OPTIMAL CONVERGENCE RATES FOR NESTEROV ACCELERATION

Jean François Aujol, Charles H Dossal, Aude Rondepierre

OPTIMAL CONVERGENCE RATES FOR NESTEROV ACCELERATION

J-F. AUJOL [1], CH. DOSSAL [2] AND A. RONDEPIERRE [2,3]

[1] UNIV. BORDEAUX, BORDEAUX INP, CNRS, IMB, UMR 5251, F-33400 TALENCE, FRANCE.
[2] IMT, UNIV. TOULOUSE, INSA TOULOUSE, FRANCE.
[3] LAAS, UNIV. TOULOUSE, CNRS, TOULOUSE, FRANCE.

JEAN-FRANCOIS.AUJOL@MATH.U-BORDEAUX.FR,

{CHARLES.DOSSAL,AUDE.RONDEPIERRE}@INSA-TOULOUSE.FR

Abstract. In this paper, we study the behavior of solutions of the ODE associated to Nesterov acceleration. It is well-known since the pioneering work of Nesterov that the rate of convergence $O(1/t^2)$ is optimal for the class of convex functions. In this work, we show that better convergence rates can be obtained with some additional geometrical conditions, such as Łojasiewicz property. More precisely, we prove the optimal convergence rates that can be obtained depending on the geometry of the function F to minimize. The convergence rates are new, and they shed new light on the behavior of Nesterov acceleration schemes. We prove in particular that the classical Nesterov scheme may provide convergence rates that are worse than the classical gradient descent scheme on sharp functions: for instance, the convergence rate for strongly convex functions is not geometric for the classical Nesterov scheme (while it is the case for the gradient descent algorithm). This shows that applying the classical Nesterov acceleration on convex functions without looking more at the geometrical properties of the objective functions may lead to sub-optimal algorithms.

Key-words. Lyapunov functions, rate of convergence, ODEs, optimization, Łojasiewicz property.

1. Introduction. The motivation of this paper lies in the minimization of a differentiable function F with at least one minimizer. Inspired by Nesterov pioneering work [25], we study the following ODE

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla F(x(t)) = 0 \tag{1.1}$$

where $\alpha > 0$, with $t_0 > 0$, $x(t_0) = x_0$ and $\dot{x}(t_0) = v_0$. This ODE is associated to FISTA [12] or Accelerated Gradient Method [25] :

$$x_{n+1} = y_n - h\nabla F(y_n) \text{ and } y_n = x_n + \frac{n}{n+\alpha}(x_n - x_{n-1}) \tag{1.2}$$

with h and α positive parameters. This equation, including or not a perturbation term, has been widely studied in the literature [7, 27, 17, 11, 24]. This equation belongs to a set of similar equations with various viscosity terms. It is impossible to mention all works related to the heavy ball equation or other viscosity terms. We refer the reader to the following recent works [13, 21, 24, 18, 4, 26, 3] and the references that can be found in these articles.

It was proved in [5] that if F is convex with Lipschitz gradient and if $\alpha > 3$, the trajectory $F(x(t))$, where x is the solution of (1.1), converges to the minimum F^* of F. It is also known that for $\alpha \geqslant 3$ and F convex we have

$$F(x(t)) - F^* = O\left(t^{-2}\right) \tag{1.3}$$

Extending to the continuous setting the proof of Chambolle-Dossal [19] of the convergence of iterates of FISTA, Attouch-Chbani-Peypouquet-Redont [5] proved that for $\alpha > 3$ the trajectory x weakly converges to a minimizer of F. Su et al. [27] proposed some new results, proving the integrability of $t \mapsto t(F(x(t)) - F^*)$ when $\alpha > 3$, and they gave more accurate bounds on $F(x(t)) - F^*$ in the case of strong convexity. Always in the case of the strong convexity of F, Attouch, Chbani, Peypouquet and Redont proved in [5] that the trajectory $x(t)$ satisfies

$F(x(t)) - F^* = O\left(t^{-\frac{2\alpha}{3}}\right)$ for any $\alpha > 0$. More recently several studies including a perturbation term [5, 10, 9, 1] have been proposed.

In this work, we focus on the decay of $F(x(t)) - F^*$ depending on more general geometries of F around its set of minimizers than strong convexity. Roughly speaking, we consider functions behaving like $\|x - x^*\|^\gamma$ around the minimizer for any $\gamma \geqslant 1$. Our aim is to show the optimal convergence rates that can be obtained depending on this local geometry. In particular we prove that if F is strongly convex with a Lipschitz continuous gradient, the decay is actually better than $O\left(t^{-\frac{2\alpha}{3}}\right)$. We also prove that the actual decay for quadratic functions is $O\left(t^{-\alpha}\right)$. These results rely on two geometrical conditions: a first one ensuring that the function is sufficiently flat around the set of minimizers, and a second one ensuring that it is sufficiently sharp.

The paper is organized as follows. In Section 2, we introduce the geometrical hypotheses we consider for the function F, and their relation with Łojasiewicz property. We then recap the state of the art results on the ODE (1.1) in Section 3. We present the contributions of the paper in Section 4: depending on the geometry of the function F and the value of the damping parameter α, we give optimal rates of convergence. The proofs of the theorems are given in Section 5. Some technical proofs are postponed to Appendix A.

2. Local geometry of convex functions. In this section we introduce two notions describing the geometry of a convex function around its minimizers.

DEFINITION 2.1. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function, $X^* := \operatorname{argmin} F \neq \emptyset$ and: $F^* := \inf F$.*

(i) Let $\gamma \geqslant 1$. The function F satisfies the hypothesis $\boldsymbol{H}_1(\gamma)$ if, for any critical point $x^* \in X^*$, there exists $\eta > 0$ such that:

$$\forall x \in B(x^*, \eta), \quad 0 \leqslant F(x) - F^* \leqslant \frac{1}{\gamma}\langle \nabla F(x), x - x^* \rangle.$$

(ii) Let $r \geqslant 1$. The function F satisfies the growth condition $\boldsymbol{H}_2(r)$ if for any critical point $x^* \in X^*$, there exists $K > 0$ and $\varepsilon > 0$, such that:

$$\forall x \in B(x^*, \varepsilon), \quad K d(x, X^*)^r \leqslant F(x) - F^*.$$

The $\mathbf{H}_1(\gamma)$ hypothesis has already been used in [17] and later in [27, 10]. This is a mild assumption, requesting slightly more than the convexity of F in the neighborhood of its minimizers. In particular, observe that any convex function automatically satisfies $\mathbf{H}_1(1)$ and that any differentiable function F ensuring that $(F - F^*)^{\frac{1}{\gamma}}$ is convex for some $\gamma \geq 1$, satisfies $\mathbf{H}_1(\gamma)$. Nevertheless having a better intuition of the geometry of convex functions satisfying $\mathbf{H}_1(\gamma)$ for some $\gamma \geq 1$, requires a little more effort:

LEMMA 2.2. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function with $X^* = \operatorname{argmin} F \neq \emptyset$, and $F^* = inf\ F$. If F satisfies $\boldsymbol{H}_1(\gamma)$ for some $\gamma \geq 1$, then:*

1. *F satisfies $\boldsymbol{H}_1(\gamma')$ for all $\gamma' \in [1, \gamma]$.*
2. *For any minimizer $x^* \in X^*$, there exists $M > 0$ and $\eta > 0$ such that:*

$$\forall x \in B(x^*, \eta),\ F(x) - F^* \leqslant M\|x - x^*\|^\gamma. \tag{2.1}$$

