A GENERAL SYSTEM OF DIFFERENTIAL EQUATIONS TO MODEL FIRST ORDER ADAPTIVE ALGORITHMS.

ANDRÉ BELOTTO DA SILVA[*†] AND MAXIME GAZEAU[*‡]

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 prove global well-posedness of the system and we investigate the numerical time convergence of its forward Euler approximation. We study, furthermore, 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 and deterministic loss function, we introduce a suitable Lyapunov functional which allow 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.

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

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

Gradient descent [15], which only depends on the partial derivatives of f, is the simplest discrete algorithm to address the optimization problem above

$$\theta_{k+1} = \theta_k - s\nabla f(\theta_k). \tag{1.1}$$

Date: November 1, 2018.

[*] Alphabetical order and equal contribution.

[†] Université Aix-Marseille, Institut de Mathématiques de Marseille, UMR CNRS 7373, Centre de Mathématiques et Informatique, 39, rue F. Joliot Curie, 13013 Marseille, France (andre-ricardo.belotto-da-silva@univ-amu.fr).

[‡] Borealis AI (maxime.gazeau@borealisai.com).

Another popular iterative approach to solve the above smooth optimization is the *proximal* point algorithm [42, 44]

$$\theta_{k+1} = \operatorname{argmin}_u \left(\frac{1}{2s} \|u - \theta_k\|^2 + f(u) \right) \tag{1.2}$$

These discrete methods can be studied solely from the standpoint of optimization performance. It can be proved that both algorithms converge to a critical point ($\nabla f(\theta_k) \to 0$ as $k \to \infty$) [39] but also almost surely to a local minimizer [32, 33]. For convex functionals with globally Lipschitz gradient, both algorithms converges at linear rate $f(\theta_k) - f(\theta_\star) = \mathcal{O}(1/(sk))$, where $\theta_\star$ is a minimal point of f [39, 42, 44]. These results give important guarantees on the convergence of each method.

For small and constant learning rate s, gradient descent (1.1) (resp. *proximal* point algorithm (1.2)) corresponds to the *forward* (resp. *backward*) Euler's discretization of the gradient flow system

$$\dot{\theta}(t) = -\nabla f(\theta(t)), \qquad \theta_0 = \theta(0), \tag{1.3}$$

under the time scaling $t = ks$ [24, 47]. The stable equilibria of this continuous system are given by the the set of strict (local) minima of the loss function f and if the level sets of f are bounded (f coercive for example), then its trajectories asymptotically converge to a critical point in the sense that $\nabla f(\theta(t)) \to 0$ as $t \to 0$. Moreover for convex functions, a linear rate of convergence $f(\theta(t)) - f(\theta_\star) = \mathcal{O}(1/t)$ holds, which is analogue to those obtained for both gradient descent and proximal point algorithm.

The study of the continuous dynamical system is very useful. The well-behaved convergence properties of the gradient flow (1.3) allows an important number of choices to design an optimization algorithm [47]. It, furthermore, provides valuable intuition to prove convergence of discrete systems: for example, continuous Lyapunov functional can be often adapted to the discrete counterparts.

For large scale machine learning, a stochastic version of gradient descent (SGD) is very popular in practice [10, 27, 43]

$$\theta_{k+1} = \theta_k - s \nabla f(\theta_k, \xi_k), \tag{1.4}$$

where $\nabla f(\theta_k, \xi_k)$ is an unbiased estimator of the true gradient $\nabla f(\theta_k)$. It is well known that this algorithm does not always converge and theoretical analysis provide conditions that guarantee convergence to a critical point [10, 37]. In particular, the learning rate should be decreasing and converging to zero. Under this condition, it feels natural to conjecture that the long time behavior of SGD is closely related to asymptotic behavior of the trajectories of equation (1.3). This method, called the *ODE method*, was introduced by Ljung [35] and extensively studied after [7, 30]. Even in the stochastic setting, a good understanding of the underlying continuous dynamical system is important.

The emergence of deep learning has spawned the recent popularity of a special class of optimizers: first order *adaptive* optimization algorithms (RMSprop [49], Adagrad[18, 19], Adadelta [52], Adam [28]) were originally designed to solve unconstrained optimization problem for a stochastic cost function (minimizing empirical risk in supervised learning). These optimizers became very popular in deep learning for both supervised and unsupervised learning tasks as it is commonly observed that the value of the training loss decays faster than for stochastic gradient descent. However, few positive results can be found in the literature, and limitations have been found. For example, it has been shown empirically that Adam and RMSprop do do not always converge in the stochastic setting, even for a convex loss function [46].

1.1. **Motivation and main results.** Inspired by the history of gradient descent and stochastic gradient descent, we analyze discrete *adaptive* optimization algorithms by introducing their continuous time counterparts, with a focus on Adam. The techniques and analysis are similar for the other algorithms (and include classical accelerated methods).

Adaptive Moment estimation (Adam) [28] is an iterative method generating a sequence $(\theta_k, m_k, v_k) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d_+$. In addition to the parameters θ, it computes the exponential moving average of the gradient and the squared gradient. The algorithm can be formulated 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 $k \geq 1$

$$\begin{cases} g_k = \nabla f(\theta_{k-1}, \xi_{k-1}) \\ m_k = \mu_k m_{k-1} + (1-\mu_k) g_k \\ v_k = \nu_k v_{k-1} + (1-\nu_k) g_k^2 \\ \theta_k = \theta_{k-1} - s\ m_k / \sqrt{v_k + \varepsilon}. \end{cases} \tag{1.5}$$

where the two parameters for the moving average, depending on the iterations, are given by

$$\begin{cases} \mu_k = \beta_1 (1-\beta_1^{k-1})/(1-\beta_1^k) \\ \nu_k = \beta_2 (1-\beta_2^{k-1})/(1-\beta_2^k). \end{cases} \tag{1.6}$$

Because the coefficients depend on k, the underlying dynamical system for Adam must be non-autonomous. In what follows, we show that Adam is a discretization of a class of differential equations and the connection is established in section 3.3. In this work, we address the following questions

(1) Is there a general random continuous dynamical system underlying adaptive algorithms? Is this system wellposed?
(2) Can adaptive optimization algorithms be formulated as a discretization (possibly an Euler discretization) of a continuous system? If yes, does this numerical approximation converge (as a numerical method)? With which order of convergence?

(3) What are the asymptotic behavior of the continuous trajectories for the deterministic counterpart? What intrinsic property make them well adapted to the specificity of deep learning landscape?

More precisely, we introduce in section 3 a general continuous dynamical system (3.1) whose *forward* Euler approximation matches a large class of first order methods. The connection with adaptive algorithms is made more precise in section 3.3 and with accelerated methods in section 3.4. Our analysis is supplemented by several comments about practical use of adaptive algorithms (see subsection 3.5). Section 4 contains the assumptions and the statement of our main results. We provide conditions under which the random differential equation is well-posed in section 4.1 and we prove that its forward Euler discretization is convergent in section 4.2. Finally, we study the asymptotic behavior of the continuous deterministic trajectories in section 4.3. In the non-convex setting we prove, under mild assumptions, that the trajectories converge to the critical locus of f (see Theorem 4.4). This result is supplemented with the analysis of the necessary conditions in order to avoid convergence to saddle or local maximum points of f (see Theorem 4.7). For convex functions, we design a Lyapunov functional and obtain a rate of convergence to a neighborhood of the critical locus (see Theorem 4.10), which depends on the behavior over time of $\nabla f, v$ and θ. In particular, this indicates that the efficiency of adaptive algorithms is highly dependant on the problem. In sections 5, we specialize the convergence results to adaptive algorithms and accelerated methods. Finally, most proofs supporting the paper are postponed to the Appendix.

This work is intended to serve as a solid foundation for the posterior study in the discrete and stochastic settings. The deterministic convergence analysis confirms that the trajectories converge to the critical locus of the objective function, which leads to natural conjectures on the convergence in the discrete and stochastic setting. In particular, we believe that the Lyapunov functional, used in section 4.3, can be adapted to the stochastic discrete methods. 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 [46]. This behavior is induced by the stochasticity in the algorithm and it is important to use the right framework to analyze the convergence of Adam in the stochastic case. This framework is well understood for the stochastic gradient descent. For this counterexample [46], stochastic gradient descent blows up for constant learning rate but converges for decaying learning rates.

2. Related work

The rate of convergence for gradient descent is not optimal and depending on the class of functions f belongs to, more efficient algorithms can be designed [9, 11, 12, 39, 40]. For smooth convex or strongly convex functions, Nesterov [39] introduced an accelerated gradient algorithm which was proven to be optimal (a lower bound matching an upper bound is provided)

$$v_{k+1} = \theta_k - s\nabla f(\theta_k) \tag{2.1}$$

$$\theta_{k+1} = v_{k+1} + \frac{k}{k+3}(v_{k+1} - v_k). \tag{2.2}$$

However, the key mechanism for acceleration is not well understood and have many interpretations [13, 26, 34]. A particular interesting interpretation of acceleration is through the lens of a second order differential equation of the form

$$\ddot{\theta} = a(t)\dot{\theta} + \nabla f(\theta), \qquad \theta(0) = \theta_0, \quad \dot{\theta}(0) = \psi_0, \tag{2.3}$$

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 (2.3) 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 (2.3). For the *Heavy Ball method*, the function a is constant and is called the damping parameter [2, 3]. In [1, 14, 21], conditions on the rate of decay of a and its limit are given in order for the trajectories of (2.3) to converge to a critical point of f. This analysis highlights situations where (2.3) 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 [48] and the authors draw interesting connections between (2.3) and *Nesterov's algorithm* (2.1). 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 $t = \sqrt{s}k$ [48]. Extension of this work are proposed in [50, 51] in which the authors studied acceleration from a different continuous equation having theoretically exponential rate of convergence. However, a naïve discretization looses the nice properties of this continuous system and current work consists on finding a better one preserving the symplectic structure of the continuous flow [8].

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 [19] consists on multiplying the gradient by a diagonal preconditioning matrix, depending on previous 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. Later on, two new adaptive algorithms RMSprop [49] and Adam [28] were proposed. The preconditioning matrix is an exponential moving average of the previous 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 [28], is therefore not correct [46]. A new algorithm Amsgrad proposed in [46] consists on modifying the preconditioning updates to recover this property. While converging, this algorithm looses the essence of the Adam's algorithm. Adam is such a mysterious algorithm that many works have been devoted to understand its behavior. Variants of Adam have been proposed [53] as well as convergence analysis towards a critical point [6, 16]. However, conditions for convergence are very restrictive and not easy to verify in practice.

3. Presentation of the model and connection to existing optimization algorithms

In this section, we briefly introduce some notations on vector's operations used in the paper. In section 3.2, we present a general system of differential equations as well as a possible discretization of it. In sections 3.3 and 3.4, we explicitly make the connection between this numerical approximation and first order optimization methods. In section 3.5, we make some observations on the behavior of the deterministic version of Adam that we illustrate on toy problems.

3.1. Compact notation.

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:

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

3.2. Presentation of the continuous time model.

Throughout this paper we study the following dynamical system (whose connection with various numerical optimization methods is established via the Euler approximation scheme (3.3)).

$$\left\{ \begin{array}{l} \dot{\theta}(t) = -m(t)/\sqrt{v(t) + \varepsilon} \\ \dot{m}(t) = h(t)\nabla f(\theta(t)) - r(t)m(t) \\ \dot{v}(t) = p(t)\left[\nabla f(\theta(t))\right]^2 - q(t)v(t), \end{array} \right. \tag{3.1}$$

where $\varepsilon \geq 0$, $(\theta, m, v, t) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d_{\geq 0} \times \mathbb{R}_{>0}$, (if $\varepsilon = 0$, then $v \in \mathbb{R}^d_{>0}$). The above system has a momentum term m (or memory term depending on the values for h and r). We also consider the alternative system

$$\begin{cases} \dot{\theta}(t) = -\nabla f(\theta)/\sqrt{v(t) + \varepsilon} \\ \dot{v}(t) = p(t) \left[\nabla f(\theta(t))\right]^2 - q(t)v(t), \end{cases} \tag{3.2}$$

The analysis of this second differential equations is similar (and simpler) to the first. However, it is interesting here to study the impact of the *rescaling term* v. In this work, we always make the following hypotheses

Assumption 1. The objective function f is assumed to be a C^1 function defined in $\mathbb{R}^d$ whose gradient is locally Lipschitz. The functions h, r, p and q are non-negative and non-increasing C^1-functions defined over $\mathbb{R}_{>0}$. We also require that:

$$h(t) \not\equiv 0,\ r(t) \not\equiv 0, \text{ and either } p(t) \not\equiv 0, \text{ or } p(t) \equiv q(t) \equiv 0$$

Additional assumptions about the regularity of those functions will be given in section 4.1. Moreover, we won't make, at any point, the assumption that the ODE (3.1) has globally Lipschitz coefficients. This assumption would be too strong and is not satisfied for quadratic functions, for example. This constitutes an important technical difficulty in the remaining of the paper.

The system is supplemented with initial conditions $\mathbf{x}_0 = (\theta_0, m_0, v_0)$ at time $t = t_0 \geq 0$. We denote by $\mathbf{x}(t, t_0, \mathbf{x}_0) = (\theta(t), m(t), v(t))$ a solution of (3.1) with initial condition $\mathbf{x}(t_0, t_0, \mathbf{x}_0) = \mathbf{x}_0$. In order to establish a relation between the continuous and the optimization algorithms, we study the finite difference approximation of (3.1) by the forward Euler method

$$\begin{cases} \theta_{k+1} = \theta_k - s m_k/\sqrt{v_k + \varepsilon} \\ m_{k+1} = (1 - s r(t_{k+1})) m_k + s h(t_{k+1}) \nabla f(\theta_{k+1}) \\ v_{k+1} = (1 - s q(t_{k+1})) v_k + s p(t_{k+1}) \left[\nabla f(\theta_{k+1})\right]^2 \end{cases} \tag{3.3}$$

where $t_k = ks$. We chose this method because it fits well with Adam discrete system. Moreover its convergence can be proven under certain extra assumptions (see subsection 4.2). However it is certainly not the only choice of discretization. In particular, more efficient quadrature rules can be considered to get more accurate numerical integration [20, 31] in the case of singular functions p, q, r, h.

The connections between our model and optimization algorithms is made precise in the next sections. Indeed, we show that these algorithms exactly match with the forward method, which is a good approximation of the continuous trajectories.

