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 with Lipschitz gradient. 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.

AMS subject classifications. 34D05, 65K05, 65K10, 90C25, 90C30

1. Introduction. The motivation of this paper lies in the minimization of a differentiable function $F : \mathbb{R}^n \to \mathbb{R}$ with at least one minimizer. Inspired by Nesterov pioneering work [24], we study the following ordinary differential equation (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 the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA)[12] or the Accelerated Gradient Method [24] :

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

with h and α positive parameters. This equation, including or not a perturbation term, has been widely studied in the literature [8, 26, 15, 11, 23]. 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, 19, 23, 16, 4, 25, 3] and the references therein.

Throughout the paper, we assume that, for any initial conditions $(x_0, v_0) \in \mathbb{R}^n \times \mathbb{R}^n$, the Cauchy problem associated with the differential equation (1.1), has a unique global solution x satisfying $(x(t_0), \dot{x}(t_0)) = (x_0, v_0)$. This is guaranteed for instance when the gradient function ∇F is Lipschitz on bounded subsets of $\mathbb{R}^n$.

In this work we investigate the convergence rates of the values $F(x(t)) - F^*$ for the trajectories of the ODE (1.1). It was proved in [6] that if F is convex with Lipschitz gradient and if $\alpha > 3$, the trajectory $F(x(t))$ 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 work of Chambolle-Dossal [17] of the convergence of iterates of FISTA, Attouch et al. [6] proved that for $\alpha > 3$ the trajectory x converges (weakly in infinite-dimensional Hilbert space) to a minimizer of F. Su et al. [26] 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 [6] 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 [6, 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. Indeed, Attouch et al. in [6] proved that if F is convex then for any $\alpha > 0$, $F(x(t)) - F^*$ tends to 0 when t goes to infinity. Combined with the coercivity of F, this convergence implies that the distance $d(x(t), X^*)$ between $x(t)$ and the set of minimizers X^* tends to 0. To analyse the asymptotic behavior of $F(x(t)) - F^*$ we can thus only assume hypotheses on F only on the neighborhood of X^* and may avoid the tough question of the convergence of the the trajectory $x(t)$ to a point of X^*.

More precisely, we consider functions behaving like $\|x - x^*\|^\gamma$ around their set of minimizers 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. In this paper, we will show that both conditions are important to get the expected convergence rates: the flatness assumption ensures that the function is not too sharp and may prevent from bad oscillations of the solution, while the sharpness condition ensures that the magnitude of the gradient of the function is not too low in the neighborhood of the minimizers.

The paper is organized as follows. In Section 2, we introduce the geometrical hypotheses we consider on 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. Throughout the paper we assume that the ODE (1.1) is defined in $\mathbb{R}^n$ 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^* and radius $r > 0$ while $\bar{B}(x^*, r)$ denotes the closed euclidean ball with center x^* and radius $r > 0$.

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 minimizer $x^* \in X^*$, there exists $\eta > 0$ such that:

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

(ii) Let $r \geqslant 1$. The function F satisfies the growth condition $\boldsymbol{H}_2(r)$ if, for any minimizer

$x^* \in X^*$, there exist $K > 0$ and $\varepsilon > 0$, such that:

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

The hypothesis $\mathbf{H}_1(\gamma)$ has already been used in [15] and later in [26, 10]. This is a mild assumption, requesting slightly more than the convexity of F in the neighborhood of its minimizers. Observe that any convex function automatically satisfies $\mathbf{H}_1(1)$ and that any differentiable function F for which $(F - F^*)^{\frac{1}{\gamma}}$ 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 $x^* \in X^*$. There then exists $\eta > 0$ such that:

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

Let $\eta' \in (0, \eta)$. 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^* and the real constant η'.

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:

$$\forall x \in \bar{B}(x^*, \eta') \;\text{ with } x \neq x^*,\; \forall t \in [0,1],\; g_x(t) \leqslant g_x(1)t^\gamma \leqslant Mt^\gamma. \tag{2.4}$$

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 minimizer (any critical point in the non-convex case). It is sometimes also called r-conditioning [18] or Hölderian error bounds [14]. This assumption is motivated by the fact that, when F is convex, $\mathbf{H}_2(r)$ is equivalent to the famous Łojasiewicz inequality [21, 22], a key tool in the mathematical analysis of continuous (or discrete) subgradient dynamical systems, with exponent $\theta = 1 - \frac{1}{r}$:

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:*

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

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

When the set X^* of the minimizers is a connected compact set, the Łojasiewicz inequality turns into a geometrical condition on F around its set of minimizers X^*, usually referred to as Hölder metric subregularity [20], and whose proof can be easily adapted from [2, Lemma 1]:

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, C^1 subanalytic functions or semi-algebraic functions [21, 22]. 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:

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

where $\mu > 0$ denotes the parameter of strong convexity and x^* the unique minimizer of F. By extension, uniformly convex functions of order $p \geqslant 2$ satisfy the global version of the hypothesis $\mathbf{H}_2(p)$ [18].

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. Let $\gamma > 1$ and consider the function defined by: $F : x \in \mathbb{R} \mapsto |x|^\gamma$. 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 uniqueness of the minimizer for functions of the form:

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

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 $\gamma = 1 + \frac{K}{2L}$. ∎

Remark 2.7. Observe that Lemma 2.6 can be easily extended to the case of convex differentiable functions with a ν-Hölder continuous gradient. Indeed, let F be a convex differentiable functions with a ν-Hölder continuous gradient for some $\nu \geqslant 1$. If F also satisfies the growth condition $\mathbf{H}_2(1 + \nu)$ (for some constant $K > 0$), then F automatically satisfies $\mathbf{H}_1(\gamma)$ with $\gamma = 1 + \frac{\alpha K}{(1+\nu)L^{\frac{1}{\nu}}}$. This result is based on a notion of generalized co-coercivity for functions having a Hölder continuous gradient.

3. Related results. In this section, we recall some classical state of the art results on the convergence properties of the trajectories of the ODE (1.1).

Let us first recall that as soon as $\alpha > 0$, $F(x(t))$ converges to F^* [10, 7], but a larger value of α is required to show the convergence of the trajectory $x(t)$. More precisely, if F is convex and $\alpha > 3$, or if F satisfies $\mathbf{H}_1(\gamma)$ hypothesis and $\alpha > 1 + \frac{2}{\gamma}$ then:

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

and the trajectory $x(t)$ converges (weakly in an infinite dimensional space) to a minimizer x^* of F [26, 10, 23]. This last point generalizes what is known on convex functions: thanks to the additional hypothesis $\mathbf{H}_1(\gamma)$, the optimal decay $\frac{1}{t^2}$ can be achieved for a damping parameter α smaller that 3.

