critical point of f. This analysis highlights situations where solutions to Equation (1.2) are fit (or not) for optimization. Intuitively, if a decays too fast to zero (like $1/t^2$) the system will oscillate and won't converge to a critical point. The case $a(t) = 3/t$ was studied more specifically in

Su et al. (2016) and the authors draw interesting connections between (1.2) and *Nesterov's algorithm.* The convergence rates obtained are $\mathcal{O}(1/(sk^2))$ and $\mathcal{O}(1/t^2)$ respectively, which match with the discrete algorithms by using the time identification $t = \sqrt{s}k$ (Su et al., 2016). Wibisono and Wilson (2015); Wibisono et al. (2016) extended this work and studied acceleration from a different continuous equation having a theoretically exponential rate of convergence. However, a naïve discretization loses the nice properties of this continuous system and current work consists of finding a better one preserving the symplectic structure of the continuous flow (Betancourt et al., 2018).

By nature, first-order adaptive algorithms have iterates that are non-linear functions of the gradient of the objective function. The analysis of convergence is, therefore, more complex, potentially because the rate of convergence might depend on the function itself. The first known algorithm Adagrad (Duchi and Singer, 2013) consists of multiplying the gradient by a diagonal preconditioning matrix, depending on previously squared gradients. The key property to prove the convergence of this algorithm is that the elements of the preconditioning matrix are positive and non-decreasing (Ward et al., 2019; Duchi and Singer, 2013; Chen et al., 2019). Later on, two new adaptive algorithms RMSprop (Tieleman and Hinton, 2012) and Adam (Kingma and Ba, 2014) were proposed. The preconditioning matrix is an exponential moving average of the previously squared gradients. As a consequence, it is no longer non-decreasing. The proof of convergence, relying on this assumption and given in the form of a regret bound in Kingma and Ba (2014), is therefore not correct (S. Reddi and Kumar, 2018). A new algorithm Amsgrad proposed by S. Reddi and Kumar (2018) consists of modifying the preconditioning updates to recover this property. While converging, this algorithm loses the essence of the Adam's algorithm.Adam is such a mysterious algorithm that many works have been devoted to understanding its behavior. Variants of Adam have been proposed by Zhang et al. (2017) as well as convergence analysis towards a critical point (De et al., 2018; Chen et al., 2019). However, conditions for convergence seem very restrictive and not easy to verify in practice.

In what follows, we use several times the same non standard operations on vectors. It is convenient to fix the notation of these operations. Given two vectors $u = (u_1, \dots, u_d)$ and $v = (v_1, \dots, v_d)$ of $\mathbb{R}^d$ and constants a, $\varepsilon \in \mathbb{R}$, we use the following notation:

$$\begin{aligned} u + \varepsilon &:= (u_1 + \varepsilon, \dots, u_d + \varepsilon) \\ u/v &:= (u_1/v_1, \dots, u_d/v_d) \\ u^a &:= (u_1^a, \dots, u_d^a) \end{aligned}$$

2. Presentation of the Model

Throughout this paper we study the following dynamical system

$$\begin{cases} \dot{\theta}(t) = -m(t)/\sqrt{v(t) + \varepsilon} \\ \dot{m}(t) = h(t)\nabla f(\theta(t)) - r(t)m(t) \\ \dot{v}(t) = p(t)\left[\nabla f(\theta(t))\right]^2 - q(t)v(t), \end{cases} \tag{2.1}$$

where $\varepsilon \geq 0$, the functions $h(t)$, $r(t)$, $p(t)$ and $q(t)$ are C^1-functions defined over $\mathbb{R}_{>0}$ and $(\theta, m, v, t) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d_{\geq 0} \times \mathbb{R}_{>0}$; if $\varepsilon = 0$, then $v \in \mathbb{R}^d_{>0}$. The above system has a momentum term m and a memory term v. The system (2.1) is supplemented with initial conditions $\mathbf{x}_0 = (\theta_0, m_0, v_0)$ at time $t = t_0 \geq 0$. We denote by $\mathbf{x}(t) = \mathbf{x}(t, t_0, \mathbf{x}_0) = (\theta(t), m(t), v(t))$ a solution of (2.1) with initial condition $\mathbf{x}(t_0, t_0, \mathbf{x}_0) = \mathbf{x}_0$, and interval of definition $t \in [t_0, t_\infty[$.

We always make the following hypotheses.

Assumption 1 *The objective function f is assumed to be a C^2 function defined in $\mathbb{R}^d$. The functions h, r, p and q are non-negative and non-increasing C^1-functions defined over $\mathbb{R}_{>0}$, and $h(t) \not\equiv 0$, $r(t) \not\equiv 0$. We also require that one of the following is satisfied:*

Either $p(t) \not\equiv 0$, in which case we say that the system is adaptive*;*

Or $p(t) \equiv q(t) \equiv 0$, in which case we say that the system is non-adaptive.

The choices of functions h, r, p and q, which yield a good optimization algorithm, are not unique and should adapt to the local properties of the loss function. We provide a list of known choices in table 1 and give some guidelines on how to choose them in §§ 6. The momentum term m accelerates the convergence of the algorithm depending on the choice of the coefficients h and r. Intuitively, the addition of the momentum m implies the existence of a special energy functional for ODE (2.1) (see Equation 2.3 below) which works as a "funnel" around minimum points of f. The trajectories of the ODE can, therefore, be "accelerated" without the risk of diverging to ∞, at the cost of an oscillatory behavior (just as water in a funnel). An adaptive system has a non-trivial dynamic induced by the addition of the memory term v used to rescale the learning rate component by component. The variable v is controlled by the history of the trajectory $\mathbf{x}(t)$ and the loss function f. When the dynamics traverses a region of high gradient, their values will accumulate in v and the dynamics will slow down. The functions p and q determine how much memory of past gradients to keep. Based on these two intuitions, we can expect that the choice of h and r controls how fast these algorithms converge, while p and q may only slow down the algorithm in general, but accelerate it for certain "classes" of loss functions. This intuition turns out to be precise when dealing with convex functions, as we discuss in §§ 6.

Remark 1 *When the system does not contain a momentum term m, as it is the case for* Adagrad *or* RMSprop*, one must consider the alternative simpler system*

$$\begin{cases} \dot\theta(t) = -\nabla f(\theta)/\sqrt{\omega(t)+\varepsilon} \\ \dot\omega(t) = p(t)\left[\nabla f(\theta(t))\right]^2 - q(t)\omega(t). \end{cases}$$

The convergence analysis is similar (and simpler) to Equation (2.1)*, but can not be derived directly from it and can be found in Belotto da Silva and Gazeau (2018).*

2.1. The Discrete-Time Model

In order to establish a relation between the continuous and the optimization algorithms, we study the finite difference approximation of Equation (2.1) by the forward Euler method

$$\begin{cases} \theta_{k+1} = \theta_k - s m_k/\sqrt{v_k+\varepsilon} \\ m_{k+1} = (1 - s r(t_{k+1}))m_k + s h(t_{k+1})\nabla f(\theta_{k+1}) \\ v_{k+1} = (1 - s q(t_{k+1}))v_k + s p(t_{k+1})\left[\nabla f(\theta_{k+1})\right]^2 \end{cases} \tag{2.2}$$

Algorithm	**Equation** (2.1)	**Equation**(2.2)
Adam (4.3)	$g(t,\lambda,\alpha)=\dfrac{1-e^{-\lambda\alpha}}{\lambda(1-e^{-t\alpha})}$ $h\equiv r\equiv g(t,\lambda,\alpha_1)$ $p\equiv q\equiv g(t,\lambda,\alpha_2)$	$\lambda=s,\ \beta_i=e^{-\lambda/\alpha_i},\ i=1,2$ $\beta_1\frac{1-\beta_1^k}{1-\beta_1^{k+1}}=\mu_{k+1}=1-sg(t_{k+1},\lambda,\alpha_1)$ $\beta_2\frac{1-\beta_2^k}{1-\beta_2^{k+1}}=\nu_{k+1}=1-sg(t_{k+1},\lambda,\alpha_2)$
Adam without rescaling	$h\equiv 1, r\equiv 1/\alpha_1$ $p\equiv q\equiv 1/\alpha_2$	$\beta_1=1-s/\alpha_1$ $\beta_2=1-s/\alpha_2$
AdaFom (4.6)	$h\equiv r\equiv g(t,\lambda,\alpha_1)$ $p\equiv q\equiv 1/t$	$\beta_1\frac{1-\beta_1^k}{1-\beta_1^{k+1}}=\mu_{k+1}=1-sg(t_{k+1},\lambda,\alpha_1)$ $sp(t_{k+1})=1/(k+1)$
Heavy Ball (4.7)	$h\equiv 1, r\equiv\gamma$ $p\equiv q\equiv 0$	$\beta=1-s\gamma,\ n_k=sm_k,\ \alpha=s^2$ 0
Nesterov (4.10)	$h\equiv 1, r\equiv 3/t$ $p\equiv q\equiv 0$	$h=1,\quad r=3/(k+1)$ 0

Table 1: Connections between Equations (2.1) and (2.2).

where $t_k = ks$. We chose this method because it fits well with Adam discrete system. It follows from the theory of Euler discretization that the approximation error between the discrete system (2.2) and continuous system (2.1) tends to zero (with order one) when the learning rate goes to zero. However this choice of discretization is of course non-unique, and more efficient quadrature rules could lead to more accurate numerical integration (Duris and Lyness, 1975; Kythe and Puri, 2002). The connections between our model and the discrete optimization algorithms is summarized by table 1, and the proof of these relations is postponed to Section 4.

2.2. An Energy Functional of ODE (2.1) and a Natural Assumption

A crucial property in the study of ODE (2.1) is the existence of an energy functional, which is inspired from Alvarez (2000):

$$E(t,\theta,m,v)=f(\theta)+\frac{1}{2h(t)}\left\|\frac{m}{[v+\varepsilon]^{1/4}}\right\|^2. \tag{2.3}$$

This functional plays a crucial role in the study of the convergence of ODE (2.1) in §3. By direct computation of its derivative (see Equation B.1 in the Appendix), we obtain that:

$$\frac{d}{dt}E(t,\theta,m,v)\leq-\frac{1}{2h(t)}\left[2r(t)-\frac{q(t)}{2}+\frac{h'(t)}{h(t)}\right]\left\|\frac{m}{[v+\varepsilon]^{1/4}}\right\|^2. \tag{2.4}$$

This leads us to the following natural hypothesis, assumed almost everywhere in the paper:

Assumption 2 *There exists $\tilde{t}>0$ such that for every $t>\tilde{t}$ we have that:*

$$2r(t)-\frac{q(t)}{2}+\frac{h'(t)}{h(t)}\geq 0.$$

In practice, this is a mild assumption in the hyper-parameters of the model. In terms of the algorithms in table 1, it is always verified by AdaFom, Heavy Ball and Nesterov (with $\tilde{t}=0$); for Adam (and Adam with rescaling) it leads to the following condition on the hyper-parameters (which is usually respected by practitioners): $3+\beta_2 > 4\beta_1$. Now, under assumption 2 the derivative of $E(t,\theta,m,v)$ is non-positive, which immediately yields the following result.

Lemma 2 *Suppose that assumptions 1 and 2 are verified. Given a solution $\mathbf{x}(t)$ of the ODE* (2.1) *such that $t_0 \geq \tilde{t}$, we have that:*

$$E(t,\mathbf{x}(t)) \leq E(t_0,\mathbf{x}_0), \quad \forall\, t \in [t_0, t_\infty[.$$

In particular, if f is coercive, then the curve $\theta(t)$ is bounded. Furthermore, if f is a function bounded from below, say by f_*, then:

$$\frac{1}{2h(t)}\left\|\frac{m}{[v+\varepsilon]^{1/4}}\right\|^2 \leq E(t_0,\mathbf{x}_0) - f_*, \quad \forall\, t \in [t_0, t_\infty[$$

The above result shows the importance of the energy functional and has important implications. For example, if f is coercive we guarantee that $\theta(t)$ is bounded (c.f. Assumption 4 below).

2.3. Existence and Uniqueness of a Solution to ODE (2.1)

We continue our analysis by showing that all solutions of ODE (2.1) are well-defined on the interval $[t_0,\infty[$. Recall that our functions $h(t)$, $r(t)$, $p(t)$ and $q(t)$ are allowed to have poles at the origin (c.f. table 1) in order to capture the phenomena present in both accelerated and adaptive methods (c.f. Nesterov and Adam). This imposes some technical difficulties similar to the analysis of the Nesterov's differential equation (Su et al., 2016).

We, therefore, need to demand extra assumptions on the coefficients of our model and the initial conditions in order to guarantee the existence and uniqueness of the solution at time $t_0 = 0$:

Assumption 3 *We assume one of the following conditions:* (1) *the functions h, r, p, q have a simple pole at $t=0$, or* (2) *$h \in C^1([0,+\infty))$ (resp. $p \in C^1([0,+\infty))$), then r (resp. q) can have a simple pole at zero; or* (3) *all functions are assumed to be C^1 on $[0,\infty)$. In cases* (1) *and* (2)*, furthermore, we demand the following two extra-conditions:*

(a) the initial conditions must be taken as:

$$m_0 = \nabla f(\theta_0) \lim_{t\to 0^+} h(t)/r(t), \quad v_0 = [\nabla f(\theta_0)]^2 \lim_{t\to 0^+} p(t)/q(t).$$

(b) We assume that $p(t) \not\equiv 0$ and that there exists a small time $\hat{t}$ such that

$$2r(t) - q(t) \geq 0, \qquad \forall t < \hat{t},$$

In terms of the algorithms in table 1, the assumption is always verified for Heavy Ball, Nesterov, AdaFom and Adam without rescaling. For Adam, nevertheless, it is necessary to add mild assumptions on the hyper-parameters such as $\beta_1 \geq 0.21$. We are now ready to enunciate the existence and uniqueness result. The proof is postponed to appendix A.

Theorem 3 (Existence and uniqueness) *Suppose that the ODE* (2.1) *satisfies assumption 1, and that either f is bounded from below and assumption 2 is satisfied with $\tilde{t} = 0$ or $p(t) \not\equiv 0$. Then for any $t_0 > 0$ and admissible initial condition $\mathbf{x}_0$, there exists a unique global solution to Equation* (2.1) *such that:*

$$\theta \in C^2([t_0, \infty); \mathbb{R}^d) \quad \textit{and} \quad m,\, v \;\in C^1([t_0, \infty); \mathbb{R}^d).$$

Suppose, furthermore, that assumption 3 is also satisfied. Then, there exists a unique global solution to Equation (2.1) *such that:*

$$\theta \in C^2((0, \infty); \mathbb{R}^d) \cap C^1([0, \infty); \mathbb{R}^d) \quad \textit{and}$$
$$m,\, v \in C^1((0, \infty); \mathbb{R}^d) \cap C^0([0, \infty); \mathbb{R}^d).$$

3. Convergence Analysis

In this section, we study the asymptotic behavior of the solutions of (2.1). Our analysis is divided in the following three steps:

(0) *Gradient convergence*: Find sufficient conditions on the functions f and p, q, r, h in order for $\nabla f(\theta(t)) \to 0$ when $t \to \infty$.

(1) *Topological convergence*: Find sufficient conditions on the functions f and p, q, r, h in order for the solutions of Equation (2.1) to converge to a critical value of f. In particular we do not require f to be convex.

(2) *Avoiding local maximum and saddles*: We want to strengthen the result of part (1) and give sufficient conditions so that the dynamics avoid local maximum and saddles and only converge to a local minimum. In other words, fix $t_0 > 0$ and denote by S_{t_0} the set of initial conditions $\mathbf{x}_0 = (\theta_0, m_0, v_0)$ such that the limit set of the associated solution $\theta(t)$ contains a critical point $\theta_\star$ which is *not* a local minimum. We give, in subsection 3.4, sufficient conditions for the set S_{t_0} to have Lebesgue measure zero.

(3) *Rate of convergence*: Under the convexity assumption, find the rate of convergence of f to a local minimum.

In the remainder of this section, we give precise statements for all of the three steps, and we will make appropriate assumptions on the objective function.

