Convergence rates of the Heavy-Ball method for quasi-strongly convex optimization

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

April 16, 2020

Abstract

In this paper, we study the behavior of solutions of the ODE associated to the Heavy Ball method. Since the pioneering work of B.T. Polyak [25], 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 the C^2 assumption is dropped. Depending on the geometry of the function to minimize, we obtain optimal convergence rates for the class of convex functions with some additional regularity such as quasi-strong convexity or strong convexity. We perform this analysis in continuous time for the ODE, and then we transpose these results for discrete optimization schemes. In particular, we propose a variant of the Heavy Ball algorithm which has the best state of the art convergence rate for first order methods to minimize strongly, composite non smooth convex functions.

Key-words Lyapunov function, rate of convergence, ODEs, optimization, strong convexity, Heavy Ball method.

1 Introduction

Let $F : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function admitting at least one minimizer. Let: $F^* = \inf F$ and $X^* = \operatorname{argmin} F$. In this paper, we are interested in the class of unconstrained optimization problems:

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

In many application fields like image processing, data science or deep learning among many others, there is a need for efficient optimization techniques. Due to the large dimension of the data, it is not possible to resort to second order information (e.g. the Hessian matrix as in Newton's method). This is the reason why first order methods are used, and there is therefore a need for developing accelerated first order methods.

Since the seminal work by B.T. Polyak [25] in 1964, the Heavy Ball algorithm is one of the main accelerated algorithm for minimizing C^2 strongly convex functions with Lipschitz gradient. From a mechanical point of view, the Heavy Ball system in continuous time corresponds to the ordinary differential equation (ODE) describing the motion of a body in the potential field F subject to a viscous friction force:

$$\ddot{x}(t) + \alpha\dot{x}(t) + \nabla F(x(t)) = 0 \tag{2}$$

where $\alpha > 0$. The physical intuition is that, as time evolves, the trajectory $x(t)$ of the body will reach a minimum of the potential F and solve the optimization problem (1) while reducing the oscillations by benefiting from friction.

In [25] Polyak proved that if F belongs to the class $\mathcal{S}^{2,1}_{\mu,L}$ of μ-strongly convex functions of class C^2 admitting a L-Lipschitz gradient, the solution of the ODE (2) converges linearly to the minimizer of F. Observe that under the same hypothesis on F the solution of the Gradient Flow:

$$\dot{x}(t) + \nabla F(x(t)) = 0 \tag{3}$$

converges linearly to the minimizer of F. It turns that this linear convergence occurs under much weaker hypotheses. In [11] Begout et al. proved that if F satisfies a Łojasiewicz property with exponent $\theta = \frac{1}{2}$, a linear decay is achieved similarly to the gradient flow. These linear decays have their analogous in the optimization setting since the discrete Heavy Ball algorithm and the Gradient descent algorithm ensure a linear decay of $F(x_n) - F^*$ when $F \in \mathcal{S}^{2,1}_{\mu,L}$ or when it belongs to more larger sets. Thus, some natural questions arise:

1. Are all these linear decays similar ?
2. Is there any benefit in using an inertial algorithm for functions for which the Gradient Descend is already linear ?
3. If the decay is linear, is the convergence really fast in practice ?
4. Can we give more accurate bounds ?

The response to the first question is simple: no, all these linear decays are not similar. Indeed, if $F \in \mathcal{S}^{2,1}_{\mu,L}$ and if x is a solution of the Heavy Ball ODE (2) for $\alpha = 2\sqrt{\mu}$, we have $F(x(t)) - F(x) = \mathcal{O}(e^{-2\sqrt{\mu}t})$ [25, Theorem 9], whereas $F(x(t)) - F(x) = \mathcal{O}(e^{-\mu t})$ if x is a solution of the Gradient Flow equation (3). And we can easily prove that these rate are achieved for quadratic functions. This remark gives an answer to the second question: if μ is very small, which is the case in many large scale problems, the inertia of the Heavy Ball method ensures a much better convergence rate. We will see that this square root also appears in the algorithm and it may explain the various practical behaviors of algorithms that are all linear.

In large scale problems, μ may be so small that the linear decay may not be visible. In many image processing problems, or statistical problems, one can observe that the convergence is very slow and FISTA [10] is better. This slowness is due to the smallness of μ. A typical example is the linear convergence of the Forward-Backward algorithm applied to the LASSO problem i.e. when $F(x) = \frac{1}{2}\left\|Ax - b\right\|^2 + \lambda\left\|x\right\|_1$. We refer the reader to Section 3.3 for more details.

The main contribution of this paper is to provide answers to the last question : yes, we can give more accurate bounds depending on the geometrical assumptions on F. More precisely most of these decays are exponential of the form $\mathcal{O}(e^{\delta\sqrt{\mu}t})$. It turns out that the value of δ highly depends on the precise geometrical hypotheses made on F and it is different if F is quadratic or only satisfies a Łojasiewicz property with parameter $\theta = \frac{1}{2}$.

In this paper, we focus on the class of convex functions being strongly convex, or quasi-strongly convex [20] which is a relaxation of strong convexity. A differentiable function F is said μ-quasi-strongly convex if:

$$\forall x \in \mathbb{R}^n, \langle \nabla F(x), x - x^* \rangle \geqslant F(x) - F(x^*) + \frac{\mu}{2}\|x - x^*\|^2 \tag{4}$$

where x^* denotes the projection of x onto the set of minimizers.

In the literature, several hypotheses have been proposed: being quadratic ($\mathcal{Q}_\mu$), belonging to the class $\mathcal{S}^{2,1}_{\mu,L}$ (resp. $\mathcal{S}^{1,1}_{\mu,L}$) of μ-strongly convex functions of class C^2 (resp. C^1) admitting a L-Lipschitz gradient, being just μ-strongly convex $\mathcal{S}_\mu$, μ-quasi-strongly convex $q\mathcal{S}_\mu$, satisfying the Polyak-Łojasiewicz ($\mathcal{PL}_\mu$) property with some constant $\mu > 0$. The exact definition of all these properties will be given in Section 2, but we draw the reader's attention to the fact that all these conditions are not equivalent and and that they characterize sub-classes of functions satisfying the Polyak-Łojasiewicz property. More precisely, if F has a unique minimizer we have:

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

Let us mention that quasi-strongly convex functions are not necessarily convex.

In this paper, we investigate both the continuous and the discrete case. We provide some convergence rates for the values $F(x(t)) - F^*$ along the trajectories of the heavy ball ODE (2) for functions F having a unique minimizer and being quasi-strongly convex. In a second time we extend these results to include perturbations and in a third time to the associated monotone inclusion to deal with non differentiable functions and especially composite functions. Finally we apply these results to provide a new optimization algorithm whose decay rate for the class of strongly convex functions is better than the state of art.

Let us now summarize from the literature and the present paper some of the main results of decay rates for the values $F(x(t)) - F^*$ along the solutions of (2) in the following table:

Hypotheses on F	References	values of α	Exponential rate of $F(x(t)) - F^*$
$\mathcal{S}^{2,1}_{\mu}$	Polyak [25]	$(0, 2\sqrt{\mu})$	α
$\mathcal{S}^{2,1}_{\mu}$	Polyak [25]	$(2\sqrt{\mu}, +\infty)$	$\alpha - \sqrt{\alpha^2 - 4\mu}$
$\mathcal{S}^{1,1}_{\mu}$	Siegel [29]	$2\sqrt{\mu}$	$\sqrt{\mu}$
$q\mathcal{S}_\mu$ and uniqueness of the minimizer	ADR, Th. 1	$(0, 3\sqrt{\frac{\mu}{2}}]$	$\frac{2\alpha}{3}$
$q\mathcal{S}_\mu$ and uniqueness of the minimizer	ADR Th. 1	$[3\sqrt{\frac{\mu}{2}}, +\infty)$	$\alpha - \sqrt{\alpha^2 - 4\mu}$
$q\mathcal{S}_{\mu,L}$ and uniqueness of the minimizer	ADR Th. 2	$(0, 3\sqrt{\frac{\mu}{2}}]$	$\frac{2}{3}\alpha\left(1 + \frac{2}{3}\frac{9\mu - 2\alpha^2}{9L + 3\mu - \frac{2}{3}\alpha^2}\right)$

For example, the third line of the table asserts that if $F \in \mathcal{S}^{1,1}_{\mu}$, then choosing $\alpha = 2\sqrt{\mu}$ in the ODE (2) ensures that $F(x(t)) - F^* = \mathcal{O}(e^{-\sqrt{\mu}t})$. The Figure 1 illustrates some of the results provided in this first table.

Similarly we can summarize some of the main results of decay rates of optimization algorithms related to the heavy ball method. These inertial algorithms will be described in Section 4. All of them ensure an exponential decay:

$$F(x_n) - F^* = \mathcal{O}(q^n) \tag{5}$$

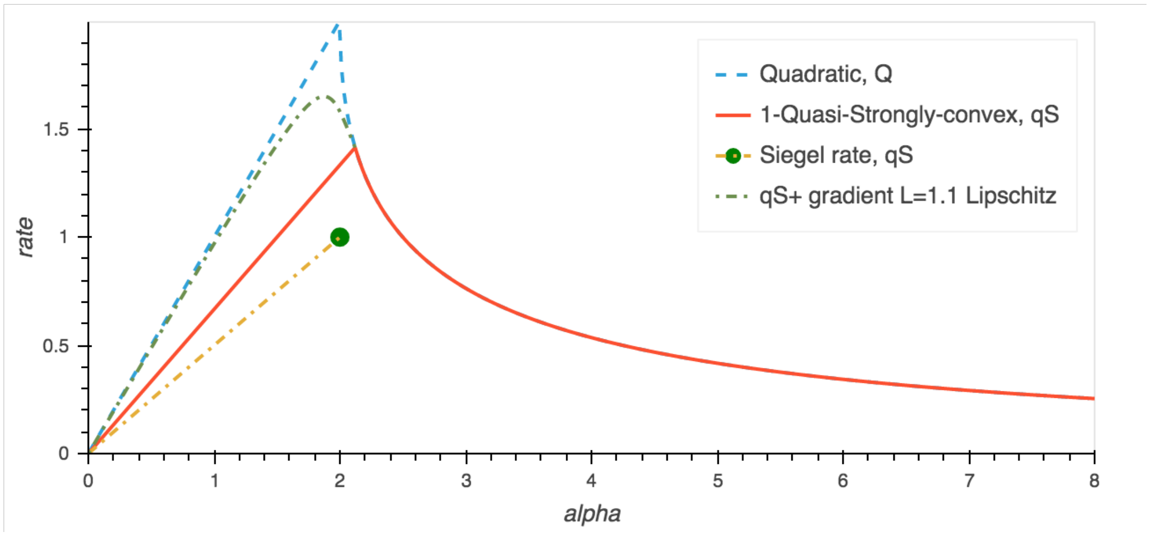

Figure 1: Decay rates that can be achieved depending on the geometrical hypotheses made on F for $\mu = 1$ the parameter of quasi-strong convexity. The dashed blue curve illustrates the results of Polyak [25], the green point illustrates the results of Siegel [29], the solid line the result given in Theorem 1 and the dashed-doted line the result of Theorem 2 for $L = 1.1$.

where the value of q depends on the condition number $\kappa = \frac{\mu}{L}$ and is different for each algorithm. Note that the Gradient Descend ensures a decay with $q = 1 - \kappa$ when $F \in \mathcal{S}^{2,1}_{\mu,L}$.

Hypothesis on F	References	Values of q	Remarks
$\mathcal{S}^{2,1}_{\mu,L}$	Polyak [25]	$\left(\frac{1-\sqrt{\kappa}}{1+\sqrt{k}}\right)^2$	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 [21]	$1 - \sqrt{\kappa} + \mathcal{O}(\kappa)$	Global convergence
$\mathcal{S}^{1,1}_{\mu,L}$	GFJ [14]	$1 - \kappa$	Global convergence
$\mathcal{S}^{2,1}_{\mu,L}$	SFL [30]	$1 - 2\sqrt{\kappa} + \mathcal{O}(\kappa)$	Global convergence, three points method
$\mathcal{S}^{1,1}_{\mu,L}$	Siegel [29]	$1 - \sqrt{\kappa}$	Global convergence, can be extended to non differentiable functions
$\mathcal{S}^{1,1}_{\mu,L}$	ADR, Th. 6	$1 - \sqrt{(2-\varepsilon)\kappa}$ $+\mathcal{O}(\kappa)$	Global convergence, can be extended to non differentiable functions

The contributions of the paper illustrated on the Figure 1 are the following: in the continuous setting, our first result is new non-asymptotic and optimal rates and global convergence for the class of quasi-strongly convex functions having a unique minimizer. We also prove that these rates can be improved if F additionally have a L-Lipschitz gradient. In particular we prove that the optimal parameter α in (2) is not $\alpha = 2\sqrt{\mu}$ for these sub-classes of functions. A second contribution is to propose stability results, namely integrability sufficient conditions on the perturbation in order to preserve the previous convergence rates. A third contribution is then to extend these results to the non differentiable case (namely to the monotone inclusion) and to prove the optimality of these exponential decays building functions for which these rates are reached. In the discrete setting, our main contribution is to provide a new optimization algorithm for minimizing functions of $\mathcal{S}^{1,1}_{\mu,L}$ with better convergence rate than the classical scheme of Nesterov built for $\mathcal{S}^{1,1}_{\mu,L}$. We finally extend this scheme to composite and non differentiable functions.

The paper is organized as follows. In Section 2, we recall the different geometric assumptions considered in this work for the function to minimize. Section 3 is then devoted to the analysis of the ODE associated to the Heavy Ball algorithm. We propose numerical schemes associated to this analysis in Section 4. We illustrate the results of the paper with numerical examples in Section 5. Eventually, we detail the proofs of the Theorems of Sections 3 and 4 in Section 6.

2 Preliminaries: relaxing strong convexity

Throughout the paper, we are interested in the class of unconstrained optimization problems:

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

where $F:\mathbb{R}^n\to\mathbb{R}$ is a convex function admitting at least one minimizer. 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:

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

In this paper we revisit the Heavy Ball method for the class of strongly and quasi-strongly convex functions. Let us first recall the definition of strong convexity:

Definition 1 (Strong convexity). *A function $F:\mathbb{R}^n\to\mathbb{R}$ is μ-strongly convex if and only if the function $F-\frac{\mu}{2}\|\cdot\|^2$ is convex. If F is differentiable, F is μ-strongly convex if and only if:*

$$\forall (x,y)\in\mathbb{R}^n\times\mathbb{R}^n,\ F(y)\geqslant F(x)+\langle\nabla F(x),y-x\rangle+\frac{\mu}{2}\|y-x\|^2.$$

We have a special interest in a more general class of quasi-strongly convex functions introduced by I. Necoara and al. in [20]:

Definition 2 (Quasi-strong convexity [20, 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^*.

