FISTA is an automatic geometrically optimized algorithm for strongly convex functions

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

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

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

Abstract: In this work, we are interested in the famous FISTA algorithm. We show that FISTA is an automatic geometrically optimized algorithm for functions satisfying a quadratic growth assumption. This explains why FISTA works better than the standard Forward-Backward algorithm (FB) in such a case, although FISTA is known to have a polynomial asymptotical convergence rate while FB is exponential. We provide a simple rule to tune the α parameter within the FISTA algorithm to reach an ε-solution with an optimal number of iterations. These new results highlight the efficiency of FISTA algorithms, and they rely on new non asymptotic bounds for FISTA.

Key-words: Nesterov acceleration, ODE, first order scheme, optimization.

1 Introduction

Let $F = f + h$ be a composite convex function defined from $\mathbb{R}^N$ to $\mathbb{R}$ whose set X^* of minimizers is not empty, such that ∇f is $L-$Lipschitz, and the proximal operator of h can be easily computed. All along the paper, the function F is assumed to also satisfy a quadratic growth condition i.e it exists $\mu > 0$ such that

$$F(x) - F(x^*) \geqslant \frac{\mu}{2} \|x - x^*\|^2$$

where x^* is the minimizer of F. This condition is actually in force when F is μ-strongly convex, but it is more general as it will be shown later.

Under these hypotheses, the Forward-Backward algorithm (FB) defines a sequence $(x_n)_{n\geqslant 0}$ satisfying $F(x_n) - F(x^*) = \mathcal{O}(e^{-\frac{\mu}{L}n})$, i.e the number n_ε^{FB} of iterations necessary to get an ε-solution satisfies $n_\varepsilon^{FB} = \mathcal{O}(\frac{\mu}{L}\log(\frac{1}{\varepsilon}))$.

Several accelerations of FB have been proposed when F is convex or strongly convex. The FISTA algorithm proposed by Beck and Teboulle [9] using an inertial scheme from Yurii Nesterov [18] was built to improve the convergence rate of FB under a convexity assumption. Nesterov and many others [18, 23, 6, 8, 22, 25] proposed various schemes dedicated to the strongly convex case. All these schemes necessitate an estimation $\tilde{\mu}$ of μ such that $\tilde{\mu} \leqslant \mu$, and they reach an ε-solution with at most $n_\varepsilon = \mathcal{O}\sqrt{\frac{\tilde{\mu}}{L}}\log(\frac{1}{\varepsilon})$ iterations. The constant hidden is the $\mathcal{O}$ depends on the exact algorithm and on the exact hypotheses made on F. It was shown in [20] that the constant in the $\mathcal{O}$ cannot be larger than 2. In [18] a scheme (dubbed NSC later in the paper) is proposed with a constant 1 in the $\mathcal{O}$. Notice that [25] reaches the factor 2 in the $\mathcal{O}$ (and thus the optimality of the convergence rate [20]), but at the cost of the additional hypothesis that the function F is differentiable. In the non differentiable case, [6] proposes a variant of the Heavy-ball scheme with a convergence rate of $\sqrt{2}$ within the $\mathcal{O}$.

In a large dimension setting $\kappa := \frac{\mu}{L} \ll 1$ and the square root gain may be crucial. It explains why inertial algorithms are widely used, and why they behave numerically better than FB, especially when $\tilde{\mu}$ is close to μ.

The main contribution of this paper is to show that the version of FISTA proposed by Chambolle and Dossal [12] and Su, Boyd and Candès [24] can reach an ε-solution with at most $n_\varepsilon^{FISTA} = \mathcal{O}\sqrt{\frac{\mu}{L}}\log(\frac{1}{\varepsilon})$ iterations under a quadratic growth condition which is weaker than a strong-convexity assumption, without any estimation of μ. This result applies to the LASSO problem which may not be strongly convex and for which the estimation of the growth parameter μ can not be tackled easily.

This bound on n_ε^{FISTA} especially explains the better performance of FISTA comparing to FB on problems such like the LASSO problem, on which FB is known to reach an exponential rate and where the previous bounds on FISTA indicate that its rate was only polynomial.

Even though FISTA was built to produce acceleration in a convex setting, it has two main advantages comparing to algorithms built for strongly convex functions.

- There is no need to estimate μ and the convergence rate does not suffer from any underestimation of μ. The bounds on n_ε^{FISTA} depends on μ and not on $\tilde{\mu}$ (an estimation of the true μ).
- The bound on n_ε^{FISTA} applies under a growth condition which is a weaker assumption than strong convexity, and which extends the field of applications of FISTA.

In Section 2, notations and definitions are given, the various algorithms and the notion of $\varepsilon-$solutions are detailed. The main contribution, Theorem 1, is stated and the comparison with the state of the art is done.

Section 3 is devoted to the study of the continuous dynamic associated to FISTA and new finite time bounds are provided for the solution of this dynamic.

In Section 4, the continuous analysis is applied to a non asymptotic study of FISTA which provides finite time bounds allowing to prove Theorem 1.

2 Number of iterations to reach an ε-solution

2.1 Framework and notations

In this paper we focus on the class of composite functions: $F = f + h$ where f is a convex, differentiable function having a L-Lipschitz gradient and h is a proper lower semicontinuous (l.s.c.) convex function whose proximal operator is known. The proximal operator of h is denoted by prox_h and defined by:

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

For this class of functions a classical minimization algorithm is the Forward-Backward algorithm (FB) whose iterations are described by:

$$x_{n+1} = \text{prox}_{sh}(x_n - s\nabla f(x_n)),\ \ s \in \left(0, \frac{2}{L}\right). \tag{2}$$

Beck and Teboulle, based on the ideas of the Nesterov acceleration, propose an accelerated version FISTA (Fast Iterative Shrinkage-Thresholding Algorithm)[13]:

$$y_n = x_n + \frac{n}{n+3}(x_n - x_{n-1}), \quad x_{n+1} = \text{prox}_{sh}(y_n - s\nabla f(y_n)). \tag{3}$$

In this paper we consider the variant of FISTA proposed by Chambolle and Dossal in [12] and denoted by FISTA:

$$y_n = x_n + \frac{n}{n+\alpha}(x_n - x_{n-1}), \quad x_{n+1} = \text{prox}_{sh}(y_n - s\nabla f(y_n)) \tag{4}$$

that ensures in addition the weak convergence of the iterates (when $\alpha > 3$). We assume moreover that the function F satisfies some global quadratic growth property $\mathcal{G}^2_\mu$ namely:

$$\exists\mu > 0,\ \forall x \in \mathbb{R}^N, F(x) - F^* \geqslant \frac{\mu}{2} d(x, X^*)^2 \tag{5}$$

where: $X^* = \arg\min F$ denotes the set of minimizers of F and $F^* = \inf\ F$. Classically the quadratic growth condition $\mathcal{G}^2_\mu$ can be seen as a relaxation of the strong convexity, and it is equivalent in the convex setting to a global Łojasiewicz property with an exponent $\frac{1}{2}$ [15, 16, 10]. Observe also that in the case when f satisfies $\mathcal{G}^2_\mu$, then so does F.

For this class of functions, the Forward-Backward algorithm ensures an exponential decay whereas FISTA classically only ensures a polynomial asymptotic convergence rate. The main contribution of this paper is to provide a non-asymptotic analysis of the FISTA algorithm and to compare the convergence rate in finite time to state-of-the art algorithms like Forward-Backward and the Nesterov accelerated algorithm for strongly convex functions [20]. Analyzing these algorithms in finite time provides a different insight on these convergence rates.

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

$$g(x) = L(x - x^+) := L\left(x - \text{prox}_{\frac{1}{L}h}(x - \frac{1}{L}\nabla f(x))\right),\ x \in \mathbb{R}^N, \tag{6}$$

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

Definition 1 (ε-solution) *Let ε be the expected accuracy. The iterate x_n is said to be an ε-solution of the problem* $\min_{x\in\mathbb{R}^N} F(x)$ *if:*

$$\|g(x_n)\| \leqslant \varepsilon. \tag{7}$$

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:

$$\|g(x_n)\| = \|\nabla F(x_n)\| \leqslant \varepsilon. \tag{8}$$

The notion of ε-solution can be seen as a good stopping criterion for an algorithm solving the composite optimization problem for mainly three reasons: first it is numerically quantifiable. Secondly controlling the norm of the composite gradient mapping is roughly equivalent to having a control on the values of the objective function. Indeed using the following property of the composite gradient mapping proven in [19, Theorem 1] and [7]:

$$\forall x \in \mathbb{R}^N,\ \frac{1}{2L}\|g(x)\|^2 \leqslant F(x) - F(x^+), \tag{9}$$

we can prove that the composite gradient mapping is controlled by the values of the objective function:

$$\forall x \in \mathbb{R}^N,\ \frac{1}{2L}\|g(x)\|^2 \leqslant F(x) - F^*. \tag{10}$$

Conversely, as shown in [7, Lemma 3.1], we also have:

$$\forall x \in \mathbb{R}^N,\ F(x^+) - F^* \leqslant \frac{2}{\mu}\|g(x)\|^2. \tag{11}$$

Thirdly, using (7) as a stopping criterion will enable us to analyze and compare algorithms in terms of the number of iterations needed to reach a given accuracy ε.

2.2 Analysing state-of-the-art optimization algorithms in terms of ε-solution

2.2.1 FB and FISTA without the growth condition $\mathcal{G}^2_\mu$

Let us first recall the Forward-Backward algorithm (FB) described by Algorithm 1:

Algorithm 1 FB: Forward-Backward algorithm to minimize $F = f + h$.