3.1. Convergence of the Gradient

In the case of non-convex functional, it is often the case that the trajectory $\theta(t)$ of the system is unbounded, even for the gradient descent. For example, if we consider:

$$f(\theta) = \frac{\theta^2 - 1}{(\theta^2 + 10)^2} \quad \text{and initial condition} \quad |\theta_0| > 5. \tag{3.1}$$

This is unavoidable and can not be solved without appropriately setting the initial conditions or via a well-chosen regularization technique (see the discussion in §§ 3.2). In this general setting, it is possible to obtain results concerning the asymptotic of the gradient ∇f, provided that the gradient is globally Lipschitz and bounded. More precisely, we prove the following result, which can be seen as a continuous counter-part of Theorem 1 in Zaheer et al. (2018) in the analysis of RMSProp.

Theorem 4 (Gradient convergence) *Suppose that assumptions 1 and 2 are satisfied and that $\lim_{t\to\infty} p(t) \neq 0$ and $\lim_{t\to\infty} q(t) \neq 0$. If f is a function bounded from below whose gradient ∇f is globally Lipschitz and bounded, then*

$$\int_{t_0}^{\infty} \|\nabla f(\theta(t))\|^2 dt < \infty$$

and $\nabla f(\theta(t)) \to 0$ when $t \to \infty$.

We postpone the proof of the above Theorem to Appendix B. We note that Adam satisfies all the assumptions of the above theorem, provided that the hyper-parameters satisfy $3 + \beta_2 > 4\beta_1$. Finally, note that example 3.1 belongs to the setting of Theorem 4.

3.2. On a Compactness Assumption

The following assumption over-arches several points of our analysis:

Assumption 4 *The solution $\theta(t)$ of the ODE* (2.1) *is bounded.*

We recall that Assumption 4 is automatically satisfied when the loss function f is coercive. More generally, by Lemma 2, Assumption 4 is automatically satisfied when the initial condition θ_0 is in a connected and bounded component of the *lower-level set* of the loss function f, that is, if the connected components of $\{\theta \in \mathbb{R}^d;\ f(\theta) \leq f(\theta_0)\}$ are bounded. This is the case, for example, when ℓ^2 regularization of the weights is used as a regularization technique. Verifying when Assumption 4 is satisfied under other regularization techniques, or under well-chosen initial conditions, remains an important open problem which should be addressed in future works.

3.3. Topological Convergence

We now search a stronger convergence result under the compactness assumption 4. In order to do so, we make an additional assumption on the asymptotic behaviour of the coefficients, which is designed to simplify the proof while still covering Adam:

Assumption 5 *Suppose that $\varepsilon > 0$. Consider the functions:*

$$H(\tau) = h(1/\tau), \quad R(\tau) = r(1/\tau), \quad P(\tau) = p(1/\tau), \quad Q(\tau) = q(1/\tau),$$

and suppose that these functions are C^1 in $[0, \infty)$, $H(0) > 0$ and $4R(0) > Q(0)$.

Note that Assumption 5 is satisfied, essentially, when the coefficients of $h(t)$ and $r(t)$ do not converge to zero at infinity. Hence, it holds for Adam, AdaFom, and the Heavy Ball differential equations, c.f. table 1. It also has the interesting feature of being almost completely *independent from the functions* $p(t)$ *and* $q(t)$, a flexibility which should be explored when trying to design new algorithms. Under this assumption, we prove the convergence of the dynamics in the following sense:

Theorem 5 (Topological Convergence) *Suppose that assumptions 1, 2, 4 and 5 are verified. Then* $f(\theta(t)) \to f_\star$ *and* $m(t) \to 0$ *when* $t \to \infty$*, where* $f_\star$ *is a critical value of* f*. Furthermore, if either* $Q(0) > 0$ *or* $p(t) \equiv q(t) \equiv 0$ *and* $v_0 = 0$*, then* $v(t) \to 0$*.*

The proof of Theorem 5 is postponed in Appendix C. Our method is inspired by the work of Alvarez (2000), based on the energy functional of the system (2.3). We use elementary topological techniques of qualitative theory of ODE's (à la Poincaré-Bendixson). On the one hand, this approach avoids most estimates and analytical arguments, which are typically necessary for this kind of study, and can be easily reproduced in other systems. As an immediate advantage, we do not need assumptions such as convexity of the loss function or globally Lipschitz properties of the differential equation. On the other hand, the assumption is not optimal. For example, it is not satisfied by Nesterov's acceleration Equation (4.10). We believe that the optimal threshold to guarantee convergence of ODE (2.1) should be given by inequality in terms of poles of order at most one for the functions H and R. This idea is supported by the results in Cabot et al. (2009) which show that the function R can not be a polynomial function of order bigger than 1 in the case of the dissipative system related to accelerated dynamics.

3.4. Avoiding Local Maximum and Saddles

In this section, we make the following extra assumption:

Assumption 6 *A critical point* $\theta_\star$ *of* f *is either a local-minimum or it satisfies the two following properties:*

(a) it is a strict saddle (following Definition 1 in Lee et al. (2019)), that is, there exists a strictly negative eigenvalue of the Hessian $\mathcal{H}_f(\theta_\star)$ *of* f *at* $\theta_\star$*.*

(b) it is an isolated critical point, that is, there is a neighbourhood U *around* $\theta_\star$ *that does not contain any other critical points.*

We provide a discussion about this assumption in Remark 22, in the Appendix.

Now, fix a time $t_0 > 0$ and recall that the topological limit of a curve $\theta(t)$, called ω-limit, is given by:

$$\omega(\theta(t)) = \bigcap_{\tau > t_0} \overline{\theta([\tau, \infty))}.$$

Consider the set of initial conditions such that the limit set of the associated orbit contains a critical point which is not a local minimum

$$S_{t_0} := \{\mathbf{x}_0 = (\theta_0, m_0, v_0);\, \omega(\theta(t)) \ni \theta_\star, \text{ where } \theta_\star \text{ is a strict saddle}\}$$

The main result of this subsection is the following:

Theorem 6 (Avoiding Saddle and Local Maximum points) *Suppose that assumptions 1, 2, 4, 5 and 6 are satisfied. If either $Q(0) > 0$ or $p(t) \equiv q(t) \equiv 0$, then the set S_{t_0} has Lebesgue measure zero for every $t_0 > 0$.*

It follows that, if $\mathbf{x}_0 = (\theta_0, m_0, v_0)$ is a random initial condition, then the solution $\mathbf{x}(t, t_0, \mathbf{x}_0) = (\theta(t), m(t), \theta(t))$ converges to a local minimum of f with total probability. Similar results are proved for discrete systems having isolated critical points in (Lee et al., 2016, 2019), using essentially the same method as in here. More precisely, we use the theory of *central-stable* manifold (for vector-fields).

3.5. Rate of Convergence

The study of the rate of convergence of $f(\theta(t))$ to the minimum value $f(\theta_\star)$ usually relies on a convexity assumption and a Lyapunov energy functional as in Su et al. (2016); Alvarez et al. (2002); Alvarez (2000) and Gadat and Panloup (2014). It is therefore natural to make the following assumption.

Assumption 7 *The function f is convex and admits a minimum point, that is, there exists $\theta_\star$ such that $f(\theta) \geq f(\theta_\star)$ for every $\theta \in \mathbb{R}^d$.*

Now, strictly speaking, we do not find a Lyapunov functional for (2.1), but a natural functional which allow us to prove convergence to a least a neighbourhood of a local minimum. For accelerated methods, the proposed functional corresponds to the standard Lyapunov energy used in many other works as in Su et al. (2016); Alvarez et al. (2002); Alvarez (2000); Gadat and Panloup (2014). More precisely, let $t_0 > \tilde{t}$ (as defined in Assumption 2) and consider the following functions

$$\mathcal{A}(t) = \int_{t_0}^{t} h(\tau)\mathcal{B}(\tau)d\tau, \qquad \mathcal{B}(t) = e^{\int_{t_0}^{t} r(\tau)d\tau} \int_{t}^{\infty} e^{-\int_{t_0}^{s} r(u)du} d\tau$$

$$\mathcal{C}(t) = \frac{1}{h(t)} \int_{t_0}^{t} h(\tau)\mathcal{B}(\tau)d\tau.$$

The expressions of $\mathcal{A}(t)$, $\mathcal{B}(t)$ and $\mathcal{C}(t)$ are simple to compute for all the expressions in table 1, as we show in §4. Note, furthermore, that these functions only depend on h and r (which re-enforce the heuristic that p and q can be chosen in a very flexible way). We are ready to introduce the energy functional used in this section:

$$\begin{aligned}\mathcal{E}(t, m, v, \theta) =& \mathcal{A}(t)\left(f(\theta) - f(\theta_\star)\right) \\ & + \frac{1}{2}\left\|[v+\varepsilon]^{1/4}(\theta - \theta_\star)\right\|^2 - \mathcal{B}(t)\left\langle \theta - \theta_\star, m\right\rangle + \frac{\mathcal{C}(t)}{2}\left\|\frac{m}{[v+\varepsilon]^{1/4}}\right\|^2.\end{aligned} \tag{3.2}$$

We now need the following assumptions in order to control the behaviour of this functional:

Assumption 8 *We make the following two assumptions:*

(a) $\lim_{t\to\infty} \int_{t_0}^{t} e^{-\int_{t_0}^{\tau} r(u)du} d\tau < +\infty$

(b) There exists $\tilde{t} > t_0$ such that for all $t \geq \tilde{t}$

$$\mathcal{B}^2(t) \leq \mathcal{C}(t) \quad \text{and} \quad 3\mathcal{B}(t) \leq \mathcal{C}(t)\left(2r(t) - \frac{q(t)}{2} + \frac{h'(t)}{h(t)}\right).$$

Note that Assumption 8(a) is necessary for the function $\mathcal{B}(t)$ to be well-defined, and that imposes an important constraint in the asymptotic behaviour of the function $r(t)$. More precisely, the limit $\lim_{t\to\infty} t^{1+\epsilon} r(t)$ must be zero for every $\epsilon > 0$, which implies that $r(t)$ has at most a pole of order 1 at infinity. Assumption 8(b) provides the asymptotic control on the derivative of the energy functional (3.2), and should be compared with Assumption 2. Once again, note that it is independent of the function p, and almost independent of q. We are now ready to state the main theorem of this section:

Theorem 7 *We assume that Assumptions 1, 2, 7 and 8 are all satisfied. Then for all $t \geq \tilde{t}$, where $\tilde{t}$ is given in Assumption 2, we have*

$$f(\theta) - f(\theta_\star) \leq \frac{1}{4\mathcal{A}(t)}\left[4\mathcal{E}(\tilde{t}, m(\tilde{t}), v(\tilde{t}), \theta(\tilde{t})) + \int_{\tilde{t}}^{t} p(u)\left\langle \frac{[\nabla f(\theta)]^2}{[v+\varepsilon]^{1/2}}, [\theta - \theta_\star]^2 \right\rangle du\right]$$

where $\mathcal{E}(t, m, v, \theta)$ is the Lyapunov functional (3.2). Furthermore, if either the system is non-adaptive (that is, $p(t) \equiv q(t) \equiv 0$) or if $\lim_{t\to\infty} p(t)/q(t) < \infty$ and assumption 4 is satisfied, then there exist two positive and finite constants $\mathcal{K}_1$ and $\mathcal{K}_2$ (which depend on f, θ_0, v_0 and ε) such that for all $t \geq \tilde{t}$:

$$f(\theta(t)) - f(\theta_\star) \leq \frac{1}{\mathcal{A}(t)}\left[\mathcal{E}(\tilde{t}, m(\tilde{t}), v(\tilde{t}), \theta(\tilde{t})) + \mathcal{K}_1 + \mathcal{K}_2 \int_{\tilde{t}}^{t} q(u)du\right].$$

It follows that the ODE (2.1) converges to the minimum point with rate of convergence of order at least:

$$\max\left\{1, \int_{t_0}^{t} q(u)du\right\} / \mathcal{A}(t).$$

It is true that for convex functional, classical convergence results of first-order methods do not require assumption 4 because most algorithms are linear in the gradient and convexity inequalities allow for complete control of the derivative of the Lyapunov energy. This is also the case in the above Theorem 7 for non-adaptive algorithms. But, this is not the case for adaptive algorithms for which the updates are highly non-linear in the gradient (in both v and θ). As a consequence, the upper bound of the energy's derivative depends on the history of the trajectories and assumption 4 is a sufficient condition to obtain the convergence. We recall that if f is strongly convex then it is coercive and the assumption is automatically satisfied. Obtaining convergence results without the compactness assumption 4 remains an important open question.

From Theorem 7, we observe that the rate of convergence to the global minimum depends on: the choices of h, r which in turn define the function $\mathcal{A}$; the choices of p, q which, without extra assumption on the loss function, degrade the rate of convergence. It follows from the first inequality of Theorem 7 that the rate of convergence of ODE (2.1) depends in an essential way from the asymptotic behavior of the term:

$$\left\| \frac{\nabla f(\theta)}{[v+\varepsilon]^{1/2}} \right\|. \tag{3.3}$$

This has been independently remarked in Chen et al. (2019), where the authors back up this intuition by numerical simulations. The authors also propose AdaFom, whose coefficients are chosen such that the sum of the terms (3.3) is telescopic, so they can be controlled, allowing convergence results. Note that our Theorem 7 recovers the expected (in the deterministic setting) rate of convergence of AdaFom (see Corollary 9). In general, the term (3.3) can not be controlled in a similar fashion, and the second inequality of Theorem 7 controls the term (3.3) only in terms of the functions h, r and q.

4. Convergence Results: Application to First Order Algorithms

In this section, we specify the choice of functions h, p, q, r corresponding to different optimization methods and apply each convergence theorem to these algorithms. All proofs are postponed to Appendix F. We start by a brief discussion on the assumptions which appear in this section, and we then move on to present differential equations and convergence results on Adam, AdaFom, Heavy Ball and Nesterov. Our results on Adam are new and we give full details of the derivations. Some of our results on AdaFom and Heavy Ball are, up to our knowledge, also new. Their proofs can be easily formalized by repeating the arguments used for Adam, and we present a "sketch of proofs" in order to provide a guideline on the necessary changes. Finally, we recover several known results of non-adaptive algorithms, such as sharp estimates for the rate of convergence of Nesterov (see Corollary 11).

4.1. On the Different Assumptions

In the convergence analysis, we recurrent make assumptions which were introduced in § 3. We briefly recall their meaning and situations where they are satisfied:

Assumption 4 states that the trajectory $\theta(t)$ is bounded. We recall that there are some very practical situations where this assumption is always satisfied; for example in the case of coercive objective functions (see the discussion after the definition of assumption 4).

Assumption 6 gives a condition on the nature and the degeneracy of the critical points of the objective function. It is used in the study of saddle points and local maximum points of the loss functions, and it also appears in Lee et al. (2019). Note that this assumption is satisfied for generic functions (e.g. Morse functions). See Remark 22 for further discussion.

Assumption 7 states that the loss function f is convex and admits a minimum point.

4.2. Adam

Adaptive Moment Estimation (Adam) proposed in Kingma and Ba (2014) is a famous variant of RMSprop that incorporates a momentum equation. We recall that Adam has three hyper-parameters: the learning rate s and the exponential rate of decay for the moment estimates $\beta_1, \beta_2 \in (0,1)$. The parameter ε is usually set to 10^{-8} to avoid dividing by zero. This parameter is typically not tuned. The algorithm reads as follows: for any constants $\beta_1, \beta_2 \in (0,1)$, $\varepsilon > 0$ and initial vectors $\theta_0 \in \mathbb{R}^d, m_0 = v_0 = 0$ and for all $k \geq 1$

$$\begin{cases} g_k = \nabla f(\theta_{k-1}) \\ m_k = \mu_k m_{k-1} + (1-\mu_k) g_k \\ v_k = \nu_k v_{k-1} + (1-\nu_k) g_k^2 \\ \theta_k = \theta_{k-1} - s\ m_k/(\sqrt{v_k} + \varepsilon). \end{cases} \tag{4.1}$$

where the two parameters for the moving average are given by

$$\begin{cases} \mu_k = \beta_1 (1-\beta_1^{k-1})/(1-\beta_1^k) \\ \nu_k = \beta_2 (1-\beta_2^{k-1})/(1-\beta_2^k). \end{cases}$$

We rewrite the update for the parameters θ such that

$$\theta_k = \theta_{k-1} - s\ m_k/\sqrt{v_k + \varepsilon}.$$

This change does not change the behavior of the algorithm. By modifying the order of the updates and the value of the initial conditions, we can rewrite the above algorithm in a more suitable way for our analysis. Indeed, let $\theta_0 \in \mathbb{R}^d$ be such that $\nabla_\theta f(\theta_0) \neq 0$ and $m_0 = \nabla_\theta f(\theta_0),\ v_0 = \nabla_\theta f(\theta_0)^2$, then the following recursive update rules are equivalent to Adam for all $k \geq 0$

$$\begin{cases} \theta_{k+1} = \theta_k - s\ m_k/\sqrt{v_k + \varepsilon} \\ g_{k+1} = \nabla f(\theta_{k+1}) \\ m_{k+1} = \mu_{k+2} m_k + (1-\mu_{k+2}) g_{k+1} \\ v_{k+1} = \nu_{k+2} v_k + (1-\nu_{k+2}) g_{k+1}^2 \end{cases} \tag{4.2}$$

As a consequence, the initial velocity is $\dot{\theta}_0 = -\operatorname{sign}(\nabla f(\theta_0))$.