Remark 3.1 (On the stochastic model). In subsections 4.1 and 4.2 below, we enunciate our results in terms of the stochastic version of (3.1)

$$\begin{cases} \dot{\theta}(t) = -m(t)/\sqrt{v(t)+\varepsilon} \\ \dot{m}(t) = h(t)\nabla f(\theta(t),\xi(t)) - r(t)m(t) \\ \dot{v}(t) = p(t)\left[\nabla f(\theta(t),\xi(t))\right]^2 - q(t)v(t), \end{cases} \tag{3.4}$$

where $\xi(t)$ is a stochastic process with continuous sample paths. In this case, the forward Euler discretization yields:

$$\begin{cases} \theta_{k+1} = \theta_k - sm_k/\sqrt{v_k+\varepsilon} \\ m_{k+1} = (1-sr(t_{k+1}))m_k + sh(t_{k+1})\nabla f(\theta_{k+1},\xi_{k+1}) \\ v_{k+1} = (1-sq(t_{k+1}))v_k + sp(t_{k+1})\left[\nabla f(\theta_{k+1},\xi_{k+1})\right]^2 \end{cases} \tag{3.5}$$

3.3. **Adaptive optimization algorithms.**

3.3.1. Adagrad *differential equation.* Adagrad [18] was designed to incorporate knowledge of the geometry of the data previously observed during the training. For all $k \in \mathbb{N}$,

$$\begin{cases} v_{k+1} = \displaystyle\sum_{j=0}^{k} \left[\nabla f(\theta_j)\right]^2 \\ \theta_{k+1} = \theta_k - s\nabla_\theta f(\theta_k)/\sqrt{v_{k+1}}. \end{cases} \tag{3.6}$$

with initial conditions $v_0 = 0$ and $\theta_0 \in \mathbb{R}^d$. Here, we recall that $[\nabla_\theta f(\theta)]^2$ denotes the element-wise product $\nabla_\theta f(\theta) \odot \nabla_\theta f(\theta)$. The adaptive part in the algorithm comes from the term $\sqrt{v_k}$ which is precisely the preconditioning matrix used to scale the gradients. The algorithm (3.6) can be equivalently described by

$$\begin{cases} \theta_{k+1} = \theta_k - s\nabla f(\theta_k)/\sqrt{G_k} \\ G_{k+1} = G_k + [\nabla f(\theta_{k+1})]^2. \end{cases} \tag{3.7}$$

with initial condition θ_0 and $G_0 = \nabla f(\theta_0)^2$. By setting $\alpha = s^2$ and $\omega_k = \alpha G_k$, it is easy to conclude that the Adagrad's update rule can be written as:

$$\begin{cases} \theta_{k+1} = \theta_k - \alpha\nabla f(\theta_k)/\sqrt{\omega_k} \\ \omega_{k+1} = \omega_k + \alpha[\nabla f(\theta_{k+1})]^2 \end{cases} \tag{3.8}$$

with the re-scaled initial condition $\omega_0 = \alpha[\nabla f(\theta_0)]^2$ and $\theta_0 \in \mathbb{R}^d$. It is now easy to see that (3.8) is a forward Euler discretization of the system of differential equations (3.2) (which we call the Adagrad differential equation) with initial condition given by $\theta_0 \in \mathbb{R}^d$, $\omega_0 = \alpha[\nabla f(\theta_0)]^2$ and with $q \equiv 0, \varepsilon = 0$ and $p \equiv 1$.

3.3.2. RMSProp *and* Adam *differential equations.* The only difference between these two optimizers and Adagrad is how the preconditioning matrix is computed. In RMSProp, it consists of an exponentially decaying moving average rather than a sum of the previous gradients

$$\begin{cases} v_{k+1} = \beta v_k + (1-\beta)\nabla_\theta f(\theta_k)^2 \\ \theta_{k+1} = \theta_k - s\nabla_\theta f(\theta_k)/\sqrt{v_{k+1}}. \end{cases} \tag{3.9}$$

Adaptive Moment estimation (Adam) [28] is a famous variant of RMSProp that incorporates a momentum equation. More precisely, it computes the exponential moving average of the gradient and the square gradient. This method combines the advantages of RMSProp [49] in addition to the running average for the gradient. We recall that Adam has three hyperparameters: 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. As already formulated in the Introduction, 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 $k \geq 1$

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

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

$$\theta_k = \theta_{k-1} - s\ m_k/\sqrt{v_k + \varepsilon}. \tag{3.10}$$

This modification does not change anything in 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 $k \geq 0$

$$\begin{cases} \theta_{k+1} = \theta_k - s\ m_k/\sqrt{v_k + \varepsilon} \\ g_{k+1} = \nabla f(\theta_{k+1}) \\ m_{k+1} = \mu_{k+2} m_k + (1-\mu_{k+2}) g_{k+1} \\ v_{k+1} = \nu_{k+2} v_k + (1-\nu_{k+2}) g_{k+1}^2 \end{cases} \tag{3.11}$$

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

Remark 3.2 (On the parameters μ and ν). In the original formulation of the algorithm (as stated in (1.5)), 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 $\nu = \beta_2$ and this does not change the conclusion of our work. In particular, let us consider the system (3.1) with

$$h \equiv r \equiv 1/\alpha_1, \qquad p \equiv q \equiv 1/\alpha_2$$

where α_1, α_2 are two positive constants. It can be easily checked that the choice of $\beta_1 = (1 - s/\alpha_1)$ and $\beta_2 = (1 - s/\alpha_2)$ leads to the ADAM optimizer (3.11) without rescaling.

Consider now the three parameter family of differential equations

$$\left\{ \begin{array}{l} \dot{\theta} = -m/\sqrt{v+\varepsilon} \\ \dot{m} = g_1^A(t,\lambda,\alpha_1,\alpha_2)\,(\nabla f(\theta) - m) \\ \dot{v} = g_2^A(t,\lambda,\alpha_1,\alpha_2)\,\left(\nabla f(\theta)^2 - v\right) \end{array} \right. \tag{3.12}$$

where the coefficients in (3.1) are given by

$$h \equiv r \equiv g_1^A(t,\lambda,\alpha_1,\alpha_2), \qquad p \equiv q \equiv g_2^A(t,\lambda,\alpha_1,\alpha_2), \tag{3.13}$$

and $(\lambda, \alpha_1, \alpha_2)$ are positive real numbers and:

$$g_i^A(t,\lambda,\alpha_1,\alpha_2) = \frac{1 - e^{-\lambda/\alpha_i}}{\lambda\left(1 - e^{-t/\alpha_i}\right)}, \qquad i = 1, 2. \tag{3.14}$$

Note that both functions have a simple pole at $t = 0$ and, furthermore, satisfy assumption 3 below. Now, let us consider the associated discretization (3.3) 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

$$\lambda = s, \qquad \beta_i = e^{-\lambda/\alpha_i},\ i = 1, 2. \tag{3.15}$$

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

$$s g_i^A((k+1)s, \lambda, \alpha_1, \alpha_2) = 1 - \beta_1 \frac{1 - \beta_1^k}{1 - \beta_1^{k+1}} = 1 - \mu_{k+1},$$

which recovers ADAM's discrete system (3.11) (apart from small difference in the evaluation of μ). Therefore, ADAM is an Euler discretization of system (3.1) for the choice of function (3.13)-(3.14) and parameters (3.15).

3.4. **Accelerated optimization algorithms.** In this section, we show that our framework encompasses previous studies of accelerated methods.

3.4.1. *Heavy Ball differential equation.* We consider the Heavy ball second order differential equation [2]

$$\ddot{x} + \gamma\dot{x} + \nabla f(x) = 0, \tag{3.16}$$

where $\gamma > 0$. By taking $\theta = x$ and $m = -\dot{x}$ (and $v \equiv 1$), we obtain the system (3.1) with

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

Equation (3.3) simplifies to

$$\begin{cases} \theta_{k+1} = \theta_k - sm_k \\ m_{k+1} = (1 - s\gamma)m_k + s\nabla f(\theta_{k+1}) \end{cases} \tag{3.17}$$

which corresponds to the classical Heavy ball methods with damping coefficient $\beta = 1 - s\gamma$, momentum variable $n_k = sm_k$ and learning rate $\alpha = s^2$. Implicit discretization has also been considered in [2].

3.4.2. *Nesterov differential equation.* Following [48], we consider the Nesterov second order differential equation, parametrized by the constant $r > 0$,

$$\ddot{x} + \frac{r}{t}\dot{x} + \nabla f(x) = 0. \tag{3.18}$$

Similarly as in the Heavy Ball case, we define $\theta = x$ and $m = -\dot{x}$ and write the above equation as a system (3.1) with

$$h(t) \equiv 1, \qquad r(t) = r/t, \qquad \textit{and} \qquad p(t) \equiv q(t) \equiv 0.$$

In [48], 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 $s \to 0$. Similar analysis holds here.

3.5. **Considerations on adaptive algorithms.** We make here empirical observations and discuss some limitations on adaptive algorithms. On 2D toy problems, we emphasize some facts limiting the applicability of such algorithms in practice.

3.5.1. *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 depend on the local curvature of the function f near the optimum.

Proposition 3.3 (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* (3.11)*.*

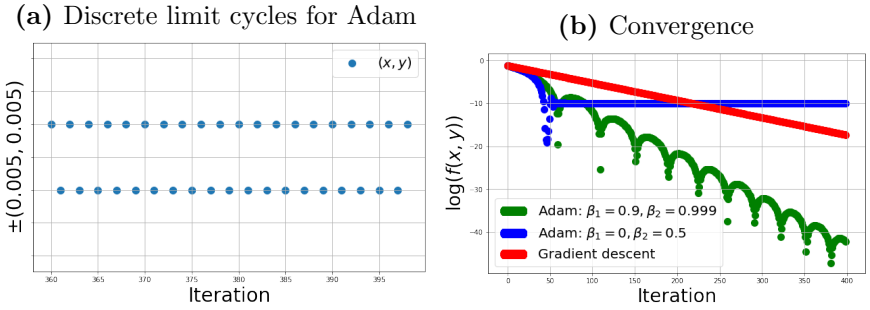

Figure 1. Illustration of discrete limit cycles for the ADAM's algorithm with $\varepsilon = 10^{-8}, \beta_2 = 0.5, s = 10^{-2}$. **a)** Limit cycle of period two for ADAM. The algorithm oscillates between two points $(0.005, 0.005)$ and $(-0.005, -0.005)$. **b)** Plot of the logarithm of f versus the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

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

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 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 will reduce this gap but cannot remove it.

3.5.2. *For* ADAM *and* RMSPROP*, the hyper-parameters β_1, β_2 are functions of learning rate s.* 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 3.3, the parameters β_1 and β_2 were chosen as functions of the learning rate s. 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 α_i constant. The coefficients must be modified according to the formula (3.15), which we recall here

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

In plot **b)**, Figure 2, we compute the logarithm of the error between different trajectories

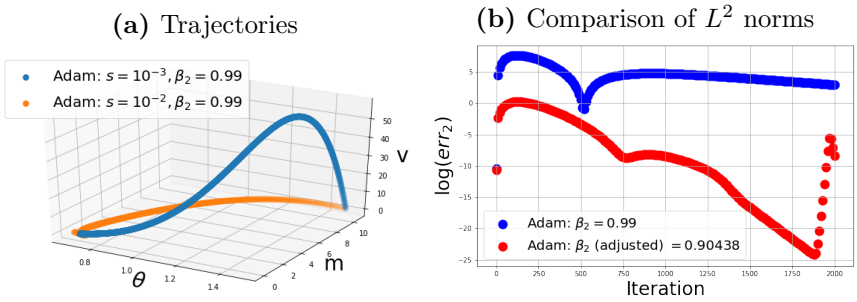

Figure 2. Fixing β_2 and changing the learning rate s lead to different dynamics. **a)** Trajectories of ADAM (1) & (2) when only the learning rate is changed. **b)** Comparison of the error between trajectories (1) & (2) and (1) & (3). As expected the discrepancies between (1) & (3) is very small.

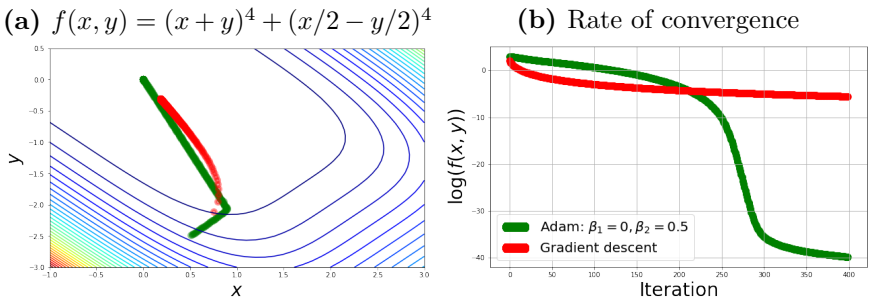

Figure 3. Comparison between gradient descent and ADAM. Gradient Descent converges faster initially when the gradients are large but ADAM outperforms Gradient descent after entering the flat region. Both trajectories start from the point $(0.5, -2.5)$.

(1) This is the reference dynamics: $\beta_2 = 0.99$ and $s = 0.001$. According to formula (3.15), $\alpha_2 = -0.001/\log(0.99) \approx 0.0995$
(2) The second dynamics: $\beta_2 = 0.99$ and $s = 0.01$
(3) For the third dynamics, we keep the same learning rate $s = 0.01$ but adjust $\beta_2 = \exp(-0.01/0.0995) = 0.90438$.

3.5.3. *On the strength of* ADAM *and* RMSPROP *on flat surfaces.* The convergence analysis in the convex case (see Section 4.3) seems to indicate that ADAM and RMSPROP are rather slow algorithms and the convergence is only guaranteed in a neighborhood of the global minimum. However, there are situations where they seem to perform consistently well. This is the case for flat surfaces (see figure 3).

3.5.4. ADAM *and* RMSPROP *are fundamentally different from* ADAGRAD *and* AMSGRAD. It is common to consider adaptive algorithms as a whole.

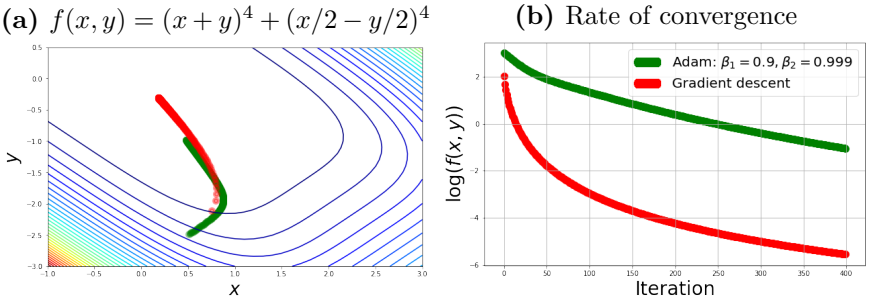

Figure 4. Comparison between gradient descent and ADAM. Gradient Descent outperforms ADAM in this example because β_1, β_2 are large and ADAM keeps memory of the past large gradients. Both trajectories start from the point $(0.5, -2.5)$.

However, we do not believe this should be done. The variable v has different asymptotic and it is easily seen on the continuous dynamics that the equilibrium are different. In the strictly convex setting, for ADAM (with $\varepsilon > 0$), the equilibrium is given by $(\theta_\star, 0, 0)$ (and respectively $(\theta_\star, 0)$ for RMSPROP), while for ADAGRAD it is given by $(\theta_\star, c)$ for some (large) constant c. In particular because v is non-decreasing, we believe that AMSGRAD should be considered as a modification of ADAGRAD rather than ADAM.