In the sub-critical case (namely when $\alpha < 3$), it has been proven in [7, 10] that if F is convex, the convergence rate is then given by:

$$F(x(t)) - F^* = O\left(\frac{1}{t^{\frac{2\alpha}{3}}}\right), \tag{3.1}$$

but we can no longer prove the convergence of the trajectory $x(t)$.

The purpose in this paper is to prove that by exploiting the geometry of the function F, better rates of convergence can be achieved for the values $F(x(t)) - F^*$.

Consider first the case when F is convex and $\alpha \leqslant 1 + \frac{2}{\gamma}$. A first contribution in this paper is to provide convergence rates for the values when F only satisfies $\mathbf{H}_1(\gamma)$. Although we can no longer prove the convergence of the trajectory $x(t)$, we still have the following convergence rate for $F(x(t)) - F^*$:

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

and this decay is optimal and achieved for $F(x) = |x|^\gamma$ for any $\gamma \geqslant 1$. These results have been first stated and proved in the unpublished report [10] by Aujol and Dossal in 2017 for convex differentiable functions satisfying $(F - F^*)^{\frac{1}{\gamma}}$ convex. Observe that this decay is still valid for $\gamma = 1$ i.e. with the sole assumption of convexity as shown in [7], and that the constant 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).

Consider now the case when $\alpha > 1 + \frac{2}{\gamma}$. In that case, with the sole assumption $\mathbf{H}_1(\gamma)$ on F for some $\gamma \geqslant 1$, it is not possible to get a bound on the decay rate like $O(\frac{1}{t^\delta})$ with $\delta > 2$. Indeed as shown in [6, Example 2.12], 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. More conditions are thus needed to obtain a decay faster than $O\left(\frac{1}{t^2}\right)$, which is the uniform rate that can be achieved for $\alpha \geqslant 3$ for convex functions.

Our main contribution is to show that a flatness condition $\mathbf{H}_1$ associated to classical sharpness conditions such as the Łojasiewicz property provides new and better decay rates on the values $F(x(t)) - F^*$, and to prove the optimality of these rates in the sense that they are achieved for instance for the function $F(x) = |x|^\gamma$, $x \in \mathbb{R}$, $\gamma \geqslant 1$.

We will then confront our results to well-known results in the literature. In particular we will focus on the case when F is strongly convex or has a strong minimizer [15]. In that case,

Attouch Chbani, Peypouquet and Redont in [6] following Su, Boyd and Candes [26] proved that for any $\alpha > 0$ we have:

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

(see also [7] for more general viscosity term in that setting). In Section 4, we will prove the optimality of the power $\frac{2\alpha}{3}$ in [5], and that if F has additionally a Lipschitz gradient then the decay rate of $F(x(t)) - F^*$ is always strictly better than $O\left(t^{-\frac{2\alpha}{3}}\right)$.

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

$$\dot{x}(t) + \nabla F(x(t)) = 0, \tag{3.3}$$

or the ODE associated to the heavy ball method

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

under geometrical conditions such that the Łojasiewicz property have been proposed, see for example Polyak-Shcherbakov [25]. 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 exponential 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 $\alpha > 0$. If F satisfies $\boldsymbol{H}_1(\gamma)$ and if $\alpha \leqslant 1 + \frac{2}{\gamma}$ then:*

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

Note that a proof of the Theorem 4.1 has been proposed in the unpublished report [10]. The obtained decay is proved to be optimal in the sense that it can be achieved for some explicit functions F for any $\gamma < 1$. As a consequence one cannot expect a $o(t^{-\frac{2\gamma\alpha}{\gamma+2}})$ decay when $\alpha < 1 + \frac{2}{\gamma}$.

Let us now consider the case when $\alpha > 1 + \frac{2}{\gamma}$. The second result in this paper provides optimal convergence rates for functions whose geometry is sharp, with a large friction coefficient:

THEOREM 4.2. *Let $\gamma \geqslant 1$ and $\alpha > 0$. If F satisfies $\boldsymbol{H}_1(\gamma)$ and $\boldsymbol{H}_2(2)$ for some $\gamma \leqslant 2$, if F has a unique minimizer and if $\alpha > 1 + \frac{2}{\gamma}$ then*

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

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 Theorem 4.2 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 2.5). The optimality of the convergence rate result is precisely stated in the next Proposition:

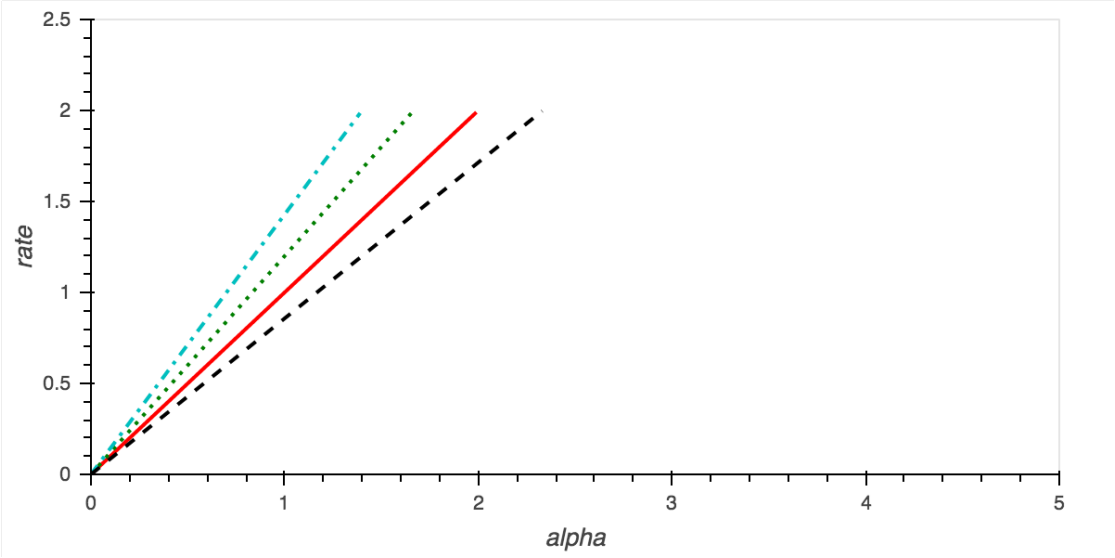

FIG. 1. *Decay rate* $r(\alpha,\gamma)=\frac{2\alpha\gamma}{\gamma+2}$ *depending on* α *when* $\alpha\leqslant 1+\frac{2}{\gamma}$ *and when* F *satisfies* $\boldsymbol{H}_1(\gamma)$ *(as in Theorem 4.1) for four values* γ*:* $\gamma_1=1.5$ *dashed line,* $\gamma_2=2$*, solid line,* $\gamma_3=3$ *dotted line and* $\gamma_4=5$ *dashed-dotted line.*