4.2.1. Adam Differential Equation

Consider now the three parameter family of differential equations

$$\begin{cases} \dot{\theta} = -m/\sqrt{v + \varepsilon} \\ \dot{m} = g(t, \lambda, \alpha_1)\left(\nabla f(\theta) - m\right) \\ \dot{v} = g(t, \lambda, \alpha_2)\left(\nabla f(\theta)^2 - v\right) \end{cases} \tag{4.3}$$

where the coefficients in ODE (2.1) are given by

$$h \equiv r \equiv g(t, \lambda, \alpha_1) = \frac{1 - e^{-\lambda/\alpha_1}}{\lambda\left(1 - e^{-t/\alpha_1}\right)}, \qquad p \equiv q \equiv g(t, \lambda, \alpha_2) = \frac{1 - e^{-\lambda/\alpha_2}}{\lambda\left(1 - e^{-t/\alpha_2}\right)},$$

where $(\lambda, \alpha_1, \alpha_2)$ are positive real numbers. Note that the coefficients have a simple pole at $t = 0$ and, therefore, satisfy assumption 3. Now, let us consider the associated discretization

(2.2) with learning rate s and a sub-family of discrete models parametrized by $(\beta_1, \beta_2) \in (0,1) \times (0,1)$ which are given by

$$\lambda = s, \qquad \beta_i = e^{-\lambda/\alpha_i},\ i = 1, 2. \tag{4.4}$$

It easily follows that for $i = 1, 2$

$$sg((k+1)s, \lambda, \alpha_1) = 1 - \beta_1 \frac{1 - \beta_1^k}{1 - \beta_1^{k+1}} = 1 - \mu_{k+1},$$

which recovers Adam's discrete system (4.2), apart from small difference in the evaluation of μ. Therefore, Adam is an Euler discretization of system (4.3).

4.2.2. Adam without Rescaling Differential Equation

In the original formulation of the algorithm (as stated in (4.1)), the parameters μ and ν depends on the iterations k to correct for the bias induced by the moving average. These coefficients can also be taken constant $\mu = \beta_1$ and $\nu = \beta_2$, in which case we say that the algorithm is Adam *without rescaling.* In this case, it is easy to verify that the differential equation (4.3), with $g(t, \lambda, \alpha) = 1/\alpha$, is the continuous counter-part of the algorithm when we consider the sub-family given by $\beta_1 = (1 - s/\alpha_1)$ and $\beta_2 = (1 - s/\alpha_2)$.

4.2.3. Convergence of Adam

The next corollary contains the main results about the convergence of Adam.

Corollary 8 (Convergence of Adam) *Suppose that f is a C^2 function, $\varepsilon > 0$ and*

$$3 + \beta_2 > 4\beta_1, \qquad \textit{where} \qquad \beta_i = \exp(-\lambda/\alpha_i), \quad i = 1, 2.$$

Then the following convergence results for equation (4.3) *hold true.*

(0) Convergence of the gradient: *Suppose that the loss function f is bounded from below and its gradient ∇f is globally Lipschitz and bounded. Then $\nabla f(\theta(t)) \to 0$ when $t \to \infty$.*

(I) Topological convergence: *Under assumption 4, $f(\theta(t)) \to f_\star$, $m(t) \to 0$ and $v(t) \to 0$ when $t \to \infty$, where $f_\star$ is a critical value of f.*

(II) Non-local minimum avoidance: *Suppose that assumptions 4 and 6 are satisfied. Fix $t_0 > 0$ and denote by S_{t_0} the set of initial conditions $(\theta_0, m_0, v_0) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d_{\geq 0}$ such that $\theta_\star \in \omega(\theta(t))$, where $\theta_\star$ is not a local-minimum of f. Then S_{t_0} has Lebesgue measure zero.*

(III) Rate of convergence: *Under assumptions 4 and 7, there exists a constant $\mathcal{K} > 0$ which depends on f, θ_0 and v_0, so that:*

$$\lim_{t\to\infty} f(\theta(t)) - f(\theta_\star) < \mathcal{K} \frac{1 - e^{-\lambda/\alpha_2}}{\alpha_1(1 - e^{-\lambda/\alpha_1})} = \mathcal{K} \ln(1/\beta_1) \frac{1 - \beta_2}{s(1 - \beta_1)}.$$

The rate of convergence to this neighbourhood, furthermore, is of order $\mathcal{O}(1/t)$.

The proof of this result is given in Appendix F. Note that there is an apparent paradox between point (I) and (III) of the Corollary: the dynamics are convergent by (I), but the "fast" rate of convergence (of order $\mathcal{O}(1/t)$) can only be guaranteed to a neighborhood. This is no paradox, nevertheless, because Adam might converge very *slowly* once it attains the neighborhood given by point (III). In particular, the discrete version of Adam may not converge even in the deterministic case (see Proposition 12 below).

Point (III) has two other surprising consequences. First, it justifies the usual choice of practitioners to take the hyper parameter β_2 as close to 1 as possible while there is more flexibility in the choice of β_1 (even if it is convenient to take it close to 1) because:

$$\lim_{\beta_2 \to 1} 1 - \beta_2 = 0, \quad \lim_{\beta_1 \to 1} \frac{\ln(1/\beta_1)}{1-\beta_1} = 1.$$

Second, the size of the neighbourhood of fast convergence (of order $\mathcal{O}(1/t)$) is controlled by a constant $\mathcal{K}$ which only depends on f and the initial conditions. This indicates that the success of Adam for certain loss functions (e.g. loss functions in deep learning) might be associated to *features* on the "class" of loss functions considered.

4.2.4. A Variant of Adam: AdaFom

In Chen et al. (2019), the authors propose a variation of Adam algorithm which can be guaranteed to have good convergence rate. We provide the differential equation associated to this algorithm. We recover its expected deterministic convergence rate following Corollary 3.2 in Chen et al. (2019). The algorithm reads as follows: for any constants $\beta_1 \in (0,1)$, $\varepsilon \geq 0$ and initial vectors $\theta_0 \in \mathbb{R}^d, m_0 = v_0 = 0$ and for all $k \geq 1$

$$\begin{cases} g_k = \nabla f(\theta_{k-1}) \\ m_k = \mu_k m_{k-1} + (1-\mu_k) g_k \\ v_k = (1 - 1/k) v_{k-1} + 1/k\, g_k^2 \\ \theta_k = \theta_{k-1} - s\ m_k / (\sqrt{v_k} + \varepsilon). \end{cases} \tag{4.5}$$

where the parameter for the moving average are given by (just as in Adam):

$$\mu_k = \beta_1 (1 - \beta_1^{k-1}) / (1 - \beta_1^k).$$

Consider now the two parameter family of differential equations

$$\begin{cases} \dot{\theta} = -m/\sqrt{v + \varepsilon} \\ \dot{m} = \dfrac{1 - e^{-\lambda/\alpha}}{\lambda \left(1 - e^{-t/\alpha}\right)} \left(\nabla f(\theta) - m\right) \\ \dot{v} = \dfrac{1}{t} \left(\nabla f(\theta)^2 - v\right) \end{cases} \tag{4.6}$$

where (λ, α) are positive real numbers. Just as in the case of Adam, consider the associated discretization (2.2) with learning rate s and a sub-family of discrete models parametrized by $\lambda = s$ and $\beta = e^{-\lambda/\alpha}$. The reader may verify, following the same steps of the analysis of Adam that the discretization of this sub-family recovers AdaFom algorithm (4.5). We now turn to the convergence analysis :

Corollary 9 (Convergence of AdaFom) *Suppose that f is a C^2 function, $\varepsilon > 0$ and assumption 4 is satisfied. Then the following convergence results for* (4.6) *hold true*

(I) Topological convergence: $f(\theta(t)) \to f_\star$ *and* $m(t) \to 0$ *when* $t \to \infty$, *where* $f_\star$ *is a critical value of* f.

(III) Rate of convergence: *Under assumption 7,* $f(\theta(t)) \to f(\theta_\star)$ *with the rate* $\mathcal{O}(\ln(t)/t)$.

Note that Theorems 4 and 6 can not be applied to ODE (4.6) in a direct way because $\lim_{t\to 0} p(t) = \lim_{t\to 0} q(t) = \lim_{t\to 0} 1/t = 0$.

4.3. Accelerated Methods

In this section, we study accelerated methods via our generalized dynamical system (2.1). We recover known results from the literature.

4.3.1. Heavy Ball

Following Alvarez (2000), we consider the Heavy ball second order differential equation

$$\ddot{x} + \gamma \dot{x} + \nabla f(x) = 0, \tag{4.7}$$

where $\gamma > 0$. By taking $\theta = x$ and $m = -\dot{x}$ (and $v \equiv 1$), we obtain system (2.1) with

$$h(t) \equiv 1, \qquad r(t) \equiv \gamma, \qquad \text{and} \qquad p(t) \equiv q(t) \equiv 0.$$

Note that Equation (2.2) simplifies to

$$\begin{cases} \theta_{k+1} = \theta_k - s m_k \\ m_{k+1} = (1 - s\gamma) m_k + s \nabla f(\theta_{k+1}) \end{cases} \tag{4.8}$$

which corresponds to the classical Heavy ball methods with damping coefficient $\beta = 1 - s\gamma$, momentum variable $n_k = s m_k$ and learning rate $\alpha = s^2$. Implicit discretization has also been considered in Alvarez (2000). From our analysis on the continuous system (4.7), we recover results given in Lee et al. (2019) and Ghadimi et al. (2015) for the discrete update rules (4.8).

Corollary 10 (Convergence of Heavy Ball) *Suppose that f is a C^2 function. Then the following convergence results for equation* (4.7) *hold true*

(I) Topological convergence: *Under assumption 4,* $f(\theta(t)) \to f_\star$ *and* $m(t) \to 0$ *when* $t \to \infty$, *where* $f_\star$ *is a critical value of* f.

(II) Non-local minimum avoidance: *Suppose that assumptions 4 and 6 are satisfied. Fix* $t_0 > 0$ *and denote by* S_{t_0} *the set of initial conditions* $(\theta_0, m_0) \in \mathbb{R}^d \times \mathbb{R}^d$ *such that* $\theta_\star \in \omega(\theta(t))$, *where* $\theta_\star$ *is not a local-minimum of* f. *Then* S_{t_0} *has Lebesgue measure is zero.*

(III) Rate of convergence: *Under assumption 7,* $f(\theta(t)) \to f(\theta_\star)$ *with the rate* $\mathcal{O}(1/t)$.

4.3.2. Nesterov

Following Su et al. (2016), we consider the Nesterov's second order differential equation, parametrized by the constant $r > 0$,

$$\ddot{x} + \frac{r}{t}\dot{x} + \nabla f(x) = 0. \tag{4.9}$$

Similarly as in the Heavy Ball case, we define $\theta = x$ and $m = -\dot{x}$ and write the above equation as a specialization of system (2.1) given by:

$$\begin{cases} \dot{\theta} & = -m \\ \dot{m} & = \nabla f(\theta) - r/t \cdot m \end{cases} \tag{4.10}$$

where $r > 0$. In its standard formulation $r = 3$. In Su et al. (2016), the authors studied a slightly different forward Euler scheme and proved that the difference between the numerical scheme and the Nesterov algorithm goes to zero in the limit $s \to 0$. So ODE (4.10) can be considered as the continuous counterpart of Nesterov discrete algorithm. We are ready enunciate the main convergence result for Nesterov's differential equation, which have been previously proved in Su et al. (2016); Attouch et al. (2019); Attouch et al. (2018).

Corollary 11 (Convergence Rate of Nesterov) *Suppose that f is a C^2 function and that assumption 7 is satisfied. Then the following convergence result for equation* (4.10) *hold true: $f(\theta) \to f(\theta_\star)$ when $t \to \infty$ with rate of convergence:*

$$\mathcal{O}(1/t^2), \text{ if } r \geq 3$$
$$\mathcal{O}(1/t^{2r/3}), \text{ if } r \leq 3.$$

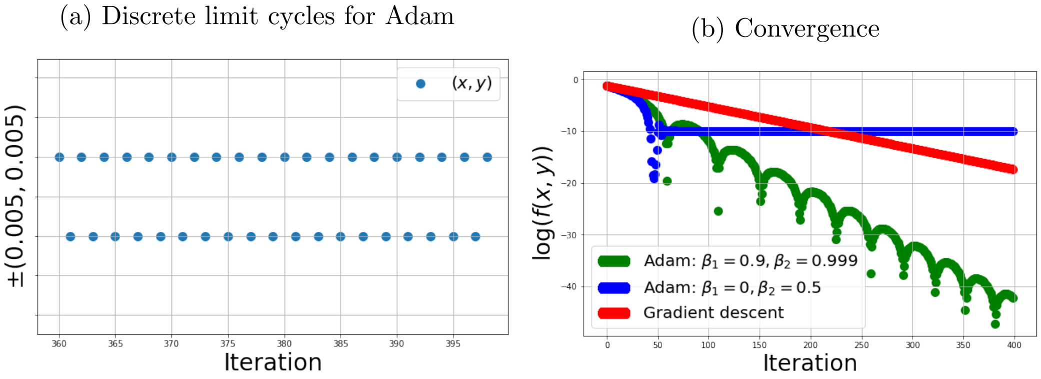
5. Considerations about the Discrete Algorithm (2.2)

In this section, we draw insights on the discrete algorithm (2.2) from our analysis on the continuous dynamical system.

5.1. The Discrete Dynamics does not Necessarily Converge.

One strong limitation of Adam is the existence of discrete limit cycles in the sense that the algorithm produces oscillations that never damp out. If the discrete dynamics reaches such an equilibrium, the difference $f(\theta_k) - f(\theta_\star)$ can not converge arbitrarily close to zero with an increasing number of steps. However, it reaches a neighborhood of the critical point whose radius is determined by the learning rate s. Decaying the learning rate is, therefore, necessary to obtain convergence of the dynamics Numerically, we found that Adam with $\beta_1 > 0$ suffers from the same phenomena but the limit cycles are more difficult to establish. We believe that the existence of such cycles depends on the local curvature of the function f near the optimum.

Proposition 12 (Existence of a discrete limit cycle for Adam) *Let $\beta_1 = 0$ and $f(\theta) = \theta^2/2$. Then there exists a discrete limit cycle for Equation* (4.2)*.*


(a) Discrete limit cycles for Adam

(b) Convergence

Figure 1: Illustration of discrete limit cycles for the Adam's algorithm with $\varepsilon = 10^{-8}, \beta_2 = 0.5, s = 10^{-2}$. **a)** Limit cycle of period two for Adam. The algorithm oscillates between two points $(0.005, 0.005)$ and $(-0.005, -0.005)$. **b)** Plot of the logarithm of f versus the number of iterations. The loss plateau after 50 iterations.