Proof. The proof of the first point of Lemma 2.2 is straightforward. The second point relies on the following elementary result in dimension 1: let $g : \mathbb{R} \to \mathbb{R}$ be a convex differentiable function such that $0 \in \operatorname{argmin} g$, $g(0) = 0$ and:

$$\forall t \in [0, 1],\ g(t) \leq \frac{t}{\gamma} g'(t),$$

for some $\gamma \geqslant 1$. Then the function $t \mapsto t^{-\gamma} g(t)$ is monotonically increasing on $[0,1]$ and:

$$\forall t \in [0,1],\ g(t) \leqslant g(1) t^{\gamma}. \tag{2.2}$$

Consider now any convex differentiable function $F : \mathbb{R}^n \to \mathbb{R}$ satisfying the condition $\mathbf{H}_1(\gamma)$, and $x^* \in X^*$. So there exists $\eta > 0$ such that:

$$\forall x \in \bar{B}(x^*, \eta), \quad 0 \leqslant F(x) - F^* \leqslant \frac{1}{\gamma} \langle \nabla F(x), x - x^* \rangle.$$

For any $x \in \bar{B}(x^*, \eta)$ with $x \neq x^*$, we introduce the following univariate function:

$$g_x : t \in [0,1] \mapsto F\left(x^* + t\eta \frac{x - x^*}{\|x - x^*\|} \right) - F^*.$$

First, observe that, for all $x \in \bar{B}(x^*, \eta)$ with $x \neq x^*$ and for all $t \in [0,1]$, we have: $x^* + t\eta \frac{x - x^*}{\|x - x^*\|} \in \bar{B}(x^*, \eta)$. Since F is continuous on the compact set $\bar{B}(x^*, \eta)$, we deduce that:

$$\exists M > 0,\ \forall x \in \bar{B}(x^*, \eta) \ \text{ with } x \neq x^*,\ \forall t \in [0,1],\ g_x(t) \leq M. \tag{2.3}$$

Note here that the constant M only depends on the point x^*.

Then, by construction, g_x is a convex differentiable function satisfying: $0 \in \operatorname{argmin}(g_x)$, $g_x(0) = 0$ and:

$$\begin{aligned} \forall t \in (0,1],\ g_x'(t) &= \left\langle \nabla F\left(x^* + t\eta \frac{x - x^*}{\|x - x^*\|} \right), \eta \frac{x - x^*}{\|x - x^*\|} \right\rangle \\ &\geqslant \frac{\gamma}{t} \left(F\left(x^* + t\eta \frac{x - x^*}{\|x - x^*\|} \right) - F^* \right) = \frac{\gamma}{t} g_x(t) \end{aligned}$$

Thus, using the one dimensional result (2.2) and the uniform bound (2.3), we get:

$$\forall x \in \bar{B}(x^*, \eta) \ \text{ with } x \neq x^*,\ \forall t \in [0,1],\ g_x(t) \leqslant g_x(1) t^{\gamma} \leqslant M t^{\gamma} \tag{2.4}$$

Finally by choosing $t = \frac{1}{\eta} \|x - x^*\|$, we obtain the expected result. □

In other words, the hypothesis $\mathbf{H}_1(\gamma)$ can be seen as a "flatness" condition on the function F in the sense that it ensures that F is sufficiently flat (at least as flat as $x \mapsto |x|^{\gamma}$) in the neighborhood of its minimizers.

The hypothesis $\mathbf{H}_2(r)$, $r \geqslant 1$, is a growth condition on the function F around any critical point. It is sometimes also called r-conditioning [20] or Hölderian error bounds [16]. This assumption is motivated by the fact that, when F is convex, $\mathbf{H}_2(r)$ is equivalent to the famous Łojasievicz inequality [22, 23], a key tool in the mathematical analysis of continuous (or discrete) dynamical systems, with exponent $\theta = 1 - \frac{1}{r}$ [14, 15]:

DEFINITION 2.3. *A differentiable function $F : \mathbb{R}^n \to \mathbb{R}$ is said to have the Łojasiewicz property with exponent $\theta \in [0,1)$ if, for any critical point x^*, there exist $c > 0$ and $\varepsilon > 0$ such that:*

$$\forall x \in B(x^*, \varepsilon),\ \|\nabla F(x)\| \geqslant c |F(x) - F^*|^{\theta}. \tag{2.5}$$

where: $0^0 = 0$ when $\theta = 0$ by convention.

Observe that the inequality (2.5) is automatically satisfied at any non-critical point, so that the Łojasiewicz property is in fact a geometrical condition on the function F around any critical points. Moreover, when the set X^* of the minimizers is a connected compact set, the Łojasiewicz

inequality turns into a geometrical condition on F around X^* as stated in [2, Lemma 1] whose proof can be adapted to establish the following result:

LEMMA 2.4. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function satisfying the growth condition $\boldsymbol{H}_2(r)$ for some $r \geqslant 1$. Assume that the set $X^* = \operatorname{argmin} F$ is compact. Then there exist $K > 0$ and $\varepsilon > 0$ such that, for all $x \in \mathbb{R}^n$:*

$$d(x, X^*) \leqslant \varepsilon \Rightarrow K d(x, X^*)^r \leqslant F(x) - F^*.$$

Typical examples of functions having the Łojasiewicz property are real-analytic functions and C^1 subanalytic functions [22], or semialgebraic functions [2]. Strongly convex functions satisfy a global Łojasiewicz property with exponent $\theta = \frac{1}{2}$ [2], or equivalently a global version of the hypothesis $\mathbf{H}_2(2)$, namely:

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

where $\mu > 0$ denotes the parameter of strong convexity. By extension, uniformly convex functions of order $p \geqslant 2$ satisfy the global version of the hypothesis $\mathbf{H}_2(p)$ [20].

Let us now present two simple examples of convex differentiable functions to illustrate situations where the hypothesis $\mathbf{H}_1$ and $\mathbf{H}_2$ are satisfied. Consider the function $F : x \in \mathbb{R} \mapsto |x|^\gamma$ for some $\gamma > 1$. We easily check that F satisfies the hypothesis $\mathbf{H}_1(\gamma')$ for some $\gamma' \geq 1$ if and only if $\gamma' \in [1, \gamma]$. By definition, F also naturally satisfies $\mathbf{H}_2(r)$ if and only if $r \geqslant \gamma$. Same conditions on γ' and r can be derived without the uniqueness of the minimizer for functions of the form:

$$F(x) = \begin{cases} ||x| - a|^\gamma & \text{if } |x| \geqslant a, \\ 0 & \text{otherwise,} \end{cases} \tag{2.6}$$

with $a > 0$, and whose set of minimizers is: $X^* = [-a, a]$, since conditions $\mathbf{H}_1(\gamma)$ and $\mathbf{H}_2(r)$ only make sense around the extremal points of X^*.

