Internship Report - ENSAE, Third Year

Gradient descent, Nesterov's acceleration and Polyak's heavy ball method in continuous time: convergence rate analysis under geometric conditions and perturbations

Othmane Sebbouh

Supervisor: Charles Dossal

June - September 2018

Institut de Mathématiques de Toulouse

Acknowledgements

I thank my tutor, Charles Dossal, from the Institut de Mathématiques de Toulouse, for his great availability, the clarity of his explanations, and the experience he has given me. The four months I spent with him confirmed my desire to continue my journey in scientific research.

I would also like to thank the entire Modelisation, Simulation, Optimisation team from the Institut de Mathématiques de Toulouse.

Gradient descent, Nesterov's acceleration and Polyak's heavy ball method in continuous time: convergence rate analysis under geometric conditions and perturbations

Othmane Sebbouh

October 1, 2018

Abstract

In this report, I outlay the work done while being an intern during four months at the Institut Mathématique de Toulouse under the supervision of Charles Dossal. The general subject of my internship was the relation between first-order convex optimization algorithms and ordinary differential equations (ODE). We focused on the study of these differential equations, with a particular emphasis on the role played by the geometry of the optimized function on the convergence rates we derived. We were also interested in perturbed versions of the ODEs we considerd.

Contents

1 Introduction

1.1 Motivation

In the family of first-order convex optimization algorithms, the foundation is the steepest descent method, or gradient descent, which dates back to Cauchy. Considering a convex and differentiable function f with an L-smooth gradient, one iteration of this algorithm can be written as: for all $k \geqslant 0$

$$x_{k+1} = x_k - s\nabla f(x_k) \tag{1.1}$$

where $s \leqslant \frac{1}{L}$ is a step parameter, also called learning rate. Under convexity and smoothness assumptions, noting $f^* = \min f$ we can prove that for all $k \geqslant 0$, we have

$$f(x_k) - f^* = O(\frac{1}{k}) \tag{1.2}$$

The k-th step of this algorithm can also be written: $x_{k+1} - x_k = -s\nabla f(x_k)$. One well-known fact about this algorithm is that it corresponds, with a particular time scale, in the limit case where the step is infinitesimal, to the following ODE often referred to as the gradient flow equation: for all $t \geqslant 0$

$$\dot{x}(t) = -\nabla f(x(t)) \tag{1.3}$$

Let x be a solution of this ODE. Remarkably, if f is convex, we get an analogous convergence rate:

$$f(x(t)) - f^* = O(\frac{1}{t}) \tag{1.4}$$

The proofs of the two previous convergence rates (of the discrete algorithm and its associated ODEs) are notoriously easy. However, deriving the convergence rate for the gradient flow is substantially easier, as we will show in the second section of this report.

But these convergence rates are now known to be suboptimal for first-order optimization algorithms for the class of convex and L-smooth functions. Nesterov's accelerated gradient algorithm (NAG) [1] achieves the optimal rate of $O(\frac{1}{k^2})$. One way of writing the algorithm developed by Nesterov is:

$$\begin{aligned} x_k &= y_{k-1} - s\nabla f(y_{k-1}) \\ y_k &= x_k + \frac{k-1}{k+c-1}(x_k - x_{k-1}) \end{aligned} \tag{1.5}$$

with $c \geqslant 3$. As announced, for convex functions with an L-lipschitz gradient, we have for all $k \geqslant 0$:

$$f(x_k) - f^* = O(\frac{1}{k^2}) \tag{1.6}$$

The original proof of this result by Nesterov is tedious and somewhat mysterious. Guided by our remark about gradient descent and the gradient flow, one may wonder if there is an ODE associated with NAG, for which we can easily derive a convergence rate, and take inspiration from the proof in continuous time to derive an easier proof for the discrete algorithm. In fact, from 1983 to 2015, such a result wasn't known. However, in 2015, Su, Boyd and Candès [12] showed that Nesterov's accelerated gradient descent algorithm corresponds, with a particular time scale, when the descent step s is infinitesimal, to the following second-order ODE:

$$\ddot{x}(t) + \frac{c}{t}\dot{x}(t) + \nabla f(x(t)) = 0 \tag{1.7}$$

In their paper, they further showed that for x being a solution of this second-order ODE, we have an analogous convergence rate:

$$f(x(t)) - f^* = O(\frac{1}{t^2}) \tag{1.8}$$

The proof of this result is very easy. Based on it, they provided an easier proof for the convergence rate in the discrete case, which we report in Section 3.

This paper by Su, Boyd and Candès [12] fueled the research in the relation between optimization and the analysis of differential equations in several directions. To cite a few, Attouch et al [4], based on a work by Dossal and Chambolle [7], proved that the convergence rate of Nesterov's AGD is in fact $o(\frac{1}{k^2})$ using a Lyapunov function analysis similar to [12]. Wilson et al [5] proposed to start from very general differential equations to derive optimization algorithms. França et al [6] proved that ADMM also corresponds to a particular first order ODE and that its accelerated version corresponds to a second-order ODE.

This report proposes to study the second order ODE associated to Nesterov's accelerated gradient algorithm depending on the geometry of the convex function we consider. This analysis, which is in line with Dossal, Aujol and Rondepierre [9], allows to better understand the behaviour of Nesterov's AGD algorithm. The geometrical conditions we will consider will be explained in the third subsection of this report. Our contributions are three-fold:

1. We extend the results of [9] to the perturbed case $\ddot{x}(t)+\frac{c}{t}\dot{x}(t)+\nabla f(x(t)) = g(t)$, where g is a perturbation function which can correspond to approached calculations of the gradients: the new results are **Theorem** 3.5 and **Theorem** 3.6.

2. We study convergence rates for solutions of the following ODE $\ddot{x}(t) + \frac{c}{(1+t)^{\alpha}}\dot{x}(t) + \nabla f(x(t)) = g(t)$ depending on the geometry of the function. This equation bridges the gap between the ODE corresponding to POlyak's heavy ball method [17] : $\ddot{x}(t)+c\dot{x}(t)+\nabla f(x(t)) = 0$ and the one corresponding to Nesterov's AGD method $\ddot{x}(t) + \frac{c}{t}\dot{x}(t) + \nabla f(x(t)) = 0$. The new results are **Theorem** 4.2 and **Theorem** 4.3.

3. For the ODE $\ddot{x}(t)+\frac{c}{(1+t)^{\alpha}}\dot{x}(t)+\nabla f(x(t)) = 0$, convergence rates results in the litterature are all based on an asymptotic analysis. In this work, we derive an upper-bound on $f(x(t)) - f^*$ for all $\alpha \in [0, 1]$, thereby effectively bridging the gap between the heavy-ball method and Nesterov's AGD, for which upper-bounds existed (only in the ergodic form for the heavy-ball method). The new result is **Theorem** 4.1.

But first, let us present the central work this report relies on.

1.2 Presentation of Su, Boyd and Candès (2016)

First, we present a heuristic derivation of the ODE associated with Nesterov's AG algorithm. Then, we present an interpretation of the second-orde ODE. Afterwards, we present the Lyapunov function on which the proof of the $O(\frac{1}{t^2})$ convergence rate is built, and show how this Lyapunov function

(i) helps derive an easier proof of the $O(\frac{1}{k^2})$ convergence rate of NAG,

(ii) relates to the proof of the convergence of the iterates of NAG shown a year earlier by Chambolle and Dossal [7].

Derivation of the second-order ODE associated with NAG. Let f be a convex function with an L-lipschitz gradient. We introduce the time scaling $t = \frac{k}{\sqrt{s}}$, where s is the step parameter of NAG, which should be chosen such that $s \leqslant \frac{1}{L}$. Then $x_k \approx X(k\sqrt{s})$ for some smooth curve $X(t)$, where $t \geqslant 0$. Recall that the second equation defining NAG is the following:

$$y_k = x_k + \frac{k-1}{k+c-1}(x_k - x_{k-1}) \tag{1.9}$$

But since $x_{k+1} = y_k - s\nabla f(x_k)$, the previous equation rewrites, after scaling by $\sqrt{s}$:

$$\frac{x_{k+1}-x_k}{\sqrt{s}} = \frac{k-1}{k+c-1}\frac{x_k - x_{k-1}}{\sqrt{s}} - \sqrt{s}\nabla f(y_k) \tag{1.10}$$

Now, using our approximations, we can write $x_{k+1} = x_{\frac{t+\sqrt{s}}{\sqrt{s}}} \approx X(t+\sqrt{s})$ and the same way $x_{k-1} \approx X(t-\sqrt{s})$. Using second-order Taylor approximations, we have:

$$\begin{aligned} \frac{x_{k+1}-x_k}{\sqrt{s}} &= \frac{X(t+\sqrt{s})-X(t)}{\sqrt{s}} = \dot{X}(t) + \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s}) \\ \frac{x_k - x_{k-1}}{\sqrt{s}} &= \frac{X(t)-X(t-\sqrt{s})}{\sqrt{s}} = \dot{X}(t) - \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s}) \\ \sqrt{s}\nabla f(y_k) &= \sqrt{s}\nabla f(X(t)) + o(\sqrt{s}) \end{aligned} \tag{1.11}$$

Where we also used the fact that $y_k = X(t) + o(1)$. Now, noting that $\frac{k-1}{k+c-1} = 1 - \frac{c}{k+c-1} \approx 1 - \frac{c}{k} \approx 1 - \frac{c\sqrt{s}}{t}$, (1.10) rewrites:

$$\dot{X}(t) + \frac{1}{2}\ddot{X}(t)\sqrt{s} = \Big(1 - \frac{c\sqrt{s}}{t}\Big)\Big(\dot{X}(t) - \frac{1}{2}\ddot{X}(t)\sqrt{s}\Big) - \sqrt{s}\nabla f(X(t))\sqrt{s} + o(\sqrt{s}) \tag{1.12}$$

Then identifying the coefficients of $\sqrt{s}$, we are left with the desired ODE:

$$\ddot{X}(t) + \frac{c}{t}\dot{X}(t) + \nabla f(X(t)) = 0 \tag{1.13}$$

with initial conditions $X(0) = x_0$ and $\dot{X}(t) = 0$.

In view of this ODE, the trajectory of the iterates of NAG can be viewed as obeying to a damping system, where the friction parameter vanishes as time goes. Let's take a look at the the trajectory of a solution to this ODE. Here, we reproduce the results of Su, Boyd and Candès. To highlight the particularities of NAG as compared to GD, we compare the trajectories of a solution to the ODE associated with NAG to a solution to the gradient flow equation, starting from the same initialization. We consider the function: $f(x) = 0.02x_1^2 + 0.001x_2^2$

Fig1: Gradient Flow vs NAG ODE

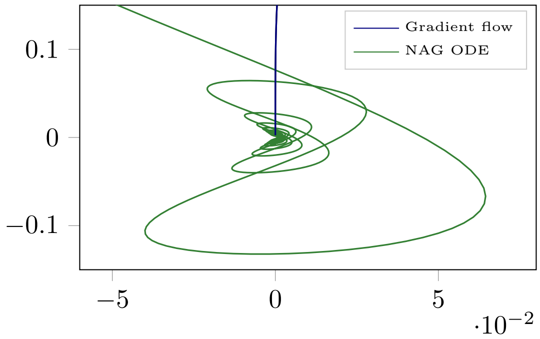
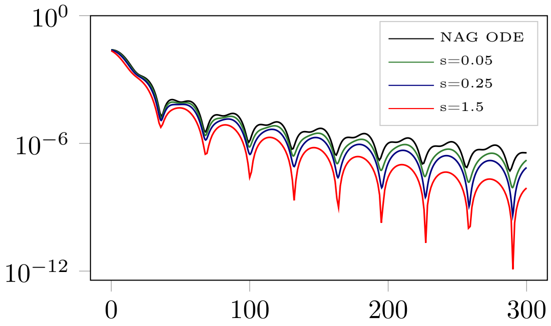


Fig2: NAG vs NAG ODE



Notice how the oscillations around the minimizer only start after a certain number of iterations, when the trajectory approaches the minimizer. This can be intuitively explained by an analysis of the ODE. Specifically, the damping parameter $\frac{c}{t}$ plays a major role in this oscillatory behaviour. In the first iterations, corresponding to a low value of t, the damping parameter is high, therefore the trajectory moves without oscillations towards the equilibrium, which is the minimizer. In the next iterations, which correspond to a high t, the friction becomes lower and tends to vanish, which produces the oscillations around the minimizer.

Hence, as we saw, one benefit of the ODE analysis for NAG is that it sheds light on the oscillatory behavior of the trajectory of the iterates. A second benefit is the way the convergence rate analysis in the continuous case provides an easier proof for the discrete case and further insights on the convergence of the iterates of NAG. We will present the proof in Section 3, but we can already look at the central elements and results of this proof to analyse it.

The central tool in the analysis of the second order ODE associated to NAG is the following Lyapunov function:

$$\mathcal{E}(t) = \frac{2t^2}{c-1}(f(x(t)) - f^*) + (c-1)\|x(t) - x^* + \frac{t}{c-1}\dot{x}(t)\|^2 \tag{1.14}$$

By simply differentiating this function and proving its decrease solely based on the convexity assumption, we can deduce an upper bound on $f(x(t)) - f^*$ and the desired converge rate $O(\frac{1}{t^2})$. Motivated by this very simple approach, Su, Boyd and Candès [12] proposed to analyze the following sequence in the discrete case:

$$\mathcal{E}(k) = \frac{2(k+c-2)^2 s}{c-1}(f(x_k) - f^*) + (c-1)\|z_k - x^*\|^2 \tag{1.15}$$

where $z_k = \frac{k+c-1}{c-1}y_k - \frac{k}{c-1}x_k$. They prove that this sequence is decreasing. This proof, which we do not report for brievity, is much simpler than the original proof by Nesterov [1]. This settles the simple convergence rate proof.

Another subject for which the new approach of [12] made things simpler to understand is the convergence proof of the iterates of NAG. Indeed, in the continuous case, [12] prove that if $c > 3$:

$$\int_0^{+\infty} t(f(X(t)) - f^*)dt \leqslant \frac{(c-1)^2\|x_0 - x^*\|^2}{2(c-3)} \tag{1.16}$$

Studying the analogous discrete Lyapunov function, they also demonstrate that:

$$\sum_{k=1}^{+\infty}(k+r-1)(f(x_k) - f^*) \leqslant \frac{(c-1)^2\|x_0 - x^*\|^2}{2s(c-3)} \tag{1.17}$$

In fact, this result had already been demonstrated in [7], but the demonstration was much more difficult. This result is decisive in the proof of the convergence iterates of NAG. It was also central in [4], which demonstrated that the convergence rate of NAG is actually $o(\frac{1}{k^2})$.