Proof Let us assume that there exists a k such that $\theta_k = s/2$ and that $v_k = (s/2)^2$, where s is the learning rate. It easily follows from the update rules that

$$\begin{aligned}\theta_{k+1} &= \theta_k - s\frac{\nabla f(\theta_k)}{\sqrt{v_k}} = \frac{s}{2} - s = -\theta_k \\ v_{k+1} &= (s/2)^2\end{aligned}$$

Therefore $\theta_{k+2} = -\frac{s}{2} + s = \theta_k$ and the system has entered a discrete equilibrium. ■

We illustrate this behavior in Figure 1 on the strongly convex toy function $f(x, y) = x^2 + y^2$.

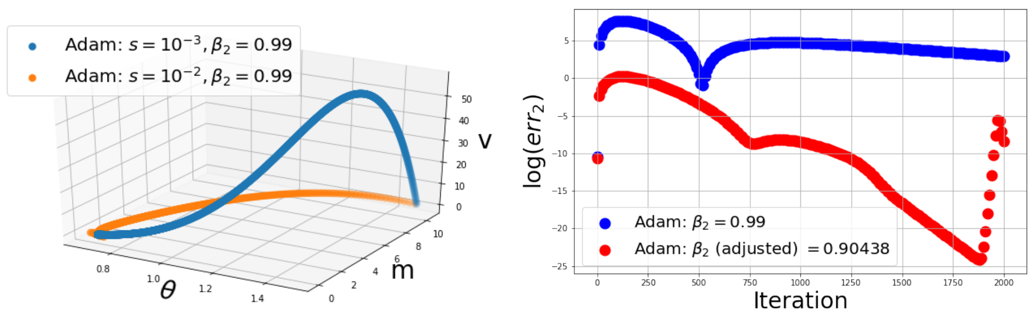
It is important to note that the value of the gap between $f(\theta_k)$ and $f(\theta_\star)$ depends on the learning rate. Choosing a smaller learning rate reduces the gap, but doesn't remove it.

5.2. The Hyper-Parameters β_1, β_2 in Adam Should be Tuned in Terms of the Learning Rate

The second observation is related to the hyper-parameters of the optimizers and give important guidance on how to tune them. As observed in Section 4.2, the parameters β_1 and β_2 are chosen as functions of the learning rate s and parameters α_1 and α_2. It is often the case in practice (in particular in stochastic optimization) to decay the learning rate during the training process. By doing so, the discrete dynamics is completely modified unless the β's are adjusted to keep the parameters α_i constant. Therefore, once a particular choice of hyper-parameters seems promising, a decay in the learning rate should be accompanied by changing the hyper-parameters β_i according to the formula (4.4), which we recall here

$$\beta_i = e^{-s/\alpha_i},\ i = 1, 2.$$

Indeed, by doing so the underlying dynamics is preserved. We illustrate this in plot **b)**, Figure 2, where we compute the logarithm of the error between different trajectories



(a) Trajectories

(b) Comparison of L^2 norms

Figure 2: Fixing β_2 and changing the learning rate s lead to different dynamics. **a)** Trajectories of Adam (1) & (2) when only the learning rate is changed. **b)** Comparison of the error between trajectories (1) & (2) and (1) & (3). As expected the discrepancies between (1) & (3) is very small.

1. The reference dynamics: $\beta_2 = 0.99$ and $s = 0.001$. According to formula (4.4), $\alpha_2 = -0.001/\log(0.99) \approx 0.0995$
2. Second dynamics: $\beta_2 = 0.99$ and $s = 0.01$
3. Third dynamics: same learning rate as the second dynamics $s = 0.01$ but we adjust the hyper-parameter: $\beta_2 = \exp(-0.01/0.0995) = 0.90438$.

5.3. Convergence Properties of Adam Depend on The Class of Loss Functions

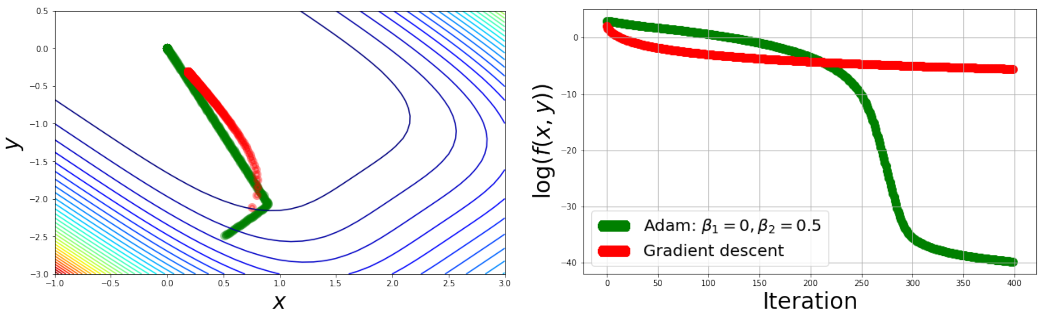
The convergence analysis in the convex case, given in Theorem 7, seem to indicate that Adam is a rather slow algorithm because quick convergence (in the order $o(1/t)$) is only guaranteed in a neighborhood of the global minimum. Under a closer look, however, we note that the size of the neighborhood might be very small depending on the class of loss functions because of its impact on the term (3.3) and the constant $\mathcal{K}$. In particular, we believe that there are situations where Adam should perform consistently better than other algorithms, provided that the hyper-parameters are well-chosen. We believe this is the case for flat loss functions as illustrated in Figure 3. We intend to deepen this remark in forthcoming works.

6. Guidelines for Future Adaptive Algorithms

Existence and uniqueness of solutions for ODE (2.1) holds for almost arbitrary functions $h(t)$, $r(t)$, $p(t)$ and $q(t)$. Our convergence analysis, nevertheless, imposes important restriction on the choice of these functions (see Assumptions 5, 6, 8), which we now discuss. From Theorem 7, the rate of convergence to the global minimum is at least given by the decay of the function

$$\max\left\{1, \int_{t_0}^{t} q(u)du\right\} / \mathcal{A}(t),$$

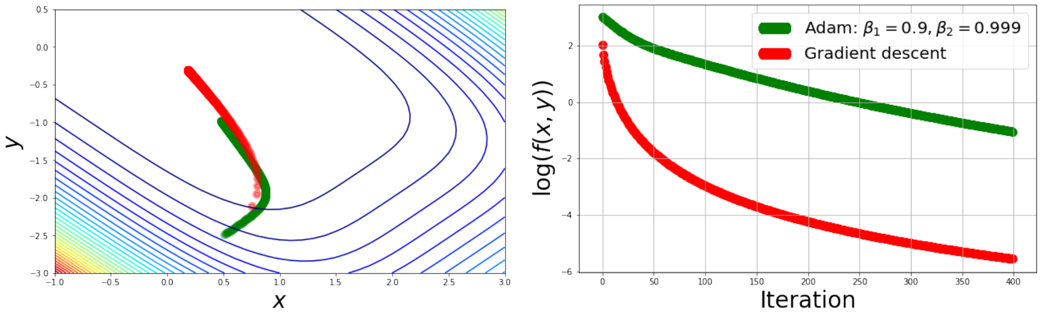
where $\mathcal{A}$ depends only on h and r. It is natural to seek for functions $h(t)$, $r(t)$, $p(t)$ and $q(t)$ such that this upper-bound decays faster to zero, while satisfying the conditions 2 and 8.



(a) $f(x, y) = (x + y)^4 + (x/2 - y/2)^4$

(b) Rate of convergence

Figure 3: Comparison between gradient descent and Adam. Gradient Descent converges faster initially when the gradients are large but Adam outperforms Gradient descent after entering the flat region. Both trajectories start from the point $(0.5, -2.5)$.



(a) $f(x, y) = (x + y)^4 + (x/2 - y/2)^4$

(b) Rate of convergence

Figure 4: Comparison between gradient descent and Adam. Gradient Descent outperforms Adam in this example because β_1, β_2 are large and Adam keeps memory of the past large gradients. Both trajectories start from the point $(0.5, -2.5)$.

Note that in Theorem 7, we obtained tighter estimates on the rate of convergence depending on the history of the gradient and the variables v and θ. This suggests that the efficiency of the algorithm does not only depend on the choice of the functions $h(t)$, $r(t)$, $p(t)$ and $q(t)$ but also on the path the dynamics is taking and therefore on properties of the loss.

On the flexibility of the coefficients $h(t)$ and $r(t)$. We start by recalling that Assumption 8(a) implies a strong constraint on the function $r(t)$. Indeed, it is necessary that $\lim_{t\to\infty} r(t) \cdot t^{1+\epsilon} = 0$ for all $\epsilon > 0$. This is a strong restriction, which have consequences to the choice of $h(t)$. In order to illustrate this, let us consider two examples $r_1(t) = r_1$ and $r_2(t) = r_2/t$ with $r_2 > 1$, which are the natural allowed examples. With these choices of functions, we seek for h such that $\mathcal{A}$ has at least linear growth. In the notation of §§ 3.5, we have that:

$$\mathcal{B}_1(t) = 1/r_1, \quad \mathcal{B}_2(t) = t/(r_2 - 1)$$

This imposes strong restrictions on $h(t)$, indeed:

$$\lim_{t\to\infty} h_1(t) > 0, \quad \lim_{t\to\infty} t \cdot h_2(t) > 0.$$

It follows that the natural reasonable choices are one of the following three:

1. $r(t) = r$, $h(t) = h$ (c.f. Heavy Ball, Adam and AdaFom). Under these conditions, one can expect a convergence rate of order at least $\mathcal{O}(1/t)$ and convergence guarantees (including avoidance of saddle points) even in the non-convex setting;

2. $r(t) = r/t$, $h(t) = h$ (c.f Nesterov). Under these conditions, one can expect a fast convergence rate of order at least $\mathcal{O}(1/t^2)$, but we obtain fewer guarantees in the non-convex setting because of the convergence to zero of the function r;

3. $r(t) = r/t$ and $h(t) = h/t$. Under these conditions, one can expect a convergence rate of order at least $\mathcal{O}(1/t)$, but we obtain fewer guarantees in the non-convex setting;

Note that Assumptions 2 and 8 impose further relations on the hyper-parameters r and h.

On the flexibility of the coefficients $p(t)$ and $q(t)$. In sharp contrast with the previous analysis, the coefficients $p(t)$ and $q(t)$ require fewer restrictions. In particular, note that assumptions 2, 5 and 8 are almost independent on these coefficients, besides mild constraints on $q(t)$. Our analysis leads to an interesting dilemma, nevertheless, which deserves further investigation. At the one hand, Theorem 6 guarantess avoidance of saddle points under the necessary condition that $\lim_{t\to\infty} q(t) > 0$ (c.f. Adam). On the other hand, in order to obtain faster convergence, Theorem 7 indicates that the function $q(t)$ should have a fast decay to zero to control the growth of $\int_{t_0}^{t} q(u)du$. Indeed:

(i) if $\lim_{t\to\infty} q(t) > 0$ (c.f Adam), then the denominator $\int_{t_0}^{t} q(u)du$ has linear growth which degrades the rate of convergence of the algorithm.

(ii) if $\lim_{t\to\infty} t \cdot q(t) > 0$ (c.f. AdaFom) then the denominator has logarithmic growth.

(iii) if $\lim_{t\to\infty} t^{1+\epsilon} \cdot q(t) = 0$ for some $\epsilon > 0$, then there is no loss in the expected convergence rate.

It is, of course, interesting that an optimization algorithm avoids saddle points and converges as fast as possible. This is an intriguing point, which we feel deserves further empirical investigation.

Final remarks. From the analysis outlined above, we feel that the combination of choices 1.*ii* (AdaFom), 1.*iii* and 2.*ii* are promising, at least from the perspective of our current analysis, and deserve further empirical investigation. Moreover, we hope to explore further different choices of functions $p(t)$ and $q(t)$ as well as the design of hybrid algorithms, which are also supported by this analysis.

7. Conclusion and Final Discussion

The main objective of this work is to provide a theoretical framework to study adaptive algorithms. The proposed continuous dynamical system (2.1) is flexible enough to encompass commonly used adaptive algorithms (as we show in Section 4), but stays specific enough to allow simple proofs and guidelines. Our work shows that adaptive dynamics converge to a critical locus of the loss function but possibly at a slower rate than non-adaptive algorithms. Due to the nature of the adaptivity, the convergence rate is not just a non-increasing function of time but also depends on the gradient history. The performance of adaptive algorithms is linked to the trajectory taken by the dynamics and properties of the loss function. We also analyze how different choices of coefficients h, r, p and q impact on the convergence of the dynamics and we suggested several possible algorithms to be tested (see §§ 6). It supports our interest in linking specific choices of adaptive algorithms (more precisely, specific choices of the coefficients h, r, p and q) with properties from the loss function. We intend to pursue this direction in future works.

The deterministic convergence analysis leads to natural conjectures on the convergence in the discrete and stochastic setting. In particular, we believe that the Lyapunov functional (3.2) can be adapted to the stochastic discrete framework (Gadat et al., 2018). We note that, nevertheless, a precise correspondence between results valid for a continuous ODE and the stochastic discrete counterparts is far from being obvious. Indeed, recall that Adam and RMSprop are not always converging in the stochastic setting, even for a convex loss function (S. Reddi and Kumar, 2018). We expect, therefore, new restrictions on the coefficients $h(t)$, $r(t)$, $p(t)$ and $q(t)$, as well as on the loss function and the learning rate. We believe that those conditions will be different compared to SGD.

Acknowledgments

We would like to thank Daniel Panazzolo for a private communication concerning central-stable manifolds, and Ioana Gavra for pointing out the reference Zaheer et al. (2018) to us. This work was supported by Borealis AI. We would further like to thank all of the referee's for the very useful suggestions, which greatly improved the presentation of the paper.

Appendix A. Existence and Uniqueness of Solutions

A.1. The Cauchy Problem for $t_0 > 0$ under Assumption 2.

We compute elementary bounds for the solutions of the ODE (2.1), under the additional assumption 2 with $\tilde{t} = 0$ and that f is bounded from below (or under Assumption 4).

Indeed, let $t_0 > 0$ and an initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$ be fixed. Because of Assumption 1, we are in conditions of Picard Theorem, so there exists a solution $\mathbf{x}(t)$ with initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$ and interval of definition $[t_0, t_\infty[$. The crucial point with this assumption 2 with $\tilde{t} = 0$, is that we can apply Lemma 2 in order to find a constant $\mathcal{K}(\mathbf{x}_0, t_0)$ which depends on the initial condition, such that:

$$\left\| \frac{m(t)}{\sqrt{v(t) + \varepsilon}} \right\| \leq \mathcal{K}(\mathbf{x}_0, t_0), \quad \forall t \in [t_0, t_\infty[, \tag{A.1}$$

which implies that $\dot{\theta}(t)$ is uniformly bounded. We conclude that:

$$\|\theta(t)\| \leq \|\theta_0\| + \mathcal{K}(\mathbf{x}_0, t_0)(t - t_0), \quad \forall t \in [t_0, t_\infty[. \tag{A.2}$$

It follows that: either $t_\infty = \infty$, in which case we are done, or $t_\infty < \infty$ and $\theta(t)$ satisfies assumption 4. In order to conclude, it is enough to treat this last case:

Lemma 13 *Let $t_0 > 0$ and $\mathbf{x}_0 = (\theta_0, m_0, v_0)$ be fixed. Under assumption 1, suppose that the gradient $\nabla f(\theta(t))$ is bounded, that is, $L_g = \sup\{\|\nabla f(\theta(t))\|\,; t \geq t_0\}$ is well-defined (this is verified whenever assumption 4 is satisfied). Then there exists an unique solution $\mathbf{x}(t) = (\theta(t), m(t), v(t))$ of (2.1) with initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$, and which is defined for all t in $[t_0, \infty)$. Furthermore, we have $v(t) \geq 0$ for all $t \in [t_0, \infty)$ and:*

$$\|m(t)\| \leq \|m(t_0)\| + L_g d \int_{t_0}^{t} h(s)ds, \quad \|v(t)\| \leq \|v(t_0)\| + L_g^2 d \int_{t_0}^{t} p(s)ds \tag{A.3}$$

where we recall that d stands for the dimension of the space. If we suppose that $r(t) \not\equiv 0$ and $q(t) \not\equiv 0$, furthermore, then:

$$\|m(t)\| \leq \|m(t_0)\| + L_g d \sup_{s\in[t_0,t]} \left\{\frac{h(s)}{r(s)}\right\}, \quad \|v(t)\| \leq \|v(t_0)\| + L_g^2 d \sup_{s\in[t_0,t]} \left\{\frac{p(s)}{q(s)}\right\}$$

Proof By assumption 1 and classical ODE's, there exists a solution $\mathbf{x}(t) = (\theta(t), m(t), v(t))$ of system (2.1) with maximal interval of definition $[t_0, T)$ and initial conditions $\mathbf{x}(t_0) = \mathbf{x}_0 = (\theta_0, m_0, v_0)$. Now, consider the functions:

$$a(t) = \exp\left(\int_{t_0}^{t} r(s)ds\right), \quad b(t) = \exp\left(\int_{t_0}^{t} q(s)ds\right)$$

which are increasing functions bigger than 1 (for all $t \geq t_0$). We note that:

$$\frac{d}{dt}(m \cdot a(t)) = a(t)h(t)\nabla f(\theta), \quad \frac{d}{dt}(v \cdot b(t)) = b(t)p(t)\nabla f(\theta)^2.$$

Next, by hypothesis we can assume that $|\nabla f(\theta(t))| \leq L_g$ for some positive real number L_g. We easily get inequalities (A.3) and in turn conclude that $T = \infty$. Finally, if $r(t) \not\equiv 0$, we get by direct integration:

$$\begin{aligned}|m_i(t)| &\leq |m_i(t_0)| + L_g \frac{1}{a(t)} \int_{t_0}^{t} r(s)a(s)\frac{h(s)}{r(s)}ds \\ &= |m_i(t_0)| + L_g \sup_{s\in[t_0,t]} \left\{\frac{h(s)}{r(s)}\right\} \frac{a(t) - a(t_0)}{a(t)} \leq L \sup_{s\in[t_0,t]} \left\{\frac{h(s)}{r(s)}\right\}\end{aligned}$$

A similar computation holds whenever $q(t) \not\equiv 0$, which concludes the Lemma. ■

In order to prove Theorem 3 without assumption 2 with $\tilde{t} = 0$, it is necessary to obtain the estimate (A.1). This is possible whenever $p(t) \not\equiv 0$, as we will show in Lemma 15 below.