PROPOSITION 4.3. *Let* $\gamma\in(1,2]$*. Let us assume that* $\alpha>0$*. Let* x *be a solution of* (1.1) *with* $F(x)=|x|^\gamma$*,* $|x(t_0)|<1$ *and* $\dot{x}(t_0)=0$ *where* $t_0>\sqrt{\max(0,\frac{\alpha\gamma(\alpha-1-2/\gamma)}{(\gamma+2)^2})}$*. There exists* $K>0$ *such that for any* $T>0$*, there exists* $t\geqslant T$ *such that*

$$F(x(t))-F^*\geqslant\frac{K}{t^{\frac{2\gamma\alpha}{\gamma+2}}}. \tag{4.1}$$

Let us make several observations: first, to apply Theorem 4.2, more conditions are needed than for Theorem 4.1: the hypothesis $\mathbf{H}_2(2)$ and the uniqueness of the minimizer are needed 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 [26]. The uniqueness of the minimizer is crucial in the proof of Theorem 4.2, but it is still an open problem to know if this uniqueness is a necessary condition. In particular, observe that if $\dot{x}(t_0)=0$, then for all $t\geqslant t_0$, $x(t)$ belongs to $x_0+\mathrm{Im}(\nabla F)$ where $\mathrm{Im}(\nabla F)$ stands for the vector space generated by $\nabla F(x)$ for all x in $\mathbb{R}^n$. As a consequence, Theorem 4.2 still holds true as long as the assumptions are valid in $x_0+\mathrm{Im}(\nabla F)$.

Remark 4.4 (The Least-Square problem). 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 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 $x_0+\mathrm{Im}(A^*)$. Since we have uniqueness of the solution on $x_0+\mathrm{Im}(A^*)$, Theorem 4.2 can be applied.

We can also remark that if F is a quadratic function in the neighborhood of x^*, then F satisfies $\mathbf{H}_1(\gamma)$ for any $\gamma\in[1,2]$. Consequently, Theorem 4.2 applies with $\gamma=2$ and thus:

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

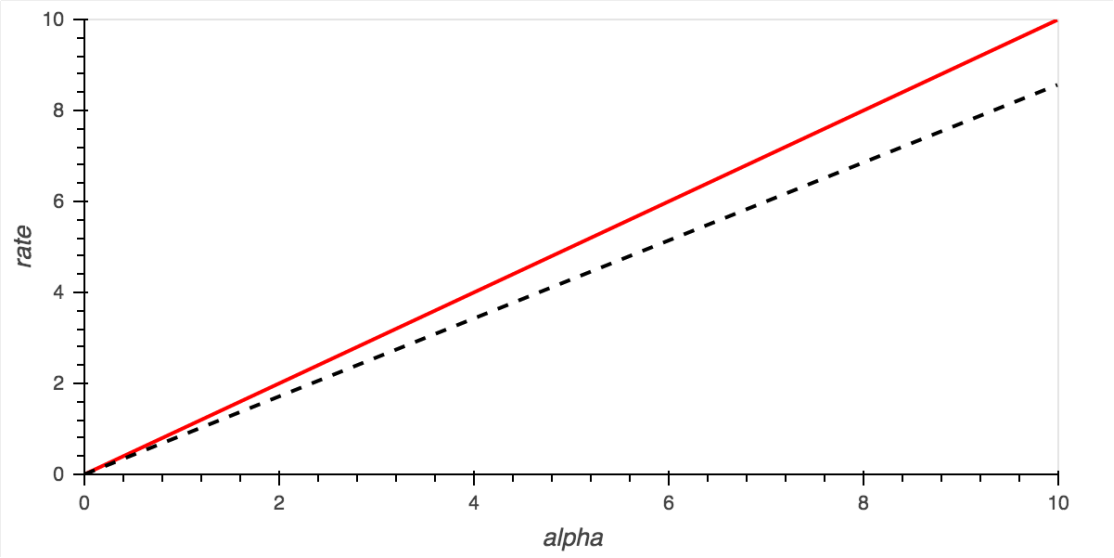

FIG. 2. *Decay rate* $r(\alpha,\gamma) = \frac{2\alpha\gamma}{\gamma+2}$ *depending on the value of* α *when* F *satisfies* $\boldsymbol{H}_1(\gamma)$ *and* $\boldsymbol{H}_2(2)$ *(as in Theorem 4.2) with* $\gamma \leqslant 2$ *for two values* γ *:* $\gamma_1 = 1.5$ *dashed line,* $\gamma_2 = 2$*, solid line.*

Observe that the optimality result provided by the Proposition 4.3 ensures that we cannot expect an exponential decay of $F(x(t)) - F^*$ for quadratic functions whereas this exponential decay can be achieved for the ODE associated to Gradient descent or Heavy ball method [25].

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 assumption $\mathbf{H}_1(\gamma)$ with some $1 < \gamma \leqslant 2$ as shown by Lemma 2.6, and Theorem 4.2 applies with $\gamma > 1$.

Finally if F is strongly convex or has a strong minimizer, then F naturally satisfies $\mathbf{H}_1(1)$ and a global version of $\mathbf{H}_2(2)$. Since we prove the optimality of the decay rates given by Theorem 4.2, a consequence of this work is also the optimality of the power $\frac{2\alpha}{3}$ in [5] for strongly convex functions and functions having a strong minimizer.

In both cases, we thus obtain convergence rates which are strictly better than $O(t^{-\frac{2\alpha}{3}})$ that is proposed for strongly convex functions by Su et al. [26] and Attouch et al. [6]. Finally it is worth noticing that the decay for strongly convex functions is not exponential while it is the case for the classical gradient descent scheme (see e.g. [18]). 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.