4. Statement of the main results

In this section, we state the main theoretical results of this paper: existence and uniqueness of solutions for the continuous equation (3.4), order of convergence for the *forward* Euler scheme (3.5) and convergence analysis for equation (3.1) (in the deterministic case).

4.1. Existence and uniqueness of solutions.

In this section we present the main result about existence and solutions for the random ordinary differential equation (3.4). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space where $\mathcal{F}$ is a σ-algebra on Ω and $\mathbb{P}$ is a probability measure. A continuous random differential equation is an ordinary differential equation for almost every realization $\omega \in \Omega$ and we enjoy the fact that the techniques for the deterministic and stochastic system are essentially the same. However, the solution to a random differential equation has to be a stochastic process defined on an ω-independent time interval. We ask for the following assumptions on the noise process, the function f and its gradient.

Assumption 2.

(1) Let $\xi : [0, T] \times \Omega \to \mathbb{R}^m$ be an $\mathbb{R}^m$-valued stochastic process with continuous sample paths.
(2) For almost all $\omega \in \Omega$, $\nabla f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ is a continuous function.

(3) For all $\nu \in \mathbb{R}^d$, the function $f(\cdot, \nu) : \mathbb{R}^d \to \mathbb{R}$ is C^1 and its gradient is locally Lipschitz i.e for all ν and all $x, y, \in \mathbb{R}^d$, there exists a constant L, depending on the norm of x and y, such that almost surely

$$\|\nabla f(x, \nu) - \nabla f(y, \nu)\| \leq L\left(\|x\|, \|y\|\right) \|x - y\|.$$

The function L is almost surely bounded on the bounded sets.

(4) There exist a positive constant B such that for all $\nu \in \mathbb{R}^m$ almost surely

$$\|\nabla f(0, \nu)\| \leq B.$$

Therefore

$$\|\nabla f(x, \nu)\| \leq L\left(\|x\|\right) \|x\| + B.$$

The assumption on the continuity of the stochastic process $(\xi(t))_t$ is only necessary to prove that the solution $\mathbf{x}$ is continuously differentiable. It can be relaxed if the continuity of $\mathbf{x}$ is sufficient. Moreover, note that the ODE (3.4) is measurable in ω and continuous in (t, x), since ∇f is continuous in both variables, and $\xi_t(\omega)$ is measurable in ω and has continuous sample paths. Finally, we assume that the gradient is almost surely locally Lipschitz. It corresponds to locally Lipschitz in the deterministic case and is a necessary condition in order to use the Banach contraction mapping theorem.

In order to study the existence of the solution at $t = 0$, we demand the following from the coefficients, which are allowed to have a pole at zero.

Assumption 3. We assume one of the following condition

(1) The functions h, r, p, q have a simple pole at $t = 0$.
(2) If $h \in C^1([0, +\infty))$ (resp $p \in C^1([0, +\infty))$), then r (resp. q) can have (at most) a simple pole at zero.
(3) In any other cases, all functions are assumed to be C^1 on $[0, \infty)$. The initial conditions can be taken arbitrarily.

In cases (1) and (2), furthermore, we demand that there exists a small time $\hat{t}$ such that

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

and that the initial conditions must be taken as:

$$m_0 = \nabla f(\theta_0, \xi_0) \lim_{t \to 0^+} h(t)/r(t), \quad v_0 = \left[\nabla f(\theta_0, \xi_0)\right]^2 \lim_{t \to 0^+} p(t)/q(t).$$

Note that these assumptions are verified by Adam's (under certain assumptions on the parameters, e.g. $\beta_1 \geq 0.21$) and Nesterov equations (3.12) and (3.18), respectively. The potential singularity in time makes the analysis more technical and we rely on a compactness argument to prove wellposedness at $t = 0$. In order to illustrate this technical difficulty, let us consider the differential equation:

$$\dot{\theta}(t) = \frac{L}{t}\theta(t), \quad \theta(0) = 0$$

Note that there always exist a solution given by $\theta(t) \equiv 0$ of the above system. Now the solutions of this equation with initial condition at $t_0 = 1$ are given

by $\theta(t) = \theta(1) \cdot t^L$ (for $L \neq 1$). Therefore, uniqueness only holds if $L < 0$. Indeed, if $L \geq 0$, for every $\theta(1) \in \mathbb{R}$ the solution $\theta(t) = \theta(1) \cdot t^L$ converges to 0 when $t \to 0$. Finally we assume that the local variance is bounded, which is needed to prove existence and uniqueness at $t = 0$. More precisely:

Assumption 4. There exists two constants σ_1 and σ_2 such that for all $t > 0$

$$\mathbb{E}\left(\sup_{0<u\leq t} \|h(u)\left(\nabla f(x, \xi(u)) - \nabla f(x, \xi_0)\right)\|^2\right) \leq \sigma_1$$

and

$$\mathbb{E}\left(\sup_{0<u\leq t} \left\|p(u)\left([\nabla f(x, \xi(u))]^2 - [\nabla f(x, \xi_0)]^2\right)\right\|^2\right) \leq \sigma_2.$$

for all initial conditions $x \in \mathbb{R}^d$ and $\xi_0 \in \mathbb{R}^m$.

We recall that the system (3.1) is supplemented with initial conditions. We say that they are admissible at t_0 if $\mathbf{x}(t_0) = (\theta_0, m_0, v_0) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d_{\geq 0}$. If $\varepsilon = 0$, then $v \in \mathbb{R}^d_{>0}$. We are now ready to enunciate our main result:

Theorem 4.1. *Suppose that the ODE* (3.4) *satisfies assumptions 1 and 2. In addition, for any $t_0 > 0$ and admissible initial condition $\mathbf{x}(t_0)$, there exists a unique global solution to equation* (3.4) *such that $\mathbb{P}$-almost surely:*

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

Suppose, furthermore, that assumptions 3 and 4 are also satisfied. Then, there exists a unique global solution to equation (3.4) *such that $\mathbb{P}$-almost surely:*

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

The proof is postponed in Appendix B and is divided as follows: we first study existence and uniqueness for $t > 0$ using the Banach fixed point theorem and a cut-off argument. Then we study a regularized system and from a compactness argument, we obtain that the system is well-posed for $t = 0$.

4.2. **Convergence of the Euler discretization.** We now study the validity of the approximation given by the discrete system (3.5) of the differential equation (3.4). As in the previous section, we work on the stochastic equations and we give some additional regularity assumptions on the noise process $(\xi_t)_{t\in\mathbb{R}_+}$.

In the traditional finite-time convergence (and order of convergence) theory for numerical methods, one usually requires the vector field to be globally Lipschitz, which is not the case here (see Assumptions 2). This represents a technical difficulty, and we rely on weaker assumptions, such as the one's requested in [25], to prove our results. Furthermore, as pointed out in [23, 29, 38], the stochastic driving process has usually at most Hölder

continuous sample paths (and therefore are not continuously differentiable) and standard numerical analysis do not directly apply.

In order to be precise, let us now fix the notation. Fix a (final) time $T > t_0$ and consider an interval $[t_0, T]$ on which we study the approximation of the solution of (3.4). We recall that the step size is given by $s > 0$. Consider $K_{t,s} = [(t - t_0)/s]$, the integer part of $(t - t_0)/s$. We write $t_k = t_0 + ks$ for any $k \in [\![0, K_{t,s}]\!]$ and $\pi = \{t_0 < t_1 < \cdots < t_K = T\}$ is a homogeneous partition of the interval $[t_0, T]$. Now, let $\mathbf{x}_0 = (\theta_0, m_0, v_0)$ be a fixed initial condition and consider:

- The sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ given by the discrete system (3.5) with initial condition $\mathbf{x}_0$;
- The sequence $\widetilde{\mathbf{x}}_k := \mathbf{y}(t_k)$ for all $k \in \mathbb{N}$, where $\mathbf{y}(t)$ is the exact solution of ODE (3.4) with initial condition $\mathbf{y}(t_0) = \mathbf{x}_0$.

We study the almost sure order of convergence of the Euler approximation. Following [23] we prove that the strong order of convergence in probability is determined by the order of the Hölder continuity of the sample paths of the driving process. More precisely, as in [29, page 2931], we make the following assumption on the regularity in time of the stochastic process $(\xi_t)_{t\in\mathbb{R}_+}$.

Assumption 5. The stochastic process $(\xi_t)_{0\le t\le T}$ is assumed to have sample paths which are locally Hölder continuous with the same exponent i.e. there exists a $\gamma \in (0, 1]$ such that almost surely and for all T, there exists a constant

$$\left\|\xi(t) - \xi(t')\right\| \le C_T \left|t - t'\right|^\gamma, \qquad \forall t, t' \le T.$$

Moreover, we introduce the modulus of continuity of the gradient of f on $[0, T]$

$$\omega_f(s, x, T) = \sup_{\substack{t\neq u,\\ t_0\le t,u\le T,\\ |t-u|\le s}} \|\nabla f(x, \xi(t)) - \nabla f(x, \xi(u))\|$$

and we make the following assumption

Assumption 6. There is a positive constant α and a positive constant C, bounded on bounded sets, such that

$$\omega_f(s, x, T) \le C(\|x\|) \sup_{\substack{t\neq u,\\ t_0\le t,u\le T,\\ |t-u|\le s}} \|\xi(t) - \xi(u)\|^\alpha$$

These two conditions will determine the order of approximation of the Euler scheme. Following [23], we get:

Theorem 4.2. *Let $T > 0$ and suppose that $t_0 > 0$. Let us consider a compact set A_0 of $\mathbb{R}^d_{\ge 0} \times \mathbb{R}^d_{\ge 0} \times \mathbb{R}_{>0}$ and assume that the ODE* (3.4) *and discretization* (3.5) *satisfies assumptions 1, 2, 5 and 6. Then, there exists a constant $C(T, A_0)$ (which only depends on T and the compact A_0) such*

that for any admissible initial condition $\mathbf{x}(t_0) \in A_0$*, the numerical scheme satisfies*

$$\max_{k=0,\cdots,K} \|\mathbf{x}_k - \widetilde{\mathbf{x}}_k\| \leq C(T, A_0) s^{\min(\alpha\gamma,1)}.$$

4.3. **Convergence analysis in continuous time.** In this section, we study the asymptotic behavior of the solutions of (3.1). Our analysis is based on the following three steps:

(1) *Topological convergence*: Find sufficient conditions on the functions f and p, q, r, h in order for the solutions of equation (3.1) to converge to a critical value of f, that is, $\nabla f(\theta(t)) \to 0$ when $t \to \infty$. 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 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 4.3.2, 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.

Remark 4.3 (On the convexity assumption). For non-convex functions, in a neighborhood of a local minimum, the function is not always locally convex. Consider, for example, the function $f(x) = x_1^2 x_2^2$ which has the origin as a local minimum. It is not locally convex at 0 since there is no sufficiently small neighborhood of the origin where f is convex. However, consider the space $C^\infty(\mathbb{R}^d, \mathbb{R})$ of all C^∞ function $f : \mathbb{R}^d \to \mathbb{R}$ and fix a compact set K. For *almost* every function $f \in C^\infty(\mathbb{R}^d, \mathbb{R})$, if $x_0 \in K$ is a local minimum of f then f is locally convex at x_0 (see Remark F.3 for a precise statement). Therefore a small perturbation of the function $f(x) = x_1^2 x_2^2$ can induce local convexity. Consider the one-parameter family $f_\lambda(x) = x_1^2 x_2^2 + \lambda^2(x_1^2 + x_2^2)$. Then for every $\lambda \in \mathbb{R} \setminus \{0\}$, the function $f_\lambda(x)$ is locally convex at 0. This is illustrated in Figure 5.

In the remaining of this section, we give precise statements for all three steps. For each step, we will make appropriate assumptions on the objective function. However, the following two assumptions must be satisfied in all cases

Assumption 7. There exists $\widetilde{t} \in [t_0, \infty)$ such that:

$$2r(t) - \frac{q(t)}{2} + \frac{h'(t)}{h(t)} \geq 0, \qquad \forall\, t > \widetilde{t}$$

Assumption 8. The solution $\theta(t)$ of the ODE (3.1) is bounded.

We start by the topological convergence analysis.

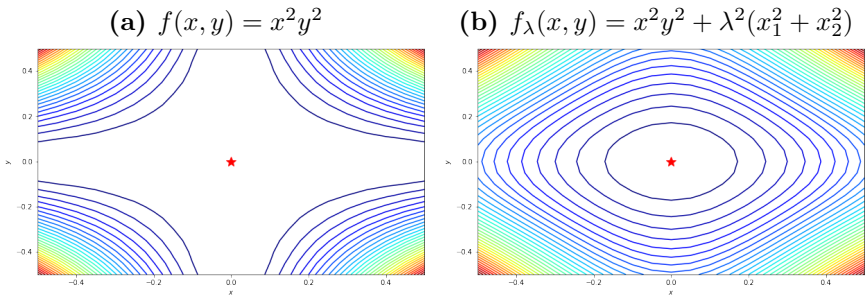

Figure 5. On locally convex C^2 functional.

4.3.1. *Topological convergence.* In this part, we make an additional assumption on the asymptotic behavior of the coefficients.

Assumption 9. Suppose that $\varepsilon > 0$. Consider the functions:

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

and suppose that these functions are C^1 in $[0, \infty)$ such that $H(0) > 0$; $R(0) > 0$; $4R(0) > Q(0)$ and either $P(s) \equiv Q(s) \equiv 0$, or $Q(0) > 0$.

Let us observe that this assumption is satisfied when the coefficients do not converge to zero at infinity. Hence, it holds for Adam (3.12) and the Heavy ball differential equation (3.16) but not for the Nesterov's acceleration equation (3.18). Under this assumption, we prove the convergence of the dynamics in the following sense:

Theorem 4.4 (Topological Convergence)**.** *Suppose that assumptions 1, 7, 8 and 9 are verified (and that $v_0 = 0$ and $\varepsilon = 1$ if $p(t) \equiv q(t) \equiv 0$). Then $f(\theta(t)) \to f_\star$, $m(t) \to 0$ and $v(t) \to 0$ when $t \to \infty$, where $f_\star$ is a critical value of f.*

The proof of this result is postponed in Appendix D. Our method is inspired by the work of Alvarez [2], based on the energy functional of the system. However, we use a different argument, based on the Poincaré-Bendixson type arguments, which does not rely on convexity. In the next part, we improve the above result and we prove that the dynamics converge to a local minimum for almost all initial condition.

4.3.2. *Avoiding local maximum and saddles.* Before stating the main hypothesis of this section, we need introduce the definition of isolated critical points and strict saddle points

Definition 4.5. A critical point $\theta_\star$ is called isolated if there is a neighborhood U around $\theta_\star$ that does not contain any other critical points.

Definition 4.6. (Following [32, Definition 1]) A critical point $\theta_\star$ of a C^2 function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be a *strict saddle* if there exists a strictly negative eigenvalue of the Hessian $\mathcal{H}_f(\theta_\star)$ of f at $\theta_\star$.

