Strong Convergence of FISTA Iterates under Hölderian and Quadratic Growth Conditions

Jean-François Aujol, Charles Dossal, Hippolyte Labarrière, Aude Rondepierre

Strong Convergence of FISTA Iterates under Hölderian and Quadratic Growth Conditions

J.-F. Aujol* C. Dossal† H. Labarrière‡ A. Rondepierre†§

July 23, 2024

Abstract

Introduced by Beck and Teboulle in [10], FISTA (for Fast Iterative Shrinkage-Thresholding Algorithm) is a first-order method widely used in convex optimization. Adapted from Nesterov's accelerated gradient method for convex functions [29], the generated sequence guarantees a decay of the function values of $\mathcal{O}\left(n^{-2}\right)$ in the convex setting. We show that for coercive functions satisfying some local growth condition (namely a Hölderian or quadratic growth condition), this sequence strongly converges to a minimizer. This property, which has never been proved without assuming the uniqueness of the minimizer, is associated with improved convergence rates for the function values. The proposed analysis is based on a preliminary study of the Asymptotic Vanishing Damping system introduced by Su et al. in [33] to model Nesterov's accelerated gradient method in a continuous setting. Novel improved convergence results are also shown for the solutions of this dynamical system, including the finite length of the trajectory under the aforementioned geometry conditions.

1 Introduction

Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) is a well-known scheme introduced by Beck and Teboulle in [10] for the minimization of convex composite functions. Considering a convex function $F : \mathcal{H} \to \mathbb{R}$ where $\mathcal{H}$ is a Hilbert space, F is called composite if it can be written $F = f + h$ where f is a convex differentiable function having a L-Lipchitz gradient and h is is a proper lower semicontinuous (l.s.c.) convex function.

This method uses inertia to achieve acceleration, based on the ideas proposed by Nesterov in the convex setting [29]. While the classical proximal gradient method (also called Forward-Backward [16]) guarantees a decrease of the error of order $\mathcal{O}\left(n^{-1}\right)$, FISTA builds a sequence $(x_n)_{n\in\mathbb{N}}$ which ensures that if F is a convex composite function, then

$$F(x_n) - F^* \leqslant \frac{2L\|x_0 - x^*\|^2}{(n+1)^2}, \tag{1}$$

for any minimizer x^* of X^* where $F^* = \min_{x\in\mathcal{H}} F(x)$. The question of the convergence of FISTA iterates remained unanswered for a few years before Chambolle and D. show in [15] that for a slightly modified inertial term depending on a non negative real number $\alpha > 3$, the sequence $(x_n)_{n\in\mathbb{N}}$ weakly converges to a minimizer of F. This $\mathcal{O}\left(n^{-2}\right)$ rate can be improved to a $o\left(n^{-2}\right)$ rate for this variation of FISTA proposed by Chambolle and D. as demonstrated by Attouch and Peypouquet in [5] but no first order method can guarantee a decrease of the error faster than this rate for this class of functions as shown in [27].

*Univ. Bordeaux, Bordeaux INP, CNRS, IMB, UMR 5251, F-33400 Talence, France
†IMT, Univ. Toulouse, INSA Toulouse, Toulouse, France
‡MaLGa, DIBRIS, Università di Genova, Genoa, Italy
§LAAS, Univ. Toulouse, CNRS, Toulouse, France

Better convergence guarantees can be proven by making stronger assumptions on the function F and by conveniently adjusting the inertial parameter. Su et al. [33] show that FISTA iterates can achieve a rate of $F(x_n) - F^* = \mathcal{O}\left(n^{-3}\right)$ for strongly convex functions. Attouch and Cabot [4] improve this result as they prove that the error decreases as $\mathcal{O}\left(n^{-\frac{2\alpha}{3}}\right)$ for any $\alpha > 0$ as long as F has a strong minimum, i.e. F has a unique minimizer x^* and a **global quadratic growth**:

$$\exists \mu > 0,\ \forall x \in \mathcal{H},\ \frac{\mu}{2}\|x - x^*\|^2 \leqslant F(x) - F^*. \tag{2}$$

An additional flatness condition which requires the differentiability of F allows to strengthen this convergence guarantee as shown in [1, 8]. In [1], Apidopoulos et al. give an improved convergence rate of the error under the aforementioned flatness condition, a uniqueness assumption on the minimizer x^* of F and a **Hölderian error bound** hypothesis:

$$\exists \gamma > 2,\ \exists K > 0,\ \forall x \in B(x^*, \varepsilon),\ K\|x - x^*\|^\gamma \leqslant F(x) - F^*. \tag{3}$$

The works mentioned above mainly focus on finding the fastest convergence rate and since every improved result relies on the hypothesis that F has a unique minimizer x^*, the strong convergence of FISTA iterates is actually trivial (under these hypotheses, $F(x_n) - F^* \to 0$ implies that $\|x_n - x^*\| \to 0$).

This observation is also true when considering the study of the corresponding ordinary differential equation (ODE) i.e. Asymptotic Vanishing Damping system (AVD) defined by

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

Introduced by Su et al. in [33] as a system which can be discretized to recover Nesterov's accelerated gradient method, this ODE shares most of its convergence properties with FISTA iterates. Several papers (see [3, 7, 8, 26]) are devoted to its analysis under geometry assumptions and most of the fast convergence results require F to have a unique minimizer x^* which automatically guarantees that $\|x(t) - x^*\| \to 0$.

In this paper, we analyse theoretically the behavior of FISTA iterates and its corresponding ODE under Hölderian and quadratic growth assumptions without any hypothesis on the uniqueness of the minimizer. Indeed in this geometrical setting, the strong convergence of the iterates $(x_n)_{n\in\mathbb{N}}$ (resp the trajectory $x(\cdot)$) is no longer a consequence of the decay of $(F(x_n))_{n\in\mathbb{N}}$ (resp $F(x(\cdot))$) but a consequence of the bounds of $(\|x_n - x_{n-1}\|)_{n\in\mathbb{N}}$ (resp $\|\dot{x}(\cdot)\|$). The main contributions are the following :

1. Strong convergence of FISTA iterates for functions having a local Hölderian growth (3) with parameter $\gamma > 2$ (for a well-chosen inertial parameter). In addition, we prove that the error $F(x_n) - F^*$ decreases as $\mathcal{O}\left(n^{-\frac{2\gamma}{\gamma-2}}\right)$.

2. Strong convergence of FISTA iterates for functions having a quadratic growth (2) (for a well-chosen inertial parameter) and non-asymptotic bound on the error if this assumption is global. We recover the convergence rate proved if F has a unique minimizer i.e.

$$F(x_n) - F^* = \mathcal{O}\left(n^{-\frac{2\alpha}{3}}\right),$$

   for α sufficiently large.

3. Finite trajectory of the solution of (AVD) under Hölderian or quadratic growth without a uniqueness assumption on the minimizers of F. We show that if the set of minimizers X^* is sufficiently regular, the error along the trajectories decreases respectively as $\mathcal{O}\left(t^{-\frac{2\gamma}{\gamma-2}}\right)$ or $\mathcal{O}\left(t^{-\frac{2\alpha}{3}}\right)$ for the aforementioned assumptions if α is sufficiently large.

The paper is organized as follows. Section 2 presents key geometry concepts used in the paper before giving an overview of the literature on FISTA and the Asymptotic Vanishing Damping system. The main results on the strong convergence of FISTA iterates are then stated and discussed in Section 3. Section 4 contains the analogous convergence results obtained for the trajectories of the Asymptotic Vanishing Damping system. The proofs of the main theorems are given in Section 5 while the other demonstrations are postponed to Appendix A and Appendix B.

2 Preliminaries and State of the Art

Let $\mathcal{H}$ be a Hilbert space. This work focuses on the class $\mathcal{C}$ of composite functions defined by:

Definition 1. *Let $\mathcal{C}$ be the class of convex functions F defined from $\mathcal{H}$ to $\mathbb{R} \cup \{+\infty\}$ such that $F = f + h$, where f is a convex differentiable function having a L-Lipschitz gradient, and h is a convex function whose proximal operator is known. The set of minimizers X^* of F is non-empty but not necessarily reduced to one point.*

This set $\mathcal{C}$ depends on the non negative real number L, but to lighten the notation and because there is no ambiguity, we choose the simple notation $\mathcal{C}$.

2.1 Geometry of convex functions

In this paper we consider the general class of convex composite functions satisfying some growth condition in the neighborhood of their sets of minimizers:

Definition 2 (*Local growth conditions*)**.** *Let $F : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous convex function with a non-empty set of minimizers X^*. Let $F^* = \min_{x \in \mathcal{H}} F(x)$. The function F is said to satisfy a Hölderian growth condition $\mathcal{G}^{\gamma}_{loc}$ for some $\gamma > 2$ if there exist $K > 0$ and $\varepsilon > 0$ such that for all $x \in \mathcal{H}$ satisfying $d(x, X^*) \leqslant \varepsilon$, we have:*

$$K d(x, X^*)^{\gamma} \leqslant F(x) - F^*. \tag{4}$$

Moreover, the function F satisfies a local quadratic growth condition $\mathcal{G}^2_{\mu,loc}$ for some $\mu > 0$ if there exists $\varepsilon > 0$ such that for all $x \in \mathcal{H}$ satisfying: $d(x, X^*) \leqslant \varepsilon$, we have:

$$\frac{\mu}{2} d(x, X^*)^2 \leqslant F(x) - F^*. \tag{5}$$

In the context of finite-time analysis, we also introduce the global version of these growth conditions:

Definition 3 (*Global growth conditions*)**.** *Let $F : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous convex function with a non-empty set of minimizers X^*. Let $F^* = \min_{x \in \mathcal{H}} F(x)$. The function F satisfies the growth condition $\mathcal{G}^{\gamma}$ for some $\gamma > 2$ if there exists $K > 0$ such that:*

$$\forall x \in \mathcal{H}, \quad K d(x, X^*)^{\gamma} \leqslant F(x) - F^*. \tag{6}$$

Moreover, the function F satisfies a quadratic growth condition $\mathcal{G}^2_{\mu}$ for some $\mu > 0$ if:

$$\forall x \in \mathcal{H}, \quad \frac{\mu}{2} d(x, X^*)^2 \leqslant F(x) - F^*. \tag{7}$$

The growth conditions $\mathcal{G}^{\gamma}_{\text{loc}}$ ($\gamma \geqslant 2$) can be seen as sharpness assumptions on the function F characterizing functions behaving at least as $\|\cdot\|^{\gamma}$ in the neighborhood of their minimizers. In the convex setting, the class of functions satisfying some growth condition is a subclass of the functions having a Łojasiewicz property [24, 25], a key tool for the mathematical analysis of continuous and discrete dynamical systems. Initially introduced to prove the convergence of the trajectories for the gradient flow of analytic functions, an extension to nonsmooth functions has been proposed by Bolte et al. in [11, 12]:

Definition 4 (The Łojasiewicz property). *Let $F : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous convex function with a non-empty set of minimizers X^*. Let $F^* = \min_{x\in\mathcal{H}} F(x)$. The function F has a Łojasiewicz property if for any minimizer x^*, there exist $\theta \in [0, 1)$, $c > 0$, $\varepsilon > 0$ such that:*

$$\forall x \in B(x^*, \varepsilon),\ c\left(F(x) - F^*\right)^\theta \leqslant d(0, \partial F(x)). \tag{8}$$

Let us finally introduce the notion of flatness characterizing differentiable functions that are at least as flat as $\|\cdot\|^\gamma$ with $\gamma > 1$:

$$\forall x^* \in X^*, \quad \forall x \in \mathcal{H}, \quad F(x) - F^* \leqslant \frac{1}{\gamma} \langle \nabla F(x), x - x^* \rangle, \tag{$\mathcal{F}_\gamma$}$$

where $F^* = \min_{x\in\mathcal{H}} F(x)$. Note that if F is convex, then it satisfies $(\mathcal{F}_\gamma)$ for $\gamma = 1$. This notion is recalled here to enable latter comparisons, particularly with the convergence results presented in [1].

To conclude this section, observe that in the context of local growth assumptions, the convergence of the sequence of $(F(x_n) - F^*)_{n\in\mathbb{N}}$ to 0 does not trivially imply the convergence of a given sequence of iterates $(x_n)_{n\in\mathbb{N}}$ to the set of minimizers X^*. The coercivity of F is needed to conclude:

Lemma 1. *Let $F \in \mathcal{C}$ be a coercive function satisfying a local growth condition $\mathcal{G}^\gamma_{loc}$ for some real parameters $\gamma \geqslant 2$ and $K > 0$. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence of iterates generated by a given algorithm $\mathcal{A}$.*

If the sequence $(F(x_n) - F^*)_{n\in\mathbb{N}}$ converge to 0, then $\left(d(x_n, X^*)\right)_{n\in\mathbb{N}}$ converges to 0 and:

$$\exists N \in \mathbb{N},\ \forall n \geqslant N,\ F(x_n) - F^* \geqslant K d(x_n, X^*)^\gamma.$$

Proof. Assume that the sequence $\left(d(x_n, X^*)\right)_{n\in\mathbb{N}}$ does not converge to 0. Thus, there exists $\varepsilon > 0$ and a non-decreasing function $\phi : \mathbb{N} \to \mathbb{N}$ such that the sub-sequence $(x_{\phi(n)})_{n\in\mathbb{N}}$ satisfies:

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

Since the sequence $\left(F(x_n) - F^*\right)_{n\in\mathbb{N}}$ is assumed to converge to 0 , it is also bounded. Combined with the coercivity of F, this implies that the sequence $(x_n)_{n\in\mathbb{N}}$ is bounded too. Therefore, there exists a closed bounded set C containing X^* such that

$$\{x_n,\ n \in \mathbb{N}\} \subset C. \tag{9}$$

Let $K_\varepsilon = C \cap \{x \in \mathcal{H},\ d(x, X^*) \geqslant \varepsilon\}$. By construction, K_ε is a weakly compact subset of $\mathcal{H}$ and $K_\varepsilon \cap X^* = \emptyset$. Moreover, for all $n \in \mathbb{N}$, we have $x_{\phi(n)} \in K_\varepsilon$ so there exists a weakly convergent sub-sequence $\left(x_{\psi\circ\phi(n)}\right)_{n\in\mathbb{N}}$ whose weak limit denoted by $\tilde{x}$ belongs to K_ε and thus $\tilde{x} \notin X^*$.

Consequently, since F is convex and lower semi-continuous (we remind the reader that when F is convex, then F is weak lsc if and only if F is strong lsc, see e.g. [14]),we have

$$\liminf F\left(x_{\psi\circ\phi(n)}\right) - F^* \geqslant F(\tilde{x}) - F^*. \tag{10}$$

Since the whole sequence $\left(F(x_n) - F^*\right)_{n\in\mathbb{N}}$ tends to 0 when $n \to +\infty$, and since $F(\tilde{x}) \geqslant F^*$, it implies that $F(\tilde{x}) - F^* = 0$ which is impossible since $\tilde{x} \notin X^*$. Thus, the sequence $\left(d(x_n, X^*)\right)_{n\in\mathbb{N}}$ converges to 0 as $n \to +\infty$. □

This technical lemma will be useful throughout the paper to establish new convergence rates for the class of composite functions satisfying certain growth conditions, without assuming the uniqueness of the minimizer.

2.2 FISTA and its variants

To solve the minimization problem

$$\min_{x\in\mathcal{H}} F(x), \tag{11}$$

where F is a convex composite function in the class $\mathcal{C}$ (see Definition 1), a classical algorithm is the **Proximal Gradient method** also called **Forward-Backward** [16]. Before defining properly this scheme, it is necessary to introduce the notion of proximal operator. Considering $h : \mathcal{H} \to \mathbb{R}\cup\{+\infty\}$ a proper lower semicontinuous convex function, its proximal operator denoted prox_h is defined for all $x \in \mathcal{H}$ as

$$\text{prox}_h(x) = \arg\min_{y\in\mathcal{H}} h(y) + \frac{1}{2}\|x-y\|^2. \tag{12}$$

Given an initialization $x_0 \in \mathcal{H}$, the iterates of the Proximal Gradient method are defined as

$$\forall n \in \mathbb{N}, \quad x_{n+1} = \text{prox}_{sh}\left(x_n - s\nabla f(x_n)\right), \tag{13}$$

where the step size $s > 0$ should be chosen smaller than $\frac{1}{L}$ to ensure that $F(x_n) - F^* = \mathcal{O}\left(n^{-1}\right)$. In 2009 Beck and Teboulle introduce in [10] the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) for the same class of functions. While the Proximal Gradient method is the composite extension of the Gradient Descent method in the differentiable setting, FISTA is a generalization of Nesterov's accelerated gradient method for convex functions [29]. Indeed, the iterates of FISTA are defined in the following way:

$$x_0 \in \mathcal{H}, \quad \forall n \in \mathbb{N}, \quad \begin{cases} y_n = x_n + \alpha_n\left(x_n - x_{n-1}\right) \\ x_{n+1} = \text{prox}_{sh}\left(y_n - s\nabla f(y_n)\right), \end{cases} \tag{14}$$

where $x_{-1} = x_0$ and the sequence $(\alpha_n)_{n\in\mathbb{N}}$ is that defined by Nesterov in [29] as:

$$\alpha_0 = 0, \quad \forall n \in \mathbb{N}, \begin{cases} t_{n+1} = \dfrac{1+\sqrt{1+4t_n^2}}{2} \\ \alpha_{n+1} = \dfrac{t_n - 1}{t_{n+1}}, \end{cases} \tag{15}$$

where $t_0 = 1$. The authors prove that $F(x_n) - F^* = \mathcal{O}\left(n^{-2}\right)$ for $s \in \left(0, \frac{1}{L}\right)$ (and in particular (1) for $s = \frac{1}{L}$). Although this convergence rate reveals a significant improvement over Proximal Gradient method, the authors do not show the weak convergence of the iterates.