Let us now investigate the relation between the parameters γ and r in the general case: any convex differentiable function F satisfying both $\mathbf{H}_1(\gamma)$ and $\mathbf{H}_2(r)$, has to be at least as flat as $x \mapsto \|x\|^\gamma$ and as sharp as $x \mapsto \|x\|^r$ in the neighborhood of its minimizers. Combining the flatness condition $\mathbf{H}_1(\gamma)$ and the growth condition $\mathbf{H}_2(r)$, we consistently deduce:

LEMMA 2.5. *If a convex differentiable function satisfies both $\boldsymbol{H}_1(\gamma)$ and $\boldsymbol{H}_2(r)$ then necessarily $r \geqslant \gamma$.*

Finally, we conclude this section by showing that an additional assumption of the Lipschitz continuity of the gradient provides additional information on the local geometry of F: indeed, for convex functions, the Lipschitz continuity of the gradient is equivalent to a quadratic upper bound on F:

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n,\ F(x) - F(y) \leqslant \langle \nabla F(y), x - y \rangle + \frac{L}{2} \|x - y\|^2. \tag{2.7}$$

Applying (2.7) at $y = x^*$, we then deduce:

$$\forall x \in \mathbb{R}^n,\ F(x) - F^* \leqslant \frac{L}{2} \|x - x^*\|^2, \tag{2.8}$$

which indicates that F is at least as flat as $\|x - x^*\|^2$ around X^*. More precisely:

LEMMA 2.6. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function with a L-Lipschitz continuous gradient for some $L > 0$. Assume also that F satisfies the growth condition $\boldsymbol{H}_2(2)$ for some constant $K > 0$. Then F automatically satisfies $\boldsymbol{H}_1(\gamma)$ with $\gamma = 1 + \frac{K}{2L} \in (1, 2]$.*

Proof. Since F is convex with a Lipschitz continuous gradient, we have:

$$\forall (x,y) \in \mathbb{R}^n, F(y) - F(x) - \langle \nabla F(x), y - x\rangle \geqslant \frac{1}{2L}\|\nabla F(y) - \nabla F(x)\|^2,$$

hence:

$$\forall x \in \mathbb{R}^n, F(x) - F^* \leqslant \langle \nabla F(x), x - x^*\rangle - \frac{1}{2L}\|\nabla F(x)\|^2.$$

Assume in addition that F satisfies the growth condition $\mathbf{H}_2(2)$ for some constant $K > 0$. Then F has the Łojasiewicz property with exponent $\theta = \frac{1}{2}$ and constant $c = \sqrt{K}$. Thus:

$$\left(1 + \frac{K}{2L}\right)|F(x) - F^*| \leqslant \langle \nabla F(x), x - x^*\rangle,$$

in the neighborhood of its minimizers, which means that F satisfies $\mathbf{H}_1(\gamma)$ with $\gamma = 1 + \frac{K}{2L}$. □

3. Related results. In this section, we recall some classical state of the art results related to the ODE (1.1).

Let us first recall that as soon as $\alpha > 0$, then $F(x(t))$ converges to F^* [10, 6]. As recalled in Section 1, a larger value of α is required to show the convergence of the trajectory $x(t)$.

If F is convex and $\alpha > 3$ or if F satisfies $\mathbf{H}_1(\gamma)$ hypothesis and $\alpha > 1 + \frac{2}{\gamma}$ then

$$F(x(t)) - F^* = o\left(\frac{1}{t^2}\right), \tag{3.1}$$

and the trajectory $x(t)$ (weakly in an infinite dimensional space) converges to a minimizer x^* of F [27, 10, 24]. This means that thanks to the additional hypothesis $\mathbf{H}_1(\gamma)$, the damping parameter α can be chosen smaller.

Now, if F satisfies $\mathbf{H}_1(\gamma)$ and $\alpha \leqslant 1 + \frac{2}{\gamma}$, then we can no longer prove the convergence of the trajectory $x(t)$ but we still have the following convergence rate for $F(x(t))$:

$$F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2\gamma\alpha}{2+\gamma}}}\right). \tag{3.2}$$

Moreover, this decay is optimal and reached for $F(x) = |x|^\gamma$ if $\gamma \geqslant 1$ (see [6] for with the sole assumption of convexity which corresponds to the case $\gamma = 1$ and [10] for the general case and the optimality of the rate). Notice that for convex functions, the decay is $O\left(\frac{1}{t^{\frac{2\alpha}{3}}}\right)$. We can also notice that the bound hidden in the big O is explicit and available also for $\gamma < 1$, that is for non convex functions (for example for functions whose square is convex).

If F is the function $F(x) = |x|^\gamma$, where $x \in \mathbb{R}$, with $\gamma > 2$ and $\alpha \geqslant \frac{\gamma+2}{\gamma-2}$, then the ODE (1.1) admits an explicit solution of the form $x(t) = \frac{K}{t^{\frac{2}{\gamma-2}}}$ [6]. This simple calculation will show the optimality of the convergence rate we obtain in Theorem 4.3.

Moreover, if F is strongly convex, then for any $\alpha > 0$ we have [27]

$$F(x(t)) - F^* = 0\left(\frac{1}{t^{\frac{2\alpha}{3}}}\right). \tag{3.3}$$

This shows that by assuming more on the geometry of the function F (in this case, strong convexity), better rates of convergence can be achieved.

Eventually several results about the convergence rate of the solutions of ODE associated to the classical gradient descent :

$$\dot{x}(t) + \nabla F(x(t)) = 0 \tag{3.4}$$

or the ODE associated to the heavy ball method

$$\ddot{x} + \alpha\dot{x}(t) + \nabla F(x(t)) = 0 \tag{3.5}$$

under geometrical conditions such that Łojasiewicz properties have been proposed, see for example Polyak-Shcherbakov [26]. The authors prove that if the function F satisfies $\mathbf{H}_2(2)$ and some other conditions, the decay of $F(x(t)) - F^*$ is exponential for the solutions of both previous equations. These rates are the continuous counterparts of the exponential decay rate of the classical gradient descent algorithm and the heavy ball method algorithm for strongly convex functions.

In the next section we will prove that this geometric rate is not true for solutions of (1.1) even for quadratic functions, and we will prove that from an optimization point of view, the classical Nesterov acceleration may be less efficient than the classical gradient descent.

4. Contributions. In this section, we state the optimal convergence rates that can be achieved when F satisfies hypotheses such as $\mathbf{H}_1(\gamma)$ and/or $\mathbf{H}_2(r)$. The first result gives optimal control for functions whose geometry is sharp :

THEOREM 4.1. *Let $\gamma \geqslant 1$ and $\alpha > 0$.*

1. *If F satisfies the hypothesis $\boldsymbol{H}_1(\gamma)$ and if $\alpha \leqslant 1 + \frac{2}{\gamma}$, then*

$$F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2\gamma\alpha}{\gamma+2}}}\right) \tag{4.1}$$

2. *If F satisfies the hypotheses $\boldsymbol{H}_1(\gamma)$ and $\boldsymbol{H}_2(2)$ and if F has a unique minimizer, if $\alpha > 1 + \frac{2}{\gamma}$ then*

$$F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2\gamma\alpha}{\gamma+2}}}\right) \tag{4.2}$$

Moreover this decay is optimal in the sense that for any $\gamma \in (1, 2]$ this rate is achieved for the function $F(x) = |x|^\gamma$.