We now make the following assumption

Assumption 10. The function f is C^2. Moreover each critical point $\theta_\star$ which is not a local-minimum is a strict saddle and is isolated (as a critical point).

Now, fix a time $t_0 > 0$ and recall that the topological limit of a curve $\theta(t)$ 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 4.7 (Avoiding Saddle and Local Maximum points)**.** *Suppose that assumptions 1, 7, 8, 9 and 10 are satisfied. Then, for every $t_0 > 0$, the set S_{t_0} has Lebesgue measure zero. More precisely, the Hausdorff dimension of S_{t_0} is smaller or equal to $3d - 1$.*

Remark 4.8. 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.

Our method to prove the above result relies on the central(-stable) manifold theory of differential equations (see [45, Chapter 1] for detailed exposition). Similar results are proved for discrete systems having isolated critical points in [32, 33]. An extension of this result to non-isolated critical points can be found in [41]. In this work, we assume that critical points are isolated in order to exclude pathological differences between local and global center-stable manifold theory (see example 4.9 below). The general theory of central-stable manifolds (c.f. [45, Theorem 3.2] and [41, Lemma 5]) can be applied if the differential equation has globally Lipschitz coefficients and assumption 10 is not mandatory. However, the global Lipschitz assumption is not met in our case. We believe, nevertheless, that assumption 10 can be relaxed to prove saddle point avoidance in the case of non-isolated critical points, c.f [41].

Example 4.9. Consider the differential equation:

$$\begin{aligned}\dot{x} &= y - (y^2 - \sin(x)^2)\sin(x)\cos(x)\\ \dot{y} &= \sin(x)\cos(x) + (y^2 - \sin(x)^2))y.\end{aligned}$$

and let us consider the orbits of the differential equation whose limit set contain the singularity $(\pi, 0)$. Note the following contrast:

- Local: consider a (very) small neighborhood U of $(\pi, 0)$, then the only solutions which contain $(\pi, 0)$ in their limit set have initial conditions in:
$$(x(t_0), y(t_0)) \in U \cap \{y^2 = \sin(x)^2\}$$
and all other solutions "leave" U in finite time.
- Global: Every solution $(x(t), y(t))$ with initial condition
$$(x(t_0), y(t_0)) \in \{-\pi < x < \pi,\, y^2 < \sin(x)^2\} \setminus \{(0, 0)\}$$
converges to the set $\{-\pi \leq x \leq \pi,\, y^2 = \sin(x)^2\}$, which contains the singular point $(\pi, 0)$.

4.3.3. *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 (c.f. Remark 4.3) and a Lyapunov energy functional (see [2, 3, 21, 48]). Strictly speaking, we do not find a Lyapunov functional for (3.1), but a natural functional which allow us to prove convergence to a least a neighborhood of a local minimum. For accelerated methods, the proposed functional corresponds to the standard Lyapunov energy used in many other works c.f. [2, 3, 21, 48]. This approach relies on the following assumptions:

Assumption 11. 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 $\theta \in \mathbb{R}^d$.

Assumption 12. Let $t_0 > \tilde{t}$ (as defined in assumption 7) and suppose that

$$\lim_{t\to\infty} \int_{t_0}^{t} e^{-\int_{t_0}^{s} r(u)du} ds < +\infty \tag{4.1}$$

$$\text{either } \lim_{t\to\infty} \frac{p(t)}{q(t)} < \infty \text{ or } p(t) \equiv q(t) \equiv 0. \tag{4.2}$$

Furthermore, we define the following three functions (which only depend on h, k, r and q)

$$\mathcal{A}(t) = \int_{t_0}^{t} h(s)\mathcal{B}(s)ds \tag{4.3}$$

$$\mathcal{B}(t) = e^{\int_{t_0}^{t} r(s)ds} \int_{t}^{\infty} e^{-\int_{t_0}^{s} r(u)du} ds \tag{4.4}$$

$$\mathcal{C}(t) = \frac{\mathcal{A}(t)}{h(t)} = \frac{1}{h(t)} \int_{t_0}^{t} h(s)\mathcal{B}(s)ds \tag{4.5}$$

and we suppose that for all $t \geq t_0$

$$\mathcal{B}^2(t) \leq \mathcal{C}(t) \tag{4.6}$$

$$3\mathcal{B}(t) \leq \mathcal{C}(t) \left(2r(t) - \frac{q(t)}{2} + \frac{h'(t)}{h(t)}\right). \tag{4.7}$$

We are now ready to state the main theorem of this section:

Theorem 4.10. *We assume that assumptions 1, 7, 11 and 12 are all satisfied. Then*

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

where $\mathcal{E}(t, m, v, \theta)$ is a Lyapunov functional for the system (3.1)*. In particular under assumption 8, 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 $t \geq t_0$:*

$$f(\theta(t)) - f(\theta_\star) \leq \frac{\mathcal{E}(t_0, m_0, v_0, \theta_0)}{\mathcal{A}(t)} + \frac{\mathcal{K}_1 + \mathcal{K}_2 \int_{t_0}^t q(u)du}{\mathcal{A}(t)}.$$

It follows that the ODE (3.1) *converges to the minimum point with rate of convergence of order at most:*

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

5. 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 them. We start by a brief discussion on the assumptions which appear in this section.

5.1. On the different assumptions.

In the convergence analysis, we made several assumptions on the objective function and the coefficients. We briefly discuss their meaning and situations where they are satisfied.

- Assumption 7 gives an asymptotic relationship between the coefficients. In what follows, we discuss how this assumption affects the choice of the parameters of each optimization algorithm.
- Assumption 8 states that the trajectory $\theta(t)$ is bounded. While this may not always be true, there are some very practical situations where this assumption is always satisfied. For example in the case of coercive objective functions (see Lemma D.1).
- Assumption 9 is satisfied, roughly, when the coefficients do not converge to zero at infinity. It is worth noticing that this assumption holds for Adam (3.12) and the Heavy ball differential equation (3.16) but not for the Nesterov's acceleration equation (3.18). This represents a limitation of our analysis.
- Assumption 10 gives a condition on the nature and the degeneracy of the critical points of the objective function. As discussed in Remark F.3, this assumption is satisfied for a large class of functions. Indeed, every Morse function f satisfies assumption 10. This implies that, if we fix a compact set K, for *almost* every function $f \in C^\infty(\mathbb{R}^d, \mathbb{R})$, the restriction of f to K satisfies assumption 10 (the meaning of almost is made more precise in Remark (F.3)).

5.2. **Convergence of Adam.** In this section, we state convergence results in the specific case of the ADAM's differential equation (3.12).

Corollary 5.1 (Convergence of ADAM)**.** *Suppose $\varepsilon > 0$ and let assumptions 1 and 8 be satisfied for equation* (3.12)*. Moreover, we assume*

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

Then the following convergence results hold true

- *(I)* Topological convergence: *$f(\theta(t)) \to f_\star$, $m(t) \to 0$ and $v(t) \to 0$ when $t \to \infty$, where $f_\star$ is a critical value of f.*
- *(II)* Non-local minimum avoidance: *Assume that assumption 10 holds true.Then the set S_{t_0} has Lebesgue measure zero.*
- *(III)* Rate of convergence: *Under the additional assumption 11 , there exists a constant $\mathcal{K} > 0$ which depends on f, θ_0 and v_0, so that:*

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

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

Remark 5.2 (Existence and Uniqueness of Solution for Adam differential equation). In order to guarantee that assumption (3) is satisfied (and therefore, there exists an unique solution of Adams differential equation at $t_0 = 0$), it is enough to demand $\beta_1 \geq 0.21$. Indeed, in this case we cna guarantee that $q(t) - 2r(t) < 0$ for small enough t.

Proof of Corollary 5.1. The proof of *(I)* directly follows from Theorem 4.4 provided that assumptions 7 and 9 are satisfied. Hence, the proof simply consists on 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

$$h \equiv r \equiv g_1^A(t, \lambda, \alpha_1, \alpha_2), \qquad p \equiv q \equiv g_2^A(t, \lambda, \alpha_1, \alpha_2),$$

and $(\lambda, \alpha_1, \alpha_2)$ are positive real numbers and:

$$g_i^A(t, \lambda, \alpha_1, \alpha_2) = \frac{1 - e^{-\lambda/\alpha_i}}{\lambda\left(1 - e^{-t/\alpha_i}\right)}, \qquad i = 1, 2.$$

It is easy to check that assumptions 7 and 9 are satisfied if there exists a t, large enough, such that

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

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

$$1 - e^{-\lambda/\alpha_1} > \frac{1 - e^{-\lambda/\alpha_2}}{4}.$$

We conclude using the expressions of β_1 and β_2. The proof of *(II)* follows directly from Theorem 4.7 since assumptions 7 and 9 are satisfied under the

condition $3 + \beta_2 > 4\beta_1$. In order to prove *(III)*, let us check the hypotheses of Theorem 4.10. 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:

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

and it easily follows that assumption 12 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. □

5.3. **Convergence of Adagrad.** We start by recalling the differential equation related to ADAGRAD, which was derived in subsection 3.3.1:

$$\begin{cases} \dot{\theta}(t) = -\nabla f(\theta(t)) / \sqrt{\omega(t)} \\ \dot{\omega}(t) = \left[\nabla f(\theta(t))\right]^2, \end{cases} \tag{5.1}$$

with initial condition given by $\theta_0 \in \mathbb{R}^d$ and $\omega_0 = \alpha[\nabla f(\theta_0)]^2$. Now, every consideration made for (3.1) can be specialized to this differential equation via the following functional:

$$\begin{aligned}
E(t, \theta, \omega) &= f(\theta) + \frac{1}{2} \left\| \omega^{1/4} \right\|^2 \\
\mathcal{E}(t, \theta, \omega) &= t \left[f(\theta) - f(\theta_\star) \right] + \frac{1}{2} \left\| [\omega]^{1/4} \odot (\theta - \theta_*) \right\|^2
\end{aligned}$$

Indeed, mutatis mutandis, the same considerations (in a much simpler form) made in the appendixes give rise to the following results:

Theorem 5.3. *Suppose that assumptions 1 and 8 are satisfied for equation* (5.1)*. Then*

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

(II) Non-local minimum avoidance: *We assume the additional hypothesis 10 on the objective function. Fix $t_0 > 0$ and denote by S_{t_0} the set of initial conditions $(\theta_0, \omega_0) \in \mathbb{R}^d \times \mathbb{R}^d_{\geq 0}$ such that $\theta_\star \in \omega(\theta(t))$, where $\theta_\star$ is not a local-minimum of f. Then the Lebesgue measure of S_{t_0} is zero.*

(III) Rate of convergence: *Under the additional assumption 11, $f(\theta(t)) \to f(\theta_\star)$ with the rate $\mathcal{O}(1/t)$.*

5.4. **Convergence of Heavy Ball.** In this section, we recover some of the classical convergence results for the Heavy Ball method. Let us first recall that the Heavy Ball differential equation (3.16) is a special case of the ODE (3.1) with the choice

$$h(t) = 1 \qquad r(t) = \gamma \qquad p(t) = q(t) = 0.$$

Under assumptions 1, 8, 10 and 11, it is easy to check that all hypotheses of Theorems 4.4, 4.7 and 4.10 are satisfied. Moreover, a direct computation gives the rate of convergence $\mathcal{O}(1/t)$ since

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

Analogous results for the discrete update rules (3.17) are given in [22, 32].

5.5. **Convergence of Nesterov's differential equation.** Recall that the differential equation of Nesterov (3.18), presented in subsection 3.4.2, is given by

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

where $r > 0$. The condition 9 is not satisfied and the Theorems 4.4 and 4.7 can not be applied here. We are ready enunciate the main convergence result for Nesterov's differential equation. These results exist and can be found in [4, 5, 48].

Corollary 5.4 (Convergence Rate of Nesterov)**.** *Suppose that equation* (3.18) *satisfies assumptions 1, 8 and 11 (for example, if f is coercive C^2 and convex). Then $f(\theta) \to f(\theta_\star)$ when $t \to \infty$ with rate of convergence:*

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

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

Proof. The proof for $r \leq 3$ is given in subsection E.2. Therefore, we assume that $r \geq 3$ in here. We recall that:

$$h(t) = 1 \qquad r(t) = r/t \qquad p(t) = q(t) = 0$$

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 12 are verified. □

Acknowledgements. We would like to thank Daniel Panazzolo for a private communication concerning central-stable manifolds.

Appendix A. Preliminaries

This appendix gives some preliminary results used in the proof of the main theorems. We start this appendix by giving nonlinear versions of the Gronwall Lemma.

A.1. **Gronwall's Lemmas.**

Lemma A.1 (Gronwall's Lemma)**.** *Let $T > 0$, $\lambda \in L^1(0,T)$, $\lambda \geq 0$ almost everywhere and $C_1, C_2 \geq 0$. Let $\varphi \in L^1(0,T)$, $\varphi \geq 0$ almost everywhere, be such that $\lambda\varphi \in L^1(0,T)$ and*

$$\varphi(t) \leq C_1 + C_2 \int_0^t \lambda(s)\varphi(s)ds$$

for almost every $t \in (0,T)$. Then we have

$$\varphi(t) \leq C_1 \exp\left(C_2 \int_0^t \lambda(s)ds\right)$$

Lemma A.2. *Let $\varphi : [t_0, t_1]$ be absolutely continous stricly non-negative function and suppose φ obeys the differential inequality for $0 \leq \alpha \leq 1$*

$$\varphi'(t) \leq \beta(t)\varphi^\alpha(t)$$

for almost every $t \in [t_0, t_1]$, where β is continuous. Then for all $t \in [t_0, t_1]$ and all $0 \leq \alpha < 1$

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

If $\alpha = 1$ then

$$\varphi(t) \leq e^{\int_{t_0}^t \beta(s)ds}\varphi(t_0).$$

Proof.

$$[\varphi^{1-\alpha}]' = (1-\alpha)\varphi^{-\alpha}\varphi' \leq (1-\alpha)\varphi^{-\alpha}(t)\beta(t)\varphi^\alpha(t) = (1-\alpha)\beta(t)$$

Integrating over time gives

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

□

Lemma A.3. *Let $\varphi : [t_0, t_1]$ be absolutely continous stricly non-negative function and suppose φ obeys the differential inequality for $0 \leq \alpha < 1$*

$$\varphi'(t) \leq \gamma(t)\varphi(t) + \beta(t)\varphi^\alpha(t)$$

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.

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

and we conclude using the standard Gronwall Lemma applied to $\varphi^{1-\alpha}$. □

A.2. **A priori estimates under boundedness assumption.** In this section, we compute elementary bounds for the solutions of the ODE (3.1), under the additional assumption 8. These bounds are used in the proof of convergence Theorems 4.4 and 4.10. More precise bounds are studies in section B, where the assumption 8 is not verified.