This property of the sequence $(x_n)_{n\in\mathbb{N}}$ is proved by Chambolle and Dossal in [15] for a slightly different version of FISTA, choosing $(\alpha_n)_{n\in\mathbb{N}}$ as $\alpha_n = \frac{n}{n+\alpha}$ with $\alpha > 3$. Attouch and Peypouquet show in [5] that this choice for α ensures that $F(x_n) - F^* = o\left(n^{-2}\right)$.

Remark 1. *The sequence $(\alpha_n)_{n\in\mathbb{N}}$ introduced by Chambolle and Dossal (defined as $\alpha_n = \frac{n}{n+\alpha}$ with $\alpha > 3$) and that given by Nesterov (i.e. (15)) have a similar behavior when $\alpha = 3$. In practice, the Chambolle-Dossal formulation is more convenient to draw a parallel with the continuous setting (see Section 4) and to obtain improved convergence properties under additional geometry assumptions, while the Nesterov formulation facilitates the implementation of linesearch strategies.*

Note that Kim and Fessler introduce Optimized Gradient Method in [20] (and a proximal version in [21]) which also ensures a decrease of the error of order $\mathcal{O}\left(n^{-2}\right)$ but with a tightest and optimal bound in the differentiable case.

Remark 2 (Why so many names ?)**.** *When introduced by Beck and Teboulle, FISTA is presented as an accelerated version of Iterative Shrinkage-Thresholding algorithms [17] (ISTA) which are methods solving problems of the form:*

$$\min_{x\in\mathcal{H}} F(x) := \|Ax - b\|^2 + \lambda\|x\|_1.$$

The appellation ISTA (and consequently FISTA) comes from the fact that the proximal operator of $\|\cdot\|_1$ is the soft-thresholding operator. However, the function F defined in that way only belongs to a subclass of composite convex functions that FISTA can actually minimize.

This confusion may explain the numerous names given to FISTA such as Nesterov's Accelerated Forward-Backward [5], Accelerated Proximal Gradient Descent [23] or Inertial Forward-Backward [5]. It also occurs that FISTA refers to (14) *where the sequence $(\alpha_n)_{n\in\mathbb{N}}$ is set constant in time equal to $\alpha \in (0,1)$ (see [22]), a method also called V-FISTA by Beck in [9].*

2.3 Convergence under additional geometry assumptions

In this section, we give an overview of the known convergence properties of the Proximal Gradient Methods and of the Chambolle-Dossal formulation of FISTA for convex composite functions satisfying an additional growth assumption.

Proximal Gradient Method under geometry assumptions The convergence of Proximal Gradient Method has been studied under several growth conditions in particular by Garrigos et al. in [18]. In this paper, the authors prove that the iterates of the Proximal Gradient Method converge strongly to a minimizer of F if the function is p-Łojasiewicz with $p \geqslant 1$ without any uniqueness assumption on the set of minimizers. The Łojasiewicz property can be linked to the growth assumptions stated in Section 2.1 and the strong convergence result holds if F satisfies $\mathcal{G}_\mu^2$ or $\mathcal{G}^\gamma$. Moreover, if F satisfies $\mathcal{G}_\mu^2$, then

$$F(x_n) - F^* = \mathcal{O}\left(e^{-\frac{\mu}{4L}n}\right)$$

and if F has an Hölderian growth i.e. $\mathcal{G}^\gamma$ with $\gamma > 2$ then

$$F(x_n) - F^* = \mathcal{O}\left(n^{-\frac{\gamma}{\gamma-2}}\right).$$

FISTA under geometry assumptions As stated previously, it is well known (see [15, 33]) that in a convex setting the iterates of the Chambolle-Dossal formulation of FISTA i.e. $\alpha_n = \frac{n}{n+\alpha}$, satisfy:

$$F(x_n) - F^* \leqslant \frac{(\alpha-1)^2\|x_0 - x^*\|^2}{2s(n+\alpha-2)^2}, \tag{16}$$

for any $x^* \in X^*$ as long as $s \leqslant \frac{1}{L}$ and $\alpha \geqslant 3$. The following works show that additional assumptions on F allow to guarantee better convergence properties. The paragraph is summarized in Table 1.

First, Su, Boyd and Candès show in [33] that this rate can be improved to $\mathcal{O}\left(n^{-3}\right)$ for strongly convex functions if $\alpha \geqslant \frac{9}{2}$. Attouch and Cabot strengthen this result in [3] by proving that $F(x_n) - F^* = \mathcal{O}\left(n^{-\frac{2\alpha}{3}}\right)$ for $\alpha > 0$ when F has a strong minimizer, i.e. F has a quadratic growth $\mathcal{G}_\mu^2$ and a unique minimizer. The understanding of FISTA in this setting is then enhanced by Aujol et al. in [8] as the authors provide non-asymptotical results enlightening the dependency in α. Apidopoulos et al. also give improved guarantees for functions having a Hölderian and quadratic growth in [1].

Reference	**Assumption on F**	**Parameter range**	**Convergence rate of $F(x_n) - F^*$**
Su et al. [33]	Strong convexity	$\alpha \geqslant \frac{9}{2}$	$\mathcal{O}\left(n^{-3}\right)$
Attouch, Cabot [3]	$\mathcal{G}_\mu^2$ and uniqueness of the minimizer	$\alpha > 0$	$\mathcal{O}\left(n^{-\frac{2\alpha}{3}}\right)$
Apidopoulos et al. [1] Aujol et al. [8]	$\mathcal{F}_\gamma$ and $\mathcal{G}_\mu^2$, $\gamma \geqslant 1$ Uniqueness of the minimizer	$\alpha \geqslant 1 + \frac{2}{\gamma}$	$\mathcal{O}\left(n^{-\frac{2\alpha\gamma}{\gamma+2}}\right)$
Apidopoulos et al. [1]	$\mathcal{F}_{\gamma_1}$ and $\mathcal{G}^{\gamma_2}$ where $\gamma_2 \geqslant \gamma_1 > 2$ Uniqueness of the minimizer	$\alpha \geqslant \frac{\gamma_1+2}{\gamma_1-2}$	$\mathcal{O}\left(n^{-\frac{2\gamma_2}{\gamma_2-2}}\right)$

Table 1: Convergence rate of $F(x_n) - F^*$ for FISTA under geometry assumptions on F.

The convergence results stated above give strong guarantees but they all rely on the hypothesis that F has a unique minimizer. Similarly, this assumption appears in [34] when proving the linear convergence of FISTA iterates for a LASSO problem. We can observe that in each aforementioned case, this condition allows to prove trivially the strong convergence of FISTA iterates towards the unique minimizer x^* of F: we know that $\lim_{n\to+\infty} F(x_n) - F^* = 0$ and $K\|x_n - x^*\|^\gamma \leqslant F(x_n) - F^*$ for some $\gamma \geqslant 2$ due to the considered growth assumption. Hence, $\|x_n - x^*\| \to 0$ when $n \to +\infty$.

2.4 The Asymptotic Vanishing Damping (AVD) system

In the seminal work by Su et al. [33], the authors demonstrate that the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA), within a differentiable framework, can be interpreted as the discretization of the following ordinary differential equation (ODE) called Asymptotic Vanishing Damping (AVD) system [29, 33]

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

where $\alpha = 3$.

The connection between inertial algorithms and ODEs dates back to the pioneering work of Polyak [30] on Heavy Ball schemes. In Polyak's observations, the following equation describes the evolution of a particle subject to a force field described by ∇F and a potentially time-dependent friction term $\alpha(t)$:

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) + \nabla F(x(t)) = 0. \tag{17}$$

If F is μ-strongly convex, Polyak demonstrates that the optimal friction is constant, depending on μ, ensuring an exponential decay of $F(x(t)) - F^*$.

Attouch et al. [2, 4] provide a comprehensive study of the solution to the ODE (17) based on the properties of F and the friction $\alpha(t)$, in particular analyzing the ODE (AVD). In both papers, the authors provide convergence rates for $F(x(t)) - F^*$ in the strongly convex case and for functions growing quadratically with a unique minimizer. Specifically, they show that:

$$F(x(t)) - F^* = O\left(t^{-\frac{2\alpha}{3}}\right). \tag{18}$$

Aujol et al. [7] demonstrate that these convergence rates can be improved by introducing an assumption of flatness, also known as quasar convexity. Under weaker growth conditions and quasar convexity, Aujol et al. [7] and later Luo et al. [26] provide new convergence rates for the solution of (AVD).

All these results assume that the function F to be minimized admits a unique minimizer. These findings are summarized in Table 2.

Weak convergence of $x(\cdot)$ towards a minimizer of F has been demonstrated by Attouch et al. [4] by adapting the convergence strategy proposed for the iterates of FISTA by Chambolle et al. [15]. Under the assumption of convexity of F, strong convergence is straightforward if F is strongly convex or if F grows quadratically with a unique minimizer, but less clear without these assumptions. In their work, the authors propose several sets of assumptions, such as the parity of F or the non-emptiness of the interior of the set of minimizers of F, to ensure strong convergence of $x(\cdot)$ towards a minimizer x^* of F.

In Section 4, we present new results on convergence rates under growth assumptions without assuming uniqueness of the minimizer. The strong convergence of the trajectory towards a minimizer of F is also proved by showing its finite length.

Reference	**Assumption on F**	**Parameter range**	**Convergence rate of $F(x(t)) - F^*$**
Su et al. [33]	$\mathcal{S}_\mu$	$\alpha \geqslant \frac{9}{2}$	$\mathcal{O}\left(t^{-3}\right)$
Attouch et al. [4]	$\mathcal{S}_\mu$	$\alpha > 3$	$\mathcal{O}\left(t^{-\frac{2\alpha}{3}}\right)$
Aujol et al. [7, 8]	$\mathcal{F}_\gamma$ and $\mathcal{G}_\mu^2$ Uniqueness of the minimizer	$\alpha > 1 + \frac{2}{\gamma}$	$\mathcal{O}\left(t^{-\frac{2\alpha\gamma}{\gamma+2}}\right)$
Aujol et al. [7]	$\mathcal{F}_{\gamma_1}$ and $\mathcal{G}^{\gamma_2}$ where $\gamma_2 \geqslant \gamma_1 > 2$ F coercive	$\alpha \geqslant \frac{\gamma_1+2}{\gamma_1-2}$	$\mathcal{O}\left(t^{-\frac{2\gamma_2}{\gamma_2-2}}\right)$
Luo, Xiao [26]	$\mathcal{F}_{\gamma_1}$ and $\mathcal{G}^{\gamma_2}$ where $\gamma_2 \geqslant \gamma_1 > 2$ Uniqueness of the minimizer	$\alpha \in \left(\frac{\gamma_1+2}{\gamma_1}, \frac{\gamma_1+2}{\gamma_1} \cdot \frac{\gamma_2}{\gamma_2-2}\right)$	$\mathcal{O}\left(t^{-\frac{2\alpha\gamma_1}{\gamma_1+2}}\right)$

Table 2: Convergence rate of $F(x(t)) - F^*$ where x is solution of (AVD) under geometry assumptions on F.

3 Strong convergence of FISTA iterates

In this section, we establish the **strong convergence of FISTA iterates to a minimizer** of a composite function $F \in \mathcal{C}$ (see Definition 1) if this function has a Hölderian or quadratic growth. Recall that iterates of FISTA are defined as:

$$x_0 \in \mathcal{H}, \quad \forall n \in \mathbb{N}, \begin{cases} y_n = x_n + \alpha_n (x_n - x_{n-1}) \\ x_{n+1} = \operatorname{prox}_{sh}(y_n - s\nabla f(y_n)), \end{cases} \tag{19}$$

where $x_{-1} = x_0$ and we choose the Chambolle-Dossal definition of $(\alpha_n)_{n\in\mathbb{N}}$ i.e. $\alpha_n = \frac{n}{n+\alpha}$ with $\alpha > 3$.

This property stated in Theorem 1, Corollary 1 and Theorem 3 relies on asymptotic controls of the sequence $(\|x_n - x_{n-1}\|)_{n\in\mathbb{N}}$ ensuring that the trajectory described by FISTA iterates has a finite length. Worst-case convergence rates for the error are given based on Lyapunov analyses and using the links between FISTA and (AVD). We also provide convergence guarantees in the continuous setting under similar assumptions in Section 4.

3.1 Hölderian growth condition

We first consider functions satisfying the local Hölderian growth condition $\mathcal{G}^{\gamma}_{loc}$ for $\gamma > 2$ and give convergence rates for FISTA iterates.

Theorem 1. *Let $F \in \mathcal{C}$ be a coercive composite function having a Hölderian growth i.e. satisfying $\mathcal{G}^{\gamma}_{loc}$ for some $\gamma > 2$. Then for $\alpha > 5 + \frac{8}{\gamma-2}$, the sequence $(x_n)_{n\in\mathbb{N}}$ provided by (19) with $s = \frac{1}{L}$ satisfies:*

$$F(x_n) - F^* = \mathcal{O}\left(n^{-\frac{2\gamma}{\gamma-2}}\right), \quad \|x_n - x_{n-1}\| = \mathcal{O}\left(n^{-\frac{\gamma}{\gamma-2}}\right). \tag{20}$$

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

The proof of Theorem 1 is detailed in Section 5.1. Note that this theorem can be seen as a discrete version of Theorem 4 giving properties of the solution of the ODE associated to Nesterov and presented in Section 4.

Several comments can be made about Theorem 1. First note that the strong convergence of the sequence $(x_n)_{n\in\mathbb{N}}$ is a consequence of the summability of $(\|x_n - x_{n-1}\|)_{n\in\mathbb{N}}$. Also observe that the convergence rate (20) is faster than the one achieved by the Proximal Gradient descend, see Section 2.3 for more details. Hence, FISTA provides an improvement for the class of convex functions satisfying a local Hölderian growth condition. Similar bounds have been established by Apidopoulos et al. [1] but the assumptions of Theorem 1 are weaker: no flatness hypothesis and no uniqueness of the minimizer are required.

Lastly, the conclusions of Theorem 1 hold if the composite function F satisfies $\mathcal{G}^{\gamma}_{\text{loc}}$ for $\gamma > 2$ and for $\alpha > 5 + \frac{8}{\gamma-2}$. Remarking that F satisfies $\mathcal{G}^{\gamma'}_{\text{loc}}$ for any $\gamma' \geqslant \gamma$ and that Theorem 1 thus holds for any $\gamma' > \max(\gamma, \frac{8}{\alpha-5} + 2)$, we deduce the following Corollary :

Corollary 1. *Let $F \in \mathcal{C}$ be a coercive composite function having a Hölderian growth i.e. satisfying $\mathcal{G}^{\gamma}_{loc}$ for $\gamma > 2$. Then, for any $\alpha > 5$, the sequence $(x_n)_{n\in\mathbb{N}}$ provided by (19) converges strongly to a minimizer of F.*

Finally, observe that the growth properties required in Theorem 1 are only local and thus, the decays are asymptotic. Even if the proof of Theorem 1 relies on a Lyapunov analysis, it seems technically difficult in this Hölderian setting to exhibit explicit bounds for a given number of iteration n.

3.2 Quadratic growth condition

In this section, we consider that F has a quadratic growth (denoted by $\mathcal{G}^2_{\mu}$ for the global growth condition and $\mathcal{G}^2_{\mu,loc}$ for the local one) with parameter $\mu > 0$. This assumption is more restrictive than the Hölderian growth condition considered in Section 3, and allows to derive stronger convergence results.

Theorem 2. *Let $F \in \mathcal{C}$ be a composite coercive function satisfying a quadratic growth condition $\mathcal{G}^2_{\mu}$ for some real parameter $\mu > 0$. Let $\alpha \geqslant 3 + \frac{3}{\sqrt{2}}$ and $\kappa = \frac{\mu}{L}$. Then there exist $\kappa_0 > 0$ such that for any $0 < \kappa \leqslant \kappa_0$, the sequence $(x_n)_{n\in\mathbb{N}}$ generated by FISTA with $s = \frac{1}{L}$ satisfies:*

$$\forall n \geqslant \frac{3\alpha}{\sqrt{\kappa}},\ F(x_n) - F^* \leqslant \frac{9}{4}e^{-2}M_0\left(\frac{8e}{3\sqrt{\kappa}}\alpha\right)^{\frac{2\alpha}{3}} n^{-\frac{2\alpha}{3}}, \tag{21}$$

where $M_0 = F(x_0) - F^*$ denotes the potential energy of the system at initial time.