In this section, we remained purposefully elusive, as it is intended as a way to read a corpus of three papers: [7], [12] and [4], who are intertwined, and to demonstrate how [12] gave important insights on NAG.

In the next section, we will dive deeper into what will be the core of this report, which is the analysis of the second order ODE associated to NAG and other ODEs, under geometrical assumptions on the functions we consider.

1.3 Geometrical conditions considered

We first define a mild assumption on the geometry of convex functions around their minimizers, see e.g. [9]. The following condition requires the function to be flat enough around its minimizers.

Definition 1.1. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function with $X^* = \operatorname{argmin} F \neq \emptyset$, and $F^* = \inf f$. Let $\gamma \geq 1$. The function F satisfies the condition $\mathbf{H_1}(\gamma)$ if for any critical point $x^* \in X^*$, there exists $\eta > 0$ and $\epsilon > 0$ such that:*

$$\forall x \in B(x^*, \epsilon), 0 \leqslant F(x) - F^* \leqslant \frac{1}{\gamma}\langle \nabla F(x), x - x^* \rangle \tag{1.18}$$

Notice that $\forall \gamma \geqslant 1$, the condition $\mathbf{H_1}(\gamma)$ is verified whenever the function $(F - F^*)^{\frac{1}{\gamma}}$ is convex.

We also define a geometrical growth condition that can be found for example in [9]. This condition forces the considered function to be sharp around its minimizers.

Definition 1.2. *Lef $F : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function with $X^* = \operatorname{argmin} F \neq \emptyset$, and $F^* = \inf f$. Let $r \geq 1$. The function F satisfies the growth condition $\mathbf{H_2}(r)$ if for any critical point $x^* \in X^*$, there exists $K > 0$ and $\epsilon > 0$ such that:*

$$\forall x \in B(x^*, \epsilon), \;\; K d(x, X^*)^r \leq F(x) - F^* \tag{1.19}$$

We define another well-known geometrical condition, the Lojasewicz condition, which we will prove to be equivalent, in the convex setting, to a growth condition.

Definition 1.3. *Lef $F : \mathbb{R}^n \to \mathbb{R}$ be a proper differentiable function. Let $\theta \in [0, 1)$. F is said to have the Lojasewicz property (and we note $\mathbf{Loj}(\theta)$) if, for any critical point x^*, there exists $c > 0$ and $\epsilon > 0$ such that:*

$$\forall x \in B(x^*, \epsilon), \;\; \|\nabla F(x)\| \geq c(F(x) - F^*)^\theta \tag{1.20}$$

The next lemma states that the growth condition in the convex case implies the Lojasewicz property.

Lemma 1.1. *Lef $F : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function with $X^* = \operatorname{argmin} F \neq \emptyset$, and $F^* = \inf f$. Let $r \geq 1$. If F satisfies the growth condition $\mathbf{H_2}(r)$, then F has the Lojasewicz property $\mathbf{Loj}(\theta)$ with $\theta = 1 - \frac{1}{r}$.*

Proof. Let $r \geq 1$. Assume F satisfies $\mathbf{H_2}(r)$. Then there exists $\forall x \in B(x^*, \epsilon)$:

$$K\mathrm{d}(x, X^*)^r \leq F(x) - F^* \tag{1.21}$$

Let $x \in B(x^*, \epsilon)$. Since F is convex, we have:

$$F(x) - F^* \leq \langle \nabla F(x), x - x^* \rangle \tag{1.22}$$

And by Cauchy-Schwarz's inequality:

$$F(x) - F^* \leq \|\nabla F(x)\| \|x - x^*\| \tag{1.23}$$

The last equality is satisfied for all $x^* \in X^*$, hence:

$$F(x) - F^* \leq \|\nabla F(x)\| d(x, X^*) \tag{1.24}$$

And since F satisfies $\mathbf{H_2}(r)$:

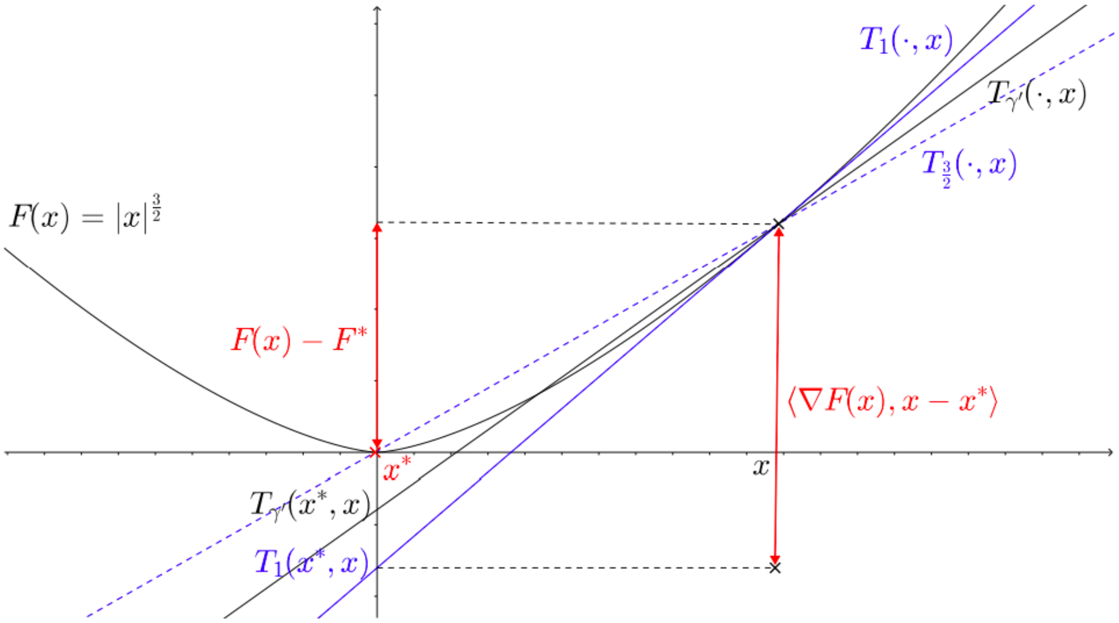
$$(F(x) - F^*)^r \leq K\|\nabla F(x)\|^r (F(x) - F^*) \tag{1.25}$$

Hence:

$$(F(x) - F^*)^{1-\frac{1}{r}} \leq K\|\nabla F(x)\| \tag{1.26}$$

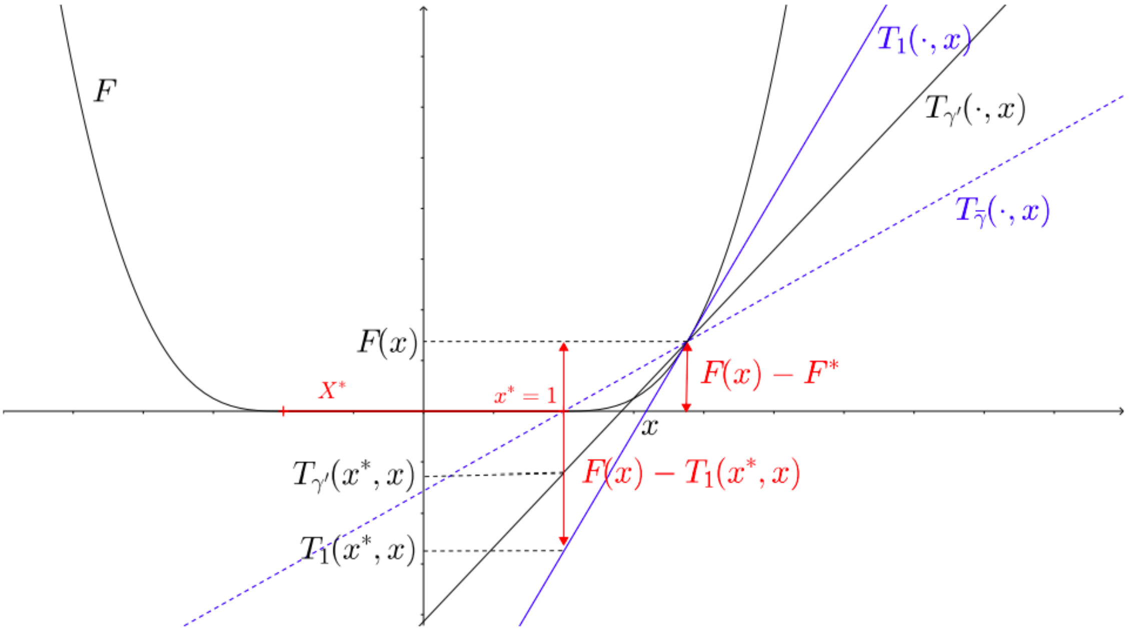
□

To give a better sense of these conditions, one can notice that the one-dimensional function $x :\to |x|^{\gamma}$ verifies the condition $\mathbf{H}_1(\gamma)$ for all $\gamma \geqslant 1$. To visualize these conditions, we present plots of different functions verifying different geometric conditions. These plots come from [9].



The function $F : x \mapsto |x|^{\frac{3}{2}}$ *satisfies* $\boldsymbol{H}_1(\gamma')$ *and* $\boldsymbol{H}_2(r)$ *if and only if* $1 \leq \gamma' \leq \gamma = \frac{3}{2}$ *and* $r \geq \frac{3}{2}$.

Fig3



The function F *defined by* (2.3) *with* $\gamma = 3$ *satisfies* $\boldsymbol{H}_1(\gamma')$ *and* $\boldsymbol{H}_2(r)$ *iff* $\gamma' \in [1, \gamma]$ *and* $r \geqslant 3$.

Fig4

1.4 Organization of the report

We present here the organization of the remainder of the report.

In Section 2, we will introduce the reader to the analysis of the gradient flow equation in presence of perturbations and with additional geometrical assumptions.

In Section 3, we will present the proofs of several results concerning the second-order ODE associated with NAG. This includes some new results.

In Section 4, we explore a general equation which includes both the ODE associated with NAG and the ODE associated with Polyak's Heavy Ball Method. As for the two previous sections, we will analyze convergence rates taking into account the possible presence of perturbations and additional geometrical assumptions. This allows to shed light on how particular, and in a sense difficult, the analysis of NAG is. Moreover, this section contains a new upper bound on the convergence of the function towards its minimum.

2 First-order ODE associated with gradient descent

We consider a convex and differentiable function $F : \mathbb{R}^n \to \mathbb{R}$ and a function $g : [t_0, +\infty) \to \mathbb{R}^n$, where $t_0 \geqslant 0$. In this section, we will study the following ODE, which corresponds to the perturbed gradient flow equation:

$$\dot{x}(t) + \nabla F(x) = g(t) \tag{2.1}$$

The most common way to think of the perturbation g is an error in the computation of the gradient, for example when calculating the gradient is costly and is consequently approximated. We will often refer to some results in the discrete case to show the analogy between the gradient flow equation and gradient descent. In the discrete case, we will assume that F has an L-smooth gradient. This assumption is not needed in the continuous case as the steps we take are by nature infinitesimal.

We first analyze the convergence rates without geometrical assumptions. Then we include geometrical assumptions to analyze the perturbation-free gradient flow.

2.1 Unperturbed case

In this section, we consider the case where $g(t) = 0$. One can show that any solution x of the ODE (2.1) satisfies a convergence rate of $O(t^{-1})$ of $F(x(t))$ to a minimum of F.

Theorem 2.1. *For any solution x to the ODE (2.1) with $g(t) = 0$, we have:*

$$F(x(t)) - F^* = O(t^{-1}) \tag{2.2}$$

More precisely, we have for all $t \geqslant 0$:

$$F(x(t)) - F^* \leq \frac{\|x_0 - x^*\|^2}{2t} \tag{2.3}$$

Proof. The proof of this proposition relies on verifying that the following energy is a Lyapunov function:

$$\mathcal{E}(t) = t(F(x(t)) - F^*) + \frac{1}{2}\|x(t) - x^*\|^2 \tag{2.4}$$

Indeed, since x verifies (2.1):

$$\mathcal{E}'(t) = F(x(t)) - F^* - t\|\nabla F(x)\|^2 - \langle \nabla F(x(t)), x(t) - x^* \rangle \tag{2.5}$$

Then, since F is convex:

$$\mathcal{E}'(t) \leq -t\|\nabla F(x)\|^2 \tag{2.6}$$

Hence $\mathcal{E}$ is decreasing, therefore: for any $t > 0$, $\mathcal{E}(t) \leq \mathcal{E}(0)$, Hence:

$$F(x(t)) - F^* \leq \frac{\|x_0 - x^*\|^2}{2t} \tag{2.7}$$

□

Remark 2.1. *Note that this bound on the error for the gradient flow equation logically compares with similar bounds derived for the gradient descent algorithm. Indeed for the k-th iterate of the gradient descent algorithm, one can easily show that:*

$$F(x_k) - F^* \leqslant \frac{2L\|x_0 - x^*\|^2}{k+4} \tag{2.8}$$

2.2 Perturbed case

In this section, we will determine under which conditions on g the convergence rate $O(t^{-1})$ can be preserved. The following proposition shows that it suffices for the function $t \to \|g(t)\|$ to be integrable.

Theorem 2.2. *Assume that $\int_{t_0}^{+\infty} \|g(t)\|dt < +\infty$. Then for any x satisfying (2.1), noting $\overline{x(t)} = \frac{1}{t}\int_0^t x(t)$:*

$$F(\overline{x(t)}) - F^* = O(t^{-1}) \tag{2.9}$$

Notice that contrary to the unperturbed case, we now give an error bound on the objective function taken at the average value of the trajectory. This makes sense since taking the error bound on a single time point can make the error bound very pessimistic as the error $g(t)$ can be particularly large at that time point while still verifying the integrability condition.