Let us now focus on flat geometries i.e. geometries associated to $\gamma > 2$. Note that the uniqueness of the minimizer is not need anymore:

THEOREM 4.5. *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:*

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

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

COROLLARY 4.6. *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:

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

and

$$\|\dot{x}(t)\| = O\left(\frac{1}{t^{\frac{\gamma}{\gamma-2}}}\right). \tag{4.2}$$

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

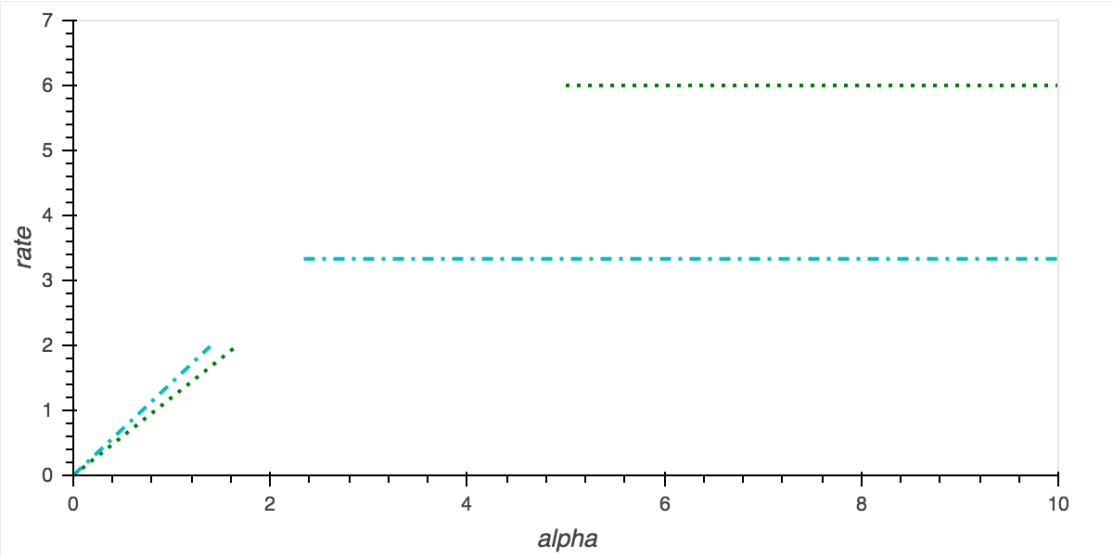

FIG. 3. *Decay rate $r(\alpha,\gamma) = \frac{2\gamma}{\gamma-2}$ depending on the value of α when $\alpha \geqslant \frac{\gamma+2}{\gamma-2}$ when F satisfies $\boldsymbol{H}_1(\gamma)$ (as in Theorem 4.5) for two values γ: $\gamma_3 = 3$ dotted line and $\gamma_4 = 5$ dashed-dotted line.*

Observe that the decay obtained in Corollary 4.6 is optimal since Attouch et al. proved that it is achieved for the function $F(x) = |x|^\gamma$ in [6].

From Theorems 4.1, 4.2 and 4.5, we can make the following comments: first in Theorems 4.2 and 4.5, both conditions $\mathbf{H}_1$ and $\mathbf{H}_2$ are used to get a decay rate and it turns out that these two conditions are important.

With the sole hypothesis $\mathbf{H}_2(\gamma)$ it seems difficult to establish optimal rate. Consider for instance the function $F(x) = |x|^3$ which satisfies $\mathbf{H}_1(3)$ and $\mathbf{H}_2(3)$. Applying Theorem 4.5 with $\gamma_1 = \gamma_2 = 3$, we know that for this function with $\alpha = \frac{\gamma_1+2}{\gamma_1-2} = 5$, we have $F(x(t)) - F^* = O\left(\frac{1}{t^6}\right)$. But, 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 4.2 we know that we cannot achieve a decay better than $\frac{1}{t^{\frac{2\alpha\gamma}{\gamma+2}}} = \frac{1}{t^5}$ for $\alpha = 5$.

Consider now a convex function F behaving like $\|x - x^*\|^\gamma$ in the neighborhood of its unique minimizer x^*. The decay of $F(x(t)) - F^*$ then depends directly on α if $\gamma \leqslant 2$, but it does not depend on α for large α if $\gamma > 2$. Moreover for such functions the best decay rate of $F(x(t)) - F^*$

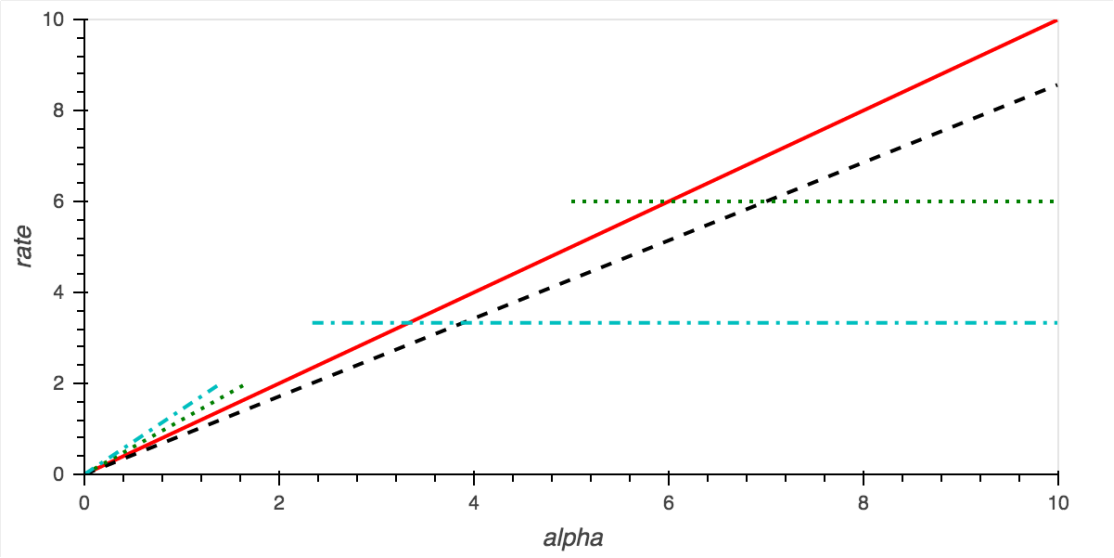

FIG. 4. *Decay rate $r(\alpha,\gamma)$ depending on the value of α if F satisfies $\boldsymbol{H}_1(\gamma)$ and $\boldsymbol{H}_2(r)$ with $r=\max(2,\gamma)$ for four values γ : $\gamma_1=1.5$ dashed line, $\gamma_2=2$, solid line, $\gamma_3=3$ dotted line and $\gamma_4=5$ dashed-dotted line.*

is $O\left(\frac{1}{t^\alpha}\right)$ and is achieved for $\gamma=2$ i.e. for quadratic like functions around the minimizer. If $\gamma<2$, it seems that the oscillations of the solution $x(t)$ prevent us from getting 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.