Theorem 2, whose proof is detailed in Section 5.2, is an extension of [8, Theorem 6] to the class of composite functions with a set of minimizers not reduced to a single point. Similar results can be demonstrated by assuming that F is coercive and only satisfies some local quadratic growth condition. Indeed, the worst-case convergence rate of FISTA (16) is well known (see [33, 15]) and in particular, we know that the sequence $(F(x_n) - F^*)_{n\in\mathbb{N}}$ converges to 0. Then, according to

Lemma 1, so does the distance $(d(x_n, X^*))_{n\in\mathbb{N}}$ of the iterates to the set of minimizers. Thus, all the inequalities used and demonstrated in the proof of Theorem 2 remain valid for n large enough and the obtained convergence rates thus hold asymptotically. Our main contribution is to show that under local quadratic growth assumption and without minimizer uniqueness assumption, the trajectory of FISTA iterates is of finite length and strongly converges to a minimizer of F:

Theorem 3. *Let $F \in \mathcal{C}$ be a composite coercive function satisfying a local quadratic growth condition $\mathcal{G}^2_{\mu,loc}$ for some real parameter $\mu > 0$. Then for any $\alpha \geqslant 3 + \frac{3}{\sqrt{2}}$, the sequence $(x_n)_{n\in\mathbb{N}}$ of iterates provided by (19) with $s = \frac{1}{L}$, satisfies:*

$$F(x_n) - F^* = \mathcal{O}\left(n^{-\frac{2\alpha}{3}}\right), \quad \|x_n - x_{n-1}\| = \mathcal{O}\left(n^{-\frac{\alpha}{3}}\right). \tag{22}$$

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

Thus, under the quadratic growth property, we find the rate of convergence in $\mathcal{O}\left(n^{-\frac{2\alpha}{3}}\right)$ known until now only for FISTA under uniqueness of the minimizer. Moreover, observe that if the quadratic growth hypothesis is assumed to be global, Theorem 2 provides explicit non-asymptotic bounds that can be used to parameterize FISTA as it was done in [8].

More precisely, let $\varepsilon > 0$. The minimizers of the composite function F can be characterized by the optimality condition $0 \in \partial F(x)$, or equivalently $g(x) = 0$ where:

$$g(x) = L(x - x^+) := L\left(x - \operatorname{prox}_{\frac{1}{L}h}(x - \frac{1}{L}\nabla f(x))\right), \ x \in \mathcal{H}, \tag{23}$$

denotes the composite gradient mapping and $x^+ := \operatorname{prox}_{\frac{1}{L}h}(x - \frac{1}{L}\nabla f(x))$. This last formulation is convenient for defining an approximate solution to the composite problem, and thus to deduce a tractable stopping criterion:

Definition 5 (ε-solution)**.** *Let ε be the expected accuracy. The iterate x_n is said to be an ε-solution of the problem $\min_{x\in\mathcal{H}} F(x)$ if:*

$$\|g(x_n)\| \leqslant \varepsilon. \tag{24}$$

Observe that in the differentiable case (i.e. when $h = 0$), we have: $g(x) = \nabla f(x)$ so that an ε-solution is nothing more than an iterate x_n satisfying:

$$\|g(x_n)\| = \|\nabla F(x_n)\| \leqslant \varepsilon. \tag{25}$$

The notion of ε-solution can be seen as a good stopping criterion for an algorithm solving the composite optimization problem for the following reasons. It is numerically quantifiable and in addition, controlling the norm of the composite gradient mapping is roughly equivalent to having a control on the values of the objective function. Indeed using [28, Theorem 1] and [8, Lemma 3.1], we can prove that the composite gradient mapping is controlled by the values of the objective function:

$$\forall x \in \mathbb{R}^N, \ \frac{1}{2L}\|g(x)\|^2 \leqslant F(x) - F^*. \tag{26}$$

Hence, from Theorem 2, a sufficient condition to reach an ε-solution is:

$$\frac{9L}{2}e^{-2}M_0\left(\frac{8e}{3\sqrt{\kappa}}\alpha\right)^{\frac{2\alpha}{3}} n^{-\frac{2\alpha}{3}} \leqslant \varepsilon^2, \tag{27}$$

which amounts to

$$n \geqslant \left(\sqrt{\frac{LM_0}{2}}\frac{3}{e\,\varepsilon}\right)^{\frac{3}{\alpha}} \frac{8e}{3\sqrt{\kappa}}\alpha. \tag{28}$$

Minimizing the number of iterations to reach an ε-solution with respect to the friction parameter α, we thus deduce that choosing

$$\alpha = \alpha_\varepsilon := 3 \log \left(\frac{3}{e\,\varepsilon} \sqrt{\frac{LM_0}{2}} \right), \tag{29}$$

will ensure to reach an ε-solution in at most:

$$n_\varepsilon := \frac{8e^2}{\sqrt{\kappa}} \log \left(\frac{3}{e\,\varepsilon} \sqrt{\frac{LM_0}{2}} \right) \tag{30}$$

iterations. In other words, for a fixed precision $\varepsilon > 0$, it is possible to parameterize FISTA such that the number of iterations to reach an ε-solution is comparable to the number of iterations required by an algorithm with an exponential decay.

Notice that in the case of FISTA with the assumption of a unique minimizer [8], for the exact same choice of $\alpha = \alpha_\varepsilon$ (which is not the optimized choice stated in [8, Theorem 3]), the number of iterations (denoted by $n_\varepsilon^{FISTA,uniq}$) to reach an ε-solution is then:

$$n_\varepsilon^{FISTA,uniq} \quad = \quad \frac{8e^2}{3\sqrt{\kappa}} \alpha_\varepsilon = \frac{8e^2}{\sqrt{\kappa}} \log \left(\frac{5\sqrt{LM_0}}{e\sqrt{2}\varepsilon} \right), \tag{31}$$

which is better than that given by (30) for FISTA without the minimizer uniqueness assumption:

$$n_\varepsilon = n_\varepsilon^{FISTA,uniq} + \frac{8e^2}{\sqrt{\kappa}} \log \left(\frac{3\sqrt{2}}{5} \right) > n_\varepsilon^{FISTA,uniq}. \tag{32}$$

Remark 3. *The convergence rate stated in Theorem 2 can be strengthened if there exists $\gamma > 1$ such that some flatness condition is satisfied:*

$$\forall x \in \mathcal{H},\ F(x) - F^* \leqslant \frac{1}{\gamma} \langle \nabla F(x), x - x^* \rangle, \tag{33}$$

for any minimizer $x^* \in X^*$, as it was done in [8, Theorem 4].

4 Asymptotic Vanishing Damping system under geometry conditions

Let us now consider the AVD system

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

which has been widely studied in the literature, in particular using Lyapunov-type approaches (see e.g. Table 2 for a short overview). Let us mention the references [33, 4, 7, 8] that introduce the following energy:

$$\mathcal{E}(t) = t^2 \left(F(x(t)) - F^* \right) + \frac{1}{2} \| \lambda (x(t) - x^*) + t\dot{x}(t) \|^2 \tag{34}$$

with different values of $\lambda > 0$, depending on a given minimizer x^* which is supposed to be constant in time. The uniqueness assumption of the minimizer is not necessary to obtain the results proved by Su, Boyd and Candès [33] and Attouch, Chbani, Peypouquet and Redont [4] in the convex case. On the other hand, when assuming an additional growth property, the fact that these energies depend on a fixed $x^* \in X^*$ is limiting for determining improved convergence rates. Our approach

to extend classical analysis without the uniqueness assumption (similar to that in [6]) consists in slightly modifying the Lyapunov energy (34) as follows:

$$\mathcal{E}(t) = t^2\left(F(x(t)) - F^*\right) + \frac{1}{2}\|\lambda(x(t) - x^*(t)) + t\dot{x}(t)\|^2 + \frac{\xi}{2}\|x(t) - x^*(t)\|^2 \tag{35}$$

where $x^*(t)$ denotes the projection of the trajectory $x(t)$ onto the set of minimizers X^*:

$$x^*(t) = P_{X^*}(x(t)) := \arg\min_{x^* \in X^*} \|x(t) - x^*\|^2.$$

Note that since F is assumed to be continuous and convex, the set X^* is actually a closed convex set and the projection onto X^* is thus well defined. This modification of the energy $\mathcal{E}$ leads to a question when attempting to conduct the Lyapunov analysis: is $t \mapsto x^*(t)$ differentiable?

The smoothness of $t \mapsto x^*(t)$ is related to the smoothness of P_{X^*}. In fact, if P_{X^*} is directionally differentiable then $t \mapsto x^*(t)$ is right-differentiable (and left-differentiable) and its right-hand derivative is equal to $P'_{X^*}(x(t), \dot{x}(t))$. We refer the reader to Appendix A.1 for more insightful explanations.

In [13, Theorem 7.2], Bonnans et al. prove that if a closed convex set $\mathcal{S} \subset \mathcal{X}$ is second order regular at $P_{\mathcal{S}}(x)$ for some $x \in \mathcal{X}$, then $P_{\mathcal{S}}$ is directionally differentiable at x.

Definition 6. *[32, Definition 2.1] A set S is said second order regular at a point $\bar{x} \in S$ if for any sequence $(x_n)_{n\in\mathbb{N}}$ in S of the form: $x_n = \bar{x} + t_n h + \frac{1}{2}t_n^2 r_n$, where $(t_n)_{n\in\mathbb{N}}$ is monotonically non-increasing to 0, $t_n r_n \to 0$ and $h \in \mathcal{H}$, it follows that:*

$$\lim_{n\to+\infty} d(r_n, T_S^2(\bar{x}, h)) = 0,$$

where $T_S^2(\bar{x}, h)$ denotes the inner second order tangent set to S in the direction h:

$$T_S^2(\bar{x}, h) := \left\{ w \in \mathcal{H} \ : \ d(\bar{x} + th + \frac{1}{2}t^2 w, S) = o\left(t^2\right) \right\}.$$

The set S is said second order regular if it is second order regular at every point.

We refer the reader to [13, 32] to have a complete understanding of the complex notion of second order regularity. Keep in mind that sets having a C^2 boundary [19] (in the sense that their boundary is locally a C^2 sub-manifold of $\mathcal{H}$) and polyhedral sets [32] are second-order regular, so that the projection onto these sets is actually directionally differentiable.

Assuming that the set of minimizers X^* is second order regular instead of the classical uniqueness assumption, Theorem 4 provides new bounds on $F(x(t)) - F(x^*)$ and on $\|\dot{x}(t)\|$ under Hölderian growth conditions. The proof is detailed in Appendix A.2.

Theorem 4. *Let F be a convex differentiable function with a non-empty second order regular set of minimizers X^*. If F is coercive and satisfies a Hölderian growth condition $\mathcal{G}^\gamma_{loc}$ for some $\gamma > 2$. Then, for any $\alpha > \frac{9}{2} + \frac{6}{\gamma-2}$, the trajectories provided by* (AVD) *satisfy*

$$F(x(t)) - F^* = \mathcal{O}\left(t^{-\frac{2\gamma}{\gamma-2}}\right), \ \|\dot{x}(t)\| = \mathcal{O}\left(t^{-\frac{\gamma}{\gamma-2}}\right), \tag{36}$$

and strongly converge to a minimizer of F.

Unlike Aujol et al. [7] and Luo et al. [26], no flatness condition on F or uniqueness of the minimizer is needed here. The only added hypothesis is the regularity of the set of minimizers. This hypothesis may be technical, but seems difficult to remove. Note that the bound on $\|\dot{x}(\cdot)\|$ implies that the trajectory $x(\cdot)$ has a finite length and strongly converges to a minimizer of F.

Finally, we consider the class of convex differentiable functions having a quadratic growth. Applying the strategy described at the beginning of this section and in Appendix A.1, we propose an extension of [8, Theorem 5] to functions having a set of minimizers not reduced to a single point, and complement this theorem with a result on $\|\dot{x}(\cdot)\|$ ensuring that the trajectory $x(\cdot)$ has finite length and thus strongly converges to a minimizer x^* of F.

Theorem 5. *Let F be a convex differentiable function with a non-empty second order regular set of minimizers X^*. Assume that F is coercive and satisfies a local quadratic growth condition $\mathcal{G}^2_{\mu,loc}$ for some $\mu > 0$. Let x be a solution of* (AVD) *for some $t_0 \geqslant 0$ and $\alpha > 0$. If $\alpha > 3$ and μ is small enough then we have:*

$$F(x(t)) - F^* = \mathcal{O}\left(t^{-\frac{2\alpha}{3}}\right),\ \|\dot{x}(t)\| = \mathcal{O}\left(t^{-\frac{\alpha}{3}}\right). \tag{37}$$

and the trajectory $x(\cdot)$ strongly converges to a minimizer of F.

Note that this rate in $\mathcal{O}\left(t^{-\frac{2\alpha}{3}}\right)$ was already known but for classes of functions satisfying stronger geometric assumptions, in particular for strongly convex functions in [33, Theorem 8] and for convex functions having a strong minimizer in [2, Theorem 3.12].

Assuming now that F satisfies a global quadratic growth hypothesis, explicit bounds on the decay of the functional can be calculated. This will subsequently allow for an optimized choice of friction parameter values α:

Proposition 1. *Let F be a convex differentiable function with a non-empty second order regular set of minimizers X^*. Assume that F satisfies a global quadratic growth condition $\mathcal{G}^2_\mu$ for some $\mu > 0$. Let x be a solution of* (AVD) *for some $t_0 \geqslant 0$ and $\alpha > 0$. If $\alpha > 3$ and μ is small enough then we have:*

$$\forall t \geqslant \frac{\alpha r^*}{3\sqrt{\mu}} \geqslant t_0,\ F(x(t)) - F^* \leqslant C_1 e^{\frac{2}{3}C_2(\alpha-3)} M_0 \left(\frac{\alpha r^*}{3t\sqrt{\mu}}\right)^{\frac{2\alpha}{3}}, \tag{38}$$

where $M_0 = F(x(t_0)) - F^* + \frac{1}{2}\|\dot{x}(t_0)\|^2$, $r^* \simeq 3$ is the unique positive real root of the polynomial: $r \mapsto r^3 - r^2 - 2(1+\sqrt{2})r - 4$ and

$$C_1 = 1 + \frac{2}{r^*} + \frac{4}{r^{*2}},\ C_2 = \frac{1}{r^*} + \frac{1+\sqrt{2}}{r^{*2}} + \frac{4}{3r^{*3}}.$$

We give a simplified analysis of this bound by removing some of the constants in the bound (38) for more readability. Let $\varepsilon > 0$ be the desired precision on the functional decay $F(x(t)) - F^*$. For any $\alpha > 3$, the minimum time t to reach the precision ε is at least in:

$$\left(\frac{\alpha}{t\sqrt{\mu}}\right)^{\frac{2\alpha}{3}} \leqslant \varepsilon \iff t \geqslant \frac{\alpha}{\sqrt{\mu}}\left(\frac{1}{\varepsilon}\right)^{\frac{3}{2\alpha}}$$

which corresponds to the polynomial rate stated in Theorem 5. Choosing now $\alpha = C\log\left(\frac{1}{\varepsilon}\right)$ for a well-chosen real constant $C > 0$, the minimum time t to reach an ε-solution is at least in:

$$\left(\frac{\alpha}{t\sqrt{\mu}}\right)^{\frac{2\alpha}{3}} \leqslant \varepsilon \iff t \geqslant \frac{Ce^{\frac{3}{2C}}}{\sqrt{\mu}}\log\left(\frac{1}{\varepsilon}\right)$$

which is comparable to a fast exponential decay of the trajectory.

5 Proofs of Theorem 1 and Theorem 2

The proofs of Theorems 1, 2 and 3 are based on a Lyapunov analysis involving similar terms. In particular, the convergence proofs of Theorems 2 and 3 are built around

$$E_n = \frac{2n^2}{L}(F(x_n) - F^*) + \left\|\lambda(x_{n-1} - x^*_{n-1}) + n(x_n - x_{n-1})\right\|^2, \tag{39}$$

where $\lambda > 0$, while we consider the following discrete Lyapunov energy for Theorem 1:

$$\mathcal{E}_n = \frac{2n^2}{L}(F(x_n) - F^*) + \|\lambda(x_n - x^*_n) + n\alpha_n(x_n - x_{n-1})\|^2 + \xi\|x_n - x^*_n\|^2 + \lambda n\alpha_n^2\|x_n - x_{n-1}\|^2, \tag{40}$$

where $\lambda > 0$, $\xi < 0$ and x_n^*, $n \in \mathbb{N}$, denotes the projection of x_n onto the set of minimizers X^*. For the sake of clarity, we introduce the following notations:

$$\begin{aligned} w_n = \frac{2}{L}(F(x_n) - F^*),\ h_n = \|x_n - x_n^*\|^2,\ \delta_n = \|x_n - x_{n-1}\|^2, \\ \gamma_n^* = \|x_n^* - x_{n-1}^*\|^2,\ \alpha_n = \frac{n}{n+\alpha}. \end{aligned} \tag{41}$$

Both convergence proofs rely on two technical lemma. The first one, whose proof is given in Section B.1, is crucial for handling the non-uniqueness of the minimizer:

Lemma 2. *For all $n \in \mathbb{N}^*$, the following equalities hold:*

1. $\langle x_n - x_n^*, x_n - x_{n-1}\rangle = \frac{1}{2}(h_n - h_{n-1} + \delta_n - \gamma_n^*) + \langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle$.
2. $\langle x_{n-1} - x_{n-1}^*, x_n - x_{n-1}\rangle = \frac{1}{2}(h_n - h_{n-1} - \delta_n + \gamma_n^*) + \langle x_n - x_n^*, x_n^* - x_{n-1}^*\rangle$,

The second one encodes the fact that the sequence $(x_n)_{n\in\mathbb{N}}$ is provided by (19). Its proof is based on a descent lemma proved in [15] and is detailed in Section B.2.