- *Initialization:* $x_0 \in \mathbb{R}^N$, $\varepsilon > 0$.
- *Iterations* $(n \geq 0)$*:* update x_n as follows:

$$x_{n+1} = \text{prox}_{\frac{1}{L}h}(x_n - \frac{1}{L}\nabla f(x_n)) \tag{12}$$

until $\|g(x_n)\| \leq \varepsilon$.

The FB algorithm provides the following bound when F is convex [20, 9]:

$$F(x_n) - F(x^*) \leq \frac{2L\|x_0 - x^*\|^2}{n}. \tag{13}$$

Using (10), this implies that a number of iterations of the order $\mathcal{O}\left(\frac{L^2}{\varepsilon^2}\right)$ is required to get an ε-solution.

A. Beck and M. Teboulle propose in [9] an accelerated version of FB, known as FISTA (Fast Iterative Shrinkage-Thresholding Algorithm). In this paper we focus on the version proposed by Chambolle and Dossal [12] and Su, Boyd and Candes [24] and simply called FISTA from now on.

Algorithm 2 FISTA: Nesterov accelerated algorithm for convex functions $F = f + h$

- *Initialization:* $x_0 \in \mathbb{R}^N$, $x_{-1} = x_0$, $\varepsilon > 0$, $\alpha \geq 3$.
- *Iterations* $(n \geq 0)$*:* update x_n and y_n as follows:

$$\begin{cases} y_n = x_n + \dfrac{n}{n+\alpha}(x_n - x_{n-1}) \\ x_{n+1} = \text{prox}_{\frac{1}{L}h}(y_n - \frac{1}{L}\nabla f(y_n)) \end{cases} \tag{14}$$

until $\|g(x_n))\| \leqslant \varepsilon$.

The FISTA algorithm provides the following bound when F is convex [20, 9]:

$$F(x_n) - F(x^*) \leq \frac{2L\|x_0 - x^*\|^2}{(n+1)^2}. \tag{15}$$

Using (10), this implies that a number of iterations of the order $\mathcal{O}\left(\frac{L}{\varepsilon}\right)$ is required to get an ε-solution.

2.2.2 FB and FISTA with the growth condition $\mathcal{G}^2_\mu$

In the rest of the paper, we assume that F additionally satisfies a quadratic growth condition $\mathcal{G}^2_\mu$ for some real parameter $\mu > 0$ i.e.:

$$\exists \mu > 0,\ \forall x \in \mathbb{R}^N, F(x) - F^* \geqslant \frac{\mu}{2}d(x, X^*)^2 \tag{16}$$

where: $X^* = \arg\min F$ denotes the set of minimizers of F and $F^* = \inf\ F$.

Classically the FB method provides then an ε-solution in $\mathcal{O}\left(\frac{1}{\kappa}\log\left(\frac{1}{\varepsilon}\right)\right)$ iterations [14], which is of course much better than the previous case (without the $\mathcal{G}^2_\mu$ assumption). More precisely:

Theorem 1 *Let $F = f + g$ where f is a convex differentiable function having a L-Lipschitz gradient for some $L > 0$, and g a proper convex l.s.c. function. Assume additionally that F satisfies a quadratic growth condition $\mathcal{G}^2_\mu$ for some real parameter $\mu > 0$.*

Let $\varepsilon > 0$ and:

$$n_\varepsilon^{FB} = \frac{1}{|\log(1-\kappa)|}\log\left(\frac{2LM_0}{\varepsilon^2}\right) \tag{17}$$

where $\kappa = \frac{\mu}{L}$ and $M_0 = F(x_0) - F^*$ denotes the potential energy at initial time. Let $(x_n)_{n\in\mathbb{R}^N}$ be a sequence of iterates generated by the FB algorithm. If $n \geq n_\varepsilon^{FB}$ then the iterate x_n is a ε-solution.

To provide bounds on the number of iterations to get an ε solution, asymptotic bounds on n are not sufficient. The dependencies of these bounds on α and $\kappa = \frac{\mu}{L}$ are also crucial. The main results of the paper (Theorems 3 and 4) are based on new explicit and non asymptotic bounds developed in Part 4. These finite time bounds on FISTA are based on the Lyapunov analysis of the continuous dynamic studied in Part 3.

2.2.3 Algorithms devoted to strongly convex functions

Consider now the Nesterov scheme designed for strongly convex functions [20, Algorithm 2.2.11] which is known to ensure obtaining an ε-solution at most in $\mathcal{O}\left(\frac{1}{\sqrt{\kappa}}\log\left(\frac{1}{\varepsilon}\right)\right)$ iterations [20, Theorem 2.2.3]. More precisely:

Algorithm 3 NSC: Nesterov accelerated algorithm for strongly convex functions

---

- *Initialization:* $x_0 \in \mathbb{R}^N$, $x_{-1} = x_0$, $s \in (0, \frac{1}{L})$.
- *Iterations ($n \geq 0$):* update x_n and y_n as follows:

$$\begin{cases} y_n = x_n + \dfrac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}(x_n - x_{n-1}) \\ x_{n+1} = y_n - \frac{1}{L}\nabla F(y_n) \end{cases} \tag{18}$$

  until $\|g(x_n)\| \leq \varepsilon$.

Theorem 2 *Let F be a convex differentiable function having a L-Lipschitz gradient and admitting a unique minimizer x^*. Assume additionally that F is μ-strongly convex for some real parameter $\mu > 0$. Let $\varepsilon > 0$. Then for $\kappa = \frac{\mu}{L}$ small enough,*

$$\forall n \in \mathbb{N},\ F(x_n) - F(x^*) \leqslant 2(1-\sqrt{\kappa})^n\left(F(x_0) - F(x^*)\right),$$

which means that an ε-solution can be obtained in at most:

$$n_\varepsilon^{NSC} = \frac{1}{|\log(1-\sqrt{\kappa})|}\log\left(\frac{4LM_0}{\varepsilon^2}\right). \tag{19}$$

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

A crucial point to keep in mind with this result, is that iterations (18) depends on μ, and thus, μ must be estimated a priori. In practice, μ may be unknown and only an estimation $\tilde{\mu}$ of μ can be used to define the sequence $(x_n)_{n\geqslant 0}$. To apply the previous theorem we must have $\tilde{\mu}\leqslant\mu$ and the previous bound becomes :

$$n_\varepsilon^{NSC}=\frac{1}{\left|\log(1-\sqrt{\frac{\tilde{\mu}}{L}})\right|}\log\left(\frac{4LM_0}{\varepsilon^2}\right). \tag{20}$$

where $\tilde{\mu}\leqslant\mu$ is the one used to define $\kappa:=\frac{\tilde{\mu}}{L}$ in (18). If $\tilde{\mu}\ll\mu$, this bound may be much higher than the one given in Theorem 3.

The field of accelerated methods for strongly convex functions is a very active one. The fastest one is the one proposed in [25] (which improves NSC with a 2 factor within the exponential decay). The references [22] (with the additional hypothesis of the differentiability of the function F), and [6] (without additional assumption) propose a $\sqrt{2}$ factor improvement with respect to NSC. Note that the case of $F=f+g$ with F convex satisfying a growth condition $\mathcal{G}_\mu^2$ for some real parameter $\mu>0$ has recently been addressed in [23, 6, 8].

2.2.4 FISTA restart

Restarting FISTA is another way to get a linear convergence in the case when F is convex and satisfies the growth condition $\mathcal{G}_\mu^2$ (see e.g. [17] or [21]). In particular, there has been recent works where the growth parameter μ is estimated on the fly by the algorithms, see [1, 7, 13].

We give here the result of [7]: the number of iterations to get an ε-solution is bounded by

$$n_\varepsilon^{Restart}=\frac{7.2}{\sqrt{\kappa}}\left(4.5+\log\left(1+1.3\frac{LM_0}{\varepsilon^2}\right)\right) \tag{21}$$

so that $n_\varepsilon^{Restart}$ is of order

$$\mathcal{O}\left(\frac{1}{\sqrt{\kappa}}\log\left(\frac{1}{\varepsilon}\right)\right). \tag{22}$$

2.3 FISTA is an automatic geometrically optimized algorithm

FISTA with a varying parameter α applied to strongly convex functions, or satisfying a quadratic growth condition, has already been studied and is known to have a polynomial decay. In [24, Theorem 9], the authors proved that if $F=f+h$ and if f is strongly convex then for $\alpha\geqslant\frac{9}{2}$

$$F(x_n)-F(x^*)\leqslant C(\alpha)L\sqrt{\frac{L}{\mu}}\frac{\|x_0-x^*\|^2}{n^3}\leqslant C(\alpha)\left(\frac{\sqrt{L}}{\sqrt{\mu}n}\right)^3(F(x_0)-F(x^*)). \tag{23}$$

In [4] and in [5], Attouch et al and Aujol et al. proved the following asymptotic decay :

$$F(x_n)-F(x^*)=\mathcal{O}\left(\frac{1}{n^{\frac{2\alpha\gamma}{\gamma+2}}}\right) \tag{24}$$

which coincides with the previous bound (23) if $\alpha=\frac{9}{2}$ and $\gamma=1$, when F satisfies some flatness hypothesis $\mathcal{H}_\gamma$:

$$\forall x\in\mathbb{R}^N,\ F(x)-F(x^*)\leqslant\frac{1}{\gamma}\langle\nabla F(x),x-x^*\rangle. \tag{25}$$

To provide bounds on the number of iterations to get an ε solution, asymptotic bounds on n are not sufficient. The dependencies of these bounds on α and $\kappa=\frac{\mu}{L}$ are also crucial. The following Theorem is based on new explicit and non asymptotic bounds developed in Part 4. These finite time bounds on FISTA are based on a Lyapunov analysis of the continuous dynamic studied in Part 3.