Finally, observe that as shown in Figures 3 and 4, the case when $1+\frac{2}{\gamma}<\alpha<\frac{\gamma+2}{\gamma-2}$ is not covered by our results. Although we did not get a better convergence rate than $\frac{1}{t^2}$ in that case, we can prove that there exist some initial conditions for which the convergence rate can not be better than $t^{-\frac{2\gamma\alpha}{\gamma+2}}$:

PROPOSITION 4.7. *Let $\gamma>2$ and $1+\frac{2}{\gamma}<\alpha<\frac{\gamma+2}{\gamma-2}$. Let x be a solution of (3.4) with $F(x)=|x|^\gamma$, $|x(t_0)|<1$ and $\dot{x}(t_0)=0$ for any given $t_0>0$. Then there exists $K>0$ such that for any $T>0$, there exists $t\geqslant T$ such that:*

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

Numerical Experiments. In the following numerical experiments, the optimality of the decays given in all previous theorems, are tested for various choices of α and γ.

More precisely we use a discrete Nesterov scheme to approximate the solution of (1.1) for $F(x)=|x|^\gamma$ on the interval $[t_0,T]$ with $t_0=0$ and $\dot{x}(t_0)=0$, see [26].

If $\gamma\geqslant 2$, ∇F is a Lipschitz function and we define the sequence $(x_n)_{n\in\mathbb{N}}$ as follows:

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

where $h\in(0,1)$ is a time step.

If $\gamma < 2$, we use a proximal step :

$$x_n = prox_{hF}(y_n) \text{ with } y_n = x_n + \frac{n}{n+\alpha}(x_n - x_{n-1}).$$

It has been shown that $x_n \approx x(n\sqrt{h})$ where the function x is a solution of the ODE (1.1).

In the following numerical experiments the sequence $(x_n)_{n\in\mathbb{N}}$ is computed for various pairs (γ, α). The step size is always set to $h = 10^{-7}$.

We define the function $rate(\alpha, \gamma)$ as the expected rate given in all the previous theorems and Proposition 4.7, that is:

$$rate(\alpha,\gamma) := \begin{cases} \dfrac{2\alpha\gamma}{\gamma+2} & \text{if } \gamma \leqslant 2 \text{ or if } \gamma > 2 \text{ and } \alpha \leqslant 1 + \frac{2}{\gamma}, \\ \dfrac{2\gamma}{\gamma-2} & \text{if } \gamma > 2 \text{ and } \alpha \geqslant \frac{\gamma+2}{\gamma-2}, \\ \dfrac{2\alpha\gamma}{\gamma+2} & \text{if} \gamma > 2 \text{ and } \alpha \in (1 + \frac{2}{\gamma}, \frac{\gamma+2}{\gamma-2}). \end{cases}$$

If the function $z(t) := (F(x(t)) - F(x^*))\, t^\delta$ is bounded but does not tend to 0, we can deduce that δ is the largest value such that $F(x(t)) - F(x^*) = O\left(t^{-\delta}\right)$. We define

$$z_n := (F(x_n) - F(x^*)) \times (n\sqrt{h})^{rate(\alpha,\gamma)} \approx (F(x(t)) - F(x^*))t^{rate(\alpha,\gamma)},$$

and if the function $rate(\alpha, \gamma)$ is optimal we expect that the sequence $(z_n)_{n\in\mathbb{N}}$ is bounded but do not decay to 0. The following figures give for various choices of (α, γ) the trajectory of the sequence $(z_n)_{n\in\mathbb{N}}$. The values are re-scaled such that the maximum is always 1. In all these numerical examples, we will observe that the sequence $(z_n)_{n\in\mathbb{N}}$ is bounded and does not tend to 0.

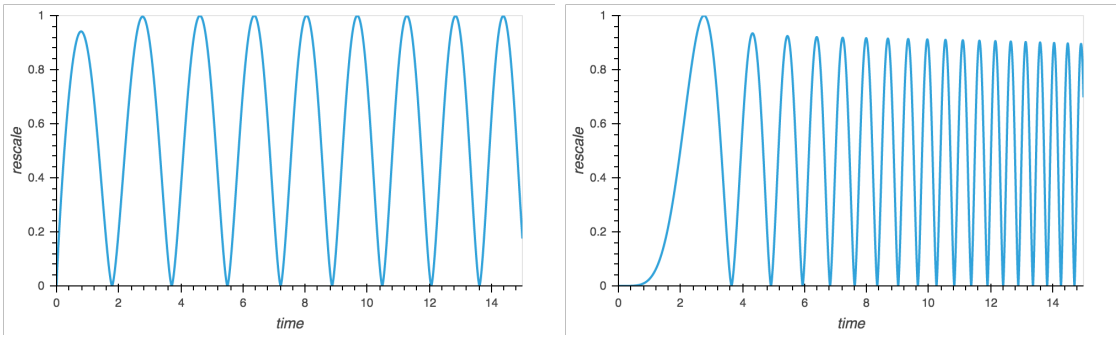

FIG. 5. *Case when* $\gamma = 1.5$. *On the left* $\alpha = 1$ *and* $rate(\alpha,\gamma) = \frac{2\alpha\gamma}{\gamma+2} = \frac{6}{7}$. *On the right* $\alpha = 6$ *and* $rate(\alpha,\gamma) = \frac{2\alpha\gamma}{\gamma+2} = \frac{36}{7}$

- The Figures 5 and 6 with $\gamma = 1.5$ and $\gamma = 2$ illustrate Theorem 4.1, Theorem 4.2 and Proposition 4.3. Indeed for sharp functions (i.e for $\gamma \leqslant 2$) the rate is proved to be optimal.
- In the case $\gamma = 3$ and $\alpha = 1$, the fact that $(F(x(t)) - F(x^*))t^{rate(\alpha,\gamma)}$ is bounded is also a consequence of Theorem 4.1. The optimality of this rate is not proven but the experiments show that it numerically is.

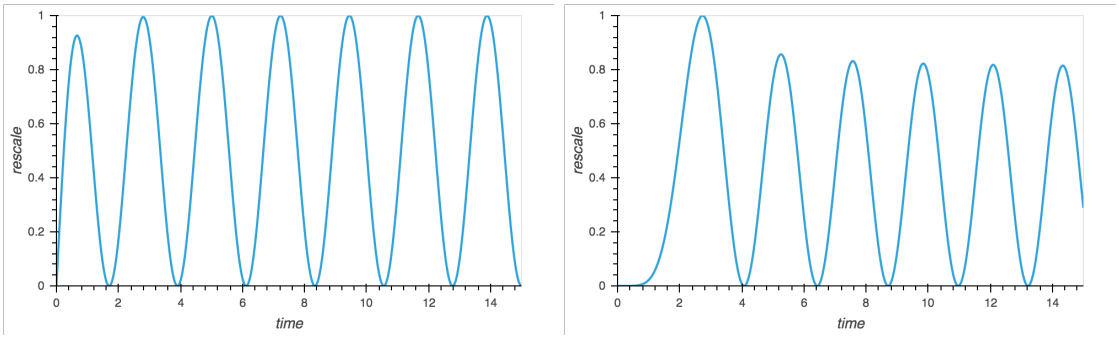

FIG. 6. *Case when* $\gamma = 2$. *On the left* $\alpha = 1$ *and* $rate(\alpha, \gamma) = \frac{2\alpha\gamma}{\gamma+2} = 1$. *On the right* $\alpha = 6$ *and* $rate(\alpha, \gamma) = \frac{2\alpha\gamma}{\gamma+2} = 6$