Lemma 3. *Let $(x_n)_{n\in\mathbb{N}}$ be the sequence provided by (19) with $s = \frac{1}{L}$. Then, for any $n \in \mathbb{N}^*$,*

$$w_{n+1} - w_n \leqslant \alpha_n^2 \delta_n - \delta_{n+1}, \tag{42}$$

and

$$\begin{aligned} w_{n+1} \leqslant (1+\alpha_n)h_n + (\alpha_n^2 + \alpha_n)\delta_n - \alpha_n h_{n-1} - h_{n+1} - \gamma_{n+1}^* - \alpha_n\gamma_n^* \\ + 2\alpha_n\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle - 2\langle x_{n+1} - x_{n+1}^*, x_{n+1}^* - x_n^*\rangle, \end{aligned} \tag{43}$$

where $\alpha_n = \frac{n}{n+\alpha}$.

We would like to point out that several controls can be deduced from the properties of the projection onto a convex. Indeed, if C is a closed convex set such that $C \subset E$, then for any $x \in E$ and $y \in C$,

$$\langle x - p, y - p\rangle \leqslant 0,$$

where p denotes the projection of x onto C. This property directly guarantees inequalities such as

$$\langle x_n - x_n^*, x_n^* - x_{n-1}^*\rangle \geqslant 0 \quad \text{and} \quad \langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle \leqslant 0.$$

5.1 Proof of Theorem 1

5.1.1 Sketch of the proof

Recall that our analysis relies on the following discrete Lyapunov energy:

$$\mathcal{E}_n = \frac{2n^2}{L}(F(x_n) - F^*) + \|\lambda(x_n - x_n^*) + n\alpha_n(x_n - x_{n-1})\|^2 + \xi\|x_n - x_n^*\|^2 + \lambda n\alpha_n^2\|x_n - x_{n-1}\|^2, \tag{44}$$

where $\lambda > 0$ and $\xi < 0$. Given the notations introduced in (41), it can be rewritten:

$$\mathcal{E}_n = n^2 w_n + b_n + \xi h_n + \lambda n\alpha_n^2\delta_n, \tag{45}$$

where:

$$\begin{aligned} w_n = \frac{2}{L}(F(x_n) - F^*),\ h_n = \|x_n - x_n^*\|^2,\ \delta_n = \|x_n - x_{n-1}\|^2, \\ \gamma_n^* = \|x_n^* - x_{n-1}^*\|^2,\ \alpha_n = \frac{n}{n+\alpha},\ b_n = \|\lambda(x_n - x_n^*) + n\alpha_n(x_n - x_{n-1})\|^2. \end{aligned} \tag{46}$$

The strategy underlying this proof is to show that this Lyapunov energy behaves asymptotically as $n^{-\frac{4}{\gamma-2}}$. Note that this does not directly guarantee the desired convergence results since $\xi < 0$. The local growth condition $\mathcal{G}^\gamma$ satisfied by F is necessary to reach the conclusion.

In order to study the asymptotic behavior of $\mathcal{E}_n$, we define $\mathcal{J}_n = n^p\mathcal{E}_n$ where $p = 1 + \frac{4}{\gamma-2}$. The proof then follows several steps:

- Using the properties of FISTA and the convexity of F, we show that for a well-chosen set of parameters (α, λ, ξ) and n sufficiently large:

$$\mathcal{J}_{n+1} - \mathcal{J}_n \leqslant A(n+1)^{p+1} w_{n+1} + B(n+1)^{p-1} h_{n+1}, \tag{47}$$

  for some constants $A < 0$ and $B > 0$.

- Given the previous inequality and the growth condition satisfied by F, we prove that for n sufficiently large:

$$\mathcal{J}_n \leqslant Cn, \tag{48}$$

  for some constant $C > 0$. This inequality ensures that $\mathcal{E}_n$ decreases asymptotically as $n^{-\frac{4}{\gamma-2}}$.

- By coming back to the definition of $\mathcal{J}_n$ and $\mathcal{E}_n$ and using the assumption $\mathcal{G}^\gamma$ satisfied by F, we show that $n^{p+1} w_n$ and $\alpha_n^2 n^{p+1} \delta_n$ are bounded which leads to the desired results.

5.1.2 A technical Lemma before the proof of Theorem 1

In the proof of Theorem 1, the geometry of the function F will be useful to control the distance of the FISTA iterates to the set of minimizers by the decay of F along the trajectory of iterates.

Lemma 4. *Let F satisfy $\mathcal{G}^\gamma_{loc}$ for some $\gamma > 2$ and real constant $K > 0$. If $p = 1 + \frac{4}{\gamma-2}$, then for n sufficiently large,*

$$n^{p-1} h_n \leqslant \left(\frac{L}{2K}\right)^{\frac{2}{\gamma}} \left(n^{p+1} w_n\right)^{\frac{2}{\gamma}}, \tag{49}$$

where: $w_n = \dfrac{2}{L}(F(x_n) - F^*)$ and $h_n = d(x_n, X^*)^2$.

Proof. Assume that F satisfies some local Hölderian growth condition $\mathcal{G}^\gamma_{\text{loc}}$ for some $\gamma > 0$. It is well known (see [33, 15]) that the iterates of FISTA with $s = \frac{1}{L}$ and $\alpha \geqslant 3$ satisfy the following inequality

$$\forall n \in \mathbb{N},\ F(x_n) - F^* \leqslant \frac{(\alpha-1)^2 L}{2(n+\alpha-2)^2} \|x_0 - x^*\|^2,$$

which implies that the sequence $(F(x_n) - F^*)_{n\in\mathbb{N}}$ converges to 0. Applying Lemma 1, we thus deduce that the sequence $(d(x_n, X^*))_{n\in\mathbb{N}}$ converges to 0 as $n \to +\infty$ and that there exist $K > 0$ and $N \in \mathbb{N}$ such that:

$$\forall n \geqslant N,\ K d(x_n, X^*)^\gamma \leqslant F(x_n) - F^*. \tag{50}$$

or, equivalently:

$$\forall n \geqslant N,\ h_n \leqslant \left(\frac{L}{2K}\right)^{\frac{2}{\gamma}} w_n^{\frac{2}{\gamma}}.$$

Choosing $p = 1 + \frac{4}{\gamma-2}$, the expected inequality (49) holds for any $n \geqslant N$. □

5.1.3 Proof of Theorem 1

Let $(x_n)_{n\in\mathbb{N}}$ be the sequence provided by (19) and $(\mathcal{E}_n)_{n\in\mathbb{N}}$ be the Lyapunov energy defined in (44). The first step of the proof is to get an upper bound on $\mathcal{E}_{n+1} - \mathcal{E}_n$. We provide such an inequality in the following lemma which is proved in Section B.3.

Lemma 5. *Let $\xi = \lambda(\lambda + 1 - \alpha)$. For any $n \in \mathbb{N}^*$,*

$$\begin{aligned}\mathcal{E}_{n+1} - \mathcal{E}_n \leqslant{}& ((2-\lambda)n + 1)\, w_{n+1} + B_1(n) b_{n+1} + B_2(n) h_{n+1} + B_3(n)\delta_{n+1} \\ &- B_4(n)\left(\gamma^*_{n+1} - 2\langle x_n - x^*_n, x^*_{n+1} - x^*_n\rangle\right),\end{aligned} \tag{51}$$

where :

- $B_1(n) = \frac{2(\lambda+1-\alpha)}{n+1} + \frac{\alpha(2\lambda+2-\alpha)}{(n+1)^2}$,
- $B_2(n) = -\frac{2\lambda^2(\lambda+1-\alpha)}{n+1} + \frac{\alpha\lambda^2(\alpha-2\lambda-2)}{(n+1)^2}$,
- $B_3(n) = \alpha(\lambda+2) - (2\lambda^2+2\lambda+1) + \alpha^2\frac{n(\lambda-2)+\lambda-2-2\alpha}{(n+1-\alpha)^2}$,
- $B_4(n) = -2\lambda(\lambda+1-\alpha) - \frac{\alpha^2\lambda}{n+1+\alpha}$.

We introduce $\mathcal{J}_n = n^p\mathcal{E}_n$ with $p = 1 + \frac{4}{\gamma-2}$. The next step is to show the following inequality.

Lemma 6. *Let $\xi = \lambda(\lambda+1-\alpha)$. If $\lambda \leqslant \alpha - 1$, then for any $n \in \mathbb{N}^*$,*

$$\begin{aligned}\mathcal{J}_{n+1} - \mathcal{J}_n \leqslant & \left((2-\lambda+p)(n+1)^{p+1} + R_1(n)\right) w_{n+1} \\ & + \left((2(\lambda+1-\alpha)+p)(n+1)^{p-1} + R_2(n)\right) b_{n+1} \\ & + \left(\lambda(\lambda+1-\alpha)(p-2\lambda)(n+1)^{p-1} + R_3(n)\right) h_{n+1} \\ & - n^p B_4(n)\left(\gamma^*_{n+1} - 2\langle x_n - x^*_n, x^*_{n+1} - x^*_n\rangle\right),\end{aligned} \tag{52}$$

where:

$$\begin{aligned} R_1(n) &= (\lambda - 1 + p(\lambda-2))\, n^p + p(\lambda-2)n^{p-1} \\ R_2(n) &= ((4\lambda + 6 + 2p - \alpha)\alpha + 2\lambda p)\,(n+1)^{p-2} + \alpha^2(6\lambda+8+p)(n+1)^{p-3} \\ &\quad + 2\alpha^3(\lambda+2)(n+1)^{p-4} \\ R_3(n) &= \left(\lambda^2(\alpha^2 + 2\alpha + 4\lambda p + p + 1) + \lambda(\alpha-1)(p-1)\right)(n+1)^{p-2} \\ &\quad + \lambda^2\alpha\,(2p\lambda + 2p + 6\lambda\alpha + 2\alpha)\,(n+1)^{p-3} + 2\alpha^3\lambda^2(\lambda+2)(n+1)^{p-4}. \end{aligned}$$

The proof is detailed in Section B.4. By setting $\lambda = \alpha - 1 - p$, we get that $\lambda \leqslant \alpha - 1$ and:

$$\begin{aligned} C_1 &:= 2 - \lambda + p = 3 + 2p - \alpha, \\ C_2 &:= 2(\lambda + 1 - \alpha) + p = -p, \\ C_3 &:= \lambda(\lambda+1-\alpha)(p-2\lambda) = p(\alpha-1-p)(2\alpha-2-3p), \end{aligned} \tag{53}$$

which implies that for $\alpha > 3 + 2p = 5 + \frac{8}{\gamma-2}$,

$$C_1 < 0, \quad C_2 < 0, \quad C_3 > 0. \tag{54}$$

Considering the order of $R_1(n)$, $R_2(n)$ and $R_3(n)$, this guarantees that for n sufficiently large:

$$\begin{cases} C_1(n+1)^{p+1} + R_1(n) < \dfrac{C_1}{2}(n+1)^{p+1}, \\ C_2(n+1)^{p-1} + R_2(n) < 0, \\ C_3(n+1)^{p-1} + R_3(n) < 2C_3(n+1)^{p-1}. \end{cases} \tag{55}$$

In addition, for the choice $\lambda = \alpha - 1 - p$ we have that $B_4(n) = 2p(\alpha-1-p) - \alpha^2\frac{\alpha-1-p}{n+1+\alpha}$ which is positive for $\alpha > 5 + \frac{8}{\gamma-2}$ and n sufficiently large. As $\gamma^*_{n+1} \geqslant 0$ and $\langle x_n - x^*_n, x^*_{n+1} - x^*_n\rangle \leqslant 0$, this ensures that

$$n^p B_4(n)\left(\gamma^*_{n+1} - 2\langle x_n - x^*_n, x^*_{n+1} - x^*_n\rangle\right) \geqslant 0. \tag{56}$$

Hence, if $\alpha > 5 + \frac{8}{\gamma-2}$, then for n sufficiently large:

$$\mathcal{J}_{n+1} - \mathcal{J}_n \leqslant \frac{C_1}{2}(n+1)^{p+1}w_{n+1} + 2C_3(n+1)^{p-1}h_{n+1}. \tag{57}$$

1^{st} *step: Proving that* $F(x_n) - F^* = \mathcal{O}\left(n^{-\frac{2\gamma}{\gamma-2}}\right)$. To obtain bounds on the decay of F along the FISTA iterates, we take advantage of the geometry of the function F to minimize. Assuming that F satisfies a local Hölderian growth condition, Lemma 4 combined with (57) ensure that for n sufficiently large:

$$\mathcal{J}_{n+1} - \mathcal{J}_n \leqslant \frac{C_1}{2}(n+1)^{p+1}w_{n+1} + 2C_3\left(\frac{L}{2K}\right)^{\frac{2}{\gamma}}\left((n+1)^{p+1}w_{n+1}\right)^{\frac{2}{\gamma}}, \tag{58}$$

which ensures that there exists $M_0 \in \mathbb{R}$ such that $\mathcal{J}_{n+1} - \mathcal{J}_n \leqslant M_0$.

Thus, there exists $n_0 \in \mathbb{N}$ such that for all $n \geqslant n_0$, $\mathcal{J}_n \leqslant nM_0 + M_1$ where $M_1 = \mathcal{J}_{n_0} - n_0M_0$. Consequently, we get that for n sufficiently large, $\mathcal{J}_n \leqslant 2nM_0$. Coming back to the definition of $\mathcal{J}$, this implies that:

$$n^{p-1}\left(n^2w_n + b_n + \xi h_n + \lambda n\alpha_n^2\delta_n\right) \leqslant 2M_0. \tag{59}$$

Noticing that $\xi = \lambda(\lambda + 1 - \alpha) < 0$, this ensures that for n sufficiently large:

$$n^{p+1}w_n - |\xi|n^{p-1}h_n \leqslant 2M_0, \tag{60}$$

and according to Lemma 4:

$$n^{p+1}w_n - |\xi|\left(\frac{L}{2K}\right)^{\frac{2}{\gamma}}\left(n^{p+1}w_n\right)^{\frac{2}{\gamma}} \leqslant 2M_0. \tag{61}$$

The following lemma guarantees that for n sufficiently large, $n^{p+1}w_n$ is bounded.

Lemma 7. *Let* $x \in \mathbb{R}^+$, $\delta \in (0,1)$, $K_1 > 0$ *and* $K_2 > 0$. *Then,*

$$x^\delta(x^{1-\delta} - K_1) \leqslant K_2 \quad \Longrightarrow \quad x \leqslant \left(K_2^{1-\delta} + K_1\right)^{\frac{1}{1-\delta}}.$$

As a consequence, there exists $M_2 > 0$ such that for n sufficiently large, $n^{p+1}w_n \leqslant M_2$ and considering the value of p we have that:

$$F(x_n) - F^* \leqslant \frac{LM_2}{2n^{\frac{2\gamma}{\gamma-2}}}. \tag{62}$$

This proves our first claim: $F(x_n) - F^* = \mathcal{O}\left(n^{-\frac{2\gamma}{\gamma-2}}\right)$.

2^{nd} *step: Proving that the trajectory of FISTA iterates has a finite length* Let us come back to the inequality (59) which implies that for n sufficiently large:

$$n^{p-1}\left(n^2w_n + b_n - |\xi|h_n\right) \leqslant 2M_0. \tag{63}$$

By applying the inequality $\|u\|^2 \leqslant 2\|u+v\|^2 + 2\|v\|^2$ to $u = \alpha_n(x_n - x_{n-1})$ and $v = \lambda(x_n - x_n^*)$, we get:

$$b_n \geqslant \frac{n^2\alpha_n^2}{2}\delta_n - \lambda^2 h_n. \tag{64}$$

Combining this inequality with (63) leads to:

$$\left(n^{p+1}w_n - (\lambda^2 + |\xi|)n^{p-1}h_n\right) + \frac{\alpha_n^2}{2}n^{p+1}\delta_n \leqslant 2M_0. \tag{65}$$

Then, Lemma 4 gives us that

$$n^{p+1}w_n - (\lambda^2 + |\xi|)n^{p-1}h_n \geqslant n^{p+1}w_n - \left(\frac{L}{2K}\right)^{\frac{2}{\gamma}}\left(n^{p+1}w_n\right)^{\frac{2}{\gamma}}. \tag{66}$$

The study of the variations of $\varphi : x \mapsto x - \left(\frac{L}{2K}\right)^{\frac{2}{\gamma}} x^{\frac{2}{\gamma}}$ shows that there exists a real constant $M_3 \in \mathbb{R}$ such that φ is bounded from below by M_3. Hence for n large enough: $n^{p+1}w_n - (\lambda^2 + |\xi|)n^{p-1}h_n \geqslant M_3$, and:

$$\delta_n \leqslant \frac{4M_0 - 2M_3}{\alpha_n^2 n^{p+1}}, \tag{67}$$

and therefore: $\|x_n - x_{n-1}\| = \mathcal{O}\left(n^{-\frac{\gamma}{\gamma-2}}\right)$.

