Convergence rates of the Heavy-Ball method with Łojasiewicz property

J-F. Aujol [1], Ch. Dossal [2] and A. Rondepierre [2,3]

[1] Univ. Bordeaux, Bordeaux INP, CNRS, IMB, UMR 5251, F-33400 Talence, France.
[2] IMT, Univ. Toulouse, INSA Toulouse, France.
[3] LAAS, Univ. Toulouse, CNRS, Toulouse, France.

Jean-Francois.Aujol@math.u-bordeaux.fr, {Charles.Dossal,Aude.Rondepierre}@insa-toulouse.fr

Abstract In this paper, a joint study of the behavior of solutions of the Heavy Ball ODE and Heavy Ball type algorithms is given. Since the pioneering work of B.T. Polyak [38], it is well known that such a scheme is very efficient for C^2 strongly convex functions with Lipschitz gradient. But much less is known when only growth conditions are considered. Depending on the geometry of the function to minimize, convergence rates for convex functions, with some additional regularity such as quasi-strong convexity, or strong convexity, were recently obtained in [12]. Convergence results with much weaker assumptions are given in the present paper: namely, linear convergence rates when assuming a growth condition (which amounts to a Łojasiewicz property in the convex case). This analysis is firstly performed in continuous time for the ODE, and then transposed for discrete optimization schemes. In particular, a variant of the Heavy Ball algorithm is proposed, which converges geometrically whatever the parameters choice, and which has the best state of the art convergence rate for first order methods to minimize composite non smooth convex functions satisfying a Łojasiewicz property.

Key-words Lyapunov function, rate of convergence, ODEs, optimization, Łojasiewicz property, Heavy Ball method.

1 Introduction

Let us consider the following unconstrained minimization problem

$$\min_{x\in\mathbb{R}^n} F(x) \tag{1}$$

where $F : \mathbb{R}^n \to \mathbb{R}$ is a convex function having a non-empty set of minimizers X^*. Depending on the assumptions on F, there exist plenty of algorithms to solve this problem. The convergence of these algorithms and their rate of convergence may highly depend on the properties of F. If F is differentiable with a L-Lipschitz continuous gradient, a classical algorithm is the explicit Gradient Descent (GD) which ensures the convergence to a minimizer x^* of F. The convergence rate of (GD) i.e. the decay of $F(x_n)-F(x^*)$, may be quite slow for some convex functions but it is actually exponential when F is strongly convex. Most algorithms such as inertial algorithms (Heavy Ball or Nesterov acceleration schemes), ensure a better convergence rate if F is known to be strongly convex. Most of the time the parameter μ of strong convexity is a parameter of the algorithm.

Nevertheless, the strong convexity assumption may not be satisfied in many numerical problems. Strong convexity is indeed a strong hypothesis. It turns out that many functions are not strongly convex but still have good properties for optimization. A weaker property ensuring an exponential convergence of the Gradient Descent is actually the Łojasiewicz property with an exponent $\theta = \frac{1}{2}$. This property is equivalent for convex functions to a quadratic growth condition around the set of minimizers. Roughly speaking a function F satisfies the Łojasiewicz property with parameter $\theta \in [0, 1)$ if around its set of minimizers we have:

$$F(x) - F^* \geqslant \frac{\mu}{2} d(x, X^*)^{\frac{1}{\theta}}. \tag{2}$$

that is, if $\theta = \frac{1}{2}$, F grows at most quadratically around the set of minimizers. In the present work we focus on the class of convex functions satisfying a Łojasiewicz property with a special interest to the subclass of functions satisfying a Łojasiewicz property with an exponent $\frac{1}{2}$, which gathers a large set of functions used in image processing and statistics. A well-known example of such a function is the LASSO function:

$$F(x) = \frac{1}{2} \|Ax - y\|^2 + \lambda \|x\|_1 \tag{3}$$

If $Ker(A) \neq \{0\}$, the function F is not strongly convex and it may not be quasi-strongly convex [31, Definition 1] but it belongs to the class of convex functions having the Łojasiewicz property with an exponent $\theta = \frac{1}{2}$ [18].

Following Polyak [38] a joint study of the dynamical system (4)

$$\ddot{x}(t) + \alpha \dot{x}(t) + \nabla F(x(t)) = 0 \tag{4}$$

and of an associated algorithm that is a suitable discretization of the ODE (4), is proposed.

The study of the continuous dynamic to have of a better understanding of algorithms is actually a cornerstone of many recent works in optimization, see e.g. [42, 6, 11, 2, 3, 41, 21, 10, 12].

The analysis of Polyak [38] rests on a C^2 hypothesis on F and most following studies by Nesterov [34], Ghadimi [24], Siegel [41] or by the authors of the present work [12] necessitate a strong or at least weak strong convexity assumption. These previous results do not apply under a sole growth condition. Moreover the value of the parameter α in (4) must be chosen accordingly to strong convexity parameter of F to ensure the convergence and the rate of the algorithm.

The main contributions of this work can be summarized as follows.

1. Linear convergence rates for the ODE (4) associated to the Heavy ball in the case of a growth condition and uniqueness of a minimizer (which are the best known rates under these hypotheses), see Theorems 1 and 2.

2. New convergence rates for an algorithm inspired from ODE (4) in the case of a convex function with a L-Lipschitz gradient (or in the composite case) in the case when F satisfies any Łojasiewicz property, whatever the parameters choice, see Theorems 6 and 8.

3. Linear convergence rates for an algorithm inspired from ODE (4) under the additional assumption of uniqueness of the minimizer (which are the best known rates under these hypotheses), see Theorems 7 and 9.

More precisely, we extend the analysis of the first point with a non vanishing perturbation term and to the associated monotone inclusion. A conclusion of this first study is that the value of the friction parameter α in (4) ensuring the best decay rate of $F(x(t)) - F(x^*)$ is different from the one obtained by Polyak [38] for C^2 strongly convex functions and obtained in [12] for weak strong convex functions.

The second contribution is algorithmic since we propose an inertial algorithm which is different from the classical Heavy Ball algorithm of Polyak and which convergence is ensured for a large set of parameters. No lower bounds on the growth parameters are needed to ensure the convergence of the iterates or a fast decay rate of $F(x_n) - F(x^*)$ under Łojasiewicz properties. This is a main difference with most previous works in which the parameter α must be chosen small enough with respect to the strong convexity parameter of F. The given decay rate of $F(x_n) - F(x^*)$ are polynomials and are the one we expect according to the continuous analysis of Bolte et al. [17] when the Łojasiewicz parameter θ is smaller than $\frac{1}{2}$ and is linear when $\theta = \frac{1}{2}$.

To understand the third contribution, the reader must be aware of a key point : all exponential decay rates of various algorithms are not equivalent. We refer the reader to Definitions 5 and 6 in Section 3 for the exact definition of decay rate in the case of exponential convergence. In many practical problems, the value of the parameter μ in the growth condition (2) is very small. When F is continuously differentiable with a L-Lipschitz gradient, the condition number $\kappa = \frac{\mu}{L}$ may be very small. This number drives the actual rate of many algorithms minimizing F. For the gradient descent we can expect that $F(x_n) - F^* = \mathcal{O}(q^n)$ with $q = 1 - \kappa$. This rate is actually achieved for quadratic functions and we cannot expect a better rate for the gradient descent in general. If κ is really small, this decay may be really slow and not visible on numerical experiments. It is one of the reasons why the Forward Backward algorithm that can be seen as a extension of Gradient descent to composite functions $F = f + g$, is numerically slow on the LASSO problem even if the decay is exponential.

The main advantage of inertial methods is that they are able to provide rates such that $1 - q \approx C\sqrt{\kappa}$ which is much better when κ is close to 0. We will say that the exponential decay is slow if $1 - q \approx C\kappa$ and is fast if $1 - q \approx C\sqrt{\kappa}$.

In Theorems 7 and 9, we achieve a fast exponential decay for a suitable choice of parameters depending on the growth parameter μ. Hence, these last theorems ensure better decay rates under stronger assumptions than Theorems 6 and 8 and complete the analysis under growth conditions.

The paper is structured as follows: in Section 2 , the definitions and the links between all the classical geometrical hypotheses on functions F are presented. In Section 3, we present the state of the art and extensive results for the dynamical system (4). These results are extended to the associated differential inclusion. Section 4 is dedicated to the discrete scheme, and we propose here a new scheme that has the best convergence rate for minimizing non smooth convex functions satisfying a Łojasiewicz property. Section 5 presents some numerical experiments. Most of the proofs are postponed to Section 6.

2 Preliminaries: local geometry of convex functions

Let us first recall some basic notations and definitions. We assume that $\mathbb{R}^n$ is equipped with the Euclidean scalar product $\langle \cdot, \cdot \rangle$ and the associated norm $\| \cdot \|$. As usual $B(x^*, r)$ denotes the open Euclidean ball with center $x^* \in \mathbb{R}^n$ and radius $r > 0$. For any real subset $X \subset \mathbb{R}^n$, the Euclidean distance d is defined as:

$$\forall x \in \mathbb{R}^n, \; d(x, X) = \inf_{y \in X} \|x - y\|.$$

If the function $F : \mathbb{R}^n \to \mathbb{R}$ is differentiable, its gradient is denoted by $\nabla F(x)$. When F is assumed only convex, its (convex) subdifferential denoted by $\partial F(x)$, is defined by:

$$\partial F(x) = \{s \in \mathbb{R}^n \mid \forall y \in \mathbb{R}^n,\ F(y) \geqslant F(x) + \langle s, y - x\rangle\}.$$

We remind here that $\partial F(x)$ is a closed convex set at any $x \in int(dom(F))$. If F is additionally proper lower semicontinuous then for all $x \in int(dom(F))$, its subdifferential $\partial F(x)$ is also non-empty and bounded.

In this paper we consider the general class of convex functions having a Łojasiewicz property [29, 30], a key tool in the mathematical analysis of continuous and discrete dynamical systems. Initially introduced to prove the convergence of the trajectories for the gradient flow of analytic functions, an extension to nonsmooth functions has been proposed by Bolte et al. in [16, 17]:

Definition 1 (The Łojasiewicz property for convex functions)**.** *Let $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous convex function with $X^* = \operatorname{argmin} F \neq \emptyset$. The function F has a Łojasiewicz property if for any minimizer x^*, there exist $\theta \in [0, 1)$, $c > 0$, $\varepsilon > 0$ such that:*

$$\forall x \in B(x^*, \varepsilon),\ c\,(F(x) - F(x^*))^{\theta} \leqslant d(0, \partial F(x)), \tag{5}$$

or equivalently if F satisfies a local growth condition i.e. if for any minimizer $x^* \in X^*$, there exist $r \geqslant 1$ and $\mu > 0$ such that

$$\exists \varepsilon > 0, \forall x \in B(x^*, \varepsilon),\ F(x) - F(x^*) \geqslant \frac{\mu}{2} d(x, X^*)^r. \tag{6}$$

The equivalence between the Łojasiewicz property (5) and the local growth condition (6) has been proved in [17, Theorem 5]. More generally, for convex functions, a growth condition on an arbitrary set $\Omega \subset \mathbb{R}^n$ implies a Łojasiewicz property on Ω, but the equivalence is guaranteed only on ∂F-invariant sets, see [23, Proposition 3.2] for more details.

In this paper we consider a slightly stronger growth condition, namely a growth condition defined in the neighborhood of the set of minimizers as follows:

Definition 2 (Growth condition $\mathcal{G}^r_\mu$)**.** *Let $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous convex function such that $X^* = \operatorname{argmin} F \neq \emptyset$. The function F satisfies the growth condition $\mathcal{G}^r_\mu$ for some exponent $r \geqslant 1$ and some real constant $\mu > 0$ if there exists $\varepsilon > 0$ such that for all $x \in \mathbb{R}^n$, we have:*

$$d(x, X^*) \leqslant \varepsilon \ \Rightarrow\ \frac{\mu}{2} d(x, X^*)^r \leqslant F(x) - F^*.$$

The condition $\mathcal{G}^r_\mu$ is said to be global if for all $x \in \mathbb{R}^n$, we have: $F(x) - F^* \geqslant \frac{\mu}{2} d(x, X^*)^r$.

The growth condition $\mathcal{G}^r_\mu$ can be seen as a sharpness assumption on the function F characterizing functions behaving at least as $\|\cdot\|^r$ in the neighborhood of their minimizers. Note that $\mathcal{G}^r_\mu$ implies the growth condition $\mathcal{G}^{r'}_\mu$ for all $r' > r \geqslant 1$. If F admits a unique minimizer, the growth condition $\mathcal{G}^r_\mu$ is nothing more than the local growth condition (6). More generally, when the set X^* of the minimizers is a compact set (which is guaranteed assuming for example F coercive), a straightforward extension of [4, Lemma 1] ensures that the local growth condition (6) and the condition $\mathcal{G}^r_\mu$ are actually equivalent.

More precisely combining [23, Proposition 3.2] and [4, Lemma 1], we have:

Lemma 1. *Let $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous and coercive convex function satisfying a Łojasiewicz property, or equivalently some local growth condition. Then* *there exist $r \geqslant 1$ and $\mu > 0$ such that F satisfies the growth condition $\mathcal{G}_\mu^r$ for some $\varepsilon > 0$ and has the Łojasiewicz property with the exponent $\theta = 1 - \frac{1}{r}$ on the set $\{x \in \mathbb{R}^n \mid d(x, X^*) \leqslant \varepsilon\}$ i.e.:*

$$\forall x \in \mathbb{R}^n,\ d(x, X^*) \leqslant \varepsilon \ \Rightarrow c\,(F(x) - F(x^*))^\theta \leqslant d(0, \partial F(x))$$

where $c = r\left(\frac{2}{\mu}\right)^{\frac{1}{r}}$.

Throughout the paper, we have a special interest in the class $\mathcal{G}_\mu^2$ of convex functions satisfying a quadratic growth condition $\mathcal{G}_\mu^2$ and the sub-class $\mathcal{PL}_\mu$ of functions satisfying a global version of $\mathcal{G}_\mu^2$, or equivalently the Łojasiewicz property with an exponent $\theta = \frac{1}{2}$:

Definition 3 (The Polyak-Łojasiewicz property $\mathcal{PL}_\mu$)**.** *Let $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous convex function with $X^* = \operatorname{argmin} F \neq \emptyset$. Let $F^* = \inf F$. The function F has the Polyak-Łojasiewicz property $\mathcal{PL}_\mu$ for some $\mu > 0$ if and only if:*

$$\forall x \in \mathbb{R}^n\ , d(0, \partial F(x))^2 \geqslant 2\mu\,(F(x) - F^*)\,.$$

Note that convex functions having a strong minimizer in the sense of [5, Section 3.3], satisfy the $\mathcal{PL}_\mu$ property.

Let us now recall some state-of-the-art functional classes in the analysis of the Heavy Ball system: the set $\mathcal{Q}_\mu$ of quadratic functions, the classes $\mathcal{S}_{\mu,L}^{p,1}$ for $p \in \{1, 2\}$, of μ-strongly convex functions of class C^p having a L-Lipschitz gradient and the class of just μ-strongly convex $\mathcal{S}_\mu$. We will also refer to the class $q\mathcal{S}_\mu$ of μ-quasi-strongly convex functions introduced by I. Necoara et al. in [31]:

Definition 4 (Quasi-strong convexity [31, Definition 1])**.** *A continuously differentiable function $F : \mathbb{R}^n \to \mathbb{R}$ is μ-quasi-strongly convex if for any $x \in \mathbb{R}^n$:*

$$\langle \nabla F(x), x - x^* \rangle \geqslant F(x) - F(x^*) + \frac{\mu}{2}\|x - x^*\|^2$$

where x^* denotes the projection of x onto the set $X^* = \operatorname{argmin} F \neq \emptyset$.

The quasi-strong convexity is a relaxation of the strong convexity, but it does not imply the convexity of F or the uniqueness of the minimizer. However observe that the quasi-strong convexity of F implies that F satisfies the Polyak-Łojasiewicz property $\mathcal{PL}_\mu$ [31]:

Lemma 2. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function with $X^* = \operatorname{argmin} F \neq \emptyset$. If F is μ-strongly quasi-convex then F satisfies the Polyak-Łojasiewicz property $\mathcal{PL}_\mu$.*

All these sets of functions are sub-classes of functions satisfying a quadratic growth condition $\mathcal{G}_\mu^2$. The relations between them are extensively detailed in [31, Section 3] and can be summarized as follows: if F is a differentiable convex function admitting a unique minimizer then:

$$\mathcal{Q}_\mu \Longrightarrow \mathcal{S}_{\mu,L}^{2,1} \Longrightarrow \mathcal{S}_{\mu,L}^{1,1} \Longrightarrow \mathcal{S}_\mu \Longrightarrow q\mathcal{S}_\mu \Longrightarrow \mathcal{PL}_\mu \Longrightarrow \mathcal{G}_\mu^2. \tag{7}$$

Note that the quadratic growth condition alone is not equivalent to the quasi-strong convexity. More information on the geometry of the function F to minimize, such as a global quadratic growth condition, is needed. For example we can prove that functions satisfying both a "flatness" assumption and a global quadratic growth condition $\mathcal{G}_\mu^2$ (i.e. functions at least as flat as $\|\cdot\|^\gamma$ with $\gamma > 1$, and as sharp as $\|\cdot\|^2$) as in [40], are quasi-strongly convex:

Lemma 3. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a convex continuously differentiable function with $X^* = \text{argmin}\, F \neq \emptyset$ and $F^* = \inf\ F$. If F satisfies a flatness assumption:*

$$\forall x \in \mathbb{R}^n,\ F(x) - F^* \leqslant \frac{1}{\gamma}\langle \nabla F(x), x - x^* \rangle \tag{8}$$

for some $\gamma > 1$ and the Polyak-Łojasiewicz property $\mathcal{PL}_\mu$ for some real constant $\mu > 0$ then necessarily $\gamma \leqslant 2$ and F is $(\gamma - 1)\mu$-quasi-strongly convex.

Proof. The proof is straightforward. Observe that combining the flatness assumption (8) with the quadratic growth condition, we get:

$$\begin{aligned}\forall x \in \mathbb{R}^n,\ \langle \nabla F(x), x - x^* \rangle \ &\geqslant\ F(x) - F^* + (\gamma - 1)\left(F(x) - F^*\right)\\ &\geqslant\ F(x) - F^* + (\gamma - 1)\frac{\mu}{2} d(x, X^*)^2.\end{aligned}$$

□

Finally we introduce the class $\mathcal{G}^2_{\mu,L}$ of functions satisfying the growth condition $\mathcal{G}^2_\mu$ and having a L-Lipschitz continuous gradient that will considered in the analysis of the discrete scheme in Section 4. Without further assumptions, $\mathcal{G}^2_{\mu,L}$ is not a sub-class of quasi-strongly convex functions that are extensively considered in [12] for the analysis of the Heavy ball system. However assuming a little more, namely that F satisfies a global quadratic growth condition, we have:

Lemma 4. *[12, Lemma 2] Let $F : \mathbb{R}^n \to \mathbb{R}$ be a convex continuously differentiable function with $X^* = \text{argmin}\, F \neq \emptyset$. If F has a L-Lipschitz continuous gradient for some $L > 0$ and has the Polyak-Łojasiewicz property $\mathcal{PL}_\mu$ for some $\mu > 0$ then F is $\frac{\mu^2}{L}$-quasi-strongly convex.*

3 The Heavy Ball dynamical system

In his seminal work [38] B.T. Polyak studies the Heavy Ball dynamical system (4), the properties of its solutions and the rates of convergence of $F(x(t)) - F^*$ for functions belonging to the class $\mathcal{S}^{2,1}_{\mu,L}$ of μ-strongly convex functions of class C^2 having a L-Lipschitz gradient. This first analysis is the key to analyze the convergence of the optimization scheme described in the same paper. More recently Su et al. [42] show that the Nesterov acceleration scheme can be seen as a discretization of another ODE:

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla F(x(t)) = 0. \tag{9}$$

Following this approach a large amount of papers have proposed to study ODEs in different settings to provide some new properties or convergence rates of the existing algorithms, or even to provide new algorithms. An important point to keep in mind is that it is easy to find an ODE ensuring a fast decay of $F(x(t)) - F^*$ for any set of functions using time scaling for example. The difficulty in this joint study of ODE and algorithm is to be able to find a discrete and explicit scheme ensuring the same decay rate that the one reached in the continuous setting. To set some definitions we would say that:

Definition 5. *A dynamical system provides a fast exponential decay on a functional set $\mathcal{F}_\mu$ if there exists $\delta > 0$ such that for any $F \in \mathcal{F}_\mu$ and any solution x of the associated ODE*

$$F(x(t)) - F^* = \mathcal{O}(e^{-\delta\sqrt{\mu}t}).$$

Set $\mathcal{F}_\mu$	References	Optimal values of α	Exponential rate of $F(x(t)) - F^*$
$\mathcal{S}_\mu^{2,1}$	Polyak [38]	$2\sqrt{\mu}$	$2\sqrt{\mu}$
$\mathcal{S}_\mu^{1,1}$	Siegel [41]	$2\sqrt{\mu}$	$\sqrt{\mu}$
$q\mathcal{S}_\mu$ and uniqueness of minimizer	ADR[12]	$3\sqrt{\frac{\mu}{2}}$	$\sqrt{2\mu}$
C^2 and $\mathcal{G}_\mu^2$	Bégout et al. [15]	$-$	μ

Table 1: Rate of convergence for the Heavy Ball ODE

Definition 6. *A numerical scheme provides a low exponential decay on a functional set $\mathcal{F}_\mu$ if there exists $\delta > 0$ such that for any $F \in \mathcal{F}_\mu$ and any solution x of the associated ODE*

$$F(x(t)) - F^* = \mathcal{O}(e^{-\delta\mu t}).$$

It turns out that a good discretization of a dynamical system with a fast decay may provide a fast optimization algorithm. Table 1 summarizes the decays that are achieved for the Heavy ball system on various sets of functions. We can observe that on the classes $\mathcal{S}_\mu^{2,1}$, $\mathcal{S}_\mu^{1,1}$ and $q\mathcal{S}_\mu$ with uniqueness of the minimizer, the given decays are fast exponential decays. In Polyak [38] and ADR [12] the exponential decays are also given for any α respectively on the classes $\mathcal{S}_\mu^{2,1}$ and $\mathcal{S}_\mu^{1,1}$ and the decays are proved to be optimal: [12] exhibits some functions for which these rates are reached.