Our main result is to prove that under some quadratic growth condition, the number of iterations of FISTA to reach an ε-solution is actually in $\mathcal{O}\left(\frac{1}{\sqrt{\kappa}}\log(\frac{1}{\varepsilon})\right)$:

Theorem 3 *Let $F = f + h$ where f is a convex differentiable function having a L-Lipschitz gradient for some $L > 0$, and h a proper convex l.s.c. function. Assume additionally that F satisfies a quadratic growth condition $\mathcal{G}_\mu^2$ for some real parameter $\mu > 0$.*

Let $\varepsilon > 0$ and

$$\alpha_{1,\varepsilon} := 3\log\left(\frac{5\sqrt{LM_0}}{e\varepsilon}\right) \quad \textit{and} \quad n_{1,\varepsilon}^{FISTA} := \frac{8e^2}{3\sqrt{\kappa}}\alpha_{1,\varepsilon} = \frac{8e^2}{\sqrt{\kappa}}\log\left(\frac{5\sqrt{LM_0}}{e\varepsilon}\right), \tag{26}$$

where $\kappa = \frac{\mu}{L}$ and $M_0 = F(x_0) - F^*$ denotes the mechanical energy at initial time. Let $(x_n)_{n\in\mathbb{R}^N}$ be a sequence of iterates generated by the FISTA algorithm with parameter $\alpha_{1,\varepsilon}$. There exists $\kappa_\varepsilon \in (0,1)$ such that for any $\kappa \le \kappa_\varepsilon$, if $n \ge n_{1,\varepsilon}^{FISTA}$ then the iterate x_n is an ε-solution.

Unlike Algorithm 3, the parameter of FISTA does not depend on μ and it follows that $n_{1,\varepsilon}^{FISTA}$ depends on the *real* value of μ, not on any estimation. To set α, one only needs to define an accuracy ε, the value of L (which is supposed to be known, also to define the step), and the value of M_0. The value of M_0 should be chosen to bound the mechanical energy of the system. In several situations, simple bounds can be found for M_0 for instance when $F(x^*)$ is known (least square problems) or can be estimated. Moreover since M_0 appears in the logarithm, $n_{1,\varepsilon}^{FISTA}$ is not very sensitive to M_0. But we must keep in mind that a bound on M_0 must be given to set α. It is not surprising that α depends on the ratio $\frac{M_0}{\sqrt{\varepsilon}}$ because this ratio measures the decay we want to reach on the value of $F(x_n) - F(x^*)$.

Additionally we assume now that F satisfies some flatness assumption $\mathcal{H}_\gamma$ ensuring that F is not too sharp: let $\gamma \geqslant 1$. For any minimizer x^*, we assume that:

$$\forall x \in \mathbb{R}^N,\ F(x) - F(x^*) \leqslant \frac{1}{\gamma}\langle \nabla F(x), x - x^*\rangle. \tag{27}$$

Any convex differentiable function automatically satisfies $\mathcal{H}_1$. To have a better intuition of the geometry of functions satisfying $\mathcal{H}_\gamma$ for some $\gamma > 0$, observe that the flatness property (27) implies that for any minimizer $x^* \in X^*$, there exists a real constant $M > 0$ such that:

$$\forall x \in \mathbb{R}^N,\ F(x) - F^* \leqslant M\|x - x^*\|^\gamma, \tag{28}$$

see [5, Lemma 2.2]. Thus any convex differentiable function satisfying both $\mathcal{G}_\mu^2$ and $\mathcal{H}_2$ for some $\mu > 0$ can be thus seen as almost quadratic. For this subclass of functions, Theorem 3 can be slightly improved:

Theorem 4 *Let F be a convex differentiable function having a L-Lipschitz gradient and admitting a unique minimizer x^*. Assume that F satisfies a quadratic growth condition $\mathcal{G}_\mu^2$ for some real parameter $\mu > 0$ and a flatness assumption $\mathcal{H}_2$.*

Let $\varepsilon > 0$ and

$$\alpha_{2,\varepsilon} = 2\log\left(\frac{3\sqrt{LM_0}}{e\sqrt{2\varepsilon}}\right) \quad \textit{and} \quad n_{2,\varepsilon}^{FISTA} = \frac{11e^2}{4\sqrt{\kappa}}\alpha_{2,\varepsilon} = \frac{11e^2}{2\sqrt{\kappa}}\log\left(\frac{3\sqrt{LM_0}}{e\sqrt{2\varepsilon}}\right) \tag{29}$$

where $\kappa = \frac{\mu}{L}$ and $M_0 = F(x_0) - F^*$ denotes the mechanical energy at initial time. Let $(x_n)_{n\in\mathbb{R}^N}$ be a sequence of iterates generated by the FISTA algorithm with parameter $\alpha_{2,\varepsilon}$. There exists $\kappa_\varepsilon \in (0,1)$ such that for any $\kappa \le \kappa_\varepsilon$, if $n \ge n_{2,\varepsilon}^{FISTA}$ then the iterate x_n is a ε-solution.

Proof of Theorems 3 and 4 The proof of Theorems 3 and 4 is based on new non-asymptotic bounds for FISTA provided by Theorems 6 and 7 and of the form:

$$\forall n \geqslant \frac{3\alpha}{\sqrt{\kappa}},\ F(x_n) - F^* \leqslant C_\gamma(\alpha)^{\frac{2\alpha\gamma}{\gamma+2}}\left(n\sqrt{\kappa}\right)^{-\frac{2\alpha\gamma}{\gamma+2}} M_0$$

where $M_0 = F(x_0) - F^*$, which means that an ε-solution can be obtained at most in

$$\frac{1}{\sqrt{\kappa}}C_\gamma(\alpha)\left(\frac{2LM_0}{\varepsilon^2}\right)^{\frac{2+\gamma}{2\alpha\gamma}}$$

iterations where:

$$C_1(\alpha) = \frac{2e}{3}(4\alpha - 3)\left(\frac{5}{e\sqrt{2}}\right)^{\frac{3}{\alpha}}, \quad C_2(\alpha) = \frac{e}{4}(11\alpha - 6)\left(\frac{3}{2e}\right)^{\frac{2}{\alpha}}.$$

Thus with these bounds, the optimized numbers $n_{\gamma,\varepsilon}$, $\gamma \in \{1,2\}$ of iterations to reach an ε-solution with FISTA, are respectively obtained for:

$$\alpha_{1,\varepsilon} = 3\log\left(\frac{5\sqrt{LM_0}}{e\varepsilon}\right), \quad \alpha_{2,\varepsilon} = 2\log\left(\frac{3\sqrt{LM_0}}{e\sqrt{2}\varepsilon}\right). \tag{30}$$

For these choices of α, the number of iterations to reach an ε-solution is respectively given by:

$$n_{1,\varepsilon}^{FISTA} = \frac{8e^2}{3\sqrt{\kappa}}\alpha_{1,\varepsilon} = \frac{8e^2}{\sqrt{\kappa}}\log\left(\frac{5\sqrt{LM_0}}{e\varepsilon}\right) \tag{31}$$

$$n_{2,\varepsilon}^{FISTA} = \frac{11e^2}{4\sqrt{\kappa}}\alpha_{2,\varepsilon} = \frac{11e^2}{2\sqrt{\kappa}}\log\left(\frac{3\sqrt{LM_0}}{e\sqrt{2}\varepsilon}\right). \tag{32}$$

■

2.4 Comparisons

2.4.1 Comparison with FB

Let us now compare the Forward-Backward algorithm to the FISTA scheme for a given accuracy $\varepsilon > 0$. For a condition number κ small enough and choosing:

$$\alpha = \alpha_{1,\varepsilon} = 3\log\left(\frac{5\sqrt{LM_0}}{e\varepsilon}\right),$$

we easily check that FISTA requires fewer iterations than the FB algorithm to reach an ε-solution of $\min_{x\in\mathbb{R}^N} F(x)$. Indeed,

$$n_{1,\varepsilon}^{FISTA} = \frac{8e^2}{\sqrt{\kappa}}\log\left(\frac{5\sqrt{LM_0}}{e\varepsilon}\right) \leqslant n_{\varepsilon}^{FB} = \frac{1}{|\log(1-\kappa)|}\log\left(\frac{2LM_0}{\varepsilon^2}\right). \tag{33}$$

Observe that the smaller κ is, the better FISTA is compared to FB.

Note also that since $n_{2,\varepsilon}^{FISTA} \leqslant n_{1,\varepsilon}^{FISTA}$, the comparison between FISTA and FB remains in favor of FISTA for small κ:

$$n_{2,\varepsilon}^{FISTA} = \frac{11e^2}{2\sqrt{\kappa}}\log\left(\frac{3\sqrt{LM_0}}{e\varepsilon\sqrt{2}}\right) \leqslant n_{\varepsilon}^{FB} = \frac{1}{|\log(1-\kappa)|}\log\left(\frac{2LM_0}{\varepsilon^2}\right). \tag{34}$$

Even though FB is known to have an exponential convergence [14], and FISTA a polynomial one [5], choosing the α parameter for FISTA enables to have a much better convergence rate in term of ε-solution then FB.

2.4.2 Comparison with NSC

On this subsection only the comparison between FISTA and the inertial algorithm dedicated to strongly convex functions proposed by Nesterov [20, Algorithm 2.2.11] is detailed but comparisons with any other algorithms built for strongly convex functions described in [20, 23, 6, 8, 25] will lead to the same conclusions.