3^{rd} *step: Proving that the FISTA iterates strongly converge to a minimizer of* F. The strong convergence of FISTA iterates can be deduced from the summability of $\|x_n - x_{n-1}\|$ since $\frac{\gamma}{\gamma-2} > 1$ for any $\gamma > 2$.

5.2 Proof of Theorem 2

The proof of Theorem 2 is an adaptation of the proof of [8, Theorem 6] without the assumption that F has a unique minimizer. Its structure is similar despite the involvement of additional terms linked to the relaxed setting. The tricky technical aspect is to control these additional terms in order to recover inequalities obtained in the case of uniqueness of the minimizer.

Recall that we consider the discrete Lyapunov energy defined in (39) with the notations (41):

$$E_n = n^2 w_n + \lambda^2 h_{n-1} + n^2 \delta_n + 2\lambda n \langle x_{n-1} - x_{n-1}^*, x_n - x_{n-1} \rangle, \tag{68}$$

where $\alpha > 3$, $\lambda = \frac{2\alpha}{3}$ and:

$$w_n = \frac{2}{L}(F(x_n) - F^*),\ h_n = \|x_n - x_n^*\|^2,\ \delta_n = \|x_n - x_{n-1}\|^2,\ \alpha_n = \frac{n}{n+\alpha}. \tag{69}$$

Applying the second claim of Lemma 2, the Lyapunov energy (68) can be rewritten as:

$$E_n = n^2 w_n + \lambda n h_n + (\lambda^2 - \lambda n) h_{n-1} + (n^2 - \lambda n)\delta_n + \lambda n \gamma_n^* + 2\lambda n \langle x_n - x_n^*, x_n^* - x_{n-1}^* \rangle.$$

For any $n \in \mathbb{N}^*$, we have:

$$\begin{aligned}
E_{n+1} - \left(1 - \frac{\lambda - 2}{n}\right) E_n = {} & (n+1)^2 w_{n+1} - \left(1 - \frac{\lambda - 2}{n}\right) n^2 w_n \\
& + \left((n+1)^2 - \lambda(n+1)\right) \delta_{n+1} - \left(1 - \frac{\lambda - 2}{n}\right)(n^2 - \lambda n)\delta_n \\
& + \left(\lambda^2 - \lambda(n+1) - \lambda n \left(1 - \frac{\lambda - 2}{n}\right)\right) h_n + \lambda(n+1) h_{n+1} \\
& - (\lambda^2 - \lambda n)\left(1 - \frac{\lambda - 2}{n}\right) h_{n-1} \\
& + \lambda(n+1)\gamma_{n+1}^* + 2\lambda(n+1)\langle x_{n+1} - x_{n+1}^*, x_{n+1}^* - x_n^* \rangle \\
& - \lambda(n - \lambda + 2)\gamma_n^* - 2\lambda(n - \lambda + 2)\langle x_n - x_n^*, x_n^* - x_{n-1}^* \rangle.
\end{aligned}$$

Elementary computations give that:

$$(n+1)^2 w_{n+1} - \left(1 - \frac{\lambda - 2}{n}\right) n^2 w_n = n\,(n - \lambda + 2)\,(w_{n+1} - w_n) + (\lambda n + 1) w_{n+1}.$$

Consequently, Lemma 3 ensures that for all $n \in \mathbb{N}^*$:

$$\begin{aligned}
& (n+1)^2 w_{n+1} - \left(1 - \frac{\lambda - 2}{n}\right) n^2 w_n \\
& \leqslant n\,(n - \lambda + 2)\left(\alpha_n^2 \delta_n - \delta_{n+1}\right) \\
& + (\lambda n + 1)\left((1 + \alpha_n) h_n + (\alpha_n^2 + \alpha_n)\delta_n - \alpha_n h_{n-1} - h_{n+1} - \gamma_{n+1}^* - \alpha_n \gamma_n^*\right) \\
& + 2(\lambda n + 1)\left(\alpha_n \langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^* \rangle - \langle x_{n+1} - x_{n+1}^*, x_{n+1}^* - x_n^* \rangle\right).
\end{aligned}$$

It follows that:

$$\begin{aligned} E_{n+1} - \left(1 - \frac{\lambda - 2}{n}\right) E_n \leqslant & A_1(n,\alpha)\delta_n + A_2(n,\alpha)\delta_{n+1} + B_1(n,\alpha)h_{n-1} \\ & + B_2(n,\alpha)h_n + B_3(n,\alpha)h_{n+1} + D_1(n,\alpha)\gamma^*_{n+1} \\ & + D_2(n,\alpha)\gamma^*_n + D_3(n,\alpha)\langle x_{n+1} - x^*_{n+1}, x^*_{n+1} - x^*_n\rangle \\ & + D_4(n,\alpha)\langle x_n - x^*_n, x^*_n - x^*_{n-1}\rangle \\ & + D_5(n,\alpha)\langle x_{n-1} - x^*_{n-1}, x^*_n - x^*_{n-1}\rangle, \end{aligned} \tag{70}$$

where:

- $A_1(n,\alpha) = \frac{17\alpha^2}{9} - \frac{8\alpha}{3} + 2 - \alpha\frac{(10\alpha^2-18\alpha+9)n+7\alpha^3-12\alpha^2+6\alpha}{3(n+\alpha)^2}$,
- $A_2(n,\alpha) = 1 - \frac{2\alpha}{3}$,
- $B_1(n,\alpha) = -\frac{2\alpha^2}{9} + \frac{4\alpha}{3} - 1 + \frac{3\alpha-2\alpha^3}{3(n+\alpha)} + \frac{8\alpha^3-24\alpha^2}{27n}$,
- $B_2(n,\alpha) = \frac{2\alpha^2}{9} - 2\alpha + 2 - \frac{3\alpha-2\alpha^3}{3(n+\alpha)}$,
- $B_3(n,\alpha) = \frac{2\alpha}{3} - 1$,
- $D_1(n,\alpha) = \frac{2\alpha}{3} - 1$,
- $D_2(n,\alpha) = -\frac{4\alpha}{3}n - 1 - \frac{4\alpha}{3} + \frac{10\alpha^2}{9} + \frac{3\alpha-2\alpha^3}{3(n+\alpha)}$,
- $D_3(n,\alpha) = \frac{4\alpha}{3} - 2$,
- $D_4(n,\alpha) = -\frac{4\alpha}{3}n - \frac{8\alpha}{3} + \frac{8\alpha^2}{9}$,
- $D_5(n,\alpha) = \frac{4\alpha}{3}n + 2 - \frac{4\alpha^2}{3} + \frac{\alpha(4\alpha^2-6)}{3(n+\alpha)}$.

Noticing that $B_3(n,\alpha) = -A_2(n,\alpha) = D_1(n,\alpha) = \frac{1}{2}D_3(n,\alpha)$ and:

$$B_1(n,\alpha) + B_2(n,\alpha) + B_3(n,\alpha) = \frac{8\alpha^2}{27}\frac{\alpha-3}{n} = \frac{4\alpha K(\alpha)}{3n},$$

where $K(\alpha) = \frac{2\alpha(\alpha-3)}{9}$, we get that

$$\begin{aligned} E_{n+1} - \left(1 - \frac{\lambda - 2}{n}\right) E_n \leqslant & \frac{4\alpha K(\alpha)}{3n}h_n + A_1(n,\alpha)\delta_n + B_1(n,\alpha)(h_{n-1} - h_n) \\ & + B_3(n,\alpha)(h_{n+1} - h_n - \delta_{n+1}) + B_3(n,\alpha)\gamma^*_{n+1} \\ & + D_2(n,\alpha)\gamma^*_n + 2B_3(n,\alpha)\langle x_{n+1} - x^*_{n+1}, x^*_{n+1} - x^*_n\rangle \\ & + D_4(n,\alpha)\langle x_n - x^*_n, x^*_n - x^*_{n-1}\rangle \\ & + D_5(n,\alpha)\langle x_{n-1} - x^*_{n-1}, x^*_n - x^*_{n-1}\rangle. \end{aligned} \tag{71}$$

We apply the following technical lemma that is an extension of [8, Lemma 4]. The proof can be found in Section B.5.

Lemma 8. *Let $n > \lambda$ and $(A,B) \in \mathbb{R}^2$. The following two claims hold:*

1.

$$\delta_n \leqslant \frac{2}{(n-\lambda)^2}b_n + \frac{8\alpha^2}{9(n-\lambda)^2}h_n, \tag{72}$$

where $b_n = \|\lambda(x_{n-1} - x^*_{n-1}) + n(x_n - x_{n-1})\|^2$ for any $n \in \mathbb{N}^*$.

2.

$$A\delta_n + B(h_{n-1} - h_n) \leqslant \left(2|A+B| + \frac{\sqrt{2}|B|}{\sqrt{\kappa}}\right)\left(1 + \frac{4\alpha^2}{9\kappa n^2}\right)\frac{E_n}{(n-\lambda)^2} - B\gamma_n^* + 2B\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle. \tag{73}$$

Inequality (73) ensures that for any $n > \lambda$:

$$\frac{4\alpha K(\alpha)}{3}\frac{h_n}{n} + A_1(n,\alpha)\delta_n + B_1(n,\alpha)(h_{n-1} - h_n) \leqslant \frac{\hat{C}_1(n,\alpha,\kappa)E_n}{(n-\lambda)^2} - B_1(n,\alpha)\gamma_n^* + 2B_1(n,\alpha)\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle,$$

and

$$B_3(n,\alpha)(h_{n+1} - h_n - \delta_{n+1}) \leqslant \frac{\hat{C}_2(n,\alpha,\kappa)E_{n+1}}{(n+1-\lambda)^2} - 2B_3(n,\alpha)\langle x_n - x_n^*, x_{n+1}^* - x_n^*\rangle + B_3(n,\alpha)\gamma_{n+1}^*,$$

where

$$\hat{C}_1(n,\alpha,\kappa) = 2|\frac{5}{3}\alpha^2 - \frac{4\alpha}{3} + 1 + R(n,\alpha)| + \sqrt{2}\left(\frac{|-\frac{2\alpha^2}{9} + \frac{4\alpha}{3} - 1 + Q(n,\alpha)|}{\sqrt{\kappa}}\right)\left(1 + \frac{4\alpha^2}{9\kappa n^2}\right) + \frac{4\alpha K(\alpha)}{3\kappa n}$$

with:

$$|R(\alpha,n)| = \left|A_1(n,\alpha) + B_1(n,\alpha) - (\frac{5}{3}\alpha^2 - \frac{4\alpha}{3} + 1)\right| \leqslant \frac{8\alpha^3}{n}$$

$$|Q(\alpha,n)| = \frac{\alpha^3}{3n}\left|n\frac{3-2\alpha^2}{\alpha^2(n+\alpha)} + 8\frac{\alpha-3}{9\alpha}\right| \leqslant \frac{\alpha^3}{n},$$

and $\hat{C}_2(n,\alpha,\kappa) = \left(\frac{2\alpha}{3} - 1\right)\left(4 + \frac{\sqrt{2}}{\sqrt{\kappa}}\right)\left(1 + \frac{4\alpha^2}{9\kappa(n+1)^2}\right)$. Coming back to (71), we get that:

$$\begin{aligned} E_{n+1} - \left(1 - \frac{\lambda-2}{n}\right)E_n \leqslant & \frac{\hat{C}_1(n,\alpha,\kappa)E_n}{(n-\lambda)^2} + \frac{\hat{C}_2(n,\alpha,\kappa)E_{n+1}}{(n+1-\lambda)^2} \\ & + 2B_3(n,\alpha)\gamma_{n+1}^* + 2B_3(n,\alpha)\langle x_{n+1} - x_{n+1}^*, x_{n+1}^* - x_n^*\rangle \\ & - 2B_3(n,\alpha)\langle x_n - x_n^*, x_{n+1}^* - x_n^*\rangle + (D_2(n,\alpha) - B_1(n,\alpha))\gamma_n^* \\ & + D_4(n,\alpha)\langle x_n - x_n^*, x_n^* - x_{n-1}^*\rangle \\ & + (D_5(n,\alpha) + 2B_1(n,\alpha))\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle. \end{aligned} \tag{74}$$

Note that for all $n \in \mathbb{N}^*$,

$$\gamma_n^* + \langle x_n - x_n^*, x_n^* - x_{n-1}^*\rangle - \langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle = \langle x_n - x_{n-1}, x_n^* - x_{n-1}^*\rangle,$$

and thus,

$$2B_3(n,\alpha)\left(\gamma_{n+1}^* + \langle x_{n+1} - x_{n+1}^*, x_{n+1}^* - x_n^*\rangle - \langle x_n - x_n^*, x_{n+1}^* - x_n^*\rangle\right) = 2(\lambda-1)\langle x_{n+1} - x_n, x_{n+1}^* - x_n^*\rangle.$$

Moreover, we can show that for any $n \geqslant \lambda - 2$,

$$D_4(n,\alpha) \leqslant D_2(n,\alpha) - B_1(n,\alpha) \leqslant -(D_5(n,\alpha) + 2B_1(n,\alpha)) \leqslant -2\lambda\left(n - 2\left(\lambda - 1\right)\right). \tag{75}$$

Since

$$\gamma_n^* \geqslant 0,\ \langle x_n - x_n^*, x_n^* - x_{n-1}^*\rangle \geqslant 0,\ \langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle \leqslant 0,$$

we get that

$$(D_2(n,\alpha) - B_1(n,\alpha))\gamma_n^* + D_4(n,\alpha)\langle x_n - x_n^*, x_n^* - x_{n-1}^*\rangle + (D_5(n,\alpha) + 2B_1(n,\alpha))\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle \leqslant -2\lambda\left(n - 2\left(\lambda - 1\right)\right)\langle x_n - x_{n-1}, x_n^* - x_{n-1}^*\rangle,$$

which is negative if $n \geqslant 2\lambda - 2$. By taking $n > \max\{\lambda, 2\lambda - 2\} = 2\lambda - 2$ (since $\lambda = \frac{2\alpha}{3}$ and $\alpha > 3$), we can combine the above inequality with (74)

$$E_{n+1} - \left(1 - \frac{\lambda - 2}{n}\right) E_n \leqslant \frac{\hat{C}_1(n,\alpha,\kappa)E_n}{(n-\lambda)^2} + \frac{\hat{C}_2(n,\alpha,\kappa)E_{n+1}}{(n+1-\lambda)^2} + 2(\lambda-1)\langle x_{n+1} - x_n, x^*_{n+1} - x^*_n\rangle. \tag{76}$$

As $\langle x_{n+1} - x_n, x^*_{n+1} - x^*_n\rangle \leqslant \delta_{n+1}$, for any $n > 2\lambda - 2$,

$$E_{n+1} - \left(1 - \frac{\lambda - 2}{n}\right) E_n \leqslant \frac{\hat{C}_1(n,\alpha,\kappa)E_n}{(n-\lambda)^2} + \frac{\hat{C}_2(n,\alpha,\kappa)E_{n+1}}{(n+1-\lambda)^2} + 2(\lambda-1)\delta_{n+1}.$$

Then, according to the first claim of Lemma 8 and the quadratic growth condition that can be rewritten with our notation as $h_n \leqslant \frac{E_n}{\kappa n^2}$ for any $n \in \mathbb{N}$, we get the following:

$$\delta_{n+1} \leqslant \frac{2}{(n+1-\lambda)^2} b_{n+1} + \frac{8\alpha^2}{9(n+1-\lambda)^2} h_{n+1} \leqslant \frac{2}{(n+1-\lambda)^2}\left(1 + \frac{4\alpha^2}{9\kappa(n+1)^2}\right) E_{n+1}.$$

Hence,

$$E_{n+1} - \left(1 - \frac{\lambda - 2}{n}\right) E_n \leqslant \frac{\tilde{C}_1(n,\alpha,\kappa)E_n}{(n-\lambda)^2} + \frac{\tilde{C}_2(n,\alpha,\kappa)E_{n+1}}{(n+1-\lambda)^2}, \tag{77}$$

where $\tilde{C}_1(n,\alpha,\kappa) = \hat{C}_1(n,\alpha,\kappa)$ and

$$\begin{aligned}\tilde{C}_2(n,\alpha,\kappa) &= \hat{C}_2(n,\alpha,\kappa) + 4(\lambda-1)\left(1 + \frac{4\alpha^2}{9\kappa(n+1)^2}\right) \\ &= \left(\frac{2\alpha}{3} - 1\right)\left(8 + \frac{\sqrt{2}}{\sqrt{\kappa}}\right)\left(1 + \frac{4\alpha^2}{9\kappa(n+1)^2}\right).\end{aligned}$$