A.2. A Priori Estimates and Global Solution

We state the following variant of the Gronwall's Lemma:

Lemma 14 *Let $\varphi : [t_0, t_1] \to \mathbb{R}_{>0}$ be absolutely continuous strictly non-negative function and suppose φ obeys the differential inequality $\varphi'(t) \leq \gamma(t)\varphi(t) + \beta(t)\varphi^{\alpha}(t)$ for $0 \leq \alpha < 1$ and for almost every $t \in [t_0, t_1]$, where β, γ are continuous. Then for all $t \in [t_0, t_1]$*

$$\varphi(t) \leq \left[e^{(1-\alpha)\int_{t_0}^{t}\gamma(s)ds}\varphi(t_0)^{1-\alpha} + \int_{t_0}^{t}(1-\alpha)e^{(1-\alpha)\int_{s}^{t}\gamma(u)du}\beta(s)ds\right]^{1/(1-\alpha)}.$$

Proof It is enough to apply Gronwall's Lemma to the following inequality:

$$\begin{aligned}[\varphi^{1-\alpha}]' = (1-\alpha)\varphi^{-\alpha}\varphi' &\leq (1-\alpha)\varphi^{-\alpha}(t)\left(\gamma(t)\varphi(t) + \beta(t)\varphi^{\alpha}(t)\right) \\ &= (1-\alpha)\gamma(t)\varphi^{1-\alpha}(t) + (1-\alpha)\beta(t).\end{aligned}$$

■

We now turn to more precise estimates of the functions $\mathbf{x}(t) = (\theta(t), m(t), v(t))$ in order to prove existence and uniqueness of the solution of Equation (2.1). We start by controlling the derivative of θ:

Lemma 15 (Estimate of $\dot{\theta}$) *Let $0 < t_0 < T < \infty$ be fixed and suppose that $p(t) \not\equiv 0$. For any $s, t \in [t_0, T]$ such that $s \leq t$, we have that:*

$$\left\| \frac{m(t)}{\sqrt{v(t)+\varepsilon}} \right\|^2 \leq e^{\int_s^t q(u) - 2r(u) du} \left\| \frac{m(s)}{\sqrt{v(s)+\varepsilon}} \right\|^2 + d \int_s^t e^{\int_u^t q(a) - 2r(a) da} \frac{h^2(u)}{p(u)} du$$

Proof It follows from direct computation that:

$$\begin{aligned}\frac{d}{dt}\frac{1}{2}\left\|\frac{m}{\sqrt{v+\varepsilon}}\right\|^2 = h(t)&\left\langle \frac{m}{\sqrt{v+\varepsilon}}, \frac{\nabla f(\theta)}{\sqrt{v+\varepsilon}} \right\rangle - r(t)\left\|\frac{m}{\sqrt{v+\varepsilon}}\right\|^2 \\ &- \frac{1}{2}p(t)\left\|\frac{m\nabla f(\theta)}{v+\varepsilon}\right\|^2 + \frac{q}{2}\left\|\frac{m\sqrt{v}}{v+\varepsilon}\right\|^2\end{aligned}$$

Now, note that by completing the square

$$\begin{aligned}h(t)\left\langle \frac{m}{\sqrt{v+\varepsilon}}, \frac{\nabla f(\theta)}{\sqrt{v+\varepsilon}} \right\rangle - \frac{p(t)}{2}\left\|\frac{m\nabla f(\theta)}{v+\varepsilon}\right\|^2 &= h(t)\sum_{i=1}^d \frac{m_i \partial_i f(\theta)}{v_i+\varepsilon} - \frac{p(t)}{2}\sum_{i=1}^d \left(\frac{m_i\partial_i f(\theta)}{v_i+\varepsilon}\right)^2 \\ &= -\frac{p(t)}{2}\left\|\frac{m\nabla f(\theta)}{v+\varepsilon} - \frac{h(t)}{p(t)}\right\|^2 + \frac{h^2(t)}{2p(t)}d\end{aligned}$$

where d is the dimension of the state space. Hence,

$$\begin{aligned}\frac{d}{dt}\frac{1}{2}\left\|\frac{m}{\sqrt{v+\varepsilon}}\right\|^2 &= -\frac{p(t)}{2}\left\|\frac{m\nabla f(\theta)}{v+\varepsilon} - \frac{h(t)}{p(t)}\right\|^2 + \frac{h^2(t)}{2p(t)}d - r(t)\left\|\frac{m}{\sqrt{v+\varepsilon}}\right\|^2 + \frac{q(t)}{2}\left\|\frac{m\sqrt{v}}{v+\varepsilon}\right\|^2 \\ &\leq \frac{h^2(t)}{2p(t)}d + \left(\frac{q(t)}{2} - r(t)\right)\left\|\frac{m}{\sqrt{v+\varepsilon}}\right\|^2\end{aligned}$$

and we easily conclude from Gronwall's Lemma. ■

The above Lemma allow us to control $\theta(t)$. The next Lemma replaces Assumption 4 in the existence proof:

Lemma 16 (Estimate of θ) *Let $0 < t_0 < T < \infty$ be fixed. For any $s, t \in [t_0, T]$ such that $s \leq t$:*

$$\|\theta(t)\|^2 \leq \left[\|\theta(s)\| + \int_s^t \left\|\frac{m}{\sqrt{v+\varepsilon}}\right\| du\right]^2.$$

and, in particular, if $p(t) \not\equiv 0$:

$$\begin{aligned}\|\theta(t)\| \leq \|\theta(s)\| &+ \left\|\frac{m(s)}{\sqrt{v(s)+\varepsilon}}\right\| \int_s^t e^{\frac{1}{2}\int_s^u q(a)-2r(a)da} du \\ &+ \sqrt{d}\int_s^t \left(\int_s^u e^{\int_a^u q(b)-2r(b)db}\frac{h^2(a)}{p(a)}da\right)^{1/2} du\end{aligned}$$

Proof From the Cauchy-Schwarz inequality, we obtain

$$\frac{d}{dt}\frac{1}{2}\|\theta(t)\|^2 = -\left\langle \theta, \frac{m}{\sqrt{v+\varepsilon}}\right\rangle \leq \|\theta\| \left\|\frac{m}{\sqrt{v+\varepsilon}}\right\|.$$

We apply Lemma 14 to $\varphi(t) = \frac{1}{2}\|\theta(t)\|^2$, $\beta(t) = \sqrt{2}\left\|\frac{m}{\sqrt{v+\varepsilon}}\right\|$ and $\alpha = 1/2$ in order to get the first inequality. The second inequality follows from the first together with the estimate of Lemma 15. ■

Lemma 17 (Estimate of m and v) *Let $0 < t_0 < T < \infty$ be fixed. For any $s, t \in [t_0, T]$ such that $s \leq t$:*

$$\begin{aligned}\|m(t)\|^2 &\leq \left[e^{-\int_{t_0}^t r(s)ds}\|m(t_0)\| + \int_{t_0}^t e^{-\int_s^t r(u)du} h(s)\|\nabla f(\theta)\| ds\right]^2 \\ \left\|\sqrt{v(t)}\right\|^2 &\leq e^{-\int_{t_0}^t q(s)ds}\left\|v^{1/2}(t_0)\right\|^2 + \int_{t_0}^t e^{-\int_s^t q(u)du} p(s)\|\nabla f(\theta)\|^2 ds, \\ \|v(t)\| &\geq e^{-\int_{t_0}^t q(s)ds}\|v_0\|.\end{aligned}$$

Proof From Cauchy Schwarz,

$$\frac{d}{dt}\frac{1}{2}\|m(t)\|^2 = h(t)\langle m, \nabla f(\theta)\rangle - r(t)\|m\|^2 \leq h(t)\|m\|\|\nabla f(\theta)\| - r(t)\|m\|^2,$$

and we just need to apply Lemma 14 in order to conclude the first inequality. Next, we apply Gronwall's lemma to

$$\frac{d}{dt}\left\|v^{1/2}(t)\right\|^2 = \left\langle \frac{p(t)\left[\nabla f(\theta)\right]^2 - q(t)v}{v^{1/2}}, v^{1/2}\right\rangle = p(t)\|\nabla f(\theta)\|^2 - q(t)\left\|v^{1/2}\right\|^2$$

in order to get the second inequality. Finally, it is enough to apply Gronwall's lemma to $\dot{v} = p(t)[\nabla f(\theta)]^2 - q(t)v \geq -q(t)v$, in order to obtain the third inequality. ■

A.3. Existence and Uniqueness: Proof of Theorem 3 when $t_0 > 0$

Indeed, let $t_0 > 0$ and an initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$ be fixed. Because of Assumption 8, we are in conditions of Picard Theorem, so there exists a solution $\mathbf{x}(t)$ with initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$ and interval of definition $[t_0, t_\infty[$. If $t_\infty = \infty$, we are done, so suppose by contradiction that $t_\infty < \infty$. Then: if $p(t) \not\equiv 0$, by Lemma 16 we conclude that $\theta(t)$ is bounded; otherwise, assumption 2 with $\tilde{t} = 0$ is satisfied and f is bounded by below, and by inequality (A.2) we conclude that $\theta(t)$ is bounded. In either case we are in conditions to apply Lemma 13 which implies that $t_\infty = \infty$, a contradiction.

A.4. Existence and Uniqueness for $t_0 = 0$

In the previous section, we proved that for all $T > 0$, there exists a unique solution to the system (2.1) in the space $C^1([t_0, T]; \mathbb{R}^n)$ for any strictly positive time t_0. The purpose of this section is to extend this result to solutions starting at $t_0 = 0$. Classical results on differential equations do not apply directly here because the functions h, r, k, q are allowed to have a pole of order one at $t = 0$ (see Assumption 3).

We follow a standard argument in dynamical systems: we will approximate the solution of ODE (2.1) by a sequence of functions with good convergence properties. In this section, we limit ourselves to describing which is the sequence of functions, and we leave the "good convergence properties" for later on. Indeed, we consider the orbits $\mathbf{x}_\delta$, for $\delta > 0$, which are solutions to the equation

$$\begin{cases} \dot{\theta_\delta}(t) = -m_\delta(t)/\sqrt{v_\delta(t) + \varepsilon} \\ \dot{m_\delta}(t) = h_\delta(t)\nabla f(\theta_\delta(t)) - r_\delta(t) m_\delta(t) \\ \dot{v_\delta}(t) = p_\delta(t)\left[\nabla f(\theta_\delta(t)\right]^2 - q_\delta(t) v_\delta(t), \end{cases} \tag{A.4}$$

where

$$h_\delta(t) = h(\max(\delta, t))$$

and similar formulas hold for r_δ, p_δ and q_δ. Those functions are continuous and locally Liptchitz in time, and defined for every $t > 0$ by the previous section. Note that for every $T > 0$, $\mathbf{x}_\delta \in C([0, T]; \mathbb{R}^n)$, and is C^1 everywhere outside $t = \delta$.

In order to show that this family of functions converges, we use Arzela-Ascoli Theorem. The next section is dedicated to proving that the hypotheses of Arzela-Ascoli are verified.

A.4.1. Equicontinuity and Uniform Boundedness

We prove in this section, that the family of functions $\mathbf{x}_\delta$ is equicontinuous and uniformly bounded, where $\mathbf{x}_\delta$ is the solution to (A.4). This allows us to apply Arzela-Ascoli at the end of this subsection in order to get the candidate for a solution of ODE (2.1).

The key result is the following proposition whose proof is left to subsection A.4.3

Proposition 18 *If assumptions 3 is satisfied, then there exists a positive constant $C_2(T)$, independent of δ, such that for all $t, s \in [0, T]$*

$$\|\mathbf{x}_\delta(t) - \mathbf{x}_\delta(s)\|^2 \leq C_2(T)(t - s)^2.$$

As a consequence of the previous Proposition, we can control the norm of the solution $\mathbf{x}_\delta$; this is done in terms of a special norm (instead of the usual one). More precisely, let us recall the notion of fractional Sobolev space. For a real number $0 < \delta < 1$ and $p \geq 1$, we denote by $W^{\alpha,p}([0,T])$ the fractional Sobolev space of functions $u \in L^p(0,T)$ satisfying

$$\int_0^T \int_0^T \frac{\|u(t) - u(s)\|^p}{|t-s|^{\delta p+1}} dsdt < +\infty$$

The space $C^\gamma([t_0,T];\mathbb{R}^{3d})$ is the space of Hölder continuous function of order $\gamma > 0$ on $[t_0,T]$ with values in $\mathbb{R}^{3d}$. It follows that

Lemma 19 *If assumptions 3 is satisfied, there exists a positive constant $C_3(T)$, independent of δ, such that*

$$\|\mathbf{x}_\delta\|^2_{W^{\gamma,2}} \leq C_3(T)$$

for any $\gamma < 1$.

Proof The proof is a direct consequence of Lemma 18. Indeed

$$\int_0^T \int_0^T \frac{\|\mathbf{x}_\delta(t) - \mathbf{x}_\delta(s)\|^2}{|t-s|^{2\gamma+1}} dsdt \leq C_2(T) \int_0^T \int_0^T \frac{(t-s)^2}{|t-s|^{2\gamma+1}} dsdt < +\infty$$

where the last inequality holds if and only if $\gamma < 1$. ■

We now use the Sobolev embedding $W^{\gamma,2}([0,T]) \hookrightarrow C^\alpha([0,T])$ for $\gamma - \alpha > 1/2$ and $\gamma < 1$, which implies $\alpha < 1/2$. It follows that the family $\mathbf{x}_\delta \in C^\alpha([0,T],\mathbb{R}^n)$. From Lemma 19, we conclude that the family is uniformly bounded. Finally, the family is equi-continuous because of the definition of the norm in C^α and its uniform bound in δ.

Applying Arzela Ascoli Theorem, we deduce that there exists a converging sub-sequence (still denoted $\mathbf{x}_\delta$) in $C([0,T],\mathbb{R}^n)$. We denote by $\widehat{\mathbf{x}}$ its limit and we prove in the next section that $\widehat{\mathbf{x}}$ satisfies Equation (2.1).