The Nesterov algorithm for strongly convex functions (Algorithm 3) necessitates an estimation $\widetilde{\mu}$ of the strong convexity parameter μ that is usually not known. If μ is underestimated (i.e. $\widetilde{\mu} \leqslant \mu$), the NSC algorithm will be miscalibrated and therefore slowed down. One of the strengths of FISTA, revealed by our non-asymptotic analysis, is that FISTA is self-adaptative and will have better performances than its NSC variant for strongly convex functions when the strong convexity parameter is not known. Indeed, by choosing the α friction parameter only according to the desired accuracy ε, FISTA will generate a sequence of iterates until reaching an ε-solution without the need of any estimation of the strong convexity parameter μ. Note also that the number of iterations to reach an ε-solution for FISTA actually depends on the true value of μ whether it is known or not.

If the strong parameter μ of F is known, it is clear that for small enough κ, Algorithm 3 is faster than FISTA i.e.

$$n_\varepsilon^{NSC} = \frac{1}{|\log(1-\sqrt{\kappa})|}\log\left(\frac{4LM_0}{\varepsilon^2}\right) \leqslant n_{1,\varepsilon}^{FISTA} = \frac{4e^2}{\sqrt{\kappa}}\log\left(\frac{25LM_0}{e^2\varepsilon^2}\right). \tag{35}$$

Hence the number of iterations needed to reach an $\varepsilon-$solution is smaller for the scheme built for strongly convex functions and using explicitly this parameter μ at each step (18) (since $\kappa = \frac{\mu}{L}$) of the algorithm than the one necessary for FISTA to get the same accuracy. This better behavior was indeed expected.

On the other hand if μ is not perfectly known, which is often the case in large dimension problem, μ should be estimated by $\tilde{\mu}$ and to ensure that the exponential decay of these inertial algorithms are in force, $\tilde{\mu}$ must be chosen such that $\tilde{\mu} \leqslant \mu$ and

$$n_\varepsilon^{NSC} = \frac{1}{\left|\log(1-\sqrt{\frac{\tilde{\mu}}{L}})\right|}\log\left(\frac{4LM_0}{\varepsilon^2}\right) \geqslant \frac{1}{|\log(1-\sqrt{\kappa})|}\log\left(\frac{4LM_0}{\varepsilon^2}\right) \tag{36}$$

If $\tilde{\mu}$ is close to μ, one can expect that $n_\varepsilon^{NSC} \leqslant n_{1,\varepsilon}^{FISTA}$. But if $\tilde{\mu} \leqslant \frac{1}{16e^4}\mu$ then

$$n_\varepsilon^{NSC} \geqslant n_{1,\varepsilon}^{FISTA} \tag{37}$$

and thus if μ is not known with a good accuracy, it may be better to use FISTA.

In practice, FISTA may outperform Algorithm 3 even for much smaller underestimation of μ. The bound given in Theorem 3 may be pessimistic. We illustrate this lower performance of this inertial algorithm with fixed friction term in the numerical experiments comparing to FISTA in the subsection dedicated to numerical experiments. These experiments confirm that even for $\tilde{\mu} = \frac{\mu}{10}$ the loss may be huge and FISTA is actually better for a large set of accuracies ε.

Indeed iterations of FISTA use the value of a parameter α which is defined from ε and not from an estimation of μ which implies that $n_{1,\varepsilon}^{FISTA}$ does not depend on any estimation of μ. The fact that n_ε^{NSC} may be larger than $n_{1,\varepsilon}^{FISTA}$ when μ is not well estimated is a small surprise since FISTA was not built for strongly convex functions. Moreover bounds on $n_{1,\varepsilon}^{FISTA}$ apply under a quadratic growth property and then extend the potential application of this result to a larger set of functions F such as the LASSO :

$$F(x) = \frac{1}{2}\|Ax-y\|^2 + \lambda\|x\|_1 \tag{38}$$

widely used in Statistics and Image and Signal processing. This function F is not strongly convex when $\lambda > 0$ but it satisfies some growth properties, see for example [11].

To be fair in the comparison between both algorithm, we must emphasize that for FISTA, the parameter α depends on the targeted accuracy ε, and FISTA may be better than the other inertial algorithms for this specific accuracy. If n goes to infinity, the algorithms built with a fixed inertia (here $\frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}$) may have a better behavior than FISTA, since for any fixed α, FISTA as a polynomial decay rate, see for example [2, 3].

2.4.3 Comparison with FISTA restart

Under the assumption of quadratic growth it appears that the FISTA restart scheme has a convergence rate similar to FISTA with optimal parameter. In both cases the bounds on the number of iterations needed to get an ε-solution is proportional to $\frac{1}{\sqrt{\kappa}}\log(\frac{\sqrt{ML}}{\varepsilon})$. Preliminary numerical experiments seem to indicate that the best solution may depend on the function F to minimize and both approaches deserve to be tested. In Theorem 1, F is supposed to have a unique minimizer, while one can notice that such an assumption is not needed for FISTA restart, even if in practice, it does not seem to have any impact on the convergence rate. One can also observe that the parameter of FISTA must be chosen according to the wished accuracy ε ; for this accuracy FISTA will be efficient and almost optimal all along the trajectory. The main inconvenient of FISTA restart scheme is its relative complexity where FISTA is really simple to implement.

When ε tends to 0, according to (21):

$$n_\varepsilon^{Restart} \sim \frac{14.4}{\sqrt{\kappa}}\log\left(\frac{\sqrt{LM_0}}{\varepsilon}\right) \tag{39}$$

while

$$n_{1,\varepsilon}^{FISTA} \sim \frac{8e^2}{\sqrt{\kappa}}\log\left(\frac{\sqrt{LM_0}}{\varepsilon}\right) \tag{40}$$

$$n_{2,\varepsilon}^{FISTA} \sim \frac{11e^2}{2\sqrt{\kappa}}\log\left(\frac{\sqrt{LM_0}}{\varepsilon}\right). \tag{41}$$

Hence

$$n_{1,\varepsilon}^{FISTA} \approx \frac{59.1}{\sqrt{\kappa}}\log\left(\frac{\sqrt{LM_0}}{\varepsilon}\right) \tag{42}$$

$$n_{2,\varepsilon}^{FISTA} \approx \frac{40.6}{\sqrt{\kappa}}\log\left(\frac{\sqrt{LM_0}}{\varepsilon}\right). \tag{43}$$

i.e. the bounds for FISTA restart [7] are always slightly better then the one proposed here in the paper. However, we will see in the next subsection that this slight theoretical edge for the worst case analysis may not always prevail in practice. Moreover, it can be argued that FISTA restart needs in general more calls to the function to minimize then classical FISTA (which has a negative impact on its speed), and that it is slightly more complicated to code.

2.4.4 Numerical comparisons

Let us first consider a function $F(x) = \|Ax-b\|^2$ where A is 100×100 gaussian matrix width independent and identically distributed components. On that example $\kappa \approx 4.7\times 10^{-7}$. On Figure 1, are given the values of $\log(\|\nabla F(x_n)\|)$ along the trajectory for various algorithms.

The blue curve corresponds to the Gradient descend, the red curve yo FISTA with $\alpha = 8$, the yellow curve to FISTA with $\alpha = 30$, the green one to the Algorithm 3 of Nesterov for strongly convex functions using the exact value of μ computed from the realisation of the matrix A. The grey curve corresponds to Algorithm 3 with $\tilde{\mu} = \frac{1}{10}\mu$. The graph is displaid in Figure 1. From this graph, several comments can be done :

- For small ε, Algorithm 3 with the precise value of μ (the green curve), seems to be the best one, which was expected.
- If μ is underestimated (grey curve), even of a 10 factor, the decay may be much small smaller with the FISTA schemes.
- For $\varepsilon \leqslant e^{-15}$, the parameter $\alpha = 8$ (red curve) seems to be better than $\alpha = 30$ for FISTA.
- For $\varepsilon \geqslant e^{-15}$, the parameter $\alpha = 30$ (yellow curve) seems to be better than $\alpha = 8$ for FISTA.

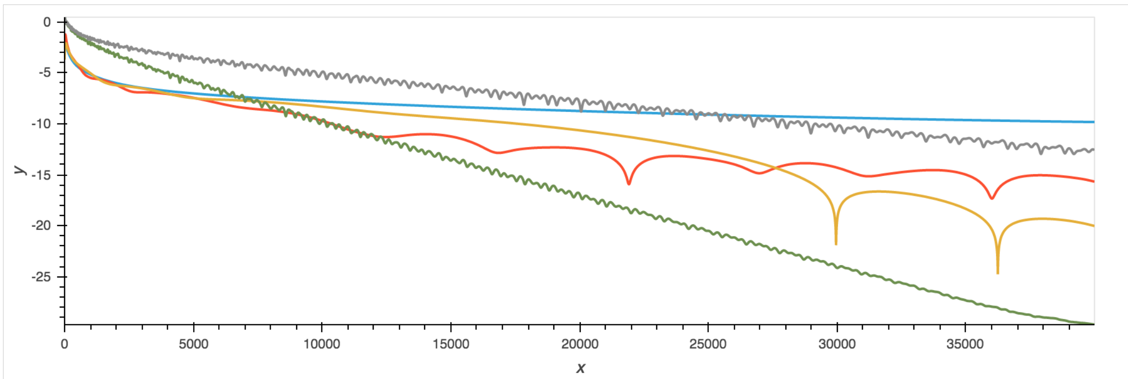

Figure 1: Example on the least square problem. Green is NSC, grey is NSC with a 10 factor underestimate of μ, red is FISTA with $\alpha = 8$ and yellow is FISTA with $\alpha = 30$.

More complete numerical experiments show in practice that the optimal value of α for a given accuracy, that is the value of α ensuring the minimum number iteration to reach an ε-solution, is actually a non increasing function of the accuracy ε, which illustrates Theorem 3.