We refer the reader to [20, Section 3] for a complete review of relations between several functional classes relaxing the strong convexity properties. Note that strongly convex functions are a subclass of the class of quasi-strongly convex functions. But the quasi-strong convexity does not imply the convexity of F and does not ensure the uniqueness of the minimizer.

The class of quasi-strongly convex functions is a subclass of functions having the Polyak-Łojasiewicz property, namely the Łojasiewicz property [18, 19] with an exponent equal to $\frac{1}{2}$. In the convex setting the Polyak-Łojasiewicz property is equivalent to a quadratic growth condition which is another relaxation of the strong convexity:

Lemma 1. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function with $X^* = \operatorname{argmin} F \neq \emptyset$ and $F^* = \inf\ F$. If F is μ-strongly-quasi convex then F also satisfies the Polyak-Łojasiewicz ($\mathcal{PL}_\mu$) property:*

$$\forall x \in \mathbb{R}^n, \|\nabla F(x)\|^2 \geqslant 2\mu(F(x) - F^*). \tag{6}$$

If in addition F is convex then F satisfies the growth condition $\mathcal{G}(2)$:

$$\forall x \in \mathbb{R}^n, F(x) - F^* \geqslant \frac{\mu}{2} d(x, X^*)^2. \tag{7}$$

The quadratic growth condition $\mathcal{G}(2)$ can be seen as a sharpness assumption ensuring that the magnitude of the gradient is not too low in the neighborhood of the minimizers, see [8] for more details. Roughly speaking, any function F satisfying $\mathcal{G}(2)$ is at least as sharp as $\|x\|^2$ in the neighborhood of its set of minimizers. Note that when F has a unique minimizer, the quadratic growth condition $\mathcal{G}(2)$ is exactly the characterisation of the notion of strong minimizer of F introduced in [2, Section 3.3].

Finally observe that any differentiable convex function with a Lipschitz continuous gradient and satisfying a quadratic growth condition, is quasi-strongly convex:

Lemma 2. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a convex continuously differentiable function with $X^* = \operatorname{argmin} F \neq \emptyset$ and $F^* = \inf\ F$. If F has a L-Lipschitz continuous gradient for some $L > 0$ and satisfies the growth condition $\mathcal{G}(2)$ then F is $\frac{\mu^2}{L}$-quasi-strongly convex.*

Proof. Using the assumption that F has a Lipschitz continuous gradient, the Polyak-Łojasiewicz inequality (6) and then the quadratic growth condition, we have:

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

□

3 The continuous case

Let $F : \mathbb{R}^n \to \mathbb{R}$ be a convex continuously differentiable function admitting at least one minimizer. In this section we study the convergence rates in finite time for the values $F(x(t)) - F^*$ along the trajectories of the perturbed second-order ordinary differential equation:

$$\ddot{x}(t) + \alpha \dot{x}(t) + \nabla F(x(t)) = g(t) \tag{8}$$

for any $t \geqslant t_0$, 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, for any given initial conditions

$(x_0, v_0) \in \mathbb{R}^n \times \mathbb{R}^n$, the Cauchy problem associated with the ODE (8), 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 $\mathbb{R}^n$ [15, 16].

In his seminal work [25], B.T. Polyak gives a global convergence rate for strongly convex functions of class C^2 with a Lipschitz continuous gradient. In [11], P. Bégout, J. Bolte and M.A. Jendoubi prove the strong convergence of the trajectory at an exponential decay (depending on the Łojasiewicz exponent) provided that F is real-analytic. More recently B.T. Polyak and P. Shcherbakov prove the exponential decay of the values of $F(x(t)) - F^*$ for the class of C^2 functions satisfying the Polyak-Łojasiewicz inequality (6) without assuming the uniqueness of the minimizer in [26].

Let us recall that the exponential decay of solutions of the Gradient Descend flow (3) for functions in $(\mathcal{PL}_\mu)$ is straightforward. Indeed defining $\mathcal{E}(t) = F(x(t)) - F(x^*)$, we get

$$\mathcal{E}'(t) = -\left\|\nabla F(x(t))\right\|^2 \leqslant -\mu(F(x(t)) - F(x^*)) \tag{9}$$

which ensures that

$$F(x(t)) - F(x^*) \leqslant (F(x(t_0)) - F(x^*))\, e^{-\mu(t-t_0)}. \tag{10}$$

We can thus observe that the exponent is proportional to μ and not $\sqrt{\mu}$.

Proving convergence rates for the Heavy Ball method in the case of an objective function which is not C^2 is an active field of research. Without the C^2 assumption, for the class of differentiable strongly convex functions, a suitable choice of the friction parameter α provides an exponential decay of the values $F(x(t)) - F^*$ [5, 29]. More precisely, if F is assumed μ-strongly convex differentiable, J.W. Siegel proves in [29] that for $\alpha = 2\sqrt{\mu}$:

$$F(x(t)) - F^* \leqslant 2\,(F(x_0) - F^*)\, e^{-\sqrt{\mu}t} \tag{11}$$

where $x(\cdot)$ is solution of the ODE (8) with $g = 0$ and $(x(0), \dot{x}(0)) = (x_0, 0)$. Applying [27, Theorem 3] with $\alpha = \delta\sqrt{\mu}$ for δ large enough, an exponential decay with an exponent proportional to $\sqrt{\mu}$ has also be obtained in [27] for the class of convex differentiable functions satisfying the Polyak-Łojasiewicz property. Let us also mention the work on higher order ODEs by Shi et al. [28] where under the same hypotheses as [29], the authors get a convergence rate of the order $\mathcal{O}\left(e^{-\frac{\sqrt{\mu}}{4}t}\right)$ (which is of course worse).

The goal of this section is to provide non-asymptotic convergence rates for the values $F(x(t)) - F^*$ that can be achieved for the class of quasi-strongly convex functions and the sub-class of quasi-strongly convex functions having a Lipschitz continuous gradient. In particular, we provide optimal convergence rates for the class of differentiable quasi-strongly convex functions.

3.1 Convergence rates in the unperturbed case

Let us first consider the unperturbed ODE:

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

where $F : \mathbb{R}^n \to \mathbb{R}$ is a differentiable convex function admitting a unique minimizer x^*. Assume that F is additionally quasi-strongly convex. Based on a Lyapunov analysis, the aim of this paragraph is to establish non-asymptotic convergence rates for the values $F(x(t)) - F^*$ along the trajectory $x(t)$ solution of (12). The following Lyapunov energy plays a central role in our whole analysis:

$$\mathcal{E}_{\lambda,\xi}(t) = F(x(t)) - F^* + \frac{1}{2}\left\|\lambda(x(t)-x^*) + \dot{x}(t)\right\|^2 + \frac{\xi}{2}\left\|x(t)-x^*\right\|^2 \tag{13}$$

where λ and ξ are two real parameters. To establish non-asymptotic results, we will need the following technical lemma:

Lemma 3. *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 $(x(t_0), \dot{x}(t_0)) = (x_0, v_0)$ and $M_0 = F(x_0) - F^* + \frac{1}{2}\|v_0\|^2$. Then:*

$$\forall t \geqslant t_0,\ F(x(t)) - F^* \leqslant M_0, \quad \|\dot{x}(t)\|^2 \leqslant 2M_0.$$

Assume in addition that F satisfies the growth condition $\mathcal{G}(2)$ for some constant $\mu > 0$. Then for any minimizer $x^* \in X^*$, for all $\lambda \geqslant 0$ and for all $\xi \leqslant 0$, we have:

$$\forall t \geqslant t_0,\ \mathcal{E}_{\lambda,\xi}(t) \leqslant \left(1 + \left(\frac{\lambda}{\sqrt{\mu}} + 1\right)^2\right) M_0.$$

Proof. Introducing the Lyapunov energy: $W(t) = F(x(t)) - F^* + \frac{1}{2}\|\dot{x}(t)\|^2$ and using the ODE (12), we easily prove:

$$W'(t) = \langle \nabla F(x(t)), \dot{x}(t)\rangle + \langle \ddot{x}(t), \dot{x}(t)\rangle = -\alpha\|\dot{x}(t)\|^2 \leqslant 0.$$

The energy W is so non increasing: $\forall t \geqslant t_0, W(t) \leqslant W(t_0)$. We then deduce:

$$\forall t \geqslant t_0,\ F(x(t)) - F^* \leqslant M_0, \quad \|\dot{x}(t)\|^2 \leqslant 2M_0.$$

where: $M_0 = W(t_0) = F(x_0) - F^* + \frac{1}{2}\|\dot{v}_0\|^2 > 0$. Assuming that F satisfies the growth condition $\mathcal{G}(2)$ with a constant $\mu > 0$, we get:

$$\begin{aligned} \|\lambda(x(t)-x^*) + \dot{x}(t)\|^2 &\leqslant \lambda^2\|x(t)-x^*\|^2 + \|\dot{x}(t)\|^2 + 2\lambda\|x(t)-x^*\|\|\dot{x}(t)\| \\ &\leqslant \left(\frac{2}{\mu}\lambda^2 + 2 + 2\lambda\sqrt{\frac{2}{\mu}}\sqrt{2}\right) M_0 = 2\left(\frac{\lambda}{\sqrt{\mu}} + 1\right)^2 M_0. \end{aligned}$$

Since $\xi \leqslant 0$ we finally get the expected inequality. ∎

Our first result provides a non-asymptotic convergence rate for the values $F(x(t)) - F^*$ for the class of quasi-strongly convex functions.

Theorem 1. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable 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 $M_0 = F(x_0) - F^* + \frac{1}{2}\|v_0\|^2$.*

Assume that F is μ-quasi-strongly convex for some $\mu > 0$ and admits a unique minimizer x^*.

- *If $\alpha \leqslant 3\sqrt{\frac{\mu}{2}}$ then:*

$$\forall t \geqslant t_0,\ F(x(t)) - F^* \leqslant M_0 \frac{9\mu + (2\alpha + 3\sqrt{\mu})^2}{3\mu} e^{-\frac{2\alpha}{3}(t-t_0)}.$$

- *If $\alpha > 3\sqrt{\frac{\mu}{2}}$ then:*

$$\forall t \geqslant t_0,\ F(x(t)) - F^* \leqslant M_0 \frac{\mu + \left(\alpha + \sqrt{\mu} - \sqrt{\alpha^2 - 4\mu}\right)^2}{3\mu - \alpha^2 + \alpha\sqrt{\alpha^2 - 4\mu}} e^{-(\alpha - \sqrt{\alpha^2 - 4\mu})(t - t_0)}.$$

This rate is optimal in the sense that it is achieved for the function $F(x) = \frac{\mu}{2}\|x\|^2$.

Let us focus on the optimality of the proposed decay rates. Consider the quadratic function $F(x) = \frac{\mu}{2}\|x\|^2$, $x \in \mathbb{R}^n$. By definition F is μ-strongly convex, therefore μ-quasi-strongly convex, and has a μ-Lipschitz gradient. In that case, the ODE (12) is a second order linear differential equation with constant coefficients whose solutions can be easily computed. A straightforward computation shows that the convergence rates on the values $F(x(t)) - F^*$ are given by:

$$F(x(t)) - F^* = \begin{cases} \mathcal{O}\left(e^{-\alpha t}\right) & \text{if } \alpha \leqslant 2\sqrt{\mu} \\ \mathcal{O}\left(e^{-(\alpha - \sqrt{\alpha^2 - 4\mu})t}\right) & \text{if } \alpha > 2\sqrt{\mu}. \end{cases} \tag{14}$$

We so conclude that the convergence rate with the exponent $\alpha - \sqrt{\alpha^2 - 4\mu}$ given in Theorem 1 when $\alpha > 3\sqrt{\frac{\mu}{2}}$, is optimal for the class of quasi-strongly convex functions. On the other hand, this suggests that the convergence rates given in Theorem 1 when $\alpha < 3\sqrt{\frac{\mu}{2}}$ may be improved at least for the sub-class of quasi-strongly convex functions having a Lipschitz continuous gradient as investigated hereafter.

In fact, we can prove that the exponential rate $\frac{2\alpha}{3}$ is actually optimal for the class $\mathcal{S}_\mu$ of strongly convex functions. It also proves that for such functions, $\alpha = 2\sqrt{\mu}$ is not the parameter that ensures the best decay rate.

Proposition 1. *Let $\mu > 0$ and $\alpha < 3\sqrt{\frac{\mu}{2}}$. The exponent $\frac{2\alpha}{3}$ in the exponential rate is optimal on the class $\mathcal{S}_\mu$ of strongly convex functions in the sense that for any $\delta \in (\frac{2\alpha}{3}, \frac{4\alpha}{3})$ and any $r \in (1, \frac{3\delta}{2\alpha})$, if $F(x) = |x|^r + \frac{\mu}{2}|x|^2$, the solution x of the ODE (12) satisfies*

$$\limsup_{t > t_0} \left(F(x(t)) - F^*\right) e^{\delta t} > 0.$$

Proof. Let us consider the Lyapunov energy $\mathcal{E}_{\lambda,\xi}$ defined by (13). Differentiating $\mathcal{E}_{\lambda,\xi}$ as done in the proof of Theorem 1 (see Subsection 6.1) with $\lambda = \frac{2\alpha}{3}$ and $\xi = -\frac{2\alpha^2}{9}$ we have the following equality:

$$\forall t \geqslant t_0,\ \mathcal{E}'(t) + \frac{2\alpha}{3}\mathcal{E}(t) = \frac{2\alpha^3}{27}\|x(t) - x^*\|^2 + \frac{2\alpha}{3}\left(F(x(t)) - F^* - \langle \nabla F(x(t)), x(t) - x^* \rangle\right). \tag{15}$$

Observe now that for $F(x) = |x|^r + \frac{\mu}{2}|x|^2$, we have: $F'(x)(x - x^*) = r|x|^r + \mu x$. Thus applying (15) we get for this specific function:

$$\forall t \geqslant t_0,\ \mathcal{E}'(t) + \frac{2\alpha}{3}\mathcal{E}(t) = \frac{2\alpha^3}{27}|x(t)|^2 + \frac{2\alpha}{3}\left((1 - r)|x(t)|^r - \frac{\mu}{2}|x(t)|^2\right).$$

It then follows that for any $\delta \in (\frac{2\alpha}{3}, \frac{4\alpha}{3})$

$$\forall t \geqslant t_0,\ \mathcal{E}'(t) + \delta\mathcal{E}(t) \geqslant \left(\frac{2\alpha^3}{27} - \frac{\alpha\mu}{3} - \frac{\alpha^2}{9}(\delta - \frac{2\alpha}{3})\right)|x(t)|^2 + (\delta - \frac{2\alpha r}{3})|x(t)|^r.$$

If $r \in (1, \frac{3\delta}{2\alpha})$ then the right member is non negative for t sufficiently large. It turns out that there exits t_1 and $K > 0$ such that

$$\forall t \geqslant t_1, \quad \mathcal{E}(t) \geqslant K e^{-\delta t}.$$

We conclude following the proof of Theorem 5 (see Subsection 6.3) by showing that $y(t) = e^{\delta t}|x(t)|^r$ cannot tend to 0 when t tends to $+\infty$. □

To conclude this section, we finally show that the convergence rates for the values $F(x(t)) - F^*$ obtained for the general class of quasi-strongly convex functions in Theorem 1, can be improved in the case $\alpha < 3\sqrt{\frac{\mu}{2}}$ if more information about the geometry of F is available.