Proving this proposition requires a very useful lemma, called the Grönwall-Bellman Lemma:

Lemma 2.1. *Let $m \in L_1(t_0, T; \mathbb{R})$ such that $m \geq 0$ a.e on (t_0, T). Suppose $w : [t_0; T] \to \mathbb{R}$ is continuous. Let c be a nonnegative constant. Suppose that:*

$$\frac{1}{2}w^2(t) \leq \frac{1}{2}c^2 + \int_{t_0}^t m(s)w(s)ds \tag{2.10}$$

Then, for all $t \in [t_0, T]$:

$$|w(t)| \leq c + \int_{t_0}^t m(s)ds \tag{2.11}$$

We are now in position to prove the proposition.

Proof. Let $T \geq t \geq t_0 > 0$. To prove the proposition, we will show that the following function is a Lyapunov function:

$$\mathcal{E}(t) = t\left[F(\overline{x(t)}) - F^*\right] + \frac{1}{2}\|x(t) - x^*\|^2 + \int_t^T \langle g(u), x(u) - x^*\rangle du \tag{2.12}$$

Noting $\overline{x(t)} = \frac{1}{t}\int_0^t x(t)$, we have:

$$\mathcal{E}'(t) = F\big(\overline{x(t)}\big) - F^* + \frac{t}{t^2}\langle \nabla F\big(\overline{x(t)}\big), tx(t) - \int_0^t x(u)du\rangle + \langle x(t) - x^*, \dot{x}(t)\rangle - \langle g(t), x(t) - x^*\rangle \tag{2.13}$$

Hence, using the fact that x verifies (2.1):

$$\dot{\mathcal{E}}(t) = F\big(\overline{x(t)}\big) - F^* + \langle \nabla F\big(\overline{x(t)}\big), x(t) - \overline{x(t)}\rangle - \langle \nabla F(x(t)), x(t) - x^*\rangle \tag{2.14}$$

By convexity of F, we have $-\langle \nabla F(x(t)), x(t) - x^*\rangle \leq -(F(x(t)) - F^*) \leq 0$. Hence:

$$\mathcal{E}'(t) \leq F\big(\overline{x(t)}\big) - F(x(t)) + \langle \nabla F\big(\overline{x(t)}\big), x(t) - \overline{x(t)}\rangle \tag{2.15}$$

Again using the convexity of F, we have: $\dot{\mathcal{E}}(t) \leq 0$. In particular, $\mathcal{E}(t) \leq \mathcal{E}(t_0)$. Hence, noting $c = \|x(t_0) - x^*\|$:

$$t\left[F\Big(\frac{1}{t}\int_0^t x(t)\Big) - F^*\right] + \frac{1}{2}\|x(t) - x^*\|^2 \leq \frac{1}{2}c^2 + \int_{t_0}^t \langle g(u), x(u) - x^*\rangle du \tag{2.16}$$

Hence, using Cauchy-Schwarz inequality:

$$t\left[F\Big(\frac{1}{t}\int_0^t x(t)\Big) - F^*\right] + \frac{1}{2}\|x(t) - x^*\|^2 \leq \frac{1}{2}c^2 + \int_{t_0}^t \|g(u)\|\|x(u) - x^*\|du \tag{2.17}$$

In particular, we have:

$$\frac{1}{2}\|x(t) - x^*\|^2 \leq \frac{1}{2}c^2 + \int_{t_0}^t \|g(u)\|\|x(u) - x^*\|du \tag{2.18}$$

Hence, using **Lemma 2.1**:

$$\|x(t) - x^*\| \leq c + \int_{t_0}^t \|g(u)\|du \tag{2.19}$$

And since $\int_0^{+\infty} \|g(t)\|dt < +\infty$:

$$\sup_t \|x(t) - x^*\| < +\infty \tag{2.20}$$

Back to equation (2.17), we conclude that:

$$t\left[F\Big(\frac{1}{t}\int_0^t x(t)\Big) - F^*\right] \leq \frac{1}{2}c^2 + \Big(c + \int_{t_0}^{+\infty} \|g(t)\|dt\Big)\int_0^{+\infty} \|g(t)\|dt \tag{2.21}$$

which yields the desired result □

Remark 2.2. *Schmidt, Bach and Leroux [11] proved a similar result for the gradient descent algorithm. Noting $e_k \geqslant 0$ the error in the calculation of the gradient at the i-th iteration, they found the following bound:*

$$F\Big(\frac{1}{k}\sum_{i=1}^{k} x_k\Big) - F^* \leqslant \frac{L}{2k}\Big(\|x_0 - x^*\| + \frac{2}{L}\sum_{i=1}^{k}\|e_i\|\Big)^2 \tag{2.22}$$

Hence, to keep the $O(\frac{1}{k})$ convergence rate, it is required that $(\|e_k\|)_k$ be summable. Moreover, looking more precisely at the proof of this upper bound, we can notice that the authors of [11] use a discrete version of the Grönwall Bellman lemma we used in our proof. Again, this suggests that studying the continuous case, whose analysis is much simpler, serves as an excellent guide to propose and demonstrate results in the discrete case.

2.3 Unperturbed case with geometry

In this section, we will derive convergence rates for the values of F for trajectories satisfying the ODE (2.1) and the growth condition $\mathbf{H_2}(r)$ for $r \geq 2$, i.e. for functions behaving at best as $x \to x^2$, at worst as $x \to |x|$ around their minimizers. These results, already known from Garrigos [10], are presented in a simpler way here.

Theorem 2.3. *Let $r > 2$. If F satisfies $\mathbf{H_2}(r)$ and x is a solution to the ODE 2.1, then:*

$$F(x(t)) - F^* = O(t^{-\frac{r}{r-2}}) \tag{2.23}$$

Proof. Define $\forall t > 0,\ h(t) = -\frac{1}{1-2\theta}(F(x(t)) - F^*)^{1-2\theta}$. Then:

$$\begin{aligned} h'(t) &= -(F(x(t)) - F^*)^{-2\theta}\langle \nabla F(x(t)), \dot{x}(t)\rangle \\ &= (F(x(t)) - F^*)^{-2\theta}\|\nabla F(x(t))\|^2 \end{aligned} \tag{2.24}$$

Since F satisfies $\mathbf{H_2}(r)$, we have by the lemma 1 that F satisfies $\mathbf{Loj}(\theta)$ with $\theta = 1 - \frac{1}{r} \in (\frac{1}{2}, 1)$, there exists $c > 0$ and $\epsilon > 0$ such that $\forall x \in B(x^*, \epsilon)$:

$$(F(x(t)) - F^*)^{2\theta}\|\nabla F(x(t))\|^2 \geq c \tag{2.25}$$

Hence $h'(t) \geq c$ and $h(t) - h(0) = \int_0^t h'(u)du \geq ct$. Then:

$$\frac{1}{2\theta - 1}(F(x(t)) - F^*)^{1-2\theta} \geq ct + h(0) \tag{2.26}$$

Therefore:

$$F(x(t)) - F^* \leq (c(2\theta - 1))^{\frac{1}{1-2\theta}}(ct + h(0))^{\frac{1}{1-2\theta}} \tag{2.27}$$

Which yields the desired result □

Theorem 2.4. *If F satisfies $\mathbf{H_2}(2)$, then there exists $c > 0$ such that:*

$$F(x(t)) - F^* = O(e^{-ct})$$

Proof. Define $\forall t \geq 0,\ h(t) = F(x(t)) - F^*$. Then: $h'(t) = -\|\nabla F(x(t))\|^2$. Since F satisfies $\mathbf{H_2}(2)$, it satisfies $\mathbf{Loj}(\frac{1}{2})$. Hence, there exists $c > 0$ and $\epsilon > 0$ such that $\forall x \in B(x^*, \epsilon)$:

$$\|\nabla F(x(t))\|^2 \geq c(F(x(t)) - F^*) \tag{2.28}$$

Hence: $h'(t) \leq -ch(t)$. Therefore $F(x(t)) - F^* \leq h(0)e^{-ct}$, i.e.

$$F(x(t)) - F^* = O(e^{-ct}) \tag{2.29}$$

□

3 Second-order ODE associated to the Nesterov acceleration

We are now going to study the ODE associated to the Nesterov acceleration. In the introduction of the report, we presented how this ODE can be derived from NAG. In this section, we will dive deeper in the proof of the convergence rates results in the continuous case. The proof for NAG, which is analogous to the ones presented for the continuous case, is not reported here for brevity. For $c > 0$, we consider the following equation:

$$\ddot{x}(t) + \frac{c}{t}\dot{x}(t) + \nabla F(x) = g(t) \tag{3.1}$$

The goal is to study this equation similarly to what we did for the first-order equation. The new results will concern the perturbed case with geometry.

3.1 Unperturbed case

This case was studied extensively after the publication of Su, Boyd, Candès [12]. We eluded in the introduction to the Lyapunov functions used in the demonstration of the $O(\frac{1}{t^2})$ convergence rate result. We will know see how simple this demonstration is. One very important parameter in the convergence analysis is c, which appears in the damping coefficient. In fact, unless $c \geqslant 3$, we can't obtain the $O(\frac{1}{t^2})$ convergence rate.

Theorem 3.1. *If x is a solution to the ODE (3.1) with $c \geqslant 3$*

$$F(x(t)) - F^* = O(t^{-2}) \tag{3.2}$$

Proof. Consider the following energy function:

$$\mathcal{E}(t) = \frac{2t^2}{c-1}(F(x(t)) - F^*) + (c-1)\|x(t) + \frac{t}{c-1}\dot{x}(t) - x^*\|^2 \tag{3.3}$$

Then

$$\mathcal{E}'(t) = \frac{4t}{\alpha - 1}(F(x(t)) - F^*) + 2t^2\langle\nabla F(x(t)), \dot{x}(t)\rangle + \langle x(t) + \frac{t}{\alpha-1}\dot{x}(t) - x^*, \alpha\dot{x}(t) + t\ddot{x}(t)\rangle \tag{3.4}$$

And since x is a solution to the ODE (3.1):

$$\begin{aligned}\mathcal{E}'(t) &= \frac{4t}{c-1}(F(x(t)) - F^*) - 2t\langle\nabla F(x(t)), x(t) - x^*\rangle \\ &\leqslant -\frac{2(c-3)t}{c-1}(F(x(t)) - F^*)\end{aligned} \tag{3.5}$$

Where the inequality follows from the convexity of F. Since $c > 3$, $\mathcal{E}$ is decreasing. Then

$$\frac{2t^2}{c-1}(F(x(t)) - F^*) \leqslant \mathcal{E}(t) \leqslant \mathcal{E}(0) = (c-1)\|x_0 - x^*\|^2 \tag{3.6}$$

Which yields the desired result. □

3.2 Perturbed case

As we did with the gradient flow equation, we need to determine under which conditions on the perturbations we can still achieve the $O(\frac{1}{t^2})$ convergence rate. The results on the perturbed case can be found in several articles including Attouch, Chbani [3].

Theorem 3.2. *Assume that $\int_{t_0}^{+\infty} s\|g(s)\|ds < +\infty$. If x is a solution to the ODE (3.1) with $c \geqslant 3$, then*

$$F(x(t)) - F^* = O\left(\frac{1}{t^2}\right) \tag{3.7}$$

Proof. We define the following energy function inspired by the one proposed by Su, Boyd, Candès [12]

$$\mathcal{E}(t) = \frac{2t^2}{c-1}(F(x(t)) - F^*) + (c-1)\|x(t) + \frac{t}{c-1}\dot{x}(t) - x^*\|^2 + 2\int_t^T s\langle(\lambda(x(s) - x^*) + s\dot{x}(s)), g(s)\rangle ds \tag{3.8}$$

Differentiating and using the fact that x is a solution to the ODE (3.1) and that F is convex, one can show that:

$$\mathcal{E}(t) \leqslant -\frac{2t(c-3)}{c-1}(F(x(t)) - F^*) \leqslant 0 \tag{3.9}$$

Hence, $\mathcal{E}$ is decreasing. In particular for all $t \geqslant t_0$, $\mathcal{E}(t) \leqslant \mathcal{E}(t_0)$, which yields, noting $K = \frac{2t_0^2}{c-1}(F(x(t_0)) - F^*) + (c-1)\|x(t_0) + \frac{t_0}{c-1}\dot{x}(t_0) - x^*\|^2$

$$\frac{2t^2}{c-1}(F(x(t)) - F^*) + (c-1)\|x(t) + \frac{t}{\alpha - 1}\dot{x}(t) - x^*\|^2 \leqslant K + 2\int_{t_0}^{t} s\langle(\lambda(x(s) - x^*) + s\dot{x}(s)), g(s)\rangle ds \quad (3.10)$$

In particular

$$\|x(t) + \frac{t}{c-1}\dot{x}(t) - x^*\|^2 \leqslant \frac{K}{2(c-1)} + \frac{1}{c-1}\int_{t_0}^{t} s\langle(\lambda(x(s) - x^*) + s\dot{x}(s)), g(s)\rangle ds \quad (3.11)$$

Then, applying the Grönwall-Bellman (2.1), we have:

$$\|x(t) + \frac{t}{c-1}\dot{x}(t) - x^*\| \leqslant \left(\frac{K}{c-1}\right)^{\frac{1}{2}} + \frac{1}{c-1}\int_{t_0}^{t} sg(s)ds \quad (3.12)$$

Since $\int_{t_0}^{t} sg(s)ds < +\infty$, we then have

$$\sup_t \ \|x(t) + \frac{t}{c-1}\dot{x}(t) - x^*\| \leqslant \left(\frac{K}{c-1}\right)^{\frac{1}{2}} + \frac{1}{c-1}\int_{t_0}^{+\infty} sg(s)ds \quad (3.13)$$

Hence:

$$\frac{2t^2}{c-1}(F(x(t)) - F^*) \leqslant C + 2\left(\left(\frac{K}{c-1}\right)^{\frac{1}{2}} + \frac{1}{c-1}\int_{t_0}^{+\infty} sg(s)ds\right)\int_{t_0}^{+\infty} sg(s)ds \quad (3.14)$$

Which yields the desired result. □

Remark 3.1. *Again, we can notice that in [11], the authors derived a similar upper bound. With $e_k \geqslant 0$ the error on the calculation of the gradient at the k-th iteration, we have:*

$$F(x_k) - F^* \leqslant \frac{2L}{(k+1)^2}\left(\|x_0 - x^*\| + \frac{2}{L}\sum_{i=1}^{k} i\|e_i\|\right)^2 \quad (3.15)$$

3.3 Unperturbed case with geometry

We now want to study the ODE 3.1 with additional geometrical assumptions to see if we can obtain convergence rates beyond the optimal $O(\frac{1}{t^2})$ for the class of convex and smooth function. Dossal, Aujol, Rondepierre [9] proved that it is indeed the case, and provided optimal convergence rates for the classes of functions they consider. We report here their results without demonstration, as we will use similar reasoning in the derivation of our new results in the following sections.