Lemma A.4. *Let $t_0 > 0$ and $\mathbf{x}_0 = (\theta_0, m_0, v_0)$ be fixed. Under assumptions 1 and 8, there exists an unique solution $\mathbf{x}(t) = (\theta(t), m(t), v(t))$ of (3.1) 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, denoting by $L_g = \sup\{\|\nabla f(\theta(t))\| ; t \geq t_0\}$, we get:*

$$\begin{aligned}
\|m(t)\| &\leq \|m(t_0)\| + L_g d \int_{t_0}^t h(s)ds, \\
\|v(t)\| &\leq \|v(t_0)\| + L_g^2 d \int_{t_0}^t p(s)ds
\end{aligned} \tag{A.1}$$

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:

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

Proof. By assumption 1 and classical ODE's, there exists a solution $\mathbf{x}(t) = (\theta(t), m(t), v(t))$ of system (3.1) 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:

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

In particular, we easily conclude that

$$m(t) = \frac{1}{a(t)}\left(m_0 + \int_{t_0}^{t} a(s)h(s)\nabla f(\theta)ds\right)$$
$$v(t) = \frac{1}{b(t)}\left(v_0 + \int_{t_0}^{t} b(s)p(s)\nabla f(\theta)^2 ds\right)$$

Next, under assumption 8, we can assume that $|\nabla f(\theta(t))| \leq L_g$ for some positive real number L_g. It is now easy to get inequalities (A.1), which lead to:

$$|m_i(t)| \leq |m_i(t_0)| + L_g\, h(t_0)\,(t - t_0)$$
$$|v_i(t)| \leq |v_i(t_0)| + L_g^2\, p(t_0)\,(t - t_0)$$
$$|v_i(t)| \geq \frac{1}{b(t)} v_0 > 0$$

and we easily conclude that $T = \infty$. Finally, if $r(t) \not\equiv 0$, we get by direct integration:

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

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

Appendix B. Existence and uniqueness of solutions

B.1. **The Cauchy problem for $t_0 > 0$.** The proof of this result relies on a standard cut off argument to construct a local solution. This is done to handle the fact that the nonlinearities are not assumed to be globally Lipschitz (see Assumption 2). Global existence follows by standard applications of the Gronwall Lemma and of the alternative on the existence time. For consistency, we give the main steps of the proof.

We denote by $\mathbf{x}$ the vector function $\mathbf{x} = (\theta, m, v) \in \mathbb{R}^{3d}$ with $\theta, m, v \in \mathbb{R}^d$. Let $\Theta \in C_c^\infty(\mathbb{R})$ such that $\Theta \geq 0$ with support $\mathrm{supp}(\Theta) \in [0, 2]$ and $\Theta \equiv 1$

on $[0,1]$. For all $\ell \in \mathbb{N}_*$, we set $\Theta_\ell(x) = \Theta(x/\ell)$ and we denote

$$F_\ell(t, \mathbf{x}(t), \xi) = \begin{pmatrix} h(t)\Theta_\ell\left(\|\theta(t)\|^2\right)\nabla f(\theta(t), \xi(t)) - r(t)m(t) \\ p(t)\Theta_\ell\left(\|\theta(t)\|^2\right)\left[\nabla f(\theta(t), \xi(t))\right]^2 - q(t)v(t) \\ -\Theta_\ell\left(\|\mathbf{x}(t)\|^2\right)\frac{m(t)}{\sqrt{v(t)+\varepsilon}} \end{pmatrix}$$
$$= \begin{pmatrix} F_{1,\ell}(t, \mathbf{x}(t), \xi) \\ F_{2,\ell}(t, \mathbf{x}(t), \xi) \\ F_{3,\ell}(t, \mathbf{x}(t), \xi) \end{pmatrix}.$$

Then we consider the equation

$$\dot{\mathbf{x}}_\ell = F_\ell(t, \mathbf{x}(t), \xi)$$

and its equivalent integral formulation, by assumption 2,

$$\mathbf{x}_\ell(t) = \mathbf{x}_0 + \int_0^t F_\ell(s, \mathbf{x}_\ell(s), \xi(s))ds. \tag{B.1}$$

In the following proposition, we state that there is a unique solution with trajectories in $\mathcal{E}_c^T = C([t_0, T], \mathbb{R}^{3d})$, the space of continuous functions from $[t_0, T]$ into $\mathbb{R}^{3d}$. We endowed this space with the norm

$$\|\mathbf{x}\|_{\mathcal{E}_c^T} = \sup_{t\in[t_0,T]} \|\mathbf{x}(t)\|,$$

where $\|.\|$ denotes the euclidean norm in $\mathbb{R}^{3d}$.

Proposition B.1. *Assume Assumption 2 holds and let $T > t_0$; then Equation* (B.1) *has a unique solution $\mathbf{x}_\ell \in L^2(\Omega, \mathcal{E}_c^T)$.*

Proof of Proposition B.1. We prove this result by using a fixed point argument in the Banach space $L^2(\Omega, \mathcal{E}_c^T)$ for sufficiently small time T depending on both ℓ, δ and the initial condition. To do so we will need the following Lemma, whose proof is postponed in section B.4

Lemma B.2. *Assume Assumption 2 holds. Then the function $[\nabla f(\cdot, \xi)]^2$ is locally Lipschitz and for all $x, y \in \mathbb{R}^d$*

$$\left\|[\nabla f(x, \xi)]^2 - [\nabla f(y, \xi)]^2\right\| \leq M\left(\|x\|, \|y\|\right)\|x - y\|$$

where the positive constant M is given by

$$M\left(\|x\|, \|y\|\right) = \sqrt{2}L\left(\|x\|, \|y\|\right)\left(B + \max\left(L\left(\|x\|\right)\|x\|, L\left(\|y\|\right)\|y\|\right)\right).$$

Moreover the mapping $m/\sqrt{v+\varepsilon}$ is almost surely locally Lipschitz for all $\varepsilon > 0$.

For $\varepsilon = 0$, if v is a.s. lower bounded by a strictly positive constant then the mapping $m/\sqrt{v}$ is almost surely locally Lipschitz.

Note that from equation (3.1) and the choice of initial condition, the stochastic mapping v is a.s bounded from below. Indeed, almost surely and for all time

$$\dot{v} = p(t)[\nabla f(\theta, \xi)]^2 - q(t)v \geq -q(t)v,$$

and we obtain by an application of the Gronwall Lemma that

$$v(t) \geq e^{-\int_{t_0}^t q(s)ds} v_0.$$

Hence, it follows that

$$\left\| \frac{m}{\sqrt{v}} - \frac{n}{\sqrt{y}} \right\| \leq \frac{\|m-n\|}{v_0^{1/2}} e^{\frac{1}{2}\int_{t_0}^t q(s)ds} + \frac{\|n\,(y-v)\|}{2v_0^{3/2}} e^{\frac{3}{2}\int_{t_0}^t q(s)ds}.$$

Given $\mathbf{x}_\ell \in L^2\left(\Omega, \mathcal{E}_c^T\right)$, we denote by $\varphi(\mathbf{x}_\ell)$ the right hand side of Equation (B.1). From assumption 2 and Lemma B.2, we deduce that φ maps $L^2(\Omega, \mathcal{E}_c^T)$ into itself. Moreover, from assumption 2, Lemma B.2 and the same argument as in [17] for the cut-off, we obtain

$$\mathbb{E}\left(\|\varphi\left(\mathbf{x}_\ell(t)\right) - \varphi\left(\mathbf{y}_\ell(t)\right)\|_{\mathcal{E}_c^T}^2 \right)^{1/2} \leq TC(\ell, T)\mathbb{E}\left(\|\mathbf{x}_\ell(t) - \mathbf{y}_\ell(t)\|_{\mathcal{E}_c^T}^2 \right)^{1/2}.$$

where the constant C depends on the radius ℓ, the Lipschitz constant L and M of ∇f and ∇f^2, the final time T and the constant c. Hence, φ is a contraction mapping in $L^2(\Omega, \mathcal{E}_c^T)$ if T is chosen such that $TC(\ell, T) < 1$. We conclude that φ has a unique fixed point in $L^2(\Omega, \mathcal{E}_c^T)$ which is the unique solution to Equation (B.1) . Since the existence time only depends on ℓ, the solution can be extended globally for all $T > t_0$, which concludes the proof of Proposition B.1. □

Our aim is to prove global existence for the process $\mathbf{x}$, solution to equation (3.4), which will be constructed from the previous results. From the integral formulation, we can easily prove that τ_ℓ is non-decreasing with ℓ and $\mathbf{x}_\ell(t) = \mathbf{x}_{\ell'}(t)$ on $[0, \tau_\ell]$ for any $\ell < \ell'$. As a consequence, we set $\tau^* = \lim_{\ell \to +\infty} \tau_\ell$ and we define a local solution $\mathbf{x}$ to Equation (3.4) on $[t_0, \tau^*(\mathbf{x}(0)))$ by setting $\mathbf{x}(t) = \mathbf{x}_\ell(t)$ on $[t_0, \tau_\ell]$. By construction of τ^*, it is clear that the following alternative holds almost surely

$$\tau^*(\mathbf{x}(0)) = +\infty \tag{B.2}$$

$$\tau^*(\mathbf{x}(0)) < +\infty \quad \text{and} \quad \lim_{t \to \tau^*} \|\mathbf{x}_\ell(t)\| = +\infty \tag{B.3}$$

B.2. **A priori estimates and global solution.** Our aim is now to get global existence for the process $\mathbf{x}$ solution to (3.4). A key observation is given in the next lemma which states that the growth of the norm $m/\sqrt{v+\varepsilon}$ is uniform in ε.

Lemma B.3. *For any $s, t \in [t_0, T \wedge \tau_\ell]$ such that $t_0 \leq s \leq t \leq T \wedge \tau_\ell$*

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

Proof. The first inequality follows by integrating the equation in time. On the other hand, from the Young inequality

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

We notice that

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

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

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

and we easily conclude. □

Lemma B.4. *For all $t_0 \le s \le t \le T\wedge\tau_\ell$,*

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

Therefore

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

Proof. From the Cauchy-Schwarz inequality, we obtain

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

We apply Lemma A.2 to $\varphi(t) = \frac{1}{2}\|\theta(t)\|^2$, $\beta(t) = \sqrt{2}\left\|\frac{m}{\sqrt{v+\varepsilon}}\right\|$ and $\alpha = 1/2$. □

Lemma B.5. *For all $t_0 \le s \le t \le T \wedge \tau_\ell$, we have*

$$\|m(t)\|^2 \le \left[e^{-\int_{t_0}^t r(s)ds}\|m(t_0)\| + \int_{t_0}^t e^{-\int_s^t r(u)du} h(s)\|\nabla f(\theta,\xi)\|\,ds\right]^2.$$

Proof. From Cauchy Schwarz,

$$\begin{aligned}\frac{d}{dt}\frac{1}{2}\|m(t)\|^2 &= h(t)\langle m, \nabla f(\theta,\xi)\rangle - r(t)\|m\|^2 \\ &\le h(t)\|m\|\,\|\nabla f(\theta,\xi)\| - r(t)\|m\|^2\end{aligned}$$

Then solving the equation gives

$$\|m(t)\|^2 \le \left[e^{-\int_{t_0}^t r(s)ds}\|m(t_0)\| + \int_{t_0}^t e^{-\int_s^t r(u)du} h(s)\|\nabla f(\theta,\xi)\|\,ds\right]^2.$$

□

Lemma B.6. *For all $t_0 \le s \le t \le T \wedge \tau_\ell$, we have*

$$\left\|v^{1/2}(t)\right\|^2 \le e^{-\int_{t_0}^t q(s)ds}\left\|v^{1/2}(t_0)\right\|^2 + \int_{t_0}^t e^{-\int_s^t q(u)du} p(s)\|\nabla f(\theta,\xi)\|^2\,ds.$$

Proof. Note that

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

and the Lemma easily follows from the Gronwall lemma. □

From the alternative (B.2) and estimates in section B.2, we conclude that $\tau^* = +\infty$ and that there exists a unique solution $\mathbf{x}$ of (3.4) with path almost surely in $C^1((t_0,T];\mathbb{R}^n) \cap C([t_0,T];\mathbb{R}^n)$ for any $T > t_0$.

B.3. **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 (3.4) in the space $C^1((t_0,T];\mathbb{R}^n) \cap C([t_0,T];\mathbb{R}^n)$ for any strictly positive time t_0. The purpose of this section is to extend this result to solutions starting at $t = 0$. 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).

For that reason, we introduce a smoothed process $\mathbf{x}_\delta$, for $\delta > 0$, solution to the equation

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

where

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

and similar formulas hold for r_δ, p_δ and q_δ. Those functions are continous in time and from the previous section, it is obvious that there exists a global solution to this equation in $C^1((0,T];\mathbb{R}^n) \cap C([0,T];\mathbb{R}^n)$.

B.3.1. *Equicontinuity and uniform boundedness.* We now prove existence and uniqueness of system (3.4) using a compactness argument. 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 (B.4). Then applying the Arzela-Ascoli theorem, we extract a converging subsequence, and we prove that its limit is unique and satisfy (3.4). This is a standard argument in dynamical system and has been used for example in [48]. The key result is the following proposition which is proved in Appendix B.4

Proposition B.7. *Moreover assume Assumptions 3 and 4 are satisfied. Then there exists a positive constant $C_2(T)$, independent of δ, such that for all $t, s \in [0,T]$*

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

The proof is straightforward when none of the functions p, q, r, h have a singularity at zero. It will become apparent in the proof that the pole can only be of order one.

Before stating the next Lemma, we introduce 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 B.8. *Under the same assumptions as Lemma B.7, there exists a positive constant $C_3(T)$, independent of δ, such that for all $t, s \in [0,T]$*

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

for any $\gamma < 1$.

Proof. The proof is a consequence of Lemma B.7. Indeed

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

if $\gamma < 1$. □

We 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 $\mathbf{x}_\delta \in C^{\alpha}([0,T],\mathbb{R}^n)$ and is therefore equi-continuous and uniformly bounded. Applying Arzela Ascoli theorem, we deduce that there exists a converging subsequence (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 (3.4).

B.3.2. *Identification of the limit and uniqueness of the solution.* From Arzela-Ascoli, we can extract a sub-sequence converging to $\widehat{\mathbf{x}}$ in $C([0,T],\mathbb{R}^n)$. We prove in this section that $\widehat{\mathbf{x}}$ is the unique solution to the equation (3.4) with initial condition $\widehat{\mathbf{x}}_0 = \mathbf{x}_0$. The proof heavily relies on the fact that the function $\sqrt{v+\varepsilon}$ is lower bounded (under assumption 3) and we can deduce the uniform convergence of $1/\sqrt{v_\delta+\varepsilon}$ to $1/\sqrt{\widehat{v}+\varepsilon}$ from the uniform convergence of v_δ to $\widehat{v}$. The proof of existence and uniqueness also relies on the following lemma

Lemma B.9. *There exists two constants K_1 and K_2 such that for all $t \in [0,T]$ and all $\delta > 0$*

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

Existence. The convergence of the initial conditions are a direct consequence of the uniform convergence. For any $0 < t \leq T$, we can prove that $\widehat{\mathbf{x}}$ satisfies equation (3.4) from the convergence of $\mathbf{x}_\delta$ towards $\widehat{\mathbf{x}}$, assumption 2 and Lemma B.2 on the local Lipschitz property of the gradient and the squared gradient, Lemma B.9 and the following inequality

$$\left\| \frac{\widehat{m}}{\sqrt{\widehat{v}+\varepsilon}} - \frac{m_\delta}{\sqrt{v_\delta+\varepsilon}} \right\| \leq \left\| \frac{m_\delta - \widehat{m}}{\sqrt{\widehat{v}+\varepsilon}} \right\| + \left\| \frac{m_\delta}{\sqrt{v_\delta+\varepsilon}} \right\| \left\| \frac{\widehat{v} - v_\delta}{\sqrt{\widehat{v}+\varepsilon}\left(\sqrt{v_\delta+\varepsilon}+\sqrt{\widehat{v}+\varepsilon}\right)} \right\|.$$

Uniqueness. We proceed by contradiction. Assume there exist two solutions $\mathbf{x} = (\theta, m, v)$ and $\mathbf{y} = (\psi, n, y)$ to the system (3.4).