Assume that F additionally has a L-Lipschitz continuous gradient. It turns out that the Lyapunov energy (13) does not allow to choose $\lambda > \frac{2}{3}\alpha$ because the term $\left(\frac{3\lambda}{2} - \alpha\right)\|\dot{x}(t)\|^2$ in its derivative would be non negative. We propose to add the following mechanic energy to the previous one:

$$\mathcal{E}_m(t) = F(x(t)) - F(x^*) + \frac{1}{2}\|\dot{x}(t)\|^2. \tag{16}$$

Let $\beta \geqslant 0$. We consider the energy $\mathcal{E}(t) = \mathcal{E}_{\lambda,\xi}(t) + \beta\mathcal{E}_m(t)$ defined by

$$\mathcal{E}(t) = (1+\beta)(F(x(t)) - F^*) + \frac{1}{2}\|\lambda(x(t) - x^*) + \dot{x}(t)\|^2 + \frac{\xi}{2}\|x(t) - x^*\|^2 + \frac{\beta}{2}\|\dot{x}(t)\|^2. \tag{17}$$

Using Lemma 3, a straightforward computation enables to show that the new energy $\mathcal{E}$ is uniformly bounded on the time interval $[t_0, +\infty[$. More precisely, if F satisfies a global growth condition $\mathcal{G}(2)$ for $\mu > 0$ then for any minimizer $x^* \in X^*$, for all $\lambda \geqslant 0$ and for all $\xi \leqslant 0$,

$$\forall t \geqslant t_0,\ \mathcal{E}(t) \leqslant \left(1 + 2\beta + \left(\frac{\lambda}{\sqrt{\mu}} + 1\right)^2\right) M_0 \tag{18}$$

where: $M_0 = F(x_0) - F^* + \frac{1}{2}\|v_0\|^2$. We then have:

Theorem 2. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable 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 $(x(t_0), \dot{x}(t_0)) = (x_0, v_0)$.*

Assume that F is μ-quasi-strongly convex for some $\mu > 0$ with a L-Lipschitz gradient and admits a unique minimizer x^*. Let $M_0 = F(x_0) - F^* + \frac{1}{2}\|v_0\|^2$. If $\alpha \leqslant 3\sqrt{\frac{\mu}{2}}$ then:

$$\forall t \geqslant t_0,\ F(x(t)) - F^* \leqslant \left[1 + 2\frac{1+\beta}{(1-\beta)^2}\right] \frac{\mu(1+2\beta) + (\sqrt{\mu} + \lambda)^2}{\mu(1+\beta)} M_0 e^{-\lambda(t-t_0)}$$

with: $\beta = \frac{\mu}{L} - \frac{2\alpha^2}{9L}$ and $\lambda = 2\alpha\frac{1+\beta}{3+\beta} = \frac{2}{3}\alpha\left(1 + \frac{2}{3}\frac{9\mu - 2\alpha^2}{9L + 3\mu - \frac{2}{3}\alpha^2}\right)$.

Note that the Theorem 2 only provides an upper bound on the actual convergence rate for the values $F(x(t)) - F^*$ since the value $\beta = \frac{\mu}{L} - \frac{2\alpha^2}{9L}$ is actually a lower bound of the value β^* ensuring the best convergence rate. Following the proof of Theorem 2 detailed in Section 6, the theoretical value β^* satisfies:

$$\frac{\mu}{L} - \frac{2\alpha^2}{9L} \leqslant \beta^* < \frac{\mu}{L}$$

and can be numerically evaluated as the smallest root of the polynomial $\beta \mapsto 2\alpha^2(1-\beta^2) - (3+\beta)^2(\mu - \beta L)$ inside the open interval $[0, \frac{\mu}{L})$. In the case when $\alpha = 3\sqrt{\frac{\mu}{2}}$ then Theorem 2 applies with $\beta = 0$ and we find exactly the control provided by Theorem 1 in $\mathcal{O}(e^{-\frac{2\alpha}{3}t})$.

3.2 Convergence analysis under perturbations

In this section we extend our convergence analysis to the solutions of the perturbed differential equation:

$$\ddot{x}(t) + \alpha\dot{x}(t) + \nabla F(x(t)) = g(t). \tag{19}$$

Our main contribution is to provide integrability sufficient conditions on the perturbation g in order to guarantee that the convergence properties previously established are preserved. All our analysis is based on the same Lyapunov energy as in the unperturbed case:

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

where λ and ξ are two real parameters and x^* denotes a minimizer of F. To deal with the perturbation term, we choose to add an integral term in the energy $\mathcal{E}$ as done in [4, 9, 7, 27] and the references therein:

$$\mathcal{G}(t) = \mathcal{E}(t) + \int_t^T \langle \lambda(x(s)-x^*) + (1+\beta)\dot{x}(s)), g(s)\rangle ds. \tag{20}$$

As previously done all the results stated in this section are non-asymptotic and based on the following lemma extending Lemma 3 to the perturbed case:

Lemma 4. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable 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 perturbed ODE (19) for given initial conditions $(x(t_0), \dot{x}(t_0)) = (x_0, v_0)$.*

Assume that $\int_{t_0}^{+\infty} \|g(s)\|ds < +\infty$. Then:

$$\begin{array}{llll} \forall t \geqslant t_0, & \|\dot{x}(t)\| & \leqslant & \sqrt{2M_0} + I_0 \\ & F(x(t)) - F^* & \leqslant & M_0 + (\sqrt{2M_0} + I_0)I_0 \end{array}$$

where $M_0 = F(x_0) - F^* + \frac{1}{2}\|v_0\|^2$ and $I_0 = \int_{t_0}^{+\infty} \|g(s)\|ds < +\infty$.

If in addition F satisfies the growth condition $\mathcal{G}(2)$ for some $\mu > 0$, then for any minimizer $x^* \in X^*$, for all $\lambda \geqslant 0$ and $\xi \leqslant 0$, we have for all $t \geqslant t_0$

$$\mathcal{E}_{\lambda,\xi}(t) \leqslant M_0 + (\sqrt{2M_0} + I_0)I_0 + \left(\sqrt{M_0} + \frac{I_0}{\sqrt{2}} + \frac{\lambda}{\sqrt{\mu}}\sqrt{M_0 + (\sqrt{2M_0} + I_0)I_0}\right)^2.$$

Proof. Let $T > 0$. We introduce the following energy:

$$W(t) = F(x(t)) - F^* + \frac{1}{2}\|\dot{x}(t)\|^2 + \int_t^T \langle g(s), \dot{x}(s)\rangle ds.$$

Using the ODE (19), we easily prove: $W'(t) = -\alpha\|\dot{x}(t)\|^2 \leqslant 0$, so that the energy W is non increasing: $\forall t \geqslant t_0, W(t) \leqslant W(t_0)$. We then deduce:

$$\forall t \geqslant t_0,\ F(x(t)) - F^* + \frac{1}{2}\|\dot{x}(t)\|^2 \ \leqslant\ M_0 + \int_{t_0}^t \langle g(s), \dot{x}(s)\rangle ds \leqslant M_0 + \int_{t_0}^t \|g(s)\|\|\dot{x}(s)\|ds$$

where: $M_0 = W(t_0) = F(x_0) - F^* + \frac{1}{2}\|\dot{v}_0\|^2 > 0$. Hence:

$$\forall t \geqslant t_0,\ \frac{1}{2}\|\dot{x}(t)\|^2 \leqslant M_0 + \int_{t_0}^{t} \|g(s)\|\|\dot{x}(s)\|ds.$$

Let: $I_0 = \int_{t_0}^{+\infty} \|g(s)\|ds$. Applying the Grönwall-Bellman Lemma [12, Lemma A.5], we obtain:

$$\forall t \geqslant t_0,\ \|\dot{x}(t)\| \leqslant \sqrt{2M_0} + \int_{t_0}^{t} \|g(s)\|ds \leqslant \sqrt{2M_0} + I_0$$

Hence for all $t \geqslant t_0$, we have: $F(x(t)) - F^* \leqslant M_0 + (\sqrt{2M_0} + I_0)I_0$. Assuming now that F satisfies the growth condition $\mathcal{G}(2)$ with the constant μ, we have:

$$\begin{aligned} \|\lambda(x(t)-x^*) + \dot{x}(t)\|^2 &\leqslant \lambda^2\|x(t)-x^*\|^2 + \|\dot{x}(t)\|^2 + 2\lambda\|x(t)-x^*\|\|\dot{x}(t)\| \\ &\leqslant (\lambda\|x(t)-x^*\| + \|\dot{x}(t)\|)^2 \\ &\leqslant 2\left(\sqrt{M_0} + \frac{I_0}{\sqrt{2}} + \frac{\lambda}{\sqrt{\mu}}\sqrt{M_0 + (\sqrt{2M_0}+I_0)I_0}\right)^2. \end{aligned}$$

Since $\xi \leqslant 0$ we finally get the expected inequality. □

Assuming now some integrability conditions on the perturbation g, we prove that the exponential decays stated in the unperturbed case for the class of quasi-strongly convex functions and its subclass of quasi-strongly convex functions having a Lipschitz continuous gradient, are preserved.

Theorem 3. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable 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 perturbed ODE (19) for given initial conditions $(x(t_0), \dot{x}(t_0)) = (x_0, v_0)$. Let:*

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

Assume that F is μ-quasi-strongly convex for some $\mu > 0$ and admits a unique minimizer x^*.

- *Assume that $\alpha \leqslant 3\sqrt{\frac{\mu}{2}}$. If*

  $$J_0(\beta) = \int_{t_0}^{+\infty} e^{2\alpha\frac{1+\beta}{3+\beta}t}\|g(t)\|dt < +\infty$$

  then for all $t \geqslant t_0$:

  $$F(x(t)) - F^* \leqslant \frac{e^{\lambda t_0}E_0(\lambda) + (\sqrt{2(1+\beta)E_0(\lambda)} + (1+\beta)I_0)J_0(\beta)}{1+\beta}\left[1 + 2\frac{1+\beta}{(1-\beta)^2}\right]e^{-\lambda t}$$

  where $\lambda = 2\alpha\frac{1+\beta}{3+\beta}$ and $\beta = 0$. If F additionally has a L-Lipschitz continuous gradient then the decay rate can be improved by choosing:

  $$\beta = \frac{\mu}{L} - \frac{2\alpha^2}{9L} \quad \text{and} \quad \lambda = 2\alpha\frac{1+\beta}{3+\beta} = \frac{2}{3}\alpha\left(1 + \frac{2}{3}\frac{9\mu - 2\alpha^2}{9L + 3\mu - \frac{2}{3}\alpha^2}\right).$$

- *Assume that* $\alpha > 3\sqrt{\frac{\mu}{2}}$. *If:*

$$J_0 = \int_{t_0}^{+\infty} e^{(\alpha-\sqrt{\alpha^2-4\mu})t}\|g(t)\|dt < +\infty$$

then:

$$\forall t \geqslant t_0,\ F(x(t)) - F^* \leqslant 2\mu\frac{e^{\lambda t_0}E_0(\lambda) + (\sqrt{2E_0(\lambda)} + I_0)J_0}{2\mu - (\alpha - \sqrt{\alpha^2 - 4\mu})^2}e^{-(\alpha-\sqrt{\alpha^2-4\mu})t}.$$

3.3 The non-differentiable case

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

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

To study some optimization algorithms dedicated to non smooth functions, it may be useful to understand the behavior of solutions of (21). For example, to solve the LASSO problem:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \beta\|x\|_1 \tag{22}$$

proximal algorithms such as the Forward Backward can be used. It is known that on such problems inertial algorithms like FISTA may be used. It is shown in [1] that the behavior of FISTA is linked with the behavior of solutions of:

$$0 \in \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \partial F(x(t)). \tag{23}$$

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. If F is non smooth but strongly convex or quasi-strongly convex, it may be interesting to understand how the Heavy Ball algorithm can be used, and how to choose the parameter α. Since the C^2 assumption is irrelevant here, an analysis with weaker assumptions of the differential inclusion may be useful. Actually, we will see in the part dedicated to the optimization scheme, that the previous analysis applies to non smooth functions.

3.3.1 Solutions of the differential inclusion

The differential inclusion problem (21) admits a shock solution [24, 1] and it is known [3, 13] that for any solution x of (21), $F(x(t)) - F^*$ 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 is 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 (21), we use exactly the same energy defined to study the Heavy Ball ODE :

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

This time these Lyapunov energies may not be differentiable. Fortunately, the shock solutions [24, 1] of the differential inclusion (21) are obtained as limit of C^2 functions, where the the subdifferential ∂F is replaced by its Moreau Yosida approximation [1].

Let us recall the definition of shock solution for the differential inclusion (21):

Definition 3 (Shock solution [24, 1])**.** *A function $x : [t_0, +\infty) \longrightarrow \mathbb{R}^n$ is an energy-conserving shock solution of the differential inclusion* (21) *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:

$$F_\gamma(x) = \min_y \left(F(y) + \frac{1}{2\gamma}\|x - y\|^2 \right) \tag{24}$$

and the following approximating ODE:

$$\begin{aligned} &\ddot{x}_\gamma(t) + \alpha\dot{x}_\gamma(t) + \nabla F_\gamma(x_\gamma(t)) = 0 \\ &x_\gamma(t_0) = x_0 \quad \dot{x}_\gamma(t_0) = v_0. \end{aligned} \tag{25}$$

The differential equation (25) 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 [1, Theorems 3.2 and 3.3], we have the following result:

Theorem 4. *Assume F to be a lower semi continuous convex function. Let $\{F_\gamma\}_{\gamma>0}$ the Moreau-Yosida approximations of F. There exists a sub-sequence $\{x_\gamma\}_{\gamma>0}$ of solutions of* (25) *that converges to a shock solution of* (21) *according to the following scheme :*

- $x_\gamma \underset{\gamma\to 0}{\longrightarrow} x$ *uniformly on $[t_0, T]$ for all $T > t_0$.*
- $\dot{x}_\gamma \underset{\gamma\to 0}{\longrightarrow} \dot{x}$ *in $L^p([t_0, T]; \mathbb{R}^n)$, for all $p \in [1, +\infty)$ and $T > t_0$.*
- $F_\gamma(x_\gamma) \underset{\gamma\to 0}{\longrightarrow} F(x)$ *in $L^p([t_0, T]; \mathbb{R}^n)$, for all $\forall p \in [1, +\infty)$ and $T > t_0$.*

From Corollary 3.6 of [1], we also have:

Corollary 1. *If dom$(F) = \mathbb{R}^n$, then the differential inclusion* (21) *admits a shock solution x , such that :*

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

It turns out that all the results shown for the Heavy ball ODE remain valid for the differential inclusion (21). Indeed, the approximated solutions x_γ of Theorem 4 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 (21) also satisfies these properties (see e.g. [1] 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 (22):

Corollary 2. *Let us set $F(x) = \frac{1}{2}\|Ax - b\|^2 + \beta\|x\|_1$. Assume that $Ker(A) = \{0\}$. Then F is μ-strongly convex, where μ is the minimal spectral value of A^*A, there is a solution of the differential inclusion* (21) *such that the conclusions of Theorem 1 hold.*

3.3.2 Optimality of the decays

In Theorem 1, we assert that if $F \in q\mathcal{S}_\mu^{1,1}$ i.e. if F is a continuously differentiable μ-quasi-strongly convex function, and if $\alpha < 3\sqrt{\frac{\mu}{2}}$, we can ensure that:

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

In this section, we show that this decay also applies to some solutions of the associated differential inclusion (21) for functions in $q\mathcal{S}_\mu$. A natural question arises : is this rate optimal, or can we expect a better rate on $q\mathcal{S}_\mu$? The Theorem 5 answers to this question by proving that the exponential rate $\frac{2\alpha}{3}$ cannot be improved on the class $\mathcal{S}_\mu$ for any $\alpha < 3\sqrt{\frac{\mu}{2}}$. Its proof detailed in Section 6.3 lies on lower bounds of suitable Lyapunov energies.