Theorem 3.3. *Let $\gamma \geqslant 1$ and $c > 0$.*

1. *If F satisfies the hypothesis $\mathbf{H_1}(\gamma)$ and if $c \leqslant 1 + \frac{2}{\gamma}$, then*

$$F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2\gamma c}{\gamma+2}}}\right) \quad (3.16)$$

2. *If F satisfies the hypotheses $\mathbf{H_1}(\gamma)$ and $\mathbf{H_2}(2)$ and if F has a unique minimizer, if $c > 1 + \frac{2}{\gamma}$, then*

$$F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2\gamma c}{\gamma+2}}}\right) \quad (3.17)$$

This theorem presents two results:

1. As recalled in the beginning of the first subsection, we can't obtain a $O(\frac{1}{t^2})$ convergence rate unlesse $c \geqslant 3$. Another well-known result is that if we take $c \in [1, 3]$, we achieve $O(\frac{1}{t^{\frac{2c}{3}}})$. The first result states that assuming F satisfies $\mathbf{H}_1(\gamma)$ with $\gamma \geqslant 1$, we can reach slightly better convergence rates than the $O(\frac{1}{t^{\frac{2c}{3}}})$ with $c \leqslant 1 + \frac{2}{\gamma}$

2. To understand the second point simply, notice that strongly convex functions satisfy the hypothesis $\mathbf{H}_2(\gamma)$. A very noticeable fact about this result is that it states the following: one can take c very high leads to the best asymptotic convergence rate. One has to keep in mind that for strongly convex functions, there exists a Nesterov acceleration scheme which achieves a linear convergence rate, where the damping parameter is chosen to be $\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$. However, achieving such a rate requires knowing the strong convexity parameter, which is not always easy to compute. Also, the rates we present here are asymptotic. Setting a very high value for c might not be a good idea if NAG is run for a few iterations.

Theorem 3.4. *Let $\gamma_1 > 2$, $\gamma_2 > 2$. If F is coercive and satisfies $\mathbf{H_1}(\gamma_1)$ and $\mathbf{H_2}(\gamma_2)$,*

1. *If $\gamma_1 \leqslant \gamma_2$ and $c \geqslant \frac{\gamma_1+2}{\gamma_1-2}$, then*

$$F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2\gamma_2}{\gamma_2-2}}}\right) \tag{3.18}$$

2. *If $\gamma_2 \leqslant \gamma_1$, then for any $\gamma \in [\gamma_2, \gamma_1]$, if $\alpha \geqslant \frac{\gamma+2}{\gamma-2}$, then*

$$F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2\gamma}{\gamma-2}}}\right) \tag{3.19}$$

To explain what this theorem tells us, we recall the reader that NAG, unlike, gradient descent, doesn't take advantage of very sharp functions. Instead, the main well-known benefit of NAG is the fact that it is much faster than gradient descent when the magnitude of the gradient is low. Conversely, when the function is very sharp, NAG tends to exhibit more oscillations, and hence becomes slower. In this theorem, we assume that our function is flat enough such that we don't have too many oscillations, but also that it is sharp enough such that the magnitude of the gradient isn't too low. Under these assumptions, we get very favorable convergence rates.

3.4 Perturbed case with geometry

In this section, we will extend the results of [9] to the perturbed case. As usual, we are trying to find under which conditions on g we can keep the convergence rates presented in the previous subsection. The results presented in this section are new.

Theorem 3.5. *Let $\gamma \geq 1$ and $c > 0$. Let $\eta = \frac{\gamma c}{\gamma+2}$ and assume that $\int_{t_0}^{+\infty} t^\eta g(t)dt < +\infty$*

1. *If F satisfies the hypothesis $\mathbf{H_1}(\gamma)$ and if $c \leqslant 1 + \frac{2}{\gamma}$, then*

$$F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2\gamma c}{\gamma+2}}}\right) \tag{3.20}$$

2. *If F satisfies the hypotheses $\mathbf{H_1}(\gamma)$ and $\mathbf{H_2}(2)$ and if F has a unique minimizer, if $c > 1 + \frac{2}{\gamma}$, then*

$$F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2\gamma c}{\gamma+2}}}\right) \tag{3.21}$$

Note that the first point is not new, it was demonstrated in Aujol and Dossal [8]. We give a proof of the second point. To do so, we will use the following variation of the Grönwall-Bellman lemma:

Lemma 3.1. *Let $m \in L_1(t_0, T; R)$ such that $m \geq 0$ a.e on (t_0, T). Suppose $w : [t_0; T] \to \mathbb{R}$ is continuous. Let K be a nonnegative constant. Suppose that:*

$$w(t) \leq K + \int_{t_0}^{t} m(s)w(s)ds \tag{3.22}$$

Then, for all $t \in [t_0, T]$:

$$w(t) \leq K \ \exp\Big(\int_{t_0}^{t} m(s)ds\Big) \tag{3.23}$$

The proof of our results will rely on the following energy function:

$$\mathcal{G}(t)=\mathcal{H}(t)+\int_t^T\langle s^{\frac{p}{2}}(\lambda(x(s)-x^*)+s\dot{x}(s)),s^{\frac{p+2}{2}}g(s)\rangle ds \tag{3.24}$$

Where $\mathcal{H}(t)=t\mathcal{E}(t)$ and

$$\mathcal{E}(t)=t^2(F(x(t))-F^*)+\frac{1}{2}\|\lambda(x(t)-x^*)+t\dot{x}(t)\|^2+\frac{\xi}{2}\|x(t)-x^*\|^2 \tag{3.25}$$

Noting

$$\begin{aligned}a(t)&=t(F(x(t))-F^*)\\ b(t)&=\frac{1}{2t}\|\lambda(x(t)-x^*)+t\dot{x}(t)\|^2\\ c(t)&=\frac{1}{2t}\|x(t)-x^*\|^2\end{aligned} \tag{3.26}$$

We have:

$$\mathcal{E}(t)=a(t)+b(t)+\xi c(t) \tag{3.27}$$

Our proofs are based on the following lemma:

Lemma 3.2. *Let $\gamma\geqslant 1$. If F satisfies the hypothesis $\mathbf{H}_1(\gamma)$ and if $\xi=\lambda(\lambda+1-c)$, then*

$$\mathcal{G}'(t)\leqslant t^p(2+p-\gamma\lambda)a(t)+(p+2\lambda+2r-2c)b(t)+\lambda(\lambda+1-c)c(t)) \tag{3.28}$$

This lemma is proved in the appendix A. We now turn to the proof of the theorem.

Proof. Then, chosing $p=\frac{2\gamma c}{\gamma+2}-2$ and $\lambda=\frac{2c}{\gamma+2}$, we are left with

$$\mathcal{G}'(t)\leqslant K_1t^pc(t) \tag{3.29}$$

where $K_1=\lambda(\lambda+1-c)(-2\lambda+p)$. Since K_1 is not always negative for any value of γ when $c>1+\frac{2}{\gamma}$, we cannot conclude that the energy function $\mathcal{G}$ is decreasing. Hence we will use the same arguments as in [9].

Using the uniqueness of the minimizer and the fact that F satisfies $\mathbf{H}_2(2)$, there exists $K>0$ such that:

$$Kt\|x(t)-x^*\|^2\leqslant t(F(x(t))-F^*)=a(t) \tag{3.30}$$

Hence:

$$c(t)\leqslant\frac{1}{2Kt^2}a(t) \tag{3.31}$$

And since $\xi<0$

$$\mathcal{H}(t)\geqslant t^{p+1}(a(t)+\xi c(t))\geqslant t^{p+1}(1+\frac{\xi}{2Kt^2})a(t) \tag{3.32}$$

Then, there exists $t_1\geqslant t_0$ such that for all $t\geqslant t_1$, $\mathcal{H}(t)\geqslant 0$ and

$$\mathcal{H}(t)\geqslant\frac{1}{2}t^{p+1}a(t) \tag{3.33}$$

Now from (3.29), (3.31) and (3.33), we have

$$\mathcal{G}'(t)\leqslant\frac{K_1}{K}\frac{\mathcal{H}(t)}{t^3} \tag{3.34}$$

And since $\mathcal{H}'(t)=\mathcal{G}'(t)+\langle\lambda(x(t)-x^*)+t\dot{x}(t),t^{p+1}g(t)\rangle$:

$$\begin{aligned}\mathcal{H}'(t)&\leq\frac{K_1}{K}\frac{\mathcal{H}(t)}{t^3}+\langle t^{\frac{p}{2}}(\lambda(x(t)-x^*)+t\dot{x}(t)),t^{\frac{p+2}{2}}g(t)\rangle\\ &\leq\frac{K_1}{K}\frac{\mathcal{H}(t)}{t^3}+t^{\frac{p}{2}}\|\lambda(x(t)-x^*)+t\dot{x}(t)\|t^{\frac{p+2}{2}}\|g(t)\|\\ &=\frac{K_1}{K}\frac{\mathcal{H}(t)}{t^3}+\sqrt{2}(t^{p+1}b(t))^{\frac{1}{2}}t^{\frac{p+2}{2}}\|g(t)\|\end{aligned} \tag{3.35}$$

But since $\mathcal{H}(t) = t^{p+1}b(t) + t^{p+1}(a(t) + \xi c(t))$ and as demonstrated by [9], there exists $t_1 \geqslant t_0$ such that for all $t \geq t_1$, $t^{p+1}(a(t) + \xi c(t)) \geq \frac{1}{2}t^{p+1}a(t)$, which means that $t^{p+1}b(t) + \frac{1}{2}t^{p+1}a(t) \leq \mathcal{H}(t)$. Hence: $t^{p+1}b(t) \leq \mathcal{H}(t)$. Therefore:

$$\mathcal{H}'(t) \leq \frac{K_1}{K}\frac{\mathcal{H}(t)}{t^3} + \sqrt{2}\mathcal{H}(t)^{\frac{1}{2}}t^{\frac{p+2}{2}}\|g(t)\| \tag{3.36}$$

Then, dividing both sides of the quality by $2\mathcal{H}(t)^{\frac{1}{2}}$:

$$\frac{\mathcal{H}'(t)}{2\mathcal{H}(t)^{\frac{1}{2}}} \leq \frac{K_1}{2K}\frac{\mathcal{H}(t)^{\frac{1}{2}}}{t^3} + \frac{\sqrt{2}}{2}t^{\frac{p+2}{2}}\|g(t)\| \tag{3.37}$$

Then, integrating between t_1 and t:

$$\mathcal{H}(t)^{\frac{1}{2}} \leq \frac{K_1}{2K}\int_{t_1}^{t}\frac{\mathcal{H}(s)^{\frac{1}{2}}}{s^3}ds + \frac{\sqrt{2}}{2}\int_{t_1}^{t}s^{\frac{p+2}{2}}\|g(s)\|ds \tag{3.38}$$

Since $\int_{t_1}^{+\infty} s^{\frac{p+2}{2}}\|g(s)\|ds < +\infty$:

$$\mathcal{H}(t)^{\frac{1}{2}} \leq \frac{K_1}{2K}\int_{t_1}^{t}\frac{\mathcal{H}(s)^{\frac{1}{2}}}{s^3}ds + \frac{\sqrt{2}}{2}\int_{t_1}^{+\infty}s^{\frac{p+2}{2}}\|g(s)\|ds \tag{3.39}$$

Noting $\beta = \frac{\sqrt{2}}{2}\int_{t_1}^{+\infty} s^{\frac{p+2}{2}}\|g(s)\|ds$, we have:

$$\mathcal{H}(t)^{\frac{1}{2}} \leq \beta + \int_{t_1}^{t}\frac{K_1}{2Ks^3}\mathcal{H}(s)^{\frac{1}{2}}ds \tag{3.40}$$

Then, using **Lemma** (3.1):

$$\mathcal{H}(t) \leq \beta^2\ \exp\Big(\frac{K_1}{2K}\int_{t_1}^{+\infty}\frac{1}{s^3}ds\Big) \tag{3.41}$$

Hence we found $A > 0$ such that for all $t \geq t_1$, $\mathcal{H}(t) \leq A$. Since $\frac{1}{2}t^{p+2}(F(x(t)) - F^*) = \frac{1}{2}t^{p+1}a(t) \leq \mathcal{H}(t)$, we have the desired result. □

Theorem 3.6. *Let $\gamma_1 > 2$, $\gamma_2 > 2$. Let $\eta = \frac{\gamma_1}{\gamma_1 - 2}$ and assume that $\int_{t_0}^{+\infty} t^\eta g(t)dt < +\infty$. If F admits a unique minimizer and satisfies $\mathbf{H_1}(\gamma_1)$ and $\mathbf{H_2}(\gamma_2)$, if $\gamma_2 \leqslant \gamma_1$, then for any $\gamma \in [\gamma_2, \gamma_1]$, if $c \geqslant \frac{\gamma+2}{\gamma-2}$, then*

$$F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2\gamma}{\gamma-2}}}\right) \tag{3.42}$$

To prove this theorem we will need the following lemma, for which we give a proof in the appendix C.

Lemma 3.3. *Let $K > 0$, $q, r \leqslant 1$. Let $T > t_0 > 0$. Let $f \in L_1(t_0, T; \mathbb{R})$ such that $f \geq 0$ a.e on (t_0, T). Suppose $m : [t_0; T] \to \mathbb{R}$ is continuous. Suppose that:*

$$m(t) \leqslant \beta + \int_{t_0}^{t}[m^q(s) + m^r(s)]f(s)ds \tag{3.43}$$

Then

$$x(t) \leqslant \beta + A(t)\exp\big(\int_{t_0}^{t}B(s)ds\big) \tag{3.44}$$

where, for any $K > 0$:

$$\begin{aligned} A(t) &= \int_{t_0}^{t}[(1-q)K^q + qK^{q-1}c + (1-r)K^r + rK^{r-1}c]f(s)ds \\ B(t) &= (qK^{q-1} + rK^{r-1})f(t) \end{aligned} \tag{3.45}$$

We are now in position to prove the theorem.