As $\kappa \in (0,1]$, for any $n \geqslant \frac{4\alpha}{3\sqrt{\kappa}}$, we have that $\dfrac{1}{n-\lambda} = \dfrac{1}{n - \frac{2\alpha}{3}} \leqslant \dfrac{1}{n}\left(1 + \sqrt{\kappa}\right)$ and thus, for any $n \geqslant \frac{4\alpha}{3\sqrt{\kappa}}$,

$$E_{n+1} - \left(1 - \frac{\frac{2\alpha}{3} - 2}{n}\right) E_n \leqslant (1+\sqrt{\kappa})^2 \left(\tilde{C}_1(n,\alpha,\kappa)\frac{E_n}{n^2} + \tilde{C}_2(n,\alpha,\kappa)\frac{E_{n+1}}{(n+1)^2}\right). \tag{78}$$

Observe that this inequality is identical to the one obtained in [8, Proof of Lemma 1] under the assumption that F has a unique minimizer. The value of $\tilde{C}_1(n,\alpha,\kappa)$ does not change while $\tilde{C}_2(n,\alpha,\kappa)$ is slightly larger (in the case of uniqueness of the minimizer, $\tilde{C}_2(n,\alpha,\kappa)$ is equal to $\hat{C}_2(n,\alpha,\kappa)$). As a consequence, the bounds computed for $\tilde{C}_1(n,\alpha,\kappa)$ in [8] are still valid and in particular, there exist some real constants $\tilde{c}_1$ and $\tilde{c}_2$ such that for any $\alpha \geqslant 3 + \frac{3}{\sqrt{2}}$ and any $n \geqslant \frac{4\alpha}{3\sqrt{\kappa}}$,

$$\tilde{C}_1(n,\alpha,\kappa) \leqslant \frac{5}{4}\sqrt{\frac{2}{\kappa}} P(\alpha)\left(1 + \tilde{c}_1\sqrt{\kappa} + \tilde{c}_2\kappa\right), \tag{79}$$

where $P : \alpha \mapsto \frac{2}{9}(\alpha - 3)(\frac{8}{5}\alpha - 3) - 1$. Moreover, note that for any $n \geqslant \frac{4\alpha}{3\sqrt{\kappa}}$ and $\alpha \geqslant 3$,

$$\begin{aligned}\widetilde{C}_2(n,\alpha,\kappa) &= \left(\frac{2\alpha}{3} - 1\right)\left(8 + \frac{\sqrt{2}}{\sqrt{\kappa}}\right)\left(1 + \frac{4\alpha^2}{9\kappa(n+1)^2}\right) \\ &\leqslant \frac{5}{4}\sqrt{\frac{2}{\kappa}}\left(\frac{2\alpha}{3} - 1\right)\left(1 + 4\sqrt{2\kappa}\right).\end{aligned}$$

Hence, for any $\alpha \geqslant 3 + \frac{3}{\sqrt{2}}$:

$$\forall n \geqslant \frac{4\alpha}{3\sqrt{\kappa}},\ E_{n+1} - \left(1 - \frac{\frac{2\alpha}{3} - 2}{n}\right) E_n \leqslant \frac{\mathbf{C_1}(\alpha,\kappa)E_n}{n^2} + \frac{\mathbf{C_2}(\alpha,\kappa)E_{n+1}}{(n+1)^2}, \tag{80}$$

where:

- $\mathbf{C_1}(\alpha,\kappa) = \frac{5}{4}\sqrt{\frac{2}{\kappa}}\left[\frac{2}{9}(\alpha-3)\left(\frac{8}{5}\alpha-3\right)-1\right](1+\sqrt{\kappa})^2\,(1+\tilde{c}_1\sqrt{\kappa}+\tilde{c}_2\kappa)$,
- $\mathbf{C_2}(\alpha,\kappa) = \frac{5}{4}\sqrt{\frac{2}{\kappa}}\left(\frac{2\alpha}{3}-1\right)(1+\sqrt{\kappa})^2(1+4\sqrt{2\kappa})$.

From there, we refer the reader to [8] since the last steps of this proof are detailed in the proof of [8, Theorem 6]. We first integrate inequality (80) with the following lemma which is a slightly modified version of [8, Lemma 2].

Lemma 9. *Let $\alpha \geqslant 3$ and $n_0 \geqslant \frac{4\alpha}{3\sqrt{\kappa}}$. If the energy E_n satisfies* (80) *then:*

$$\forall n \geqslant n_0,\ E_n \leqslant E_{n_0}\left(\frac{n}{n_0}\right)^{-\left(\frac{2\alpha}{3}-2\right)} e^{\phi(n_0)}, \tag{81}$$

where $\phi(n_0) = \frac{5}{6n_0}\sqrt{\frac{2}{\kappa}}(\alpha-3)\left(\frac{16}{15}\alpha-1\right)\left(1+c\kappa^{\frac{1}{4}}\right)$ and $c>0$ is independent to α.

The proof of this lemma is identical to the proof of [8, Lemma 2] despite $\mathbf{C_2}(\alpha,\kappa)$ being larger than $C_2(\alpha,\kappa)$ in the other version. This difference is absorbed in the constant $c>0$.

Since $F(x_n)-F^* \leqslant \frac{L}{2n^2}E_n$, we get that for any $n \geqslant \frac{4\alpha}{3\sqrt{\kappa}}$,

$$F(x_n)-F^* \leqslant \frac{L}{2}\left(n_0^{\frac{2\alpha}{3}-2}e^{\phi(n_0)}\right)E_{n_0}n^{-\frac{2\alpha}{3}}.$$

It is then essential to choose a relevant value for n_0 to get a control as tight as possible on $F(x_n)-F^*$. This discussion is already detailed in [8] leading to the choice

$$n_0 = \frac{5}{4}\sqrt{\frac{2}{\kappa}}\left(\frac{16}{15}\alpha-1\right)\left(1+c\kappa^{\frac{1}{4}}\right),$$

which ensures that if κ is sufficiently small, then

$$\forall n \geqslant \frac{3\alpha}{\sqrt{\kappa}},\ F(x_n)-F^* \leqslant \frac{9}{4}e^{-2}M_0\left(\frac{8e}{3\sqrt{\kappa}}\alpha\right)^{\frac{2\alpha}{3}} n^{-\frac{2\alpha}{3}}, \tag{82}$$

where $M_0 = F(x_0)-F^*$.

We now prove the second claim of Theorem 2. According to (72), for any $n>\lambda$,

$$\delta_n \leqslant \frac{2}{(n-\lambda)^2}b_n + \frac{8\alpha^2}{9(n-\lambda)^2}h_n, \tag{83}$$

where $b_n = \|\lambda(x_{n-1}-x^*_{n-1})+n(x_n-x_{n-1})\|^2$. Considering the definition of the Lyapunov energy E_n, we have for any $n>\lambda$, $b_n \leqslant E_n$, hence:

$$\delta_n \leqslant \frac{2}{(n-\lambda)^2}E_n + \frac{8\alpha^2}{9(n-\lambda)^2}h_n, \tag{84}$$

Since F is assumed to satisfy a global quadratic growth condition $\mathcal{G}^2_\mu$ which implies that $h_n \leqslant \frac{E_n}{\kappa n^2}$ for any $n\in\mathbb{N}$, we get:

$$\forall n>\lambda,\ \delta_n \leqslant \frac{2}{(n-\lambda)^2}\left(1+\frac{4\alpha^2}{9\kappa n^2}\right)E_n. \tag{85}$$

Hence, for any $n \geqslant \frac{4\alpha}{3\sqrt{\kappa}}$, $\delta_n \leqslant \frac{5}{2(n-\lambda)^2}E_n$. By applying Lemma 9, we get that there exists some real constant $K>0$ such that $\delta_n \leqslant \frac{K}{n^{\frac{2\alpha}{3}}}$, which ensures that

$$\|x_n-x_{n-1}\| = \mathcal{O}\left(n^{-\frac{\alpha}{3}}\right). \tag{86}$$

Finally, the strong convergence of FISTA iterates in the case when F satisfies some global quadratic growth condition, follows from the summability of $\|x_n-x_{n-1}\|$ since $\alpha \geqslant 3+\frac{3}{\sqrt{2}} > 3$.

A Appendix

A.1 Handling non-uniqueness of the minimizers in the continuous setting

In this section we assume that F is a convex differentiable function having a L-Lipschitz gradient and a non-empty set of minimizers X^*. We introduce the following Lyapunov energy:

$$\mathcal{E}(t) = t^2\left(F(x(t)) - F^*\right) + \frac{1}{2}\|\lambda(x(t) - x^*(t)) + t\dot{x}(t)\|^2 + \frac{\xi}{2}\|x(t) - x^*(t)\|^2, \tag{87}$$

where for all $t \geqslant t_0$, $x^*(t)$ denotes the projection of $x(t)$ onto X^*, i.e

$$x^*(t) = \arg\inf_{x^* \in X^*} \|x(t) - x^*\|^2.$$

Assume additionally that X^* is second-order regular in the sense of Definition 6 to that the projection $t \mapsto x^*(t)$ onto X^* is right differentiable, as well as $\mathcal{E}$, and the right-hand derivative of x^* is equal to $P'_{X^*}(x(t), \dot{x}(t))$. For the sake of simplicity, let $\dot{x^*}$ and $\dot{\mathcal{E}}$ denote the corresponding right-hand derivatives. We can then write that:

$$\dot{\mathcal{E}}(t) = D(t) - (\lambda^2 + \xi)\langle x(t) - x^*(t), \dot{x^*}(t)\rangle - \lambda t\langle \dot{x}(t), \dot{x^*}(t)\rangle, \tag{88}$$

where

$$\begin{aligned} D(t) =& 2t\left(F(x(t)) - F^*\right) + t^2\langle \nabla F(x(t)), \dot{x}(t)\rangle + \langle \lambda(x(t) - x^*(t)) + t\dot{x}(t), (\lambda+1)\dot{x}(t) + t\ddot{x}(t)\rangle \\ &+ \xi\langle x(t) - x^*(t), \dot{x}(t)\rangle. \end{aligned}$$

Observe that D is exactly equal to $\dot{\mathcal{E}}$ if F has a unique minimizer x^*. The objective is then to control the additional terms $\langle x(t) - x^*(t), \dot{x^*}(t)\rangle$ and $\langle \dot{x}(t), \dot{x^*}(t)\rangle$. We introduce Figure 1 to give an intuition of the behavior of these terms.

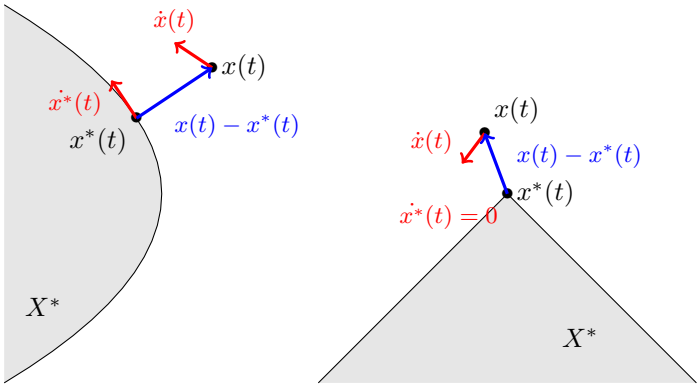

Figure 1: Behavior of $\dot{x^*}$ for a set of minimizers having a C^2 bound (on the left) and a polyhedral set of minimizers (on the right).

We can first prove that $\langle \dot{x}(t), \dot{x^*}(t)\rangle$ is positive by using the expression $\dot{x^*}(t) = \lim_{h\to 0} \frac{x^*(t+h) - x^*(t)}{h}$ and the property of the projection onto a convex set. Indeed, as X^* is a closed convex set, for any $x \in \mathcal{H}$ and $u \in X^*$:

$$\langle x - P_{X^*}(x), u - P_{X^*}(x)\rangle \leqslant 0.$$

Thus, for any $h > 0$ we have:

$$\begin{aligned} \langle x(t+h) - x(t), x^*(t+h) - x^*(t)\rangle &= \langle x(t+h) - x^*(t+h), x^*(t+h) - x^*(t)\rangle \\ &\quad + \|x^*(t+h) - x^*(t)\|^2 \\ &\quad + \langle x(t) - x^*(t), x^*(t) - x^*(t+h)\rangle \\ &\geqslant 0. \end{aligned}$$

By considering h tending towards 0 we can deduce that $\langle \dot{x}(t), \dot{x^*}(t) \rangle \geqslant 0$.

In [13, Theorem 7.2] the authors give an expression of the directional derivative $P'_{\mathcal{S}}(x,d)$ for a closed convex set $\mathcal{S} \subset \mathcal{X}$ being second order regular at $P_{\mathcal{S}}(x)$ for some $x \in \mathcal{X}$. This directional derivative satisfies:

$$\langle x - P_{\mathcal{S}}(x), P'_{\mathcal{S}}(x,d) \rangle = 0.$$

Considering the assumptions made on X^* we can deduce that $\langle x(t) - x^*(t), \dot{x^*}(t) \rangle = 0$ for all $t \geqslant t_0$.

These results ensure that for any choices of parameters $\lambda > 0$ and $\xi \in \mathbb{R}$, we have that $\dot{\mathcal{E}}^*(t) \leqslant D(t)$. From this point, it is sufficient to apply the following lemma to extend the desired convergence results to the non-unique case. A proof is given in Section B.6.

Lemma 10. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a continuous function which is right-differentiable. Assume that*

$$\forall t \geqslant t_0, \ \phi_+(t) \leqslant \psi(t), \tag{89}$$

where $\phi_+(t) = \lim\limits_{h \to 0, \ h>0} \dfrac{\phi(t+h) - \phi(t)}{h}$ denotes the right derivative of ϕ at t. Then,

$$\forall t \geqslant t_0, \ \phi(t) \leqslant \phi(t_0) + \int_{t_0}^{t} \psi(u) du. \tag{90}$$

A.2 Proof of Theorem 4 under Hölderian growth condition

We focus our analysis on the following Lyapunov energy introduced in [31]:

$$\mathcal{J}(t) = t^p \left(t^2 (F(x(t)) - F^*) + \frac{1}{2} \| \lambda(x(t) - x^*(t)) + t\dot{x}(t) \|^2 + \frac{\xi}{2} \| x(t) - x^*(t) \|^2 \right), \tag{91}$$

where $p = 1 + \frac{4}{\gamma - 2}$ and $\lambda > 0$. We use the following notations:

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

The Lyapunov function can be rewritten as follows:

$$\mathcal{J}(t) = t^{p+1} \left(a(t) + b(t) + \xi c(t) \right).$$

Following the discussion on the derivability of $\mathcal{E}^*$ defined in (87) in Section A.1, we can say that under the assumption made on X^*, $\mathcal{E}^*$ is right differentiable. Noticing that $\mathcal{J}(t) = t^p \mathcal{E}^*(t)$, this is also true for $\mathcal{J}$. For the sake of simplicity, the right derivative of $\mathcal{J}$ is denoted $\mathcal{J}'$. By adapting [31, Lemma 4.4] to our case, we get that if $\xi = \lambda(\lambda + 1 - \alpha)$, then

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

Let $\lambda = \alpha - 1 - \frac{p}{2}$. Under the condition $\alpha > \frac{9}{2} + \frac{6}{\gamma - 2}$, we have that

$$\begin{cases} 2 + p - \lambda < 0, \\ 2(\lambda + 1 - \alpha) + p = 0, \\ \lambda(\lambda + 1 - \alpha)(p - 2\lambda) > 0. \end{cases}$$

As a consequence, we can write that:

$$\mathcal{J}'(t) \leqslant t^p \left(A a(t) + C c(t) \right),$$

where $A = 3 - \alpha + \frac{3p}{2} < 0$ and $C = p \left(\alpha - 1 - \frac{p}{2} \right) (\alpha - 1) > 0$. We can apply [31, Lemma 4.5] which we recall below.