A.4.2. Identification of the Limit and Uniqueness of the Solution

Existence. The convergence of the initial conditions are a direct consequence of the uniform convergence (which implies point-wise convergence at every point). Now fix $T > 0$; it is clear that the ODE (A.4) converges uniformely to ODE (2.1) when $\delta \to 0$ in a neighbourhood of $t = T$ (indeed, for $\delta << T$, the two differential equations are equal). Since x_δ converges uniformly to $\hat{x}$, we conclude that $\hat{x}$ is a solution of of ODE (2.1) in a neighbourhood of $t = T$. Since $T > 0$ was arbitrary, we conclude the result.

Uniqueness. We proceed by contradiction. Assume there exist two solutions $\mathbf{x} = (\theta, m, v)$ and $\mathbf{y} = (\psi, n, w)$ to the system (2.1). An easy computation shows that for all $0 \leq t \leq T$ (because v and w are lower bounded, see Lemma 17)

$$\|\theta(t) - \psi(t)\| \leq \int_0^t \left\| \frac{m}{\sqrt{v+\varepsilon}} - \frac{n}{\sqrt{w+\varepsilon}} \right\| ds \leq C \int_0^t \|m-n\| + \|n\| \, \|w-v\| \, ds$$

By continuity of the solution of equation (2.1) on $[0, T]$, we know that there exists a constant $\widetilde{C}$ such that for all $s \leq t$, $\|n(s)\| \leq \widetilde{C}$ and therefore

$$\|\theta(t) - \psi(t)\| \leq C \int_0^t \|m - n\| + \widetilde{C} \|y - v\| \, ds. \tag{A.5}$$

Let us consider $a_\eta(t) = \exp\left(\int_\eta^t r(s)ds\right)$. Computing the time derivative of $m \cdot a_\eta(t)$ and integrating, we easily conclude that

$$m(t) = \frac{1}{a_\eta(t)} \left(m(\eta) + \int_\eta^t a_\eta(s) h(s) \nabla f(\theta) ds \right)$$

It follows from Assumptions 3 and inequality (A.5), that for all $\eta \leq t \leq T$,

$$\begin{aligned}
\|m(t) - n(t)\| &= \left\| \frac{1}{a_\eta(t)} \left(m(\eta) - n(\eta)\right) + \frac{1}{a_\eta(t)} \int_\eta^t a_\eta(s) h(s) \left(\nabla f(\theta) - \nabla f(\psi)\right) ds \right\| \\
&\leq \|m(\eta) - n(\eta)\| + C_1 \int_\eta^t h(s) \int_0^s \|m - n\| + \widetilde{C} \|v - w\| \, du ds \\
&\leq \|m(\eta) - n(\eta)\| + C_1 \left(\sup_{0 \leq u \leq t} \|m - n\| + \widetilde{C} \sup_{0 \leq u \leq t} \|v - w\| \right) \int_\eta^t s \cdot h(s) ds
\end{aligned}$$

By continuity of the process m and n, the fact that $m_0 = n_0$ and the continuity of $s \mapsto sh(s)$ on $[0, t]$, we obtain by taking the limit when η goes to zero that, apart from increasing C_1,

$$\|m(t) - n(t)\| \leq C_1 t \left(\sup_{0 \leq u \leq t} \|m - n\| + \widetilde{C} \sup_{0 \leq u \leq t} \|v - w\| \right).$$

Similarly we introduce $b_\eta(t) = \exp\left(\int_\eta^t q(s)ds\right)$. Then, we prove that there is a constant C_2 such that

$$\|v(t) - w(t)\| \leq C_2 t \left(\sup_{0 \leq u \leq t} \|m - n\| + \widetilde{C} \sup_{0 \leq u \leq t} \|v - w\| \right).$$

Hence, by combining all bounds, there exists two constants, still denoted C_1 and C_2, such that

$$\|m(t) - n(t)\| + \|v(t) - w(t)\| + \|\theta(t) - \psi(t)\| \leq C_1 t \sup_{0 < u \leq t} \|m - n\| + C_2 t \sup_{0 < u \leq t} \|v - w\|.$$

Since there exists a $t > 0$ such that $C_1 t$ and $C_2 t$ are strictly smaller than 1, this inequality yields a contradiction. We conclude that the solution must be unique.

A.4.3. Proof of Proposition 18

We start by a preliminary estimate, which extends Lemma 15 to an uniform bound in terms of δ:

Lemma 20 *Suppose that $p(t) \not\equiv 0$. There exists two constants K_1 and K_2 such that for all $t \in [0,T]$ and all $\delta > 0$ sufficiently small*

$$\left\| \frac{m_\delta(t)}{\sqrt{v_\delta(t)+\varepsilon}} \right\|^2 \leq K_1 \left\| \frac{m_0}{\sqrt{v_0+\varepsilon}} \right\|^2 + K_2.$$

Proof From Lemma 15 and assumption 3 (which implies that $\delta h_\delta(\delta)$, $\delta q_\delta(\delta)$, $\delta r_\delta(\delta)$ and $\frac{h_\delta(\delta)}{p_\delta(\delta)}$ are bounded for $\delta < 1$), there exits a constants $K_1 \geq 0$ and $K_2 \geq 0$ such that for every $\delta \leq 1$ and $t < \delta$, we have:

$$\begin{aligned}
\left\| \frac{m_\delta(t)}{\sqrt{v_\delta(t)+\varepsilon}} \right\|^2 &\leq e^{\int_0^\delta q_\delta(\delta)-2r_\delta(\delta)du} \left\| \frac{m_0}{\sqrt{v_0+\varepsilon}} \right\|^2 + d\int_0^\delta e^{\int_u^\delta q_\delta(\delta)-2r_\delta(\delta)da} \frac{h_\delta^2(\delta)}{p_\delta(\delta)} du \\
&= e^{\delta(q_\delta(\delta)-2r_\delta(\delta))} \left\| \frac{m_0}{\sqrt{v_0+\varepsilon}} \right\|^2 + d\frac{e^{\delta(q_\delta(\delta)-2r_\delta(\delta))}-1}{q_\delta(\delta)-2r_\delta(\delta)} \frac{h_\delta^2(\delta)}{p_\delta(\delta)} \qquad \text{(A.6)} \\
&\leq K_1 \left\| \frac{m_0}{\sqrt{v_0+\varepsilon}} \right\|^2 + K_2.
\end{aligned}$$

Moreover from Lemma 15 and assumption 3 (which implies that $q_\delta(u) - 2r_\delta(u) < 0$, $h_\delta(t)/p_\delta(u)$ and $h_\delta(t)/r_\delta(u)$ are bounded for δ and u small), there exits a constants $\widetilde{K}_1 \geq 0$ and $\widetilde{K}_2 \geq 0$ such that for every $\delta > 0$ small enough and $t \geq \delta$, we have:

$$\begin{aligned}
\left\| \frac{m_\delta(t)}{\sqrt{v_\delta(t)+\varepsilon}} \right\|^2 &\leq e^{\int_\delta^t q_\delta(u)-2r_\delta(u)du} \left\| \frac{m_\delta(\delta)}{\sqrt{v_\delta(\delta)+\varepsilon}} \right\|^2 + d\int_\delta^t e^{\int_u^t q_\delta(a)-2r_\delta(a)da} \frac{h_\delta^2(u)}{p_\delta(u)} du \\
&\leq \left\| \frac{m_\delta(\delta)}{\sqrt{v_\delta(\delta)+\varepsilon}} \right\|^2 + d \sup_{\delta<u<t} \frac{h_\delta(u)}{p_\delta(u)} \sup_{\delta<u<t} \left| \frac{h_\delta(u)}{q_\delta(u)-2r_\delta(u)} \right| \qquad \text{(A.7)} \\
&\leq \widetilde{K}_1 \left\| \frac{m_\delta(\delta)}{\sqrt{v_\delta(\delta)+\varepsilon}} \right\|^2 + \widetilde{K}_2,
\end{aligned}$$

We conclude combining the two inequalities. ■

We are now ready to prove Proposition 18. The proof uses the integral formulation and weighted space. First, we define the following norm for all $0 < t \leq T$

$$\begin{aligned}
N(t,\delta) = &\sup_{0<u\leq t} \|h_\delta(u)\nabla f(\theta_\delta(u)) - r_\delta(u)m_\delta(u)\| \\
&+ \sup_{0<u\leq t} \left\| p_\delta(u)\left[\nabla f(\theta_\delta(u))\right]^2 - q_\delta(u)v_\delta(u) \right\| + \sup_{0<u\leq t} \left\| \frac{m_\delta(u)}{\sqrt{v_\delta(u)+\varepsilon}} \right\|.
\end{aligned}$$

We claim that there exists a constant $C(T)$ (independent of δ) such that $N(t,\delta) \leq C(T)$ for all $t \in (0,T]$. Note that Proposition 18 immediately follows from the claim and the following inequality

$$\|\mathbf{x}_\delta(t) - \mathbf{x}_\delta(s)\|^2 \leq \left(\int_s^t \|\dot{\mathbf{x}}_\delta(u)\| du \right)^2 \leq N(T,\delta)^2 (t-s)^2.$$

We now turn to the proof of the claim.

The case $t \leq \delta$. For all $t \leq \delta$, the function r_δ is constant and the equation for m_δ, given by system (A.4), has the equivalent Duhamel formulation given by

$$m_\delta(t) = e^{-tr_\delta(\delta)}m_0 + e^{-tr_\delta(\delta)} \int_0^t e^{ur_\delta(\delta)} h_\delta(\delta) \nabla f(\theta_\delta(u)) du \tag{A.8}$$

A similar equation holds for v_δ. From Lemma 20, we know that $\left\| m_\delta(t)/\sqrt{v_\delta(t)+\varepsilon} \right\|^2$ is uniformly bounded with respect to δ. Moreover, $\|\theta_\delta(t)\|$ is uniformly bounded; indeed

$$\|\theta_\delta(t) - \theta_0\| \leq \int_0^t \left\| \frac{m_\delta(u)}{\sqrt{v_\delta(u)+\varepsilon}} \right\| du \leq t \left(K_1 \left\| \frac{m_0}{\sqrt{v_0+\varepsilon}} \right\| + K_2 \right). \tag{A.9}$$

Next, consider the first term which appears in $N(t,\delta)$. From the Duhamel formulation (A.8), the triangle inequality and that the initial condition $m_0 = \nabla f(\theta_0) \lim_{t\to 0^+} h(t)/r(t)$, we obtain an upper bound of the form

$$\|h_\delta(\delta)\nabla f(\theta_\delta(t)) - r_\delta(\delta) m_\delta(t)\| \leq N_1 + N_2 + N_3$$

where:

$$N_1 = \|h_\delta(\delta)\left(\nabla f(\theta_\delta(t)) - \nabla f(\theta_0)\right)\|, \qquad N_2 = \left\| r_\delta(\delta) e^{-tr_\delta(\delta)} \left(m_0 - \frac{h_\delta(\delta)}{r_\delta(\delta)} \nabla f(\theta_0) \right) \right\|$$
$$N_3 = \left\| r_\delta(\delta) e^{-tr_\delta(\delta)} \int_0^t e^{ur_\delta(\delta)} h_\delta(\delta) \left(\nabla f(\theta_\delta(u)) - \nabla f(\theta_0)\right) du \right\|$$

We now show that each one of these terms are bounded uniformly in terms of δ.

The term N_1 is bounded because f is C^2 (and therefore, the gradient is locally Lipschitz) and $\theta_\delta(t)$ is uniformly bounded by inequality (A.9); in particular, denote by L the Lipschitz constant of ∇f in the compact set containing all solutions $\theta_\delta(t)$ for bounded t. More precisely, by the Duhamel formula (A.8) and Lemma 20

$$N_1 \leq \delta h_\delta(\delta) L \left(K_1 \left\| \frac{m_0}{\sqrt{v_0+\varepsilon}} \right\|^2 + K_2 \right),$$

and we easily conclude that N_1 is uniformly bounded by assumption 3.

The term N_2 is bounded from the choice of the initial condition and the fact that $h(t)/r(t)$ is a C^1 function. More precisely

$$N_2 \leq \delta r_\delta(\delta) e^{-tr_\delta(\delta)} \|\nabla f(\theta_0)\| \left| \frac{h_\delta(\delta)}{r_\delta(\delta)} - \lim_{t\to 0} \frac{h(t)}{r(t)} \right| \delta^{-1}.$$

and we can easily conclude that N_2 is uniformly bounded by assumption 3.

The term N_3 is bounded in a similar way as N_1 using assumption 3, inequality (A.9) and Lemma 20. More precisely:

$$N_3 \leq \frac{\delta^2 r_\delta(\delta) h_\delta(\delta)}{2} L \left(K_1 \left\| \frac{m_0}{\sqrt{v_0+\varepsilon}} \right\|^2 + K_2 \right).$$

Gathering all bounds, we easily conclude that there exists a constant C_1 such that:

$$\sup_{0<u\leq t} \|h_\delta(u)\nabla f(\theta_\delta(u)) - r_\delta(u)m_\delta(u)\| \leq C_1, \quad \forall\, \delta \leq 1.$$

From a similar argument, we obtain that there exists a constant C_2 such that:

$$\sup_{0<u\leq t} \left\| p_\delta(u)\left[\nabla f(\theta_\delta(u))\right]^2 - q_\delta(u)v_\delta(u)\right\| \leq C_2 \quad \forall\, \delta \leq 1.$$

We conclude that there exists a constant C such that $N(t,\delta) < C$ for every $t \leq \delta$ and $\delta \leq 1$.

The case $t > \delta$. The proof uses the same arguments as in the case of $t \leq \delta$ using the appropriate integral formulation and Lemma 20. We omit the details here.

Appendix B. Proof of Theorem 4

From the fact that f is bounded from below, we conclude that the energy functional $E(t,\theta,m,v)$ introduced in (2.3) is bounded from below. Under assumption 2, this energy is non-increasing, and we conclude that there exists $E_\star \in \mathbb{R}$ such that:

$$\lim_{t\to\infty} E(t,\theta,m,v) = E_\star$$

Next, by direct computation we obtain that:

$$\frac{d}{dt}E(t,\theta,m,v) = -\frac{1}{h(t)}\left(r(t) + \frac{h'(t)}{2h(t)}\right)\left\|\frac{m}{[v+\varepsilon]^{1/4}}\right\|^2 + \sum_{i=1}^{d}\frac{m_i^2\{q(t)v_i - p(t)\cdot[\partial_{\theta_i}f(\theta)]^2\}}{4h(t)\cdot(v_i+\varepsilon)^{3/2}}, \tag{B.1}$$

and it follows from assumption 2, that:

$$\frac{d}{dt}E(t,\theta,m,v) \leq -\frac{p(t)}{4h(t)}\left\|\frac{\nabla f(\theta)}{(v+\epsilon)^{3/4}}\right\|^2$$

and, since the gradient of f is bounded, $\lim_{t\to\infty} p(t) \neq 0$ and $\lim_{t\to\infty} q(t) \neq 0$ by hypothesis, we conclude by Lemma 13 that v is bounded from above. Furthermore, from the fact that h is decreasing, we get that there exists a constant $C > 0$ such that:

$$\frac{d}{dt}E(t,\theta,m,v) \leq -C\,\|\nabla f(\theta)\|^2 \implies 0 \leq C\int_{t_0}^{\infty}\|\nabla f(\theta(s))\|^2\,ds \leq E(t_0,\theta_0,m_0,v_0) - E_*$$

It is therefore clear that $\liminf_{t\to\infty}\|\nabla f(\theta)\|^2 = 0$. We claim that the same is true for the limit sup, because of the hypothesis of ∇f being globally Lipschitz, which avoid sharp oscillations. Indeed, suppose by contradiction that $\limsup_{t\to\infty}\|\nabla f(\theta)\|^2 > K > 0$. It follows that there exists a sequence of points $(t_i)_{i\in\mathbb{N}}$ which diverges to infinity, and such that $\|\nabla f(\theta(t_i))\| \geq \sqrt{K}/2$. Now, by Lemma 2, we get that $\nabla f(\theta(t))$ is a globally Lipschitz for every $t \in [\tilde{t},\infty[$ and, without loss of generality, we assume $t_0 \geq \tilde{t}$. It follows that there exists $\delta > 0$ such that $\|\nabla f(\theta(t))\| \geq \sqrt{K}/4$ whenever $t \in B_\delta(t_i)$, for every $i \in \mathbb{N}$. Apart from shrinking $\delta > 0$, we can assume that all the balls $B_\delta(t_i)$ are disjoint, allowing us to conclude that:

$$\int_{t_0}^{\infty}\|\nabla f(\theta(s))\|^2\,ds \geq \sum_{i=1}^{\infty}\int_{B_\delta(t_i)}\|\nabla f(\theta(s))\|^2\,ds \geq \frac{K^2}{16}\sum_{i=1}^{\infty}2\delta = \infty$$

which contradicts the fact that the energy is bounded from below. We therefore conclude that $\limsup_{t\to\infty} \|\nabla f(\theta)\|^2 = 0$, finishing the proof.