We then illustrate the paper with a second example, where an inpainting problem is solved using a LASSO formulation.

$$F(x) = \frac{1}{2} \|Mx - Mx^o\|^2 + \lambda \|Tx\|_1 \tag{44}$$

where x^o is a target image, M a random masking operator and T an orthogonal wavelet transform. Figure 2 shows an example, with from the left to right, the original image x^o, the masked image Mx^o and the solution x^* minimizer of F.

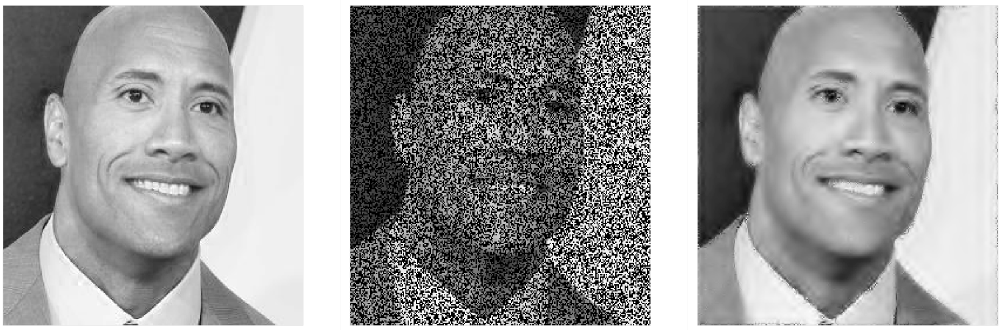

Figure 2: An inpainting example (left: original image, middle: degraded image, right: inpainted result)

Figure 3 displays the curves of the trajectory of $\log(L \|x_n - x_n^+\|)$ for several algorithms. The blue curve corresponds to the Forward Backward algorithm. The red curve corresponds to FISTA-restart scheme described in [7], the yellow curve to FISTA with $\alpha = 3$ (which is the classical FISTA algorithm), the green curve to FISTA with $\alpha = 12$ and the grey curve to FISTA with $\alpha = 30$.

Several comments can be done :

- All inertial schemes are better than FB (blue curve) for any precision ε.
- If we compare the three FISTA algorithms we observe that for large ε, $\alpha = 3$ seems to be better, for small ε, $\alpha = 30$ seems to be the better choice and in between $\alpha = 12$ is better

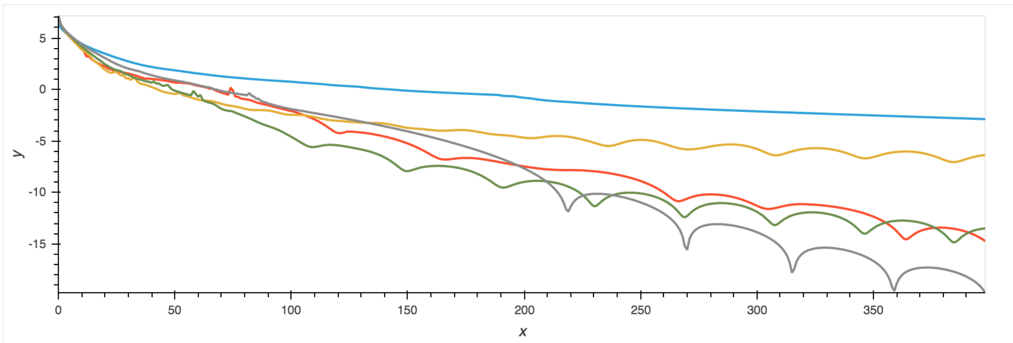

Figure 3: Example on a LASSO problem: FB is in blue, FISTA-restart in red, FISTA with $\alpha = 3$ in yellow, FISTA with $\alpha = 12$ in green, FISTA with $\alpha = 30$ in grey.

than the two others. That is what was expected from Theorem 3 : the optimal value of α is a non increasing function of ε.

- The restart FISTA [7] behaves quite well for any precision and its efficiency seems to be close to FISTA with the best parameter for all accuracy.

3 The continuous case

Let $F : \mathbb{R}^N \to \mathbb{R}$ be a convex differentiable function admitting at least one minimizer x^* and $X^* = \operatorname{argmin} F$. In this section, we study the convergence rates in finite time for the values $F(x(t)) - F(x^*)$ along the trajectories of the well-known ordinary differential equation:

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla F(x(t)) = 0 \tag{45}$$

for any $t \geqslant t_0$ with $t_0 > 0$, associated to the Nesterov scheme. We assume that, for any initial conditions $(x_0, v_0) \in \mathbb{R}^N \times \mathbb{R}^N$, the Cauchy problem associated with the ODE (45) admits a unique global solution satisfying $(x(t_0), \dot{x}(t_0)) = (x_0, v_0)$.

Theorem 5 *Let $F : \mathbb{R}^N \to \mathbb{R}$ be a convex differentiable function admitting a unique minimizer x^*. Assume that F satisfies both a quadratic growth condition $\mathcal{G}^2_\mu$ and a flatness condition $\mathcal{H}_\gamma$:*

$$\forall x \in \mathbb{R}^N,\ F(x) - F(x^*) \leqslant \frac{1}{\gamma}\langle \nabla F(x), x - x^*\rangle \tag{46}$$

for some $\mu > 0$ and $\gamma \geqslant 1$. If $\alpha > 1 + \frac{2}{\gamma}$ and for μ small enough, we have:

$$\forall t \geqslant \frac{\alpha r^*}{(\gamma+2)\sqrt{\mu}} \geqslant t_0,\ F(x(t)) - F(x^*) \leqslant C_1 e^{\frac{2\gamma}{\gamma+2}C_2(\alpha-1-\frac{2}{\gamma})} E_m(t_0) \left(\frac{\alpha r^*}{t(\gamma+2)\sqrt{\mu}}\right)^{\frac{2\alpha\gamma}{\gamma+2}}$$

where $E_m(t) = F(x(t)) - F(x^*) + \frac{1}{2}\|\dot{x}(t)\|^2$ denotes the mechanical energy of the system, r^* is the unique positive real root of the polynomial: $r \mapsto r^3 - r^2 - 2(1+\sqrt{2})r - 4$ and

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

Theorem 5 provides an explicit bound on $F(x(t)) - F(x^*)$ decaying like $t^{-\frac{2\alpha}{3}}$ when $\gamma = 1$ (respectively like $\frac{1}{t^\alpha}$ when $\gamma = 2$). This bounds depends on the growth parameter μ and the friction coefficient α and is valid for sufficiently large enough t namely for

$$t \geqslant t_{\gamma,\alpha,\mu} := \frac{\alpha r^*}{(\gamma+2)\sqrt{\mu}}. \tag{47}$$

Actually this restriction is not really a problem. First because it is possible to reduce $t_{\gamma,\alpha,\mu}$ in the proof, which would lead to not as good asymptotic bounds and secondly because inequality such that $F(x(t)) - F(x^*) \leqslant \frac{K}{t^{2\alpha/3}}$ may not provide interesting bounds for small t. It turns out that for most inertial algorithm, the decay of the $F(x(t)) - F(x^*)$ is not significant if $t \ll \frac{1}{\sqrt{\mu}}$. Indeed, even if finite time bounds are valid from $t = t_0 = 0$, they only provide accurate bounds for $t \geqslant \frac{1}{\sqrt{\mu}}$.

We can also observe that the bound given by Theorem 5 for a given t, is not a decaying function of α which explains why it is not relevant to choose α as large as possible if we consider the ODE on an interval $[t_0, T]$.

Finally observe that the choice of the parameter α can be optimized for a given t by choosing:

$$\alpha_{opt} = t\frac{(\gamma+2)\sqrt{\mu}}{r^*}e^{-1-C_2} \tag{48}$$

which implies a fast exponential decay rate on the values:

$$\begin{aligned} F(x(t)) - F(x^*) &\leqslant C_1 E_m(t_0)\left(e^{-1-C_2}\right)^{\frac{2\gamma\alpha_{opt}}{\gamma+2}} e^{\frac{2\gamma}{\gamma+2}C_2(\alpha_{opt}-1-\frac{2}{\gamma})} \\ &\leqslant C_1 E_m(t_0) e^{-(1+\frac{2}{\gamma})C_2} e^{-\frac{2\gamma}{r^*}e^{-1-C_2}\sqrt{\mu}t}. \end{aligned}$$

This optimal choice depends also on μ that can unknown in practice but we can remark that if $\alpha \approx t\sqrt{\mu}$, FISTA ensures a fast exponential decay which ensures the best possible decay for the bound given in Theorem 5.