Theorem 5. *If $F(x) = |x| + \frac{\mu}{2}|x|^2$ and if $\alpha < 3\sqrt{\frac{\mu}{2}}$ then any solution of* (21) *and any $\delta > \frac{2\alpha}{3}$ we have*

$$\limsup_{t \geqslant t_0} e^{\delta t}(F(x(t)) - F^*) > 0. \tag{26}$$

4 The discrete case

In this section we present a new inertial scheme to minimize a function $F \in \mathcal{S}_{\mu,L}^{1,1}$ i.e. μ-strongly convex, differentiable, whose gradient is L-Lipschitz continuous. In a second time, this scheme is extended to a sum of two convex functions $F = f + h$ using an inertial proximal gradient algorithm. These schemes can be seen as discretizations of the Heavy Ball ODE and they are variations of the schemes proposed by B.T. Polyak [25], Y. Nesterov [21] and J.W. Siegel [29].

Most gradient algorithms, classical gradient descent or inertial algorithms, ensure a linear decay of $F(x_n) - F^*$ when $F \in \mathcal{S}_{\mu,L}^{1,1}$. This decay depends mostly on the condition number κ:

$$\kappa = \frac{\mu}{L}.$$

In his book [23, Theorem 2.1.13], Y. Nesterov has shown that for such functions, any first order method cannot ensure in general a better decreasing rate than

$$F(x_n) - F^* = \mathcal{O}\left(\frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}\right)^{2n}. \tag{27}$$

When κ is small, we have

$$\left(\frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}\right)^2 = 1 - 4\sqrt{\kappa} + o(\sqrt{\kappa}). \tag{28}$$

We will see that several algorithms inspired by the continuous Heavy Ball ODE have been proposed to minimize functions belonging to $\mathcal{S}^{1,1}_{\mu,L}$. In the next section we provide a new scheme achieving the optimal rate for the continuous ODE.

The seminal Heavy Ball algorithm proposed by Polyak in [25] was designed for functions in $\mathcal{S}^{2,1}_{\mu,L}$ that are C^2 and strongly convex. It turns out that the C^2 hypothesis is crucial to ensure the convergence and the rate of the method. Since we do not make this C^2 assumption, we will not compare extensively our algorithm to the classical Heavy Ball algorithm.

4.1 The Differentiable case

Several algorithms to minimize 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:

$$\ddot{x}(t) + \alpha\dot{x}(t) + \nabla F(x(t)) = 0 \tag{29}$$

rewritten as the following first order differential system:

$$\begin{cases} \dot{x}(t) = v(t) \\ \dot{v}(t) = -\alpha v(t) - \nabla F(x(t)). \end{cases} \tag{30}$$

The first one was proposed by Polyak in [25] for functions in $\mathcal{S}^{2,1}_{\mu,L}$:

$$\begin{cases} x_{n+\frac{1}{2}} &= x_n + \left(\frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}\right)^2 (x_n - x_{n-1}) \\ x_{n+1} &= x_{n+\frac{1}{2}} - s^2\nabla F(x_n) \end{cases} \tag{31}$$

with $s = \frac{2}{\sqrt{L}+\sqrt{\mu}}$ which can bee 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 [14] for example. It is worth mentioning that Ghadimi et al. in [14, Theorem 4] prove 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:

$$F(x_n) - F^* = \mathcal{O}((1-\kappa)^n) \tag{32}$$

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 [23], Nesterov proposes a scheme which is quite similar:

$$\begin{cases} x_{n+\frac{1}{2}} = & x_n + \left(\dfrac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}\right)(x_n - x_{n-1}) \\ x_{n+1} = & x_{n+\frac{1}{2}} - s^2\nabla F(x_{n+\frac{1}{2}}) \end{cases} \tag{33}$$

with $s = \frac{1}{\sqrt{L}}$. This scheme can also be seen as 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 (33) for functions in $\mathcal{S}^{1,1}_{\mu,L}$ and he gives a convergence rate:

$$F(x_n) - F^* = \mathcal{O}((1-\sqrt{\kappa})^n). \tag{34}$$

Notice that another variant of this algorithm with the same asymptotic decrease rate was also introduced by Y. Nesterov in [22] with an extension to non differentiable functions (but still strongly convex). An application of this last scheme to image processing can be found in [6].

The schemes (31) and (33) are called two points schemes since the computation of x_{n+1} needs the two previous points x_n and x_{n-1}. These schemes can also be written in another building the point x_{n+1} from the previous iterate x_n and a direction v_n. Hence the Heavy Ball algorithm can be written as a two steps scheme where the input is a pair (x_n, v_n):

$$\begin{cases} x_{n+\frac{1}{2}} = & x_n + s v_n \\ v_{n+\frac{1}{2}} = & v_n \\ x_{n+1} = & x_{n+\frac{1}{2}} - s^2\nabla F(x_n) \\ v_{n+1} = & v_n + s(-2\sqrt{\mu}\frac{1}{1+\sqrt{\kappa}}v_n - \nabla F(x_n)) \end{cases} \tag{35}$$

with $s = \frac{2}{\sqrt{L}+\sqrt{\mu}}$. The link between the scheme (35) and the ODE (29) with $\alpha = 2\sqrt{\mu}$ appears in the set up of the variable v_{n+1}: when κ tends to 0, we can see that: $v_{n+1} \approx v_n + s(-2\sqrt{\mu}v_n - \nabla F(x_n))$. Similarly the Nesterov scheme can be written as:

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

with $s = \frac{1}{\sqrt{L}}$. Once again the link between the scheme and the ODE can be seen in the expression of v_{n+1}. Later J.W. Siegel in [29] proposed a variation of these schemes:

$$\begin{cases} x_{n+\frac{1}{2}} = & x_n + s v_n \\ v_{n+\frac{1}{2}} = & (1+\sqrt{\kappa})^{-2}(v_n - s\nabla F(x_{n+\frac{1}{2}})) \\ x_{n+1} = & x_{n+\frac{1}{2}} - s^2\nabla F(x_{n+\frac{1}{2}}) \\ v_{n+1} = & v_{n+\frac{1}{2}} + s(1+\sqrt{\kappa})^{-1}\sqrt{\kappa}\nabla F(x_{n+\frac{1}{2}}) \end{cases} \tag{37}$$

with $s = \frac{1}{\sqrt{L}}$. Notice that the scheme (37) provides the same convergence rate as the scheme proposed by Nesterov, see (34). The expression of v_{n+1} can be stated as follows:

$$v_{n+1} = v_n + \frac{s}{(1+\sqrt{\kappa})^2}\left[-\alpha\left(1-\frac{\sqrt{\kappa}}{2}\right)v_n + (-1+\sqrt{\kappa}+\kappa)\nabla F(x_{n+\frac{1}{2}})\right].$$

One can observe in this last expression that the sequence $(v_n)_{n\in\mathbb{N}}$ is a particular discretization of the variable $v = \dot{x}$ in (30). It turns out that this discretization allows to reach a decay rate similar to the Nesterov scheme (34).

We can also remark that for the three given schemes the choice of $v_{n+\frac{1}{2}}$ is arbitrary since x_{n+1} does not directly depend on $v_{n+\frac{1}{2}}$. From an algorithmic point of view $v_{n+\frac{1}{2}}$ is actually hidden and it has no real interest in the Polyak and the Nesterov schemes. The main issue in defining $v_{n+\frac{1}{2}}$ is in the theoretical analysis of these inertial algorithms. It turns out that the definition of $v_{n+\frac{1}{2}}$ simplifies the Lyapunov analysis of the scheme introduced by Siegel and will be useful to analyze the scheme we introduce now inspired by the previous ones:

$$\begin{cases} x_{n+\frac{1}{2}} = & x_n + sv_n \\ v_{n+\frac{1}{2}} = & (1+\frac{3\lambda s}{2})^{-1}(v_n - s\nabla F(x_{n+\frac{1}{2}})) \\ x_{n+1} = & x_{n+\frac{1}{2}} - s^2\nabla F(x_{n+\frac{1}{2}}) \\ v_{n+1} = & v_{n+\frac{1}{2}} + (1+\lambda s)^{-1}\lambda s^2 \nabla F(x_{n+\frac{1}{2}}) \end{cases} \tag{38}$$

with $s = \frac{1}{\sqrt{L}}$. The careful reader can check that the sequence $(v_n)_{n\in\mathbb{N}}$ is yet another discretization of the variable $v = \dot{x}$ in (30) with $\alpha = \frac{3\lambda}{2}$. The interest of this new scheme (38) is that it allows to provide a better decay rate of $F(x_n) - F^*$ that is asymptotically better than the previous ones for suitable choices of λ :

Theorem 6. *If F is μ-strongly convex, if $\lambda < \sqrt{2\mu}$, if F is differentiable, with gradient $L-$ Lipschitz, the sequence $(x_n)_{n\in\mathbb{N}}$ provided by* (38) *with $s = \frac{1}{\sqrt{L}}$ satisfies:*

$$F(x_n) - F^* \quad \leqslant \quad \left(1 - \frac{\lambda^2}{2\mu}\right)^{-1} \left(F(x_0) - F^* + \frac{1}{2}\|\lambda(x_0 - x^*) + \left(1 + \frac{\lambda}{\sqrt{L}}\right) v_0\|^2\right) \left(1 + \sqrt{\frac{\lambda^2}{L} - \frac{3\lambda^2}{2L}}\right)^{-n}$$

One can observe that for κ small enough, we have: $(1 + \sqrt{\frac{\lambda^2}{L} - \frac{3\lambda^2}{2L}})^{-1} < 1 - \sqrt{\kappa}$, that is the decay is better than the ones achieved by the schemes of Nesterov (33) and Siegel (37). More precisely, the new scheme (38) allows to get a rate:

$$F(x_n) - F^* = \mathcal{O}(1 + \sqrt{2\kappa} - 3\kappa - \varepsilon)^{-n}$$

for any $\varepsilon > 0$ and for κ small enough, choosing λ depending on ε. Notice that we have

$$(1 + \sqrt{2\kappa} - 3\kappa)^{-1} = 1 - \sqrt{2\kappa} + o(\sqrt{\kappa}). \tag{39}$$

Hence we see that our new scheme improves over Nesterov's rate (34) by a factor $\sqrt{2}$.

Recently, a triple momentum method has been introduced in [30] with the following rate:

$$F(x_n) - F^* \leq \frac{C}{\kappa}\left(1 - \sqrt{\kappa}\right)^{2n}. \tag{40}$$

When κ is small, we have

$$(1 - \sqrt{\kappa})^2 = 1 - 2\sqrt{\kappa} + o(\sqrt{\kappa}). \tag{41}$$

This provides a better asymptotic rate. But one should notice that $\frac{C}{\kappa}$ explodes when $\kappa \to 0^+$. Hence it may not be the best choice when one is interested in finite error bounds. Moreover, as far as we know, this method cannot be extended to the case of composite optimization with $F = g + h$ where g is a L-Lipschitz gradient convex function and h a possibly non smooth convex lower semi-continuous function. The function F to be minimized needs to be differentiable to use the scheme of [30], contrary to the results presented in this paper (see Theorem 7).

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:

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

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:

$$0 \in \ddot{x}(t) + \alpha\dot{x}(t) + \partial F(x(t)). \tag{42}$$

This inclusion problem admits a shock solution (see [1] and [24]) 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 [29] proposes an extension of the discrete scheme built for differentiable function. In the following section we prove that our scheme can be directly extended to such function F using the Forward-Backward algorithm, also called Proximal Gradient Operator.

We recall the definition of the proximal operator:

$$\operatorname{prox}_h(x) = \operatorname{argmin}_y \left(h(x) + \frac{1}{2}\|x - y\|^2 \right).$$

Using the optimality condition, we have the equivalence:

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

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.

If F is convex, where $F = f + h$, with f differentiable, ∇f is L-Lipschitz 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:

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

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 $\tilde{x}$ of F.

The operator T shares many properties with the gradient descent. The algorithm FISTA of Beck and Teboulle [10] can be seen as a Nesterov acceleration to this operator T. Following Siegel [29] we modify the previous scheme so that it can be used with $F = f + h$ with f a smooth strongly convex function with L-Lipschitz gradient and h a possibly non smooth convex function: we replace $g = \frac{1}{\lambda}\nabla F(x_{n+\frac{1}{2}})$ in (38) by

$$\tilde{g} = \frac{\lambda}{t^2}\left(x_{n+\frac{1}{2}} - \operatorname{prox}_{s^2 h}(x_{n+\frac{1}{2}} - s^2\nabla f(x_{n+\frac{1}{2}}))\right) \tag{43}$$

where $t := \lambda s$. More precisely the new scheme can be written as:

$$\begin{cases} \lambda x_{n+\frac{1}{2}} &= \lambda x_n + t v_n \\ v := v_{n+\frac{1}{2}} &= (1 + \frac{3t}{2})^{-1}(v_n - t\tilde{g}) \\ \lambda x_{n+1} &= \lambda x_{n+\frac{1}{2}} - t^2\tilde{g} \\ v_{n+1} &= v + (1+t)^{-1}t^2\tilde{g} \end{cases} \tag{44}$$

that is exactly the original scheme (38) replacing g by $\tilde{g}$. This new scheme shares the same properties as the previous one:

Theorem 7. *Let $F = f + h$. If f is $\mu-$strongly convex, differentiable with gradient L-Lipschitz, if h is convex, proper and lower semi-continuous, if $\lambda < \sqrt{2\mu}$, the sequence $(x_n)_{n\in\mathbb{N}}$ provided by (44) with $s = \frac{1}{\sqrt{L}}$ satisfies :*

$$F(x_n) - F^* \quad \leqslant \quad \left(1 - \frac{\lambda^2}{2\mu}\right)^{-1} \left(F(x_0) - F^* + \frac{1}{2}\|\lambda(x_0 - x^*) + (1 + \frac{\lambda}{\sqrt{L}})v_0\|^2\right) \left(1 + \sqrt{\frac{\lambda^2}{L} - \frac{2\lambda^2}{L}}\right)^{-n}.$$

This Theorem applies then to the LASSO problem when the function $x \mapsto \|Ax - b\|^2$ is strongly convex, i.e. when $\ker(A) \neq \{0\}$, and it ensures that in this setting we can expect an exponential decay $\mathcal{O}(1 + \sqrt{(2-\varepsilon)\kappa})^{-n}$ for any ε and κ sufficiently small. As far as we kwnow, this is the best rate that can be found in the literature.