Proof. Consider the energy previously defined:

$$\mathcal{G}(t) = \mathcal{H}(t) + \int_t^T \langle s^{\frac{p}{2}}(\lambda(x(s)-x^*) + s\dot{x}(s)), s^{\frac{p+2}{2}} g(s)\rangle ds \tag{3.46}$$

Set $\lambda = \frac{2}{\gamma_1-2}$ and $p = \frac{4}{\gamma_1-2}$. Since $c \geqslant \frac{\gamma_1+2}{\gamma_1-2}$, we have from (3.28):

$$\mathcal{G}'(t) \leqslant 2t^{\frac{4}{\gamma_1-2}}\Big(\frac{\gamma_1+2}{\gamma_1-2} - c\Big) b(t) \leqslant 0 \tag{3.47}$$

Hence, in particular for all $t \geqslant t_0$, $\mathcal{G}(t) \leqslant \mathcal{G}(t_0)$. Then:

$$\mathcal{H}(t) \leqslant \mathcal{H}(t_0) + \int_{t_0}^t \langle s^{\frac{p}{2}}(\lambda(x(s)-x^*) + s\dot{x}(s)), s^{\frac{p+2}{2}} g(s)\rangle ds \tag{3.48}$$

Hence, we have:

$$t^{p+1}a(t) + t^{p+1}b(t) \leqslant \mathcal{H}(t_0) + |\xi| t^{p+1}c(t) + \int_{t_0}^t \langle s^{\frac{p}{2}}(\lambda(x(s)-x^*) + s\dot{x}(s)), s^{\frac{p+2}{2}} g(s)\rangle ds \tag{3.49}$$

Since F satisfies $\mathbf{H}_2(\gamma_2)$ and admits a unique minimizer, there exists $\omega > 0$ such that:

$$\|x(t) - x^*\|^2 \leq \omega (F(x(t)) - F^*)^{\frac{2}{\gamma_2}} \tag{3.50}$$

Then, $t^{p+1}c(t) = \frac{1}{2}t^p\|x(t)-x^*\|^2 \leqslant \frac{\omega}{2}t^p(F(x(t))-F^*)^{\frac{2}{\gamma_2}} = \frac{\omega}{2}t^{p-\frac{2}{\gamma_2}}a(t)^{\frac{2}{\gamma_2}}$. But $p - \frac{2}{\gamma_2} = \frac{2}{\gamma_2}\frac{2\gamma_2-\gamma_1+2}{\gamma_1-2}$. And since $\gamma_2 \leq \gamma_1$, $p - \frac{2}{\gamma_2} \leq \frac{2}{\gamma_2}\frac{\gamma_1+2}{\gamma_1-2} = \frac{2}{\gamma_2}(p+1)$. Hence there exists some $t_1 \geqslant t_0$ such that $\forall t \geqslant t_1, t^{p+1}c(t) \leqslant \frac{\omega}{2}(t^{p+1}a(t))^{\frac{2}{\gamma_2}}$. Therefore:

$$t^{p+1}a(t) + t^{p+1}b(t) \leqslant \mathcal{H}(t_0) + |\xi|\frac{\omega}{2}(t^{p+1}a(t))^{\frac{2}{\gamma_2}} + \int_{t_0}^t \langle s^{\frac{p}{2}}(\lambda(x(s)-x^*) + s\dot{x}(s)), s^{\frac{p+2}{2}} g(s)\rangle ds \tag{3.51}$$

Writing $|\xi|\frac{\omega}{2}(t^{p+1}a(t))^{\frac{2}{\gamma_2}} = \int_{t_0}^t |\xi|\frac{\omega}{2}(s^{p+1}a(s))^{\frac{2}{\gamma_2}}\delta_t(s)ds$, $\|t^{\frac{p}{2}}(\lambda(x(t)-x^*) + s\dot{x}(s))\| = (2t^{p+1}b(t))^{\frac{1}{2}}$ and using the Cauchy-Schwartz inequality, we have:

$$t^{p+1}a(t) + t^{p+1}b(t) \leqslant \mathcal{H}(t_0) + \int_{t_0}^t |\xi|\frac{\omega}{2}(s^{p+1}a(s))^{\frac{2}{\gamma_2}}\delta_t(s) + (2s^{p+1}b(s))^{\frac{1}{2}}\|s^{\frac{p+2}{2}}g(s)\|ds \tag{3.52}$$

Since all the terms are positive in the integral, we have:

$$t^{p+1}a(t) + t^{p+1}b(t) \leqslant \mathcal{H}(t_0) + \int_{t_0}^t \big((s^{p+1}a(s))^{\frac{2}{\gamma_2}} + (s^{p+1}b(s))^{\frac{1}{2}}\big)(|\xi|\frac{\omega}{2}\delta_t(s) + \sqrt{2}\|s^{\frac{p+2}{2}}g(s)\|)ds \tag{3.53}$$

Noting $d(t) = t^{p+1}a(t) + t^{p+1}b(t)$, we have again since all the terms are positive:

$$d(t) \leqslant \mathcal{H}(t_0) + \int_{t_0}^t \big(d(s)^{\frac{2}{\gamma_2}} + d(s)^{\frac{1}{2}}\big)(|\xi|\frac{\omega}{2}\delta_t(s) + \sqrt{2}\|s^{\frac{p+2}{2}}g(s)\|)ds \tag{3.54}$$

Then, with the notations of **Lemma** 3.3 and $f(s) = |\xi|\frac{\omega}{2}\delta_t(s) + \sqrt{2}\|s^{\frac{p+2}{2}}g(s)\|$, $q = \frac{2}{\gamma_2}$, $r = \frac{1}{2}$, we have:

$$d(t) \leqslant \mathcal{H}(t_0) + A(t)\exp\Big(\int_{t_0}^t B(s)ds\Big) \tag{3.55}$$

Let $K > 0$. Since all the terms in the integrals are positive and $\int_{t_0}^{+\infty}\|s^{\frac{p+2}{2}}g(s)\|ds < +\infty$, we have, noting $c_1 = (1-q)K^q + qK^{q-1}\omega + (1-r)K^r + rK^{r-1}\omega$ and $c_2 = qK^{q-1} + rK^{r-1}$:

$$\begin{aligned} A(t) &\leqslant \int_{t_0}^{+\infty} c_1 f(s)ds \\ \int_{t_0}^t B(s)ds &\leqslant \int_{t_0}^{+\infty} c_2 f(s)ds \end{aligned} \tag{3.56}$$

Hence:

$$d(t) \leqslant \mathcal{H}(t_0) + \Big(\int_{t_0}^{+\infty} c_1 f(s) ds \Big) \exp \Big(\int_{t_0}^{+\infty} c_2 f(s) ds \Big) \tag{3.57}$$

Hence, we have found $A > 0$ such that $\forall t \geqslant t_1$:

$$t^{p+1} a(t) + t^{p+1} b(t) \leqslant A \tag{3.58}$$

In particular:

$$t^{\frac{2\gamma_1}{\gamma_1 - 2}} (F(x(t)) - F^*) \leqslant A \tag{3.59}$$

Which is the desired result. □

As noticed in [9], if $\gamma_1 = \gamma_2$, we have furthermore the convergence of the trajectory:

Corollary 3.1. *Let $\gamma > 2$ and $\eta = \frac{\gamma}{\gamma - 2}$. Suppose $\int_{t_0}^{+\infty} t^\eta \|g(t)\| dt < +\infty$. If F admits a unique minimizer and satisfies $\mathbf{H}_1(\gamma)$ and $\mathbf{H}_2(\gamma)$, and if $c \geqslant \frac{\gamma+2}{\gamma-2}$ then*

$$\|\dot{x}(t)\| = O\left(\frac{1}{t^{\frac{\gamma}{\gamma-2}}} \right) \tag{3.60}$$

Proof. From the proof of the previous theorem we have that there exists $A > 0$ such that $\frac{t^{\frac{4}{\gamma-2}}}{2} \|\lambda(x(t) - x^*) + t\dot{x}(t)\|^2 = t^{p+1} b(t) \leqslant A$. Hence, there exists $A_1 > 0$ such that:

$$\|\lambda(x(t) - x^*) + t\dot{x}(t)\| \leqslant \frac{A_1}{t^{\frac{2}{\gamma-2}}} \tag{3.61}$$

This yields the fact that:

$$t\|\dot{x}(t)\| \leqslant \lambda \|x(t) - x^*\| + \frac{A_1}{t^{\frac{2}{\gamma-2}}} \tag{3.62}$$

But since F satisfies $\mathbf{H}_2(\gamma)$

$$\|x(t) - x^*\| \leqslant c(F(x(t)) - F^*)^{\frac{1}{\gamma}} \tag{3.63}$$

and by the using the previous theorem, we have $F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2\gamma}{\gamma-2}}} \right)$, i.e exists $A_2 > 0$ such that:

$$\|x(t) - x^*\| \leqslant \frac{A_2}{t^{\frac{2}{\gamma-2}}} \tag{3.64}$$

Hence, back to (3.62), we have

$$t\|\dot{x}(t)\| \leqslant \frac{\lambda A_2 + A_1}{t^{\frac{2}{\gamma-2}}} \tag{3.65}$$

Which means, with $A_3 = \lambda A_2 + A_1 > 0$

$$\|\dot{x}(t)\| \leqslant \frac{A_3}{t^{\frac{\gamma}{\gamma-2}}} \tag{3.66}$$

Hence $\|\dot{x}(t)\|$ is integrable, therefore the trajectory is finite. □

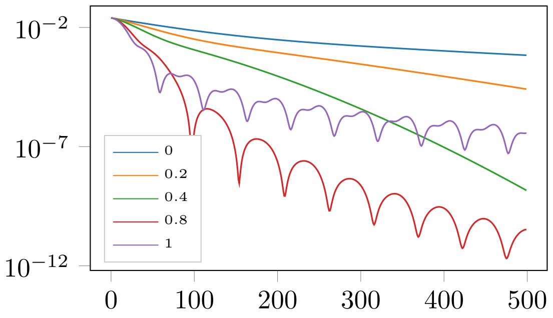
4 Heavy Ball system with a vanishing damping term

We will now be interested in the following ODE:

$$\ddot{x}(t) + \frac{c}{(1+t)^\alpha} \dot{x}(t) + \nabla F(x) = g(t) \tag{4.1}$$

where $c > 0$ and $\alpha \in [0, 1]$ We can notice that with this equation we can study both the ODE associated with NAG and the ODE associated with the Heavy Ball Method. But as we will see, studying it with a general $\alpha \in [0, 1]$ is in fact trickier, as the cases $\alpha = 0$ and especially $\alpha = 1$ sometimes require particular attention. For x a solution to the ODE (4.2), we plot $f(x(t))$ for different values of α with $c = 3$ with $f(x) = 0.02x_1^2 + 0.001x_2^2$.

Fig4: HB ODE depending on α



4.1 An upper bound for the HB equation

May and Balti's [14] proved convergence rates for the solutions of the ODE (4.2) with and without perturbations. However, their results are asymptotic. The question is: can we derive a non-asymptotic bound on the ODE (4.2)? In this note, we prove that this result can be achieved, and we notice as expected that we require a condition on the friction parameter c to achieve such a bound.

Theorem 4.1. *Let F a convex and differentiable function on a Hilbert space $\mathcal{H}$. Let $c > 0$ and $\alpha \in [0, 1]$ Let x be a solution of the following ODE with the initial conditions $x(0) = x_0 \in \mathcal{H}$ and $\dot{x}(0) = 0$:*

$$\ddot{x}(t) + \frac{c}{(1+t)^{\alpha}}\dot{x}(t) + \nabla F(x(t)) = 0 \tag{4.2}$$

If $c \geqslant 2\alpha + 1$, then there exists some constant $b > 0$ such that:

$$F(x(t)) - F^* \leqslant \frac{2(F(x_0) - F^*) + b\|x_0 - x^*\|^2}{(1+t)^{1+\alpha}} \tag{4.3}$$

Proof. Let $\beta \in [0, 1]$. We define functions $\lambda_\beta(.)$ and $\xi_\beta(.)$ to be two real-valued functions, with $\lambda_\beta(.)$ non-negative. Consider the following function:

$$\mathcal{E}_\beta(t) = (1+t)^{2\beta}(F(x(t)) - F^*) + \frac{1}{2}\|\lambda_\beta(t)(x(t) - x^*) + (1+t)^{\beta}\dot{x}(t)\|^2 + \xi_\beta(t)\|x(t) - x^*\|^2 \tag{4.4}$$

We will now demonstrate that with careful choices of $\lambda_\beta(.)$ and $\xi_\beta(.)$, we have:

$$\mathcal{E}'_\beta(t) \leqslant a_\beta(t)(1+t)^{\beta}\|\dot{x}(t)\|^2 + \Big(\frac{\xi'_\beta(t)}{2} + \lambda_\beta(t)\lambda'_\beta(t)\Big)\|x(t) - x^*\|^2 \tag{4.5}$$

Let's differentiate $\mathcal{E}_\beta$:

$$\begin{aligned}
\mathcal{E}'_\beta(t) &= 2\beta(1+t)^{2\beta-1}(F(x(t)) - F^*) + (1+t)^{2\beta}\langle \nabla F(x(t)), \dot{x}(t)\rangle \\
&+ \langle \lambda_\beta(t)(x(t) - x^*) + (1+t)^{\beta}\dot{x}(t), \lambda'_\beta(t)(x(t) - x^*) + \lambda_\beta(t)\dot{x}(t) + \beta(1+t)^{\beta-1}\dot{x}(t) + (1+t)^{\beta}\ddot{x}(t)\rangle \\
&+ \frac{\xi'_\beta(t)}{2}\|x(t) - x^*\|^2 + \xi_\beta(t)\langle \dot{x}(t), x(t) - x^*\rangle \\
&= 2\beta(1+t)^{2\beta-1}(F(x(t)) - F^*) + (1+t)^{2\beta}\langle \nabla F(x(t)), \dot{x}(t)\rangle \\
&+ \Big(\xi_\beta(t) + \lambda_\beta(t)a_\beta(t) + \lambda'_\beta(t)(1+t)^{\beta}\Big)\langle \dot{x}(t), x(t) - x^*\rangle \\
&+ a_\beta(t)(1+t)^{\beta}\|\dot{x}(t)\|^2 - (1+t)^{2\beta}\langle \nabla F(x(t)), \dot{x}(t)\rangle \\
&- \lambda_\beta(t)(1+t)^{\beta}\langle F(x(t)), x(t) - x^*\rangle + \Big(\frac{\xi'_\beta(t)}{2} + \lambda_\beta(t)\lambda'_\beta(t)\Big)\|x(t) - x^*\|^2
\end{aligned} \tag{4.6}$$

where $a_\beta(t) = \lambda_\beta(t) + \beta(1+t)^{\beta-1} - c(1+t)^{\beta-\alpha}$. Since F is convex, we have: $-\lambda_\beta(t)(1+t)^{\beta}\langle \nabla F(x(t)), x(t) - x^*\rangle \leqslant -\lambda_\beta(t)(1+t)^{\beta}(F(x(t)) - F^*)$. If we further set $\underline{\xi_\beta(t) = -\lambda_\beta(t)a_\beta(t) - \lambda'_\beta(t)(1+t)^{\beta}}$, then

$$\begin{aligned}
\mathcal{E}'_\beta(t) &\leqslant (2\beta(1+t)^{2\beta-1} - \lambda_\beta(t)(1+t)^{\beta})(F(x(t)) - F^*) \\
&+ a_\beta(t)(1+t)^{\beta}\|\dot{x}(t)\|^2 + \Big(\frac{\xi'_\beta(t)}{2} + \lambda_\beta(t)\lambda'_\beta(t)\Big)\|x(t) - x^*\|^2
\end{aligned} \tag{4.7}$$