Let us now consider the class of convex functions having a Łojasiewicz property. We first recall that the exponential decay of solutions of the Gradient Descend flow:

$$\dot{x}(t) + \nabla F(x(t)) = 0 \tag{10}$$

for convex functions in $\mathcal{PL}_\mu$ (i.e. having the Polyak-Łojasiewicz property) is straightforward. Indeed defining $\mathcal{E}(t) = F(x(t)) - F^*$, we get $\mathcal{E}'(t) = -\|\nabla F(x(t))\|^2 \leqslant -2\mu(F(x(t)) - F^*)$ ensuring that:

$$F(x(t)) - F^* \leqslant (F(x(t_0)) - F^*)\, e^{-2\mu(t-t_0)}$$

which is a slow exponential decay rate. These rates are actually exact for quadratic functions. In Section 4 we will see that the associated discrete algorithms inherit of these fast or low rates and this difference partially explains the good behavior of inertial algorithms with respect to the steepest descent in large dimension problems for which μ is really small.

For the Nesterov ODE (9), in [11] the authors prove that the decay rate for convex functions admitting a unique minimizer and satisfying a growth condition $\mathcal{G}_\mu^2$ provides only a polynomial rate. For the Heavy Ball system, there are few results. For flat geometries i.e. for convex functions satisfying any local growth condition $\mathcal{G}_\mu^r$ with $r > 2$, Bégout et al. in [15] prove the convergence of the trajectory to a minimizer x^* of F at a polynomial rate provided that F is of class C^2:

$$\|x(t) - x^*\| = \mathcal{O}\left(t^{-\frac{1}{r-2}}\right). \tag{11}$$

Without further information on the local geometry of F around its set of minimizers, applying [11, Lemma 2.2] to $\gamma = 1$ (which corresponds to the sole convexity assumption), this implies the same polynomial decay on the values $F(x(t)) - F^*$. If F behaves like $\|x\|^r$, with $r > 2$ around its set of minimizers, the convergence rate provided by [15] is more accurately in $\mathcal{O}\left(t^{-\frac{r}{r-2}}\right)$.

For sharp geometries i.e. for convex functions satisfying some quadratic growth condition $\mathcal{G}_\mu^2$, or equivalently having a Łojasiewicz property with an exponent $\frac{1}{2}$, Bégout et al. in [15] and

Polyak et al. [39] respectively prove that the decay is exponential when F is of class C^2. In [39] the exact decay is not given but our computations indicate slow exponential rates in these two approaches. However note that these two approaches do not assume the uniqueness of the minimizer of F.

The main contribution of this part is Theorem 1 stating that the Heavy Ball ODE ensures a fast decay rate on the set of differentiable convex functions satisfying a quadratic growth condition $\mathcal{G}_\mu^2$ and having a unique minimizer. We also prove that these decays are preserved for the perturbed Heavy Ball ODE assuming sufficient integrability conditions on the perturbation. Moreover, it should be noticed that the optimal value of α is still different from the previous ones. One can observe in the Figure 1 that the optimal value of parameter α in the Heavy Ball

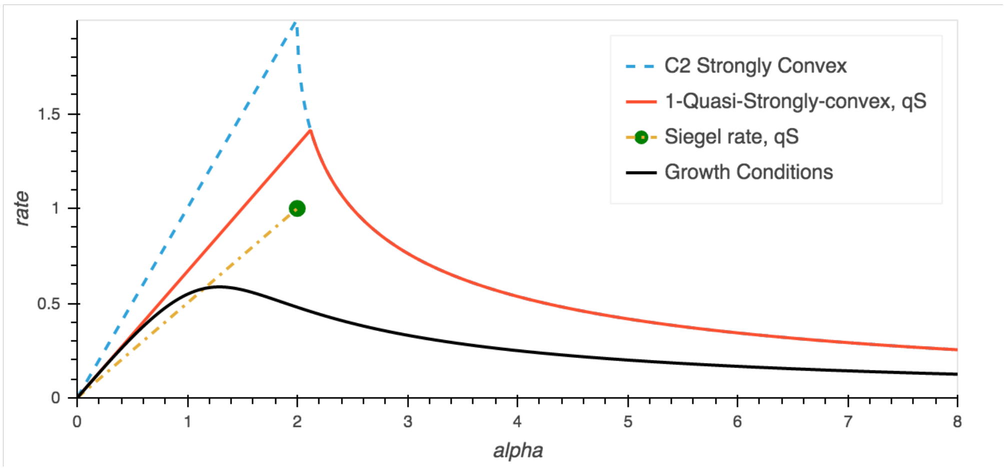

Figure 1: Decay rates that can be achieved depending on the geometrical hypotheses made on F for $\mu = 1$ the parameter of the growth condition. The stronger the regularity assumption, the better the rate. The optimal value of α depends on the regularity assumption.

ODE (12) depends on the exact hypothesis made of F. The exponential rate is lower under growth conditions than under strong convexity or quasi-strong convexity assumptions, which is not surprising since this hypothesis is weaker than the other ones.

3.1 The differentiable case

Let $F : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable convex function having a unique minimizer x^* and $F^* = \inf\ F$. We consider the Heavy Ball dynamical system:

$$\ddot{x}(t) + \alpha\dot{x}(t) + \nabla F(x(t)) = 0 \tag{12}$$

for any $t \geqslant t_0$, where $t_0 > 0$. Throughout the paper, we assume that, for any given initial conditions $(x_0, v_0) \in \mathbb{R}^n \times \mathbb{R}^n$, the Cauchy problem associated with the ODE (12), admits a unique global solution satisfying $(x(0), \dot{x}(0)) = (x_0, v_0)$. This is guaranteed in particular when the gradient of F is Lipschitz on bounded subsets of $\mathbb{R}^n$ [21].

We can now state the main theorem for the Heavy Ball ODE (12) whose proof is detailed in Section 6.1. Note that to analyze the asymptotic behavior of $F(x(t)) - F^*$, we only need local geometric assumptions on F since the trajectory x converges to some minimizer of F provided that F is a convex function of class C^1 admitting at least one minimizer and that the ODE (12) has a unique global solution, see [1, Theorem 2.1] for more details:

Theorem 1. *Let $\alpha > 0$ and $t_0 > 0$. Let $F : \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function such that: $X^* = \operatorname{argmin} F \neq \emptyset$ and $F^* = \inf\ F$. Let x be the solution of the ODE (12) for given initial conditions $(x(t_0), \dot{x}(t_0)) = (x_0, v_0)$.*

If F satisfies a quadratic growth condition $\mathcal{G}^2_\mu$ for some real constant $\mu > 0$ and admits a unique minimizer then

$$F(x(t)) - F^* = \mathcal{O}\left(e^{-\delta_{\alpha,\mu} t}\right)$$

where $\delta_{\alpha,\mu}$ is the (unique) real root of the polynomial:

$$P_{\alpha,\mu}(\delta) = \delta^3 - 3\alpha\delta^2 + (3\mu + 2\alpha^2)\delta - 2\mu\alpha$$

on the interval $[0, \frac{2\alpha}{3}[$. In particular: $\delta_{\alpha,\mu} \leqslant (2 - \sqrt{2})\sqrt{\mu}$ and the best rate $\delta_{\alpha,\mu} = (2 - \sqrt{2})\sqrt{\mu}$ is achieved for $\alpha = (2 - \frac{\sqrt{2}}{2})\sqrt{\mu}$.

Now assuming that the quadratic growth condition is global, we also get non-asymptotic convergence rates for this class of functions:

Corollary 1. *Under the assumptions of Theorem 1, if F satisfies a global quadratic growth condition $\mathcal{G}^2_\mu$ and admits a unique minimizer then for any $\delta \leqslant \delta_{\alpha,\mu}$, we have:*

$$\forall t \geqslant t_0,\ F(x(t)) - F^* \leqslant \frac{4\mu + (2\alpha - \delta + 2\sqrt{\mu})^2}{\mu(4 - \alpha\delta + \delta^2)} M_0 e^{-\delta(t-t_0)}$$

where: $M_0 = F(x_0) - F^* + \frac{1}{2}\|v_0\|^2$. If $\alpha = (2 - \frac{\sqrt{2}}{2})\sqrt{\mu}$, the best rate is given by:

$$F(x(t)) - F^* \leqslant (\frac{11}{2} - 2\sqrt{2}) M_0 e^{-(2-\sqrt{2})\sqrt{\mu}(t-t_0)}. \tag{13}$$

Remark 1. *Using the open-source Python library Sympy for symbolic mathematics, we are able to provide some explicit equivalents of the real root $\delta_{\alpha,\mu}$. For a fixed $\mu > 0$, we have:*

$$\delta_{\alpha,\mu} \sim \frac{2\alpha}{3},\ \alpha \to 0^+, \qquad \delta_{\alpha,\mu} \sim \frac{\mu}{\alpha},\ \alpha \to +\infty.$$

Moreover studying the variations of the polynomial $P_{\alpha,\mu}$ we deduce a criterion to easily compare a given exponent $\delta \in [0, \frac{2\alpha}{3}[$ with the best possible rate $\delta_{\alpha,\mu}$:

$$\delta \leqslant \delta_{\alpha,\mu} \Longleftrightarrow P_{\alpha,\mu}(\delta) \leqslant 0. \tag{14}$$

Remark 2. *Using the Python library Sympy, we actually have an explicit (but barely readable) expression of $\delta_{\alpha,\mu}$. For α large enough (numerically larger than $0.15\mu^2$), we have:*

$$\delta_{\alpha,\mu} = \mu - \frac{18^{\frac{1}{3}}}{3} \frac{\mu^2 - 3\alpha}{d(\alpha,\mu)} - \frac{12^{\frac{1}{3}}}{6} d(\alpha,\mu)$$

where: $d(\alpha,\mu) = \left(9\alpha\mu + \sqrt{3}\sqrt{108\alpha^3 - 81\alpha^2\mu^2 + 36\mu^4\alpha - 4\mu^6}\right)^{\frac{1}{3}}$. Otherwise for small values of α, we have:

$$\delta_{\alpha,\mu} = \alpha - \frac{18^{\frac{1}{3}}}{3} \frac{\alpha^2 - 3\mu}{d(\mu,\alpha)} - \frac{12^{\frac{1}{3}}}{6} d(\mu,\alpha).$$

To our best knowledge the only other result in the literature providing a fast exponential decay for the values $F(x(t))-F^*$ can be found in [40]. Indeed, applying [40, Theorem 3.2] with $K_2=2\mu$ and $\gamma=1$ (corresponding to the sole convexity assumption), we get an exponential decay in $\mathcal{O}\left(e^{-\frac{\mu}{2\alpha}}\right)$ provided that $\alpha \geqslant \sqrt{\frac{3}{2}}\sqrt{\mu}$, which can be seen as a fast exponential decay when α is chosen proportional to $\sqrt{\mu}$ as in Theorem 1. Since $P_{\alpha,\mu}(\frac{\mu}{2\alpha}) \leqslant 0$, we easily show that this rate is always worse than the one provided by Theorem 1.

Consider now the subclass $\mathcal{G}_\mu^r$ with $r\in(1,2)$ of differentiable convex functions being sharper than quadratic in the neighborhood of its minimizers. Since $\mathcal{G}_\mu^r$ is a sub-class of $\mathcal{G}_\mu^2$, Theorem 1 applies and we deduce that the exponent in the exponential decay is at least in $\delta_{\alpha,\mu}$. A natural question is then: can this convergence rate be improved for this class of functions ? In [12], the authors prove that the exponent $\frac{2\alpha}{3}$ is optimal on the sub-class $\mathcal{S}_\mu$ of strongly convex functions when $\alpha < 3\sqrt{\frac{\mu}{2}}$. In fact, we can prove that the exponential decay is at least in $\frac{2\alpha}{3}$ for any $\alpha>0$ on the sub-class $\mathcal{G}_\mu^r$ with $r\in(1,2)$, providing thus an upper bound for the convergence rate:

Proposition 1. *Let $F:\mathbb{R}^n\to\mathbb{R}$ be a continuously differentiable convex function with $X^* = \operatorname{argmin} F \neq \emptyset$ and $F^* = \inf F$. Let $\alpha>0$ and $t_0>0$. Let x be the solution of the ODE (12) for given initial conditions $(x(t_0),\dot{x}(t_0))=(x_0,v_0)$.*

Let $r\in(1,2)$. If F admits a unique minimizer and satisfies a local growth condition $\mathcal{G}_\mu^r$ for some $\mu>0$ then:

$$\forall \lambda < \frac{2\alpha}{3},\ F(x(t))-F^* = \mathcal{O}\left(e^{-\lambda t}\right).$$

Proof. Let us consider the Lyapunov energy:

$$\mathcal{E}(t) = (F(x(t))-F^*) + \frac{1}{2}\left\|\lambda(x(t)-x^*)+\dot{x}(t)\right\|^2 + \frac{\xi}{2}\left\|x(t)-x^*\right\|^2. \tag{15}$$

Differentiating the energy (15) and using the ODE (12), we have:

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

Using the parameters $\lambda=\frac{2\alpha}{3}$ and $\xi=-\frac{2\alpha^2}{9}$ we have the following equality (the careful reader will have noticed that the parameters of the Lyapunov function are not the same as the ones used in Theorem 1):

$$\forall t\geqslant t_0,\ \mathcal{E}'(t)+\frac{2\alpha}{3}\mathcal{E}(t) = \frac{2\alpha^3}{27}\|x(t)-x^*\|^2 + \frac{2\alpha}{3}\left[F(x(t))-F^*-\langle\nabla F(x(t)),x(t)-x^*\rangle\right] \tag{16}$$

It follows that if F is convex we get

$$\forall t\geqslant t_0,\ \mathcal{E}'(t)+\frac{2\alpha}{3}\mathcal{E}(t) \leqslant \frac{2\alpha^3}{27}\|x(t)-x^*\|^2.$$

Observe that using the growth condition $\mathcal{G}_\mu^r$ with $r\in(1,2)$, there exists $t_1\geqslant t_0$ such that:

$$\forall t\geqslant t_1,\ \mathcal{E}(t)\geqslant F(x(t))-F^*+\frac{\xi}{2}\|x(t)-x^*\|^2 \geqslant \frac{\mu}{2}\|x(t)-x^*\|^r - \frac{\alpha^2}{9}\|x(t)-x^*\|^2.$$

Let $\epsilon\in(0,\frac{2\alpha}{3})$. Then:

$$\begin{aligned}\forall t\geqslant t_1,\ \mathcal{E}'(t)+(\frac{2\alpha}{3}-\epsilon)\mathcal{E}(t) &\leqslant \left(\frac{2\alpha^3}{27}+\epsilon\frac{\alpha^2}{9}\right)\|x(t)-x^*\|^2 - \epsilon\frac{\mu}{2}\|x(t)-x^*\|^r\\ &\leqslant \left((\frac{2\alpha^3}{27}+\epsilon\frac{\alpha^2}{9})\|x(t)-x^*\|^{2-r}-\epsilon\frac{\mu}{2}\right)\|x(t)-x^*\|^r.\end{aligned}$$

We can then conclude that there exists $t_2 > t_1$ such that for all $t \geqslant t_2$, we have:

$$\forall \epsilon \in (0, \frac{2\alpha}{3}),\ \mathcal{E}'(t) + (\frac{2\alpha}{3} - \epsilon)\mathcal{E}(t) \leqslant 0,$$

hence: $\forall \lambda < \frac{2\alpha}{3},\ \mathcal{E}(t) = \mathcal{O}\left(e^{-\lambda t}\right)$. To deduce the control on the values $F(x(t)) - F^*$ we use the growth condition $\mathcal{G}_\mu^r$:

$$\begin{aligned} \forall t \geqslant t_2,\ \mathcal{E}(t) &\geqslant F(x(t)) - F^* + \frac{\xi}{2}\|x(t) - x^*\|^2 = F(x(t)) - F^* - \frac{2\alpha^2}{9}\|x(t) - x^*\|^2 \\ &\geqslant \left(1 - \frac{4\alpha^2}{9\mu}\|x(t) - x^*\|^{2-r}\right)(F(x(t)) - F^*) \geqslant \frac{1}{2}(F(x(t)) - F^*) \end{aligned} \tag{17}$$

for t large enough, and we finally obtain the expected convergence rate.

□

Remark 3. *Consider the particular case of convex functions behaving like $F(x) = \frac{\mu}{2}\|x\|^r$ with $r \in (1, 2)$ in the neighborhood of their minimizers. Then F naturally satisfies the growth condition $\mathcal{G}_\mu^r$, or equivalently a Łojasiewicz property with an exponent $\theta = 1 - \frac{1}{r}$. By a strategy similar to the one used in Proposition 1 with the set of parameters of Theorem 1, and replacing the convexity inequality by the equality: $\langle \nabla F(x), x - x^* \rangle = r(F(x) - F^*)$, we can prove that the decay is even better than $\frac{2\alpha}{3}$, namely:*

$$\forall \lambda < \frac{2r\alpha}{r+2},\ F(x(t)) - F^* = \mathcal{O}\left(e^{-\lambda t}\right). \tag{18}$$

Let us conclude this section by a stability study of the Heavy Ball dynamical system. We consider the perturbed version of the Heavy Ball ODE:

$$\forall t \geqslant t_0,\ \ddot{x}(t) + \alpha\dot{x}(t) + \nabla F(x(t)) = g(t) \tag{19}$$

where $t_0 > 0$ and $g : [t_0, +\infty[$ is an integrable source term that can be interpreted as an external perturbation exerted on the system. We assume that the Cauchy problem associated with the ODE (19), admits a unique global solution satisfying $(x(0), \dot{x}(0)) = (x_0, v_0)$ for any given initial conditions $(x_0, v_0) \in \mathbb{R}^n \times \mathbb{R}^n$. This is guaranteed in particular when the gradient of F is Lipschitz on bounded subsets of $\mathbb{R}^n$ [25, 28].

Theorem 2 provides sufficient integrability conditions on the perturbation g in order to guarantee that the convergence rate established in Theorem 1 is preserved. Note that as in the unperturbed case, the trajectory $x(t)$ solution of the perturbed ODE (19) converges (weakly in a Hilbert space) to a minimizer of F, see [25, Theorem 3.3], so that we only need local geometric assumptions on F to analyze the asymptotic behavior of $F(x(t)) - F^*$:

Theorem 2. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable convex function with $X^* = \operatorname{argmin} F \neq \emptyset$ and $F^* = \inf F$. Let $\alpha > 0$ and $t_0 > 0$. Let x be the solution of the ODE (19) for given initial conditions $(x(t_0), \dot{x}(t_0)) = (x_0, v_0)$.*

Assume that F satisfies a quadratic growth condition $\mathcal{G}_\mu^2$ for some real constant $\mu > 0$ and admits a unique minimizer. Let $\delta_{\alpha,\mu}$ be defined as in Theorem 1 and $\delta \leqslant \delta_{\alpha,\mu}$. If:

$$J_0 = \int_{t_0}^{+\infty} e^{\delta t}\|g(t)\|dt < +\infty,$$

then:

$$F(x(t)) - F^* = \mathcal{O}\left(e^{-\delta(t-t_0)}\right).$$

As in the unperturbed case, assuming that F additionally satisfies a global quadratic condition $\mathcal{G}^2_\mu$, we can provide non-asymptotic bounds on the values $F(x(t)) - F^*$:

Corollary 2. *Under the hypotheses of Theorem 2, if F additionally satisfies a global quadratic condition $\mathcal{G}^2_\mu$ then for any $\delta \leqslant \delta_{\alpha,\mu}$, we have:*

$$\forall t \geqslant t_0,\ F(x(t)) - F^* \leqslant \frac{4\left(E_0(\delta) + (\sqrt{2E_0(\delta)} + I_0)J_0 e^{-\delta t_0}\right)}{4 - \alpha\delta + \delta^2} M_0 e^{-\delta(t-t_0)}$$

where:

$$\begin{aligned} M_0 &= F(x_0) - F^* + \tfrac{1}{2}\|v_0\|^2, \quad I_0 = \textstyle\int_{t_0}^{+\infty} \|g(s)\| ds, \\ E_0(\delta) &= M_0 + (\sqrt{2M_0} + I_0)I_0 + \left(\sqrt{M_0} + \tfrac{I_0}{\sqrt{2}} + \tfrac{2\alpha - \delta}{2\sqrt{\mu}}\sqrt{M_0 + (\sqrt{2M_0} + I_0)I_0}\right)^2. \end{aligned}$$

3.2 The non-differentiable case

Assume now that F is a convex but non differentiable function. In that case, the Heavy Ball ODE (12) has no meaning anymore but we can consider the following differential inclusion:

$$0 \in \ddot{x}(t) + \alpha\dot{x}(t) + \partial F(x(t)). \tag{20}$$

To study some optimization algorithms dedicated to non smooth functions, it may be useful to understand the behavior of solutions of (20). This is the case for instance for the LASSO problem (3), for which proximal algorithms such as the Forward Backward (or its accelerated version FISTA) can be used. It was shown in [3] that the behavior of FISTA is linked with the behavior of solutions of

$$0 \in \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \partial F(x(t)). \tag{21}$$

It turns out that FISTA is not the only inertial algorithm that can be used to minimize the LASSO problem or any non smooth optimization problem. It is shown in [12] that a variant of the Heavy Ball algorithm can be used in the case when F is strongly convex or quasi-strongly convex, and the optimal value of the friction parameter α is given.