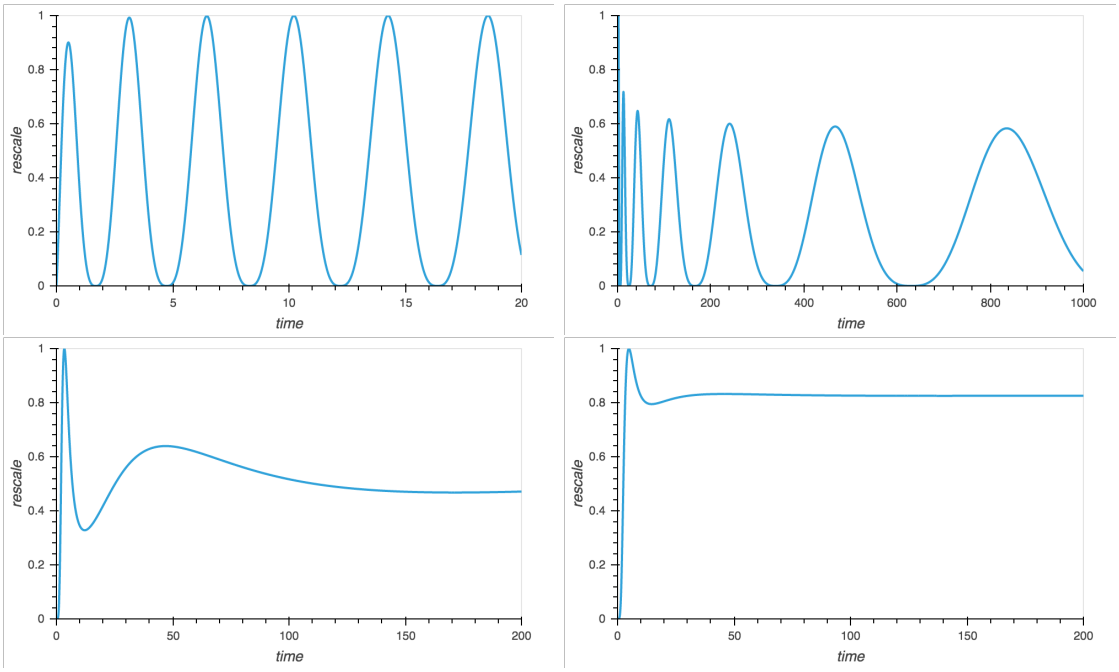

FIG. 7. *Case when* $\gamma = 3$. *On the top left* $\alpha = 1$ *and* $rate(\alpha, \gamma) = \frac{2\alpha\gamma}{\gamma+2} = 1.2$, *on the top right* $\alpha = 4$ *and* $rate(\alpha, \gamma) = \frac{2\alpha\gamma}{\gamma+2} = 4.8$, *on bottom left* $\alpha = 6$ *and* $rate(\alpha, \gamma) = \frac{2\gamma}{\gamma-2} = 6$, *on bottom right* $\alpha = 8$ *and* $rate(\alpha, \gamma) = \frac{2\gamma}{\gamma-2} = 6$

- In the case $\gamma = 3$ and $\alpha = 4$, $\alpha \in (\frac{\gamma+2}{\gamma}, \frac{\gamma+2}{\gamma-2})$ then the fact that $(F(x(t)) - F(x^*))t^{rate(\alpha,\gamma)}$ is bounded is not proved but the experiments from Figure 7 show that it numerically is. However Proposition 4.7 proves that the sequence $(z_n)_{n\in\mathbb{N}}$ does not tend to 0, which is illustrated by the experiments.
- When $\gamma = 3$ and $\alpha = 6$ or $\alpha = 8$, Theorem 4.5 ensures that the sequence $(z_n)_{n\in\mathbb{N}}$ is bounded. This rate is proved to be optimal and the numerical experiments from Figure

7 show that this rate is actually achieved for this specific choice of parameters.

5. Proofs. In this section, we detail the proofs of the results presented in Section 4, namely Theorems 4.1, 4.2 and 4.5, Propositions 4.3 and 4.7, Corollary 4.6.

The proofs of the theorems rely on Lyapunov functions $\mathcal{E}$ and $\mathcal{H}$ introduced by Su, Boyd and Candes [26], Attouch, Chbani, Peypouquet and Redont [6] and Aujol-Dossal [10] :

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

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 :

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

Using the following notations:

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

we have:

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

From now on we will choose

$$\xi = \lambda(\lambda + 1 - \alpha),$$

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

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

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

Note that this inequality is actually an equality for the specific choice $F(x) = |x|^\gamma$, $\gamma > 1$.

5.1. Proof of Theorems 4.1 and 4.2. In this section we prove Theorem 4.1 and Theorem 4.2. Note that a complete proof of Theorem 4.1, including the optimality of the rate, can be found in the unpublished report [10] under the hypothesis that $(F - F^*)^{\frac{1}{\gamma}}$ is convex. The proof of both Theorems are actually similar. The choice of p and λ are the same but, to prove the first point, 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.

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

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

From Lemma 5.1, it appears that:

$$\mathcal{H}'(t) \leqslant K_1 t^p c(t) \tag{5.1}$$

where the real constant K_1 is given by:

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

Hence:

$$K_1 = \frac{4\alpha\gamma}{(\gamma+2)^3}\left(1+\frac{2}{\gamma}-\alpha\right)\left(\alpha(-2+\gamma)-\gamma-2\right). \tag{5.2}$$

Consider first the case when: $\alpha \leqslant 1+\frac{2}{\gamma}$. In that case, we observe that: $\xi \geq 0$, so that the energy $\mathcal{H}$ is actually a sum of non-negative terms. Coming back to (5.1), we have:

$$\mathcal{H}'(t) \leqslant K_1 t^p c(t). \tag{5.3}$$

Since $\alpha \leqslant 1+\frac{2}{\gamma}$, the sign of the constant K_1 is the same as that of $\alpha(-2+\gamma)-\gamma-2$, and thus $K_1 \leqslant 0$ for any $\gamma \geqslant 1$. According to (5.3), the energy $\mathcal{H}$ is thus non-increasing and bounded i.e.:

$$\forall t \geqslant t_0,\ \mathcal{H}(t) \leqslant \mathcal{H}(t_0).$$

Since $\mathcal{H}$ is a sum of non-negative terms, it follows directly that:

$$\forall t \geqslant t_0,\ t^{p+2}(F(x(t))-F^*) \leqslant \mathcal{H}(t_0),$$

which concludes the proof of Theorem 4.1.