Note that the first point of Theorem 4.1 is already proven in [10] and that the second point of Theorem 4.1 only applies for $\gamma \leqslant 2$, since there is no function that satisfies both conditions $\mathbf{H}_1(\gamma)$ with $\gamma > 2$ and $\mathbf{H}_2(2)$ (see Lemma 2.2). The optimality of the convergence rate result is precisely stated in the next Proposition:

PROPOSITION 4.2. *Let $\gamma \in (1, 2]$. Let us assume that $\alpha > 1 + \frac{2}{\gamma}$. Let x be a solution of (1.1) with $F(x) = |x|^\gamma$ with $t_0 > \sqrt{\frac{\alpha(\gamma+2-\alpha\gamma)}{(\gamma+2)^2}}$, $|x(t_0)| < 1$ and $\dot{x}(t_0) = 0$. There exists $K > 0$ such that for any $T > 0$, there exists $t \geqslant T$ such that*

$$F(x(t)) - F^* \geqslant \frac{K}{t^{\frac{2\gamma\alpha}{\gamma+2}}}. \tag{4.3}$$

Let us make several observations: first, to apply the second point of Theorem 4.1, more conditions are needed than for the first point: the hypothesis $\mathbf{H}_2(2)$ and the uniqueness of the minimizer are needed (only for the second point) to prove a decay faster than $O(\frac{1}{t^2})$, which is

the uniform rate than can be achieved with $\alpha \geqslant 3$ for convex functions [27]. The uniqueness of the minimizer in the second point is crucial in the proof, but it is still an open problem to know if this uniqueness is a necessary condition.

We can remark that if F is a quadratic function in the neighborhood of x^*, Theorem 4.1 applies with $\gamma = 2$ and thus

$$F(x(t)) - F^* = O\left(\frac{1}{t^\alpha}\right) \tag{4.4}$$

Likewise, if F is a convex differentiable function with a Lipschitz continuous gradient, and if F satisfies the growth condition $\mathbf{H}_2(2)$, then F automatically satisfies the $\mathbf{H}_1(\gamma)$ hypothesis with some $1 < \gamma \leqslant 2$ as shown in Section 2, and Theorem 4.1 applies with $\gamma > 1$. In both cases, we thus obtain convergence rates which are strictly better than $O\left(\frac{1}{t^{\frac{2\alpha}{3}}}\right)$ that is proposed for strongly convex functions by Su et al. [27] and Attouch-Chbani-Peypouquet-Redont [5]. Finally, we can remark that the decay for quadratic or strongly convex functions is not geometric, while it is the case for the classical gradient descent scheme (see e.g. [20]). This shows that applying the classical Nesterov acceleration on convex functions without looking more at the geometrical properties of the objective functions may lead to sub-optimal algorithms.

Remark 1 (The Least-Square problem).

1. *We remark that if $\dot{x}(t_0) = 0$, then for all $t \geqslant t_0$ we have that $x(t)$ belongs to $x_0 + Im(\nabla F)$ where $Im(\nabla F)$ stands for the vectorial space generated by $\nabla F(x)$ for all x in $\mathbb{R}^n$. As a consequence, Theorem 4.1 and 4.3 still hold true as long as the assumptions are valid in $x_0 + Im(\nabla F)$.*
2. *Let us consider the classical Least-Square problem defined by:*

$$\min_{x\in\mathbb{R}^n} F(x) := \frac{1}{2}\|Ax - b\|^2,$$

where A is a bounded linear operator and $b \in \mathbb{R}^n$. If $\dot{x}(t_0) = 0$, then for all $t \geqslant t_0$, we have thus that $x(t)$ belongs to the affine subspace $x_0 + Im(A^*)$. We can therefore apply the second point of Theorem 4.1 since we have uniqueness of the solution here.

The second result deals with geometries associated to $\gamma > 2$.

Theorem 4.3. *Let $\gamma_1 > 2$ and $\gamma_2 > 2$. Assume that F is coercive and satisfies $\boldsymbol{H}_1(\gamma_1)$ and $\boldsymbol{H}_2(\gamma_2)$ with $\gamma_1 \leqslant \gamma_2$. If $\alpha \geqslant \frac{\gamma_1+2}{\gamma_1-2}$ then we have*

$$F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2\gamma_2}{\gamma_2-2}}}\right). \tag{4.5}$$

One can notice that in Theorem 4.3 the uniqueness of the minimizer is not needed anymore.

In the case when $\gamma_1 = \gamma_2$, we have furthermore the convergence of the trajectory:

Corollary 4.4. *Let $\gamma > 2$, if F is coercive and satisfies $\boldsymbol{H}_1(\gamma)$ and $\boldsymbol{H}_2(\gamma)$ and if $\alpha \geqslant \frac{\gamma+2}{\gamma-2}$ then we have*

$$F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2\gamma}{\gamma-2}}}\right) \tag{4.6}$$

and

$$\|\dot{x}(t)\| = O\left(\frac{1}{t^{\frac{\gamma}{\gamma-2}}}\right). \tag{4.7}$$

Moreover the trajectory $x(t)$ is finite and it converges to a minimizer x^* of F.

This decay is optimal since Attouch et al. proved that it is achieved for the function $F(x) = |x|^\gamma$ in [5].

Comments on the hypotheses $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$. From these two theorems we can make the following comments:

1. If the function F behaves like $\|x - x^*\|^\gamma$ in the neighborhood of its unique minimizer x^*, then the decay of $F(x(t)) - F^*$ depends directly on α if $\gamma \leqslant 2$, but it does not depend on α, for large α if $\gamma > 2$.
2. For such functions the best decay rate of $F(x(t)) - F^*$ is achieved for $\gamma = 2$, that is for quadratic like functions around the minimizer, which is $O\left(\frac{1}{t^\alpha}\right)$. If $\gamma < 2$, it seems that the oscillations of the solution $x(t)$ prevent us to get an optimal decay rate. The inertia seems to be too large for such functions. If $\gamma > 2$, for large α, the decay is not as fast because the gradient of the functions decays too fast in the neighborhood of the minimizer. For these functions a larger inertia could be more efficient.
3. In the second point of Theorem 4.1 and in Theorem 4.3 both conditions $\mathbf{H}_1$ and $\mathbf{H}_2$ are used to get a decay rate. It turns out that these two conditions are important. Condition $\mathbf{H}_1(\gamma)$ ensures that the function is not too sharp and it may prevent from bad oscillations of the solution, while condition $\mathbf{H}_2(\gamma)$ ensures that the magnitude of the gradient of the function is not too low in the neighborhood of the minimizers.
4. With the sole condition $\mathbf{H}_1(\gamma)$ on F for any $\gamma \geqslant 1$, it is impossible to get a bound on the decay rate like $O(\frac{1}{t^\delta})$ with $\delta > 2$. Indeed, for any $\eta > 2$ and for a large friction parameter α, the solution x of the ODE associated to $F(x) = |x|^\eta$ satisfies $F(x(t)) - F^* = Kt^{-\frac{2\eta}{\eta-2}}$ and the power $\frac{2\eta}{\eta-2}$ can be chosen arbitrary close to 2.
5. With the sole hypothesis $\mathbf{H}_2(\gamma)$, it seems difficult to establish optimal rate. Indeed the function $F(x) = |x|^3$ satisfies $\mathbf{H}_2(3)$. Applying Theorem 4.3 with $\gamma_1 = \gamma_2 = 3$, we know that for this function with $\alpha = \frac{\gamma_1+2}{\gamma_1-2} = 5$, we have $F(x(t)) - F^* = O\left(\frac{1}{t^6}\right)$.
   Nevertheless, with the sole hypothesis $\mathbf{H}_2(3)$, such a decay cannot be achieved. Indeed, the function $F(x) = |x|^2$ satisfies $\mathbf{H}_2(3)$, but from the optimality part of Theorem 4.1 we know that we cannot achieve a decay better than $\frac{1}{t^{\frac{2\alpha\gamma}{\gamma+2}}} = \frac{1}{t^5}$ for $\alpha = 5$.
6. It seems that the use of both hypotheses is a simple way to provide optimal decay rates.