Lemma 11. *If F satisfies the inequality (4) for some $\gamma > 2$ and $K > 0$, i.e. F satisfies $\mathcal{G}^{\gamma}_{loc}$, then there exists $t_1 \geqslant t_0$ such that for all $t \geqslant t_1$,*

$$t^{p_2+1}c(t) \leqslant \frac{K^{-\frac{2}{\gamma}}}{2}\left(t^{p_2+1}a(t)\right)^{\frac{2}{\gamma}},$$

where $p_2 = \frac{4}{\gamma-2}$.
It follows that for any $m \in \mathbb{R}$, there exists $M \in \mathbb{R}$ such that for any $t \geqslant t_1$,

$$t^{p_2+1}(mc(t) - a(t)) \leqslant M.$$

As $p = p_2+1$, this lemma ensures that there exists $M_1 \in \mathbb{R}$ such that for all $t \geqslant t_0$, $\mathcal{J}'(t) \leqslant M_1$. Then, Lemma 10 gives us that that there exists $M_2 \in \mathbb{R}$ such that $\mathcal{J}(t) \leqslant M_1 t + M_2$ and consequently

$$t^{p+1}(a(t) + \xi c(t)) \leqslant M_1 t + M_2.$$

Therefore, for t sufficiently large,

$$t^p a(t) \leqslant 2M_1 + |\xi| t^p c(t).$$

The first claim of Lemma 11 gives us that there exists $M_3 > 0$ such that:

$$t^p a(t) \leqslant 2M_1 + M_3\left(t^p a(t)\right)^{\frac{2}{\gamma}}. \tag{92}$$

Lemma 7 guarantees that there exists $M_4 > 0$ such that for t sufficiently large,

$$t^p a(t) \leqslant M_4,$$

and thus,

$$F(x(t)) - F^* \leqslant \frac{M_4}{t^{p+1}}. \tag{93}$$

As $p + 1 = \frac{2\gamma}{\gamma-2}$, the first claim is proved.
We prove the second claim by coming back to the inequality $\mathcal{J}(t) \leqslant M_1 t + M_2$. By applying the inequality $\|u\|^2 \leqslant 2\|u+v\|^2 + 2\|v\|^2$ to $u = t\dot{x}(t)$ and $v = \lambda(x(t) - x^*(t))$, we get that

$$b(t) \geqslant \frac{t}{2}\|\dot{x}(t)\|^2 - \lambda^2 c(t).$$

Consequently, for sufficiently large t we have that:

$$t^p\left(a(t) - \left(|\xi| + \lambda^2\right)c(t) + \frac{t}{2}\|\dot{x}(t)\|^2\right) \leqslant 2M_1.$$

Lemma 11 gives us that there exists $M_5 > 0$ such that:

$$t^p\left(a(t) - \left(|\xi| + \lambda^2\right)c(t)\right) \geqslant t^p a(t) - M_5\left(t^p a(t)\right)^{\frac{2}{\gamma}}.$$

Lemma 12. *Let $g : x \mapsto x - Kx^{\delta}$ for some $K > 0$ and $\delta \in (0,1)$. Then for all $x \geqslant 0$,*

$$g(x) \geqslant K(\delta-1)(\delta K)^{\frac{\delta}{1-\delta}}.$$

Lemma 12 ensures that there exists $M_6 \in \mathbb{R}$ such that $t^p\left(a(t) - \left(|\xi| + \lambda^2\right)c(t)\right) \geqslant M_6$. Hence, for t sufficiently large,

$$\frac{t^{p+1}}{2}\|\dot{x}(t)\|^2 \leqslant 2M_1 + M_6,$$

and thus:

$$\|\dot{x}(t)\| \leqslant \frac{M_7}{t^{\frac{p+1}{2}}}, \tag{94}$$

where $M_7 = 4M_1 + 2M_6 \geqslant 0$ and $\frac{p+1}{2} = \frac{\gamma}{\gamma-2}$. Thus the trajectory $t \mapsto (x(t), \dot{x}(t))$ has a finite length and $t \mapsto x(t)$ strongly converges to a minimizer of F.

A.3 Proof of Theorem 5 and Proposition 1 under a quadratic growth condition

The proof of Theorem 5 is very similar to the one of [8, Theorem 5] and is not reproduced entirely here, but we recall the essential steps of this proof. We first introduce the following Lyapunov energy:

$$\mathcal{E}(t) = t^2(F(x(t)) - F^*) + \frac{1}{2}\|\lambda(x(t) - x^*(t)) + t\dot{x}(t)\|^2,\ \lambda = \frac{2\alpha}{3} \tag{95}$$

where $x^*(t)$ denotes the projection of the trajectory $x(t)$ solution of (AVD) onto the set of minimizers. According to the discussion in Section A.1, the energy $\mathcal{E}$ is (right-)differentiable, allowing to carry out the proof of [8, Theorem 5] without any particular difficulty. The only additional challenge is to deal with the terms involving $\dot{x}^*(t)$, which is described in Section A.1).

First case: F satisfies a global quadratic growth condition (Proposition 1) Following the proof of [8, Theorem 5], we can show that the right derivative of $\mathcal{E}$ denoted $\mathcal{E}'$ satisfies:

$$\forall t \geqslant t_0, \quad \mathcal{E}'(t) + \frac{\lambda - 2}{t}\mathcal{E}(t) \leqslant \phi(t)\mathcal{E}(t),$$

where:

$$\phi : t \mapsto \frac{2\alpha\,(\alpha - 3)}{9\mu t^2}\left(\sqrt{\mu} + \frac{2\alpha}{3t}(1 + \sqrt{2}) + \frac{4\alpha^2}{9\sqrt{\mu}t^2}\right).$$

This inequality combined with Lemma 10 ensures that $t \mapsto \mathcal{E}(t)t^{\lambda-2}e^{\Phi(t)}$, where $\Phi : t \mapsto \int_t^{+\infty}\phi(s)ds$, is decreasing on $[t_0, +\infty)$. As a consequence, for any $t_1 \geqslant t_0$:

$$\forall t \geqslant t_1, \quad \mathcal{E}(t) \leqslant \mathcal{E}(t_1)\left(\frac{t_1}{t}\right)^{\lambda-2} e^{\Phi(t_1)-\Phi(t)}.$$

The next steps of the demonstration rely on showing that Φ is positive, choosing a relevant value for t_1 and bounding each term of the inequality.

Second case: F satisfies a local quadratic growth condition (Theorem 5) Similar to Lemma 1, we can use the coercivity of F and the convergence of $t \mapsto F(x(t)) - F^*$ to 0 (since it is well known that for any $\alpha > 3$, $F(x(t)) - F^* = \mathcal{O}\left(t^{-2}\right)$, see [33]) to prove the existence of $t_\varepsilon \geqslant t_0$ such that:

$$\forall t \geqslant t_\varepsilon, \quad \frac{\mu}{2}d(x(t), X^*) \leqslant F(x(t)) - F^*).$$

By following the proof in the global case and replacing t_0 by t_ε, we can easily find the desired asymptotic result:

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

Showing that the trajectory has a finite length We consider that F satisfies $\mathcal{G}^2_{\mu,loc}$. It is shown that there exist some $t_1 \geqslant t_\varepsilon$ and $K > 0$ such that

$$\forall t \geqslant t_1,\ \mathcal{E}(t) \leqslant Kt^{-\frac{2\alpha}{3}+2}. \tag{96}$$

Moreover, by applying inequality $\|u\|^2 \leqslant 2\|u + v\|^2 + 2\|v\|^2$, we obtain that:

$$\|\dot{x}(t)\|^2 \leqslant \frac{2}{t^2}\|\lambda(x(t) - x^*(t)) + t\dot{x}(t)\|^2 + \frac{2\lambda^2}{t^2}\|x(t) - x^*(t)\|^2. \tag{97}$$

Hence, the assumption $\mathcal{G}^2_{\mu,loc}$ guarantees that

$$\forall t \geqslant t_1, \quad \|\dot{x}(t)\|^2 \leqslant \frac{4}{t^2}\left(1 + \frac{\lambda^2}{\mu t^2}\right)\mathcal{E}(t). \tag{98}$$

Inequality (96) gets us to the conclusion:

$$\|\dot{x}(t)\| = \mathcal{O}\left(t^{-\frac{\alpha}{3}}\right). \tag{99}$$

Since $\alpha > 3$, we obtain that $\int_{t_1}^{+\infty} \|\dot{x}(t)\| dt < +\infty$ which implies that the trajectory $x(\cdot)$ has a finite length. Combined with the convergence rate on function values, this guarantees that $x(\cdot)$ converges to some minimizer of F. □

B Proofs of technical Lemmas 2, 3, 5, 8 and 10

B.1 Proof of Lemma 2

Let $n \in \mathbb{N}^*$. By rewriting

$$x_n - x_n^* = \frac{1}{2}\left((x_n - x_{n-1}) + (x_{n-1} - x_{n-1}^*) + (x_{n-1}^* - x_n^*) + (x_n - x_n^*)\right),$$

we get that:

$$\langle x_n - x_n^*, x_n - x_{n-1}\rangle = \frac{1}{2}\delta_n + \frac{1}{2}\langle (x_{n-1} - x_{n-1}^*) + (x_{n-1}^* - x_n^*) + (x_n - x_n^*), x_n - x_{n-1}\rangle.$$

Noticing that $x_n - x_{n-1} = (x_n - x_n^*) + (x_n^* - x_{n-1}^*) + (x_{n-1}^* - x_{n-1})$ leads to:

$$\begin{aligned}
2\langle x_n - x_n^*, x_n - x_{n-1}\rangle &= \delta_n + \langle x_{n-1} - x_{n-1}^*, x_n - x_n^*\rangle + \langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle \\
&\quad - h_{n-1} - \langle x_n^* - x_{n-1}^*, x_n - x_n^*\rangle + \langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle \\
&\quad - \gamma_n^* + \langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle - \langle x_{n-1} - x_{n-1}^*, x_n - x_n^*\rangle + h_n \\
&= h_n - h_{n-1} + \delta_n - \gamma_n^* + 2\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle.
\end{aligned}$$

The second claim is proved using the same approach. We rewrite

$$x_{n-1} - x_{n-1}^* = \frac{1}{2}\left((x_{n-1} - x_n) + (x_n - x_n^*) + (x_n^* - x_{n-1}^*) + (x_{n-1}^* - x_{n-1})\right),$$

and consequently:

$$2\langle x_{n-1} - x_{n-1}^*, x_n - x_{n-1}\rangle = -\delta_n + \langle (x_n - x_n^*) + (x_n^* - x_{n-1}^*) + (x_{n-1}^* - x_{n-1}), x_n - x_{n-1}\rangle.$$

By applying the same rewriting of $x_n - x_{n-1}$, simple calculations give that:

$$\langle x_{n-1} - x_{n-1}^*, x_n - x_{n-1}\rangle = \frac{1}{2}(h_n - h_{n-1} - \delta_n + \gamma_n^*) + \langle x_n - x_n^*, x_n^* - x_{n-1}^*\rangle.$$

□

B.2 Proof of Lemma 3

The first claim is straightforward as Lemma 3.1 of [15] ensures that:

$$F(x_{n+1}) - F(x_n) \leqslant \frac{L}{2}\left(\|y_n - x_n\|^2 - \|x_{n+1} - x_n\|^2\right).$$

By writing $y_n = x_n + \alpha_n(x_n - x_{n-1})$ and $\frac{2}{L}(F(x_{n+1}) - F(x_n)) = w_{n+1} - w_n$, we can conclude.

By applying Lemma 3.1 of [15] to an other couple of points, we get that:

$$F(x_{n+1}) - F^* \leqslant \frac{L}{2}\left(\|y_n - x_n^*\|^2 - \|x_{n+1} - x_n^*\|^2\right).$$

It follows that:

$$\begin{aligned} w_{n+1} &\leqslant \|x_n + \alpha_n(x_n - x_{n-1}) - x_n^*\|^2 - \|(x_{n+1} - x_{n+1}^*) + (x_{n+1}^* - x_n^*)\|^2 \\ &\leqslant h_n + \alpha_n^2\delta_n - h_{n+1} - \gamma_{n+1}^* + 2\alpha_n\langle x_n - x_n^*, x_n - x_{n-1}\rangle \\ &\quad - 2\langle x_{n+1} - x_{n+1}^*, x_{n+1}^* - x_n^*\rangle. \end{aligned}$$

Recall that the first claim of Lemma 2 ensures that:

$$\langle x_n - x_n^*, x_n - x_{n-1}\rangle = \frac{1}{2}(h_n - h_{n-1} + \delta_n - \gamma_n^*) + \langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle,$$

we can deduce that:

$$\begin{aligned} w_{n+1} \leqslant (1+\alpha_n)h_n + (\alpha_n^2 + \alpha_n)\delta_n - \alpha_n h_{n-1} - h_{n+1} - \gamma_{n+1}^* - \alpha_n\gamma_n^* \\ + 2\alpha_n\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle - 2\langle x_{n+1} - x_{n+1}^*, x_{n+1}^* - x_n^*\rangle. \end{aligned}$$

□

B.3 Proof of Lemma 5

Recall the definition of the discrete Lyapunov energy $\mathcal{E}$:

$$\mathcal{E}_n = n^2 w_n + b_n + \xi h_n + \lambda n\alpha_n^2\delta_n. \tag{100}$$

Observe that for any $n \in \mathbb{N}$,

$$b_n = \lambda^2 h_n + n^2\alpha_n^2\delta_n + 2\lambda n\alpha_n\langle x_n - x_n^*, x_n - x_{n-1}\rangle, \tag{101}$$

and by applying the first claim of Lemma 2 we get that:

$$\begin{aligned} b_n =& \lambda^2 h_n + \lambda n\alpha_n(h_n - h_{n-1}) + n\alpha_n(n\alpha_n + \lambda)\delta_n \\ &- \lambda n\alpha_n\gamma_n^* + 2\lambda n\alpha_n\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle. \end{aligned} \tag{102}$$

As a consequence,

$$\begin{aligned} b_{n+1} - b_n =& \lambda(\lambda + (n+1)\alpha_{n+1})(h_{n+1} - h_n) + (n+1)\alpha_{n+1}\left((n+1)\alpha_{n+1} + \lambda\right)\delta_{n+1} \\ &- \lambda n\alpha_n(h_n - h_{n-1}) - n\alpha_n(n\alpha_n + \lambda)\delta_n - \lambda(n+1)\alpha_{n+1}\gamma_{n+1}^* + \lambda n\alpha_n\gamma_n^* \\ &+ 2\lambda(n+1)\alpha_{n+1}\langle x_n - x_n^*, x_{n+1}^* - x_n^*\rangle - 2\lambda n\alpha_n\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle. \end{aligned} \tag{103}$$

On the other hand, we have that:

$$(n+1)^2 w_{n+1} - n^2 w_n = n^2(w_{n+1} - w_n) + (2n+1)w_{n+1}, \tag{104}$$

and by applying the first claim of Lemma 3:

$$(n+1)^2 w_{n+1} - n^2 w_n \leqslant n^2(\alpha_n^2\delta_n - \delta_{n+1}) + (2n+1)w_{n+1}, \tag{105}$$

By combining (103) and (105), we get that:

$$\begin{aligned} \mathcal{E}_{n+1} - \mathcal{E}_n \leqslant& (2n+1)w_{n+1} - \lambda n(\alpha_n + \alpha_n^2)\delta_n \\ &+ ((n+1)\alpha_{n+1}((n+1)\alpha_{n+1} + \lambda) + \lambda(n+1)\alpha_{n+1}^2 - n^2)\delta_{n+1} \\ &+ \lambda n\alpha_n(h_{n-1} - h_n) - \lambda\left(\lambda + (n+1)\alpha_{n+1} + \frac{\xi}{\lambda}\right)(h_n - h_{n+1}) \\ &- \lambda(n+1)\alpha_{n+1}\gamma_{n+1}^* + 2\lambda(n+1)\alpha_{n+1}\langle x_n - x_n^*, x_{n+1}^* - x_n^*\rangle \\ &+ \lambda n\alpha_n\gamma_n^* - 2\lambda n\alpha_n\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle. \end{aligned} \tag{106}$$

Observe that the second claim of Lemma 3 guarantees that:

$$\begin{aligned}-\lambda n w_{n+1} + \lambda n(h_n - h_{n+1}) + \lambda n\alpha_n(h_n - h_{n-1}) + \lambda n(\alpha_n + \alpha_n^2)\delta_n - \lambda n\gamma_{n+1}^* - \lambda n\alpha_n\gamma_n^* \\ +2\lambda n\alpha_n\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle - 2\lambda n\langle x_{n+1} - x_{n+1}^*, x_{n+1}^* - x_n^*\rangle \leqslant 0.\end{aligned} \tag{107}$$

Adding inequality (107) to (106) leads to

$$\begin{aligned}\mathcal{E}_{n+1} - \mathcal{E}_n \leqslant &((2-\lambda)n+1)w_{n+1} + ((n+1)\alpha_{n+1}((n+1)\alpha_{n+1}+\lambda) + \lambda(n+1)\alpha_{n+1}^2 - n^2)\delta_{n+1} \\ &- \lambda\left(\lambda + (n+1)\alpha_{n+1} + \frac{\xi}{\lambda} - n\right)(h_n - h_{n+1}) + \mathcal{X}_n^*,\end{aligned} \tag{108}$$

where

$$\begin{aligned}\mathcal{X}_n^* = (-\lambda(n+1)\alpha_{n+1} - \lambda n)\gamma_{n+1}^* + 2\lambda(n+1)\alpha_{n+1}\langle x_n - x_n^*, x_{n+1}^* - x_n^*\rangle \\ -2\lambda n\langle x_{n+1} - x_{n+1}^*, x_{n+1}^* - x_n^*\rangle.\end{aligned}$$

Observe that since X^* is a closed convex set, $\langle x_n - x_n^*, x_{n+1}^* - x_n^*\rangle \leqslant 0$ and $\langle x_{n+1} - x_{n+1}^*, x_{n+1}^* - x_n^*\rangle \geqslant 0$. Hence, for any $n \in \mathbb{N}$, $\mathcal{X}_n^* \leqslant 0$. By choosing $\xi = \lambda(\lambda + 1 - \alpha)$, we then obtain that:

$$\mathcal{E}_{n+1} - \mathcal{E}_n \leqslant ((2-\lambda)n+1)w_{n+1} + A_1(n)\delta_{n+1} + A_2(n)(h_n - h_{n+1}), \tag{109}$$

where:

- $A_1(n) = 2(\lambda + 1 - \alpha)n + \frac{n^2(3\alpha^2 - 3\alpha\lambda - 2\alpha + 2\lambda + 1) + n(2\alpha^3 - 2\alpha^2\lambda + 2\alpha^2 - 2\alpha\lambda - 2\alpha + 4\lambda + 2) + 1 + 2\lambda + \alpha\lambda}{(n+1+\alpha)^2}$,
- $A_2(n) = -2\lambda(\lambda + 1 - \alpha) - \frac{\alpha^2\lambda}{n+1+\alpha}$.