Consider now the case when: $\alpha > 1+\frac{2}{\gamma}$. In that case, we first observe that: $\xi < 0$, so that $\mathcal{H}$ is not a sum of non-negative functions anymore, and an additional growth condition $\mathbf{H}_2(2)$ will be needed to bound the term in $\|x(t)-x^*\|^2$. Coming back to (5.1), we have:

$$\mathcal{H}'(t) \leqslant K_1 t^p c(t). \tag{5.4}$$

Since $\alpha > 1+\frac{2}{\gamma}$, the sign of the constant K_1 is the opposite of the sign of $\alpha(\gamma-2)-(\gamma+2)$. Moreover, since $\gamma \leqslant 2$, then $\alpha(\gamma-2)-(\gamma+2) < 0$ and thus $K_1 > 0$.

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

$$Kt\,\|x(t)-x^*\|^2 \leqslant t(F(x(t))-F^*) = a(t),$$

and thus

$$c(t) \leqslant \frac{1}{2Kt^2}a(t). \tag{5.5}$$

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

$$\mathcal{H}(t) \geqslant t^{p+1}(a(t)+\xi c(t)) \geqslant t^{p+1}(1+\frac{\xi}{2Kt^2})a(t). \tag{5.6}$$

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

$$\mathcal{H}(t) \geqslant \frac{1}{2}t^{p+1}a(t). \tag{5.7}$$

From (5.4), (5.5) and (5.7), we get:

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

From the Grönwall Lemma in its differential form, there exists $A > 0$ such that for all $t \geqslant t_1$, we have: $\mathcal{H}(t) \leqslant A$. According to (5.7), we then conclude that $t^{p+2}(F(x(t)) - F^*) = t^{p+1}a(t)$ is bounded which concludes the proof of Theorem 4.2.

5.2. Proof of Proposition 4.3 (Optimality of the convergence rates). Before proving the optimality of the convergence rate stated in Proposition 4.3, 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:*

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

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 4.3: 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 $F(x) = |x|^\gamma$, $\gamma > 1$, the inequality of Lemma 5.1 is actually an equality. Using the values $p = \frac{2\gamma\alpha}{\gamma+2} - 2$ and $\lambda = \frac{2\alpha}{\gamma+2}$ of Theorems 4.1 and 4.2, we have a closed form for the derivative of function $\mathcal{H}$:

$$\mathcal{H}'(t) = K_1 t^p c(t) = \frac{K_1}{2} t^{p-1}|x(t)|^2, \tag{5.8}$$

where K_1 is the constant given in (5.2). We will now prove that it exists $\ell > 0$ such that for t large enough:

$$\mathcal{H}(t) \geqslant \ell.$$

To prove that point we consider two cases depending on the sign of $\alpha - (1 + \frac{2}{\gamma})$.

1. Case when $\alpha \leqslant 1 + \frac{2}{\gamma}$, $\xi \geqslant 0$ and $K_1 \leqslant 0$. We can first observe that $\mathcal{H}$ is a non negative and non increasing function. Moreover it exists $\tilde{t} \geqslant t_0$ such that for $t \geqslant \tilde{t}$, $|x(t)| \leqslant 1$ and:

$$t^p c(t) \leqslant \frac{t^p a(t)}{2t^2} \leqslant \frac{\mathcal{H}(t)}{t^3},$$

which implies using (5.8) that:

$$|\mathcal{H}'(t)| \leqslant |K_1|\frac{\mathcal{H}(t)}{t^3}.$$

If we denote $G(t) = \ln(\mathcal{H}(t))$ we get for all $t \geqslant \tilde{t}$,

$$|G(t) - G(\tilde{t})| \leqslant \int_{\tilde{t}}^{t} \frac{|K_1|}{s^3} ds.$$

We deduce that $|G(t)|$ is bounded below and then that it exists $\ell > 0$ such that for t large enough:

$$\mathcal{H}(t) \geqslant \ell,$$

2. Case when $\alpha > 1 + \frac{2}{\gamma}$, $\xi < 0$ and $K_1 > 0$. This implies in particular that $\mathcal{H}$ is non-decreasing. Moreover, from Theorem 4.2, $\mathcal{H}$ is bounded above. Coming back to the inequality (5.6), we observe that $\mathcal{H}(t_0) > 0$ provided that $1 + \frac{\xi}{2t_0^2} > 0$, with $K = 1$ and $\xi = \lambda(\lambda - \alpha + 1)$, i.e.:

$$t_0 > \sqrt{\frac{\alpha\gamma}{(\gamma+2)^2}(\alpha - (1 + \frac{2}{\gamma}))}.$$

In particular, we have that for any $t \geqslant t_0$

$$\mathcal{H}(t) \geqslant \ell,$$

with $\ell = \mathcal{H}(t_0)$

Hence for any $\alpha > 0$ and for t large enough

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

Moreover, since $c(t) = o(a(t))$ when $t \to +\infty$, we have that for t large enough,

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

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

$$y(t) := t^\lambda x(t),$$

where: $\lambda = \frac{2\alpha}{\gamma+2}$. From the Theorem 4.1 and Theorem 4.2, we know that $y(t)$ is bounded. Hence, from Lemma 5.2, there exists $t_1 > T$ such that

$$|\dot{y}(t_1)| \leqslant \frac{\epsilon}{t_1}. \tag{5.9}$$

But:

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

Hence using (5.9):

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

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

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

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

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

For $\epsilon = \sqrt{\frac{\ell}{2}}$ 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 4.5. We detail here the proof of Theorem 4.5.