Relation with the state of the art.. As it was explained in the introduction, Attouch, Chbani, Peypouquet and Redont in [8] following Su, Boyd and Candes [27] proved that $F(x(t)) - F^* = O\left(t^{-\frac{2\alpha}{3}}\right)$ if F is strongly convex or has a strong minimizer, see also [6] for more general viscosity term in that setting. The contribution of the present work with respect to these previous result is the following

- It enlightens the fact that a flatness hypothesis $\mathbf{H}_1$ associated to classical sharpness hypotheses like Łojasievicz properties improve the decay rate of $F(x(t)) - F^*$.
- We prove the optimality of the given decay rates. A consequence of this work is also the optimality of the power $\frac{2\alpha}{3}$ in [8] with conditions $\mathbf{H}_1(1)$ and $\mathbf{H}_2(2)$.
- We prove that if a function F satisfies $\mathbf{H}_2(2)$ and a Lipschitz gradient condition, it satisfies $\mathbf{H}_1(\gamma)$ for a $\gamma > 1$ and thus the decay rate of $F(x(t)) - F^*$ is always strictly better than $O\left(t^{-\frac{2\alpha}{3}}\right)$.
- We prove that for quadratic functions we get $F(x(t)) - F^* = O\left(t^{-\alpha}\right)$
- This optimality also ensures that we cannot expect an exponential decay of $F(x(t)) - F^*$ for quadratic functions. Let us recall that we can achieve this exponential decay for the ODE associated to Gradient descent or Heavy ball method [26]

5. Proofs. Both proofs rely on Lyapunov functions $\mathcal{E}$ and $\mathcal{H}$ introduced by Su, Boyd and Candes [27], Attouch, Chbani, Peypouquet and Redont [5] and Aujol-Dossal [10] :

$$\mathcal{E}_{\lambda,\xi}(t) = t^2(F(x(t)) - F^*) + \frac{1}{2}\left\|\lambda(x(t) - x^*) + t\dot{x}(t)\right\|^2 + \frac{\xi}{2}\left\|x(t) - x^*\right\|^2. \tag{5.1}$$

where x^* is a minimizer of F and λ and ξ are two real numbers. The function $\mathcal{H}$ is defined from $\mathcal{E}$ and it depends on another real parameter p :

$$\mathcal{H}(t) = t^p\mathcal{E}(t) \tag{5.2}$$

Using the following notations

$$a(t) = t(F(x(t)) - F^*) \tag{5.3}$$

$$b(t) = \frac{1}{2t}\left\|\lambda(x(t) - x^*) + t\dot{x}(t)\right\|^2 \tag{5.4}$$

$$c(t) = \frac{1}{2t}\left\|x(t) - x^*\right\|^2 \tag{5.5}$$

we have

$$\mathcal{E}(t) = t(a(t) + b(t) + \xi c(t)) \tag{5.6}$$

From now on we will choose

$$\xi = \lambda(\lambda + 1 - \alpha) \tag{5.7}$$

and we will use the following Lemma whose proof is postponed to Appendix A:

LEMMA 5.1. *If F satisfies the hypothesis $\boldsymbol{H}_1(\gamma)$, for any $\gamma > 0$ and if $\xi = \lambda(\lambda - \alpha + 1)$ then*

$$\mathcal{H}'(t) \leqslant t^p\left((2 - \gamma\lambda + p)a(t) + (2\lambda + 2 - 2\alpha + p)b(t) + \lambda(\lambda + 1 - \alpha)(-2\lambda + p)c(t)\right) \tag{5.8}$$

One can remark that this inequality is actually an equality for the specific choice $F(x) = |x|^\gamma$.

5.1. Proof of Theorem 4.1. In this section we prove the second point of the Theorem and we refer to [10] for a complete proof of the first point, including the optimality of the rate. The proof of this first point is actually similar to the following one but simpler. The choice of p and λ are the same, but due to the value of α, the function $\mathcal{H}$ is non increasing and sum of non negative terms, which simplifies the analysis and necessitates less hypotheses to conclude.

Proof. We choose here $p = \frac{2\gamma\alpha}{\gamma+2} - 2$ and $\lambda = \frac{2\alpha}{\gamma+2}$ and thus

$$\xi = \frac{2\alpha}{(\gamma+2)^2}(2 + \gamma(1 - \alpha)) \tag{5.9}$$

Since we consider the case when $\alpha > 1 + \frac{2}{\gamma}$, we have $\xi < 0$ and thus $\mathcal{H}$ is not defined as a sum of positive functions.

Moreover, from Lemma 5.1 it appears that

$$\mathcal{H}'(t) \leqslant K_1 t^p c(t) \quad \text{with } K_1 = \lambda(\lambda + 1 - \alpha)(-2\lambda + p). \tag{5.10}$$

Let us compute explicitely the value of K_1. We have $p = \frac{2\gamma\alpha}{\gamma+2} - 2$ and $\lambda = \frac{2\alpha}{\gamma+2}$, so that:

$$\begin{aligned} K_1 &= \frac{2\alpha}{\gamma+2}\left(\frac{2\alpha}{\gamma+2} + 1 - \alpha\right)\left(-2\frac{2\alpha}{\gamma+2} + \frac{2\gamma\alpha}{\gamma+2} - 2\right) \\ &= \frac{4\alpha}{(\gamma+2)^3}\left(2\alpha + \gamma + 2 - \alpha\gamma - 2\alpha\right)\left(-2\alpha + \gamma\alpha - \gamma - 2\right) \\ &= \frac{4\alpha}{(\gamma+2)^3}\left(\gamma + 2 - \alpha\gamma\right)\left(\alpha(-2 + \gamma) - \gamma - 2\right) \end{aligned}$$

Hence

$$K_1 = \frac{4\alpha(\gamma(1-\alpha)+2)(\alpha(\gamma-2)-(\gamma+2))}{(\gamma+2)^3} \tag{5.11}$$

Since $\alpha > 1 + \frac{2}{\gamma}$, we have $\gamma(1-\alpha)+2 < 0$. Hence the sign of K_1 is the opposit of the sign of $\alpha(\gamma-2)-(\gamma+2)$. More precisely, if $\gamma \leqslant 2$, then $\alpha(\gamma-2)-(\gamma+2) < 0$ and thus $K_1 > 0$. If $\gamma > 2$, then $\alpha(\gamma-2)-(\gamma+2) > 0$ if and only if $\alpha > \frac{\gamma+2}{\gamma-2}$. Hence $K_1 > 0$ if and only if $\alpha < \frac{\gamma+2}{\gamma-2}$.