5 Numerical results

In this section, we illustrate the theoretical results of the previous sections.

5.1 Case of an anisotropic quadratic function

To compare the Heavy ball based algorithms, we first test them on a toy example:

$$F(x_1, x_2) = \frac{1}{2}x_1^2 + 500x_2^2. \tag{45}$$

For this function, $\mu = 1$ and $L = 1000$. The starting point is set to $x_{init} = [1, 1]$.

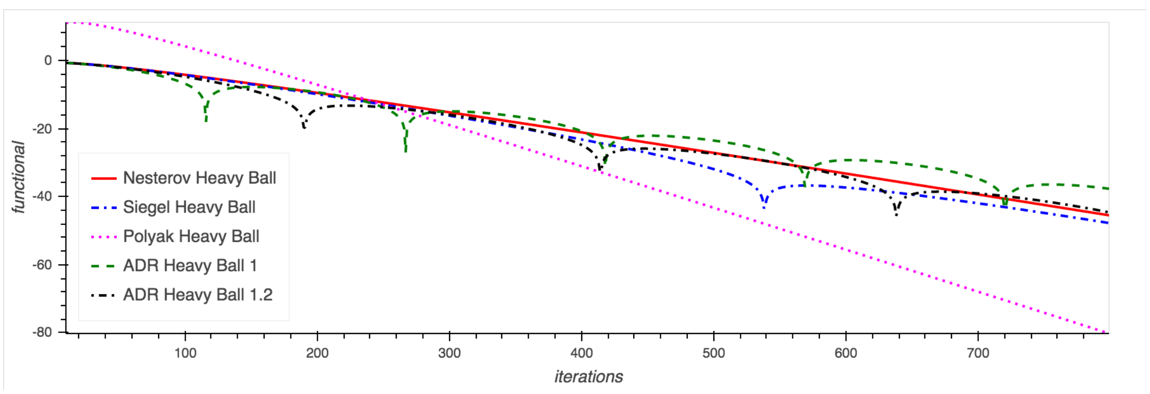

Figure 2: Comparison of various algorithms

In Figure 2, the logarithm of $\log(F(x_n) - F^*)$ is computed for each sequence provided by the various algorithms. We can first observe that Polyak's scheme [25] has the best asymptotic rate although it is slower at the beginning. Actually, the distance to the minimizer may grow during the first iterations due to the large step size $s \approx \frac{4}{L}$ when $\kappa = \frac{\mu}{L} << 1$. For high accuracy, the Polyak's scheme is the most efficient, but for a fair approximation of the minimizer it may not be the best method to use. We can see that the other algorithms share roughly the same behavior (with a better convergence than Polyak's scheme for the first iterations). Since the function F to minimize is quadratic, it is not surprising (it is known since [25] that Polyak's scheme is asymptotically optimal among quadratic functions). The schemes of Nesterov [23], Siegel [29] and the ones introduced in this paper are built to be as efficient as possible for strongly convex functions (but not for quadratic functions). Moreover all these schemes are discretizations of the same ODE, so that their behaviors are similar.

5.2 An example of divergence for Polyak's scheme

In Figure 3 we can observe the convergence of Polyak's scheme to a 3-cycle for a function $F \in \mathcal{S}^{1,1}_{\mu,L}$ but $F \notin \mathcal{S}^{2,1}_{\mu,L}$ enlightening the difference between the two classes of functions and the possible problematic behavior of Polyak's scheme for the set $F \in \mathcal{S}^{1,1}_{\mu,L}$. This example was given in Lessard et al. [17].

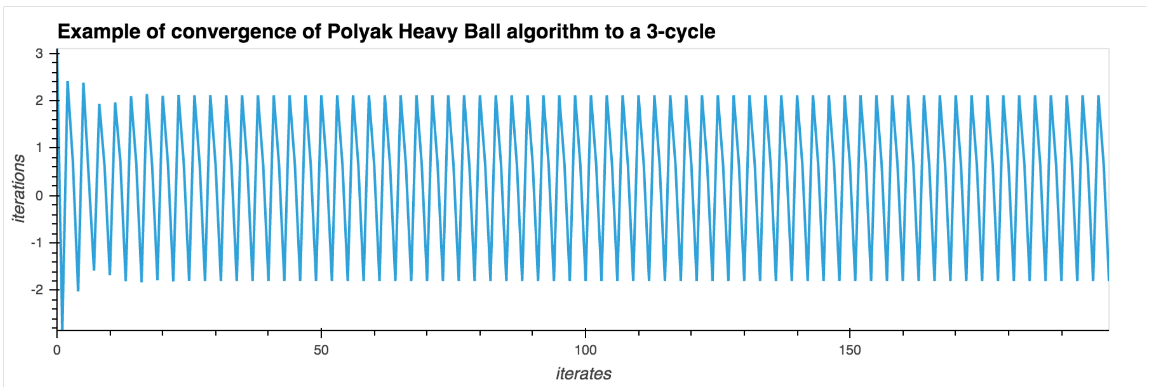

Figure 3: Convergence of Polyak's algorithm to a 3-cycle for a function $F \in \mathcal{S}^{1,1}_{\mu,L}$.

5.3 Case of a non smooth anisotropic strongly convex function

In this last example, we compare the algorithm of Siegel [29] and the one presented in this paper designed for non smooth and composite functions applied to the function:

$$F(x_1, x_2) = \frac{\mu}{2}x_1^2 + \frac{L}{2}x_2^2 + |x_1| + |x_2| \tag{46}$$

with $\mu = 10^{-2}$ and $L = 10^4$. On this example, we can observe that the convergence to the minimizer is better with our algorithm.

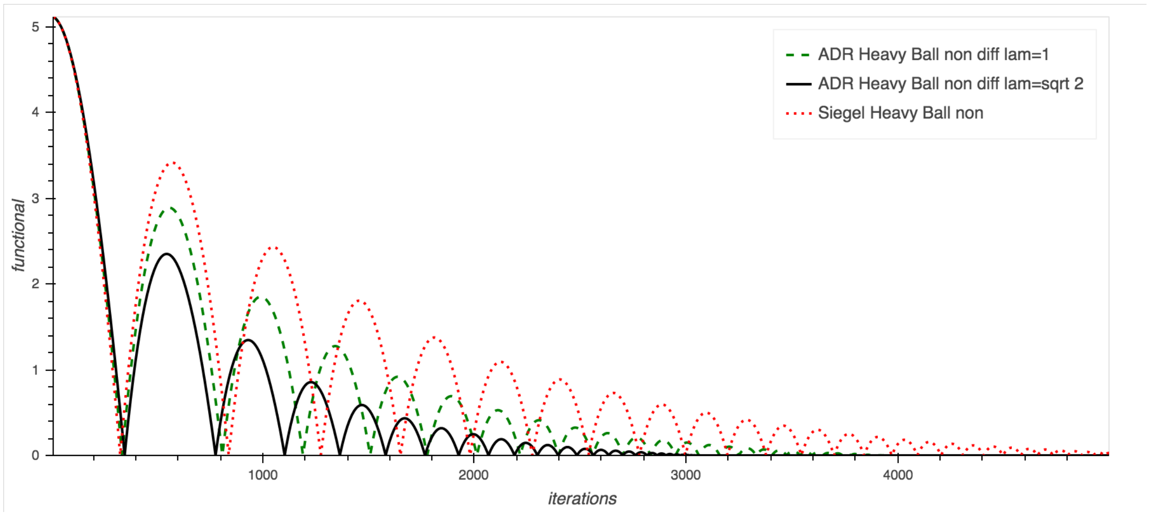

Figure 4: Comparison of our scheme with the one of Siegel applied to composite and non differentiable functions.

6 Proofs

6.1 Proofs of Theorems 1 and 2

We consider the energy $\mathcal{E}(t)$ defined by

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

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

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

Using now the μ-quasi-strong convexity of F, we get:

$$\begin{aligned}\mathcal{E}'(t) &\leqslant -\lambda(F(x(t)) - F^*) - \frac{\lambda\mu}{2}\|x(t)-x^*\|^2 + (\lambda - \alpha(1+\beta))\|\dot{x}(t)\|^2 \\ &\quad +(\xi + \lambda(\lambda-\alpha))\langle \dot{x}(t), x(t)-x^*\rangle \\ &\leqslant -\lambda\mathcal{E}(t) + \lambda\beta(F(x(t)) - F^*) + \frac{\lambda}{2}(\xi + \lambda^2 - \mu)\|x(t)-x^*\|^2 \\ &\quad + \left(\frac{3}{2}\lambda - \alpha + \beta\left(\frac{\lambda}{2} - \alpha\right)\right)\|\dot{x}(t)\|^2 + (\xi + \lambda(2\lambda - \alpha))\langle \dot{x}(t), x(t)-x^*\rangle\end{aligned}$$

From now on, there will be 2 cases: the case $\beta = 0$ corresponding to the Theorem 1 and the case $\beta \geqslant 0$ corresponding to the Theorem 2.

- If $\beta = 0$, we then have:

$$\mathcal{E}'(t)+\lambda\mathcal{E}(t) \leqslant \frac{\lambda}{2}(\xi+\lambda^2-\mu)\|x(t)-x^*\|^2+\left(\frac{3}{2}\lambda-\alpha\right)\|\dot{x}(t)\|^2+(\xi+\lambda(2\lambda-\alpha))\langle\dot{x}(t),x(t)-x^*\rangle \tag{47}$$

- If $\beta \geqslant 0$ and assuming in addition that F has a L-Lipschitz gradient, we have:

$$F((x(t)) - F^* \leq \frac{L}{2}\|x(t)-x^*\|^2,$$

hence:

$$\begin{aligned}\mathcal{E}'(t)+\lambda\mathcal{E}(t) &\leqslant \frac{\lambda}{2}(\beta L+\xi+\lambda^2-\mu)\|x(t)-x^*\|^2+(\xi+\lambda(2\lambda-\alpha))\langle\dot{x}(t),x(t)-x^*\rangle\\ &+\left(\frac{3}{2}\lambda-\alpha+\beta\left(\frac{\lambda}{2}-\alpha\right)\right)\|\dot{x}(t)\|^2\end{aligned} \tag{48}$$

In these two cases, the idea is to choose the parameters $\lambda > 0$ (as large as possible) and ξ such that the right side of the differential inequality is negative. For that, we need to control a term of the form:

$$F(a,b,c) = a\|u\|^2 + b\|v\|^2 + c\langle u,v\rangle \tag{49}$$

Lemma 5. *If $a \leq 0$, $b \leq 0$ and $c^2 \leq 4ab$, then $F(a,b,c) \leq 0$.*

Proof. Observe that if $a = 0$ then $c = 0$ and the result of the Lemma holds since $b \leq 0$. Assume now that $a < 0$.

$$F(a,b,c) = -(-a\|u\|^2 - b\|v\|^2 - c\langle u,v\rangle) \leqslant -\left(\sqrt{-a}\|u\| - \frac{|c|}{2\sqrt{-a}}\|v\|\right)^2 - \left(\frac{c^2}{4a}-b\right)\|v\|^2.$$

Hence, to have $F(a,b,c) \leq 0$, it is sufficient to have $\frac{c^2}{4a} - b \geq 0$, i.e. (since $a < 0$): $c^2 \leq 4ab$. □

Proof of Theorem 1 (case $\beta = 0$) As previously stated by (47), for the class of the μ-quasi-strongly convex functions F, we have:

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

The question now is how to choose the parameter $\lambda > 0$ as large as possible and ξ with respect to the friction coefficient α such that the right side of the inequality (47) is negative while ensuring the control of the values $F(x(t)) - F^*$ by the energy $\mathcal{E}_{\lambda,\xi}$.

According to Lemma 5, to get: $\mathcal{E}'_{\lambda,\xi}(t)+\lambda\mathcal{E}_{\lambda,\xi}(t) \leqslant 0$, it is sufficient to choose λ and ξ such that:

$$\xi+\lambda^2-\mu \leqslant 0, \quad \lambda \leqslant \frac{2}{3}\alpha, \quad \Delta(\lambda,\xi,\alpha) \leqslant 0. \tag{50}$$

where: $\Delta(\lambda,\xi,\alpha) := (\xi+2\lambda^2-\lambda\alpha)^2 - \lambda(\xi+\lambda^2-\mu)(3\lambda-2\alpha)$.

Firstly, we choose the parameter ξ that minimize the quantity $\Delta(\lambda,\xi,\alpha)$ i.e.:

$$\xi = -\frac{\lambda^2}{2} < 0. \tag{51}$$

Thus the conditions (50) can thus be rewritten as:

$$\tilde{\Delta}(\lambda,\alpha)=\Delta(\lambda,-\frac{\lambda^2}{2},\alpha)\leqslant 0,\quad \lambda\leqslant\frac{2}{3}\alpha,$$

where: $\tilde{\Delta}(\lambda,\alpha)=\frac{\lambda}{4}(3\lambda-2\alpha)(\lambda^2-2\alpha\lambda+4\mu)$.