Proof of Theorem 5 Our analysis is based on the following Lyapunov energy:

$$\mathcal{E}(t) = t^2(F(x(t) - F(x^*)) + \frac{1}{2}\left\|\lambda(x(t)-x^*) + t\dot{x}(t)\right\|^2, \quad \lambda = \frac{2\alpha}{\gamma+2} \tag{49}$$

where the parameter λ is chosen accordingly to [5]. Remember that the expected asymptotic convergence rate is polynomial in $\mathcal{O}\left(t^{-\frac{2\alpha\gamma}{\gamma+2}}\right)$ [5] with an exponent equal to $\lambda\gamma$. Differentiating the Lyapunov energy $\mathcal{E}$, we easily prove that:

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

Using the flatness assumption and replacing $\lambda = \frac{2\alpha}{\gamma+2}$, we finally get:

$$\mathcal{E}'(t) + \frac{\gamma\lambda-2}{t}\mathcal{E}(t) \leqslant K(\alpha)\left(\frac{2\alpha}{(\gamma+2)t}\|x(t)-x^*\|^2 + \langle x(t)-x^*, \dot{x}(t)\rangle\right) \tag{50}$$

where: $K(\alpha) = \frac{2\alpha\gamma}{(\gamma+2)^2}(\alpha - 1 - \frac{2}{\gamma})$. We now need to control the scalar product whose sign is unknown. Combining the following two inequalities:

$$|\langle x(t)-x^*, \dot{x}(t)\rangle| \leqslant \frac{\sqrt{\mu}}{2}\left\|x(t)-x^*\right\|^2 + \frac{1}{2\sqrt{\mu}}\left\|\dot{x}(t)\right\|^2 \tag{51}$$

where the factor $\sqrt{\mu}$ is chosen to get the tightest control on the energy, and

$$t^2\left\|\dot{x}(t)\right\|^2 \leqslant \left(1+\theta\frac{\alpha}{t\sqrt{\mu}}\right)\left\|\lambda(x(t)-x^*) + t\dot{x}(t)\right\|^2 + \lambda^2\left(1+\frac{t\sqrt{\mu}}{\theta\alpha}\right)\left\|x(t)-x^*\right\|^2 \tag{52}$$

for any $\theta > 0$, we get:

$$\begin{aligned}
\mathcal{E}'(t) + \frac{\gamma\lambda - 2}{t}\mathcal{E}(t) \quad &\leqslant \quad K(\alpha)\left[\frac{\sqrt{\mu}}{2} + \frac{2\alpha}{(\gamma+2)t}\left(1 + \frac{1}{(\gamma+2)\theta}\right) + \frac{2\alpha^2}{(\gamma+2)^2\sqrt{\mu}t^2}\right]\|x(t) - x^*\|^2 \\
&+ \frac{K(\alpha)}{2\sqrt{\mu}t^2}\left(1 + \theta\frac{\alpha}{t\sqrt{\mu}}\right)\|\lambda(x(t) - x^*) + t\dot{x}(t)\|^2 \qquad (53) \\
&\leqslant \quad \frac{2}{\mu}K(\alpha)\left[\frac{\sqrt{\mu}}{2} + \frac{2\alpha}{(\gamma+2)t}\left(1 + \frac{1}{(\gamma+2)\theta}\right) + \frac{2\alpha^2}{(\gamma+2)^2\sqrt{\mu}t^2}\right](F(x(t)) - F(x^*)) \\
&+ \frac{K(\alpha)}{2\sqrt{\mu}t^2}\left(1 + \theta\frac{\alpha}{t\sqrt{\mu}}\right)\|\lambda(x(t) - x^*) + t\dot{x}(t)\|^2 \qquad (54)
\end{aligned}$$

using the growth condition $\mathcal{G}_\mu^2$. We then choose the parameter θ to make equal the coefficients before $\frac{1}{t^3}$ in $t^2(F(x(t)) - F(x^*))$ and $\frac{1}{2}\|\lambda(x(t) - x^*) + t\dot{x}(t)\|^2$, i.e. such that:

$$\frac{2}{\mu}\frac{2\alpha}{(\gamma+2)}\left(1 + \frac{1}{(\gamma+2)\theta}\right) = \frac{\theta\alpha}{\mu} \qquad (55)$$

or equivalently:

$$(\gamma+2)^2\theta^2 - 4(\gamma+2)\theta - 4 = 0. \qquad (56)$$

A straightforward computation shows that this last equation has exactly one positive root:

$$\theta = \frac{2}{\gamma+2}(1 + \sqrt{2}) \qquad (57)$$

For these choice of parameters, we have:

$$\mathcal{E}'(t) + \frac{\gamma\lambda - 2}{t}\mathcal{E}(t) \leqslant \frac{K(\alpha)}{\mu t^2}\left(\sqrt{\mu} + \frac{2\alpha}{(\gamma+2)t}(1 + \sqrt{2}) + \frac{4\alpha^2}{(\gamma+2)^2\sqrt{\mu}t^2}\right)\mathcal{E}(t). \qquad (58)$$

Let us now define:

$$\varphi(t) := \frac{K(\alpha)}{\mu t^2}\left(\sqrt{\mu} + \frac{2\alpha}{(\gamma+2)t}(1 + \sqrt{2}) + \frac{4\alpha^2}{(\gamma+2)^2\sqrt{\mu}t^2}\right) \qquad (59)$$

and: $\Phi(t) = \int_t^{+\infty}\varphi(x)dx$. We so have:

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

Consequently the function $t \mapsto \mathcal{E}(t)t^{\lambda\gamma-2}e^{\Phi(t)}$ is non-increasing, and for any $t_1 \in \mathbb{R}$, we get:

$$\forall t \geqslant t_1,\ \mathcal{E}(t) \leqslant \mathcal{E}(t_1)\left(\frac{t_1}{t}\right)^{\lambda\gamma-2}e^{\Phi(t_1)-\Phi(t)}. \qquad (60)$$

A good choice of t_1 is one ensuring a control as tight as possible on the energy $\mathcal{E}$. For that, t_1 is chosen such that t_1 minimizes the function $u \mapsto u^{\lambda\gamma-2}e^{\Phi(u)}$ i.e. such that t_1 satisfies the equation:

$$\frac{\lambda\gamma - 2}{u} - \varphi(u) = 0 \qquad (61)$$

Noticing that: $\lambda\gamma - 2 = \frac{\gamma+2}{\alpha}K(\alpha)$ and simplifying the equation by $K(\alpha)$, the equation can be rewritten as:

$$\frac{\gamma+2}{\alpha u} = \frac{1}{\mu u^2}\left(\sqrt{\mu} + \frac{2\alpha}{(\gamma+2)u}(1 + \sqrt{2}) + \frac{4\alpha^2}{(\gamma+2)^2\sqrt{\mu}u^2}\right). \qquad (62)$$

Introducing $r = (\gamma+2)\frac{\sqrt{\mu}}{\alpha}u$, we finally have to solve:

$$r^3 - r^2 - 2(1 + \sqrt{2})r - 4 = 0. \qquad (63)$$

A straightforward computation shows that the polynomial $r \mapsto r^3 - r^2 - 2(1+\sqrt{2})r - 4$ has only one real root: $r^* \simeq 3$ (for which Python gives us an analytical value).

Defining $t_1 = \frac{\alpha}{(\gamma+2)\sqrt{\mu}} r^*$, the control on the energy is given by:

$$\forall t \geqslant t_1,\ \mathcal{E}(t) \leqslant \mathcal{E}(\frac{\alpha}{(\gamma+2)\sqrt{\mu}} r^*) \left(\frac{\alpha r^*}{t(\gamma+2)\sqrt{\mu}}\right)^{\gamma\lambda-2} e^{\Phi(t_1)-\Phi(t)}. \tag{64}$$

Observe now that the term $\mathcal{E}(\frac{\alpha}{(\gamma+2)\sqrt{\mu}} r^*)$ can be bounded by the mechanical energy of the system:

$$E_m(t) = F(x(t)) - F^* + \frac{1}{2}\|\dot{x}(t)\|^2 \tag{65}$$

Note that this energy is non-increasing since: $E_m'(t) = \langle \nabla F(x(t)) + \ddot{x}(t), \dot{x}(t)\rangle = -\frac{\alpha}{t}\|\dot{x}(t)\|^2 \leqslant 0$, hence E_m is uniformly bounded on $[t_0, +\infty[$. We then have:

$$\begin{aligned}
\mathcal{E}(t_1) &= t_1^2(F(x(t_1)) - F(x^*)) + \frac{1}{2}\left\|\frac{2\alpha}{\gamma+2}(x(t_1) - x^*) + t_1\dot{x}(t_1))\right\|^2 \\
&= t_1^2(F(x(t_1)) - F(x^*) + \frac{1}{2}\|\dot{x}(t_1)\|^2) + \frac{2\alpha^2}{(\gamma+2)^2}\|x(t_1) - x^*\|^2 + \frac{2\alpha}{\gamma+2} t_1 \langle (x(t_1) - x^*), \dot{x}(t_1)\rangle \\
&\leqslant \frac{\alpha^2}{(\gamma+2)^2\mu}(r^*+2)^2 E_m(t_1) \leqslant (1 + \frac{2}{r^*})^2 t_1^2 E_m(t_0)
\end{aligned} \tag{66}$$

using (51) again to control the scalar product.

Observe that the primitive $\Phi(t) = \int_t^{+\infty} \varphi(x)dx$ of φ has a simple analytic expression showing that Φ is non-positive and:

$$\Phi(t_1) = (\gamma+2)\frac{K(\alpha)}{\alpha}\left(\frac{1}{r^*} + \frac{1+\sqrt{2}}{r^{*2}} + \frac{4}{3r^{*3}}\right) \tag{67}$$

We finally obtain the following control on the values:

$$F(x(t)) - F(x^*) \leqslant C_1 E_m(t_0)\left(\frac{\alpha r^*}{t(\gamma+2)\sqrt{\mu}}\right)^{\frac{2\alpha\gamma}{\gamma+2}} e^{\frac{2\gamma}{\gamma+2}C_2(\alpha-1-\frac{2}{\gamma})} \tag{68}$$

where: $C_1 = \left(1 + \frac{2}{r^*}\right)^2,\ C_2 = \frac{1}{r^*} + \frac{1+\sqrt{2}}{r^{*2}} + \frac{4}{3r^{*3}}$.

4 Discrete non asymptotic analysis of FISTA

Let $F = f + h$ be a convex composite function where f is a convex, differentiable function having a L-Lipschitz gradient and h is a l.s.c. convex function whose proximal operator is known. Let $X^* = \arg\min F$ and $F^* = \inf\ F$.