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

$$Kt \left\| x(t) - x^* \right\|^2 \leqslant t(F(x(t)) - F^*) = a(t), \tag{5.12}$$

and thus

$$c(t) \leqslant \frac{1}{2Kt^2} a(t). \tag{5.13}$$

Since $\xi < 0$ with our choice of parameters, we get:

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

It follows that it exists t_1 such that for all $t \geqslant t_1$, $\mathcal{H}(t) \geqslant 0$ and

$$\mathcal{H}(t) \geqslant \frac{1}{2} t^{p+1} a(t). \tag{5.14}$$

From (5.10), (5.13) and (5.14), we get

$$\mathcal{H}'(t) \leqslant \frac{K_1}{K} \frac{\mathcal{H}(t)}{t^3} \tag{5.15}$$

From the Gronwall Lemma it exists $A > 0$ such that $\forall t \geqslant t_1$ $\mathcal{H}(t) \leqslant A$. According to (5.14), we then conclude that $t^{p+2}(F(x(t)) - F^*) = t^{p+1}a(t)$ is bounded which concludes the proof of the point 2. of Theorem 4.1.
□

5.2. Proof of Proposition 4.2 (Optimality of the convergence rates). Before proving the optimality of the convergence rate stated in Proposition 4.2, we need the following technical lemma:

LEMMA 5.2. *Let y a continuously differentiable function with values in $\mathbb{R}$. Let $T > 0$ and $\epsilon > 0$. If y is bounded, then there exists $t_1 > T$ such that:*

$$|\dot{y}(t_1)| \leqslant \frac{\epsilon}{t_1} \tag{5.16}$$

Proof. We split the proof into two cases.

1. There exists $t_1 > T$ such that $\dot{y}(t_1) = 0$.
2. $\dot{y}(t)$ is of constant sign for $t > T$. For instance we assume $\dot{y}(t) > 0$. By contradiction, let us assume that $\dot{y}(t) > \frac{\epsilon}{t}$ $\forall t > T$. Then $y(t)$ cannot be a bounded function as assumed.

□

Let us now prove the Proposition 4.2: the idea of the proof is the following: we first show that $\mathcal{H}$ is bounded from below. Since $\mathcal{H}$ is a sum of 3 terms including the term $F - F^*$, we then

show that given $t_1 \geq t_0$, there always exists a time $t \geq t_1$ such that the value of $\mathcal{H}$ is concentrated on the term $F - F^*$.

We start the proof by using the fact that, for the function $F(x) = |x|^\gamma$, the inequality of Lemma 5.1 is actually an equality, and using the values $p = \frac{2\gamma\alpha}{\gamma+2} - 2$ and $\lambda = \frac{2\alpha}{\gamma+2}$ of the previous Theorem, we have a closed form for the derivative of function $\mathcal{H}$:

$$\mathcal{H}'(t) = K_1 t^p c(t) = \frac{K_1}{2} t^{p-1} |x(t)|^2 \tag{5.17}$$

where K_1 is a positive constant. This implies in particular that $\mathcal{H}$ is non-decreasing.

From the previous Theorem, $\mathcal{H}$ is bounded above. Observing that since $t_0 > \sqrt{\frac{\alpha(\gamma+2-\alpha\gamma)}{(\gamma+2)^2}}$, we have: $\mathcal{H}(t_0) > 0$, it follows that it exists $\ell > 0$ such that $\mathcal{H}(t) \to \ell$ when $t \to +\infty$. In particular, we have that for t large enough,

$$\mathcal{H}(t) \geqslant \frac{\ell}{2} \tag{5.18}$$

i.e.

$$a(t) + b(t) + \xi c(t) \geqslant \frac{\ell}{2t^{p+1}} \tag{5.19}$$

Moreover, since $\xi < 0$, we then have:

$$a(t) + b(t) \geqslant \frac{\ell}{2t^{p+1}} \tag{5.20}$$

Let $T > 0$ and $\epsilon > 0$. We set:

$$y(t) := t^\lambda x(t) \tag{5.21}$$

where: $\lambda = \frac{2\alpha}{\gamma+2}$. From the previous Theorem, we know that $y(t)$ is bounded. Hence, from Lemma 5.2 , there exists $t_1 > T$ such that

$$|\dot{y}(t_1)| \leqslant \frac{\epsilon}{t_1} \tag{5.22}$$

But

$$\dot{y}(t) = t^{\lambda-1} \left(\lambda x(t) + t\dot{x}(t)\right) \tag{5.23}$$

Hence using (5.22):

$$t_1^\lambda \left|\lambda x(t_1) + t_1 \dot{x}(t_1)\right| \leqslant \epsilon \tag{5.24}$$

We remind the reader that $b(t) = \frac{1}{2t} \left\|\lambda(x(t) - x^*) + t\dot{x}(t)\right\|^2$. We thus have:

$$b(t_1) \leqslant \frac{\epsilon^2}{2t_1^{2\lambda+1}} \tag{5.25}$$

Since $\gamma \leqslant 2$, $\lambda = \frac{2\alpha}{\gamma+2}$ and $p = \frac{2\gamma\alpha}{\gamma+2} - 2$, we have $2\lambda + 1 \geq p + 1$, and thus

$$b(t_1) \leqslant \frac{\epsilon^2}{2t_1^{p+1}}. \tag{5.26}$$

For $\epsilon = \sqrt{\frac{l}{2}}$ for example, there exists thus some $t_1 > T$ such that $b(t_1) \leqslant \frac{\ell}{4t_1^{p+1}}$. Then $a(t_1) \geqslant \frac{\ell}{4t_1^{p+1}}$, i.e. $F(x(t_1)) - F^* \geqslant \frac{\ell}{4t_1^{p+2}}$. Since $p + 2 = \frac{2\gamma\alpha}{\gamma+2}$, this concludes the proof.

5.3. Proof of Theorem 4.3. We detail here the proof of Theorem 4.3.

Let us consider $\gamma_1 > 2$, $\gamma_2 > 2$, and $\alpha \geqslant \frac{\gamma_1+2}{\gamma_1-2}$.

We consider here functions $\mathcal{H}$ for all x^* in the set X^* of minimizers of F and prove that these functions are uniformely bounded. More precisely for any $x^* \in X^*$ we define $\mathcal{H}(t)$ with $p = \frac{4}{\gamma_1-2}$ and $\lambda = \frac{2}{\gamma_1-2}$. With this choice of λ and p, using Hypothesis $\mathbf{H}_1(\gamma_1)$ we have from Lemma 5.1:

$$\mathcal{H}'(t) \leqslant 2t^{\frac{4}{\gamma_1-2}} \left(\frac{\gamma_1+2}{\gamma_1-2} - \alpha \right) b(t). \tag{5.27}$$

which is non positive when $\alpha \geqslant \frac{\gamma_1+2}{\gamma_1-2}$, which implies that the function $\mathcal{H}$ is bounded above. Hence for any choice of x^* in the set of minimizers X^*, the function $\mathcal{H}$ is bounded above and since the set of minimizers is bounded (F is coercive), there exists $A > 0$ and t_0 such that for all choices of x^* in X^*,

$$\mathcal{H}(t_0) \leqslant A, \tag{5.28}$$

which implies that for all $x^* \in X^*$ and for all $t \geqslant t_0$

$$\mathcal{H}(t) \leqslant A. \tag{5.29}$$

Hence for all $t \geqslant t_0$ and for all $x^* \in X^*$

$$t^{\frac{4}{\gamma_1-2}} t^2 (F(x(t)) - F^*) \leqslant \frac{|\xi|}{2} t^{\frac{4}{\gamma_1-2}} \|x(t) - x^*\|^2 + A, \tag{5.30}$$

which implies that

$$t^{\frac{4}{\gamma_1-2}} t^2 (F(x(t)) - F^*) \leqslant \frac{|\xi|}{2} t^{\frac{4}{\gamma_1-2}} d(x(t), X^*)^2 + A. \tag{5.31}$$