If F is less regular, but if it still satisfies a growth condition (i.e. if F is in $\mathcal{G}^2_\mu$), it may be interesting to understand how the Heavy Ball algorithm can be used, and how to choose the parameter α. Actually, we will see in the part dedicated to the optimization scheme that the previous analysis applies to this case.

3.2.1 Solutions of the differential inclusion

The differential inclusion problem (20) admits a shock solution [37, 3] and it is known [7, 22] that for any solution x of (20), $F(x(t)) - F^*$ converges to 0 for any $\alpha > 0$. Most of known convergence rates of $F(x(t)) - F^*$ are consequences of a Lyapunov analysis. An energy $\mathcal{E}$ is defined and it is shown to be a non increasing function of t. To prove that $\mathcal{E}$ is non increasing, the simplest way is to compute the derivative $\mathcal{E}'$ of $\mathcal{E}$. To study solutions of (20), we use exactly the same energy defined to study the Heavy Ball ODE:

$$\mathcal{E}_{\lambda,\xi}(t) = F(x(t)) - F^* + \frac{1}{2}\|\lambda(x(t) - x^*) + \dot{x}(t)\|^2 + \frac{\xi^2}{2}\|x(t) - x^*\|^2. \tag{22}$$

This time these Lyapunov energies may not be differentiable. Fortunately, the shock solutions [37, 3] of the differential inclusion (20) are obtained as limit of C^2 functions, where the subdifferential ∂F is replaced by its Moreau Yosida approximation [3].

Let us recall the definition of shock solution for the differential inclusion (20):

Definition 7 (Shock solution [37, 3])**.** *A function $x : [t_0, +\infty) \longrightarrow \mathbb{R}^n$ is an **energy-conserving shock solution** of the differential inclusion* (20) *if:*

1. $x \in \mathcal{C}^{0,1}([t_0, T]; \mathbb{R}^n)$ *for all* $T > t_0$*, i.e. x is a Lipschitz continuous function.*
2. $\dot{x} \in BV([t_0, T]; \mathbb{R}^n)$ *for all* $T > t_0$*.*
3. $x(t) \in dom(F)$ *for all* $t \geq t_0$*.*
4. *For all* $\phi \in \mathcal{C}_c^1([t_0, +\infty), \mathbb{R}^+)$ *and* $v \in \mathcal{C}([t_0, +\infty), dom(F))$*, it holds:*
$$\int_{t_0}^{T} (F(x(t)) - F(v(t)))\phi(t)dt \leq \langle \ddot{x} + \alpha\dot{x}, (v - x)\phi\rangle_{\mathcal{M}\times\mathcal{C}}.$$
5. *x satisfies the following energy equation for a.e.* $t \geq t_0$
$$F(x(t)) - F(x_0) + \frac{1}{2}\left\|\dot{x}(t)\right\|^2 - \frac{1}{2}\left\|v_0\right\|^2 + \int_{t_0}^{t} \alpha \left\|\dot{x}(s)\right\|^2 ds = 0.$$

We then consider the Moreau-Yosida approximations $\{F_\gamma\}_{\gamma>0}$ of F defined by:

$$F_\gamma(x) = \min_y \left(F(y) + \frac{1}{2\gamma}\|x - y\|^2\right) \tag{23}$$

and the following approximating ODE:

$$\begin{aligned} &\ddot{x}_\gamma(t) + \alpha\dot{x}_\gamma(t) + \nabla F_\gamma(x_\gamma(t)) = 0 \\ &x_\gamma(t_0) = x_0 \quad \dot{x}_\gamma(t_0) = v_0. \end{aligned} \tag{24}$$

The differential equation (24) falls into the classical theory of differential equations and admits a unique solution x_γ of class C^2 on $[t_0, +\infty)$ for all $\gamma > 0$. More precisely, using [3, Theorems 3.2 and 3.3], we have the following result:

Theorem 3. *Assume F to be a lower semi continuous convex function. Let $\{F_\gamma\}_{\gamma>0}$ the Moreau-Yosida approximations of F. There exists a subsequence $\{x_\gamma\}_{\gamma>0}$ of solutions of* (24) *that converges to a shock solution of* (20) *according to the following scheme:*

$$\begin{aligned} &\bullet \quad x_\gamma \underset{\gamma\to 0}{\longrightarrow} x \quad \text{uniformly on } [t_0, T] \text{ for all } T > t_0. \\ &\bullet \quad \dot{x}_\gamma \underset{\gamma\to 0}{\longrightarrow} \dot{x} \quad \text{in } L^p([t_0, T]; \mathbb{R}^n) \text{, for all } p \in [1, +\infty) \text{ and } T > t_0. \\ &\bullet \quad F_\gamma(x_\gamma) \underset{\gamma\to 0}{\longrightarrow} F(x) \quad \text{in } L^p([t_0, T]; \mathbb{R}^n)\text{, for all } \forall p \in [1, +\infty) \text{ and } T > t_0. \end{aligned} \tag{25}$$

From Corollary 3.6 of [3], we also have:

Corollary 3. *If $dom(F) = \mathbb{R}^n$, then the differential inclusion* (20) *admits a shock solution x , such that:*

$$x \in W^{2,\infty}((t_0,T);\mathbb{R}^n) \cap \mathcal{C}^1([t_0,+\infty);\mathbb{R}^n), \text{ for all } T > t_0.$$

It turns out that all the results shown for the Heavy ball ODE remain valid for the differential inclusion (20). Indeed, the approximated solutions x_γ of Theorem 3 are solutions of the Heavy ball ODE and they thus satisfy all the previous properties. By passing to the limit $\gamma \to 0^+$, the shock solutions of (20) also satisfies these properties (see e.g. [3] for more details).

We do not restate all the Theorems of the previous section for the differential inclusion case. However, we state a result for a particular case of interest, the LASSO problem recalled here:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|^2 + \beta\|x\|_1 \tag{26}$$

Let $F(x) = \frac{1}{2}\|Ax-b\|^2 + \beta\|x\|_1$. As shown in [19, Lemma 10], the function F satisfies a quadratic growth condition, or equivalently has the Łojasiewicz property with an exponent $\theta = \frac{1}{2}$, on some ℓ^1 ball.

Corollary 4. *If the LASSO problem* (26) *admits a unique minimizer, even if Ker*$(A) \neq \{0\}$*, there is a solution of the differential inclusion* (20) *such that the conclusions of Theorem 1 hold.*

3.2.2 Optimality of the decays

In Proposition 1, we have shown that if F is at least as sharp as $\|\cdot\|^r$ for some $r \in (1,2)$, then

$$\forall \lambda < \frac{2\alpha}{3},\ F(x(t)) - F^* = \mathcal{O}\left(e^{-\lambda t}\right).$$

In Theorem 4 whose proof is detailed in Subsection 6.3, it is proven that a better exponential decay than $\frac{2\alpha}{3}$ is unreachable for the solution of (20) under the quadratic growth condition $\mathcal{G}^2_\mu$, proving that in some sense this decay is reached for the function $F(x) = |x|$ for any $\alpha > 0$.

Theorem 4. *If* $F(x) = |x|$ *then for all* $\alpha > 0$ *then all solutions of* (20) *satisfy:*

$$F(x(t)) - F^* = \mathcal{O}\left(e^{-\frac{2\alpha}{3}t}\right)$$

and this decay is optimal:

$$\limsup_{t\geqslant t_0} e^{\frac{2\alpha}{3}t}(F(x(t)) - F^*) > 0.$$

4 Optimization scheme

In this section we introduce an inertial scheme to minimize a convex differentiable function F whose gradient is L-Lipschitz continuous, and satisfying any Łojasiewicz property. In a second time, this scheme is extended to a sum of two convex functions $F = f + h$ using an inertial proximal gradient algorithm. Both schemes can be seen as discretizations of the Heavy Ball ODE and they are variations of the schemes proposed by B.T. Polyak [38], Y. Nesterov [32] and J.W. Siegel [41].

If F is a convex differentiable function satisfying $\mathcal{G}^2_\mu$ and whose gradient is L-Lipschitz, then the steepest descent algorithm defined by

$$x_{n+1} = x_n - \frac{\beta}{L}\nabla F(x_n), \text{ with } \beta \in (1,2) \tag{27}$$

ensures an exponential decay:

$$F(x_n) - F(x^*) = \mathcal{O}(q^n) \tag{28}$$

Hyp. on F	References	Values of q	Remarks
$\mathcal{S}^{2,1}_{\mu,L}$	Polyak [38]	$\left(\frac{1-\sqrt{\kappa}}{1+\sqrt{k}}\right)^2$	Local convergence and optimal rate on $\mathcal{S}^{2,1}_{\mu,L}$, may diverge on $\mathcal{S}^{1,1}_{\mu,L}$
$\mathcal{S}^{1,1}_{\mu,L}$	Nesterov [32]	$1-\sqrt{\kappa}+\mathcal{O}(\kappa)$	Global convergence
$\mathcal{S}^{1,1}_{\mu,L}$	GFJ [24]	$1-\kappa$	Global convergence
$\mathcal{S}^{2,1}_{\mu,L}$	SFL [44]	$1-2\sqrt{\kappa}+\mathcal{O}(\kappa)$	Global convergence, three points method
$\mathcal{S}^{1,1}_{\mu,L}$	Siegel [41]	$1-\sqrt{\kappa}$	Global convergence, can be extended to non differentiable functions
$q\mathcal{S}^{1,1}_{\mu,L}$	ADR[12] Th. 7	$1-\sqrt{(2-\varepsilon)\kappa+\mathcal{O}(\kappa)}$	Global convergence, can be extended to non differentiable functions
$\mathcal{G}^2_\mu$ and uniqueness of the minimizer	Th. 7 and 9	$1-(2-\sqrt{2})\sqrt{\kappa}$ $+\mathcal{O}(\sqrt{\kappa})$	Global convergence, can be extended to non differentiable functions
$\mathcal{G}^2_\mu$	Th. 6 and 8	$1-\kappa+\epsilon+\mathcal{O}(\kappa)$	Global convergence, can be extended to non differentiable functions

Table 2: Rates of convergence of inertial algorithms depending on the regularity of the function to minimize

where $q=1-\kappa$ and $\kappa:=\frac{\mu}{L}$ the conditioning number.

If $F=f+h$, where f is a convex differentiable function whose gradient is L-Lipschitz, and h is convex, the Forward-Backward algorithm uses the proximal operator of h

$$\text{prox}_{sh}(z)=\operatorname*{argmin}_{x\in\mathbb{R}^n} h(x)+\frac{1}{2s}\|x-z\|^2 \tag{29}$$

and it is defined by the following operator T

$$T(x):=\text{prox}_{\frac{1}{L}h}(x-\frac{1}{L}\nabla f(x)),\quad x_{n+1}=Tx_n. \tag{30}$$

If F satisfies $\mathcal{G}^2_\mu$, the sequence generated by the Forward Backward algorithm ensures an exponential decay with $1-q=\kappa$. If $\kappa<<1$ this exponential decay may be slow. In [19] Bolte et al. proved that for the LASSO problem (26) the function F satisfies the $\mathcal{PL}_\mu$ property and thus the growth condition $\mathcal{G}^2_\mu$. It turns out that the decay is so slow that FISTA [14] which only ensures a $\mathcal{O}(\frac{1}{n^2})$ decay, is better in practice for many large scale problems.

The question arising is then: can we provide a fast exponential decay that is $F(x_n)-F^*=\mathcal{O}(q^n)$ with $1-q\approx C\sqrt{\kappa}$ for such functions ? Namely, for a single differentiable function F belonging to $\mathcal{G}^2_\mu$ or a composite function $F=f+h$ where only f is differentiable ?

If F is μ-strongly convex, it has been proved that such rates can be achieved as summarized in Table 2. But all these results do not apply directly to the set $\mathcal{G}^2_\mu$ since we do not have the strong convexity assumption. Nevertheless observe that from Lemma 4, if $F\in\mathcal{G}^2_\mu$ is differentiable with L-Lipschitz gradient, then F is also $\frac{\mu^2}{L}$-quasi strongly convex. Hence applying the rates from the Table 2, we do not reach any fast exponential decay that is with $1-q=C\sqrt{\kappa}$ where $\kappa=\frac{\mu}{L}$ with μ the parameter of growth. Moreover, only Siegel [41] and ADR [12] proposed an extension to the non differentiable case and both approaches necessitate the strong convexity of the differentiable part which is a real bottleneck. Indeed the strong convexity hypothesis of the differentiable part is a really strong hypothesis, much more than the $\mathcal{G}^2_\mu$ property and the uniqueness of the minimizer.

Consider now the LASSO problem (26). The strong convexity of the differentiable part implies that $Ker(A) = \{0\}$. If $Ker(A) \neq \{0\}$ the LASSO problem may have an infinite set of solutions but it is not the generic case. Indeed for a generic linear operator A, the LASSO problem (26) admits a unique minimizer (see for example Lemma 3 in [43] for the exact definition of genericity and more detailed results). This means that except for a very specific choice of A, the LASSO problem (26) admits a unique minimizer.

To summarize, none of the previous analysis applies to the LASSO problem (26) for a generic matrix A and none of them ensures an exponential decay such that $q = 1 - C\sqrt{\frac{\mu}{L}}$ where μ is the growth parameter for any function $F \in \mathcal{G}^2_\mu$.

In [36], the authors observe that a restarted FISTA algorithm provides a fast exponential decay if $\mathcal{S}^{1,1}_{\mu,L}$. Actually this result is also available for functions satisfying a growth condition, as remarked in [31]:

Theorem 5. *If F is a convex differentiable function whose gradient is L-Lipschitz and if $F \in \mathcal{G}^2_\mu$ then the sequence $(x_n)_{n\in\mathbb{N}}$ generated by FISTA with a restart each $k^* = E\left[\frac{2e}{\sqrt{\kappa}}\right]$ satisfies*

$$F(x_n) - F^* = \mathcal{O}\left(\left(1 - \frac{\sqrt{\kappa}}{e}\right)^n\right). \tag{31}$$

One can notice that FISTA applies to composite functions and that there is no need of uniqueness of the minimizer of F to get this fast rate. Restarted FISTA is thus a relevant algorithm to minimize such functions. One can also notice that even if the inertial parameter does not depend on μ, the starting rule depends on κ. If μ is overestimated, then the proof of convergence does not hold any more.

Another natural question with respect to the Table 2 is the behavior of each algorithm when the growth constant μ is not known. Indeed, the rates given in Table 2 are related to the choice of parameters in the algorithm that depends on μ. It is always possible to underestimate μ but at the cost of a loss of speed (the more μ is underestimated, the slower the algorithm). With non adequate parameters, some of the algorithms mentioned in Table 2 may even diverge.

In this section we propose a new algorithm (41) with a better exponential decay based on the previous analysis of the Heavy Ball ODE (4). Firstly on differentiable functions and in a second time on composite functions, see Theorems 7 and 9. This last algorithm allows to solve the LASSO problem with a fast exponential rate for generic matrix A.

It turns out that the algorithm (41) can be seen as a specific case of a more general algorithm (38) whose parameters do not depend on μ. We can show (see Theorems 6 and 8) that in fact such a scheme converges under very weak assumptions: it suffices that the function F to minimize belongs to $\mathcal{G}^r_\mu$ and has a L-Lipschitz gradient. And a (slow) geometrical decay is reached as soon as F satisfies any Łojasiewicz property.

4.1 The Differentiable case

Several algorithms to minimize strongly convex functions of $\mathcal{S}^{1,1}_{\mu,L}$ or $\mathcal{S}^{2,1}_{\mu,L}$ are inspired by the Heavy Ball ODE in the unperturbed continuous case:

$$\ddot{x}(t) + \alpha\dot{x}(t) + \nabla F(x(t)) = 0. \tag{32}$$

rewritten as the following first order differential system:

$$\begin{cases} \dot{x}(t) = v(t) \\ \dot{v}(t) = -\alpha v(t) - \nabla F(x(t)). \end{cases} \tag{33}$$

The first one was proposed by Polyak in [38] for functions in $\mathcal{S}^{2,1}_{\mu,L}$:

$$\begin{cases} x_{n+\frac{1}{2}} &= x_n + \left(\dfrac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}\right)^2 (x_n - x_{n-1}) \\ x_{n+1} &= x_{n+\frac{1}{2}} - s^2 \nabla F(x_n) \end{cases} \tag{34}$$

with $s = \frac{2}{\sqrt{L}+\sqrt{\mu}}$, which can be seen as a discretization of the Heavy Ball ODE for $\alpha = 2\sqrt{\mu}$. This algorithm is efficient for functions in $\mathcal{S}^{2,1}_{\mu,L}$ but it may diverge for some functions in $\mathcal{S}^{1,1}_{\mu,L}$, see Ghadimi et al. [24] for example. It is worth mentioning that Ghadimi et al. in [24, Theorem 4] proved the linear convergence of such a scheme for functions F in $\mathcal{S}^{1,1}_{\mu,L}$ changing the step and the inertia, but the rate in this case is

$$F(x_n) - F^* = \mathcal{O}((1-\kappa)^n) \tag{35}$$

that is the best rate that can be achieved of the gradient descent on $\mathcal{S}^{1,1}_{\mu,L}$. As we will see further, this decay is much worse than the ones that can be achieved using other schemes for small κ since for small κ, $\kappa << \sqrt{\kappa}$.

In his book [35], Nesterov proposes a scheme which is quite similar:

$$\begin{cases} x_{n+\frac{1}{2}} = & x_n + \left(\dfrac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}\right) (x_n - x_{n-1}) \\ x_{n+1} = & x_{n+\frac{1}{2}} - s^2 \nabla F(x_{n+\frac{1}{2}}) \end{cases} \tag{36}$$

with $s = \frac{1}{\sqrt{L}}$. This scheme can also be seen as a discretization of the Heavy Ball ODE with $\alpha = 2\sqrt{\mu}$, but the descent step s^2 is about four times lower. Nesterov proves the convergence of the scheme (36) for functions in $\mathcal{S}^{1,1}_{\mu,L}$ and he gives a convergence rate:

$$F(x_n) - F^* = \mathcal{O}((1-\sqrt{\kappa})^n). \tag{37}$$

Notice that another variant of this algorithm with the same asymptotic decrease rate was also introduced by Y. Nesterov in [33] with an extension to non differentiable functions (but still strongly convex). An application of this last scheme to image processing can be found in [9].

4.1.1 Convergence of the scheme

In this paper, we consider the following general scheme:

$$\begin{cases} x_{n+\frac{1}{2}} = & x_n + s v_n \\ v_{n+\frac{1}{2}} = & (1+\gamma\lambda s)^{-1}(v_n - s\nabla F(x_{n+\frac{1}{2}})) \\ x_{n+1} = & x_{n+\frac{1}{2}} - s^2 \nabla F(x_{n+\frac{1}{2}}) \\ v_{n+1} = & v_{n+\frac{1}{2}} + (1+\lambda s)^{-1} \lambda s^2 \nabla F(x_{n+\frac{1}{2}}) \end{cases} \tag{38}$$

with $\gamma > 0$ and $s = \frac{1}{\sqrt{L}}$.

Theorem 6. *Let F be a convex differentiable function having a L-Lipschitz gradient. Let $s = \frac{1}{\sqrt{L}}$. If $\gamma\lambda^2 < L$ then the following assertions hold:*

1. *The sequences $(\nabla F(x_n))_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ generated by the scheme* (38)*, converge to* 0 *when $n \to +\infty$.*

2. If F is coercive then the sequences $(F(x_n)-F^*)_{n\in\mathbb{N}}$ and $(d(x_n,X^*))_{n\in\mathbb{N}}$ converge to 0.

3. Assume that F is coercive and has some Lojasiewicz property with an exponent $\theta\in(0,1)$.

(a) If $\theta=\frac{1}{2}$ i.e if F satisfies some local quadratic growth condition $\mathcal{G}_\mu^2$, then the values $(F(x_n)-F^*)_{n\in\mathbb{N}}$ decay exponentially to 0 and the iterates $(x_n)_{n\in\mathbb{N}}$ converge to a minimizer x^* of F. More precisely for any $\varepsilon\in(0,1)$, if $\gamma=(1-\varepsilon)\frac{L^2}{\lambda^2}$, there exists $\lambda_\varepsilon>0$ such that for all $\lambda\in(0,\lambda_\varepsilon)$ then:

$$F(x_n)-F^*=\mathcal{O}\left(\left(1+\frac{\mu}{L}-\varepsilon\frac{\mu}{L+\lambda\sqrt{L}}\right)^{-n}\right). \tag{39}$$

(b) If $\theta\in(\frac{1}{2},1)$ then

$$F(x_n)-F(x^*)=\mathcal{O}\left(\frac{1}{n^{\frac{1}{2\theta-1}}}\right). \tag{40}$$

Moreover, if $\theta<\frac{3}{4}$, then the sequence $(x_n)_{n\in\mathbb{N}}$ converges to a minimizer x^* of F and the length of the trajectory is finite.

The proof of the Theorem is detailed in Subsection 6.4.