- Assume first $\alpha\leqslant 3\sqrt{\frac{\mu}{2}}$ i.e.: $\lambda=\frac{2}{3}\alpha\leqslant\sqrt{2\mu}$. In that case, we have: $\tilde{\Delta}(\frac{2}{3}\alpha,\alpha)=0$ so that the largest admissible value for λ is: $\lambda^*=\frac{2}{3}\alpha$. With this choice of parameter, we have:

  $$\forall t\geqslant t_0,\ \mathcal{E}_{\lambda,\xi}'(t)+\frac{2}{3}\alpha\mathcal{E}_{\lambda,\xi}(t)\leqslant 0,$$

  hence using Lemma 3:

  $$\forall t\geqslant t_0,\ \mathcal{E}_{\lambda,\xi}(t)\leqslant\mathcal{E}_{\lambda,\xi}(t_0)e^{-\frac{2}{3}\alpha(t-t_0)}.$$

  Let us now prove that we can control the trajectory in finite time. Observe that the quasi-strong convexity of F ensures that F satisfies the global growth condition $\mathcal{G}(2)$ with the real constant μ so that:

  $$\forall t\geqslant t_0,\ \mathcal{E}_{\lambda,\xi}(t)\geqslant(1-\frac{2\alpha^2}{9\mu})(F(x(t))-F^*)+\frac{1}{2}\left\|\lambda(x(t)-x^*)+\dot{x}(t)\right\|^2.\qquad(52)$$

  Hence:

  $$\forall t\geqslant t_0,\ \mathcal{E}_{\lambda,\xi}(t)\geqslant\frac{1}{2}\left\|\lambda(x(t)-x^*)+\dot{x}(t)\right\|^2.$$

  We now set: $y(t)=e^{\lambda t}(x(t)-x^*)$. It follows:

  $$\forall t\geqslant t_0,\ \|\dot{y}(t)\|=e^{\lambda t}\|\lambda(x(t)-x^*)+\dot{x}(t)\|\leqslant\sqrt{2\mathcal{E}_{\lambda,\xi}(t_0)}e^{\frac{\lambda}{2}t_0}e^{\frac{\lambda}{2}t}.$$

  Integrating between t_0 and t, we then get:

  $$\forall t\geqslant t_0,\ \|y(t)\|\leqslant\frac{2}{\lambda}\sqrt{2\mathcal{E}_{\lambda,\xi}(t_0)}e^{\frac{\lambda}{2}t_0}e^{\frac{\lambda}{2}t},$$

  hence:

  $$\forall t\geqslant t_0,\ \|x(t)-x^*\|^2\leqslant\frac{8}{\lambda^2}\mathcal{E}_{\lambda,\xi}(t_0)e^{-\lambda(t-t_0)}.$$

  Coming back to the definition of the energy $\mathcal{E}_{\lambda,\xi}(t)$, we have for all $t\geqslant t_0$

  $$\begin{aligned}F(x(t))-F^* &= \mathcal{E}_{\lambda,\xi}(t)-\frac{1}{2}\left\|\lambda(x(t)-x^*)+\dot{x}(t)\right\|^2+\frac{\lambda^2}{4}\left\|x(t)-x^*\right\|^2\\ &\leqslant \mathcal{E}_{\lambda,\xi}(t)+\frac{\lambda^2}{4}\left\|x(t)-x^*\right\|^2\leqslant 3\mathcal{E}_{\lambda,\xi}(t_0)e^{-\frac{2}{3}\alpha(t-t_0)}\end{aligned}$$

  which concludes the proof using Lemma 3.

- Assume now that $\alpha>3\sqrt{\frac{\mu}{2}}$ i.e.: $\frac{2}{3}\alpha>\sqrt{2\mu}$. In that case, $\frac{2}{3}\alpha$ is not an admissible value for λ anymore. Let us then discuss the sign of $\tilde{\Delta}(\lambda,\alpha)$ when $\lambda<\sqrt{2\mu}<\frac{2}{3}\alpha$, which is equivalent to study the sign the polynomial $G(\lambda,\alpha)=\lambda^2-2\alpha\lambda+4\mu$. Since $\alpha>3\sqrt{\frac{\mu}{2}}$, its discriminant $\delta=\alpha^2-4\mu$ is non negative and G admits two real roots:

$$\lambda_{\pm}=\alpha\pm\sqrt{\alpha^2-4\mu}.$$

Observe now that $\alpha+\sqrt{\alpha^2-4\mu}>\frac{2\alpha}{3}$, hence $\alpha+\sqrt{\alpha^2-4\mu}$ is not an admissible value for λ. But we easily prove that the hypothesis $\alpha>3\sqrt{\frac{\mu}{2}}$ is equivalent to: $\alpha-\sqrt{\alpha^2-4\mu}<\sqrt{2\mu}$, so that the largest admissible value for λ is given by:

$$\lambda^*=\alpha-\sqrt{\alpha^2-4\mu}.$$

As previously done, the energy $\mathcal{E}_{\lambda,\xi}$ satisfies the following differential inequality: $\mathcal{E}_{\lambda,\xi}'(t)+\lambda^*\mathcal{E}_{\lambda,\xi}(t)\leqslant 0$ so that using Lemma 3, for all $t\geqslant t_0$

$$\mathcal{E}_{\lambda,\xi}(t)\quad\leqslant\quad\mathcal{E}_{\lambda,\xi}(t_0)e^{-\lambda^*(t-t_0)}\leqslant\left[1+\left(1+\frac{\alpha}{\mu}-\sqrt{\frac{\alpha^2}{\mu^2}-4}\right)^2\right]M_0e^{-\lambda^*(t-t_0)}$$

where: $M_0=F(x_0)-F^*+\frac{1}{2}\|v_0\|^2$. Coming back now to the definition of the energy and applying the quadratic growth condition $\mathcal{G}_\mu(2)$, we get:

$$\begin{aligned}\mathcal{E}_{\lambda,\xi}(t)\quad&=\quad F(x(t))-F^*+\frac{1}{2}\left\|\lambda(x(t)-x^*)+\dot{x}(t)\right\|^2-\frac{\lambda^2}{4}\left\|x(t)-x^*\right\|^2\\&\geqslant\quad F(x(t))-F^*-\frac{\lambda^2}{4}\left\|x(t)-x^*\right\|^2\geqslant(1-\frac{\lambda^2}{2\mu})(F(x(t))-F^*).\end{aligned}$$

And the expected control on the values $F(x(t))-F^*$ is obtained straightforward since $\lambda^*=\alpha-\sqrt{\alpha^2-4\mu}<\sqrt{2\mu}$ for any $\alpha>3\sqrt{\frac{\mu}{2}}$.

Proof of Theorem 2 (case when $\beta\geqslant 0$) As previously stated for the class of the μ-quasi-strongly convex functions F having a Lipschitz continuous gradient, we have:

$$\begin{aligned}\mathcal{E}'(t)+\lambda\mathcal{E}(t)\quad&\leqslant\quad\frac{\lambda}{2}(\beta L+\xi+\lambda^2-\mu)\|x(t)-x^*\|^2+\left(\frac{3}{2}\lambda-\alpha+\beta\left(\frac{\lambda}{2}-\alpha\right)\right)\|\dot{x}(t)\|^2\\&\quad+(\xi+\lambda(2\lambda-\alpha))\langle\dot{x}(t),x(t)-x^*\rangle.\end{aligned}$$

The question now is how to choose the parameters $\lambda>0$ as large as possible, $\beta\geqslant 0$ and ξ with respect to the friction coefficient α and the Lipschitz parameter L such that:

$$\forall t\geqslant t_0,\mathcal{E}'(t)+\lambda\mathcal{E}(t)\leqslant 0.\tag{53}$$

Let us define:

$$\Delta(\lambda,\xi,\alpha,\beta):=(\xi+2\lambda^2-\lambda\alpha)^2-\lambda(\beta L+\xi+\lambda^2-\mu)\left(3\lambda-2\alpha+\beta(\lambda-2\alpha)\right).\tag{54}$$

According to Lemma 5, it is sufficient to choose λ and ξ such that:

$$\beta L+\xi+\lambda^2-\mu\leqslant 0,\quad\tfrac{3}{2}\lambda-\alpha+\beta(\tfrac{\lambda}{2}-\alpha)\leqslant 0,\quad\Delta(\lambda,\xi,\alpha,\beta)\leqslant 0$$

to ensure that the energy $\mathcal{E}$ satisfies the differential inequality (53). First, we choose for ξ the one ensuring that Δ is minimal, i.e.:

$$\xi = -\frac{\lambda^2}{2}(1-\beta) - \beta\lambda\alpha. \tag{55}$$

Re-injecting the optimal ξ into (54), we eventually obtain:

$$\begin{aligned}\widetilde{\Delta}(\lambda,\alpha,\beta) &= \Delta(\lambda, -\frac{\lambda^2}{2}(1-\beta) - \beta\lambda\alpha, \alpha, \beta)\\ &= \frac{\lambda}{4}\left((3+\beta)\lambda - 2\alpha(1+\beta)\right)\left(\lambda^2(1-\beta) - 2\alpha(1-\beta)\lambda + 4(\mu-\beta L)\right).\end{aligned}$$

Consequently, applying Lemma 5, we need to choose the parameter λ as large as possible and satisfying all the following constraints for some $\beta \in [0,1]$:

$$\begin{cases} \lambda^2(1+\beta) - 2\alpha\beta\lambda - 2(\mu-\beta L) \leqslant 0, \\ \lambda \leqslant 2\alpha\dfrac{1+\beta}{3+\beta}, \quad \widetilde{\Delta}(\lambda,\alpha,\beta) \leqslant 0 \end{cases} \tag{56}$$

and the parameter $\beta \in [0,1]$ that maximizes the value of λ. Let:

$$P_1(\lambda,\beta,\alpha) = \lambda^2(1+\beta) - 2\alpha\beta\lambda - 2(\mu-\beta L).$$

Since $\tilde{\Delta}(2\alpha\frac{1+\beta}{3+\beta}, \alpha, \beta) = 0$, the quantity $2\alpha\frac{1+\beta}{3+\beta}$ is the largest admissible value for λ if and only if:

$$P_1(2\alpha\frac{1+\beta}{3+\beta}, \beta, \alpha) \leqslant 0,$$

or equivalently, if and only if:

$$Q_\alpha(\beta) = 2\alpha^2(1-\beta^2) - (3+\beta)^2(\mu-\beta L) \leqslant 0 \tag{57}$$

for any $0 \leqslant \beta < \frac{\mu}{L}$. Let us so discuss the sign of Q_α with respect to the choice of α. To that end, observe that assuming $\alpha \leqslant 3\sqrt{\frac{\mu}{2}}$, we have:

$$Q_\alpha(0) = 2\alpha^2 - 9\mu \leqslant 0, \quad Q_\alpha(\frac{\mu}{L}) = 2\alpha^2(1-\frac{\mu}{L}) > 0.$$

Hence Q_α admits at least one real root, denoted by β^* inside the interval $[0, \frac{\mu}{L}[$ such that:

$$\forall \beta \in [0,\beta^*], \; Q_\alpha(\beta) \leqslant 0.$$

Note that the value $\beta = 0$ is admissible when $\alpha = 3\sqrt{\frac{\mu}{2}}$. In any case, the best choice for λ is: $\lambda = 2\alpha\frac{1+\beta}{3+\beta}$, and the best rate is obtained for the largest admissible β ensuring the control on the values $F(x(t)) - F^*$. And with these choices of parameters, we get the following control on the energy:

$$\mathcal{E}(t) \leqslant \mathcal{E}(t_0)e^{-\lambda(t-t_0)}. \tag{58}$$

As in the proof of Theorem 1, we first prove that we can control the trajectory $\|x(t) - x^*\|$ from the energy $\mathcal{E}$. With our choice of parameters ξ and λ and remembering that the μ-quasi-strong convexity of F ensures that F also satisfies the global growth condition $\mathcal{G}(2)$ with the constant μ (see Lemma 1), we have:

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

Observe now that for any $\beta \in [0,\beta^*]$, we have: $Q_\alpha(\beta) \leqslant 0$ which implies:

$$\alpha \leqslant (3+\beta)\sqrt{\frac{\mu-\beta L}{2(1-\beta^2)}} \leqslant (3+\beta)\sqrt{\frac{\mu}{2(1+\beta)}}.$$

It follows:

$$1+\beta-\frac{\lambda^2}{2\mu} = 1+\beta-2\alpha^2\frac{(1+\beta)^2}{\mu(3+\beta)^2} \geqslant 0$$

and:

$$\mathcal{E}(t) \geqslant \frac{1}{2(\beta+1)}\left\|\lambda(x(t)-x^*)+(\beta+1)\dot{x}(t)\right\|^2.$$

We now define: $y(t) = e^{\frac{\lambda}{1+\beta}t}(x(t)-x^*)$. Hence:

$$\begin{aligned}\|\dot{y}(t)\| &= e^{\frac{\lambda}{1+\beta}t}\|\frac{\lambda}{1+\beta}(x(t)-x^*)+\dot{x}(t)\| = \frac{e^{\frac{\lambda}{1+\beta}t}}{1+\beta}\|\lambda(x(t)-x^*)+(\beta+1)\dot{x}(t)\| \\ &\leqslant e^{\frac{\lambda}{1+\beta}t}\sqrt{\frac{2}{1+\beta}\mathcal{E}(t)} \leqslant \sqrt{\frac{2}{1+\beta}\mathcal{E}(t_0)}e^{-\frac{\lambda}{2}t_0}e^{\frac{1-\beta}{2(1+\beta)}\lambda t}.\end{aligned}$$

Remember that $\lambda = 2\alpha\frac{1+\beta}{3+\beta}$. Integrating between t_0 and t, we get:

$$\|y(t)\| \leqslant \frac{3+\beta}{\alpha(1-\beta)}\sqrt{\frac{2}{1+\beta}\mathcal{E}(t_0)}e^{-\frac{\lambda}{2}t_0}e^{\frac{1-\beta}{2(1+\beta)}\lambda t}$$

which implies:

$$\|x(t)-x^*\| \leqslant \frac{3+\beta}{\alpha(1-\beta)}\sqrt{\frac{2}{1+\beta}\mathcal{E}(t_0)}e^{-\frac{\lambda}{2}(t-t_0)}.$$

Coming back now to the definition of the energy, we have a control on the values $F(x(t))-F^*$ from the energy and the trajectory $x(t)-x^*$:

$$(1+\beta)(F(x(t))-F^*) \leqslant \mathcal{E}(t)+\frac{\lambda^2}{4}\|x(t)-x^*\|^2 \leqslant \left[1+2\frac{1+\beta}{(1-\beta)^2}\right]\mathcal{E}(t_0)e^{-\lambda(t-t_0)}.$$

Using Lemma 3 and the inequality (18), we finally obtain the expected control.