We now set:

$$v(t) := t^{\frac{4}{\gamma_2-2}} d(x(t), X^*)^2 \tag{5.32}$$

where d denotes the euclidean distance. Using (5.31) we have:

$$t^{\frac{2\gamma_1}{\gamma_1-2}} (F(x(t)) - F^*) \leqslant \frac{|\xi|}{2} t^{\frac{4}{\gamma_1-2} - \frac{4}{\gamma_2-2}} v(t) + A. \tag{5.33}$$

Using the hypothesis $\mathbf{H}_2(\gamma_2)$ applied under the form given by Lemma 2.4 (since X^* is compact), there exists $K > 0$ such that

$$K \left(t^{-\frac{4}{\gamma_2-2}} v(t) \right)^{\frac{\gamma_2}{2}} \leqslant F(x(t)) - F^*, \tag{5.34}$$

which is equivalent to

$$K v(t)^{\frac{\gamma_2}{2}} t^{\frac{-2\gamma_2}{\gamma_2-2}} \leqslant F(x(t)) - F^*. \tag{5.35}$$

Hence:

$$K t^{\frac{2\gamma_1}{\gamma_1-2}} t^{\frac{-2\gamma_2}{\gamma_2-2}} v(t)^{\frac{\gamma_2}{2}} \leqslant t^{\frac{2\gamma_1}{\gamma_1-2}} (F(x(t)) - F^*). \tag{5.36}$$

Using (5.33), we obtain:

$$Kt^{\frac{2\gamma_1}{\gamma_1-2}-\frac{2\gamma_2}{\gamma_2-2}}v(t)^{\frac{\gamma_2}{2}} \leqslant \frac{|\xi|}{2}t^{\frac{4}{\gamma_1-2}-\frac{4}{\gamma_2-2}}v(t)+A, \tag{5.37}$$

i.e.:

$$Kv(t)^{\frac{\gamma_2}{2}} \leqslant \frac{|\xi|}{2}v(t)+At^{\frac{4}{\gamma_2-2}-\frac{4}{\gamma_1-2}}. \tag{5.38}$$

Since $\gamma_1 \leqslant \gamma_2$, we deduce that v is bounded. Hence, using (5.33) there exists some positive constant B such that:

$$F(x(t))-F^* \leqslant Bt^{\frac{-2\gamma_2}{\gamma_2-2}}+At^{\frac{-2\gamma_1}{\gamma_1-2}}. \tag{5.39}$$

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

5.3.1. Proof of Corollary 4.4. We are now in position to prove Corollary 4.4.

Proof. The first point of Corollary 4.4 is just a particular instance of Theorem 4.3. In the sequel, we prove the second point of Corollary 4.4.

Let $t \geqslant t_0$ and $\tilde{x} \in X^*$ such that

$$d(x(t),\tilde{x}) = d(x(t),X^*) \tag{5.40}$$

We previously proved that it exists $A>0$ such that for any $t \geqslant t_0$ and any $x^* \in X^*$

$$\mathcal{H}(t) \leqslant A. \tag{5.41}$$

For the choice $x^*=\tilde{x}$ this inequality ensures that

$$\frac{t^{\frac{4}{\gamma-2}}}{2}\left\|\lambda(x(t)-\tilde{x})+t\dot{x}(t)\right\|^2+t^{\frac{4}{\gamma-2}}\frac{\xi}{2}d(x(t),\tilde{x})^2 \leqslant A \tag{5.42}$$

which is equivalent to

$$\frac{t^{\frac{4}{\gamma-2}}}{2}\left\|\lambda(x(t)-\tilde{x})+t\dot{x}(t)\right\|^2 \leqslant \frac{|\xi|}{2}v(t)+A \tag{5.43}$$

where $v(t)$ is defined in (5.32) with $\gamma=\gamma_2$. Using the fact that the function v is bounded (a consequence of (5.38)) we deduce that it exists a real number $A_1>0$ such that

$$\left\|\lambda(x(t)-\tilde{x})+t\dot{x}(t)\right\| \leqslant \frac{A_1}{t^{\frac{2}{\gamma-2}}}. \tag{5.44}$$

Thus:

$$t\left\|\dot{x}(t)\right\| \leqslant \frac{A_1}{t^{\frac{2}{\gamma-2}}}+|\lambda|d(x(t),\tilde{x}) = \frac{A_1+|\lambda|\sqrt{v(t)}}{t^{\frac{2}{\gamma-2}}}. \tag{5.45}$$

Using once again the fact that the function v is bounded we deduce that it exists a real number A_2 such that

$$\left\|\dot{x}(t)\right\| \leqslant \frac{A_2}{t^{\frac{\gamma}{\gamma-2}}} \tag{5.46}$$

which implies that $\|\dot{x}(t)\|$ is an integrable function. As a consequence, we deduce that the trajectory $x(t)$ is finite.

□

Acknowledgement. This study has been carried out with financial support from the French state, managed by the French National Research Agency (ANR GOTMI) (ANR-16-VCE33-0010-01). J-F. Aujol is a member of Institut Universitaire de France.

Appendix A. Proof of Lemma 5.1. We prove here Lemma 5.1. Notice that the computations are standard (see e.g. [10]).

We will make use of the following results:

LEMMA A.1. *Let $\gamma > 0$, if $g(x) = (F(x) - F(x^*))^{\frac{1}{\gamma}}$ is convex then*

$$\gamma(F(x(t)) - F(x^*)) \leqslant \langle \nabla F(x(t)), x(t) - x^* \rangle \tag{A.1}$$

Proof. Since g is convex we have

$$g(x(t)) \leqslant \langle \nabla g(x(t)), x(t) - x^* \rangle \tag{A.2}$$

and $\nabla g(x(t)) = \frac{1}{\gamma}(F(x(t) - F(x^*))^{\frac{1}{\gamma}-1}\nabla F(x(t))$. Replacing $g(x(t))$ by $(F(x(t)) - F(x^*))^{\frac{1}{\gamma}}$ we get the result.
□

LEMMA A.2.

$$\mathcal{E}'_{\lambda,\xi}(t) = 2a(t) + t(\lambda + 1 - \alpha)\left\|\dot{x}(t)\right\|^2 \tag{A.3}$$

$$+ \lambda t\langle -\nabla F(x(t)), x(t) - x^* \rangle + (\lambda(\lambda+1) - \alpha\lambda + \xi)\langle \dot{x}(t), x(t) - x^* \rangle \tag{A.4}$$