An easy computation shows that for all $0 \leq t \leq T$ (because v and y are a.s. lower bounded)

$$\begin{aligned} \|\theta(t) - \psi(t)\| &\leq \int_0^t \left\| \frac{m}{\sqrt{v+\varepsilon}} - \frac{n}{\sqrt{y+\varepsilon}} \right\| ds \\ &\leq C \int_0^t \|m-n\| + \|n(s)\| \, \|y-v\| \, ds \end{aligned}$$

By continuity of the solution of equation (3.1) 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

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

Now, consider the functions

$$a_\eta(t) = \exp\left(\int_\eta^t r(s)ds\right), \quad b_\eta(t) = \exp\left(\int_\eta^t q(s)ds\right)$$

which are increasing functions bigger than 1 (for all $t \geq \eta > 0$). We note that:

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

In particular, we easily conclude that

$$m(t) = \frac{1}{a_\eta(t)}\left(m(\eta) + \int_\eta^t a_\eta(s)h(s)\nabla f(\theta,\xi)ds\right)$$
$$v(t) = \frac{1}{b_\eta(t)}\left(v(\eta) + \int_\eta^t b_\eta(s)p(s)\nabla f(\theta,\xi)^2 ds\right)$$

It follows from Assumptions 2 and 3, inequality (B.5), and the fact that the Lipschitz constant L is bounded on bounded sets, that for all $\eta \leq t \leq T$,

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

By continuity of the process m and n, the fact that $m_0 = v_0$ and the continuity of $s \mapsto sh(s)$ on $[0,t]$, we obtain by taking the limit when η goes to zero that

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

Similarly there is a constant C_2 such that

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

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

$$\begin{aligned}
&\|m(t) - n(t)\| + \|v(t) - y(t)\| + \|\theta(t) - \psi(t)\| \\
&\qquad\qquad \leq C_1 t \sup_{0<u\leq t}\|m-n\| + C_2 t \sup_{0<u\leq t}\|y-v\|.
\end{aligned}$$

Since there exists a $t > 0$ such that $C_1 t$ and $C_2 t$ are strictly smaller than 1, we conclude on the uniqueness of the solution.

B.4. **Supporting proofs.**

Proof of Lemma B.2. We denote $\nabla f = g$. From the identity $a^2 - b^2 = (a-b)(a+b)$ and the Young inequality, we get

$$\begin{aligned}
&\left\|g^2(x,\xi) - g^2(y,\xi)\right\|^2 \\
&= \sum_{i=1}^{d} \left|g_i^2(x,\xi) - g_i^2(y,\xi)\right|^2 \\
&\leq 2\max\left(\|g(x,\xi)\|_\infty^2, \|g(y,\xi)\|_\infty^2\right) \|g(x,\xi) - g(y,\xi)\|^2
\end{aligned}$$

Therefore from the local Lipschitz assumption

$$\begin{aligned}
&\left\|g^2(x,\xi) - g^2(y,\xi)\right\| \\
&\leq \sqrt{2}L\left(\|x\|, \|y\|\right) \max\left(\|g(x,\xi)\|_\infty, \|g(y,\xi)\|_\infty\right) \|x-y\| \\
&\leq \sqrt{2}L\left(\|x\|, \|y\|\right) \left(\|g(0,\xi)\| + \max\left(L\left(\|x\|\right)\|x\|, L\left(\|y\|\right)\|y\|\right)\right) \|x-y\| \\
&\leq M\left(\|x\|, \|y\|\right) \|x-y\|
\end{aligned}$$

where

$$M\left(\|x\|, \|y\|\right) = \sqrt{2}L\left(\|x\|, \|y\|\right)\left(B + \max\left(L\left(\|x\|\right)\|x\|, L\left(\|y\|\right)\|y\|\right)\right).$$

□

Proof of Lemma B.9. From Lemma B.3 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 $\delta \leq 1$ we have:

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

Moreover from Lemma B.3 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 $\delta \leq 1$ we

have:

$$\left\|\frac{m_\delta(t)}{\sqrt{v_\delta(t)+\varepsilon}}\right\|^2 \leq e^{\int_\delta^t q_\delta(u)-2r_\delta(u)du}\left\|\frac{m_\delta(\delta)}{\sqrt{v_\delta(\delta)+\varepsilon}}\right\|^2 + d\int_\delta^t e^{\int_u^t q_\delta(a)-2r_\delta(a)da}\frac{h_\delta^2(u)}{p_\delta(u)}du$$

$$\leq \left\|\frac{m_\delta(\delta)}{\sqrt{v_\delta(\delta)+\varepsilon}}\right\|^2 + d\sup_{\delta<u<t}\frac{h_\delta(u)}{p_\delta(u)}\sup_{\delta<u<t}\left|\frac{h_\delta(u)}{q_\delta(u)-2r_\delta(u)}\right| \tag{B.7}$$

$$\leq \widetilde{K}_1\left\|\frac{m_\delta(\delta)}{\sqrt{v_\delta(\delta)+\varepsilon}}\right\|^2 + \widetilde{K}_2,$$

which (apart from increasing $\widetilde{K}_1$ and $\widetilde{K}_2$) combined with (B.7) yields a.s.:

$$\left\|\frac{m_\delta(t)}{\sqrt{v_\delta(t)+\varepsilon}}\right\| \leq \widetilde{K}_1\left\|\frac{m_0}{\sqrt{v_0+\varepsilon}}\right\| + \widetilde{K}_2.$$

□

Proof of Proposition B.7. The proof uses the integral formulation and weighted space. First, we define the following norm for all $0 < t \leq T$

$$N(t,\delta) = \mathbb{E}\left(\sup_{0<u\leq t}\|h_\delta(u)\nabla f(\theta_\delta(u),\xi(u)) - r_\delta(u)m_\delta(u)\|^2\right)^{1/2}$$
$$+ \mathbb{E}\left(\sup_{0<u\leq t}\left\|p_\delta(u)\left[\nabla f(\theta_\delta(u),\xi(u))\right]^2 - q_\delta(u)v_\delta(u)\right\|^2\right)^{1/2}$$
$$+ \mathbb{E}\left(\sup_{0<u\leq t}\left\|\frac{m_\delta(u)}{\sqrt{v_\delta(u)+\varepsilon}}\right\|^2\right)^{1/2}.$$

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 B.7 immediately follows from the claim and the following inequality

$$\mathbb{E}\left(\|\mathbf{x}_\delta(t)-\mathbf{x}_\delta(s)\|^2\right) \leq \mathbb{E}\left[\left(\int_s^t\|\dot{\mathbf{x}}_\delta(u)\|du\right)^2\right] \leq 3N(T,\delta)^2(t-s)^2.$$

We now turn to the proof of the claim.

The case $t \leq \delta$. For all $t \leq \delta$, the functions r_δ and q_δ are constant and the equations for m_δ and v_δ, given by system (B.4), have the equivalent Duhamel formulation given by

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

$$v_\delta(t) = e^{-tq_\delta(\delta)}v_0 + e^{-tq_\delta(\delta)}\int_0^t e^{uq_\delta(\delta)}p_\delta(\delta)\left[\nabla f(\theta_\delta(u),\xi(u))\right]^2 du. \tag{B.9}$$

From Lemma B.9, 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 a.s uniformly bounded so is $L\left(\|\theta_\delta(t)\|, \|\theta_0\|\right)$. Indeed,

$$\|\theta_\delta(t)-\theta_0\| \leq \int_0^t \left\| \frac{m_\delta(u)}{\sqrt{v_\delta(u)+\varepsilon}} \right\| du \leq t\left(K_1 \left\| \frac{m_0}{\sqrt{v_0+\varepsilon}} \right\| + K_2 \right) \tag{B.10}$$

Next, let us consider the first term which appears in $N(t,\delta)$. From the Duhamel formulation (B.8), the triangle inequality and the fact that the initial condition $m_0 = \nabla f(\theta_0,\xi_0)\lim_{t\to 0^+} h(t)/r(t)$, we obtain an upper bound of the form

$$\|h_\delta(\delta)\nabla f(\theta_\delta(t),\xi(t)) - r_\delta(\delta)m_\delta(t)\| \leq N_1+N_2+N_3+N_4+N_5$$

where:

$$\begin{aligned}
N_1 &= \|h_\delta(\delta)\left(\nabla f(\theta_\delta(t),\xi(t)) - \nabla f(\theta_0,\xi(t))\right)\| \\
N_2 &= \|h_\delta(\delta)\left(\nabla f(\theta_0,\xi(t)) - \nabla f(\theta_0,\xi_0)\right)\| \\
N_3 &= \left\| r_\delta(\delta)e^{-tr_\delta(\delta)}\left(m_0 - \frac{h_\delta(\delta)}{r_\delta(\delta)}\nabla f(\theta_0,\xi_0)\right)\right\| \\
N_4 &= \left\| r_\delta(\delta)e^{-tr_\delta(\delta)}\int_0^t e^{ur_\delta(\delta)}h_\delta(\delta)\left(\nabla f(\theta_\delta(u),\xi(u)) - \nabla f(\theta_0,\xi(u))\right)du\right\| \\
N_5 &= \left\| r_\delta(\delta)e^{-tr_\delta(\delta)}\int_0^t e^{ur_\delta(\delta)}h_\delta(\delta)\left(\nabla f(\theta_0,\xi(u)) - \nabla f(\theta_0,\xi_0)\right)du\right\|.
\end{aligned}$$

We now show that each one of these terms are bounded uniformly in terms of δ. The term N_1 is bounded by the local Lipschitz assumption 2 of the gradient, inequality (B.10), the Duhamel formula (B.8) and Lemma B.9

$$N_1 \leq \delta h_\delta(\delta) L\left(\|\theta_\delta(t)\|, \|\theta_0\|\right)\left(K_1\left\|\frac{m_0}{\sqrt{v_0+\varepsilon}}\right\|^2 + K_2\right),$$

and we easily conclude that N_1 is uniformly bounded by assumption 3. From the variance assumption 4, we can bound the term N_2 as follows

$$\mathbb{E}(N_2) \leq \mathbb{E}\left(\sup_{0<u\leq\delta} \|h(u)\left(\nabla f(\theta_0,\xi(u)) - \nabla f(\theta_0,\xi_0)\right)\|^2\right)^{1/2} \leq \sigma_1^{1/2}.$$

The term N_3 is bounded from the choice of the initial condition and the fact that $h(t)/r(t)$ is a C^1 function. More precisely

$$N_3 \leq \delta r_\delta(\delta)e^{-tr_\delta(\delta)}\|\nabla f(\theta_0,\xi_0)\|\left|\frac{h_\delta(\delta)}{r_\delta(\delta)} - \lim_{t\to 0}\frac{h(t)}{r(t)}\right|\delta^{-1}.$$

The term N_4 is bounded in a similar way as N_1 using assumption 3, inequality (B.10) and Lemma B.9

$$N_4 \leq r_\delta(\delta) h_\delta(\delta) L\left(\|\theta_\delta(t)\|, \|\theta_0\|\right) \left(K_1 \left\| \frac{m_0}{\sqrt{v_0+\varepsilon}} \right\|^2 + K_2 \right) \int_0^t u \, du$$
$$\leq \frac{\delta^2 r_\delta(\delta) h_\delta(\delta)}{2} L\left(\|\theta_\delta(t)\|, \|\theta_0\|\right) \left(K_1 \left\| \frac{m_0}{\sqrt{v_0+\varepsilon}} \right\|^2 + K_2 \right).$$

Finally, the term N_5 is bounded using assumption 4, in a similar way than N_2. Indeed

$$\mathbb{E}(N_5) \leq \mathbb{E}\left(r_\delta(\delta) \sup_{0\leq u\leq t} \|h_\delta(u) \left(\nabla f(\theta_0, \xi(u)) - \nabla f(\theta_0, \xi_0)\right)\| \int_0^t ds \right)$$
$$\leq \mathbb{E}\left(\delta r_\delta(\delta) \sup_{0\leq u\leq t} \|h(u) \left(\nabla f(\theta_0, \xi(u)) - \nabla f(\theta_0, \xi_0)\right)\| \right)$$
$$\leq \delta r_\delta(\delta) \sigma_1^{1/2},$$

which is bounded by assumption 3. Gathering all bounds, we easily conclude that there exists a constant C_1 such that, for every $\delta \leq 1$:

$$\mathbb{E}\left(\sup_{0<u\leq t} \|h_\delta(u) \nabla f(\theta_\delta(u), \xi(u)) - r_\delta(u) m_\delta(u)\|^2 \right)^{1/2} \leq C_1.$$

From a similar argument, we obtain that there exists a constant C_2 such that, for every $\delta \leq 1$:

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

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 B.9. We omit the details here.

□

Appendix C. Convergence of the Euler discretization

C.1. **Proof of Theorem 4.2.** The result from [25] can be adapted to random differential equations and we check that the two assumptions are satisfied for (3.4) and (3.5). We start by a proposition, concerning the strict positivity of v_k, up to numerical approximation, as long as $v_0 > 0$.

Proposition C.1. *For all $k = 0, \cdots, K-1$*

$$m_{k+1} = \sum_{i=0}^{k} s h_{i+1} \nabla f(\theta_{i+1}, \xi_{i+1}) \prod_{j=i+1}^{k} (1 - s r_{j+1}) + \prod_{j=0}^{k} (1 - s r_{j+1}) m_0$$

and

$$v_{k+1} = \sum_{i=0}^{k} sp_{i+1}[\nabla f(\theta_{i+1}, \xi_{i+1})]^2 \prod_{j=i+1}^{k} (1 - sq_{j+1}) + \prod_{j=0}^{k}(1 - sq_{j+1})v_0,$$

where we used the notation $h_i = h(t_i)$ (and similarly for the other functions). Let assume that the learning rate s satisfies $sr_1 < 1$ and $sq_1 < 1$. Hence, the numerical scheme preserves the strict positivity of v i.e if we assume $v_0 > 0$, then for all $k = 0, \cdots, K$, $v_k > 0$.

Proof. The proof is simply based on induction and the iterative formula for m and v. □

The next proposition shows that the quantity $\left\|m_k/\sqrt{v_k}\right\|$ is bounded for finite number of iterations K. Therefore the discrete solution has the same properties as the solution of the continuous system.