Despite the fact that the value of β^* is not exactly known, we can easily get a lower bound on β^* when $\alpha < 3\sqrt{\frac{\mu}{2}}$. Indeed remember that β^* is a real root of the polynomial Q_α chosen such that

$$\forall \beta \in [0,\beta^*],\ Q_\alpha(\beta) \leqslant 0 = Q_\alpha(\beta^*)$$

which is equivalent to:

$$\forall \beta \in [0, \beta^*],\ \alpha \leqslant (3+\beta)\sqrt{\frac{\mu - \beta L}{2(1-\beta^2)}}. \tag{59}$$

Observe now that for any $\beta \geqslant 0$, we have:

$$\frac{(3+\beta)^2}{2(1-\beta^2)} \geqslant \frac{9}{2}$$

so that it is sufficient to ensure: $\alpha \leqslant 3\sqrt{\frac{\mu-\beta L}{2}}$, or equivalently: $\beta \geqslant \frac{\mu}{L} - \frac{2\alpha^2}{9L}$ to obtain (59). From the lower bound $\underline{\beta} = \frac{\mu}{L} - \frac{2\alpha^2}{9L}$, we obtain a lower bound on λ:

$$\underline{\lambda} = \frac{2\alpha}{3}\left(1 + \frac{2}{3}\frac{9\mu - 2\alpha^2}{9L + 3\mu - \frac{2}{3}\alpha^2}\right) \geqslant \frac{2\alpha}{3}.$$

6.2 Proof of Theorem 3

Let λ, ξ and T three real numbers. Let x^* be a minimizer of F and x any trajectory solution of:

$$\ddot{x}(t) + \alpha\dot{x}(t) + \nabla F(x) = g(t),$$

where: $\alpha > 0$. We introduce the following Lyapunov energy:

$$\mathcal{G}(t) = \mathcal{E}(t) + \int_t^T \langle \lambda(x(s) - x^*) + (1+\beta)\dot{x}(s)), g(s)\rangle ds \tag{60}$$

where the energy $\mathcal{E}$ is defined as in the non-perturbed case by:

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

Differentiating the energy $\mathcal{G}$, we obtain the following differential inequality on which relies the whole proof of Theorem 3:

$$\begin{aligned} \mathcal{G}'(t) &= \mathcal{E}'(t) - \langle \lambda(x(t) - x^*) + (1+\beta)\dot{x}(t), g(t)\rangle \\ &\leqslant -\lambda\mathcal{E}(t) + \frac{\lambda}{2}(\beta L + \xi + \lambda^2 - \mu)\|x(t) - x^*\|^2 + \left(\frac{3}{2}\lambda - \alpha + \beta\left(\frac{\lambda}{2} - \alpha\right)\right)\|\dot{x}(t)\|^2 \\ &\quad +(\xi + \lambda(2\lambda - \alpha))\langle \dot{x}(t), x(t) - x^*\rangle. \end{aligned}$$

The parameters β, λ and ξ are chosen as in the unperturbed case:

- If $\alpha \leqslant 3\sqrt{\frac{\mu}{2}}$, we then choose:

  $$\beta = 0, \quad \lambda = \frac{2\alpha}{3}, \quad \xi = -\frac{\lambda^2}{2}$$

  for the class of quasi-strongly convex functions, and:

  $$\beta = \frac{\mu}{L} - \frac{2\alpha^2}{9L}, \quad \lambda = 2\alpha\frac{1+\beta}{3+\beta} = \frac{2}{3}\alpha\left(1 + \frac{2}{3}\frac{9\mu - 2\alpha^2}{9L + 3\mu - \frac{2}{3}\alpha^2}\right), \quad \xi = -\frac{\lambda^2}{2}(1-\beta) - \alpha\beta\lambda$$

  for the class of quasi-strongly convex functions having a Lipschitz continuous gradient.

- If $\alpha > 3\sqrt{\frac{\mu}{2}}$, we then choose:

$$\beta = 0, \quad \lambda = \alpha - \sqrt{\alpha^2 - 4\mu}, \quad \xi = -\frac{\lambda^2}{2}.$$

In both cases, with these choices of parameters and as shown in the proofs of Theorems 1 and 2, we have:

$$\begin{aligned}
\forall t \geqslant t_0,\ \mathcal{E}(t) &= (1+\beta)(F(x(t)) - F^*) + \frac{1}{2}(\xi + \frac{\beta}{1+\beta}\lambda^2)\|x(t) - x^*\|^2 \\
&\quad + \frac{1}{2(\beta+1)}\left\|\lambda(x(t) - x^*) + (\beta+1)\dot{x}(t)\right\|^2 \\
&\geqslant \frac{1}{2(\beta+1)}\left\|\lambda(x(t) - x^*) + (\beta+1)\dot{x}(t)\right\|^2 \qquad (61)
\end{aligned}$$

In particular we deduce that for all $t \geqslant t_0$, $\mathcal{G}'(t) \leqslant -\lambda\mathcal{E}(t) \leqslant 0$. The energy $\mathcal{G}$ is so non-increasing, hence: $\forall t \geqslant t_0,\ \mathcal{G}(t) \leq \mathcal{G}(t_0)$, i.e.:

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

Coming back now to the inequality (61), we have:

$$\begin{aligned}
\frac{1}{2}\|(1+\beta)\dot{x}(t) + \lambda(x(t) - x^*)\|^2 &\leqslant (1+\beta)\mathcal{E}(t) \leqslant (1+\beta)\mathcal{E}(t_0) \\
&\quad + \int_{t_0}^{t} (1+\beta)\|g(s)\|\|(1+\beta)\dot{x}(s) + \lambda(x(s) - x^*)\|ds
\end{aligned}$$

Applying the Grönwall-Bellman Lemma [12, Lemma A.5], we obtain:

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

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

$$\sup_{t \geqslant t_0} \|(1+\beta)\dot{x}(t) + \lambda(x(t) - x^*)\| \leqslant \sqrt{2(1+\beta)\mathcal{E}(t_0)} + (1+\beta)\int_{t_0}^{+\infty} \|g(s)\|ds < +\infty.$$

We set: $A = \sqrt{2(1+\beta)\mathcal{E}(t_0)} + (1+\beta)I_0$ where $I_0 = \int_{t_0}^{+\infty} \|g(s)\|ds$. The differential inequality $\forall t \geqslant t_0,\ \mathcal{G}'(t) \leqslant -\lambda\mathcal{E}(t) \leqslant 0$ then becomes:

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

Integrating between t_0 and t , we finally obtain:

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

Hence: $\forall t \geqslant t_0,\ \mathcal{E}(t) \leqslant C_0 e^{-\lambda(t-t_0)}$ where: $C_0 = \mathcal{E}(t_0) + (\sqrt{2(1+\beta)\mathcal{E}(t_0)} + (1+\beta)I_0)J_0 e^{-\lambda t_0}$. Combining the very last inequality with Lemma 4 and the control on the values provided in the proofs of Theorems 1 and 2, we finally obtained the expected inequalities. More precisely:

- If $\alpha > 3\sqrt{\frac{\mu}{2}}$ then $\beta = 0$, $\lambda = \alpha - \sqrt{\alpha^2 - 4\mu}$ and:

$$\forall t \geqslant t_0,\ F(x(t)) - F^* \leqslant \frac{2\mu}{2\mu - (\alpha - \sqrt{\alpha^2-4\mu})^2}\mathcal{E}(t) \leqslant \frac{2\mu C_0}{2\mu - (\alpha - \sqrt{\alpha^2-4\mu})^2}e^{-\lambda(t-t_0)}.$$

- If $\alpha \leqslant 3\sqrt{\frac{\mu}{2}}$, using the same argument than in Theorem 1, we can prove that:

$$\forall t \geqslant t_0,\ \|x(t) - x^*\| \leqslant \frac{3+\beta}{\alpha(1-\beta)}\sqrt{\frac{2}{1+\beta}\mathcal{E}(t)}$$

  so that:

$$\forall t \geqslant t_0,\ (1+\beta)(F(x(t)) - F^*) \quad \leqslant \quad \mathcal{E}(t) + \frac{\lambda^2}{4}\|x(t) - x^*\|^2 \leqslant \left[1 + 2\frac{1+\beta}{(1-\beta)^2}\right]\mathcal{E}(t)$$

  which concludes the proof.

6.3 Proof of Theorem 5

The proof of Theorem 5 relies on the the inequality (47) established in the proof of Theorem 1 for μ-quasi-strongly convex functions:

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

Observe that if $F(x) = |x| + \frac{\mu}{2}|x|^2$ this inequality is actually an equality:

$$\mathcal{E}'(t) + \lambda\mathcal{E}(t) = \frac{\lambda}{2}(\xi + \lambda^2 - \mu)|x(t)|^2 + \left(\frac{3}{2}\lambda - \alpha\right)|\dot{x}(t)|^2 + (\xi + \lambda(2\lambda - \alpha))\dot{x}(t)x(t)\rangle \tag{62}$$

since for this function, we actually have for all $u \in \partial F(x)$

$$\langle u, x - x^*\rangle = F(x) - F^* + \frac{\mu}{2}\left\|x - x^*\right\|^2.$$

If we choose $\lambda = \frac{2\alpha}{3}$ and $\xi = -\frac{\lambda^2}{2}$ we get

$$\mathcal{E}'(t) + \lambda\mathcal{E}(t) = \frac{\lambda}{2}(\frac{\lambda^2}{2} - \mu)|x(t)|^2. \tag{63}$$

From (63) we deduce that for any $\delta > \lambda$

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

It follows that

$$\mathcal{E}'(t) + \delta\mathcal{E}(t) \geqslant \frac{\lambda}{2}(\lambda^2 - \mu - \frac{\delta\lambda}{2})|x(t)|^2 + (\delta - \lambda)|x(t)|.$$

Since $x(t) \to 0$ when $t \to +\infty$, it follows that it exists $t_1 > t_0$ such that $\mathcal{E}(t_1) > 0$ and such that for any $t > t_1$,

$$\mathcal{E}'(t) + \delta\mathcal{E}(t) \geqslant 0 \tag{64}$$

and thus $e^{\delta t}\mathcal{E}(t) \geqslant e^{\delta t_1}\mathcal{E}(t_1) > 0$. It follows that for any $t > t_1$

$$e^{\delta t}\left(|x(t)| + \frac{1}{2}|\lambda x(t) + \dot{x}(t)|^2\right) \geqslant e^{\delta t_1}\mathcal{E}(t_1) > 0.$$

Setting $y(t) := e^{\delta t}x(t)$ we have

$$|y(t)| + \frac{1}{2}|(\lambda - \delta)y(t) + \dot{y}(t)|^2 e^{-\delta t} \geqslant e^{\delta t_1}\mathcal{E}(t_1) > 0. \tag{65}$$

Let us define: $\mathcal{H}(t) := e^{\delta t}\mathcal{E}(t)$. Hence, if it exists t_2 large enough such that for all $t \geqslant t_2$, $|y(t)| \leqslant \frac{1}{2}\mathcal{H}(t_1)$. Then there exists $t_3 \geqslant t_2$ such that:

$$\forall t \geqslant t_2,\ |\dot{y}(t)| \geqslant \sqrt{\mathcal{H}(t_1)}e^{\frac{\delta}{2}t} - \frac{\delta - \lambda}{2}\mathcal{H}(t_1) > 0.$$

Since $y \in C^1$ (thanks to Corollary 1), y is a continuous function of time. It follows that the sign of $\dot{y}(t)$ is constant on $[t_3, +\infty[$. If $\dot{y}(t_3) > 0$ then:

$$\forall t \geqslant t_3,\ y(t) \geqslant y(t_3) + \int_{t_3}^{t}\left(\sqrt{\mathcal{H}(t_1)}e^{\frac{\delta}{2}u} - \frac{\delta - \lambda}{2}\mathcal{H}(t_1)\right)du$$

which is impossible. If $\dot{y}(t_3) < 0$ then for all $t \geqslant t_3$,

$$y(t) \leqslant y(t_3) - \int_{t_3}^{t}\left(\sqrt{\mathcal{H}(t_1)}e^{\frac{\delta}{2}u} - \frac{\delta - \lambda}{2}\mathcal{H}(t_1)\right)du,$$

which is also impossible. It follows that $y(t)$ cannot tend to 0 when $t \to \infty$ which concludes the proof.

6.4 Proof of Theorem 6

The proof of Theorem 6 is based on a Lyapunov analysis inspired by the one proposed for the ODE. To study the ODE we used to following Lyapunov energy :

$$\mathcal{E}(t) = F(x(t)) - F^* + \frac{1}{2}\|\lambda(x(t) - x^*) + \dot{x}(t)\|^2 - \frac{\lambda^2}{4}\|x(t) - x^*\|^2. \tag{66}$$

To study the properties of the scheme (38) we define the sequence L_n

$$L_n := L(x_n, v_n) := F(x_n) - F^* + \frac{1}{2}\|\lambda(x_n - x^*) + (1+\lambda s)v_n\|^2 - \frac{\lambda^2}{4}\|x_n - x^*\|^2 \tag{67}$$

which can be seen as particular discretization of $\mathcal{E}$. To simplify the writing of the proof, we will use the following notations

$$t = \lambda s, \quad u = \lambda(x_{n+\frac{1}{2}} - x^*), \quad v = v_{n+\frac{1}{2}} \text{ and } g = \frac{1}{\lambda}\nabla F(x_{n+\frac{1}{2}}). \tag{68}$$

With these reduced notations, the scheme may be written:

$$\begin{cases} \lambda x_{n+\frac{1}{2}} &= \lambda x_n + t v_n \\ v := v_{n+\frac{1}{2}} &= (1+\frac{3t}{2})^{-1}(v_n - tg) \\ \lambda x_{n+1} &= \lambda x_{n+\frac{1}{2}} - t^2 g \\ v_{n+1} &= v + (1+t)^{-1}t^2 g. \end{cases} \tag{69}$$

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

$$\|\lambda(x_{n+1} - x^*) + (1+t)v_{n+1}\|^2 = \|u + (1+t)v\|^2. \tag{70}$$

Let us first compute $L_{n+\frac{1}{2}}$ and L_n using the reduced notations. We have:

$$L_{n+\frac{1}{2}} := L(x_{n+\frac{1}{2}}, v_{n+\frac{1}{2}}) := F(x_{n+\frac{1}{2}}) - F^* + \frac{1}{2}\|u + (1+t)v\|^2 - \frac{1}{4}\|u\|^2 \tag{71}$$

and L_n can be written in the following way as:

$$\begin{aligned} L_n &= F(x_n) - F^* + \frac{1}{2}\|u + \lambda(x_n - x_{n+\frac{1}{2}}) + (1+t)v_n\|^2 - \frac{1}{4}\|u + \lambda(x_n - x_{n+\frac{1}{2}})\|^2 \\ &= F(x_n) - F^* + \frac{1}{2}\left\|u + v_n\right\|^2 - \frac{1}{4}\left\|u - tv_n\right\|^2. \end{aligned}$$

Now, noticing that:

$$\frac{-\lambda^2}{4}\|x_{n+1} - x^*\|^2 = \frac{-\lambda^2}{4}\|x_{n+1} - x_{n+\frac{1}{2}} + x_{n+\frac{1}{2}} - x^*\|^2 = -\frac{1}{4}\|u\|^2 - \frac{t^4}{4}\|g\|^2 + \frac{t^2}{2}\langle u, g\rangle \tag{72}$$

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

$$L_{n+1} = L_{n+\frac{1}{2}} + F(x_{n+1}) - F(x_{n+\frac{1}{2}}) - \frac{t^4}{4}\|g\|^2 + \frac{t^2}{2}\langle u, g\rangle. \tag{73}$$

To prove Theorem 3 we demonstrated that the Lyapunov Energy defined by (66) satisfies

$$\mathcal{E}'(t) + \lambda\mathcal{E}(t) \leqslant 0. \tag{74}$$

To prove Theorem 6 we will use the following Lemma whose proof is left to Subsection 6.5:

Lemma 6.

$$\left(1 + t - \frac{3}{2}t^2\right)L_{n+1} - L_n \leq \frac{t}{4}\left(1 - \frac{2\mu}{\lambda^2} + 4t\right)\|u\|^2. \tag{75}$$

Observe that since: $t = \lambda s = \frac{\lambda}{\sqrt{L}} \leqslant \frac{\lambda}{\sqrt{\mu}}$, then

$$1 - \frac{2\mu}{\lambda^2} + 4t \quad \leqslant \quad 1 - \frac{2\mu}{\lambda^2} + 4\frac{\lambda}{\sqrt{\mu}} = 1 - 2\left(\frac{\sqrt{\mu}}{\lambda}\right)^2 + 4\frac{\lambda}{\sqrt{\mu}}. \tag{76}$$

Let $x = \frac{\sqrt{\mu}}{\lambda}$. Since $\lambda < \sqrt{2\mu}$, we have: $x > \frac{1}{\sqrt{2}}$. We easily prove that the quantity $2x^3 - x - 4$ is non negative for any $x > \frac{1}{\sqrt{2}}$, i.e.:

$$1 - \frac{2\mu}{\lambda^2} + 4t \leqslant 0$$

without any further assumption. It follows that for all $n \geqslant 0$ we have:

$$L_n \leqslant \left(1 + \frac{\lambda}{\sqrt{L}} - \frac{3\lambda^2}{2L}\right)^{-n} L_0. \tag{77}$$

Using the μ-strong convexity of F we get

$$\left(1 - \frac{\lambda^2}{2\mu}\right)(F(x_n) - F^*) \leqslant F(x_n) - F^* - \frac{\lambda^2}{4}\|x_n - x^*\|^2 \leqslant L_n \tag{78}$$

which ends the proof of Theorem 6.