Appendix C. Proof of Theorem 5

Consider the *autonomous* system associated to (2.1) (we recall that this means that we treat the time t as a variable with differential equation $\dot{t} = 1$), that is, the following vector field defined in $\mathbb{R}^{3d+1}$:

$$\begin{aligned}\partial = \partial_t - \sum_{i=1}^{d} \frac{m_i}{\sqrt{v_i+\varepsilon}}\partial_{\theta_i} + (h(t)\partial_{\theta_i} f(\theta) - r(t)m_i)\partial_{m_i} \\ + (p(t)\left[\partial_{\theta_i} f(\theta)\right]^2 - q(t)v_i)\partial_{v_i}.\end{aligned}$$

which is well-defined for every $(\theta, m, v, t) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d_{\geq 0} \times \mathbb{R}_{>0}$ (because we assume that $\varepsilon > 0$, c.f. assumption 5). In order to study the convergence of the vector field when $t \to \infty$, we perform the change of coordinates $s = 1/t$ (in this case, studying the behaviour at $t \to \infty$ is replaced to studying the behaviour when $s \to 0$), which yields:

$$\begin{aligned}\partial = -s^2\partial_s + \sum_{i=1}^{d} - \frac{m_i}{\sqrt{v_i+\varepsilon}}\partial_{\theta_i} + (H(s)\partial_{\theta_i} f(\theta) - R(s)m_i)\partial_{m_i} \\ + (P(s)\left[\partial_{\theta_i} f(\theta)\right]^2 - Q(s)v_i)\partial_{v_i}.\end{aligned} \tag{C.1}$$

and note that the above vector-field is kept autonomous. We recall that the time of the associated differential equation is now denoted by τ (that is, a solution of this vector field is a curve $\mathbf{y}(\tau) = (\theta(\tau), m(\tau), v(\tau), s(\tau))$ such that $\dot{\mathbf{y}}(\tau) = \partial(\mathbf{y}(\tau))$). We now fix an orbit

$$\mathbf{y}(\tau) = (\theta(\tau), m(\tau), v(\tau), s(\tau))$$

with initial conditions $\mathbf{y}(\tau_0) = (\theta(\tau_0), m(\tau_0), v(\tau_0), 1/\tau_0)$. By the Lemma 13, we know that $\mathbf{y}(\tau)$ is bounded and $v(\tau) > 0$ for all $\tau \in [\tau_0, \infty)$. Denote by $\omega(\mathbf{y}(\tau))$ the topological limit of $\mathbf{y}(\tau)$. By assumption 4 we know that $\theta(\tau)$ is bounded, and by Lemma 13 we conclude that $m(\tau)$ and $v(\tau)$ are also bounded. It follows that $\omega(\mathbf{y}(\tau))$ is non-empty, and that it is the union of orbits of ∂. Furthermore, from the expression $s = 1/t$ (and the fact that the solutions are defined for all $\tau \in [\tau_0, \infty)$, which implies that t takes all values in $[\tau_0, \infty)$) we know that $\omega(\mathbf{y}(\tau)) \subset (s = 0)$.

We now consider the energy functional E given in (2.3) but in this new coordinate system. More precisely, consider the functional:

$$\widetilde{E}(\mathbf{y}) = f(\theta) + \frac{1}{2H(s)}\left\|\frac{m}{[v+\varepsilon]^{1/4}}\right\|^2,$$

which by assumption 5 is everywhere well-defined (because $H(0) > 0$). It follows from direct computation that:

$$\frac{d}{d\tau}\widetilde{E}(\mathbf{y}) \leq -\frac{1}{2H(s)}\left[2R(s) - \frac{Q(s)}{2} - s^2\frac{H'(s)}{H(s)}\right]\left\|\frac{m}{[v+\varepsilon]^{1/4}}\right\|^2.$$

which is everywhere non-positive by assumption 2 and 5. Now, since $E(\mathbf{x}(\tau))$ is bounded from below (because E is continuous and $\mathbf{y}(\tau)$ is bounded), we conclude that the limit:

$$\lim_{\tau\to\infty} \widetilde{E}(\mathbf{y}(\tau)) = \widetilde{E}_\infty$$

exists. In particular, it follows that $\omega(\mathbf{y}(\tau)) \subset (\widetilde{E}(\mathbf{y}) = \widetilde{E}_\infty)$. This implies that $\omega(\mathbf{y}(\tau))$ must be contained in the set of zero derivative of $\widetilde{E}(\mathbf{y})$. By assumption 5 this implies that $\omega(\mathbf{y}(\tau)) \subset (m = 0)$. Now note that:

$$\partial \cdot m_i = H(s)\partial_{\theta_i} f(\theta) - R(s)m_i,$$

and since $H(0) \neq 0$ (assumption 5), we conclude that $\omega(\mathbf{y}(\tau)) \subset (\nabla f(\theta) = 0)$. Finally, if either $P(s) \equiv Q(s) \equiv 0$ or $Q(0) > 0$, by the expression:

$$\partial \cdot v_i = P(s)\partial_{\theta_i} f(\theta)^2 - Q(s)v_i,$$

we conclude that $\omega(\mathbf{x}(\tau)) \subset (v = 0)$. We conclude easily.

Appendix D. Proof of Theorem 6

The proof of the Theorem relies on central-stable manifold theory (see S. Chow and Wang (2009) for an introduction), that is, in the following result, which is a local version of (S. Chow and Wang, 2009, Ch. 1 Thm 4.2), by using the cut-off technique given in (S. Chow and Wang, 2009, Ch. 1 Lem. 3.1); c.f. (S. Chow and Wang, 2009, Ch. 1, Thm 1.1 and 3.2).

Theorem 21 *Consider the differential equation $\dot{x} = Ax + F(x)$ defined over $\mathbb{R}^n$, where A is a matrix which contains at least one positive eigenvalue, and $F(x)$ is a C^k function, for some $k \geq 1$, such that $F(0) = 0$ and $DF(0) = 0$. Then there exists a neighbourhood U of 0 and a C^k sub-manifold Σ (the center-stable manifold) such that:* (1) *the manifold Σ is invariant by the differential equation everywhere over U;* (2) *the manifold Σ contains the origin 0 and has dimension at most $n-1$; and* (3) *if $\mathbf{x}_0 \in U \setminus \Sigma$, then there exists $\widetilde{t_0} > t_0$ such that $\mathbf{x}(\widetilde{t_0}) \notin U$, where $\mathbf{x}(t)$ denotes the solution of the differential equation with initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$.*

Now, recall the vector field ∂ defined in (C.1), which describes the ODE (2.1). Consider the set:

$$B = \{\theta_\star \in \mathbb{R}^d;\ \nabla f(\theta_\star) = 0, \text{ and } \theta_\star \text{ is not a local minimum of } f\}$$

By assumption 6(b) the set B is discrete and, therefore, a countable union of isolated points of $\mathbb{R}^d$. It follows from Theorem 5 that the set:

$$C := \{\mathbf{y} = (\theta, m, v, s);\ \theta \in B \text{ and } \mathbf{y} \text{ is a singularity of } \partial\}$$

is a countable union of isolated points, all of each have the form $\mathbf{y}_* = (\theta_\star, 0, 0, 0)$, where $\theta_\star \in B$. We now consider the set:

$$S := \{\mathbf{y}_0 = (\theta_0, m_0, v_0, s_0);\ \omega(\mathbf{y}(\tau)) \cap C \neq \emptyset, \text{ where } \mathbf{y}(\tau_0) = \mathbf{y}_0\}$$

It follows that $\omega(\mathbf{y}(\tau))$ is a connected set, so that:

$$S := \{\mathbf{y}_0 = (\theta_0, m_0, v_0, s_0);\ \omega(\mathbf{y}(\tau)) \subset C, \text{ where } \mathbf{y}(\tau_0) = \mathbf{y}_0\}$$

We now make a local argument valid for each singular point in C in order to show that S is locally a sub-manifold; indeed, fix $\mathbf{y}_* \in S$. Consider the linearization of ∂ at the singular point $\mathbf{y}_* = (\theta_\star, 0, 0, 0)$, which is the $3d+1$ square matrix:

$$Jac(\partial)(\mathbf{y}_*) = \begin{bmatrix} 0 & -\varepsilon^{-1/2} Id & 0 & 0 \\ H(0)\mathcal{H}_f(\theta_\star) & -R(0)Id & 0 & 0 \\ 0 & 0 & -Q(0)Id & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

where Id denotes the Identity of a d-square matrix, and $\mathcal{H}_f(\theta_\star)$ is the Hessian of f at $\theta_\star$. It follows from direct computation that the eigenvalues λ of this matrix are: 0 with order 1, $-Q(0)$ with order d and the solutions of the quadratic equations:

$$\eta_i = -\frac{\varepsilon^{1/2}}{H(0)}(R(0) + \lambda)\lambda, \quad i = 1, \ldots, d \tag{D.1}$$

where $\{\eta_1, \ldots, \eta_d\}$ are the eigenvalues of $\mathcal{H}_f(\theta_\star)$. By assumption 6, we can suppose without loss of generality that $\eta_1 < 0$, and we easily conclude by equation (D.1) that there exists one strictly positive eigenvalue λ of $Jac(\partial)(\mathbf{y}_*)$. By Theorem 21, there exists an open neighbourhood $U_{\mathbf{y}_*}$ of $\mathbf{y}_*$ and a C^1 manifold $\Sigma_{\mathbf{y}_*} \subset U_{\mathbf{y}_*}$ such that every orbit $\mathbf{y}(\tau)$ with initial condition in $U_{\mathbf{y}_*} \setminus \Sigma_{\mathbf{y}_*}$, leaves $U_{\mathbf{y}_*}$ in finite time. Note that the Lebesgue measure of $\Sigma_{\mathbf{y}_*}$ is zero. Consider the set Σ given by the union of all orbits with initial conditions in $\Sigma_{\mathbf{y}_\star}$, for every $\mathbf{y}_\star \in C$. Since C is a countable set, we conclude that the Lebesgue measure of Σ must be zero (since each $\Sigma_{\mathbf{y}_*}$ has Lebesgue measure zero). Now, since $\mathbf{y}_*$ is an isolated singularity of ∂ and the ω-limit of an arbitrary orbit $\mathbf{y}(\tau)$ with initial condition in S is connected, we conclude that if $\omega(\mathbf{y}(\tau)) = \mathbf{y}_*$, then $\mathbf{y}(\tau) \subset \Sigma_{\mathbf{y}_\star}$ for $\tau >> \tau_0$. It easily follows that $S \subset \Sigma$, and we conclude that S has measure zero.

Finally, let $t_0 > 0$ be fixed and denote by $S_{t_0} = S \cap \{s = 1/t_0\}$. Now, S has volume zero and contains orbits of ∂, all of each are transverse to the set $\{s = 1/t_0\}$. It follows that the volume of $S_{t_0} \subset \mathbb{R}^{3d}$ is zero by transversality, and we conclude easily.

Remark 22 (On Assumption 6) *Assumption 6(a) was introduced in Lee et al. (2019) and has crucial technical consequences. It allows us to use the center-stable manifold theory recalled above. Without this hypothesis, the singular points of the ODE* (2.1) *at infinite (see equation* (C.1)*) can be arbitrarily degenerated, and there is no general singularity theory to treat these points in dimension higher than three. In order to relax such a hypothesis, it is necessary to develop specific singularity techniques for equation* (2.1)*, and we intend to pursue this direction in a future paper.*

Assumption 6(b) allows us to exclude pathological differences between local and global center-stable manifold theory. An alternative to this hypothesis is to add a globally Lipschitz assumption onto the system (2.1)*, and to study the relationship between the Lipschitz approximation and the Hessian of the loss function f (which would allow us to use the strong global result (S. Chow and Wang, 2009, Ch. 1 Thm 1.1)). This has been done, for example,* *in Panageas and Piliouras (2017) where the authors study the analog problem for a simpler ODE. We understand that a study in the generality of ODE* (2.1) *without condition 6*(*b*) *would demand the development of specific singularity techniques for equation* (2.1)*, and we intend to pursue this direction in a mathematical paper.*

Appendix E. Proof of Theorem 7

Let $\theta_\star$ be a minimum point of f (which exists by the convexity Assumption 7). We start by computing the derivative of the energy functional (3.2), so that we can find conditions on the functions $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$, as well as the coefficients h, p, q, r, so that $\frac{d}{dt}\mathcal{E}$ is bounded (the conditions must also guarantee that $\mathcal{E}$ is positive). In order to simplify the notation, we denote by:

$$\begin{aligned}
\mathcal{E}_1(t,m,v,\theta) &= \mathcal{A}(t)\left(f(\theta)-f(\theta_\star)\right)\\
\mathcal{E}_2(t,m,v,\theta) &= \frac{1}{2}\left\|[v+\varepsilon]^{1/4}(\theta-\theta_\star)\right\|^2 - \mathcal{B}(t)\left\langle \theta-\theta_\star, m\right\rangle + \frac{\mathcal{C}(t)}{2}\left\|\frac{m}{[v+\varepsilon]^{1/4}}\right\|^2.
\end{aligned}$$

From the convexity assumption on the objective function f, we get

$$\frac{d}{dt}\mathcal{E}_1(t,\theta) \le \mathcal{A}'(t)\left\langle \nabla f(\theta), \theta-\theta_\star\right\rangle - \mathcal{A}(t)\left\langle \nabla f(\theta), \frac{m}{[v+\varepsilon]^{1/2}}\right\rangle \tag{E.1}$$

Next, we derive each term of $\mathcal{E}_2$.