4.1.2 Improved rates of convergence

In this paper, we promote the use of the following particular scheme (where the product $\gamma\lambda$ is now fixed to be the friction parameter α):

$$\begin{cases} x_{n+\frac{1}{2}} = & x_n+sv_n\\ v_{n+\frac{1}{2}} = & (1+\alpha s)^{-1}(v_n-s\nabla F(x_{n+\frac{1}{2}}))\\ x_{n+1} = & x_{n+\frac{1}{2}}-s^2\nabla F(x_{n+\frac{1}{2}})\\ v_{n+1} = & v_{n+\frac{1}{2}}+(1+\lambda s)^{-1}\lambda s^2\nabla F(x_{n+\frac{1}{2}}) \end{cases} \tag{41}$$

with $s=\frac{1}{\sqrt{L}}$. Notice that it has the same general form as the one that we introduced in [12]. The difference with respect to the scheme of [12] will be in the choice of the parameters α and λ of the scheme. This comes from the fact that the analysis in the continuous setting has led us to different parameters in the Lyapunov function, and we now use these values in the discrete scheme. Since the parameters are different, the proof of of convergence of the new scheme here is different from the one of [12]. In [12], we used $\alpha=\frac{2\lambda}{3}$ to get a new optimal scheme for strongly convex functions. In this paper, inspired by the analysis in the continuous case detailed in Section 3, we use:

$$\alpha=\left(2-\frac{\sqrt{2}}{2}\right)\sqrt{\mu},\quad \lambda=\sqrt{\mu}\ \text{ and }\ \xi=\left(\frac{\sqrt{2}}{2}-1\right)\mu. \tag{42}$$

Let $\beta=2-\frac{\sqrt{2}}{2}$. We will use the following discrete Lyapunov function:

$$L(x_n,v_n):=F(x_n)-F^*+\frac{1}{2}\|\lambda(x_n-x^*)+(1+\lambda s)v_n\|^2-(\beta-1)\frac{\lambda^2}{2}\|x_n-x^*\|^2. \tag{43}$$

The careful reader can check that the sequence $(v_n)_{n\in\mathbb{N}}$ is a discretization of the variable $v=\dot{x}$ in (33). Indeed, if using explicit Euler schemes to discretize (33), one gets the two first lines of

(41). The third line corresponds to a classical contraction step on x_n that makes the objective function F decrease by a factor $\frac{1}{2L}\|\nabla F(x_{n+\frac{1}{2}})\|^2$ (this is the same argument that is used to prove the convergence of the explicit gradient descent scheme to minimize a differentiable function with L-Lipschitz gradient), and the fourth line keeps the middle term of the Lyapunov function (43) constant.

From (41), we can write v_{n+1} as a function of v_n:

$$\frac{v_{n+1}-v_n}{s} = -\frac{\alpha}{1+\alpha s}v_n - \frac{1-\alpha\lambda s^2}{(1+\alpha s)(1+\lambda s)}\nabla F(x_{n+\frac{1}{2}}). \tag{44}$$

We then see that (41) is indeed a discretization of (33).

The interest of the new scheme (41) is that it allows to provide a better decay rate of $F(x_n)-F^*$ that is asymptotically better than the previous ones under similar assumptions:

Theorem 7. *Assume that F is convex differentiable function with a L-Lipschitz gradient, admitting a unique minimizer and satisfying some quadratic growth condition $\mathcal{G}^2_\mu$. If we choose $\alpha = \left(2-\frac{2}{\sqrt{2}}\right)\sqrt{\mu}$ and $\lambda = \sqrt{\mu}$ in (41), then there exist $t>0$ and $K>0$ such that if $\frac{\mu}{L}\in[0,t]$, then the sequence $(x_n)_{n\in\mathbb{N}}$ provided by (41) with $s=\frac{1}{\sqrt{L}}$ satisfies:*

$$F(x_n)-F^* \leqslant K_0\left(1+\left(2-\sqrt{2}\right)\sqrt{\frac{\mu}{L}}-K\frac{\mu}{L}\right)^{-n}.$$

More precisely, remembering that $\kappa=\sqrt{\frac{\mu}{L}}$, the new scheme (41) allows to get the following rate:

$$F(x_n)-F^* = \mathcal{O}\left((1+(2-\sqrt{2})\sqrt{\kappa}-K\kappa)^{-n}\right). \tag{45}$$

Notice that we have

$$(1+(2-\sqrt{2})\sqrt{\kappa}-K\kappa)^{-1} = 1-(2-\sqrt{2})\sqrt{\kappa}+o(\sqrt{\kappa}). \tag{46}$$

The only other result under similar assumptions (quadratic growth) is the Nesterov restart algorithm of [31] (see Theorem 5), but we can see that we have a better convergence rate with Theorem 7 since $2-\sqrt{2}>\frac{1}{e}$. As far as we know, Theorem 7 gives the best convergence rate in the literature when the function F to minimize satisfies $\mathcal{G}^2_\mu$, has a L-Lipschitz gradient, and has a unique minimizer.

4.2 Discrete scheme in the non differentiable case

In many practical problems especially coming from statistics or image processing the function F to minimize is not differentiable. A classical case is the LASSO problem

$$F(x) = \frac{1}{2}\|Ax-y\|^2+\lambda\|x\|_1 \tag{47}$$

where A is a linear operator. To study the minimisation of such functions, convex but not differentiable, we cannot consider a differential equation involving F. Nevertheless we can consider the following monotone inclusion:

$$0\in\ddot{x}(t)+\alpha\dot{x}(t)+\partial F(x(t)). \tag{48}$$

This inclusion problem admits a shock solution (see [3] and [37]) and $F(x(t))-F^*$ tends to 0.

When $F = f + h$, with f differentiable, ∇f is L-Lipschitz and h is convex proper and lower semi continuous, Siegel in [41] proposed an extension of the discrete scheme built for differentiable functions. In the same spirit we can prove that our scheme can be directly extended to such composite functions using the Forward-Backward algorithm, also called Proximal Gradient Operator.

Let us first recall the definition of the proximal operator for the convex semicontinuous function h:

$$\text{prox}_h(x) = \text{argmin}_{y \in \mathbb{R}^n} \left(h(y) + \frac{1}{2}\|y - x\|^2 \right). \tag{49}$$

Using the optimality condition, we have the equivalence:

$$y = \text{prox}_h(x) \;\Leftrightarrow\; x \in \partial h(y) + y \;\Leftrightarrow y = (Id + \partial h)^{-1}(x). \tag{50}$$

The proximal operator is widely used in convex and non differentiable optimization. It is a generalization of the implicit gradient descent to convex and non differentiable function.

Let $F = f + h$, where f is convex differentiable function having a L-Lipschitz gradient and h is convex proper and lower semi continuous. A classical algorithm to minimize F is the Forward-Backward algorithm defined in the following way:

$$x_{n+1} = T(x_n), \quad \text{where } T(x) := \text{prox}_{s^2 h}\left(x - s^2 \nabla f(x)\right) \tag{51}$$

If $s^2 \leqslant \frac{1}{L}$, it can be shown that $(F(x_n) - F^*)_{n \in \mathbb{N}}$ tends to 0 and $(x_n)_{n \in \mathbb{N}}$ converges (weakly in an infinite dimension Hilbert space) to a minimizer of F.

The operator T shares many properties with the gradient descent. The algorithm FISTA of Beck and Teboulle can be seen as a Nesterov acceleration to this operator T. Following Siegel [41] we modify the previous scheme (41) so that it can be used with $F = f + h$ that is a possibly non smooth convex function: we replace the vector $\nabla F(x_{n+\frac{1}{2}})$ in (41) by

$$g_{n+\frac{1}{2}} = \frac{1}{s^2}\left(x_{n+\frac{1}{2}} - \text{prox}_{s^2 h}(x_{n+\frac{1}{2}} - s^2 \nabla f(x_{n+\frac{1}{2}}))\right) \tag{52}$$

Note that: $g_{n+\frac{1}{2}} \notin \partial F(x_{n+\frac{1}{2}})$ as we might expect. The vector $g_{n+\frac{1}{2}}$ as defined by (52) has been chosen such that the following decrease condition is satisfied:

$$\forall y \in \mathbb{R}^n, \quad F(x_{n+1}) - F(y) \leq \langle g_{n+\frac{1}{2}}, x_{n+\frac{1}{2}} - y \rangle - \frac{s^2}{2}\|g_{n+\frac{1}{2}}\|^2 \tag{53}$$

which is the only property of $g_{n+\frac{1}{2}}$ we need to establish our convergence results. With that choice the new scheme can be written in the nonsmooth case as:

$$\left\{ \begin{array}{rcc} x_{n+\frac{1}{2}} & = & x_n + s v_n \\ v := v_{n+\frac{1}{2}} & = & (1 + \gamma\lambda s)^{-1}(v_n - s g_{n+\frac{1}{2}}) \\ x_{n+1} & = & x_{n+\frac{1}{2}} - s^2 g_{n+\frac{1}{2}} \\ v_{n+1} & = & v + (1 + \lambda s)^{-1} \lambda s^2 g_{n+\frac{1}{2}} \end{array} \right. \tag{54}$$

that is exactly the original scheme (41) replacing $\nabla F(x_{n+\frac{1}{2}})$ by $g_{n+\frac{1}{2}}$. Notice that the same strategy is used in [41] and [12] to extend the convergence results of the numerical scheme in the differential case to the composite case. This new scheme shares the same properties as the previous one:

Theorem 8. *Let $F = f + h$ where f is a convex differentiable function having a L-Lipschitz gradient and h a convex proper lower semicontinuous function. Let $s = \frac{1}{\sqrt{L}}$, $\gamma > 0$ and $\lambda > 0$. If $\gamma\lambda^2 < L$ then the following assertions hold:*

1. *The sequences $(d(0,\partial F(x_n))_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ converge to 0 when $n\to+\infty$.*

2. *If F is coercive then the sequences $(F(x_n)-F^*)_{n\in\mathbb{N}}$ and $(d(x_n,X^*))_{n\in}$ converge to 0.*

3. *Assume that F is coercive and satisfies a Łojasiewicz property with an exponent $\theta\in(0,1)$.*

   (a) *If $\theta=\frac{1}{2}$ i.e if F satisfies some local quadratic growth condition $\mathcal{G}^2_\mu$, then the values $(F(x_n)-F^*)_{n\in\mathbb{N}}$ decay exponentially to 0 and the iterates $(x_n)_{n\in\mathbb{N}}$ converge to a minimizer x^* of F. More precisely for any $\varepsilon\in(0,1)$, if $\gamma=(1-\varepsilon)\frac{L^2}{\lambda^2}$, there exists $\lambda_\varepsilon>0$ such that for all $\lambda\in(0,\lambda_\varepsilon)$ then:*

$$F(x_n)-F^*=\mathcal{O}\left(\left(1+\frac{\mu}{L}-\varepsilon\frac{\mu}{L+\lambda\sqrt{L}}\right)^{-n}\right). \tag{55}$$

   (b) *If $\theta\in(\frac{1}{2},1)$ then*

$$F(x_n)-F(x^*)=\mathcal{O}\left(\frac{1}{n^{\frac{1}{2\theta-1}}}\right). \tag{56}$$

   Moreover, if $\theta<\frac{3}{4}$, then the sequence $(x_n)_{n\in\mathbb{N}}$ converges to a minimizer x^* of F and the length of the trajectory is finite.

As in the differentiable case these rates of convergence can be improved assuming that $\alpha=\gamma\lambda$ and that the objective function has a unique minimizer.

Theorem 9. *Let $F=f+h$ where f is a convex differentiable function having a L-Lipschitz gradient and h a convex proper lower semicontinuous function. Assume that F satisfies the growth condition $\mathcal{G}^2_\mu$ with parameter μ and has a unique minimizer. If $\gamma=\frac{\alpha}{\lambda}$, $\alpha=\left(2-\frac{2}{\sqrt{2}}\right)\sqrt{\mu}$ and $\lambda=\sqrt{\mu}$ in (54), then there exist $t>0$ and $K>0$ such that if $\frac{\mu}{L}\in[0,t]$, then the sequence $(x_n)_{n\in\mathbb{N}}$ provided by (54) with $s=\frac{1}{\sqrt{L}}$ satisfies:*

$$F(x_n)-F^* \quad\leqslant\quad K_0\left(1+\left(2-\sqrt{2}\right)\sqrt{\frac{\mu}{L}}-K_3\frac{\mu}{L}\right)^{-n}.$$

This Theorem applies then to the LASSO problem (26) when it admits a unique solution: it ensures that in this setting we can expect an exponential decay $\mathcal{O}\left(\left(1+\sqrt{(2-\sqrt{2})\kappa}\right)^{-n}\right)$ for κ sufficiently small. As far as we know, this is the best rate that can be found in the literature.

5 Numerical experiments

In the previous sections several convergence properties of the Heavy ball discrete scheme are proved, they are illustrated here.

More precisely the exponential convergence is ensured for most parameters λ and α. The key to get a fast exponential decay is the knowledge of the best parameter μ defining the growth condition of F. In many practical problems, μ is not known. In many theoretical results, see e.g. [38], [32] or [41], it is crucial to have a lower bound of μ to get a value of α ensuring an exponential decay. The actual decay of the algorithm depends on this lower bound estimation. The tuning of the parameter α is a problem of most Heavy Ball methods. An advantage of the proposed scheme is that even if μ is overestimated, an exponential decay is ensured: however, this decay can be quite slow. The optimal choice of α is an open and tough problem. Indeed, the estimation of the optimal α on a given problem is an ill-posed problem that may depend on the starting point of the algorithm. The relevant growth parameter is actually the one available between x and $x^* \in X^*$ for any x on the trajectory of the algorithm. In many practical problems with sparsity constraints, the points $(x_n)_{n\in\mathbb{N}}$ are nearly sparse on the trajectory of the algorithm: thus the relevant growth parameter $\tilde{\mu}$, that is the highest parameter such that for all $n \in \mathbb{N}$

$$\frac{\tilde{\mu}}{2} d(x_n, X^*)^2 \leqslant F(x_n) - F(x^*)$$

may highly depend on x_0.

Notice that choosing α and λ in our algorithm depending on the real global μ parameter may be irrelevant in some situations since the real μ may be much lower than $\tilde{\mu}$.

Liang et al [27] proved that, on a large family of optimisation problems with sparsity constraints, the trajectory of the Forward-Backward algorithm after a finite number of iterations belongs to a subspace of small dimension G depending on the set X^* of solutions. Moreover, the asymptotic exponential decay of the Forward-Backward algorithm depends on the geometrical properties of the function F to minimize on this subspace G, and not on the whole space. Even if the convergence of the trajectory of Heavy Ball schemes on such a subspace is unclear due to the inertia, numerical experiments show that the iterates are really close to sparse vectors. In the inpainting experiment, most iterates have a small number of non zero wavelet coefficients.

For all these reasons, it is today an open problem to choose the parameter of these inertial algorithms. Moreover, a choice of parameter providing a fast asymptotically decay may not be the best for a given precision and numerical experiments will show that depending on the precision, the best set of parameters may not be the same, see Figure 3.

To test these new schemes we consider two classical image processing problems. The first one is the inpainting problem and the second one is the Total Variation (TV) denoising.

5.1 Inpainting

Let x_0 be a numerical image and M a masking operator setting randomly half of the pixel to 0. We want to estimate x from $y_0 = Mx_0$. Using the assumption that natural images are sparse in a suitable wavelet basis we estimate x solving the following optimization problem.

$$\min_{x\in\mathbb{R}^n} F(x) := \frac{1}{2}\left\|Mx - y\right\|^2 + \lambda \left\|Tx\right\|_1 \tag{57}$$

where T is the orthogonal Daubechies (db2) wavelet transform. A numerical example of image inpainting with this model is shown in Figure 2.

Figure 3 compares the decay of $\log(F(x_n) - F(x^*))$ for various choices of parameters. We can observe that these new schemes can be compared to FISTA [14] for various precisions on this specific problem.

5.2 TV denoising

Let x_0 be a numerical image and n a realisation of a white Gaussian noise. We consider $y = x+n$ a noisy image and we want to estimate x from y. Using the assumption that the Total Variation of x, that is the ℓ_1 norm of the discrete gradient of x is small, we can estimate x by solving

$$\min_{x\in\mathbb{R}^n} F(x) := \frac{1}{2}\left\|x - y\right\|^2 + \lambda \left\|\nabla x\right\|_1 \tag{58}$$

Figure 2: Example of image inpainting with model (57)

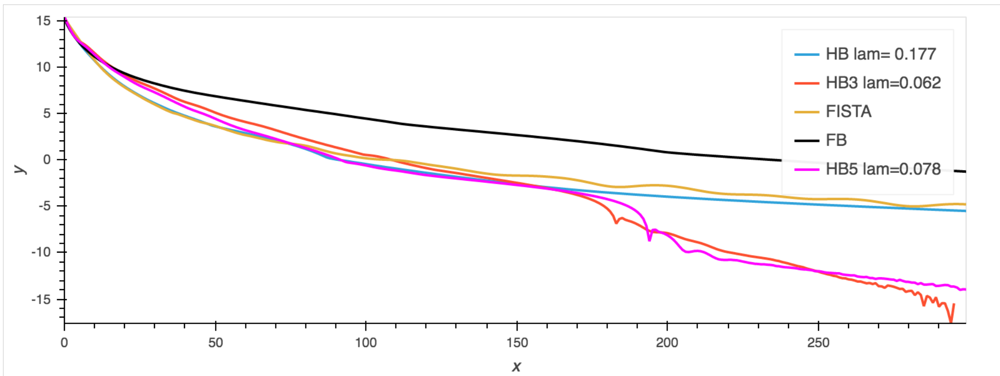

Figure 3: Comparison of several inertial algorithms for the inpainting problem (57)

Actually this problem cannot be solved directly with the Forward-Backward algorithm or its inertial versions since the proximity operator of $x \mapsto \|\nabla x\|_1$ is not explicit. A classical way to face this problem is to consider the following dual problem

$$\min_{p \in \mathbb{R}^n} G(p) := \frac{1}{2}\|\mathrm{div}(p) + y\|^2 + \iota_{\mathcal{B}_\infty(\lambda)}(p). \tag{59}$$

where $\iota_{\mathcal{B}_\infty(\lambda)}$ denotes the indicator function of the ℓ_∞ ball of radius λ. It can be shown that if p^* is a solution of the second problem, $x^* = y + \mathrm{div}(p^*)$ is a solution of the first one.

This second problem (59) can be easily solved using any composite inertial algorithm. An example of image denoising with this approach is shown in Figure 4.

Figure 4: Example of TV Denoising with model (58)

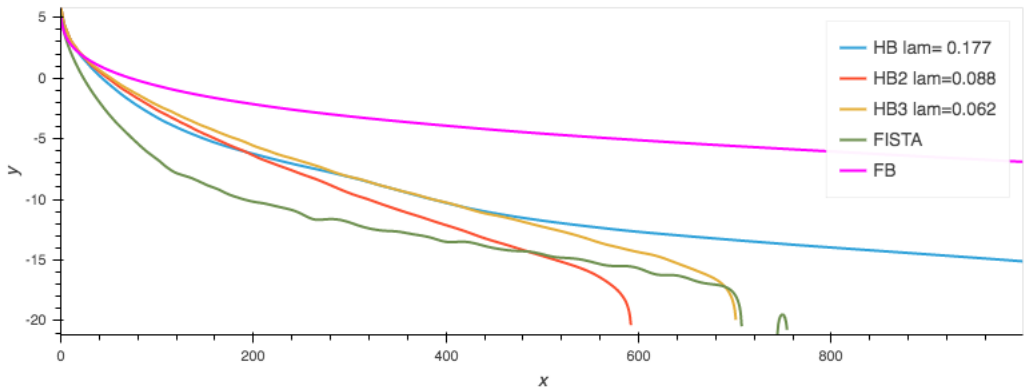

Figure 5: Comparison of several inertial algorithms for the TV denoising (59)

Figure 5 compares the decay of $\log(F(x_n) - F(x^*))$ for various choices of parameters. We can observe that these new schemes can be compared to FISTA [14] for various precisions on this specific problem.

6 Proofs

6.1 Proofs of Theorem 1 and Corollary 1

The proof of Theorem 1 and Corollary 1 relies on the following Lyapunov energy:

$$\mathcal{E}(t) = F(x(t)) - F^* + \frac{1}{2}\|\lambda(x(t) - x^*) + \dot{x}(t)\|^2 + \frac{\xi}{2}\|x(t) - x^*\|^2 \tag{60}$$

and is build in two steps: the first step consists in choosing the parameters λ and ξ such that the energy $\mathcal{E}$ satisfies some differential equation:

$$\mathcal{E}'(t) + \delta\mathcal{E}(t) \leqslant 0$$

for some suitable parameter $\delta > 0$, which implies that: $\mathcal{E}(t) = \mathcal{O}(e^{-\delta t})$. The second step will be to get the control on the values $F(x(t)) - F^*$ from the energy $\mathcal{E}(t)$. In the third step we will prove the Corollary 1: assuming that the quadratic growth condition is global, we provide an uniform bound on the energy $\mathcal{E}$ and thus non-asymptotic decay rates for the values $F(x(t)) - F^*$.