Proposition C.2. *Let assume that the learning rate s satisfies $sr_1 < 1$ and $sq_1 < 1$. For all $k = 0, \cdots, K-1$, the following bound holds*

$$\begin{aligned}\left\|\frac{m_{k+1}}{\sqrt{v_{k+1}}}\right\|^2 \leq \left\|\frac{m_0}{\sqrt{v_0}}\right\|^2 \prod_{i=0}^{k} \max\left(\frac{(1-sr_{i+1})^2}{(1-sq_{i+1})}, 1\right) \\ + sd\sum_{i=0}^{k} \frac{h_{i+1}^2}{p_{i+1}} \prod_{j=i}^{k} \max\left(\frac{(1-sr_{j+1})^2}{(1-sq_{j+1})}, 1\right).\end{aligned}$$

Moreover

$$\prod_{i=0}^{K-1} \frac{(1-sr_{i+1})^2}{(1-sq_{i+1})} \leq \exp\left(-2\int_{t_1}^{T} r(t)dt + \frac{1}{1-sq_1}\int_{t_0}^{T} q(t)dt\right).$$

Proof. From the discrete updates for m_{k+1} and v_{k+1} given by equation (3.3), we easily observe that the following identity holds true

$$\begin{aligned}&m_{j,k+1}^2 v_{j,k} - m_{j,k}^2 v_{j,k+1} \\ &= \left((1-sr_{k+1})^2 m_{j,k}^2 + s^2 h_{k+1}^2 [\partial_j f_{k+1}]^2 + 2sh_{k+1}(1-sr_{k+1})m_{j,k}\partial_j f_{k+1}\right) v_{j,k} \\ &- m_{j,k}^2(1-sq_{k+1})v_{j,k} - sp_{k+1}m_{j,k}^2[\partial_j f_{k+1}]^2 \\ &= v_{j,k}\left((1-sr_{k+1})^2 - (1-sq_{k+1})\right)\left(m_{j,k}^2 + s\frac{h_{k+1}^2}{p_{k+1}}v_{j,k}\right) \\ &+ s\frac{h_{k+1}^2}{p_{k+1}}v_{j,k+1}v_{j,k} - sp_{k+1}\left(m_{j,k}\partial_j f_{k+1} - \frac{h_{k+1}}{p_{k+1}}(1-sr_{k+1})v_{j,k}\right)^2.\end{aligned}$$

Thus, dividing both side of the previous equality by $v_{j,k+1}v_{j,k}$, we obtain for all $k \geq 0$

$$\left\| \frac{m_{k+1}}{\sqrt{v_{k+1}}} \right\|^2 - \left\| \frac{m_k}{\sqrt{v_k}} \right\|^2 = \sum_{j=1}^{d} \frac{m_{j,k+1}^2}{v_{j,k+1}} - \frac{m_{j,k}^2}{v_{j,k}}$$
$$\leq \sum_{j=1}^{d} \frac{(1-sr_{k+1})^2 - (1-sq_{k+1})}{v_{j,k+1}} \left(m_{j,k}^2 + sv_{j,k}\frac{h_{k+1}^2}{p_{k+1}} \right) + sd\frac{h_{k+1}^2}{p_{k+1}}.$$

We consider two different cases: if $(1-sr_{k+1})^2 - (1-sq_{k+1}) \leq 0$ then,

$$\left\| \frac{m_{k+1}}{\sqrt{v_{k+1}}} \right\|^2 \leq \left\| \frac{m_k}{\sqrt{v_k}} \right\|^2 + sd\frac{h_{k+1}^2}{p_{k+1}}.$$

On the other hand, if $(1-sr_{k+1})^2 - (1-sq_{k+1}) \geq 0$, then from the update rule for v_{k+1}, given by equation (3.3), and the fact that $v_{k+1} \geq (1-sq_{k+1})v_k$, we get

$$\left\| \frac{m_{k+1}}{\sqrt{v_{k+1}}} \right\|^2 - \left\| \frac{m_k}{\sqrt{v_k}} \right\|^2$$
$$\leq \sum_{j=1}^{d} \frac{(1-sr_{k+1})^2 - (1-sq_{k+1})}{(1-sq_{k+1})v_{j,k}} \left(m_{j,k}^2 + sv_{j,k}\frac{h_{k+1}^2}{p_{k+1}} \right) + sd\frac{h_{k+1}^2}{p_{k+1}}$$
$$= \frac{(1-sr_{k+1})^2 - (1-sq_{k+1})}{(1-sq_{k+1})} \left(\left\| \frac{m_k}{\sqrt{v_k}} \right\|^2 + sd\frac{h_{k+1}^2}{p_{k+1}} \right) + sd\frac{h_{k+1}^2}{p_{k+1}}.$$

Thus

$$\left\| \frac{m_{k+1}}{\sqrt{v_{k+1}}} \right\|^2 \leq \frac{(1-sr_{k+1})^2}{(1-sq_{k+1})} \left(\left\| \frac{m_k}{\sqrt{v_k}} \right\|^2 + sd\frac{h_{k+1}^2}{p_{k+1}} \right)$$

Combining the upper bounds obtained in the two cases

$$\left\| \frac{m_{k+1}}{\sqrt{v_{k+1}}} \right\|^2 \leq \max\left(\frac{(1-sr_{k+1})^2}{(1-sq_{k+1})}, 1 \right) \left(\left\| \frac{m_k}{\sqrt{v_k}} \right\|^2 + sd\frac{h_{k+1}^2}{p_{k+1}} \right)$$

By induction we get that

$$\left\| \frac{m_{k+1}}{\sqrt{v_{k+1}}} \right\|^2 \leq \left\| \frac{m_0}{\sqrt{v_0}} \right\|^2 \prod_{i=0}^{k} \max\left(\frac{(1-sr_{i+1})^2}{(1-sq_{i+1})}, 1 \right)$$
$$+ sd \sum_{i=0}^{k} \frac{h_{i+1}^2}{p_{i+1}} \prod_{j=i}^{k} \max\left(\frac{(1-sr_{j+1})^2}{(1-sq_{j+1})}, 1 \right).$$

We now prove that the previous upper bound is bounded by a constant depending only on the final time T. For all $x \in (0,1)$, we know that $-x \geq$

$\log(1-x) \geq -x/(1-x)$. Therefore

$$\log\left(\prod_{i=0}^{k} \frac{(1-sr_{i+1})^2}{(1-sq_{i+1})}\right) = \sum_{i=0}^{k} 2\log(1-sr_{i+1}) - \log(1-sq_{i+1})$$
$$\leq \sum_{i=0}^{k} -2sr_{i+1} + \frac{sq_{i+1}}{1-sq_1}.$$

From assumption 3, the functions q and r are non-increasing. Then for all $t \in [t_i, t_{i+1}]$, we have $q_i \geq q(t) \geq q_{i+1}$ and similarly for r. Integrating over $[t_i, t_{i+1}]$ and summing from zero to $K-1$ gives

$$\int_{t_0}^{T} q(t)dt \geq \sum_{i=0}^{K-1} sq_{i+1}.$$

Similarly integrating over $[t_i, t_{i+1}]$ and summing from one to $K-1$ gives

$$\int_{t_1}^{T} r(t)dt \leq \sum_{i=1}^{K-1} sr_i \leq \sum_{i=0}^{K-1} sr_{i+1}.$$

We conclude that

$$\prod_{i=0}^{K-1} \frac{(1-sr_{i+1})^2}{(1-sq_{i+1})} \leq \exp\left(-2\int_{t_1}^{T} r(t)dt + \frac{1}{1-sq_1}\int_{t_0}^{T} q(t)dt\right).$$

□

Finally, from the previous estimates, we obtain a moment bound on θ. This estimate is important since the bound only depends on the final time T, the constant p and the norm of the initial solution but not on the norm of θ itself.

Proposition C.3. *For all $k = 0, \cdots, K-1$*

$$\|\theta_{k+1}\| \leq \|\theta_0\| + s\sum_{i=0}^{k} \left\| \frac{m_i}{\sqrt{v_i+\varepsilon}} \right\|.$$

Then, from Proposition C.2, we conclude that there exists a constant $C(T) > 0$, such that for all s

$$\sup_{0\leq k\leq K_{T,s}} \|\theta_k\| \leq C(T)(1+\|\mathbf{x}_0\|).$$

Using the previous estimates given by Propositions C.1, C.2 and C.3 and the locally Lispchitz assumption 2, we obtain the following bounds for the solution of the numerical scheme (3.3).

Proposition C.4 (Bounds for the solution of the numerical scheme)**.** *Let $A_0 \subset \mathbb{R}^d_{\geq 0} \times \mathbb{R}^d_{\geq 0} \times \mathbb{R}^d_{>0}$ be a compact set. There exists a constant $C(T, A_0) > 0$, such that for all s and all initial condition $\mathbf{x}_0 \in A_0$*

$$\sup_{0\leq k\leq K_{T,s}} \|\mathbf{x}_k\| \leq C(T, A_0)\left(1+\|\mathbf{x}_0\|\right).$$

The proof follows [23]. We denote by $B_\ell = \left\{x \in \mathbb{R}^{3d}; \|x\| \leq \ell\right\}$. From section B.2 and Proposition (C.4), we know that there exists a constant ℓ such that $\mathbf{x}(t)$ and $\mathbf{x}_k$ remain in B_ℓ for all $t \in [0,T]$ and $k = 0, \cdots, K$. Moreover we considered the numerical approximation only for $t_0 > 0$ and therefore all functions p, q, r, h are continuously differentiable. The global error is controlled using the integral formulation and decomposing the error as follows

$$\begin{aligned}
&\int_{t_k}^{t_{k+1}} h(s)\nabla f(\theta(s),\xi(s)) - h(t_{k+1})\nabla f(\theta_{k+1},\xi_{k+1})ds \\
&\qquad = \int_{t_k}^{t_{k+1}} \left(h(s) - h(t_{k+1})\right)\nabla f(\theta(s),\xi(s))ds \\
&\qquad + \int_{t_k}^{t_{k+1}} h(t_{k+1})\left(\nabla f(\theta(s),\xi(s)) - \nabla f(\theta(s),\xi_{k+1})\right)ds \\
&\qquad + \int_{t_k}^{t_{k+1}} h(t_{k+1})\left(\nabla f(\theta(s),\xi_{k+1}) - \nabla f(\theta_{k+1},\xi_{k+1})\right)ds.
\end{aligned}$$

and

$$\begin{aligned}
\int_{t_k}^{t_{k+1}} \left\| \frac{m(s)}{\sqrt{v(s)+\varepsilon}} - \frac{m_k}{\sqrt{v_k+\varepsilon}} \right\| ds &\leq \int_{t_k}^{t_{k+1}} \left\| \frac{m_k - m(s)}{\sqrt{v(s)+\varepsilon}} \right\| ds \\
&+ \int_{t_k}^{t_{k+1}} \left\| \frac{m_k}{\sqrt{v_k+\varepsilon}} \right\| \left\| \frac{v(s)-v_k}{\sqrt{v(s)+\varepsilon}\left(\sqrt{v_k+\varepsilon}+\sqrt{v(s)+\varepsilon}\right)} \right\| ds.
\end{aligned}$$

We conclude the proof using again the integral formulation, assumptions 5 and 6, assumption 2 and Proposition C.2.

Appendix D. Exploring an energy functional

We consider the following energy functional, which is inspired from [2, Theorem 2.1]:

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

where we recall the notation:

$$\frac{m}{[v+\varepsilon]^{1/4}} = \left\langle \frac{m_1}{(v_1+\varepsilon)^{1/4}}, \ldots, \frac{m_d}{(v_d+\varepsilon)^{1/4}} \right\rangle.$$

The derivative of $E(t,\theta,v)$ can be easily computed:

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

and it is easy to see that

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

Therefore, under the assumption 7, the derivative of $E(t,\theta,v)$ is non-positive. In particular, this yields the following results:

Lemma D.1. *Let $\varepsilon \geq 0$. Suppose that assumptions 1 and 7 are verified. Then there exists $T \geq t_0$ such that for all $t \geq T$:*

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

In particular, if f is coercive, then assumption 8 is satisfied.

Proof. The proof is immediate from the fact that the energy has non-positive derivative. □

D.1. **Proof of Theorem 4.4.** Consider the autonomous system associated to (3.1), that is, the following vector field defined in $\mathbb{R}^{3d+1}$:

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

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 9). In order to study the convergence of the vector field when $t\to\infty$, we perform the change of coordinates $s = 1/t$ which yields:

$$\partial = -s^2\partial_s + \sum_{i=1}^{d} -\frac{m_i}{\sqrt{v_i+\varepsilon}}\partial_{\theta_i} + (H(s)\partial_{\theta_i}f(\theta) - R(s)m_i)\partial_{m_i} + (P(s)\partial_{\theta_i}f(\theta)^2 - Q(s)v_i)\partial_{v_i}. \tag{D.3}$$

Now, let us fix an orbit $x(t) = (\theta(t),m(t),v(t),s(t))$ with initial conditions $x(t_0) = (\theta(t_0),m(t_0),v(t_0),1/t_0)$. By the Lemma A.4, we know that $x(t)$ is bounded and $v(t) > 0$ for all $t \in (t_0,\infty)$. Denote by $\omega(x(t))$ (which stands for ω-limit) the topological limit of $x(t)$, that is:

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

It is well-known that the ω-limit of an orbit is a closed (therefore, under assumption 8, a non-empty, compact, connected) and invariant set by ∂. It follows from the expression of ∂ that $\omega(x(t)) \subset (s = 0)$.

We are ready to consider the energy functional E given in (D.1). More precisely, consider the functional:

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

and we note that

$$\partial(\widetilde{E}(s, \theta, m, v)) = \frac{d}{dt}\widetilde{E}(s, \theta, m, v).$$

It follows from a similar computation as the one performed in the last subsection that:

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

which is everywhere non-positive by assumption 7 and 9. Now, since $E(x(t))$ is bounded from below (because E is continuous and $x(t)$ is bounded), we conlude that the limit:

$$\lim_{t \to \infty} \widetilde{E}(x(t)) = \widetilde{E}_\infty$$

exists. In particular, we conclude that $\omega(x(t)) \subset (\widetilde{E}(s, \theta, m, v) = \widetilde{E}_\infty)$, which implies that $\omega(x(t))$ must be contained in the set of zero derivative of $\widetilde{E}(s, \theta, m, v)$. By assumption 9 this implies that $\omega(x(t)) \subset (m = 0)$. Since $H(0) \neq 0$ (assumption 9), by the expression of ∂, we conclude that $\omega(x(t)) \subset (\nabla f(\theta) = 0)$. Finally, since either $P(s) \equiv Q(s) \equiv 0$ or $Q(0) > 0$(assumption 9), by the expression of ∂, we conclude that $\omega(x(t)) \subset (v = 0)$. We conclude easily.