In this section we provide a complete non-asymptotic analysis of FISTA [12]:

$$y_n = x_n + \frac{n}{n+\alpha}(x_n - x_{n-1}), \quad x_{n+1} = \text{prox}_{sh}(y_n - s\nabla f(y_n)) \tag{69}$$

for this class of convex composite functions F satisfying additionally some global quadratic growth property $\mathcal{G}_\mu^2$:

$$\exists \mu > 0,\ \forall x \in \mathbb{R}^N, F(x) - F^* \geqslant \frac{\mu}{2}d(x, X^*)^2. \tag{70}$$

Our main contribution is this section is to provide non-asymptotic bounds on the values $F(x_n) - F^*$ along the iterates generated by FISTA, see Theorem 6. Our analysis is based on Lyapunov energies that can be seen as discretization of the Lyapunov energy introduced in the continuous setting:

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

We then prove that this bound could be slightly improve when F is nearly quadratic (i.e. satisfying both a quadratic growth condition and a flatness condition $\mathcal{H}_2$).

4.1 Non asymptotic bounds for convex composite functions

Let

$$E_n = 2sn^2(F(x_n) - F(x^*)) + \|\lambda(x_{n-1} - x^*) + n(x_n - x_{n-1})\|^2 \tag{71}$$

be a well-chosen discretization of the Lyapunov energy (49) introduced in the continuous setting. Our main result provides non-asymptotic bounds on the value $F(x_n) - F^*$ along the iterates generated by FISTA:

Theorem 6 *Let $F = f + g$ where f is a convex differentiable function having a L-Lipschitz gradient for some $L > 0$, and g a proper convex l.s.c. function. Assume additionally that F satisfies a quadratic growth condition $\mathcal{G}^2_\mu$ for some real parameter $\mu > 0$. Let $\alpha_0 > \frac{9+3\sqrt{5}}{4}$ and $\kappa = \frac{\mu}{L}$.*

Then there exist $\kappa_0 > 0$ and a real constant $C_3 > 0$ such that for any $0 < \kappa \le \kappa_0$, the sequence $(x_n)_{n\in\mathbb{N}}$ generated by FISTA satisfies:

$$\forall n \geqslant \frac{3\alpha}{\sqrt{\kappa}},\ F(x_n) - F(x^*) \leqslant \frac{25}{9}e^{-2}M_0\left(\frac{e}{3}\sqrt{\frac{2}{\kappa}}(4\alpha-3)(1+C_3\kappa^{1/4})\right)^{\frac{2\alpha}{3}} n^{-\frac{2\alpha}{3}} \tag{72}$$

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

Observe that in finite time (i.e. for a given number of iterations n), a fast exponential decay may be obtained from Theorem 6 by choosing α that minimizes the function:

$$\alpha \mapsto \frac{2\alpha}{3}\log\left(e\,\frac{4\sqrt{2}\alpha}{3n\sqrt{\kappa}}(1+C_3\kappa^{1/4})\right).$$

A straightforward computation shows that the minimum value is reached for:

$$\alpha^* := \frac{3n\sqrt{\kappa}}{4e^2\sqrt{2}\,(1+C_3\kappa^{1/4})}. \tag{73}$$

and with this choice of parameters we finally deduce:

$$F(x_n) - F(x^*) \leqslant \frac{25}{9}e^{-2}\exp\left(-\frac{n\sqrt{\kappa}}{2e^2\sqrt{2}\,(1+C_3\kappa^{1/4})}\right)M_0. \tag{74}$$

The sketch of the proof of Theorem 6 is given in Subsection 4.2, while the proofs of the technical lemmas are detailed in Appendix A.

Assuming now that F additionally satisfies a flatness condition $\mathcal{H}_2$ i.e. for any minimizer x^*:

$$\forall x \in \mathbb{R}^N,\ F(x) - F(x^*) \leqslant \frac{1}{2}\langle \nabla F(x), x - x^*\rangle. \tag{75}$$

enables us to slightly improve the bounds on the values $F(x_n) - F^*$ for composite functions that are nearly quadratic:

Theorem 7 *Let $F : \mathbb{R}^N \to \mathbb{R}$ be a convex differentiable function having a L-Lipschitz gradient and satisfying a quadratic growth condition $\mathcal{G}^2_\mu$ for some real parameters $\mu > 0$ and $L > 0$. Assume that F also satisfies a flatness condition $\mathcal{H}_2$.*

Let $\alpha_0 > 2$ and $\alpha \ge \alpha_0$. Then there exist $\kappa_0 > 0$ and a real constant $C_3 > 0$ such that for any $0 < \kappa \le \kappa_0$, the sequence $(x_n)_{n\in\mathbb{N}}$ generated by FISTA satisfies

$$\forall n \geqslant \frac{3\alpha}{\sqrt{\kappa}},\ F(x_n) - F(x^*) \leqslant \frac{9}{4}\left(e\,\frac{11\alpha-6}{4n\sqrt{2\kappa}}\left(1+C_3\kappa^{1/4}\right)\right)^{\alpha} e^{-2}M_0 \tag{76}$$

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

The proof of Theorem 7 is based on a variant of the Lyapunov energy (71):

$$E_n = 2s\,n^2(F(x_n) - F(x^*) + \left\|\frac{\alpha}{2}(x_{n-1} - x^*) + (n - \frac{\alpha}{4})(x_n - x_{n-1})\right\|^2 \tag{77}$$

and follows the exact same steps than those of the proof of Theorem 6. The sketch of the proof of Theorem 7 is given in Subsection 4.3, while the proofs of the technical lemmas are detailed in Appendix A.2.

4.2 Sketch of the proof of Theorem 6

The proof of Theorem 6 is based on the following Lyapunov energy:

$$E_n = 2sn^2(F(x_n) - F(x^*)) + \|\lambda(x_{n-1} - x^*) + n(x_n - x_{n-1})\|^2 \tag{78}$$

which can be seen as a discretization of the Lyapunov energy introduced in the continuous setting:

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

Stating $w_n = 2s(F(x_n) - F(x^*))$ and:

$$h_n = \|x_n - x^*\|^2,\ \delta_n = \|x_n - x_{n-1}\|,\ \alpha_n = \frac{n}{n+\alpha},\ \lambda = \frac{2\alpha}{3}, \tag{79}$$

the energy E_n can be rewritten as:

$$E_n = n^2 w_n + \left(\lambda^2 - \lambda n\right) h_{n-1} + \left(n^2 - \lambda n\right) \delta_n + \lambda n h_n \tag{80}$$

As in the continuous setting, the first step of the proof consists in establishing some discrete version of the differential inequality (58):

Lemma 1 *Let $\alpha_0 > \frac{9+3\sqrt{5}}{4}$ and $\kappa = \frac{\mu}{L}$. There exists $\kappa_0 > 0$ such that for any $\kappa \leqslant \kappa_0$ and for any $\alpha \geqslant \alpha_0$, there exists some real constants $\tilde{c}_1$ and $\tilde{c}_2$ such that:*

$$\forall n \geqslant \frac{\alpha}{\sqrt{\kappa}},\ E_{n+1} - \left(1 - \frac{\frac{2\alpha}{3} - 2}{n}\right) E_n \leqslant C_1(\alpha,\kappa)\frac{E_n}{n^2} + C_2(\alpha,\kappa)\frac{E_{n+1}}{(n+1)^2} \tag{81}$$

and:

$$C_1(\alpha,\kappa) = \sqrt{\frac{2}{\kappa}}(\frac{4\alpha^2}{9} - 2\alpha + 1)(1+\sqrt{\kappa})^2(1+\tilde{c}_1\sqrt{\kappa}) \tag{82}$$

$$C_2(\alpha,\kappa) = \sqrt{\frac{2}{\kappa}}(\frac{2\alpha}{3} - 1)(1+\sqrt{\kappa})^2(1+\tilde{c}_2\sqrt{\kappa}) \tag{83}$$

for some real constants $\tilde{c}_1$ and $\tilde{c}_2$.

The proof of Lemma 1 is detailed in paragraph A.1.1. The next step consists in integrating the inequality (81):

Lemma 2 *Let $\alpha_0 > \frac{9+3\sqrt{5}}{4}$ and $\alpha \geqslant \alpha_0$. Let $n_0 \geqslant \frac{\alpha}{\sqrt{\kappa}}$. If the energy E_n satisfies (81) then we have:*

$$\forall n \geqslant n_0,\ E_n \leq E_{n_0}\left(\frac{n}{n_0}\right)^{-\left(\frac{2\alpha}{3} - 2\right)} e^{\Phi(n_0)} \tag{84}$$

with

$$\Phi(n_0) = \frac{2}{9n_0}\sqrt{\frac{2}{\kappa}}\,(\alpha - 3)\,(4\alpha - 3)\left(1 + C_3\kappa^{1/4}\right) \tag{85}$$

Remembering that $F(x_n) - F(x^*) \leqslant \frac{1}{2sn^2} E_n$ for any n, we thus have:

$$\forall n \geqslant \frac{\alpha}{\sqrt{\kappa}},\ F(x_n) - F(x^*) \leqslant \frac{E_{n_0}}{2s} \left(n_0^{\frac{2\alpha}{3}-2} e^{\Phi(n_0)} \right) n^{-\frac{2\alpha}{3}}. \tag{86}$$

A good choice of n_0 is one ensuring a control as tight as possible on the values $F(x_n) - F^*$. For that, n_0 is chosen such that it minimizes the function $f : x \mapsto x^{\frac{2\alpha}{3}-2} e^{\Phi(x)}$. A straightforward computation gives:

$$n_0 = \frac{1}{3}\sqrt{\frac{2}{\kappa}}(4\alpha - 3)(1 + C_3\kappa^{1/4}). \tag{87}$$