Step 1. Differentiating the energy $\mathcal{E}$ and using the ODE (12), we have:

$$\begin{aligned} \mathcal{E}'(t) &= -\lambda\langle\nabla F(x(t)), x(t) - x^*\rangle + (\lambda - \alpha)\|\dot{x}(t)\|^2 + (\xi + \lambda(\lambda - \alpha))\langle x(t) - x^*, \dot{x}(t)\rangle \\ &\leqslant -\lambda(F(x(t)) - F^*) + (\lambda - \alpha)\|\dot{x}(t)\|^2 + (\xi + \lambda(\lambda - \alpha))\langle x(t) - x^*, \dot{x}(t)\rangle \end{aligned}$$

using the convexity of F. Let us set $\delta < \lambda$:

$$\begin{aligned} \mathcal{E}'(t) + \delta\mathcal{E}(t) &\leqslant (\delta - \lambda)(F(x(t) - F^*) + \frac{\delta}{2}(\xi + \lambda^2)\|x(t) - x^*\|^2 + (\lambda - \alpha + \frac{\delta}{2})\|\dot{x}(t)\|^2 \\ &\quad + (\xi + \lambda(\lambda - \alpha) + \delta\lambda)\langle x(t) - x^*, \dot{x}(t)\rangle \end{aligned}$$

We now choose $\xi := -\lambda(\lambda - \alpha) - \delta\lambda = \lambda(\alpha - \lambda - \delta)$ so that the inner product term disappears.

$$\mathcal{E}'(t) + \delta\mathcal{E}(t) \leqslant (\delta - \lambda)(F(x(t) - F^*) + \frac{\delta}{2}\lambda(\alpha - \delta)\|x(t) - x^*\|^2 + (\lambda - \alpha + \frac{\delta}{2})\|\dot{x}(t)\|^2 \tag{61}$$

Using the quadratic growth assumption $\mathcal{G}^2_\mu$ on F and assuming that $\delta < \lambda$, there exists $t_1 \geqslant t_0$ such that:

$$\forall t \geqslant t_1,\ \mathcal{E}'(t) + \delta\mathcal{E}(t) \leqslant (\frac{\delta}{2}\lambda(\alpha - \delta) + (\delta - \lambda)\frac{\mu}{2})\|x(t) - x^*\|^2 + (\lambda - \alpha + \frac{\delta}{2})\|\dot{x}(t)\|^2 . \tag{62}$$

Note that we can choose $t_1 = t_0$ if the quadratic growth condition is assumed to be global. Let us fix $\mu > 0$. Since we want $\mathcal{E}'(t) + \delta\mathcal{E}(t) \leqslant 0$, we need to choose λ and δ such that:

$$\frac{\delta}{2}\lambda(\alpha - \delta) + (\delta - \lambda)\frac{\mu}{2} \leqslant 0, \quad \lambda - \alpha + \frac{\delta}{2} \leqslant 0, \quad \delta < \lambda. \tag{63}$$

Let us choose $\lambda := \alpha - \frac{\delta}{2}$. We then have two constraints left:

$$\alpha^2 - \left(\frac{3\delta}{2} + \frac{\mu}{\delta}\right)\alpha + \frac{3\mu}{2} + \frac{\delta^2}{2} \leqslant 0 \tag{64}$$

$$\delta < \frac{2\alpha}{3} \tag{65}$$

The constraint (64) can be seen as a polynomial of degree 3 in δ or as a polynomial of degree 2 en α. Let us first consider (64) as a polynomial inequality of degree 3 in δ. Let $\alpha > 0$ and:

$$P_\alpha(\delta) = \delta^3 - 3\alpha\delta^2 + (3\mu + 2\alpha^2)\delta - 2\mu\alpha.$$

Noticing that: $P''_\alpha(\delta) = 6(\delta - \alpha) < 0$ for all $\delta < \frac{2\alpha}{3}$, we easily check that P'_α is non-increasing on the interval $[0, \frac{2\alpha}{3}[$.

First case: $\alpha < 3\sqrt{\frac{\mu}{2}}$**.** In that case, a straightforward computation shows that: $P'_\alpha(\frac{2\alpha}{3}) \geqslant 0$. We thus deduce that P_α is non-decreasing on the interval $[0, \frac{2\alpha}{3}[$.

Second case: $\alpha > 3\sqrt{\frac{\mu}{2}}$. In that case, a straightforward computation shows that: $P'_\alpha(\frac{2\alpha}{3}) < 0$. We thus deduce that P_α is first non-decreasing and then non-increasing on $[0, \frac{2\alpha}{3}[$.

Observe now that:

$$P_\alpha(0) = -2\alpha\mu < 0, \quad P_\alpha(\frac{2\alpha}{3}) = \frac{\alpha^3}{27} > 0.$$

This implies that in any case there exists a unique $\delta_{\alpha,\mu} \in [0, \frac{2\alpha}{3}[$ such that $P_\alpha(\delta_{\alpha,\mu}) = 0$ and:

$$\forall \delta \leqslant \delta_{\alpha,\mu},\ P_\alpha(\delta) \leqslant 0. \tag{66}$$

Let us now consider the polynomial inequality (64) as a polynomial of degree 2 in α to obtain more information on the admissible values for $\delta_{\alpha,\mu}$. If δ is fixed then the polynomial inequality (64) admits a solution in α if and only if its discriminant is positive i.e. if:

$$\Delta = \left(\frac{3\delta}{2} + \frac{\mu}{\delta}\right)^2 - 4\left(\frac{3\mu}{2} + \frac{\delta^2}{2}\right) \geqslant 0,$$

or equivalently if: $\delta^4 - 12\delta^2\mu + 4\mu^2 \geqslant 0$, i.e. if and only if:

$$\delta \leqslant (2 - \sqrt{2})\sqrt{\mu} \quad \text{or} \quad \delta \geqslant (2 + \sqrt{2})\sqrt{\mu}. \tag{67}$$

Coming back to the definition (66) of $\delta_{\alpha,\mu}$, remember that the constraint $P_\alpha(\delta) \leqslant 0$ has to be satisfied by any $\delta \in [0, \delta_{\alpha,\mu}]$ which excludes $(2+\sqrt{2})\sqrt{\mu}$ (or any larger value) from the admissible values of $\delta_{\alpha,\mu}$. We can thus conclude that for a given $\alpha > 0$:

$$\delta_{\alpha,\mu} \leqslant (2 - \sqrt{2})\sqrt{\mu}.$$

Finally, observe that for the critical value $(2 - \sqrt{2})\sqrt{\mu}$, the discriminant Δ is equal to 0 so that the polynomial $P_\alpha((2 - \sqrt{2})\sqrt{\mu})$ admits a double root in α given by:

$$\alpha = \left(2 - \frac{\sqrt{2}}{2}\right)\sqrt{\mu}. \tag{68}$$

For this particular value of α, we have:

$$\delta_{\alpha,\mu} = (2 - \sqrt{2})\sqrt{\mu}, \quad \lambda = \sqrt{\mu}, \quad \text{and } \xi = \left(\frac{\sqrt{2}}{2} - 1\right)\mu. \tag{69}$$

In any case, we have proved that with our choice of parameters, for all $\delta \leqslant \delta_{\alpha,\mu}$, we have: $\forall t \geqslant t_1,\ \mathcal{E}'(t) + \delta\mathcal{E}(t) \leqslant 0$, hence for $\delta \leqslant \delta_{\alpha,\mu}$:

$$\forall t \geqslant t_1,\ \mathcal{E}(t) \leqslant \mathcal{E}(t_1)e^{-\delta(t-t_1)}. \tag{70}$$

Step 2. The point now is to get the control on the values $F(x(t)) - F^*$ from the inequality (70). To do this, remember the definition of the energy with $\xi = -\frac{\delta\lambda}{2} = -\frac{\delta}{4}(2\alpha - \delta) < 0$:

$$\begin{aligned}\mathcal{E}(t) &= F(x(t)) - F^* + \frac{1}{2}\|\lambda(x(t) - x^*) + \dot{x}(t)\|^2 - \frac{\delta}{4}(2\alpha - \delta)\|x(t) - x^*\|^2 \\ &\geqslant F(x(t)) - F^* - \frac{\delta}{4}(2\alpha - \delta)\|x(t) - x^*\|^2\end{aligned}$$

Using the quadratic growth condition satisfied by F, we then get the control:

$$\forall t \geqslant t_1,\ \mathcal{E}(t) \geqslant \frac{1}{4}\left(\delta^2 - 2\alpha\delta + 4\mu\right)(F(x(t)) - F^*). \tag{71}$$

Moreover we can easily check that:

$$\forall \delta \leqslant \delta_{\alpha,\mu},\ \delta^2 - 2\alpha\delta + 4\mu > 0. \tag{72}$$

Indeed, observe that if $\alpha < 2\sqrt{\mu}$ then the discriminant of the polynomial $Q_\alpha(\delta) = \delta^2 - 2\alpha\delta + 4\mu$ is non positive so that the inequality (72) is satisfied for any $\delta \leqslant \delta_{\alpha,\mu}$. If $\alpha \geqslant 2\sqrt{\mu}$, the polynomial Q_α has two real roots given by $\delta_\pm = \alpha \pm \sqrt{\alpha^2 - 4\mu}$. Noticing that $P_\alpha(\alpha - \sqrt{\alpha^2 - 4\mu}) = \mu(\alpha + \sqrt{\alpha^2 - 4\mu}) > 0$, the variations of the polynomial P_α imply that necessarily: $\delta_{\alpha,\mu} < \alpha - \sqrt{\alpha^2 - 4\mu}$.

Combining the control (70) with the inequality (71), we get:

$$F(x(t)) - F^* \leqslant \frac{4}{\delta^2 - 2\alpha\delta + 4\mu}\mathcal{E}(t_1)e^{-\delta(t-t_1)}$$

which concludes the proof of Theorem 1.

Step 3. Assume now that the quadratic growth condition is global. In that case, we have: $t_1 = t_0$ and:

$$\forall t \geqslant t_0,\ F(x(t)) - F^* \leqslant \frac{4}{\delta^2 - 2\alpha\delta + 4\mu}\mathcal{E}(t_0)e^{-\delta(t-t_0)}.$$

Let $M_0 = F(x_0) - F^* + \frac{1}{2}\|v_0\|^2$. Applying [12, Lemma 3] with our choice of parameters (for which $\xi < 0$), we have an uniform bound on the energy $\mathcal{E}$ for the class of differentiable convex functions satisfying a quadratic growth condition for some constant $\mu > 0$, given by:

$$\forall t \geqslant t_0,\ \mathcal{E}(t) \leqslant \frac{\mu + \left(\lambda + \sqrt{\mu}\right)^2}{\mu}M_0 = \frac{4\mu + \left(2\alpha - \delta + 2\sqrt{\mu}\right)^2}{4\mu}M_0$$

which concludes the proof of Corollary 1.

6.2 Proofs of Theorem 2 and Corollary 2

The proof of Theorem 2 is quite standard, see [8, 13, 10, 40, 12] and the references therein.

Let $T > 0$. We consider the same Lyapunov energy as in the unperturbed case:

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

To deal with the perturbation term, the idea is to add an integral term to the energy $\mathcal{E}$ by considering the energy:

$$\mathcal{G}(t) = \mathcal{E}(t) + \int_t^T \langle\lambda(x(s) - x^*) + \dot{x}(s)), g(s)\rangle ds \tag{73}$$

in order to cancel all the terms depending on the perturbation g after derivation. Indeed differentiating the energy $\mathcal{G}$ and using the convexity of F as in Theorem 1, we get:

$$\begin{aligned}\mathcal{G}'(t) &= \mathcal{E}'(t) - \langle \lambda(x(t)-x^*) + \dot{x}(t), g(t)\rangle \\ &= -\lambda\langle \nabla F(x(t)), x(t)-x^*\rangle + (\lambda-\alpha)\left\|\dot{x}(t)\right\|^2 + (\xi + \lambda(\lambda-\alpha))\langle x(t)-x^*, \dot{x}(t)\rangle \\ &\leqslant -\lambda(F(x(t)) - F^*) + (\lambda-\alpha)\left\|\dot{x}(t)\right\|^2 + (\xi + \lambda(\lambda-\alpha))\langle x(t)-x^*, \dot{x}(t)\rangle\end{aligned}$$

Choosing now the same parameters λ, ξ and δ as in the unperturbed case, we obtain:

$$\forall t \geqslant t_0,\ \mathcal{G}'(t) + \delta\mathcal{E}(t) \leqslant 0. \tag{74}$$

Moreover observe that even if the energy $\mathcal{E}$ is not a sum of non-negative terms, according to (71) and (72), we have the following control:

$$\mathcal{E}(t) = F(x(t)) - F^* + \frac{1}{2}\left\|\lambda(x(t)-x^*) + \dot{x}(t)\right\|^2 + \frac{\xi}{2}\|x(t)-x^*\|^2 \tag{75}$$

$$\geqslant \frac{1}{2}\left\|\lambda(x(t)-x^*) + \dot{x}(t)\right\|^2 \tag{76}$$

so that: $\forall t \geqslant t_0,\ \mathcal{G}'(t) \leqslant -\lambda\mathcal{E}(t) \leqslant 0$. The energy $\mathcal{G}$ is so non-increasing, and: $\forall t \geqslant t_0,\ \mathcal{G}(t) \leq \mathcal{G}(t_0)$, i.e.:

$$\begin{aligned}\forall t \geqslant t_0,\ \mathcal{E}(t) &\leqslant \mathcal{E}(t_0) + \int_{t_0}^{t} \langle g(s), \dot{x}(s) + \lambda(x(s)-x^*)\rangle ds \\ &\leqslant \mathcal{E}(t_0) + \int_{t_0}^{t} \|g(s)\|\|(1+\beta)\dot{x}(s) + \lambda(x(s)-x^*)\| ds\end{aligned}$$

Combining the last inequality with (76), we get:

$$\forall t \geqslant t_0,\ \frac{1}{2}\|\lambda(x(t)-x^*) + \dot{x}(t)\|^2 \leqslant \mathcal{E}(t_0) + \int_{t_0}^{t} \|g(s)\|\|\lambda(x(s)-x^*) + \dot{x}(s)\| ds$$

Applying the Grönwall-Bellman Lemma [20, Lemma A.5], we obtain:

$$\forall t \geqslant t_0,\ \|\lambda(x(t)-x^*) + \dot{x}(t)\| \leqslant \sqrt{2\mathcal{E}(t_0)} + \int_{t_0}^{t} \|g(s)\| ds.$$

Assuming $\int_{t_0}^{+\infty} \|g(s)\| ds < +\infty$, we can conclude that:

$$\sup_{t \geqslant t_0} \|\lambda(x(t)-x^*) + \dot{x}(t)\| \leqslant \sqrt{2\mathcal{E}(t_0)} + \int_{t_0}^{+\infty} \|g(s)\| ds < +\infty.$$

We set: $A = \sqrt{2\mathcal{E}(t_0)} + I_0$ where $I_0 = \int_{t_0}^{+\infty} \|g(s)\| ds$. The differential inequality $\forall t \geqslant t_0,\ \mathcal{G}'(t) \leqslant -\delta\mathcal{E}(t) \leqslant 0$ can be rewritten as:

$$\begin{aligned}\forall t \geqslant t_0,\ \mathcal{E}'(t) &\leqslant -\delta\mathcal{E}(t) + \langle \lambda(x(t)-x^*) + (1+\beta)\dot{x}(t), g(t)\rangle. \\ &\leqslant -\delta\mathcal{E}(t) + A\|g(t)\|.\end{aligned}$$

Integrating between t_0 and t, we finally obtain:

$$\begin{aligned}\forall t \geqslant t_0,\ e^{\delta t}\mathcal{E}(t) &\leqslant e^{\delta t_0}\mathcal{E}(t_0) + A\int_{t_0}^{t} e^{\delta s}\|g(s)\| ds, \\ &\leqslant e^{\delta t_0}\mathcal{E}(t_0) + A\int_{t_0}^{+\infty} e^{\delta s}\|g(s)\| ds < +\infty.\end{aligned}$$

Hence:

$$\forall t \geqslant t_0,\ \mathcal{E}(t) \leqslant \left[\mathcal{E}(t_0) + (\sqrt{2\mathcal{E}(t_0)} + I_0)J_0 e^{-\delta t_0}\right] e^{-\delta(t-t_0)}.$$

Let us define:

$$\begin{aligned} M_0 &= F(x_0) - F^* + \tfrac{1}{2}\|v_0\|^2, \quad I_0 = \textstyle\int_{t_0}^{+\infty} \|g(s)\|ds, \\ E_0(\lambda) &= M_0 + (\sqrt{2M_0} + I_0)I_0 + \left(\sqrt{M_0} + \tfrac{I_0}{\sqrt{2}} + \tfrac{\lambda}{\sqrt{\mu}}\sqrt{M_0 + (\sqrt{2M_0} + I_0)I_0}\right)^2. \end{aligned}$$

Applying [12, Lemma 4] with our choice of parameters (for which $\xi < 0$), we have again an uniform bound on the energy $\mathcal{E}$ for the class of differentiable convex functions satisfying a quadratic growth condition for some constant $\mu > 0$, given by:

$$\forall t \geqslant t_0,\ \mathcal{E}(t) \leqslant M_0 + (\sqrt{2M_0} + I_0)I_0 + \left(\sqrt{M_0} + \frac{I_0}{\sqrt{2}} + \frac{2\alpha - \delta}{2\sqrt{\mu}}\sqrt{M_0 + (\sqrt{2M_0} + I_0)I_0}\right)^2$$

which concludes the proof.

6.3 Proof of Theorem 4

Using the definition of the Lyapunov function $\mathcal{E}$ and the convexity of F, a standard computation (as in the beginning of step 1 of the proof of Theorem 1) leads to

$$\mathcal{E}'(t) + \lambda\mathcal{E}(t) \leqslant \frac{\lambda}{2}(\xi + \lambda^2)\|x(t) - x^*\|^2 + \left(\frac{3}{2}\lambda - \alpha\right)\|\dot{x}(t)\|^2 + (\xi + \lambda(2\lambda - \alpha))\langle \dot{x}(t), x(t) - x^*\rangle.$$

Observe that if $F(x) = |x|$, this inequality is actually an equality:

$$\mathcal{E}'(t) + \lambda\mathcal{E}(t) = \frac{\lambda}{2}(\xi + \lambda^2)\|x(t) - x^*\|^2 + \left(\frac{3}{2}\lambda - \alpha\right)\|\dot{x}(t)\|^2 + (\xi + \lambda(2\lambda - \alpha))\langle \dot{x}(t), x(t) - x^*\rangle \quad (77)$$

since for this function, we actually have for all $u \in \partial F(x)$, $\langle u, x - x^*\rangle = F(x) - F^*$. If we choose $\lambda = \frac{2\alpha}{3}$ and $\xi = -\frac{\lambda}{2}$ we get

$$\mathcal{E}'(t) + \lambda\mathcal{E}(t) = \frac{\lambda^3}{4}|x(t)|^2. \quad (78)$$

Hence for any $\delta > 0$:

$$\mathcal{E}'(t) + \delta\mathcal{E}(t) = \frac{\lambda^3}{4}|x(t)|^2 + (\delta - \lambda)\mathcal{E}(t). \quad (79)$$

Since $x(t) \to 0$ when $t \to +\infty$, it follows that for $\delta = \frac{2\lambda}{3} < \lambda$, it exists $t_1 > t_0$ such that $\mathcal{E}(x(t_1)) > 0$ and such that for any $t > t_1$, $\mathcal{E}'(t) + \frac{2\lambda}{3}\mathcal{E}(t) \leqslant 0$. We deduce that for any $t > t_1$,

$$\mathcal{E}(t) \leqslant \mathcal{E}(t_1)e^{-\frac{2\lambda(t-t_1)}{3}}$$

with $\mathcal{E}(t_1) \geqslant 0$. It follows that it exists $A_1 \geqslant 0$ such that

$$|x(t)| - \frac{\lambda^2}{4}|x(t)|^2 \leqslant \mathcal{E}(t) \leqslant A_1 e^{-\frac{2\lambda t}{3}}$$

and thus it exists t_3 and A_2 such that for all $t \geqslant t_3$

$$|x(t)| \leqslant A_2 e^{-\frac{2\lambda t}{3}}. \quad (80)$$

Moreover from (78) and (80) we deduce that it exists t_3 and A_3 such that for any $t \geqslant t_3$,

$$\mathcal{E}'(t) + \delta\mathcal{E}(t) \leqslant A_3 e^{-\frac{4\lambda t}{3}}$$

and thus if we define: $\mathcal{H}_1(t) := e^{\frac{2\alpha t}{3}}\mathcal{E}(t)$ we get for all $t \geqslant t_3$, $\mathcal{H}_1'(t) \leqslant A_3 e^{-\frac{\lambda t}{3}}$. It follows that it exists B such that for all $t > t_0$, $\mathcal{H}(t) \leqslant B$ and thus

$$|x(t)| - \frac{\lambda^2}{4}|x(t)|^2 \leqslant \mathcal{E}(t) \leqslant B e^{-\frac{2\alpha t}{3}}$$

which implies that it exists $B_1 > 0$ such that: $|x(t)| \leqslant B_1 e^{-\frac{2\alpha t}{3}}$.

To prove the second point of the Theorem, we observe that (78) implies that $\mathcal{H}_1$ is a non decreasing function. If $\mathcal{H}(t_0) > 0$, it follows that for all $t \geqslant t_0$, $\mathcal{H}(t) \geqslant \mathcal{H}(t_0)$. Hence, $\forall t \geqslant t_0$

$$e^{\frac{2\alpha t}{3}}\left(|x(t)| + \frac{1}{2}|\lambda x(t) + \dot{x}(t)|\right) \geqslant \mathcal{H}(t_0) > 0.$$

Defining $y(t) = e^{\frac{2\alpha t}{3}}x(t)$ we have for all $t \geqslant t_0$

$$|y(t)| + \frac{1}{2}|\dot{y}(t)|^2 e^{-\frac{2\alpha t}{3}} \geqslant \mathcal{H}(t_0) > 0.$$

Hence, if it exists t_4 such that for all $t \geqslant t_4$, $|y(t)| \leqslant \frac{1}{2}\mathcal{H}(t_0)$, then $\forall t \geqslant t_4$, $|\dot{y}(t)|^2 \geqslant \mathcal{H}(t_0)e^{\frac{2\alpha t}{3}}$. Since $y \in C^1$ (thanks to Corollary 3), y is a continuous function of time. It follows that the sign of $\dot{y}(t)$ is constant on $[t_4, +\infty[$. If $\dot{y}(t_4) > 0$ then:

$$\forall t \geqslant t_4, y(t) \geqslant y_{t_4} + \int_{t_4}^{t} \sqrt{\mathcal{H}(t_0)} e^{\frac{\alpha u}{3}} du$$

which is impossible. If $\dot{y}(t_4) < 0$ then for all $t \geqslant t_4$, $y(t) \leqslant y_{t_4} - \int_{t_4}^{t} \sqrt{\mathcal{H}(t_0)} e^{\frac{\alpha u}{3}} du$, which is also impossible. It follows that $y(t)$ cannot tend to 0 when $t \to \infty$ which concludes the proof.