Let's now chose $\lambda_\beta(t) = 2\beta(1+t)^{\beta-1}$. Then

$$\mathcal{E}'_\beta(t) \leqslant a_\beta(t)(1+t)^\beta \|\dot{x}(t)\|^2 + \Big(\frac{\xi'_\beta(t)}{2} + \lambda_\beta(t)\lambda'_\beta(t)\Big)\|x(t)-x^*\|^2 \tag{4.8}$$

We now give the explicit values of $a_\beta(t)$ and $\xi_\beta(t)$:

$$\begin{aligned} a_\beta(t) &= 3\beta(1+t)^{\beta-1} - c(1+t)^{\beta-\alpha} \\ \xi_\beta(t) &= 2\beta\Big(c(1+t)^{2\beta-\alpha-1} - (4\beta-1)(1+t)^{2\beta-2}\Big) \end{aligned} \tag{4.9}$$

Notice that $\xi_\beta(t)$ is positive for all $t \geqslant 0$ if and only if $c \geqslant 4\beta - 1$. Moreover:

$$\xi'_\beta(t) = 2\beta\Big(c(2\beta-\alpha-1)(1+t)^{2\beta-\alpha-2} - 2(\beta-1)(4\beta-1)(1+t)^{2\beta-3}\Big) \tag{4.10}$$

Note $r = \frac{\alpha+1}{2} \in [0,1]$. Define for all $t \geqslant 0$, $\mathcal{E}(t) = \mathcal{E}_r(t) + \mathcal{E}_\alpha(t)$. Then from the inequality (4.8), we have

$$\mathcal{E}'(t) \leqslant \Big(a_\alpha(t)(1+t)^\alpha + a_r(t)(1+t)^r\Big)\|\dot{x}(t)\|^2 + \Big(\frac{\xi'_\alpha(t)}{2} + \lambda_\alpha(t)\lambda'_\alpha(t) + \frac{\xi'_r(t)}{2} + \lambda_r(t)\lambda'_r(t)\Big)\|x(t)-x^*\|^2 \tag{4.11}$$

Now we will examine under which conditions on c we have $\mathcal{E}'(t) \leqslant 0$. For the first term, we have

$$\begin{aligned} a_\alpha(t)(1+t)^\alpha + a_r(t)(1+t)^r &= (3\alpha(1+t)^{2\alpha-1} - c(1+t)^\alpha + 3r(1+t)^{2r-1} - c(1+t)^{2r-\alpha} \\ &= (3\alpha(1+t)^{2\alpha-1} + (3r-c)(1+t)^\alpha(1+t)^\alpha - c(1+t) \end{aligned} \tag{4.12}$$

Hence, for this term to be negative, we need to have $c \geqslant 3r - c + 3\alpha$, i.e. $\underline{c \geqslant \frac{3}{2}(r+\alpha) := \frac{3(3\alpha+1)}{2}}$.

For the second term, we have $\lambda_\alpha(t)\lambda'_\alpha(t) = 4\alpha^2(\alpha-1)(1+t)^{2\alpha-3}$ and

$$\xi'_\alpha(t) = 2\alpha\Big(c(\alpha-1)(1+t)^{\alpha-2} - 2(\alpha-1)(4\alpha-1)(1+t)^{2\alpha-3}\Big) = 2\alpha(1-\alpha)\Big(2(4\alpha-1)(1+t)^{2\alpha-3} - c(1+t)^{\alpha-2}\Big) \tag{4.13}$$

Hence:

$$\frac{\xi'_\alpha(t)}{2} + \lambda_\alpha(t)\lambda'_\alpha(t) = \alpha(1-\alpha)\Big(2(2\alpha-1)(1+t)^{2\alpha-3} - c(1+t)^{2\alpha-2}\Big) \tag{4.14}$$

We also have $\lambda_r(t)\lambda'_r(t) = 4r^2(r-1)(1+t)^{2r-3}$, and

$$\xi'_r(t) = 4r(1-r)(4r-1)(1+t)^{2r-3} \tag{4.15}$$

Hence:

$$\frac{\xi'_r(t)}{2} + \lambda_r(t)\lambda'_r(t) = 2r(1-r)(2r-1)(1+t)^{2r-3} \tag{4.16}$$

Noting that $2r-3 = \alpha-2$, $2r-1 = \alpha$ and $2(1-r) = (1-\alpha)$, this rewrites: $\frac{\xi'_r(t)}{2} + \lambda_r(t)\lambda'_r(t) = \alpha(1-\alpha)r(1+t)^{\alpha-2}$

Hence, we have:

$$\frac{\xi'_\alpha(t)}{2} + \lambda_\alpha(t)\lambda'_\alpha(t) + \frac{\xi'_r(t)}{2} + \lambda_r(t)\lambda'_r(t) = \alpha(1-\alpha)\Big(2(2\alpha-1)(1+t)^{2\alpha-3} - (c-r)(1+t)^{\alpha-2}\Big) \tag{4.17}$$

- If $\alpha \in \{0,1\}$, then $\frac{\xi'_\alpha(t)}{2} + \lambda_\alpha(t)\lambda'_\alpha(t) + \frac{\xi'_r(t)}{2} + \lambda_r(t)\lambda'_r(t) = 0$
- If $\alpha \in (0,1)$: for the last term to be negative for all $t \geqslant 0$, we need $c - r \geqslant 2(2\alpha-1)$, i.e. $\underline{c \geqslant r + 2(2\alpha-1)}$

Finally, we need to ensure that the functions $\xi_\alpha(.)$ and $\xi_r(.)$ are non-negative. Recall that this is the case if $c \geqslant 4r-1$ and $c \geqslant 4\alpha - 1$. As $r \geqslant \alpha$, this reduces to $c \geqslant 4r-1$, which rewrites $\underline{c \geqslant 2\alpha+1}$. Furthermore, it turns out that this condition on c is stronger than the previous underlined conditions on c.

Hence, if $c \geqslant 2\alpha + 1$, we have $\mathcal{E}'(t) \leqslant 0$. Therefore, for all $t \geqslant 0$, we have $\mathcal{E}(t) \leqslant \mathcal{E}(0)$. As all the terms in $\mathcal{E}(t)$ are positive, we also have $\mathcal{E}_r(t) \leqslant \mathcal{E}(0)$. In particular, since $2r = 1+\alpha$:

$$(1+t)^{1+\alpha}(F(x(t)) - F^*) \leqslant \mathcal{E}(0) \tag{4.18}$$

Expliciting the value of $\mathcal{E}(0)$ leads to the desired result.

□

Remark 4.1. *Notice in particular that:*

1. *For the case where $\alpha = 1$, we recover the condition $c \geqslant 3$ required to obtain the convergence rate $O(\frac{1}{t^2})$.*
2. *For the case where $\alpha = 0$: most of the results in the literature obtain a convergence rate $O(\frac{1}{t})$ at the average of the iterates $f(\frac{1}{t}\int_0^t x(s)ds)$. To the best of our knowledge, this is the first result in continuous time with this convergence rate at the last iterate. It only requires that $c \geqslant 1$. However, in the discrete case, [18] proved that with appropriate choices of time-dependent sequences for the corresponding discrete heavy ball method, we can obtain the rate $O(\frac{1}{k})$ at the last iterate.*

4.2 Unperturbed case with geometry

We first announce new results on the convergence of the function values at solutions of (4.2) with additional geometrical assumptions, but without perturbations. We distinguish the perturbed and unperturbed cases because the latter will require some additional assumptions.

Theorem 4.2. *Let $\gamma_1 > 2$, $\gamma_2 > 2$. Note $r = \frac{1+\alpha}{2}$. Suppose x is a solution to the ODE (4.2). If F is coercive and satisfies $\mathbf{H_1}(\gamma_1)$ and $\mathbf{H_2}(\gamma_2)$,*

1. *If $\gamma_1 \leqslant \gamma_2$ and $c \geqslant \frac{\gamma_1+2}{\gamma_1-2}r$ then*

$$F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2r\gamma_2}{\gamma_2-2}}}\right) \tag{4.19}$$

2. *If $\gamma_2 \leqslant \gamma_1$, then for any $\gamma \in [\gamma_2, \gamma_1]$, if $c \geqslant \frac{\gamma+2}{\gamma-2}r$,*

$$F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2r\gamma}{\gamma-2}}}\right) \tag{4.20}$$

Let $\lambda(.)$ and $\xi(.)$ be two real-valued functions, where $\lambda(.)$ is non-negative. With $r = \frac{1+\alpha}{2}$. Consider the following energy:

$$\mathcal{E}(t) = (1+t)^{2r}(F(x(t)) - F^*) + \frac{1}{2}\|\lambda(t)(x(t)-x^*) + (1+t)^r\dot{x}(t)\|^2 + \frac{\xi(t)}{2}\|x(t)-x^*\|^2 \tag{4.21}$$

Noting:

$$\begin{aligned} a(t) &= (1+t)^r(F(x(t)) - F^*) \\ b(t) &= \frac{1}{2(1+t)^r}\|\lambda(t)(x(t)-x^*) + (1+t)^r\dot{x}(t)\|^2 \\ c(t) &= \frac{1}{2(1+t)^r}\|x(t)-x^*\|^2 \end{aligned} \tag{4.22}$$

we have: $\mathcal{E}(t) = (1+t)^r(a(t) + b(t) + \xi(t)c(t))$. We also define:

$$\mathcal{H}(t) = (1+t)^p\mathcal{E}(t) \tag{4.23}$$

The proofs of our theorems rely on the following lemma.

Lemma 4.1. *Let $\gamma \geqslant 1$. If F satisfies the hypothesis $\mathbf{H}_1(\gamma)$ and if $\xi(t) = \lambda(t)(\lambda(t) + r(1+t)^{r-1} - c(1+t)^{r-\alpha}) - \lambda'(t)(1+t)^r$ and $\lambda(t) = \omega(1+t)^{r-1}$, then $\xi(t) = w((w+1)(1+t)^{2r-2} - c)$ and:*

$$\begin{aligned} \mathcal{H}'(t) \leqslant (1+t)^p\Bigg(&[(2r+p-\gamma_1\omega)(1+t)^{r-1}]a(t) + [(p+2\omega+2r)(1+t)^{r-1} - 2c(1+t)^{r-\alpha}]b(t) \\ &+ \omega[[p(\omega+1) + 2(2\omega+1)(r-1) - 2\omega(\omega+r)](1+t)^{3r-3} - c(p-2\omega)(1+t)^{r-1}]c(t)\Bigg) \end{aligned} \tag{4.24}$$

This lemma is proved in the appendix B.

Proof. Taking $w = \frac{2r}{\gamma_1-2}$ and $p = \frac{4r}{\gamma_1-2}$:

$$\mathcal{H}'(t) \leqslant (1+t)^p\left(2\Big(\frac{\gamma+2}{\gamma-2}r(1+t)^{r-1} - c(1+t)^{r-\alpha}\Big)b(t) + \omega(p+2)(r-1)(1+t)^{3r-3}c(t)\right) \tag{4.25}$$

Since $r \leqslant 1$, $\omega(p+2)(r-1)(1+t)^{3r-3} \leqslant 0$, hence:

$$\mathcal{H}'(t) \leqslant 2\Big(\frac{\gamma+2}{\gamma-2}r(1+t)^{r-1} - c(1+t)^{r-\alpha}\Big)b(t)(1+t)^p \tag{4.26}$$

And since $r-\alpha \geqslant r-1$ and $c \geqslant \frac{\gamma+2}{\gamma-2}r$ we have: $\frac{\gamma+2}{\gamma-2}r(1+t)^{r-1} - c(1+t)^{r-\alpha} \leqslant 0$. Hence, $\mathcal{H}'(t) \leqslant 0$ for all $t \geqslant 0$.

We will now use similar reasoning as in [9] to prove the results of our theorem.