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:

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

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^*,

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

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

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

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

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

which implies that

$$t^{\frac{4}{\gamma_1-2}} t^2 (F(x(t)) - F^*) \leqslant \frac{|\xi|}{2} t^{\frac{4}{\gamma_1-2}} d(x(t), X^*)^2 + A. \tag{5.10}$$

We now set:

$$v(t) := t^{\frac{4}{\gamma_2-2}} d(x(t), X^*)^2. \tag{5.11}$$

Using (5.10) we have:

$$t^{\frac{2\gamma_1}{\gamma_1-2}} (F(x(t)) - F^*) \leqslant \frac{|\xi|}{2} t^{\frac{4}{\gamma_1-2} - \frac{4}{\gamma_2-2}} v(t) + A. \tag{5.12}$$

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

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

which is equivalent to

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

Hence:

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

Using (5.12), we obtain:

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

i.e.:

$$Kv(t)^{\frac{\gamma_2}{2}} \leqslant \frac{|\xi|}{2}v(t)+At^{\frac{4}{\gamma_2-2}-\frac{4}{\gamma_1-2}}. \tag{5.13}$$

Since $2<\gamma_1\leqslant\gamma_2$, we deduce that v is bounded. Hence, using (5.12) there exists some positive constant B such that:

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

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

5.4. Proof of Corollary 4.6. We are now in position to prove Corollary 4.6. The first point of Corollary 4.6 is just a particular instance of Theorem 4.5. In the sequel, we prove the second point of Corollary 4.6.

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

$$\|x(t)-\tilde{x}\|=d(x(t),X^*).$$

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

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

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

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

which is equivalent to

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

where $v(t)$ is defined in (5.11) with $\gamma=\gamma_2$. Using the fact that the function v is bounded (a consequence of (5.13)) we deduce that there exists a positive constant $A_1>0$ such that:

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

Thus:

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

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

$$\|\dot{x}(t)\|\leqslant\frac{A_2}{t^{\frac{\gamma}{\gamma-2}}},$$

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

5.5. Proof of Proposition 4.7. The idea of the proof is very similar to that of Proposition 4.3 (optimality of the convergence rate in the sharp case i.e. when $\gamma \in (1,2]$).

For the exact same choice of parameters $p = \frac{2\gamma\alpha}{\gamma+2} - 2$ and $\lambda = \frac{2\alpha}{\gamma+2}$ and assuming that $1 + \frac{2}{\gamma} < \alpha < \frac{\gamma+2}{\gamma-2}$, we first show that the energy $\mathcal{H}$ is non-decreasing and then:

$$\forall t \geqslant t_0,\ \mathcal{H}(t) \geqslant \ell, \tag{5.14}$$

where: $\ell = \mathcal{H}(t_0) > 0$. Indeed, since $\gamma > 2$ and $\alpha < \frac{\gamma+2}{\gamma-2}$, a straightforward computation shows that: $\lambda^2 - |\xi| > 0$, so that:

$$\begin{aligned}\mathcal{H}(t_0) &= t_0^{p+2}|x(t_0)|^\gamma + \frac{t_0^p}{2}\left(|\lambda x(t_0) + t_0\dot{x}(t_0)|^2 - |\xi||x(t_0)|^2\right)\\ &= t_0^{p+2}|x(t_0)|^\gamma + \frac{t_0^p}{2}\left(\lambda^2 - |\xi|\right)|x(t_0)|^2 > 0,\end{aligned}$$

without any additional assumption on the initial time $t_0 > 0$.

Let $T > t_0$. We set: $y(t) = t^\lambda x(t)$. If $y(t)$ is bounded as it is in Proposition 4.3, by the exact same arguments, we prove that there exists $t_1 > T$ such that: $b(t_1) \leq \frac{\ell}{4t_1^{p+1}}$. Moreover since $\xi < 0$ we deduce from (5.14) that:

$$t_1^{p+1}(a(t_1) + b(t_1)) \geqslant \ell.$$

Hence:

$$a(t_1) = t_1(F(x(t_1) - F^*) \geqslant \frac{\ell}{4t_1^{p+1}},$$

i.e.: $F(x(t_1)) - F^* \geqslant \frac{\ell}{4t_1^{p+2}} = \frac{\ell}{4t_1^{\frac{2\alpha\gamma}{\gamma+2}}}$.

If $y(t)$ is not bounded, then the proof is even simpler: indeed, in that case, for any $K > 0$, there exists $t_1 \geqslant T$ such that: $y(t_1) \geq K$, hence:

$$F(x(t_1)) - F^* = |x(t_1)|^\gamma \geqslant \frac{K}{t_1^{\lambda\gamma}} = \frac{K}{t_1^{\frac{2\alpha\gamma}{\gamma+2}}},$$

which concludes the proof.

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) and partially supported by ANR-11-LABX-0040-CIMI within the program ANR-11-IDEX-0002-02. 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]).

Lemma A.1.

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

Proof. Let us differentiate the energy $\mathcal{E}$:

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

Using the ODE (1.1), we get:

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

Observing now that:

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

we can write:

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

Corollary A.2. *If F satisfies the hypothesis $\boldsymbol{H}_1(\gamma)$ and if $\xi = \lambda(\lambda+1-\alpha)$, then:*

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

Proof. Choosing $\xi = \lambda(\lambda+1-\alpha)$ in Lemma A.1, we get:

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

Applying now the assumption $\mathbf{H}_1(\gamma)$, we finally obtain the expected result.

One can notice that if $F(x) = |x|^\gamma$ the inequality of Lemma A.1 is actually an equality when $\xi = \lambda(\lambda+1-\alpha)$. This ensures that for this specific function F, the inequality in Lemma 5.1 is an equality.

Lemma A.3. *If $F(x) = |x|^\gamma$ and if $\xi = \lambda(\lambda+1-\alpha)$, then*

$$\begin{aligned}\mathcal{H}'(t) = t^p\big[&(2+p)a(t) + \lambda t\langle -\nabla F(x(t)), x(t)-x^*\rangle + (2\lambda+2-2\alpha+p)b(t) \\ &+\lambda(\lambda+1-\alpha)(-2\lambda+p)c(t)\big]\end{aligned}$$

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.1. ∎

In conclusion, to prove Lemma 5.1, it is sufficient to plug the assumption $\mathbf{H}_1(\gamma)$ into the equality of Lemma A.3.

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 (2018), pp. 551–574.

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

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

[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 (2002), pp. 273–306.

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

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

[7] H. Attouch, Z. Chbani, and H. Riahi, *Rate of convergence of the Nesterov accelerated gradient method in the subcritical case $\alpha \leqslant 3$*, ESAIM: COCV, (2019), https://doi.org/10.1051/cocv/2017083, https://doi.org/10.1051/cocv/2017083.

[8] 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 (2000), pp. 1–34.

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

[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 hal-01547251, (2017).

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

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

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

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

[15] 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 (2009), pp. 5983–6017.

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

[17] 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 (2015), pp. 968–982.

[18] 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).

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

[20] A. Y. Kruger, *Error bounds and hölder metric subregularity*, Set-Valued and Variational Analysis, 23 (2015), pp. 705–736.

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

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

[23] R. May, *Asymptotic for a second order evolution equation with convex potential and vanishing damping term*, Turkish Journal of Mathematics, 41 (2017), pp. 681–685.

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

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

[26] 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 (2016), pp. 1–43.