6.4 Proofs of Theorems 6 and 8

In this section we propose a detailed proof of the convergence of the iterates generated by the discrete scheme (38) in the differentiable case, and by the scheme (54) in the nonsmooth case. These two proofs being very similar we will present each results of the two theorems, detailing all the calculations in the differentiable case and then explaining how to adapt them in the non-differentiable case.

Let us remind here the general scheme in the differentiable case: let $\gamma > 0$.

$$\left\{\begin{array}{rcl} \lambda x_{n+\frac{1}{2}} & = & \lambda x_n + t v_n \\ v := v_{n+\frac{1}{2}} & = & (1+\gamma t)^{-1}(v_n - tg) \\ \lambda x_{n+1} & = & \lambda x_{n+\frac{1}{2}} - t^2 g \\ v_{n+1} & = & v + (1+t)^{-1} t^2 g. \end{array}\right. \tag{81}$$

using the reduced notations:

$$t = \lambda s, \quad v = v_{n+\frac{1}{2}} \text{ and } g = \frac{1}{\lambda}\nabla F(x_{n+\frac{1}{2}}). \tag{82}$$

The scheme in the nonsmooth case is obtained from (81) by replacing g by the vector $\tilde{g}$ defined by:

$$\tilde{g} = \frac{\lambda}{t^2}\left(x_{n+\frac{1}{2}} - \text{prox}_{\frac{t^2}{\lambda^2}h}(x_{n+\frac{1}{2}} - \frac{t^2}{\lambda^2}\nabla f(x_{n+\frac{1}{2}}))\right). \tag{83}$$

Keep in mind that to adapt the proof from the differentiable to the non-differentiable case, a key element will be to be able to compute a subgradient of F at some iterate. More precisely, exploiting the definition of $\tilde{g}$, we easily prove that:

$$\lambda\tilde{g} + \nabla f(x_{n+1}) - \nabla f(x_{n+\frac{1}{2}}) \in \partial F(x_{n+1}) \tag{84}$$

Moreover using the L-Lipschitz continuity and the co-coercivity of ∇f, observe that:

$$\begin{aligned}
\lambda\langle \tilde{g}, \nabla f(x_{n+1}) - \nabla f(x_{n+\frac{1}{2}})\rangle &\leqslant -\frac{\lambda^2}{t^2}\langle \nabla f(x_{n+1}) - \nabla f(x_{n+\frac{1}{2}}), x_{n+1} - x_{n+\frac{1}{2}}\rangle \\
&\leqslant -\frac{\lambda^2}{Lt^2}\|\nabla f(x_{n+1}) - \nabla f(x_{n+\frac{1}{2}})\|^2 \\
&\leqslant -\|\nabla f(x_{n+1}) - \nabla f(x_{n+\frac{1}{2}})\|^2
\end{aligned}$$

since: $t = \lambda s$ and $s = \frac{1}{\sqrt{L}}$. It follows that:

$$\begin{aligned}
\|\lambda\tilde{g} + \nabla f(x_{n+1}) - \nabla f(x_{n+\frac{1}{2}})\|^2 &= \lambda^2\|\tilde{g}\|^2 + \|\nabla f(x_{n+1}) - \nabla f(x_{n+\frac{1}{2}})\|^2 \\
&\quad +2\lambda\langle \tilde{g}, \nabla f(x_{n+1}) - \nabla f(x_{n+\frac{1}{2}})\rangle \\
&\leqslant \lambda^2\|\tilde{g}\|^2 - \|\nabla f(x_{n+1}) - \nabla f(x_{n+\frac{1}{2}})\|^2.
\end{aligned}$$

Hence:

$$\|\lambda\tilde{g} + \nabla f(x_{n+1}) - \nabla f(x_{n+\frac{1}{2}})\| \leqslant \lambda\|\tilde{g}\|. \tag{85}$$

6.4.1 Proof of Point 1. Global convergence of the discrete scheme

Following the line of Alvarez in [1], we introduce the following Lyapunov energy:

$$L_n = F(x_n) - F(x^*) + \frac{\rho}{2}\|v_n\|^2 \tag{86}$$

where $\rho > 0$ will be defined later.

The differentiable case. Remembering that $v_{n+1} = v + \frac{t^2}{1+t}g$, note that:

$$\begin{aligned}
v_n &= (1+\gamma t)v + tg = (1+\gamma t)v_{n+1} + t(1 - \frac{(1+\gamma t)t}{1+t})g \\
&= (1+\gamma t)v_{n+1} + tAg
\end{aligned}$$

where:

$$A = 1 - \frac{(1+\gamma t)t}{1+t} \tag{87}$$

Observe that the quantity A always satisfies $A < 1$ for any $t > 0$ and that $A > 0$ if and only if:

$$\gamma t^2 < 1.$$

Let us now prove that for a suitable choice of ρ, the sequence L_n is non increasing. A key element to prove this result is the following decrease inequality:

$$F(x_{n+1}) - F(x_n) - t\langle g, v_n\rangle \leqslant -\frac{t^2}{2}\|g\|^2 \tag{88}$$

whose proof is detailed at (134). This inequality is a combination of the convexity and the Lipschitz continuity of F. We thus have:

$$\begin{aligned} L_{n+1} - L_n &= F(x_{n+1}) - F(x_n) + \frac{\rho}{2}\left(\|v_{n+1}\|^2 - \|(1+\gamma t)v_{n+1} + tAg\|^2\right) \\ &\leqslant -\frac{t^2}{2}\|g\|^2 + t\langle g, (1+\gamma t)v_{n+1} + tAg\rangle + \frac{\rho}{2}\left(\|v_{n+1}\|^2 - \|(1+\gamma t)v_{n+1} + tAg\|^2\right) \\ &\leqslant (-\frac{1}{2} + A)t^2\|g\|^2 + t(1+\gamma t)\langle g, v_{n+1}\rangle + \frac{\rho}{2}\left(\|v_{n+1}\|^2 - \|(1+\gamma t)v_{n+1} + tAg\|^2\right) \end{aligned}$$

Let us focus on the last part of the right member of the above inequality:

$$\|v_{n+1}\|^2 - \|(1+\gamma t)v_{n+1} + tAg\|^2 = (-2\gamma t - \gamma^2 t^2)\|v_{n+1}\|^2 - 2(1+\gamma t)At\langle v_{n+1}, g\rangle - t^2A^2\|g\|^2.$$

It follows that

$$L_{n+1} - L_n \leqslant \left(-\frac{1}{2} + A - \frac{\rho A^2}{2}\right)t^2\|g\|^2 + (t(1+\gamma t) - \rho(1+\gamma t)At)\langle g, v_{n+1}\rangle - \frac{\rho}{2}(2\gamma t + \gamma^2 t^2)\|v_{n+1}\|^2.$$

Choosing $\rho = A^{-1}$ where A is given by (87), the scalar product term vanishes and we get

$$L_{n+1} - L_n \leqslant -\frac{t^3}{2}\left(\frac{1+\gamma t}{1+t}\right)\|g\|^2 - \frac{\rho}{2}(2\gamma t + \gamma^2 t^2)\|v_{n+1}\|^2. \tag{89}$$

If follows that $(L_n)_{n\in\mathbb{N}}$ is non increasing, and thus it converges in $\mathbb{R}^+$. Let us now rewrite inequality (89) as:

$$L_{n+1} - L_n \leqslant -a\|g\|^2 - b\|v_{n+1}\|^2 \tag{90}$$

where a and b are two positive constants. Hence:

$$\forall N \geqslant 1,\ a\sum_{n=0}^{N}\|g\|^2 + b\sum_{n=0}^{N}\|v_{n+1}\|^2 \leqslant L_0 - L_{N+1} \leqslant L_0.$$

Since L_n is a positive sequence, we deduce that $\sum\|g\|^2$ and $\sum\|v_{n+1}\|^2$ are converging sums. Hence both sequences $(\|g\|^2)_{n\in\mathbb{N}}$ and $(\|v_{n+1}\|^2)_{n\in\mathbb{N}}$ tend to 0. Using the L-Lipschitz continuity of the gradient, we also deduce the convergence of the sequence $(\nabla F(x_n))_{n\in\mathbb{N}}$ to 0.

The non differentiable case Consider now the non differentiable case. The calculations of this first step are unchanged replacing the vector $g = \frac{1}{\lambda}\nabla F(x_{n+\frac{1}{2}})$ by $\tilde{g}$ since we only need the decrease condition:

$$F(x_{n+1}) - F(x_n) \leq \lambda\langle \tilde{g}, x_{n+\frac{1}{2}} - x_n\rangle - \frac{t^2}{2}\|\tilde{g}\|^2. \tag{91}$$

which is actually satisfied by $\tilde{g}$ applying Lemma 8 to $x = x_{n+1}$ and $y = x_n$. We thus have:

$$L_{n+1} - L_n \leqslant -\frac{t^3}{2}\left(\frac{1+\gamma t}{1+t}\right)\|\tilde{g}\|^2 - \frac{\rho}{2}(2\gamma t + \gamma^2 t^2)\|v_{n+1}\|^2 \tag{92}$$

and it follows that the sequence $(L_n)_{n\in\mathbb{N}}$ is still non increasing, and converges in $\mathbb{R}^+$. Hence $\sum\|\tilde{g}\|^2$ and $\sum\|v_{n+1}\|^2$ are converging sums, and $(\|\tilde{g}\|^2)_{n\in\mathbb{N}}$ and $(\|v_{n+1}\|^2)_{n\in\mathbb{N}}$ tend to 0.

To prove the global convergence of our discrete scheme (i.e. that $d(0,\partial F(x_n))$ converges to 0), remember that:

$$\lambda\tilde{g} + \nabla f(x_{n+1}) - \nabla f(x_{n+\frac{1}{2}}) \in \partial F(x_{n+1}). \tag{93}$$

Thus using (85) we have:

$$d(0,\partial F(x_{n+1})) \leqslant \|\lambda\tilde{g} + \nabla f(x_{n+1}) - \nabla f(x_{n+\frac{1}{2}})\| \leqslant \lambda\|\tilde{g}\|.$$

Since $\|\tilde{g}\|$ converges to 0, $d(0,\partial F(x_{n+1}))$ also converges to 0.

6.4.2 Proof of Point 2. Convergence of $F(x_n) - F^*$ and $d(x_n, X^*)$ to 0.

Observe first that in the differentiable and non-differentiable cases we have:

$$\forall n\in\mathbb{N},\ F(x_n) - F^* = L_n - \frac{\rho}{2}\|v_n\|^2.$$

We prove in Step 1 that the sequence $(L_n)_{n\in\mathbb{N}}$ converges in $\mathbb{R}_+$ and that $(v_n)_{n\in\mathbb{N}}$ converges to 0, hence $(F(x_n)-F^*)_{n\in\mathbb{N}}$ also converges in $\mathbb{R}_+$ and is thus bounded.

Assume now that F is additionally coercive. Using the coercivity assumption combined with the boundedness of $(F(x_n)-F^*)_{n\in\mathbb{N}}$, we deduce that the sequence $(x_n)_{n\in\mathbb{N}}$ is bounded (as well as the sequence $(x_{n+\frac{1}{2}})_{n\in\mathbb{N}}$). Thus there exists a compact set C containing X^* and such that:

$$\{x_n\ ;\ n\in\mathbb{N}\}\cup\{x_{n+\frac{1}{2}}\ ;\ n\in\mathbb{N}\}\subset C. \tag{94}$$

Convergence of $F(x_n) - F^*$ to 0 in the differentiable case. Since the sequence $(\nabla F(x_n))_{n\in\mathbb{N}}$ converges to 0 as shown in Step 1, we deduce that any accumulation point of the sequence $(x_n)_{n\in\mathbb{N}}$ of iterates is a critical point, and thus a minimizer of F. By continuity of F and remembering that $(F(x_n)-F^*)_{n\in\mathbb{N}}$ converges, necessarily $(F(x_n)-F^*)_{n\in\mathbb{N}}$ converges to 0.

Convergence of $F(x_n) - F^*$ to 0 in the non-differentiable case. Since F is assumed convex and lower semicontinuous, its sub-differential is a non empty compact convex set and:

$$\forall n\in\mathbb{N},\ \exists s_n\in\partial F(x_n),\ d(0,\partial F(x_n)) = \|s_n\|. \tag{95}$$

Since $d(0,\partial F(x_n))$ converges to 0 according to Step 1, the sequence $(s_n)_{n\in\mathbb{N}}$ also converges to 0.

Let us now introduce the set:

$$\tilde{C} = \{(x,s)\in\mathbb{R}^n\times\mathbb{R}^n; x\in C,\ s\in\partial F(x)\}\subset C\times\partial F(C) \tag{96}$$

where C is the compact set defined by (94). Since F is a proper lower semicontinuous convex function, $\tilde{C}$ is a compact subset of $\mathbb{R}^n\times\mathbb{R}^n$, see [26, Propositions 6.2.1 and 6.2.2]. Observe now that by construction, $(x_n, s_n)\in\tilde{C}$ for all $n\in\mathbb{N}$. There so exist a non-decreasing function $\varphi:\mathbb{N}\to\mathbb{N}$ and $(\bar{x},\bar{s})\in\tilde{C}$ such that the sub-sequence $(x_{\varphi(n)}, s_{\varphi(n)})$ converges to $(\bar{x},\bar{s})$ when n tends to $+\infty$. Remember now that $(s_n)_{n\in\mathbb{N}}$ converges to 0 and $(\bar{x},\bar{s})\in\tilde{C}$, hence necessarily: $\bar{s}=0$ and:

$$0\in\partial F(\bar{x}). \tag{97}$$

We so proved that the sub-sequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$ converges to a minimizer of F. By local Lipschitz continuity of F on the relative interior of its domain, we then deduce that $(F(x_{\varphi(n)}))_{n\in\mathbb{N}}$ converges to $F(\bar{x}) = F^*$. Since the sequence $(F(x_n)-F^*)_{n\in\mathbb{N}}$ converges, we deduce that $(F(x_n)-F^*)_{n\in\mathbb{N}}$ converges to 0 when $n\to 0$.

Convergence of $d(x_n, X^*)$ to 0 in the differentiable and non-differentiable cases. Assume on the contrary that the sequence $d(x_n, X^*)$ does not converge to 0: there thus exist $\varepsilon > 0$ and a non-increasing function $\varphi : \mathbb{N} \to \mathbb{N}$ such that the sub-sequence $(x_{\varphi(n)})_{n\in\mathbb{N}}$ satisfies:

$$\forall n \in \mathbb{N},\ d(x_{\varphi(n)}, X^*) \geqslant \varepsilon. \tag{98}$$

Let $K_\varepsilon = C \cap \{x \in \mathbb{R}^n; d(x, X^*) \geqslant \varepsilon\}$. By construction, K_ε is a compact subset of $\mathbb{R}^n$ and $K_\varepsilon \cap X^* = \emptyset$. Moreover for all $n \in \mathbb{N}$, we have: $x_{\varphi(n)} \in K_\varepsilon$ so that there exists a convergent subsequence $(x_{\psi\circ\varphi(n)})_{n\in\mathbb{N}}$ whose limit denoted by $\bar{x}$ belongs to K_ε, and thus: $\bar{x} \notin X^*$.

To conclude, note that in the differentiable case as in the non-differentiable case, we can deduce that $F(x_{\psi\circ\varphi(n)}) - F^*$ then converges to $F(\bar{x}) - F^*$. Since the whole sequence $(F(x_n) - F^*)_{n\in\mathbb{N}}$ converges to 0 when $n \to 0$, necessarily: $F(\bar{x}) - F^* = 0$ which is impossible since $\bar{x} \notin X^*$.

6.4.3 Proof of Point 3. Convergence of the iterates to a minimizer of F.

Assume now that F is coercive and additionally satisfies any Łojasiewicz property with an exponent $\theta \in [0, 1)$. Under these assumptions, the set X^* of the minimizers is compact and according to Lemma 1 there exists $\mu > 0$ and $\varepsilon > 0$ such that:

$$d(x, X^*) \leqslant \varepsilon \Rightarrow 2\mu\left(F(x) - F^*\right)^\theta \leqslant d(0, \partial F(x)). \tag{99}$$

Proof of Point 3(a). Assume that $\theta = \frac{1}{2}$ i.e. that F satisfies the $\mathcal{PL}_\mu$ property. We now prove that the values $F(x_n) - F^*$ converge exponentially to 0.

Consider first the differentiable case. Observe that: $x_{n+1} = x_{n+\frac{1}{2}} - \frac{t^2}{\lambda} g$ and:

$$\lambda g + \nabla F(x_{n+1}) - \nabla F(x_{n+\frac{1}{2}}) = \nabla F(x_{n+1}).$$

Following exactly the same steps as those which led to inequality (85), we can prove that:

$$\|\nabla F(x_{n+1})\|^2 \leqslant \lambda^2\|g\|^2 - \|\nabla F(x_{n+1}) - \nabla F(x_{n+\frac{1}{2}})\|^2 \leqslant \lambda^2\|g\|^2$$

Since $d(x_{n+\frac{1}{2}}, X^*)$ tends to 0 when $n \to +\infty$, we now write the $\mathcal{PL}_\mu$ property satisfied by F at x_{n+1} i.e. there exists $n_0 \in \mathbb{N}$ such that:

$$\forall n \geqslant n_0,\ 2\mu\left(F(x_{n+1}) - F(x^*)\right) \leqslant \|\nabla F(x_{n+1})\|^2 \leqslant \lambda^2\|g\|^2. \tag{100}$$

that is:

$$-\|g\|^2 \leqslant -\frac{2\mu}{\lambda^2}\left(F(x_{n+1}) - F(x^*)\right). \tag{101}$$

Combined with (89), this inequality ensures that

$$\begin{aligned} L_{n+1} - L_n &\leqslant -t^3\left(\frac{1+\gamma t}{1+t}\right)\frac{\mu}{\lambda^2}\left(F(x_{n+1}) - F(x^*)\right) - \frac{\rho}{2}(2\gamma t + \gamma^2 t^2)\|v_{n+1}\|^2 \\ &\leqslant -t\left[\left(\frac{1+\gamma t}{1+t}\right)\frac{\mu}{L}\left(F(x_{n+1}) - F(x^*)\right) + \frac{\rho}{2}(2\gamma + \gamma^2 t)\|v_{n+1}\|^2\right] \end{aligned}$$

Let us define:

$$K = \min\left(\frac{\mu(1+\gamma t)}{L(1+t)}, \gamma(2+\gamma t)\right) \tag{102}$$

We can remark that we have $K > 0$ for any $t > 0$ so that:

$$L_{n+1} - L_n \leqslant -tKL_{n+1}. \tag{103}$$

This ensures that for any $n \geqslant n_0$

$$L_{n+1} \leqslant (1+tK)^{-1} L_n. \tag{104}$$

Note that in the non differentiable case, the very last inequality on the energy L_n is unchanged: the proof is exactly the same by replacing g by $\tilde{g}$ and $\nabla F(x_{n+1})$ by $\lambda\tilde{g} + \nabla f(x_{n+1}) - \nabla f(x_{n+\frac{1}{2}})$.

In both cases, the inequality (104) ensures that $(L_n)_{n\in\mathbb{N}}$ and thus $(F(x_n) - F(x^*))_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ decay exponentially. It follows in particular that the series $\sum_n v_n$ converges. Moreover, combining the second and the fourth lines of (81) in the differentiable case, we get:

$$t\left(\frac{\gamma t^2 - 1}{1+t}\right) g \quad = \quad (1+\gamma t)v_{n+1} - v_n.$$

Since $\sum v_n$ converges, the series $\sum_n g$ also converges. Noticing that $\lambda(x_{n+1} - x_n) = tv_n - t^2 g$, we finally deduce the convergence of the sequence $(x_n)_{n\in\mathbb{N}}$. Since X^* is a closed convex set and since $d(x_n, X^*)$ tends to 0 when $n \to +\infty$, necessarily $(x_n)_{n\in\mathbb{N}}$ converges to a minimizer of F.

In the non-differentiable case, the conclusions are the same since we similarly have:

$$t\left(\frac{\gamma t^2 - 1}{1+t}\right)\tilde{g} = (1+\gamma t)v_{n+1} - v_n, \qquad \lambda(x_{n+1} - x_n) = tv_n - t^2\tilde{g}$$

and thus the convergence of the series $\sum \tilde{g}$, and the convergence of $(x_n)_{n\in\mathbb{N}}$ to a minimizer of F.

Optimizing the constant tK. We now focus on optimizing the constant tK, where K is given by (102). We can see that the two functions inside the minimum of the definition of K are increasing functions of γ. Hence the larger γ, then the larger K.