D.2. **Proof of Theorem 4.7.** We start by enunciating the main result from center-manifold theory, which is a local version of [45, Ch. 1 Thm 4.2] by using the cut-off technique given in [45, Ch. 1 Lem. 3.1]; c.f. [45, Ch. 1, Thm 1.1 and 3.2]:

Theorem D.2. *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 neighborhood 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 contains the origin 0 and has dimension at most $n-1$;*
3. *If $x_0 \in U \backslash \Sigma$, then there exists $\widetilde{t}_0 > t_0$ such that $x(\widetilde{t}_0) \notin U$, where $x(t)$ denotes the solution of the differential equation with initial condition $x(t_0) = x_0$.*

We now proceed with the proof of Theorem 4.7. Recall the vector field ∂ defined in (D.3), which describes the ODE (3.1). We consider the set:

$$C = \{\theta_\star \in \mathbb{R}^d;\, \nabla f(\theta_\star) = 0, \text{ and } \theta_\star \text{ is a strict saddle of } f\}$$

By assumption 10, the set A is a countable union of isolated points of $\mathbb{R}^d$. It follows from Theorem 4.4 that the set:

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

is a countable union of isolated points, all of each have the form $(\theta_\star, 0, 0, 0)$, where $\theta_\star \in C$. We now consider the set:

$$S := \{(\theta_0, m_0, v_0, s_0);\, \text{the } \omega\text{-limit of } (\theta(t), m(t), v(t), s(t)) \subset B\}$$

where $(\theta(t), m(t), v(t), s(t))$ have initial condition equal to $(\theta_0, m_0, v_0, s_0)$.

We now make a local argument valid for each singular point in B in order to show that S is locally a manifold; indeed, fix $(\theta_\star, 0, 0, 0) \in S$. Consider the Jacobean of ∂ at the singular point $(\theta_\star, 0, 0, 0)$, which is the $3d+1$ square matrix:

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

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:

$$\eta_i = -\frac{\varepsilon^{1/2}}{H(0)}(R(0) + \lambda)\lambda, \quad i = 1, \ldots, d \tag{D.4}$$

where $\{\eta_1, \ldots, \eta_d\}$ are the eigenvalues of $\mathcal{H}_f(\theta_\star)$. By assumption 10, we can suppose without loss of generality that $\eta_1 < 0$, and we easily conclude by equation (D.4) that there exists one strictly positive eigenvalue λ of $Jac(\partial)(\theta_\star, 0, 0, 0)$. By Theorem D.2, there exists an open neighborhood $U_{\theta_\star}$ of $(\theta_\star, 0, 0, 0)$ and a C^1 manifold $\Sigma_{\theta_\star} \subset U_{\theta_\star}$ such that every orbit $(\theta(t), m(t), v(t), s(t))$ with initial condition in $U_{\theta_\star} \setminus \Sigma$, leaves $U_{\theta_\star}$ in finite time. In particular, since $(\theta_\star, 0, 0, 0)$ is an isolated critical point and the ω-limit of an arbitrary orbit $(\theta(t), m(t), v(t), s(t))$ is connected, we conclude that $S \cap U_{\theta_\star} \subset \Sigma_{\theta_\star}$ (otherwise, the ω-limit would contain points in the border of U). Now, consider the set Σ given by the union of all orbits with initial conditions in $\Sigma_{\theta_\star}$, for some $\theta_\star \in C$. It easily follows that $S \subset \Sigma$. Since C is a countable set, we conclude that the Hausdorff dimension of Σ is at most $3d$.

Finally, let $t_0 > 0$ be fixed and denote by $S_{t_0} = S \cap \{s = 1/t_0\}$. Note that $S_{t_0} \subset \Sigma_{t_0}$, where $\Sigma_{t_0} = \Sigma \cap \{s = 1/t_0\}$. Now, Σ has Hausdorff dimension $3d$ and contains orbits of ∂, all of each are transverse to the set $\{s = 1/t_0\}$.

It follows that the Hausdorff dimension of Σ_{t_0} is at most $3d-1$, and we conclude easily.

Appendix E. Exploring a (generalized) Lyapunov

E.1. **Proof of Theorem 4.10.** Let $\theta_\star$ be a minimum point of f (which exists by Assumption 11). Let us consider the following energy functional

$$\mathcal{E}(t,m,v,\theta) = \mathcal{E}_1(t,\theta) + \mathcal{E}_2(t,m,v,\theta)$$

where

$$\mathcal{E}_1(t,\theta) = \mathcal{A}(t)\left(f(\theta) - f(\theta_\star)\right)$$

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

This functional is used as a Lyapunov function to prove convergence to a neighborhood of the global minimum. We first compute its time derivative and we find conditions on the functions $\mathcal{B}$ 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. 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 $\mathcal{E}_2$.

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

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

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

By adding all of the above computations, we get that:

$$\begin{aligned}
&0 \leq \mathcal{E}_1(t,\theta),\, \mathcal{E}_2(t,m,v,\theta) \\
&\frac{d}{dt}\mathcal{E}(t,m,v,\theta) \leq \frac{p(t)}{4}\left\langle \frac{[\nabla f(\theta)]^2}{[v+\varepsilon]^{1/2}}\odot(\theta-\theta_\star), \theta-\theta_\star\right\rangle,
\end{aligned}$$

if 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 assumption 12 implies that all above conditions are satisfied. It is now immediate from the Fundamental Theorem of calculus (and the fact that $\mathcal{E}_2 \geq 0$) that:

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

Next, consider the constant:

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

This constant is finite by Lemma A.4. We now 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 (3.1) 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 result now follows from the fact that $v(t)$ is bounded by Lemma A.4.

E.2. **Proof of Corollary 5.4.** 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 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 which are inaccessible in the previous section. Indeed, using this more general energy functional, we prove a convergence result for Nesterov when $0 < r < 3$:

End of proof of Corollary 5.4. Let $t_0 = 1$ and $\mathcal{D}(t) = t^{-\alpha}$ for some positive α which satisfies $2 > \alpha > 1 - r$. Then:

$$\mathcal{B}(t) = t^r\int_{t}^{\infty} s^{-r-\alpha} = \frac{t^{1-\alpha}}{r+\alpha-1}$$

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

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. □

Appendix F. Morse functions

Let us briefly recall some basic notions about Morse functions. We follow [36]. We start by recalling the definition of degenerate critical point:

Definition F.1. A critical point x_0 of a C^2 function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be degenerated if the determinant of the Hessian matrix $\mathcal{H}_f(x)$ of f at x_0 is zero. Otherwise, x_0 is said to be non-degenerated.

We are now ready to define the notion of a Morse function:

Definition F.2. A C^2 function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be *Morse* if all of the critical points of f are non-degenerated.

The following is well-known properties of Morse functions:

Remark F.3 (On Morse functions).

(i) Let 0 be a critical point of a Morse function $f(\theta)$. The Morse Lemma states that there exists a (locally defined) coordinate system $x = (x_1, \ldots, x_d)$, and a number $0 \leq r \leq d$ such that:

$$f(\theta) = x_1^2 + \ldots + x_r^2 - x_{r+1}^2 - \ldots - x_d^2$$

in particular, every local minimum of a Morse function f is locally convex.

(ii) Consider the space $C^\infty(M, \mathbb{R})$ of all C^∞ functions $f : M \to \mathbb{R}$, where M is a compact smooth manifold. Almost every function $f \in C^\infty(M, \mathbb{R})$ is Morse and, therefore, satisfies assumption 10. More precisely, there exists a set $U \subset C^\infty(M, \mathbb{R})$, which is open and dense in the Whitney C^∞-topology, such that every $f \in U$ is Morse.

(iii) Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function in $C^\infty(\mathbb{R}^d, \mathbb{R})$ and fix a compact set $K \subset \mathbb{R}^d$. Then there exists a set $U_K \subset C^\infty(\mathbb{R}^d, \mathbb{R})$, which is open and dense in the Whitney C^∞-topology, such that for every $f \in U_K$, the restriction $f|_K$ is Morse.

References

[1] Cabot A., *Asymptotics for a gradient system with memory term*, Proceedings of the American Mathematical Society **137** (2009), no. 9, 3013–3024.

[2] F. Alvarez, *On the minimizing property of a second order dissipative system in hilbert spaces*, SIAM Journal on Control and Optimization **38** (2000), no. 4, 1102–1119.

[3] 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** (2002), no. 8, 747–779.

[4] H. Attouch, Z. Chbani, and H. Riahi, *Rate of convergence of the Nesterov accelerated gradient method in the subcritical case* $\alpha \leq 3$, to appear in ESAIM: Control, Optimisation and Calculus of Variations (2017), available at `2017083`.

[5] Hedy Attouch, Zaki Chbani, Juan Peypouquet, and Patrick Redont, *Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity*, Math. Program. **168** (March 2018), no. 1-2, 123–175.

[6] A. Basu, S. De, A. Mukherjee, and E. Ullah, *Convergence guarantees for RMSProp and ADAM in non-convex optimization and their comparison to Nesterov acceleration on autoencoders*, ArXiv e-prints (July 2018), available at `1807.06766`.

[7] M. Benaïm, *Dynamics of stochastic approximation algorithms*, Séminaire de probabilités xxxiii, 1999, pp. 1–68.

[8] M. Betancourt, M. I. Jordan, and A. C. Wilson, *On Symplectic Optimization*, ArXiv e-prints (February 2018), available at `1802.03653`.

[9] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical optimization: Theoretical and practical aspects (universitext)*, Springer-Verlag, Berlin, Heidelberg, 2006.

[10] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization Methods for Large-Scale Machine Learning*, ArXiv e-prints (June 2016), available at `1606.04838`.

[11] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, New York, NY, USA, 2004.

[12] S. Bubeck, *Convex optimization: Algorithms and complexity*, Found. Trends Mach. Learn. **8** (November 2015), no. 3-4, 231–357.

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

[14] A. Cabot, H. Engler, and S. Gadat, *On the Long Time Behavior of Second Order Differential Equations with Asymptotically Small Dissipation*, ArXiv e-prints (October 2007), available at `0710.1107`.

[15] A. L. Cauchy, *Methode generale pour la resolution des systemes d'equations simultanees*, Comptes Rendus de l'Academie des Science **25** (1847), 536–538.

[16] X. Chen, S. Liu, R. Sun, and M. Hong, *On the Convergence of A Class of Adam-Type Algorithms for Non-Convex Optimization*, ArXiv e-prints (2018), available at `1808.02941`.

[17] A. de Bouard and A. Debussche, *A stochastic nonlinear schrödinger equation with multiplicative noise*, Communications in Mathematical Physics **205** (1999Aug), no. 1, 161–181.

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

[19] J. Duchi and Y. Singer, *Proximal and first-order methods for convex optimization* (January 2013).

[20] C. S. Duris and J. N. Lyness, *Compound quadrature rules for the product of two functions*, SIAM Journal on Numerical Analysis **12** (1975), no. 5, 681–697.

[21] S. Gadat and F. Panloup, *Long time behaviour and stationary regime of memory gradient diffusions*, Ann. Inst. H. Poincar Probab. Statist. **50** (201405), no. 2, 564–601.

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

[23] L. Grüne and P. E. Kloeden, *Pathwise approximation of random ordinary differential equations*, BIT Numerical Mathematics **41** (2001Sep), no. 4, 711–721.

[24] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations; 2nd ed.*, Springer, Dordrecht, 2006.

[25] D. Higham, X. Mao, and A. Stuart, *Strong convergence of euler-type methods for nonlinear stochastic differential equations*, SIAM Journal on Numerical Analysis **40** (2002), no. 3, 1041–1063.

[26] B. Hu and L. Lessard, *Dissipativity Theory for Nesterov's Accelerated Method*, ArXiv e-prints (June 2017), available at `1706.04381`.

[27] J. Kiefer and J. Wolfowitz, *Stochastic estimation of the maximum of a regression function*, Ann. Math. Statist. **23** (195209), no. 3, 462–466.

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

[29] P. Kloeden and A. Jentzen, *Pathwise convergent higher order numerical schemes for random ordinary differential equations* **463** (200711), 2929–2944.

[30] H. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer.

[31] P. Kythe and P. Puri, *Computational methods for linear integral equations*, Birkhäuser Boston, 2002.

[32] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht, *First-order Methods Almost Always Avoid Saddle Points*, ArXiv e-prints (October 2017), available at `1710.07406`.

[33] 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).

[34] L. Lessard, B. Recht, and A. Packard, *Analysis and design of optimization algorithms via integral quadratic constraints*, SIAM Journal on Optimization **26** (2016), no. 1, 57–95.

[35] L. Ljung, *Analysis of recursive stochastic algorithms*, IEEE Transactions on Automatic Control **22** (1977August), no. 4, 551–575.

[36] J. Milnor, *Morse theory*, Based on lecture notes by M. Spivak and R. Wells. Annals of Mathematics Studies, No. 51, Princeton University Press, Princeton, N.J., 1963. MR0163331 (29 #634)

[37] E. Moulines and F. R. Bach, *Non-asymptotic analysis of stochastic approximation algorithms for machine learning*, Advances in neural information processing systems 24, 2011, pp. 451–459.

[38] T. Neckel and F. Rupp, *Random differential equations in scientific computing*, 2013.

[39] Y. Nesterov, *Introductory lectures on convex optimization: a basic course*, Kluwer Academic Publishers, 2004.

[40] J. Nocedal and S. J. Wright, *Numerical optimization*, second, Springer, New York, NY, USA, 2006.

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

[42] N. Parikh and S. Boyd, *Proximal algorithms*, Found. Trends Optim. **1** (January 2014), no. 3, 127–239.

[43] H. Robbins and S. Monro, *A stochastic approximation method*, The Annals of Mathematical Statistics **22** (1951), no. 3, 400–407.

[44] R. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization **14** (1976), no. 5, 877–898, available at `https://doi.org/10.1137/0314056`.

[45] C. Li S. Chow and D. Wang, *Normal forms and bifurcation of planar vector fields*, Cambridge University Press, 2009.

[46] S. Kale S. Reddi and S. Kumar, *Amsgrad, on the convergence of adam and beyond*, 2017.

[47] D. Scieur, V. Roulet, F. Bach, and A. d'Aspremont, *Integration methods and optimization algorithms*, Advances in neural information processing systems 30, 2017, pp. 1109–1118.

[48] 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).

[49] T. Tieleman and G. Hinton, COURSERA, 2012.

[50] A. Wibisono and A. C. Wilson, *On Accelerated Methods in Optimization*, ArXiv e-prints (September 2015), available at `1509.03616`.

[51] A. Wibisono, A. C. Wilson, and M. I. Jordan, *A variational perspective on accelerated methods in optimization*, Proceedings of the National Academy of Sciences **113** (2016), no. 47, E7351–E7358, available at `http://www.pnas.org/content/113/47/E7351.full.pdf`.

[52] M. D. Zeiler, *ADADELTA: An Adaptive Learning Rate Method*, ECCV (2013).

[53] Z. Zhang, L. Ma, Z. Li, and C. Wu, *Normalized Direction-preserving Adam*, ArXiv e-prints (September 2017), available at `1709.04546`.