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\left\|[v+\varepsilon]^{1/4}(\theta-\theta_\star)\right\|^2 = &-\left\langle m, \theta-\theta_\star\right\rangle - \frac{q(t)}{4}\left\langle \frac{v}{[v+\varepsilon]^{1/2}}(\theta-\theta_\star), \theta-\theta_\star\right\rangle\\
&+\frac{p(t)}{4}\left\langle \frac{[\nabla f(\theta)]^2}{[v+\varepsilon]^{1/2}}(\theta-\theta_\star), \theta-\theta_\star\right\rangle
\end{aligned}$$

$$\begin{aligned}
\frac{d}{dt}\mathcal{B}(t)\left\langle \theta-\theta_\star, m\right\rangle = &-\mathcal{B}(t)\left\|\frac{m}{[v+\varepsilon]^{1/4}}\right\|^2 + \mathcal{B}(t)h(t)\left\langle \nabla f(\theta), \theta-\theta_\star\right\rangle\\
&+\left(\mathcal{B}'(t)-\mathcal{B}(t)r(t)\right)\left\langle \theta-\theta_\star, m\right\rangle
\end{aligned}$$

$$\begin{aligned}
\frac{d}{dt}\frac{\mathcal{C}(t)}{2}\left\|\frac{m}{[v+\varepsilon]^{1/4}}\right\|^2 =&h(t)\mathcal{C}(t)\left\langle \nabla f(\theta), \frac{m}{[v+\varepsilon]^{1/2}}\right\rangle + \left(-r(t)\mathcal{C}(t)+\mathcal{C}'(t)/2\right)\left\|\frac{m}{[v+\varepsilon]^{1/4}}\right\|^2\\
&+\frac{\mathcal{C}(t)q(t)}{4}\left\|\frac{m\,[v]^{1/2}}{[v+\varepsilon]^{3/4}}\right\|^2 - \frac{\mathcal{C}(t)p(t)}{4}\left\|\frac{\nabla f(\theta)m}{[v+\varepsilon]^{3/4}}\right\|^2
\end{aligned}$$

By adding all of the above computations, we get that $\mathcal{E}_1(t,\theta)$ and $\mathcal{E}_2(t,m,v,\theta)$ are positive functions such that:

$$\frac{d}{dt}\mathcal{E}(t,m,v,\theta) \le \frac{p(t)}{4}\left\langle \frac{[\nabla f(\theta)]^2}{[v+\varepsilon]^{1/2}}(\theta-\theta_\star), \theta-\theta_\star\right\rangle,$$

provided that all the following sufficient conditions are satisfied

$$\mathcal{A}(t) \geq 0, \quad \mathcal{A}'(t) \geq 0, \tag{E.2}$$

$$\mathcal{A}'(t) = h(t)\mathcal{B}(t) \tag{E.3}$$

$$\mathcal{A}(t) = h(t)\mathcal{C}(t) \tag{E.4}$$

$$\mathcal{B}'(t) - \mathcal{B}(t)r(t) = -1 \tag{E.5}$$

$$\mathcal{B}(t) \leq \frac{\mathcal{C}(t)}{3}\left(2r(t) - \frac{q(t)}{2} + \frac{h'(t)}{h(t)}\right) \tag{E.6}$$

$$\mathcal{B}^2(t) \leq \mathcal{C}(t). \tag{E.7}$$

It is now easy to see that equations (E.3), (E.4) and (E.5) are equivalent to the choices of $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ chosen in Subsection 3.5, and that assumption 8 implies inequalities (E.6) and (E.7). It is now immediate from the Fundamental Theorem of calculus (and the fact that $\mathcal{E}_2 \geq 0$) that:

$$f(\theta) - f(\theta_\star) \leq \frac{\mathcal{E}(t_0, m_0, v_0, \theta_0)}{\mathcal{A}(t)} + \frac{\int_{t_0}^t p(u)\left\langle \frac{[\nabla f(\theta)]^2}{[v+\varepsilon]^{1/2}}, [\theta - \theta_\star]^2 \right\rangle du}{4\mathcal{A}(t)}.$$

which proves the first part of the Theorem. Next, under assumption 4, Lemma 13 implies that there exists a finite constant:

$$\mathcal{K} = \sup_{t \in \mathbb{R}_+} \left\| [v + \varepsilon]^{1/4} (\theta - \theta_\star) \right\|_\infty^2,$$

and we note that:

$$p(t)\left\langle \frac{[\nabla f(\theta)]^2}{[v+\varepsilon]^{1/2}}, [\theta - \theta_\star]^2 \right\rangle \leq \mathcal{K}p(t)\left\| \frac{\nabla f(\theta)}{\sqrt{v+\varepsilon}} \right\|^2 \leq \mathcal{K}p(t)\left\| \frac{\nabla f(\theta)}{\sqrt{v}} \right\|^2.$$

Now, from the expression of ODE (2.1) in v we get:

$$\frac{d}{dt}\ln(v) + q(t) = p(t)\frac{[\nabla f(\theta)]^2}{v}$$

which implies that:

$$p(t)\left\langle \frac{[\nabla f(\theta)]^2}{[v+\varepsilon]^{1/2}}, [\theta - \theta_\star]^2 \right\rangle \leq \mathcal{K}\left(d \cdot q(t) + \sum_{i=1}^d \frac{d}{dt}\ln(v_i)\right).$$

and it follows that:

$$\int_{t_0}^t p(u)\left\langle \frac{[\nabla f(\theta)]^2}{[v+\varepsilon]^{1/2}}, [\theta - \theta_\star]^2 \right\rangle du \leq \int_{t_0}^t \mathcal{K}\left(d \cdot q(t) + \sum_{i=1}^d \frac{d}{dt}\ln(v_i)\right) du$$

The second inequality now easily follows from the fact that $v(t)$ is bounded by Lemma 13.

E.1. A Refined Bound

We may also consider the slightly more general energy functional:

$$\mathcal{E}_2(t,m,v,\theta) = \frac{\mathcal{D}(t)}{2}\left\| [v+\varepsilon]^{1/4}(\theta-\theta_\star) \right\|^2 - \mathcal{B}(t)\langle \theta-\theta_\star, m\rangle + \frac{\mathcal{C}(t)}{2}\left\| \frac{m}{[v+\varepsilon]^{1/4}} \right\|^2,$$

where $\mathcal{D}(t)$ is a positive function. If we assume that $\mathcal{D}(t)$ is bounded, we are able to follow the same reasoning of the previous section. In this case, we need to add the sufficient condition $\mathcal{D}(t)' \leq 0$, and equality (E.5) and inequality (E.7) are now given by:

$$\mathcal{B}'(t) - \mathcal{B}(t)r(t) = -\mathcal{D}(t) \tag{E.8}$$

$$\mathcal{B}^2(t) \leq \mathcal{D}(t)\mathcal{C}(t) \tag{E.9}$$

In particular, this implies that:

$$\mathcal{B}(t) = e^{\int_{t_0}^t r(s)ds}\int_t^\infty \mathcal{D}(s)e^{-\int_{t_0}^s r(u)du}ds$$

while the equations for $\mathcal{A}(t)$ and $\mathcal{C}(t)$ are unchanged. Since $\mathcal{D}(t)$ has a negative derivative, in general, this computation can not lead to a stronger convergence rate than the one obtained in the previous section. Nevertheless, it does allow one to obtain convergence rates for parameters that are inaccessible in the previous section.

Appendix F. Proof of the Corollaries in Section 4

F.1. Proof of Corollary 8

The proof of *(0)* and *(I)* directly follows from Theorems 4 and 5 provided that assumptions 2 and 5 are satisfied. Hence, the proof simply consists of checking the validity of both assumptions under the condition that $3+\beta_2 > 4\beta_1$. Let us recall that the coefficients for the Adam's differential equations are given by

$$h \equiv r \equiv g_1^A(t,\lambda,\alpha_1,\alpha_2) = \frac{1-e^{-\lambda/\alpha_1}}{\lambda\left(1-e^{-t/\alpha_1}\right)}, \qquad p \equiv q \equiv g_2^A(t,\lambda,\alpha_1,\alpha_2 = \frac{1-e^{-\lambda/\alpha_2}}{\lambda\left(1-e^{-t/\alpha_2}\right)}),$$

and $(\lambda,\alpha_1,\alpha_2)$ are positive real numbers. It is easy to check that assumptions 2 and 5 are satisfied if there exists a t, large enough, such that

$$\frac{4(1-e^{-\lambda/\alpha_1})}{\lambda(1-e^{-t/\alpha_1})} - \frac{1-e^{-\lambda/\alpha_2}}{\lambda(1-e^{-t/\alpha_2})} > 0$$

Taking the limit as t goes to infinity in the above inequality gives

$$4(1-e^{-\lambda/\alpha_1}) > 1-e^{-\lambda/\alpha_2},$$

and we conclude by using the expressions of β_1 and β_2. Next, the proof of *(II)* follows directly from Theorem 6 since assumptions 2 and 5 are satisfied under the condition $3+\beta_2 > 4\beta_1$.

Finally, in order to prove *(III)*, let us check the hypotheses of Theorem 7. We compute explicitly the functions

$$\begin{aligned}\mathcal{A}(t) &= \frac{1-e^{-\lambda/\alpha_i}}{\lambda}\int_{t_0}^{t}\frac{e^{s/\alpha_1}}{e^{s/\alpha_1}-1}\mathcal{B}(s)ds\\ \mathcal{B}(t) &= (e^{t/\alpha_1}-1)\int_{t}^{\infty}\frac{1}{e^{s/\alpha_1}-1}ds\\ \mathcal{C}(t) &= \frac{e^{t/\alpha_1}-1}{e^{t/\alpha_1}}\int_{t_0}^{t}\frac{e^{s/\alpha_1}}{e^{s/\alpha_1}-1}\mathcal{B}(s)ds\end{aligned}$$

so, by direct computation via L'Hôpital's rule:

$$\lim_{t\to\infty}\mathcal{A}(t)/t = \alpha_1\frac{1-e^{-\lambda/\alpha_1}}{\lambda}\quad \lim_{t\to\infty}\mathcal{B}(t)=\alpha_1\quad \lim_{t\to\infty}\mathcal{C}(t)/t=\alpha_1$$

and it easily follows that assumption 8 is verified. Finally, by using L'Hôpital's rule, we get:

$$\lim_{t\to\infty}\int_{t_0}^{t}q(s)ds/\mathcal{A}(t) = \alpha_1^{-1}\frac{1-e^{-\lambda/\alpha_2}}{1-e^{-\lambda/\alpha_1}}$$

which yields the result.

F.2. Sketch of the Proof of Corollary 9

By the choice of functions $h(t)$, $r(t)$, $p(t)$ and $q(t)$, it is easy to see that Assumptions 2 and 5 are always verified. Part (I) is, therefore, direct consequences from Theorem 5. Next, the computation of $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are independent on $p(t)$ and $q(t)$, so they are analogous to the one's obtained for Adam. It follows that assumption 8 is verified. Finally, by Theorem 7, the rate of convergence is controlled by the asymptotic behaviour of $\int_{t_0}^{t} q(s)ds/\mathcal{A}(t) = \ln(t/t_0)/\mathcal{A}(t)$,, which can easily be verified to be of order $\mathcal{O}(\ln(t)/t)$.

F.3. Sketch of the Proof of Corollary 10

By the choice of functions $h(t)$, $r(t)$, $p(t)$ and $q(t)$, it is easy to see that Assumptions 2 and 5 are always verified. Part (I) and (II) are, therefore, direct consequences from Theorems 5 and 6. Next, by direct computation, we get:

$$\mathcal{A}(t)=\gamma(t-t_0)\quad \mathcal{B}(t)=\gamma\quad \mathcal{C}(t)=\gamma(t-t_0).$$

It follows that assumption 8 is verified. Finally, by Theorem 7, the rate of convergence is controlled by the asymptotic behaviour of $1/\mathcal{A}(t)$, which is of order $\mathcal{O}(\ln(t)/t)$.

F.4. Proof of Corollary 11

First, assume that $r \geq 3$. From direct computation, we get:

$$\mathcal{A}(t)=(t^2-t_0^2)/2(r-1)\quad \mathcal{B}(t)=t/(r-1)\quad \mathcal{C}(t)=(t^2-t_0^2)/2(r-1)$$

It easily follows that, whenever $r \geq 3$, the inequalities of assumption 8 are verified, and the result follows from Theorem 7.

Next, assume that $r < 3$. Let $t_0 = 1$ and $\mathcal{D}(t) = t^{-\alpha}$ for some positive α which satisfies $2 > \alpha > 1 - r$. Following Section E.1, we get:

$$\mathcal{B}(t) = \frac{t^{1-\alpha}}{r + \alpha - 1}, \quad \mathcal{A}(t) = \mathcal{C}(t) = \frac{t^{2-\alpha} - 1}{(2-\alpha)(r+\alpha-1)}$$

Therefore, from inequality (E.6) we get:

$$\frac{t^{1-\alpha}}{r+\alpha-1} \leq \frac{t^{2-\alpha} - 1}{(2-\alpha)(r+\alpha-1)} \left(\frac{2r}{3t}\right) \iff 2 - \frac{2r}{3} \leq \alpha$$

while from (E.9) we obtain:

$$\frac{t^{2-2\alpha}}{(r+\alpha-1)^2} \leq \frac{t^{2-2\alpha} - 1}{(2-\alpha)(r+\alpha-1)} \iff 1 - \frac{r}{2} \leq \alpha$$

In other words, it is enough to consider $\alpha = 2 - 2r/3$ for every $0 < r < 3$. This implies that $f(\theta(t)) \to f_\star$ with rate of convergence $o(1/\mathcal{A}(t)) = o(1/t^{2r/3})$ as we wanted to prove.

References

F. Alvarez. On the minimizing property of a second order dissipative system in Hilbert spaces. *SIAM Journal on Control and Optimization*, 38(4):1102–1119, 2000.

F. Alvarez, H. Attouch, J. Bolte, and P. Redont. A second-order gradient-like dissipative dynamical system with Hessian-driven damping: Application to optimization and mechanics. *Journal de mathématiques pures et appliquées*, 81(8):747–779, 2002.

H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Math. Program.*, 168(1-2):123–175, March 2018.

H. Attouch, Z. Chbani, and H. Riahi. Rate of convergence of the Nesterov accelerated gradient method in the subcritical case $\alpha \leq 3$. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:2, 2019.

A. Belotto da Silva and M. Gazeau. A general system of differential equations to model first order adaptive algorithms. *arXiv e-prints*, art. arXiv:1810.13108, October 2018.

M. Betancourt, M. I. Jordan, and A. C. Wilson. On Symplectic Optimization. *ArXiv e-prints*, art. arXiv:1802.03653, February 2018.

S. Bubeck, Y. T. Lee, and M. Singh. A geometric alternative to nesterov's accelerated gradient descent. *CoRR*, abs/1506.08187, 2015.

A. Cabot. Asymptotics for a gradient system with memory term. *Proceedings of the American Mathematical Society*, 137(9):3013–3024, 2009.

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(11):5983–6017, 2009.

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(11):5983–6017, 2009.

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.

Soham De, Anirbit Mukherjee, and Enayat Ullah. Convergence guarantees for RMSProp and ADAM in non-convex optimization and an empirical comparison to Nesterov acceleration. *arXiv e-prints*, art. arXiv:1807.06766, July 2018.

J. Duchi and Y. Singer. Proximal and first-order methods for convex optimization. January 2013. URL `https://cs.stanford.edu/~ppasupat/a9online/uploads/proximal_notes.pdf`.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, July 2011.

C. S. Duris and J. N. Lyness. Compound quadrature rules for the product of two functions. *SIAM Journal on Numerical Analysis*, 12(5):681–697, 1975.

S. Gadat and F. Panloup. Long time behaviour and stationary regime of memory gradient diffusions. *Ann. Inst. H. Poincar Probab. Statist.*, 50(2):564–601, 05 2014.

S. Gadat, F. Panloup, and S. Saadane. Stochastic Heavy Ball. *Electron. J. Statist.*, 12(1): 461–529, 2018.

E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson. Global convergence of the Heavy-ball method for convex optimization. *2015 European Control Conference (ECC)*, 2015.

K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1462–1471, Lille, France, 07–09 Jul 2015. PMLR.

B. Hu and L. Lessard. Dissipativity Theory for Nesterov's Accelerated Method. *ICML'17: Proceedings of the 34th International Conference on Machine Learning*, 70.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

P. Kythe and P. Puri. *Computational Methods for Linear Integral Equations.* Birkhäuser Boston, 2002. ISBN 9780817641924.

J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient Descent Converges to Minimizers. *29th Annual Conference on Learning Theory*, PMLR 49:1246-1257, 2016.

J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht. First-order Methods Almost Always Avoid Saddle Points. *Math. Program.*, 179:311337, 2019.

L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Y. Nesterov. *Introductory Lectures on Convex Optimization: a Basic Course*. Kluwer Academic Publishers, 2004.

I. Panageas and G. Piliouras. Gradient Descent Only Converges to Minimizers: Non-Isolated Critical Points and Invariant Regions. *ITCS*, 2017.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

C. Li S. Chow and D. Wang. *Normal Forms and Bifurcation of Planar Vector Fields*. Cambridge University Press, 2009.

S. Kale S. Reddi and S. Kumar. Amsgrad, on the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

W. Su, S. Boyd, and E. J. Candes. A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights. *Journal of Machine Learning Research*, 17(153), 2016.

T. Tieleman and G. Hinton. Lecture 6.5RmsProp: Divide the gradient by a running average of its recent magnitude, 2012.

R Ward, X Wu, and L Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97, pages 6677–6686, 2019.

A. Wibisono and A. C. Wilson. On Accelerated Methods in Optimization. *arXiv e-prints*, art. arXiv:1509.03616, September 2015.

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(47): E7351–E7358, 2016.

M. Zaheer, S. Reddi, D. Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9815 – 9825, 2018.

M. D. Zeiler. Adadelta: An adaptive learning rate method. *ECCV*, 2013.

Z. Zhang, L. Ma, Z. Li, and C. Wu. Normalized Direction-preserving Adam. *arXiv e-prints*, art. arXiv:1709.04546, September 2017.