Remembering the constraint $\gamma t^2 < 1$, i.e.: $\gamma < \frac{1}{t^2}$, the best choice is $\gamma = \frac{1-\epsilon}{t^2}$ (with $\epsilon \to 0^+$). For this choice of γ, we have:

$$tK = \min\left(\frac{\mu}{L}\left(1 - \frac{\epsilon}{1+t}\right), (1-\epsilon)\left(\frac{1+2t-\epsilon}{t^2}\right)\right). \tag{105}$$

Since $t \mapsto (1-\epsilon)\left(\frac{1+2t}{t^2} - \epsilon\right)$ is a decreasing function on $\mathbb{R}_+^*$ and goes to $+\infty$ when $t \to 0^+$, we see that the best possible constant tK is thus $\frac{\mu}{L}\left(1 - \frac{\epsilon}{1+t}\right)$. Thus for a given $\epsilon > 0$, there exists $t_\epsilon > 0$ such that:

$$tK = \frac{\mu}{L}\left(1 - \frac{\epsilon}{1+t}\right).$$

Proof of Point 3(b). Assume now that F satisfies a Łojasiewicz property with an exponent $\theta \in (\frac{1}{2}, 1)$ i.e. that there exists $\mu > 0$ and $\varepsilon > 0$ such that:

$$d(x, X^*) \leqslant \varepsilon \Rightarrow 2\mu\left(F(x) - F^*\right)^\theta \leqslant d(0, \partial F(x)). \tag{106}$$

We eventually prove that the decay of the values $F(x_n) - F^*$ is now polynomial both in the differentiable and non-differentiable cases.

The differentiable case. To lighten the notations, let us rewrite the inequality (89) as:

$$L_{n+1} - L_n \leqslant -A_1\|g\|^2 - A_2\frac{\rho}{2}\|v_{n+1}\|^2. \tag{107}$$

where A_1 and A_2 are two positive constants. Using the Łojasiewicz property applied at $x_{n+\frac{1}{2}}$ with $g = \frac{1}{\lambda}\nabla F(x_{n+\frac{1}{2}})$:

$$\frac{2\mu}{\lambda}(F(x_{n+\frac{1}{2}}) - F^*)^\theta \leqslant \|g\|.$$

and the decrease inequality (88) that implies: $F(x_{n+1}) - F^* \leqslant F(x_{n+\frac{1}{2}}) - F^*$, we get:

$$L_{n+1} - L_n \leqslant -A_1 \frac{4\mu^2}{\lambda^2}(F(x_{n+1}) - F(x^*))^{2\theta} - A_2 \frac{\rho}{2} \|v_{n+1}\|^{2-4\theta+4\theta}. \tag{108}$$

In the first step of the proof we prove that $\|v_{n+1}\|$ tends to 0. It follows that $\|v_{n+1}\|$ is bounded. Since $2 - 4\theta < 0$ we can then find $A_3 > 0$ such that for any $n \in \mathbb{N}$,

$$L_{n+1} - L_n \leqslant -A_3 \left(\frac{1}{2}(F(x_{n+1}) - F(x^*))^{2\theta} + \frac{1}{2}\left(\frac{\rho}{2}\right)^{2\theta} \|v_{n+1}\|^{4\theta}\right). \tag{109}$$

Using the convexity of the function $x \mapsto x^{2\theta}$ observe that:

$$\left(\frac{1}{2}(F(x_{n+1}) - F(x^*)) + \frac{1}{2}\left(\frac{\rho}{2}\right)\|v_{n+1}\|^2\right)^{2\theta} \leqslant \frac{1}{2}(F(x_{n+1}) - F(x^*))^{2\theta} + \frac{1}{2}\left(\frac{\rho}{2}\right)^{2\theta}\|v_{n+1}\|^{4\theta} \tag{110}$$

It follows that

$$L_{n+1} - L_n \leqslant -\frac{A_3}{4^\theta} L_{n+1}^{2\theta}. \tag{111}$$

We now apply Lemma 5 to the sequence $(L_n)_{n\in\mathbb{N}}$ to deduce that there exists $K > 0$ such that for all $n \in \mathbb{N}$, $F(x_n) - F(x^*) \leqslant L_n \leqslant \frac{K}{n^{\frac{1}{2\theta-1}}}$. Moreover we have

$$\|v_n\|^2 \leqslant \frac{2K}{\rho n^{\frac{1}{2\theta-1}}}. \tag{112}$$

If $\frac{1}{2\theta-1} > 2$, i.e. if $\theta < \frac{3}{4}$, it follows that the sequence $\sum_{n\geqslant 1}\|v_n\|$ is finite. Using the fact that

$$x_{n+1} - x_n = tv_n - t^2 g \tag{113}$$

and

$$g = \left(t\left(1 - \frac{(1+\gamma t)t}{1+t}\right)\right)^{-1}(v_n - (1+\gamma t)v_{n+1}) \tag{114}$$

we deduce that the sequence $\sum_{n\geqslant 1}\|x_{n+1} - x_n\|$ is finite and thus that the sequence $(x_n)_{n\in\mathbb{N}}$ converges to a minimizer x^* of F.

The non-differentiable case. In the non-differentiable case, the proof remains unchanged by replacing one more time g by $\tilde{g}$. The only difference is the way to obtain the inequality (108). Let us prove that the inequality (108) is still valid by replacing g by $\tilde{g}$ in the nonsmooth case.

First we apply the Łojasiewicz property at x_{n+1} choosing $\lambda\tilde{g} + \nabla f(x_{n+1}) - \nabla f(x_{n+\frac{1}{2}})$ as a subgradient of F at x_{n+1}: for n large enough, we have:

$$2\mu(F(x_{n+1}) - F^*)^\theta \leqslant \|\lambda\tilde{g} + \nabla f(x_{n+1}) - \nabla f(x_{n+\frac{1}{2}})\|.$$

Using now the inequality (85): $\|\lambda\tilde{g} + \nabla f(x_{n+1}) - \nabla f(x_{n+\frac{1}{2}})\| \leqslant \lambda\|\tilde{g}\|$, we then obtain:

$$\frac{2\mu}{\lambda}(F(x_{n+1}) - F(x^*))^\theta \leqslant \|\tilde{g}\|$$

as expected.

6.4.4 Technical result

To derive the polynomial rates stated in the point 3(b) of Theorems 6 and 8, we have used the following technical result:

Lemma 5. *Let $(u_n)_{n\in\mathbb{N}}$ be a non negative sequence. Assume that there exist $\theta > \frac{1}{2}$ and $a > 0$ such that:*

$$\forall n \in \mathbb{N},\ u_{n+1} + a u_{n+1}^{2\theta} \leqslant u_n. \tag{115}$$

Then it exists $K > 0$ such that

$$\forall n \geqslant 1,\ u_n \leqslant \frac{K}{n^{\frac{1}{2\theta-1}}}. \tag{116}$$

Proof. Let us define $w_n := \left(\frac{1}{n^{\frac{1}{2\theta-1}}} - \frac{1}{(n+1)^{\frac{1}{2\theta-1}}}\right)(n+1)^{\frac{2\theta}{2\theta-1}}$. This sequence converges to $\frac{1}{2\theta-1}$ and it is thus bounded i.e. it exists $A > 0$ such that for all $n \in \mathbb{N}$, $w_n \leqslant A$ and we can observe that

$$w_n \leqslant A \Longleftrightarrow \frac{1}{n^{\frac{1}{2\theta-1}}} \leqslant \frac{A}{(n+1)^{\frac{2\theta}{2\theta-1}}} + \frac{1}{(n+1)^{\frac{1}{2\theta-1}}}. \tag{117}$$

Let us define $f(x) = x + ax^{2\theta}$. This function is non decreasing on $(0, +\infty)$. Moreover for any $K > 0$ we have:

$$f\left(\frac{K}{(n+1)^{\frac{1}{2\theta-1}}}\right) = K\left(\frac{1}{(n+1)^{\frac{1}{2\theta-1}}} + \frac{aK^{2\theta-1}}{(n+1)^{\frac{2\theta}{2\theta-1}}}\right). \tag{118}$$

It turns out that for $K \geqslant \left(\frac{A}{a}\right)^{\frac{1}{2\theta-1}}$ we get using (117):

$$\frac{K}{n^{\frac{1}{2\theta-1}}} \leqslant K\left(\frac{1}{(n+1)^{\frac{1}{2\theta-1}}} + \frac{A}{(n+1)^{\frac{2\theta}{2\theta-1}}}\right) \leqslant K\left(\frac{1}{(n+1)^{\frac{1}{2\theta-1}}} + \frac{aK^{2\theta-1}}{(n+1)^{\frac{2\theta}{2\theta-1}}}\right) = f\left(\frac{K}{(n+1)^{\frac{1}{2\theta-1}}}\right) \tag{119}$$

It follows that if $u_n \leqslant \frac{K}{n^{\frac{1}{2\theta-1}}}$ we get using (115):

$$f(u_{n+1}) \leqslant u_n \leqslant \frac{K}{n^{\frac{1}{2\theta-1}}} \leqslant f\left(\frac{K}{(n+1)^{\frac{1}{2\theta-1}}}\right). \tag{120}$$

since f is non increasing it follows that $u_{n+1} \leqslant \frac{K}{(n+1)^{\frac{1}{2\theta-1}}}$.

Choosing $K = \max\left(u_1, \left(\frac{A}{a}\right)^{\frac{1}{2\theta-1}}\right)$, we conclude by induction on n. □

6.5 Proof of Theorem 7

The proof of Theorem 7 is based on a Lyapunov analysis inspired by the one proposed for the ODE. To study the ODE we used to following Lyapunov energy:

$$\mathcal{E}(t) = F(x(t)) - F^* + \frac{1}{2}\|\lambda(x(t) - x^*) + \dot{x}(t)\|^2 + \frac{\xi}{2}\|x(t) - x^*\|^2 \tag{121}$$

with

$$\lambda = \sqrt{\mu}\ ,\ \alpha = \beta\sqrt{\mu} = \beta\lambda\ ,\ \xi = (1-\beta)\mu = (1-\beta)\lambda^2 \tag{122}$$

where $\beta = 2 - \frac{\sqrt{2}}{2}$. Moreover, we have

$$\delta = (2 - \sqrt{2})\sqrt{\mu} = 2(\beta - 1)\lambda. \tag{123}$$

Let us recall the discrete scheme we consider:

$$\begin{cases} x_{n+\frac{1}{2}} = & x_n + sv_n \\ v_{n+\frac{1}{2}} = & (1+\alpha s)^{-1}(v_n - s\nabla F(x_{n+\frac{1}{2}})) \\ x_{n+1} = & x_{n+\frac{1}{2}} - s^2 \nabla F(x_{n+\frac{1}{2}}) \\ v_{n+1} = & v_{n+\frac{1}{2}} + (1+\lambda s)^{-1}\lambda s^2 \nabla F(x_{n+\frac{1}{2}}) \end{cases} \tag{124}$$

To simplify the writing of the proof, we introduce the following reduced notations:

$$t = \lambda s, \quad u = \lambda(x_{n+\frac{1}{2}} - x^*), \quad v = v_{n+\frac{1}{2}} \text{ and } g = \frac{1}{\lambda}\nabla F(x_{n+\frac{1}{2}}). \tag{125}$$

With these reduced notations, the scheme may be rewritten as:

$$\begin{cases} \lambda x_{n+\frac{1}{2}} &= \lambda x_n + tv_n \\ v := v_{n+\frac{1}{2}} &= (1+\beta t)^{-1}(v_n - tg) \\ \lambda x_{n+1} &= \lambda x_{n+\frac{1}{2}} - t^2 g \\ v_{n+1} &= v + (1+t)^{-1} t^2 g. \end{cases} \tag{126}$$

Remember that the value of v_{n+1} is actually chosen such that

$$\|\lambda(x_{n+1} - x^*) + (1+t)v_{n+1}\|^2 = \|u + (1+t)v\|^2. \tag{127}$$

To study the properties of the scheme (124), we define the sequence L_n:

$$L_n := L(x_n, v_n) := F(x_n) - F^* + \frac{1}{2}\|\lambda(x_n - x^*) + (1+\lambda s)v_n\|^2 - (\beta-1)\frac{\lambda^2}{2}\|x_n - x^*\|^2$$

which can be seen as particular discretization of the energy $\mathcal{E}$. Let us first compute $L_{n+\frac{1}{2}}$ and L_n using the reduced notations:

$$L_{n+\frac{1}{2}} := L(x_{n+\frac{1}{2}}, v_{n+\frac{1}{2}}) := F(x_{n+\frac{1}{2}}) - F^* + \frac{1}{2}\|u + (1+t)v\|^2 - \frac{\beta-1}{2}\|u\|^2 \tag{128}$$

and L_n can be written in the following way as:

$$\begin{aligned} L_n &= F(x_n) - F^* + \frac{1}{2}\|u + \lambda(x_n - x_{n+\frac{1}{2}}) + (1+t)v_n\|^2 - \frac{\beta-1}{2}\|u + \lambda(x_n - x_{n+\frac{1}{2}})\|^2 \\ &= F(x_n) - F^* + \frac{1}{2}\left\|u + v_n\right\|^2 - \frac{\beta-1}{2}\left\|u - tv_n\right\|^2. \end{aligned}$$

Now, noticing that:

$$\begin{aligned} \frac{-(\beta-1)\lambda^2}{2}\|x_{n+1} - x^*\|^2 &= \frac{-(\beta-1)\lambda^2}{2}\|x_{n+1} - x_{n+\frac{1}{2}} + x_{n+\frac{1}{2}} - x^*\|^2 \\ &= -\frac{\beta-1}{2}\|u\|^2 - \frac{(\beta-1)t^4}{2}\|g\|^2 + (\beta-1)t^2\langle u, g\rangle \end{aligned}$$

the energy L_{n+1} can be expressed as a variation of $L_{n+\frac{1}{2}}$:

$$L_{n+1} = L_{n+\frac{1}{2}} + F(x_{n+1}) - F(x_{n+\frac{1}{2}}) - (\beta-1)\frac{t^4}{2}\|g\|^2 + (\beta-1)t^2\langle u, g\rangle. \tag{129}$$

To prove Theorem 1 we demonstrated that the Lyapunov Energy defined by (121) satisfies

$$\mathcal{E}'(t) + \delta\mathcal{E}(t) \leqslant 0. \tag{130}$$

To prove Theorem 7 we will use the following Lemma whose proof is left to Subsection 6.5.1:

Lemma 6. *There exist $t_1 > 0$ and $K_3 > 0$ such that for any $t \in [0, t_1]$, it holds:*

$$\forall n \in \mathbb{N}, \quad \left(1 + \frac{\delta}{\lambda} t - t^2 K_3\right) L_{n+1} - L_n \leqslant 0. \tag{131}$$

Moreover, we then have for any $t \in [0, t_1]$,

$$\forall n \in \mathbb{N}, \; L_n \leqslant \left(1 + \frac{\delta}{\lambda} t - K_3 t^2\right)^{-n} L_0. \tag{132}$$

To get the control on the values $F(x_n) - F^*$ from the control on the energy L_n provided by Lemma 6, note that:

$$\begin{aligned} L_n &= F(x_n) - F^* + \frac{1}{2}\|\lambda(x_n - x^*) + (1 + \lambda s) v_n\|^2 - (\beta - 1)\frac{\lambda^2}{2}\|x_n - x^*\|^2 \\ &\geqslant F(x_n) - F^* - (\beta - 1)\frac{\mu}{2}\|x_n - x^*\|^2. \end{aligned} \tag{133}$$

since $\lambda = \sqrt{\mu}$. Observe now that assuming the uniqueness of the minimizer x^* of F, Theorem 6 ensures that the sequence of iterates $(x_n)_{n \in \mathbb{N}}$ converges to x^*. Thus using the quadratic growth condition $\mathcal{G}_\mu^2$ satisfied by F, there exists $N \in \mathbb{N}$ such that:

$$\forall n \geqslant N, \; \frac{\mu}{2}\|x_n - x^*\|^2 \leqslant F(x_n) - F^*.$$

Combining this last inequality with (133), we finally get: $(2 - \beta)\left(F(x_n) - F^*\right) \leqslant L_n$ for any $n \geqslant N$, and thus the expected control on $F(x_n) - F^*$.

6.5.1 Proof of Lemma 6

Sketch of proof: The proof of Lemma 6 is technical. This is the reason why we first give a structure of it:

1. A key descent inequality (134) used previously by Siegel is proven.
2. We give an upper bound of $L_{n+\frac{1}{2}} - L_n$.
3. From this bound and (134) we give an upper bound of $L_{n+1} - L_n$.
4. We deduce a bound on $(1 + t\delta/\lambda)L_{n+1} - L_n$ as a polynomial in t whose coefficients depend on u and v.
5. We conclude by bounding this polynomial by L_{n+1}.

Proof of Lemma 6: **Step 1.** We first prove the inequality (134)

$$F(x_{n+1}) - F(y) \leq \lambda\langle g, x_{n+\frac{1}{2}} - y\rangle - \frac{t^2}{2}\|g\|^2 \tag{134}$$

which is a key inequality used by Siegel in [41]. Since ∇F is $L = \frac{1}{s^2}$- Lipschitz and $g := \frac{1}{\lambda}\nabla F(x_{n+\frac{1}{2}}) := \frac{\lambda}{t^2}(x_{n+1} - x_{n+\frac{1}{2}})$ we get

$$F(x_{n+1}) - F(x_{n+\frac{1}{2}}) \leqslant -\frac{t^2}{2}\|g\|^2. \tag{135}$$

Since F is convex, for any $y \in \mathbb{R}^n$ we have

$$F(x_{n+\frac{1}{2}}) - F(y) \leqslant \lambda \langle g, x_{n+\frac{1}{2}} - y \rangle. \tag{136}$$

Inequality (134) holds summing the two previous inequalities.
Step 2.

$$L_{n+\frac{1}{2}} - L_n = F(x_{n+\frac{1}{2}}) - F(x_n) + \frac{1}{2}\|u + (1+t)v\|^2 - \frac{1}{2}\|u + v_n\|^2 - \frac{\beta - 1}{2}\left(\|u\|^2 - \|u - tv_n\|^2\right)$$

We use the identity (with the condition $A - B = a + b$):

$$\frac{1}{2}\|A\|^2 - \frac{1}{2}\|B\|^2 = \langle a, B \rangle + \langle b, A \rangle + \frac{1}{2}\|a\|^2 - \frac{1}{2}\|b\|^2 \tag{137}$$

with $A = u + (1+t)v,\ B = u + v_n,\ a = -tg,\ b = -(\beta - 1)tv$. We thus get with this identity:

$$\begin{aligned}
\frac{1}{2}\|u + (1+t)v\|^2 - \frac{1}{2}\|u + v_n\|^2 &= -t\langle g, u + v_n \rangle - (\beta - 1)t\langle v, u + (1+t)v \rangle + \frac{t^2}{2}\|g\|^2 \\
&\quad - \frac{(\beta - 1)^2 t^2}{2}\|v\|^2 \\
&= -t\langle g, v_n \rangle - t\langle g, u \rangle + \frac{t^2}{2}\|g\|^2 - (\beta - 1)t\langle v, u \rangle \\
&\quad - \frac{(\beta - 1)t}{2}(2 + (\beta + 1)t)\|v\|^2
\end{aligned}$$

Moreover observe that:

$$-\frac{\beta - 1}{2}\|u\|^2 + \frac{\beta - 1}{2}\|u - tv_n\|^2 = -(\beta - 1)t\langle v_n, u \rangle + (\beta - 1)\frac{t^2}{2}\|v_n\|^2 \tag{138}$$

Hence:

$$\begin{aligned}
L_{n+\frac{1}{2}} - L_n &= F(x_{n+\frac{1}{2}}) - F(x_n) - \langle tg, v_n \rangle - t\langle g, u \rangle + \frac{t^2}{2}\|g\|^2 - (\beta - 1)t\langle v, u \rangle \\
&\quad - \frac{(\beta - 1)t}{2}(2 + (\beta + 1)t)\|v\|^2 - t(\beta - 1)\langle v_n, u \rangle + \frac{\beta - 1}{2}t^2\|v_n\|^2.
\end{aligned}$$

Using the expression of $v_n = (1 + \beta t)v + tg$, we get:

$$\begin{aligned}
L_{n+\frac{1}{2}} - L_n &= F(x_{n+\frac{1}{2}}) - F(x_n) - \langle tg, v_n \rangle + \frac{t^2}{2}\|g\|^2 - t\langle g, u \rangle - t(\beta - 1)\langle u, v \rangle \\
&\quad - \frac{(\beta - 1)t}{2}(2 + (\beta + 1)t)\|v\|^2 - t(\beta - 1)\langle (1 + \beta t)v + tg, u \rangle \\
&\quad + (\beta - 1)\frac{t^2}{2}\|(1 + \beta t)v + tg\|^2 \\
&= F(x_{n+\frac{1}{2}}) - F(x_n) - \langle tg, v_n \rangle + \frac{t^2}{2}\|g\|^2 - t\left(1 + (\beta - 1)t\right)\langle g, u \rangle + \frac{\beta - 1}{2}t^4\|g\|^2 \\
&\quad - t(\beta - 1)\left(2 + \beta t\right)\langle v, u \rangle - \frac{\beta - 1}{2}t\left(2 + (\beta + 1)t - t(1 + \beta t)^2\right)\|v\|^2 \\
&\quad + (\beta - 1)(1 + \beta t)t^3\langle v, g \rangle.
\end{aligned}$$

Step 3. Using (129), we get:

$$\begin{aligned} L_{n+1} - L_n &= L_{n+1} - L_{n+\frac{1}{2}} + L_{n+\frac{1}{2}} - L_n \\ &= F(x_{n+1}) - F(x_n) - t\langle g, v_n\rangle + \frac{t^2}{2}\|g\|^2 - t\langle g, u\rangle \\ &\quad -t(\beta-1)\left(2+\beta t\right)\langle v, u\rangle - \frac{\beta-1}{2}t\left(2+(\beta+1)t - t(1+\beta t)^2\right)\|v\|^2 \\ &\quad +(\beta-1)(1+\beta t)t^3\langle v, g\rangle \end{aligned} \tag{139}$$