6.5 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 (79) 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 (79) we give an upper bound of $L_{n+1} - L_n$.
4. We deduce a bound on $(1+t)L_{n+1} - L_n$ 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 (79)

$$F(x_{n+1}) - F(y) \leq \lambda\langle g, x_{n+\frac{1}{2}} - y\rangle - \frac{\mu}{2}\|x_{n+\frac{1}{2}} - y\|^2 - \frac{t^2}{2}\|g\|^2 \tag{79}$$

which is a key inequality used by Siegel in [29].
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

$$F(x_{n+1}) - F(x_{n+\frac{1}{2}}) \leqslant -\frac{t^2}{2}\|g\|^2. \tag{80}$$

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

$$F(x_{n+\frac{1}{2}}) - F(y) \leqslant \lambda \langle g, x_{n+\frac{1}{2}} - y \rangle - \frac{\mu}{2} \|x_{n+\frac{1}{2}} - y\|^2. \tag{81}$$

Inequality (79) holds summing the two previous inequalities.
Step 2.

$$L_{n+\frac{1}{2}} - L_n = F(x_{n+\frac{1}{2}}) - F(x_n) + \frac{1}{2}\|u + (1+t)v\|^2 - \frac{1}{2}\|u + v_n\|^2 - \frac{1}{4}\|u\|^2 + \frac{1}{4}\|u - tv_n\|^2$$

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

$$\frac{1}{2}\|A\|^2 - \frac{1}{2}\|B\|^2 = \langle a, B \rangle + \langle b, A \rangle + \frac{1}{2}\|a\|^2 - \frac{1}{2}\|b\|^2 \tag{82}$$

with

$$\begin{array}{rclrcl} A &=& u + (1+t)v, & B &=& u + v_n \\ a &=& -tg, & b &=& -\frac{t}{2}v. \end{array}$$

We thus get with this identity:

$$\begin{aligned} \frac{1}{2}\|u + (1+t)v\|^2 - \frac{1}{2}\|u + v_n\|^2 &= -t\langle g, u + v_n \rangle - \frac{t}{2}\langle v, u + (1+t)v \rangle + \frac{t^2}{2}\|g\|^2 - \frac{t^2}{8}\|v\|^2 \\ &= -t\langle g, v_n \rangle - t\langle g, u \rangle - \frac{t}{2}\langle v, u \rangle - \frac{t(4+5t)}{8}\|v\|^2 + \frac{t^2}{2}\|g\|^2. \end{aligned}$$

Moreover observe that:

$$-\frac{1}{4}\|u\|^2 + \frac{1}{4}\|u - tv_n\|^2 = -\frac{t}{2}\langle v_n, u \rangle + \frac{t^2}{4}\|v_n\|^2 \tag{83}$$

Hence:

$$\begin{aligned} L_{n+\frac{1}{2}} - L_n &= F(x_{n+\frac{1}{2}}) - F(x_n) - \langle tg, v_n \rangle - t\langle g, u \rangle + \frac{t^2}{2}\|g\|^2 \\ &\quad -\frac{t}{2}\langle v, u \rangle - \frac{t(4+5t)}{8}\|v\|^2 - \frac{t}{2}\langle v_n, u \rangle + \frac{t^2}{4}\|v_n\|^2 \end{aligned}$$

Using the expression of v_n in $\langle v_n, u \rangle$, we get:

$$\begin{aligned} L_{n+\frac{1}{2}} - L_n &= F(x_{n+\frac{1}{2}}) - F(x_n) - \langle tg, v_n \rangle - t\left(1 + \frac{t}{2}\right)\langle g, u \rangle + \frac{t^2}{2}\|g\|^2 \\ &\quad -t\left(1 + \frac{3t}{4}\right)\langle v, u \rangle - \frac{t(4+5t)}{8}\|v\|^2 + \frac{t^2}{4}\|v_n\|^2 \end{aligned}$$

Step 3. Using (73), we get:

$$\begin{aligned} L_{n+1} - L_n &= L_{n+1} - L_{n+\frac{1}{2}} + L_{n+\frac{1}{2}} - L_n \\ &= F(x_{n+1}) - F(x_n) - t\langle g, v_n \rangle + \frac{t^2}{2}\|g\|^2 - t\langle g, u \rangle \\ &\quad -t\left(1 + \frac{3t}{4}\right)\langle v, u \rangle - \frac{t(4+5t)}{8}\|v\|^2 + \frac{t^2}{4}\|v_n\|^2 - \frac{t^4}{4}\|g\|^2 \end{aligned} \tag{84}$$

Then, we apply (79) with $y = x_n$ and $tv_n = \lambda(x_{n+\frac{1}{2}} - x_n)$ to get :

$$F(x_{n+1}) - F(x_n) - t\langle g, v_n\rangle \leq -\frac{t^2}{2}\|g\|^2 - \frac{t^2\mu}{2\lambda^2}\|v_n\|^2 \tag{85}$$

and (79) with $y = x^*$ to get

$$\langle g, u\rangle \geq F(x_{n+1}) - F^* + \frac{\mu}{2\lambda^2}\|u\|^2 + \frac{t^2}{2}\|g\|^2. \tag{86}$$

Combining (84), (85) and (86) we deduce :

$$\begin{aligned} L_{n+1} - L_n \quad \leq \quad & -\frac{t^2\mu}{2\lambda^2}\|v_n\|^2 - t(F(x_{n+1}) - F^*) - \frac{t\mu}{2\lambda^2}\|u\|^2 - \frac{t^3}{2}\|g\|^2 \\ & -t\left(1 + \frac{3t}{4}\right)\langle v, u\rangle - \frac{t(4+5t)}{8}\|v\|^2 + \frac{t^2}{4}\|v_n\|^2 - \frac{t^4}{4}\|g\|^2 \end{aligned}$$

As in the continuous case, we assume that: $\lambda \leq \sqrt{2\mu}$ so that $-\frac{t^2\mu}{2\lambda^2} + \frac{t^2}{4} \leq 0$. We thus deduce that:

$$\begin{aligned} L_{n+1} - L_n \quad \leq \quad & -t(F(x_{n+1}) - F^*) - \frac{t\mu}{2\lambda^2}\|u\|^2 - \frac{t^3}{2}\|g\|^2 \\ & -t\left(1 + \frac{3t}{4}\right)\langle v, u\rangle - \frac{t(4+5t)}{8}\|v\|^2 - \frac{t^4}{4}\|g\|^2 \end{aligned}$$

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

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

we eventually get:

$$\begin{aligned} (1+t)L_{n+1} - L_n \quad \leq \quad & \frac{t}{4}\left(1 - \frac{2\mu}{\lambda^2}\right)\|u\|^2 + \frac{t^2}{8}(3+4t)\|v\|^2 - \frac{t^3}{4}(2 + t - t^2)\|g\|^2 \\ & + \frac{t^2}{4}\langle u, v\rangle + \frac{t^3}{2}\langle u, g\rangle. \end{aligned}$$

We now use the inequality:

$$|\langle tg, u\rangle| \leq \frac{t^2}{2}\|g\|^2 + \frac{1}{2}\|u\|^2, \tag{87}$$

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

$$(1+t)L_{n+1} - L_n \leq \frac{t}{4}\left(1 - \frac{2\mu}{\lambda^2} + t\right)\|u\|^2 + \frac{t^2}{8}(3+4t)\|v\|^2 - \frac{t^3}{4}(2-t^2)\|g\|^2 + \frac{t^2}{4}\langle u,v\rangle \tag{88}$$

And thus, since $t \geq 0$ and $t = \frac{\lambda}{\sqrt{L}} \leqslant \sqrt{\frac{2\mu}{L}} \leqslant \sqrt{2}$:

$$(1+t)L_{n+1} - L_n \leq \frac{t}{4}\left(1 - \frac{2\mu}{\lambda^2} + t\right)\|u\|^2 + \frac{t^2}{8}(3+4t)\|v\|^2 + \frac{t^2}{4}\langle u,v\rangle \tag{89}$$

so that:

$$(1+t)L_{n+1} - L_n \leq \frac{t}{4}\left(1 - \frac{2\mu}{\lambda^2}\right)\|u\|^2 + \frac{t^2}{4}\left(\|u\|^2 + \frac{3+4t}{2}\|v\|^2 + \langle u,v\rangle\right). \tag{90}$$

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:*

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

Let us apply Lemma 7 with $x = \sqrt{1+t}v$ and $y = \frac{1}{\sqrt{1+t}}u$. We get:

$$(1+t)A\left\|v\right\|^2 + \langle u,v\rangle \leqslant \frac{2A}{1+t}\left\|u+(1+t)v\right\|^2 + \frac{2A}{1+t}\left\|u\right\|^2 \tag{91}$$

Choosing:

$$A(t) = \frac{3+4t}{2(t+1)} \tag{92}$$

which actually satisfies: $A(t) \geq \frac{1}{4}$ for any $t \geq 0$, we thus deduce that:

$$\begin{aligned} \frac{3+4t}{2}\|v\|^2 + \langle u,v\rangle &\leq \frac{3+4t}{(1+t)^2}\left\|u+(1+t)v\right\|^2 + \frac{3+4t}{(1+t)^2}\left\|u\right\|^2 \\ &\leq 3\left\|u+(1+t)v\right\|^2 + 3\left\|u\right\|^2. \end{aligned}$$

Hence:

$$(1+t)L_{n+1} - L_n \leq \frac{t}{4}\left(1 - \frac{2\mu}{\lambda^2}\right)\|u\|^2 + \frac{t^2}{4}\left(4\|u\|^2 + 3\left\|u+(1+t)v\right\|^2\right). \tag{93}$$

Moreover:

$$\begin{aligned} \frac{1}{2}\|u+(1+t)v\|^2 &= L_{n+1} - (F(x_{n+1}) - F^*) + \frac{\lambda^2}{4}\|x_{n+1} - x^*\|^2 \\ &\leq L_{n+1} + \frac{\lambda^2 - 2\mu}{4}\|x_{n+1} - x^*\|^2 \leq L_{n+1} \end{aligned}$$

using the μ-strong convexity of F and the fact that $\lambda^2 - 2\mu \leq 0$. We finally get:

$$\left(1 + t - \frac{3}{2}t^2\right)L_{n+1} - L_n \leq \frac{t}{4}\left(\left(1 - \frac{2\mu}{\lambda^2}\right) + 4t\right)\|u\|^2. \tag{94}$$

6.6 Proof of Theorem 7

The proof is essentially similar to the one of Theorem 6. The careful reader may have remarked that the only property of g that is used in the proof of Theorem 6 is the inequality (79) we recall here

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

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 (43) satisfies this inequality under the hypothesis of the Theorem 7, see also [29, Lemma 4.2] :

Lemma 8. *If $F = f + h$ is convex, if $f \in \mathcal{S}^{1,1}_{\mu,L}$, if h is convex, proper and lower semi-continuous and $s = \frac{1}{\sqrt{L}}$ then for all $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$:*

$$F(Tx) - F(y) \leq \frac{1}{s^2}\langle x - Tx, x - y\rangle - \frac{\mu}{2}\|x - y\|^2 - \frac{1}{2s^2}\|Tx - x\|^2 \tag{95}$$

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

$$h(Tx) - h(y) \leqslant \langle\frac{x - Tx}{s^2} - \nabla f(x), Tx - y\rangle \tag{96}$$

Since ∇f is $\frac{1}{s^2}$-Lipschitz

$$f(Tx) - f(x) \leqslant \langle\nabla f(x), Tx - x\rangle + \frac{1}{2s^2}\|Tx - x\|^2. \tag{97}$$

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

$$f(x) - f(y) \leqslant \langle\nabla f(x), x - y\rangle - \frac{\mu}{2}\|x - y\|^2. \tag{98}$$

Adding the three last inequalities we get :

$$F(Tx) - F(y) \leqslant \frac{1}{s^2}\langle x - Tx, Tx - y\rangle - \frac{\mu}{2}\|x - y\|^2 + \frac{1}{2s^2}\|Tx - x\|^2. \tag{99}$$

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

$$F(Tx) - F(y) \leqslant \frac{1}{s^2}\langle x - Tx, x - y\rangle - \frac{\mu}{2}\|x - y\|^2 - \frac{1}{2s^2}\|Tx - x\|^2. \tag{100}$$

□

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 7 :

$$\forall y \in \mathbb{R}^n, \quad F(x_{n+1}) - F(y) \leq \lambda\langle\tilde{g}, x_{n+\frac{1}{2}} - y\rangle - \frac{\mu}{2}\|x_{n+\frac{1}{2}} - y\|^2 - \frac{t^2}{2}\|\tilde{g}\|^2. \tag{101}$$

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] 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.

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

[3] 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.

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

[5] H. Attouch, X. Goudou, and P. Redont. The heavy ball with friction method, I. The continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. *Communications in Contemporary Mathematics*, 2(1):1–34, 2000.

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

[7] 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.

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

[9] 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.

[10] 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.

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

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

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

[20] 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.

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

[22] 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.

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

[24] 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.

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

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

[27] O. Sebbouh, Ch. Dossal, and A. Rondepierre. Convergence rates of damped inertial dynamics under geometric conditions. *SIAM Journal on Optimization*, To appear 2020.

[28] B. Shi, S.S. Du, M.I. Jordan, and W. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:1810.08907*, 2018.

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

[30] 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.