Proof.
We differentiate :

$$\mathcal{E}'_{\lambda,\xi}(t) = 2a(t) + t^2\langle \nabla F(x(t)), \dot{x}(t) \rangle \tag{A.5}$$

$$+ \langle \lambda\dot{x}(t) + t\ddot{x}(t) + \dot{x}(t), \lambda(x(t) - x^*) + t\dot{x}(t) \rangle + \xi\langle \dot{x}(t), x(t) - x^* \rangle \tag{A.6}$$

$$\mathcal{E}'_{\lambda,\xi}(t) = 2a(t) + t^2\langle \nabla F(x(t)) + \ddot{x}(t), \dot{x}(t) \rangle \tag{A.7}$$

$$+ (\lambda+1)t\left\|\dot{x}(t)\right\|^2 + \lambda t\langle \ddot{x}(t), x(t) - x^* \rangle + (\lambda(\lambda+1) + \xi)\langle \dot{x}(t), x(t) - x^* \rangle \tag{A.8}$$

$$\mathcal{E}'_{\lambda,\xi}(t) = 2a(t) + t^2\langle -\frac{\alpha}{t}\dot{x(t)}, \dot{x}(t) \rangle \tag{A.9}$$

$$+ (\lambda+1)t\left\|\dot{x}(t)\right\|^2 + \lambda t\langle \ddot{x}(t), x(t) - x^* \rangle + (\lambda(\lambda+1) + \xi)\langle \dot{x}(t), x(t) - x^* \rangle \tag{A.10}$$

$$\mathcal{E}'_{\lambda,\xi}(t) = 2a(t) + t(\lambda + 1 - \alpha)\left\|\dot{x}(t)\right\|^2 \tag{A.11}$$

$$+ \lambda t\langle \ddot{x}(t), x(t) - x^* \rangle + (\lambda(\lambda+1) + \xi)\langle \dot{x}(t), x(t) - x^* \rangle \tag{A.12}$$

And now, using (1.1), we get:

$$\mathcal{E}'_{\lambda,\xi}(t) = 2a(t) + t(\lambda + 1 - \alpha)\left\|\dot{x}(t)\right\|^2 \tag{A.13}$$

$$+ \lambda t\langle -\nabla F(x(t)) - \frac{\alpha}{t}\dot{x(t)}, x(t) - x^* \rangle + (\lambda(\lambda+1) + \xi)\langle \dot{x}(t), x(t) - x^* \rangle \tag{A.14}$$

$$\mathcal{E}'_{\lambda,\xi}(t) = 2a(t) + t(\lambda + 1 - \alpha)\left\|\dot{x}(t)\right\|^2 \tag{A.15}$$

$$+ \lambda t\langle -\nabla F(x(t)), x(t) - x^*\rangle + (\lambda(\lambda+1) - \alpha\lambda + \xi)\langle \dot{x}(t), x(t) - x^*\rangle \tag{A.16}$$

□

Lemma A.3.

$$\mathcal{E}'_{\lambda,\xi}(t) = 2a(t) + \lambda t\langle -\nabla F(x(t)), x(t) - x^*\rangle \tag{A.17}$$

$$+ (\xi - \lambda(\lambda + 1 - \alpha))\langle \dot{x}(t), x(t) - x^*\rangle \tag{A.18}$$

$$+ 2(\lambda + 1 - \alpha)b(t) - 2\lambda^2(\lambda + 1 - \alpha)c(t) \tag{A.19}$$

Proof.
We start from the result of Lemma A.2. Observing that

$$\frac{1}{t}\left\|\lambda(x(t) - x^*) + t\dot{x}(t)\right\|^2 = t\left\|\dot{x}(t)\right\|^2 + 2\lambda\langle \dot{x}(t), x(t) - x^*\rangle + \frac{\lambda^2}{t}\left\|x(t) - x^*\right\|^2 \tag{A.20}$$

we can write

$$\mathcal{E}'_{\lambda,\xi}(t) = 2a(t) + \lambda t\langle -\nabla F(x(t)), x(t) - x^*\rangle \tag{A.21}$$

$$+ (\xi - \lambda(\lambda + 1 - \alpha))\langle \dot{x}(t), x(t) - x^*\rangle \tag{A.22}$$

$$+ (\lambda + 1 - \alpha)\frac{1}{t}\left\|\lambda(x(t) - x^*) + t\dot{x}(t)\right\|^2 - \frac{\lambda^2(\lambda + 1 - \alpha)}{t}\left\|x(t) - x^*\right\|^2 \tag{A.23}$$

□

Lemma A.4. *If $\xi = \lambda(\lambda + 1 - \alpha)$, then*

$$\mathcal{E}'_{\lambda,\xi}(t) = 2a(t) + \lambda t\langle -\nabla F(x(t)), x(t) - x^*\rangle + 2(\lambda + 1 - \alpha)b(t) - 2\lambda^2(\lambda + 1 - \alpha)c(t) \tag{A.24}$$

Lemma A.5.
If F satisfies the hypothesis $\boldsymbol{H}_1(\gamma)$ and $\xi = \lambda(\lambda + 1 - \alpha)$, then:

$$\mathcal{E}'_{\lambda,\xi}(t) \leqslant (2 - \gamma\lambda)a(t) + 2(\lambda + 1 - \alpha)b(t) - 2\lambda^2(\lambda + 1 - \alpha)c(t) \tag{A.25}$$

Proof. To prove Lemma A.5, we only apply the inequality of Lemma A.1 in the equality of Lemma A.4.
□

One can notice that if $F(x) = |x|^\gamma$ the inequality of Lemma A.4 is actually an equality, which ensures that for this specific function F the inequality in Lemma 5.1 is an equality.

Lemma A.6. *If $\xi = \lambda(\lambda + 1 - \alpha)$, then*

$$\mathcal{H}'(t) = t^p\left((2 + p)a(t) + \lambda t\langle -\nabla F(x(t)), x(t) - x^*\rangle + (2\lambda + 2 - 2\alpha + p)b(t)\right. \tag{A.26}$$

$$\left. + \lambda(\lambda + 1 - \alpha)(-2\lambda + p)c(t)\right) \tag{A.27}$$

Proof.
We have $\mathcal{H}(t) = t^p\mathcal{E}(t)$.

Hence $\mathcal{H}'(t) = t^p\mathcal{E}'(t) + pt^{p-1}\mathcal{E}(t) = t^{p-1}(t\mathcal{E}'(t) + p\mathcal{E}(t))$. We conclude by using Lemma A.4. □

Proof. [Proof of Lemma 5.1] We only apply the inequality of Lemma A.1 in the equality of Lemma A.6. □

REFERENCES

[1] V. Apidopoulos, J.-F. Aujol, and C. 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 J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming.*, 116(1):5–16, 2009.

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

[4] 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.

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

[6] H. Attouch, Z. Chbani, and H. Riahi. Rate of convergence of the nesterov accelerated gradient method in the subcritical case $\alpha \leqslant 3$. *arXiv preprint arXiv:1706.05671*, 2017.

[7] 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(01):1–34, 2000.

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

[9] J.-F. Aujol and C. Dossal. Stability of over-relaxations for the forward-backward algorithm, application to FISTA. *SIAM Journal on Optimization*, 25(4):2408–2433, 2015.

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

[11] M. Balti and R. May. Asymptotic for the perturbed heavy ball system with vanishing damping term. *arXiv preprint arXiv:1609.00135*, 2016.

[12] 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.

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

[14] J. Bolte, A. Daniilidis, and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223 (electronic), 2006.

[15] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

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

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

[18] 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.

[19] A. Chambolle and C. Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization Theory and Applications*, 166(3):968–982, 2015.

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

[21] 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.

[22] 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.

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

[24] R. May. Asymptotic for a second order evolution equation with convex potential and vanishing damping term. *arXiv preprint arXiv:1509.05598*, 2015.

[25] Y. Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

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

[27] W. Su, S. Boyd, and E.J. Candes. A differential equation for modeling nesterov's accelerated gradient method: theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.