Then, we apply (134) with $y = x_n$ and $tv_n = \lambda(x_{n+\frac{1}{2}} - x_n)$ to get:

$$F(x_{n+1}) - F(x_n) - t\langle g, v_n\rangle \leq -\frac{t^2}{2}\|g\|^2 \tag{140}$$

and (134) with $y = x^*$ to get:

$$\langle g, u\rangle \geq F(x_{n+1}) - F^* + \frac{t^2}{2}\|g\|^2. \tag{141}$$

Combining (139), (140) and (141) we deduce:

$$\begin{aligned} L_{n+1} - L_n &\leq -t(F(x_{n+1}) - F^*) - t(\beta-1)\left(2+\beta t\right)\langle v, u\rangle + (\beta-1)(1+\beta t)t^3\langle v, g\rangle \\ &\quad -\frac{t^3}{2}\|g\|^2 - \frac{\beta-1}{2}t\left(2+(\beta+1)t - t(1+\beta t)^2\right)\|v\|^2. \end{aligned}$$

Step 4. Using the following expression of $F(x_{n+1}) - F^*$:

$$\begin{aligned} F(x_{n+1}) - F^* &= L_{n+1} - \frac{1}{2}\|\lambda(x_{n+1} - x^*) + (1+t)v_{n+1}\|^2 + \frac{(\beta-1)\lambda^2}{2}\|x_{n+1} - x^*\|^2 \\ &= L_{n+1} - \frac{1}{2}\|u + (1+t)v\|^2 + \frac{\beta-1}{2}\|u - t^2 g\|^2 \\ &= L_{n+1} - \frac{1}{2}\|u + (1+t)v\|^2 + \frac{\beta-1}{2}\left(t^4\|g\|^2 + \|u\|^2 - 2t^2\langle u, g\rangle\right) \\ &= L_{n+1} - \frac{2-\beta}{2}\|u\|^2 - \frac{(1+t)^2}{2}\|v\|^2 + \frac{(\beta-1)t^4}{2}\|g\|^2 - (1+t)\langle u, v\rangle - (\beta-1)t^2\langle u, g\rangle \end{aligned}$$

we eventually get, remembering that $\frac{\delta}{\lambda} = 2(\beta-1)$ and thus $1 - \frac{\delta}{\lambda} = 3 - 2\beta$, so that $t = 2(\beta-1)t + (3-2\beta)t$:

$$\begin{aligned} \left(1 + \frac{\delta}{\lambda}t\right)L_{n+1} - L_n &\leq -(3-2\beta)t(F(x_{n+1}) - F^*) - \frac{t^3}{2}\|g\|^2 - t(\beta-1)\left(2+\beta t\right)\langle v, u\rangle \\ &\quad -\frac{\beta-1}{2}t\left(2+(\beta+1)t - t(1+\beta t)^2\right)\|v\|^2 + (\beta-1)(1+\beta t)t^3\langle v, g\rangle \\ &\quad -2(\beta-1)t\Bigg(-\frac{(2-\beta)t}{2}\|u\|^2 - \frac{(1+t)^2}{2}\|v\|^2 \\ &\quad +\frac{(\beta-1)t^4}{2}\|g\|^2 - (1+t)\langle u, v\rangle - (\beta-1)t^2\langle u, g\rangle\Bigg) \\ &\leq -(3-2\beta)t(F(x_{n+1}) - F^*) - \frac{t^3}{2}\|g\|^2 - t(\beta-1)\left(2+\beta t\right)\langle v, u\rangle \\ &\quad -\frac{\beta-1}{2}t\left(2+(\beta+1)t - t(1+\beta t)^2\right)\|v\|^2 + (\beta-1)(1+\beta t)t^3\langle v, g\rangle \\ &\quad +(\beta-1)^2(2-\beta)t\|u\|^2 + (\beta-1)t(1+t)^2\|v\|^2 - (\beta-1)^2 t^5\|g\|^2 \\ &\quad -2(\beta-1)t(1+t)\langle u, v\rangle + 2(\beta-1)^2 t^3\langle u, g\rangle \end{aligned}$$

So that:

$$\left(1+\frac{\delta}{\lambda}t\right)L_{n+1}-L_n \leq -(3-2\beta)t(F(x_{n+1})-F^*)+(\beta-1)(2-\beta)t\|u\|^2$$
$$+(\beta-1)(2-\beta)t^2\langle v,u\rangle+\frac{\beta-1}{2}t^2\left(4-\beta+2(\beta+1)t+\beta^2t^2\right)\|v\|^2$$
$$+(\beta-1)(1+\beta t)t^3\langle v,g\rangle+2(\beta-1)^2t^3\langle u,g\rangle-\frac{t^3}{2}(1+2(\beta-1)^2t^2)\|g\|^2.$$

Using the growth condition and remembering that $\lambda^2=\mu$:

$$F(x_{n+1})-F(x^*)\geq\frac{\mu}{2}\|x_{n+1}-x\|^2=\frac{\mu}{2\lambda^2}\|u-t^2g\|^2=\frac{1}{2}\|u-t^2g\|^2 \tag{142}$$

we get:

$$\left(1+\frac{\delta}{\lambda}t\right)L_{n+1}-L_n \leq -\frac{3-2\beta}{2}t\|u-t^2g\|^2+(\beta-1)(2-\beta)t\|u\|^2$$
$$+(\beta-1)(2-\beta)t^2\langle v,u\rangle+\frac{\beta-1}{2}t^2\left(4-\beta+2(\beta+1)t+\beta^2t^2\right)\|v\|^2$$
$$+(\beta-1)(1+\beta t)t^3\langle v,g\rangle+2(\beta-1)^2t^3\langle u,g\rangle-\frac{t^3}{2}(1+2(\beta-1)^2t^2)\|g\|^2.$$
$$\leq -\frac{3-2\beta}{2}t\|u-t^2g\|^2+(\beta-1)(2-\beta)t\|u-t^2g\|^2$$
$$+(\beta-1)(2-\beta)t^2\langle v,u\rangle+\frac{\beta-1}{2}t^2\left(4-\beta+2(\beta+1)t+\beta^2t^2\right)\|v\|^2$$
$$+(\beta-1)(1+\beta t)t^3\langle v,g\rangle+2(\beta-1)t^3\langle u,g\rangle-\frac{t^3}{2}(1+2(\beta-1)t^2)\|g\|^2.$$

We thus have:

$$\left(1+\frac{\delta}{\lambda}t\right)L_{n+1}-L_n \leq (\beta-1)(2-\beta)t^2\langle v,u\rangle+\frac{\beta-1}{2}t^2\left(4-\beta+2(\beta+1)t+\beta^2t^2\right)\|v\|^2 \tag{143}$$
$$+(\beta-1)(1+\beta t)t^3\langle v,g\rangle+2(\beta-1)t^3\langle u,g\rangle-\frac{t^3}{2}(1+2(\beta-1)t^2)\|g\|^2.$$

We now use the inequality:

$$|\langle tg,v\rangle|\leq\frac{t^2}{2}\|g\|^2+\frac{1}{2}\|v\|^2 \tag{144}$$

so that: $t^3|\langle v,g\rangle|\leq\frac{1}{2}t^4\|g\|^2+\frac{1}{2}t^2\|v\|^2$ and:

$$\left(1+\frac{\delta}{\lambda}t\right)L_{n+1}-L_n \leq (\beta-1)(2-\beta)t^2\langle v,u\rangle+\frac{\beta-1}{2}t^2\left(5-\beta+(3\beta+2)t+\beta^2t^2\right)\|v\|^2$$
$$+2(\beta-1)t^3\langle u,g\rangle-\frac{t^3}{2}(1-(\beta-1)t-(2-3\beta+\beta^2)t^2)\|g\|^2.$$

From the previous inequality, we therefore know that there exists two positive real numbers K_1 and K_2 and two polynomials P_1 and P_2 such that

$$\left(1+\frac{\delta}{\lambda}t\right)L_{n+1}-L_n\leqslant t^2\left(K_1\langle u,v\rangle+P_1(t)\|v\|^2+K_2t\langle u,g\rangle+tP_2(t)\|g\|^2\right) \tag{145}$$

More precisely, we have:

$$K_1 = (\beta - 1)(2 - \beta)\ ,\ K_2 = 2(\beta - 1)\ ,\ P_1(t) = \frac{\beta - 1}{2}\left(5 - \beta + (3\beta + 2)t + \beta^2 t^2\right) \tag{146}$$

and

$$P_2(t) = -\frac{1}{2}(1 - (\beta - 1)t - (2 - 3\beta + \beta^2)t^2). \tag{147}$$

Step 5. This last step relies on the following technical lemma whose proof is straightforward:

Lemma 7. *If $A \geq \frac{1}{4}$, we have:*

$$A\left\|x\right\|^2 + \langle x, y\rangle \leqslant 2A(\left\|x + y\right\|^2 + \left\|y\right\|^2).$$

Applying Lemma 8 or 11 with $x = \sqrt{1+t}v$ and $y = \frac{u}{\sqrt{1+t}}$ and using the fact that for all $t \geqslant 0$, $\frac{P_1(t)}{K_1(1+t)} \geqslant \frac{1}{4}$ we deduce

$$K_1\langle u, v\rangle + P_1(t)\left\|v\right\|^2 \leqslant \frac{2K_1P_1(t)}{(1+t)}\left\|u + (1+t)v\right\|^2 + \frac{2K_1P_1(t)}{(1+t)}\left\|u\right\|^2. \tag{148}$$

Defining $P_3(t) := \frac{2K_1P_1(t)}{(1+t)}$, we get for all $t \geqslant 0$:

$$\left(1 + \frac{\delta}{\lambda}t\right)L_{n+1} - L_n \leqslant t^2\left(P_3(t)\left\|u + (1+t)v\right\|^2 + P_3(t)\left\|u\right\|^2 + K_2t\langle u, g\rangle + tP_2(t)\left\|g\right\|^2\right) \tag{149}$$

Moreover

$$\begin{aligned}
P_3(t)\left\|u\right\|^2 + K_2t\langle u, g\rangle + tP_2(t)\left\|g\right\|^2 &= P_3(t)\left\|u - t^2g\right\|^2 + (2tP_3(t) + K_2)\langle u - t^2g, tg\rangle \\
&\quad + \left(tP_2(t) - t^4P_3(t) + K_2t^2\right)\left\|g\right\|^2 \\
&\leqslant \left(P_3(t) + \frac{1}{2}|2tP_3(t) + K_2|\right)\left\|u - t^2g\right\|^2 \\
&\quad + \left(tP_2(t) - t^4P_3(t) + K_2t^2 + \frac{t^2}{2}|2tP_3(t) + K_2|\right)\left\|g\right\|^2
\end{aligned}$$

where we have used the inequality $\langle a, b\rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$.

We can remark now that $P_2(0) < 0$ which implies that it exists $t_1 > 0$ such that for any $t \in [0, t_1]$,

$$P_3(t)\left\|u\right\|^2 + K_2t\langle u, g\rangle + tP_2(t)\left\|g\right\|^2 \leqslant \left(P_3(t) + \frac{1}{2}|2tP_3(t) + K_2|\right)\left\|u - t^2g\right\|^2 \tag{150}$$

We deduce that for all $t \in [0, t_1]$,

$$\left(1 + \frac{\delta}{\lambda}t\right)L_{n+1} - L_n \leqslant t^2\left(P_3(t)\left\|u + (1+t)v\right\|^2 + \left(P_3(t) + \frac{1}{2}|2tP_3(t) + K_2|\right)\left\|u - t^2g\right\|^2\right) \tag{151}$$

Using the growth condition we get

$$\left\|u - t^2g\right\|^2 \leqslant \frac{2}{2 - \beta}\left(F(x_{n+1}) - F^* - \frac{\beta - 1}{2}\left\|u - t^2g\right\|^2\right). \tag{152}$$

We deduce that for all $t \in [0, t_1]$,

$$\left(1 + \frac{\delta}{\lambda} t\right) L_{n+1} - L_n \leqslant t^2 K_3 L_{n+1} \tag{153}$$

with $K_3 = \max_{t \in [0,t_1]} \max\left(2P_3(t), \frac{2}{2-\beta}\left(P_3(t) + \frac{1}{2}|2tP_3(t) + K_2|\right)\right)$. We can conclude that for any $t \in [0, t_1]$ and any $n \geqslant 0$,

$$L_n \leqslant \left(1 + \frac{\delta}{\lambda} t - K_3 t^2\right)^{-n} L_0. \tag{154}$$

6.6 Proof of Theorem 9

The proof is essentially similar to the one of Theorem 7. The careful reader may have remarked that the only property of g that is used in the proof of Theorem 7 is the inequality (134) we recall here

$$\forall y \in \mathbb{R}^n, \quad F(x_{n+1}) - F(y) \leq \lambda \langle g, x_{n+\frac{1}{2}} - y \rangle - \frac{t^2}{2} \|g\|^2.$$

It is used twice, once with $y = x_n$ and once with $y = x^*$. Actually, any vector g satisfying this descent property will ensure the decay described in both theorems. It turns out that the vector $\tilde{g}$ defined in (52) satisfies this inequality under the hypothesis of the Theorem 9, see also [41, Lemma 4.2]:

Lemma 8. *If $F = f + h$ is convex, if f is convex differentiable with L-Lipschitz gradient, if h is convex, proper and lower semicontinuous and $s = \frac{1}{\sqrt{L}}$ then for all $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$:*

$$F(Tx) - F(y) \leq \frac{1}{s^2} \langle x - Tx, x - y \rangle - \frac{1}{2s^2} \|Tx - x\|^2. \tag{155}$$

Proof. Since $Tx = \text{prox}_{s^2 h}\left(x - s^2 \nabla f(x)\right)$, we have $x - s^2 \nabla f(x) - Tx \in s^2 \partial h(Tx)$ that is for any $y \in \mathbb{R}^n$:

$$h(Tx) - h(y) \leqslant \langle \frac{x - Tx}{s^2} - \nabla f(x), Tx - y \rangle \tag{156}$$

Since ∇f is $\frac{1}{s^2}$-Lipschitz

$$f(Tx) - f(x) \leqslant \langle \nabla f(x), Tx - x \rangle + \frac{1}{2s^2} \|Tx - x\|^2. \tag{157}$$

Since f is convex, for all $y \in \mathbb{R}^n$

$$f(x) - f(y) \leqslant \langle \nabla f(x), x - y \rangle. \tag{158}$$

Adding the three last inequalities we get:

$$F(Tx) - F(y) \leqslant \frac{1}{s^2} \langle x - Tx, Tx - y \rangle + \frac{1}{2s^2} \|Tx - x\|^2. \tag{159}$$

Using $Tx - y = Tx - x + x - y$ we get

$$F(Tx) - F(y) \leqslant \frac{1}{s^2} \langle x - Tx, x - y \rangle - \frac{1}{2s^2} \|Tx - x\|^2. \tag{160}$$

□

Applying this Lemma to $x = x_{n+\frac{1}{2}}$ we have $Tx = x_{n+1}$ and using $\tilde{g} := \frac{\lambda}{t^2}(x - Tx)$ we get exactly the inequality needed to complete the proof of Theorem 9:

$$\forall y \in \mathbb{R}^n, \quad F(x_{n+1}) - F(y) \leq \lambda \langle \tilde{g}, x_{n+\frac{1}{2}} - y \rangle - \frac{t^2}{2} \|\tilde{g}\|^2. \tag{161}$$

Acknowledgments

J.-F. Aujol acknowledges the support of the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No777826.

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-PRC-CE23 Masdol and the support of FMJH Program PGMO 2019-0024 and from the support to this program from EDF-Thales-Orange.

References

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

[2] V. Apidopoulos, J-F. Aujol, and Ch. Dossal. On a second order differential inclusion modeling the FISTA algorithm. working paper or preprint, May 2017.

[3] V. Apidopoulos, J.-F. Aujol, and Ch. Dossal. The differential inclusion modeling FISTA algorithm and optimality of convergence rate in the case $b \leqslant 3$. *SIAM Journal on Optimization*, 28(1):551–574, 2018.

[4] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1):5–16, 2009.

[5] H. Attouch and A. Cabot. Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity. *Journal of Differential Equations*, 263(9):5412–5458, 2017.

[6] H. Attouch and A. Cabot. Convergence rates of inertial forward-backward algorithms. *SIAM Journal on Optimization*, 28(1):849–874, 2018.

[7] H. Attouch, A. Cabot, and P. Redont. The dynamics of elastic shocks via epigraphical regularization of a differential inclusion. Barrier and penalty approximations. *Advances in Mathematical Sciences and Applications*, 12(1):273–306, 2002.

[8] H. Attouch and Z. Chbani. Fast inertial dynamics and FISTA algorithms in convex optimization. Perturbation aspects. *arXiv preprint arXiv:1507.01367*, 2015.

[9] J.-F. Aujol. Some first-order algorithms for total variation based image restoration. *Journal of Mathematical Imaging and Vision*, 34(3):307–327, 2009.

[10] J-F. Aujol and Ch. Dossal. Optimal rate of convergence of an ODE associated to the fast gradient descent schemes for $b > 0$. *Hal Preprint hal-01547251*, June 2017.

[11] J.-F. Aujol, Ch. Dossal, and A. Rondepierre. Optimal convergence rates for Nesterov acceleration. *SIAM Journal on Optimization*, 29(4):3131–3153, 2019.

[12] J.-F. Aujol, Ch. Dossal, and A. Rondepierre. Convergence rates of the Heavy-Ball method for quasi-strongly convex optimization. preprint, April 2020.

[13] M. Balti and R. May. Asymptotic for the perturbed heavy ball system with vanishing damping term. *Evolution Equations & Control Theory*, 6(2):177–186, 2017.

[14] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[15] P. Bégout, J. Bolte, and M. A. Jendoubi. On damped second order gradients systems. *Journal of Differential Equation*, 259(9):3315–3143, 2015.

[16] J. Bolte, A. Daniilidis, and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[17] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

[18] J. Bolte, T.P. Nguyen, J. Peypouquet, and B.W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

[19] J. Bolte, T.P. Nguyen, J. Peypouquet, and B.W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Math. Program.*, 165(2):471–507, October 2017.

[20] H. Brezis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, volume 5. Elsevier, 1973.

[21] A. Cabot, H. Engler, and S. Gadat. On the long time behavior of second order differential equations with asymptotically small dissipation. *Transactions of the American Mathematical Society*, 361(11):5983–6017, 2009.

[22] A. Cabot and L. Paoli. Asymptotics for some vibro-impact problems with a linear dissipation term. *Journal de Mathématiques Pures et Appliquées*, 87(3):291–323, 2007.

[23] G. Garrigos, L. Rosasco, and S. Villa. Convergence of the forward-backward algorithm: Beyond the worst case with the help of geometry. *arXiv preprint arXiv:1703.09477*, 2017.

[24] E. Ghadimi, H.R. Feyzmahdavian, and M. Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European Control Conference (ECC)*, pages 310–315. IEEE, 2015.

[25] A. Haraux and M.A. Jendoubi. On a second order dissipative ODE in Hilbert space with an integrable source term. *Acta Mathematica Scientia*, 32(1):155 – 163, 2012. Mathematics Dedicated to professor Constantine M. Dafermos on the occasion of his 70th birthday.

[26] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of convex analysis.* Springer Science & Business Media, 2012.

[27] G. Peyré J. Liang, M.J. Fadili. Activity identification and local linear convergence of forward–backward-type methods. *SIAM Journal on Optimization*, 27(1), 2017.

[28] M. A. Jendoubi and R. May. Asymptotics for a second-order differential equation with nonautonomous damping and an integrable source term. *Applicable Analysis*, 94(2):435–443, 2015.

[29] S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles (Paris, 1962)*, pages 87–89. Éditions du Centre National de la Recherche Scientifique, Paris, 1963.

[30] S. Łojasiewicz. Sur la géométrie semi- et sous-analytique. *Annales de l'Institut Fourier. Université de Grenoble*, 43(5):1575–1595, 1993.

[31] I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1-2):69–107, 2019.

[32] Y. Nesterov. A method of solving a convex programming problem with convergence rate $o(\frac{1}{k^2})$. In *Soviet Mathematics Doklady*, volume 27(2), pages 372–376, 1983.

[33] Y. Nesterov. Gradient methods for minimizing composite objective function. core discussion papers 2007076, université catholique de louvain. *Center for Operations Research and Econometrics (CORE)*, 1:4–4, 2007.

[34] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[35] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[36] B. O'Donoghue and E. Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

[37] L.A. Paoli. An existence result for vibrations with unilateral constraints: case of a nonsmooth set of constraints. *Mathematical Models and Methods in Applied Sciences*, 10(06):815–831, 2000.

[38] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[39] B.T. Polyak and P. Shcherbakov. Lyapunov functions: An optimization theory perspective. *IFAC-PapersOnLine*, 50(1):7456–7461, 2017.

[40] O. Sebbouh, Ch. Dossal, and A. Rondepierre. Nesterov's acceleration and Polyak's heavy ball method in continuous time: convergence rate analysis under geometric conditions and perturbations. *Hal Preprint hal-02173978*, Jul 2019.

[41] J.W. Siegel. Accelerated first-order methods: Differential equations and Lyapunov functions. *arXiv preprint arXiv:1903.05671*, 2019.

[42] 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):1–43, 2016.

[43] R.J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of statistics*, 7:1456–1490, 2013.

[44] B. Van Scoy, R.A. Freeman, and K.M. Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1):49–54, 2017.