Case $\gamma_1 \leqslant \gamma_2$. Since $\mathcal{H}'(t) \leqslant 0$, for any choice of x^* in the set of minimizers X^*, the function $\mathcal{H}$ is bounded above and since the set of minimizers is bounded because F is coercive, there exists $A > 0$ and t_0 such that for all choices of x^* in X^*:

$$\mathcal{H}(t_0) \leqslant A \tag{4.27}$$

Hence for all $x^* \in X^*$ and $t \geqslant t_0$, $\mathcal{H}(t) \leqslant A$. Hence

$$(1+t)^{\frac{2r\gamma_1}{\gamma_1-2}}(F(x(t)) - F^*) \leqslant \frac{|\xi(t)|}{2}(1+t)^{\frac{4r}{\gamma_1-2}}\|x(t) - x^*\|^2 + A \tag{4.28}$$

We have $\xi(t) = \frac{2r}{\gamma-2}((\frac{2r}{\gamma-2}+1)(1+t)^{2r-2} - c)$. Since $c \geqslant \frac{\gamma_1+2}{\gamma_1-2}r \geqslant \frac{2r}{\gamma-2}+1$, we have $|\xi(t)| = \frac{2r}{\gamma-2}(c - (\frac{2r}{\gamma-2}+1))(1+t)^{2r-2}$. Hence, $|\xi(t)| \leqslant \frac{2rc}{\gamma-2}$. Therefore:

$$(1+t)^{\frac{2r\gamma_1}{\gamma_1-2}}(F(x(t)) - F^*) \leqslant \frac{rc}{\gamma-2}(1+t)^{\frac{4r}{\gamma_1-2}}\|x(t) - x^*\|^2 + A \tag{4.29}$$

And since this is verified for all $x^* \in X^*$:

$$(1+t)^{\frac{2r\gamma_1}{\gamma_1-2}}(F(x(t)) - F^*) \leqslant \frac{rc}{\gamma-2}(1+t)^{\frac{4r}{\gamma_1-2}}d(x(t), X^*)^2 + A \tag{4.30}$$

We set $v(t) = (1+t)^{\frac{4r}{\gamma_2-2}}d(x(t), X^*)^2$ Then

$$(1+t)^{\frac{2r\gamma_1}{\gamma_1-2}}(F(x(t)) - F^*) \leqslant \frac{rc}{\gamma-2}(1+t)^{\frac{4r}{\gamma_1-2}-\frac{4r}{\gamma_2-2}}v(t) + A \tag{4.31}$$

Since F satisfies $\mathbf{H}_2(\gamma_2)$, there exists K > 0 such that

$$K((1+t)^{-\frac{4r}{\gamma_2-2}}v(t))^{\frac{\gamma_2}{2}} \leqslant F(x(t)) - F^* \tag{4.32}$$

i.e

$$Kv(t)^{\frac{\gamma_2}{2}}(1+t)^{\frac{-2r\gamma_2}{\gamma_2-2}} \leqslant F(x(t)) - F^* \tag{4.33}$$

Hence

$$K(1+t)^{\frac{2r\gamma_1}{\gamma_1-2}}(1+t)^{-\frac{2r\gamma_2}{\gamma_2-2}}v(t)^{\frac{\gamma_2}{2}} \leqslant (1+t)^{\frac{2r\gamma_1}{\gamma_1-2}}(F(x(t)) - F^*) \tag{4.34}$$

Back to (4.31), this yields:

$$K(1+t)^{\frac{2r\gamma_1}{\gamma_1-2}}(1+t)^{-\frac{2r\gamma_2}{\gamma_2-2}}v(t)^{\frac{\gamma_2}{2}} \leqslant \frac{rc}{\gamma-2}(1+t)^{\frac{4r}{\gamma_1-2}-\frac{4r}{\gamma_2-2}}v(t) + A \tag{4.35}$$

Hence

$$Kv(t)^{\frac{\gamma_2}{2}} \leqslant \frac{rc}{\gamma-2}v(t) + A(1+t)^{\frac{4r}{\gamma_2-2}-\frac{4r}{\gamma_1-2}} \tag{4.36}$$

Which, since $\gamma_1 \leqslant \gamma_2$, means that v is bounded. Therefore, from (4.31) we deduce that there exists $B > 0$ such that:

$$F(x(t)) - F^* \leqslant B(1+t)^{\frac{-2r\gamma_2}{\gamma_2-2}} + A(1+t)^{\frac{-2r\gamma_1}{\gamma_1-2}} \tag{4.37}$$

Since $\gamma_1 \leqslant \gamma_2$, we have $\frac{-r2\gamma_2}{\gamma_2-2} \geqslant \frac{-2r\gamma_1}{\gamma_1-2}$. Hence $F(x(t)) - F^* = O(t^{-\frac{2r\gamma_2}{\gamma_2-2}})$.

Case $\gamma_2 \leqslant \gamma_1$. Since F satisfies $\mathbf{H}_1(\gamma_1)$ and $\mathbf{H}_2(\gamma_2)$, with $\gamma_2 \leqslant \gamma_1$, then for any $\gamma \in [\gamma_2, \gamma_1]$ we also have that F satisfies $\mathbf{H}_2(\gamma)$ and $\mathbf{H}_1(\gamma)$. We can thus only consider the case when $\gamma_1 = \gamma_2 = \gamma$, as done in the previous case. □

Remark 4.2. *In the previous theorem, we imposed* $c \geqslant \frac{\gamma_1+2}{\gamma_1-2}$. *In fact, this condition is only required for the case where* $\alpha = 1$ *as it ensures that our energy is decreasing in (4.26). For the other values of* $\alpha \in [0, 1)$, *the inequality (4.26) is asymptotically ensured by the fact that* $c > 0$ *and* $r - \alpha > r - 1$, *as there exists some* $t_1 \geqslant 0$ *such that for all* $t \geqslant t_1$, $\frac{\gamma+2}{\gamma-2}r(1+t)^{r-1} \leqslant c(1+t)^{r-\alpha}$.

4.3 Perturbed case with geometry

With all the previous proofs, the following theorem is almost a corollary of **Theorem** 3.6 and **Theorem** 4.2.

Theorem 4.3. *Let $\gamma_1 > 2$, $\gamma_2 > 2$. Note $r = \frac{1+\alpha}{2}$. Let $\eta = \frac{r\gamma_1}{\gamma_1 - 2}$ and assume that $\int_{t_0}^{+\infty}(1+t)^{\eta}\|g(t)\|dt < +\infty$. If F admits a unique minimizer and satisfies $\mathbf{H_1}(\gamma_1)$ and $\mathbf{H_2}(\gamma_2)$, if $\gamma_2 \leqslant \gamma_1$, then for any $\gamma \in [\gamma_2, \gamma_1]$, we have*

$$F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2r\gamma}{\gamma-2}}}\right) \tag{4.38}$$

The analysis with additional geometrical assumptions will *not* use the Lyapunov function developed by [14]. Instead, it will rely on a Lyapunov function closer to, e.g. [15].

Proof. In addition to the energy functions $\mathcal{E}$ and $\mathcal{H}$ defined in the proof of **Theorem** 4.2, we define

$$\mathcal{G}(t) = \mathcal{H}(t) + \int_t^T \langle (1+s)^{\frac{p}{2}}\big(\lambda(s)(x(s)-x^*) + (1+s)^r\dot{x}(s)\big), (1+s)^{\frac{p+2r}{2}} g(s)\rangle ds \tag{4.39}$$

We set $\xi(t) = \lambda(t)(\lambda(t) + r(1+t)^{r-1} - c(1+t)^{r-\alpha}) - \lambda'(t)(1+t)^r$. Let $\omega > 0$. Let $\lambda(t) = \omega(1+t)^{r-1}$. we then have, using the fact that F satisfies $\mathbf{H}_1(\gamma_1)$:

$$\begin{aligned}\mathcal{H}_p'(t) \leqslant (1+t)^p\Bigg(&[(2r+p-\gamma_1\omega)(1+t)^{r-1}]a(t) + [(p+2\omega+2r)(1+t)^{r-1} - 2c(1+t)^{r-\alpha}]b(t)\\ &+ \omega[[p(\omega+1) + 2(2\omega+1)(r-1) - 2\omega(\omega+r)](1+t)^{3r-3} - c(p-2\omega)(1+t)^{r-1}]c(t)\Bigg)\end{aligned} \tag{4.40}$$

We refer the reader to the appendix for a detailed derivation of equation (4.24). We demonstrate using the exact same arguments as the proof of **Theorem** 4.2 that with $p = \frac{4r}{\gamma_1-2}$ and $\lambda(t) = \frac{2r}{\gamma_1-2}t^{r-1}$, we have $\mathcal{G}'(t) \leqslant 0$. Noticing that $|\xi(t)| \leqslant \frac{2rc}{\gamma_1-2}$, the rest of the proof follows exactly as the one of **Theorem** 3.6, where we replace ξ by $\frac{2rc}{\gamma_1-2}$. □

We can also deduce a result on the fact that the trajectory of any solution x of the ODE (4.2) is finite.

A Differentiating the energy function for Theorem 3.5, proof of Lemma 3.2

1. **Differentiating $\mathcal{E}$**

   We have:

$$\begin{aligned}\mathcal{E}'(t) &= 2t(F(x(t)) - F^*) + t^2\langle \nabla F(x(t)), \dot{x}(t)\rangle\\ &+ \langle \lambda(x(t)-x^*) + t\dot{x}(t), \lambda\dot{x}(t) + \dot{x}(t) + t\ddot{x}(t)\rangle\\ &+ \xi\langle \dot{x}(t), x(t)-x^*\rangle\end{aligned} \tag{A.1}$$

   And since x verifies the ODE (3.1), we have:

$$\begin{aligned}(\lambda+1)\dot{x}(t) + t\ddot{x}(t) &= (\lambda+1)\dot{x}(t) - c\dot{x}(t) - t\nabla F(x(t)) + tg(t)\\ &= (\lambda+1-c)\dot{x}(t) - t\nabla F(x(t)) + tg(t)\end{aligned} \tag{A.2}$$

   Hence:

$$\begin{aligned}\mathcal{E}'(t) &= 2a(t) + (\xi + \lambda(\lambda+1-c)\langle \dot{x}(t), x(t)-x^*\rangle\\ &- \lambda t\langle \nabla F(x(t)), x(t)-x^*\rangle + t(\lambda+1-c)\|\dot{x}(t)\|^2\\ &+ \langle tg(t), \lambda(x(t)-x^*) + t\dot{x}(t)\rangle\end{aligned} \tag{A.3}$$

   Noticing that:

$$\frac{1}{t}\|\lambda(x(t)-x^*) + t\dot{x}(t)\|^2 = t\|\dot{x}(t)\|^2 + 2\lambda\langle \dot{x}(t), x(t)-x^*\rangle + \frac{\lambda^2}{t}\|x(t)-x^*\|^2 \tag{A.4}$$

We deduce that

$$\begin{aligned}\mathcal{E}'(t) &= 2a(t) - \lambda t\langle \nabla F(x(t)), x(t)-x^*\rangle + (\xi - \lambda(\lambda+1-c))\langle \dot{x}(t), x(t)-x^*\rangle \\ &+ \frac{\lambda+1-c}{t}\|\lambda(x(t)-x^*) + t\dot{x}(t)\|^2 - \frac{\lambda^2(\lambda+1-c)}{t}\|x(t)-x^*\|^2 \\ &+ \langle tg(t), \lambda(x(t)-x^*) + t\dot{x}(t)\rangle\end{aligned} \tag{A.5}$$

i.e.

$$\begin{aligned}\mathcal{E}'(t) &= 2a(t) - \lambda t\langle \nabla F(x(t)), x(t)-x^*\rangle + (\xi - \lambda(\lambda+1-c))\langle \dot{x}(t), x(t)-x^*\rangle \\ &+ 2(\lambda+1-c)b(t) - 2\lambda^2(\lambda+1-c)c(t) + \langle tg(t), \lambda(x(t)-x^*) + t\dot{x}(t)\rangle\end{aligned} \tag{A.6}$$

Chosing $\xi = \lambda(\lambda+1-c)$:

$$\begin{aligned}\mathcal{E}'(t) &= 2a(t) - \lambda t\langle \nabla F(x(t)), x(t)-x^*\rangle + 2(\lambda+1-c)b(t) - 2\lambda^2(\lambda+1-c)c(t) \\ &+ \langle tg(t), \lambda(x(t)-x^*) + t\dot{x}(t)\rangle\end{aligned} \tag{A.7}$$

Since F satisfies $\mathbf{H_1}(\gamma)$:

$$\begin{aligned}\mathcal{E}'(t) &\leqslant (2-\lambda\gamma)a(t) + 2(\lambda+1-c)b(t) - 2\lambda^2(\lambda+1-c)c(t) \\ &+ \langle tg(t), \lambda(x(t)-x^*) + t\dot{x}(t)\rangle\end{aligned} \tag{A.8}$$

2. **Differentiating $\mathcal{H}$ and $\mathcal{G}$**

   Recall that $\mathcal{H}(t) = t^p\mathcal{E}(t)$. Hence $\mathcal{H}'(t) = t^{p-1}(p\mathcal{E}(t) + t\mathcal{E}'(t))$. And since $\mathcal{E}(t) = t(a(t) + b(t) + \xi c(t))$, we deduce from (A.8) that:

$$\begin{aligned}\mathcal{H}'(t) &\leqslant t^p((2-\gamma\lambda+p)a(t) + (2\lambda+2-2c+p)b(t) + \lambda(\lambda+1-c)(-2\lambda+p)c(t)) \\ &+ \langle t^{p+1}g(t), \lambda(x(t)-x^*) + t\dot{x}(t)\rangle\end{aligned} \tag{A.9}$$

   And since:

$$\mathcal{G}(t) = \mathcal{H}(t) + \int_t^T \langle s^{\frac{p}{2}}(\lambda(x(s)-x^*) + s\dot{x}(s)), s^{\frac{p+2}{2}}g(s)\rangle ds \tag{A.10}$$

   We deduce that:

$$\mathcal{G}'(t) \leqslant t^p((2-\gamma\lambda+p)a(t) + (2\lambda+2-2c+p)b(t) + \lambda(\lambda+1-c)(-2\lambda+p)c(t)) \tag{A.11}$$

B Differentiating the energy function of Theorem 4.3, proof of Lemma 4.1

1. **Differentiating $\mathcal{E}$**

   Recall that:

$$\mathcal{E}(t) = (1+t)^{2r}(F(x(t)) - F^*) + \frac{1}{2}\|\lambda(t)(x(t)-x^*) + (1+t)^r\dot{x}(t)\|^2 + \frac{\xi(t)}{2}\|x(t)-x^*\|^2 \tag{B.1}$$

   Then

$$\begin{aligned}\mathcal{E}'(t) &= 2r(1+t)^{2r-1}(F(x(t)) - F^*) + (1+t)^{2r}\langle \nabla F(x(t)), \dot{x}(t)\rangle \\ &+ \langle \lambda(t)(x(t)-x^*) + (1+t)^r\dot{x}(t), \lambda'(t)(x(t)-x^*) + \lambda(t)\dot{x}(t) + r(1+t)^{r-1}\dot{x}(t) + (1+t)^r\ddot{x}(t)\rangle \\ &+ \frac{\xi'(t)}{2}\|x(t)-x^*\|^2 + \xi(t)\langle \dot{x}(t), x(t)-x^*\rangle\end{aligned} \tag{B.2}$$