Note that by rewriting (102) we get that:

$$h_{n-1} - h_n = -\frac{1}{\lambda n\alpha_n}b_n + \frac{\lambda}{n\alpha_n}h_n + \frac{n\alpha_n + \lambda}{\lambda}\delta_n - \gamma_n^* + 2\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle. \tag{110}$$

This ensures that:

$$\begin{aligned}\mathcal{E}_{n+1} - \mathcal{E}_n \leqslant &((2-\lambda)n+1)w_{n+1} - \frac{A_2(n)}{\lambda(n+1)\alpha_{n+1}}b_{n+1} + \frac{\lambda A_2(n)}{(n+1)\alpha_{n+1}}h_{n+1} \\ &+ \left(A_1(n) + \frac{(n+1)\alpha_{n+1} + \lambda}{\lambda}A_2(n)\right)\delta_{n+1} \\ &- A_2(n)\left(\gamma_{n+1}^* - 2\langle x_n - x_n^*, x_{n+1}^* - x_n^*\rangle\right),\end{aligned} \tag{111}$$

which is the desired inequality.

□

B.4 Proof of Lemma 6

Let $\xi = \lambda(\lambda + 1 - \alpha)$. Let $\mathcal{J}_n = n^p\mathcal{E}_n$ with $p = 1 + \frac{4}{\gamma - 2}$. Elementary computations show that:

$$\mathcal{J}_{n+1} - \mathcal{J}_n = n^p\left(\mathcal{E}_{n+1} - \mathcal{E}_n\right) + \left((n+1)^p - n^p\right)\mathcal{E}_{n+1}. \tag{112}$$

Observe that for any $n \in \mathbb{N}$, $(n+1)^p - n^p \in \left[pn^{p-1}, p(n+1)^{p-1}\right]$. Therefore, if we make the assumption that $\lambda \leqslant \alpha - 1$, we obtain that $\xi \leqslant 0$ and:

$$\begin{aligned}\left((n+1)^p - n^p\right)\mathcal{E}_{n+1} \leqslant &\, p(n+1)^{p+1}w_{n+1} + p(n+1)^{p-1}b_{n+1} + p\xi n^{p-1}h_{n+1} \\ &+ p\lambda(n+1)^p\alpha_{n+1}^2\delta_{n+1}.\end{aligned} \tag{113}$$

By applying Lemma 5 and the above inequality we get that:

$$\begin{aligned}\mathcal{J}_{n+1} - \mathcal{J}_n \leqslant &\left(n^p\left((2-\lambda)n+1\right) + p(n+1)^{p+1}\right)w_{n+1} \\ &+ \left(n^pB_1(n) + p(n+1)^{p-1}\right)b_{n+1} \\ &+ \left(n^pB_2(n) + p\xi n^{p-1}\right)h_{n+1} \\ &+ \left(n^pB_3(n) + p\lambda(n+1)^p\alpha_{n+1}^2\right)\delta_{n+1} \\ &- n^pB_4(n)\left(\gamma_{n+1}^* - 2\langle x_n - x_n^*, x_{n+1}^* - x_n^*\rangle\right).\end{aligned} \tag{114}$$

The inequality $\|u\|^2 \leqslant 2\|u+v\|^2+2\|v\|^2$ applied at $u = \alpha_n(x_n - x_{n-1})$ and $v = \lambda(x_n - x_n^*)$ ensures that:

$$\delta_n \leqslant \frac{2}{n^2\alpha_n^2} b_n + \frac{2\lambda^2}{n^2\alpha_n^2} h_n. \tag{115}$$

Thus,

$$\begin{aligned}
\mathcal{J}_{n+1} - \mathcal{J}_n \leqslant\ & \left(n^p\left((2-\lambda)n+1\right) + p(n+1)^{p+1}\right) w_{n+1} \\
& + \left(n^p B_1(n) + p(n+1)^{p-1} + 2\frac{n^p|B_3(n)| + p\lambda(n+1)^p\alpha_{n+1}^2}{(n+1)^2\alpha_{n+1}^2}\right) b_{n+1} \\
& + \left(n^p B_2(n) + p\xi n^{p-1} + 2\lambda^2\frac{n^p|B_3(n)| + p\lambda(n+1)^p\alpha_{n+1}^2}{(n+1)^2\alpha_{n+1}^2}\right) h_{n+1} \\
& - n^p B_4(n)\left(\gamma_{n+1}^* - 2\langle x_n - x_n^*, x_{n+1}^* - x_n^*\rangle\right).
\end{aligned} \tag{116}$$

By replacing ξ by its value and reorganizing each term, we get to the conclusion.

□

B.5 Proof of Lemma 8

Let $(A, B) \in \mathbb{R}^2$. Elementary computations show that for any $n \in \mathbb{N}^*$,

$$h_{n-1} - h_n - \delta_n = -2\langle x_n - x_{n-1}, x_n - x_n^*\rangle + 2\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle - \gamma_n^*.$$

Consequently, for any $n \in \mathbb{N}^*$,

$$\begin{aligned}
A\delta_n + B(h_{n-1} - h_n) &= (A+B)\delta_n + B(h_{n-1} - h_n - \delta_n) \\
&= (A+B)\delta_n - 2B\langle x_n - x_{n-1}, x_n - x_n^*\rangle \\
&\quad + 2B\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle - B\gamma_n^* \\
&\leqslant (A+B)\delta_n + 2|B|\,|\langle x_n - x_{n-1}, x_n - x_n^*\rangle| \\
&\quad + 2B\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle - B\gamma_n^*.
\end{aligned}$$

Moreover, note that for any $n \in \mathbb{N}^*$ and $\theta > 0$:

$$2|\langle x_n - x_{n-1}, x_n - x_n^*\rangle| \leqslant \frac{h_n}{\theta} + \theta\delta_n. \tag{117}$$

Hence,

$$A\delta_n + B(h_{n-1} - h_n) \leqslant (A + B + \theta|B|)\delta_n + \frac{|B|}{\theta} h_n + 2B\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle - B\gamma_n^*.\text{‘}$$

We define $b_n := \|\lambda(x_{n-1} - x_{n-1}^*) + n(x_n - x_{n-1})\|^2$. By developing the expression of b_n we get that:

$$\begin{aligned}
b_n &= \|\lambda(x_n - x_n^*) + (n-\lambda)(x_n - x_{n-1}) + \lambda(x_n^* - x_{n-1}^*)\|^2 \\
&= \|\lambda(x_n - x_n^*) + (n-\lambda)(x_n - x_{n-1})\|^2 + \lambda^2\gamma_n^* \\
&\quad + 2\lambda^2\langle x_n - x_n^*, x_n^* - x_{n-1}^*\rangle + 2\lambda(n-\lambda)\langle x_n - x_{n-1}, x_n^* - x_{n-1}^*\rangle.
\end{aligned}$$

By applying the following inequality to $u = (n-\lambda)(x_n - x_{n-1})$ and $v = \lambda(x_n - x_n^*)$:

$$\|u\|^2 \leqslant 2\|u+v\|^2 + 2\|v\|^2,$$

it comes that:

$$\begin{aligned}
(n-\lambda)^2\delta_n &\leqslant 2\|\lambda(x_n - x_n^*) + (n-\lambda)(x_n - x_{n-1})\|^2 + 2\lambda^2 h_n \\
&\leqslant 2b_n + \frac{8\alpha^2}{9} h_n - \Delta_n^*,
\end{aligned}$$

where $\Delta_n^* = 2\left(\lambda^2\gamma_n^* + 2\lambda^2\langle x_n - x_n^*, x_n^* - x_{n-1}^*\rangle + 2\lambda(n-\lambda)\langle x_n - x_{n-1}, x_n^* - x_{n-1}^*\rangle\right)$. As $\Delta_n^* \geqslant 0$ we get the first claim of the lemma i.e.

$$\forall n > \lambda, \quad \delta_n \leqslant \frac{2}{(n-\lambda)^2} b_n + \frac{8\alpha^2}{9(n-\lambda)^2} h_n. \tag{118}$$

This inequality implies that for any $n > \lambda$,

$$\begin{aligned} A\delta_n + B(h_{n-1} - h_n) \leqslant (|A+B| + \theta|B|)\frac{2}{(n-\lambda)^2} b_n + \left((|A+B| + \theta|B|)\frac{8\alpha^2}{9(n-\lambda)^2} + \frac{|B|}{\theta}\right) h_n \\ + 2B\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle - B\gamma_n^*. \end{aligned}$$

As F satisfies $\mathcal{G}_\mu^2$, we can write that $h_n \leqslant \frac{w_n}{s\mu}$ and thus,

$$\begin{aligned} A\delta_n + B(h_{n-1} - h_n) \leqslant (|A+B| + \theta|B|)\frac{2}{(n-\lambda)^2} b_n + \left((|A+B| + \theta|B|)\frac{8\alpha^2}{9s\mu(n-\lambda)^2} + \frac{|B|}{s\mu\theta}\right) w_n \\ + 2B\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle - B\gamma_n^*. \end{aligned}$$

By choosing $\theta = \frac{1}{\sqrt{2s\mu}}$ we can conclude that:

$$\begin{aligned} A\delta_n + B(h_{n-1} - h_n) \leqslant \left(2|A+B| + \frac{\sqrt{2}|B|}{\sqrt{s\mu}}\right)\frac{1}{(n-\lambda)^2} b_n + \left(\left(2|A+B| + \frac{\sqrt{2}|B|}{\sqrt{s\mu}}\right)\frac{4\alpha^2}{9s\mu(n-\lambda)^2} + \frac{\sqrt{2}|B|}{\sqrt{s\mu}}\right) w_n \\ + 2B\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle - B\gamma_n^*, \end{aligned}$$

and hence,

$$\begin{aligned} A\delta_n + B(h_{n-1} - h_n) \leqslant \left(2|A+B| + \frac{\sqrt{2}|B|}{\sqrt{s\mu}}\right)\left(1 + \frac{4\alpha^2}{9s\mu n^2}\right)\frac{E_n}{(n-\lambda)^2} \\ - B\gamma_n^* + 2B\langle x_{n-1} - x_{n-1}^*, x_n^* - x_{n-1}^*\rangle. \end{aligned}$$

□

B.6 Proof of Lemma 10

Let ϕ' denote the derivative of ϕ when it is well defined. According to [35], the function ϕ is differentiable except at a countable set of points. This implies that there exists $(t_i)_{i\in[\![1,N]\!]}$ and $N \in \mathbb{N}^* \cup \{+\infty\}$ such that for any $i \in [\![0, N-1]\!]$ and $t \in (t_i, t_{i+1})$, $\phi'(t)$ is well defined and equal to $\phi_+(t)$. We suppose that the sequence is ordered such that $t_0 < t_i < t_{i+1}$ for any i and that $t_N = +\infty$ when $N \neq +\infty$.
Suppose that $t \in (t_0, t_1)$.

- If ϕ is differentiable at t_0, then ϕ is differentiable on the interval $[t_0, t_1)$ and $\phi' = \phi_+$ in this interval. Consequently inequality (89) ensures that,

$$\phi(t) \leqslant \phi(t_0) + \int_{t_0}^t \psi(u)du.$$

- If ϕ is not differentiable at t_0, then inequality (89) guarantees that for $h > 0$ sufficiently small,

$$\phi(t_0 + h) \leqslant \phi(t_0) + h\psi(t_0).$$

  Then, the previous discussion allows us to say that ϕ is differentiable on $[t_0 + h, t_1)$. As a consequence, we can say that there exists $H \in (0, t - t_0)$ such that for any $h \in (0, H)$:

$$\phi(t) \leqslant \phi(t_0 + h) + \int_{t_0+h}^t \psi(u)du \leqslant \phi(t_0) + \int_{t_0}^t \psi(u)du + \int_{t_0}^{t_0+h} (\psi(t_0) - \psi(u))\, du.$$

  As this inequality is valid for any $h \in (0, H)$, we finally get the wanted inequality (90).

We now suppose that $t = t_1$. We just proved that (90) is true for all $t \in (t_0, t_1)$. Therefore, for all $t \in (t_0, t_1)$,

$$\phi(t) \leqslant \phi(t_0) + \int_{t_0}^{t_1} \psi(u)du,$$

and as ϕ is continuous we get the same inequality at $t = t_1$.
By using the same arguments, we can prove that (90) is valid for any $t > t_1$. Indeed, if $t > t_1$, then it means that $t \in (t_i, t_{i+1})$ or that $t = t_i$ for some $i \in [\![1, N]\!]$. In both cases, we get the wanted inequality by applying the above reasonings to the consecutive intervals (t_j, t_{j+1}) for $0 \leqslant j \leqslant i$.
□

Acknowledgements

This work was supported by PEPR PDE-AI and the ANR Masdol (grant ANR-PRC-CE23). HL acknowledges the financial support of the Ministry of Education, University and Research (grant ML4IP R205T7J2KP).

References

[1] V. Apidopoulos, J.-F. Aujol, C. Dossal, and A. Rondepierre. Convergence rates of an inertial gradient descent algorithm under growth and flatness conditions. *Mathematical Programming*, 187(1):151–193, 2021.

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

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

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

[5] H. Attouch and J. Peypouquet. The rate of convergence of nesterov's accelerated forward-backward method is actually faster than 1/k^2. *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.

[6] J.-F. Aujol, C. Dossal, H. Labarrière, and A. Rondepierre. Heavy ball momentum for non-strongly convex optimization. *arXiv preprint arXiv:2403.06930*, 2024.

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

[8] J.-F. Aujol, C. Dossal, and A. Rondepierre. FISTA is an automatic geometrically optimized algorithm for strongly convex functions. *Mathematical Programming*, 204(1):449–491, 2024.

[9] A. Beck. *First-order methods in optimization*. SIAM, 2017.

[10] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

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

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

[13] J. F. Bonnans, R. Cominetti, and A. Shapiro. Sensitivity analysis of optimization problems under second order regular constraints. *Mathematics of Operations Research*, 23(4):806–831, 1998.

[14] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, 2010.

[15] 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.

[16] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale modeling & simulation*, 4(4):1168–1200, 2005.

[17] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

[18] G. Garrigos, L. Rosasco, and S. Villa. Convergence of the forward-backward algorithm: beyond the worst-case with the help of geometry. *Mathematical Programming*, pages 1–60, 2022.

[19] J.-B. Hiriart-Urruty. At what points is the projection mapping differentiable? *The American Mathematical Monthly*, 89(7):456–458, 1982.

[20] D. Kim and J. A. Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical Programming*, 159(1):81–107, Sep 2016.

[21] D. Kim and J. A. Fessler. Adaptive restart of the optimized gradient method for convex optimization. *Journal of Optimization Theory and Applications*, 178(1):240–263, 2018.

[22] B. Li, B. Shi, and Y.-x. Yuan. Linear convergence of ista and fista. *arXiv preprint arXiv:2212.06319*, 2022.

[23] H. Li and Z. Lin. Accelerated proximal gradient methods for nonconvex programming. *Advances in neural information processing systems*, 28, 2015.

[24] 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.

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

[26] J.-R. Luo and T.-J. Xiao. Optimal convergence rates for damped inertial gradient dynamics with flat geometries. *Applied Mathematics & Optimization*, 87(3):53, Mar 2023.

[27] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

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

[29] Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In *Dokl. Akad. Nauk SSSR,*, volume 269, pages 543–547, 1983.

[30] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[31] O. Sebbouh, C. Dossal, and A. Rondepierre. Convergence rates of damped inertial dynamics under geometric conditions and perturbations. *SIAM Journal on Optimization*, 30(3):1850–1877, 2020.

[32] A. Shapiro. Differentiability properties of metric projections onto convex sets. *Journal of Optimization Theory and Applications*, 169(3):953–964, 2016.

[33] W. Su, S. Boyd, and E. Candes. A differential equation for modeling nesterov's accelerated gradient method: theory and insights. *Advances in neural information processing systems*, 27, 2014.

[34] S. Tao, D. Boley, and S. Zhang. Local linear convergence of ista and fista on the lasso problem. *SIAM Journal on Optimization*, 26(1):313–336, 2016.

[35] G. C. Young. A note on derivates and differential coefficients. *Acta mathematica*, 37(1):141–154, 1914.