   But since x satisfies the ODE (4.2), we have:

$$\lambda(t)\dot{x}(t) + r(1+t)^{r-1}\dot{x}(t) + (1+t)^r\ddot{x}(t) = (\lambda(t) + r(1+t)^{r-1} - c(1+t)^{r-\alpha})\dot{x}(t) - (1+t)^r\nabla F(x(t)) + (1+t)^r g(t) \tag{B.3}$$

   Hence, back to (B.2):

$$\begin{aligned}\mathcal{E}'(t) &= 2r(1+t)^{r-1}(F(x(t)) - F^*) - \lambda(t)\langle \nabla F(x(t)), x(t)-x^*\rangle + [\lambda(t)\lambda'(t) + \frac{\xi'(t)}{2}]\|x(t)-x^*\|^2 \\ &+ [\xi(t) + \lambda(t)(\lambda(t) + r(1+t)^{r-1} - (1+t)^{r-\alpha}) + \lambda'(t)(1+t)^r]\langle \dot{x}(t), x(t)-x^*\rangle \\ &+ (1+t)^r(\lambda(t) + r(1+t)^{r-1} - c(1+t)^{r-\alpha})\|\dot{x}(t)\|^2 + \langle (1+t)^r g(t), \lambda(t)(x(t)-x^*) + (1+t)^r\dot{x}(t)\rangle\end{aligned} \tag{B.4}$$

Noticing that

$$\frac{1}{(1+t)^r}\|\lambda(t)(x(t)-x^*)+(1+t)^r\dot{x}(t)\|^2 = (1+t)^r\|\dot{x}(t)\|^2 + 2\lambda(t)\langle \dot{x}(t), x(t)-x^*\rangle + \frac{\lambda^2(t)}{(1+t)^r}\|x(t)-x^*\|^2 \tag{B.5}$$

i.e.

$$(1+t)^r\|\dot{x}(t)\|^2 = 2b(t) - 2\lambda^2(t)c(t) - 2\lambda(t)\langle x(t)-x^*, \dot{x}(t)\rangle \tag{B.6}$$

and replacing in (B.4)

$$\begin{aligned}\mathcal{E}'(t) &= 2r(1+t)^{r-1}(F(x(t))-F^*) - \lambda(t)\langle \nabla F(x(t)), x(t)-x^*\rangle + 2(\lambda(t)+r(1+t)^{r-1}-(1+t)^{r-\alpha})b(t)\\ &\quad + [(1+t)^r(2\lambda(t)\lambda'(t)+\xi'(t)) - 2\lambda^2(t)(\lambda(t)+r(1+t)^{r-1}-c(1+t)^{r-\alpha})]c(t)\\ &\quad + [\xi(t)-\lambda(t)(\lambda(t)+r(1+t)^{r-1}-(1+t)^{r-\alpha})+\lambda'(t)(1+t)^r]\langle \dot{x}(t), x(t)-x^*\rangle\\ &\quad + \langle (1+t)^r g(t), \lambda(t)(x(t)-x^*)+(1+t)^r\dot{x}(t)\rangle\end{aligned} \tag{B.7}$$

Hence, setting $\underline{\xi(t) = \lambda(t)(\lambda(t)+r(1+t)^{r-1}-c(1+t)^{r-\alpha})-\lambda'(t)(1+t)^r}$:

$$\begin{aligned}\mathcal{E}'(t) &= 2r(1+t)^{r-1}(F(x(t))-F^*) - \lambda(t)\langle \nabla F(x(t)), x(t)-x^*\rangle + 2(\lambda(t)+r(1+t)^{r-1}-c(1+t)^{r-\alpha})b(t)\\ &\quad + [(1+t)^r(2\lambda(t)\lambda'(t)+\xi'(t)) - 2\lambda^2(t)(\lambda(t)+r(1+t)^{r-1}-c(1+t)^{r-\alpha})]c(t)\\ &\quad + \langle (1+t)^r g(t), \lambda(t)(x(t)-x^*)+(1+t)^r\dot{x}(t)\rangle\end{aligned} \tag{B.8}$$

Since F satisfies $\mathbf{H}_1(\gamma_1)$ and we assumed $\lambda(.)$ to be positive, we have

$$\begin{aligned}\mathcal{E}'(t) &\leqslant (2r(1+t)^{r-1}-\gamma_1\lambda(t))(F(x(t))-F^*) + 2(\lambda(t)+r(1+t)^{r-1}-c(1+t)^{r-\alpha})b(t)\\ &\quad + [(1+t)^r(2\lambda(t)\lambda'(t)+\xi'(t)) - 2\lambda^2(t)(\lambda(t)+r(1+t)^{r-1}-c(1+t)^{r-\alpha})]c(t)\\ &\quad + \langle (1{+})t^r g(t), \lambda(t)(x(t)-x^*)+(1+t)^r\dot{x}(t)\rangle\end{aligned} \tag{B.9}$$

2. **Differentiating $\mathcal{H}$ and $\mathcal{G}$.**
Recall that $\mathcal{H}(t) = (1+t)^p\mathcal{E}(t)$. Hence $\mathcal{H}'(t) = (1+t)^{p-1}(p\mathcal{E}(t)+(1+t)\mathcal{E}'(t))$ Hence

$$\begin{aligned}\mathcal{H}'(t) &= (1+t)^{p-1}(p(1+t)^r(a(t)+b(t)+\xi(t)c(t))+(1+t)\mathcal{E}'(t))\\ &= (1+t)^p(p(1+t)^{r-1}(a(t)+b(t)+\xi(t)c(t))+\mathcal{E}'(t))\end{aligned} \tag{B.10}$$

Therefore:

$$\begin{aligned}\mathcal{H}'(t) &\leqslant (1+t)^p\Big([(2r+p)(1+t)^{r-1}-\gamma_1\lambda(t)]a(t) + [p(1+t)^{r-1}+2(\lambda(t)+r(1+t)^{r-1}-c(1+t)^{r-\alpha})]b(t)\\ &\quad + [p(1+t)^{r-1}\xi(t)+(1+t)^r(2\lambda(t)\lambda'(t)+\xi'(t)) - 2\lambda^2(t)(\lambda(t)+r(1+t)^{r-1}-c(1+t)^{r-\alpha})]c(t)\Big)\\ &\quad + \langle (1+t)^{r+p} g(t), \lambda(t)(x(t)-x^*)+(1+t)^r\dot{x}(t)\rangle\end{aligned} \tag{B.11}$$

And since $\mathcal{G}(t) = \mathcal{H}(t) + \int_t^T \langle (1+s)^{\frac{p}{2}}\big(\lambda(s)(x(s)-x^*)+(1+s)^r\dot{x}(s)\big), (1+s)^{\frac{p+2r}{2}}g(s)\rangle ds$, we have

$$\begin{aligned}\mathcal{G}'(t) &\leqslant (1+t)^p\Big([(2r+p)(1+t)^{r-1}-\gamma_1\lambda(t)]a(t) + [p(1+t)^{r-1}+2(\lambda(t)+r(1+t)^{r-1}-c(1+t)^{r-\alpha})]b(t)\\ &\quad + [p(1+t)^{r-1}\xi(t)+(1+t)^r(2\lambda(t)\lambda'(t)+\xi'(t)) - 2\lambda^2(t)(\lambda(t)+r(1+t)^{r-1}-c(1+t)^{r-\alpha})]c(t)\Big)\end{aligned} \tag{B.12}$$

Let $\omega > 0$. We chose $\lambda(t) = \omega(1+t)^{r-1}$. Then replacing $\lambda(.)$ and $\xi(.)$ by their explicit values, we have:

$$\begin{aligned}\mathcal{H}'(t) &\leqslant (1+t)^p\Bigg([(2r+p-\gamma_1\omega)(1+t)^{r-1}]a(t) + [(p+2\omega+2r)(1+t)^{r-1}-2c(1+t)^{r-\alpha}]b(t)\\ &\quad + \omega[[p(\omega+1)+2(2\omega+1)(r-1)-2\omega(\omega+r)](1+t)^{3r-3}-c(p-2\omega)(1+t)^{r-1}]c(t)\Bigg)\end{aligned} \tag{B.13}$$

C Proof of the Grönwall-Bellman Lemma 3.3

For the ease of the reader, we recall here the lemma.

Lemma C.1. *Let $c > 0$, $q, r \leqslant 1$. Let $T > t_0 > 0$. Let $f \in L_1(t_0, T; \mathbb{R})$ such that $f \geq 0$ a.e on (t_0, T). Suppose $m : [t_0; T] \to \mathbb{R}$ is continuous. Suppose that:*

$$m(t) \leqslant c + \int_{t_0}^{t} [m^q(s) + m^r(s)] f(s) ds \tag{C.1}$$

Then

$$x(t) \leqslant c + A(t) \exp\big(\int_{t_0}^{t} B(s) ds\big) \tag{C.2}$$

where, for any $K > 0$:

$$\begin{aligned} A(t) &= \int_{t_0}^{t} [(1-q)K^q + qK^{q-1}c + (1-r)K^r + rK^{r-1}c] f(s) ds \\ B(t) &= (qK^{q-1} + rK^{r-1}) f(t) \end{aligned} \tag{C.3}$$

Proof. We define $w(t) = c + \int_{t_0}^{t} [m^q(s) + m^r(s)] f(s) ds$. Then $m(t) \leqslant w(t)$. Then:

$$w(t) \leqslant c + \int_{t_0}^{t} [w^q(s) + w^r(s)] f(s) ds \tag{C.4}$$

Noting $u(t) = \int_{t_0}^{t} [w^q(s) + w^r(s)] f(s) ds$. We have $w(t) \leqslant c + u(t)$. In the following, we will use this small lemma.

Lemma C.2. *Assume that $a \geq 0$, $1 \geq q \geq 0$, then for any $K > 0$*

$$a^q \leqslant qK^{q-1}a + (1-q)K^q \tag{C.5}$$

for any K > 0

Since $w(t) \leqslant c + u(t)$, we have using the previous lemma, we have for any $K > 0$

$$w^q(t) \leqslant [c + u(t)]^q \leqslant qK^{q-1}[c + u(t)] + (1-q)K^q \tag{C.6}$$

The same way:

$$w^r(t) \leqslant [c + u(t)]^r \leqslant rK^{r-1}[c + u(t)] + (1-r)K^r \tag{C.7}$$

Hence replacing in the expression of $u(t)$:

$$u(t) \leqslant \int_{t_0}^{t} [qK^{q-1}[c + u(s)] + (1-q)K^q + rK^{r-1}[c + u(s)] + (1-r)K^r \tag{C.8}$$

Which rewrites:

$$u(t) \leqslant A(t) + \int_{t_0}^{t} B(s) u(s) ds \tag{C.9}$$

where $A(t)$ and $B(t)$ are defined as in the lemma. Then using Gronwall-Bellman's Lemma:

$$u(t) \leqslant A(t) \exp\big(\int_{t_0}^{t} B(s) ds\big) \tag{C.10}$$

And since $m(t) \leqslant w(t) \leqslant c + u(t)$:

$$m(t) \leqslant c + A(t) \exp\big(\int_{t_0}^{t} B(s) ds\big) \tag{C.11}$$

Which achieves the proof of the lemma. □

References

[1] Yurii Nesterov — *A method for solving the convex programming problem with convergence rate* $O(\frac{1}{k^2})$*, Dokl. Akad. Nauk SSSR, 1983*

[2] Guillaume Garrigos, Lorenzo Rosasco, Silvia Villa — *Convergence of the Forward-Backward Algorithm: Beyond the Worst Case with the Help of Geometry, arXiv:1703.09477, 2017.*

[3] Hedy Attouch, Zaki Chbani — *Fast inertial dynamics and FISTA algorithms in convex optimization. Perturbation Aspects, arXiv:1507.01367.*

[4] Hedy Attouch, Juan Peypouquet — *The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than* $O(\frac{1}{k^2})$*, SIAM Journal on Optimization 26 (2016), no. 3, 1824-1834*

[5] Ashia C. Wilson, Benjamin Recht, Michael I. Jordan — *A Lyapunov Analysis of Momentum Methods in Optimization, arXiv:1611.02635*

[6] Guilherme França, Daniel P. Robinson, René Vidal — *ADMM and Accelerated ADMM as Continuous Dynamical Systems , ICML 2018 - Proceedings of the 35th International Conference on Machine Learning*

[7] Antonin Chambolle, Charles Dossal — *On the convergence of the iterates of FISTA , Journal of Optimization Theory and Applications, Springer Verlag, 2015, Volume 166 (Issue 3), pp.25*

[8] J-F. Aujol and C. Dossal — *Optimal rate of convergence of an ode associated to the fast gradient descent schemes for* $b > 0$*, Hal Preprint, June 2017.*

[9] J-F. Aujol, C. Dossal, A. Rondepierre — *Optimal convergence rates for Nesterov acceleration, arXiv:1805.05719*, 2018.

[10] Guillaume Garrigos — *Descent dynamical systems and algorithms for tame optimization and multi-objective problems, PhD thesis, Hal Preprint, 2015,*

[11] Mark Schmidt, Nicolas Le Roux, Francis Bach — *arXiv:1109.2415*, 2011.

[12] Weijie Su, Stephen Boyd, Emmanuel J. Candes — *A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights, arXiv:1503.01243*, 2015.

[13] Imen Ben Hassen — *Decay estimates to equilibrium for some asymptotically autonomous semilinear evolution equations, hal-00392110*, 2009.

[14] Mounir Balti, Ramzi May — *Asymptotic for the perturbed heavy ball system with vanishing damping term, arXiv:1609.00135*, 2016.

[15] Hedy Attouch, Zaki Chbani, Hassan Riahi — *Rate of convergence of the Nesterov accelerated gradient method in the subcritical case* $\alpha \leq 3$*, arXiv:1706.05671*, 2016.

[16] Fangcui Jiang, Fanwei Meng — *Explicit bounds on some new nonlinear integral inequalities with delay, Journal of Computational and Applied Mathematics 205 (2007) 479 486*, 2007.

[17] B. T. Polyak — *Some methods of speeding up the convergence of iteration methods, USSR Computational Mathematics and Mathematical Physics, 4:117*, 1964.

[18] E. Ghadimi, H. R. Feyzmahdavian and M. Johansson — *"Global convergence of the Heavy-ball method for convex optimization," 2015 European Control Conference (ECC), Linz, 2015, pp. 310-315.*