Observe that $f(n_0) = (e\, n_0)^{\frac{2\alpha}{3}-2}$ and that reducing κ_0 if needed, we get for any $\kappa < \kappa_0$:

$$\frac{\alpha}{\sqrt{\kappa}} \leqslant \frac{1}{3}\sqrt{\frac{2}{\kappa}}(4\alpha - 3)(1 + C_3\kappa^{1/4}) \leqslant 3\frac{\alpha}{\sqrt{\kappa}}. \tag{88}$$

We deduce:

$$\forall n \geqslant \frac{3\alpha}{\sqrt{\kappa}},\ F(x_n) - F(x^*) \leq \frac{E_{n_0}}{2s} n^{-\frac{2\alpha}{3}} (e\, n_0)^{\left(\frac{2\alpha}{3}-2\right)} \tag{89}$$

i.e.:

$$\forall n \geqslant \frac{3\alpha}{\sqrt{\kappa}},\ F(x_n) - F(x^*) \leq \frac{E_{n_0}}{2se^2 n_0^2} \left(\frac{e}{3}\sqrt{\frac{2}{\kappa}}(4\alpha - 3)(1 + C_3\kappa^{1/4}) \right)^{\frac{2\alpha}{3}} n^{-\frac{2\alpha}{3}} \tag{90}$$

Applying Lemma 6 to uniformly bound the energy E_{n_0} and the fact that $n_0 \geqslant \frac{\alpha}{\sqrt{\kappa}}$, we have:

$$\frac{E_{n_0}}{2s{n_0}^2} \quad \leqslant \quad \left(1 + \frac{2\alpha}{3\sqrt{\kappa} n_0}\right)^2 M_{n_0} \leqslant \frac{25}{9} M_{n_0} \tag{91}$$

where M_n is the mechanical energy: $M_n = F(x_n) - F(x^*) + \frac{1}{2}\|x_n - x_{n-1}\|^2$. Since the mechanical energy associated to the Nesterov scheme is non-increasing (see [12, Corollary 2]) and $x_{-1} = x_0$, we then get:

$$\forall n \geqslant n_0,\ F(x_n) - F(x^*) \quad \leqslant \quad \frac{25}{9} e^{-2} M_0 \left(\frac{e}{3}\sqrt{\frac{2}{\kappa}}(4\alpha - 3)(1 + C_3\kappa^{1/4}) \right)^{\frac{2\alpha}{3}} n^{-\frac{2\alpha}{3}}.$$

4.3 Sketch of the proof of Theorem 7

The proof of Theorem 7 follows the same line than the proof of Theorem 6, and is based on the following Lyapunov energy:

$$E_n = 2s\, n^2 (F(x_n) - F(x^*) + \left\| \frac{\alpha}{2}(x_{n-1} - x^*) + (n - \frac{\alpha}{4})(x_n - x_{n-1}) \right\|^2. \tag{92}$$

As in the proof of Theorem 6, the first step of this proof consists in establishing some discrete version of the differential inequality (58):

Lemma 3 *Let $\alpha_0 > 2$ and $\kappa = \frac{\mu}{L}$. There exists $\kappa_0 > 0$ such that for any $\kappa \leqslant \kappa_0$ and for any $\alpha \geqslant \alpha_0$, there exists some real constants $\tilde{c}_1$ and $\tilde{c}_2$ such that:*

$$\forall n \geqslant \frac{\alpha}{\sqrt{\kappa}},\ E_{n+1} - \left(1 - \frac{\alpha - 2}{n}\right) E_n \leqslant C_1(\alpha,\kappa)\frac{E_n}{n^2} + C_2(\alpha,\kappa)\frac{E_{n+1}}{(n+1)^2} \tag{93}$$

where:

$$C_1(\alpha,\kappa) \quad = \quad \frac{(\alpha-2)(5\alpha-2)}{4\sqrt{2\kappa}} \left(1 + \tilde{c}_1\sqrt{\kappa}\right)(1 + \sqrt{\kappa})^2 \tag{94}$$

$$C_2(\alpha,\kappa) \quad = \quad \frac{(\alpha-2)^2}{4\sqrt{2}\sqrt{\kappa}} \left(1 + \tilde{c}_2\sqrt{\kappa}\right)(1 + \sqrt{\kappa})^2. \tag{95}$$

The proof of Lemma 3 is detailed in paragraph A.2.1. The next step consists in integrating the inequality (93).

Lemma 4 *Let $\alpha_0 > 2$ and $\alpha \geqslant \alpha_0$. Let $n_0 \geqslant \frac{\alpha}{\sqrt{\kappa}}$. If E_n satisfies (93) then there exists a real constant $C_3 > 0$ such that:*

$$\forall n \geqslant n_0,\ E_n \leq E_{n_0} \left(\frac{n_0}{n}\right)^{\alpha-2} e^{\Phi(n_0)} \tag{96}$$

with

$$\Phi(n) = \frac{(\alpha-2)(11\alpha-6)}{4n\sqrt{2\kappa}} \left(1 + C_3\kappa^{1/4}\right) \tag{97}$$

A good choice for n_0 is one ensuring a control as tight as possible on the values $F(x_n) - F^*$. For that n_0 is chosen such that it minimizes the function $f : x \mapsto x^{\alpha-2}e^{\Phi(x)}$. A straightforward computation gives:

$$n_0 := \frac{11\alpha-6}{4\sqrt{2\kappa}} \left(1 + C_3\kappa^{1/4}\right). \tag{98}$$

Observe that $f(n_0) = (e\, n_0)^{\alpha-2}$ and that reducing κ_0 if needed, we get for any $\kappa < \kappa_0$:

$$\frac{\alpha}{\sqrt{\kappa}} \leqslant \frac{11\alpha-6}{4\sqrt{2\kappa}} \left(1 + C_3\kappa^{1/4}\right) \leqslant 3\frac{\alpha}{\sqrt{\kappa}}. \tag{99}$$

Hence:

$$\forall n \geqslant \frac{3\alpha}{\sqrt{\kappa}},\ F(x_n) - F(x^*) \leqslant \frac{E_n}{2sn^2} \leqslant \frac{E_{n_0}}{2sn_0^2} \left(\frac{n_0}{n}\right)^{\alpha} e^{\alpha-2} \tag{100}$$

i.e.:

$$\forall n \geqslant \frac{3\alpha}{\sqrt{\kappa}},\ F(x_n) - F(x^*) \quad \leqslant \quad \frac{E_{n_0}}{2se^2n_0^2} \left(e\,\frac{11\alpha-6}{4n\sqrt{2\kappa}} \left(1 + C_3\kappa^{1/4}\right)\right)^{\alpha} \tag{101}$$

Applying Lemma 8 to uniformly bound the energy E_{n_0} and noticing that: $\frac{\alpha}{2n_0\sqrt{\kappa}} \leqslant \frac{1}{2}$, we have:

$$\frac{E_{n_0}}{2sn_0^2} \leqslant \left(1 + \frac{\alpha}{2n_0\sqrt{\kappa}}\right)^2 M_{n_0} \leqslant \frac{9}{4}M_{n_0}$$

where M_n is the mechanical energy: $M_n = F(x_n) - F(x^*) + \frac{1}{2}\|x_n - x_{n-1}\|^2$. Since the mechanical energy associated to the Nesterov scheme is non-increasing (see [12, Corollary 2]) and $x_{-1} = x_0$, we then get:

$$\forall n \geqslant \frac{3\alpha}{\sqrt{\kappa}},\ F(x_n) - F(x^*) \leqslant \frac{9}{4}\left(e\,\frac{11\alpha-6}{4n\sqrt{2\kappa}} \left(1 + C_3\kappa^{1/4}\right)\right)^{\alpha} e^{-2}M_0 \tag{102}$$

Let us now reformulate this very last inequality. Let $\varepsilon > 0$. For κ_0 small enough and for any $0 < \kappa < \kappa_0$, we have:

$$\begin{aligned}\forall n \geqslant \frac{3\alpha}{\sqrt{\kappa}},\ F(x_n) - F(x^*) \quad &\leqslant \quad \frac{9}{4}\left(e\,\frac{11\alpha-6}{4n\sqrt{2\kappa}}(1+\varepsilon)\right)^{\alpha} e^{-2}M_0 \\ &\leqslant \quad \frac{9}{4}\left(e\,\frac{11\alpha}{4n\sqrt{2\kappa}}(1+\varepsilon)\right)^{\alpha} e^{-2}M_0\end{aligned}$$

To find a good value of α with respect to n, we are led to minimize

$$\alpha \mapsto \alpha \log\left(e\,\frac{11\alpha}{4n\sqrt{2\kappa}}(1+\varepsilon)\right).$$

A straightforward computation shows that the minimum value is reached for:

$$\alpha^* := \frac{4n\sqrt{2\kappa}}{11e^2(1+\varepsilon)}. \tag{103}$$

We therefore deduce

$$\forall n \geqslant \frac{3\alpha}{\sqrt{\kappa}},\ F(x_n) - F(x^*) \leqslant \frac{9}{4}e^{-2}\exp\left(-\frac{4n\sqrt{2\kappa}}{11e^2(1+\varepsilon)}\right) M_0. \tag{104}$$

Acknowledgments

J.-F. Aujol acknowledges the support of the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No777826. The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-PRC-CE23 MaSDOL and the support of FMJH Program PGMO 2019-0024 and from the support to this program from EDF-Thales-Orange.

A Technical results of Section 4

A.1 Technical Lemmas for Theorem 6

The proof of Theorem 6 is based on the following Lyapunov energy:

$$E_n = 2sn^2(F(x_n) - F(x^*)) + \|\lambda(x_{n-1} - x^*) + n(x_n - x_{n-1})\|^2 \tag{105}$$

which can be rewritten as:

$$E_n = n^2 w_n + \left(\lambda^2 - \lambda n\right) h_{n-1} + \left(n^2 - \lambda n\right) \delta_n + \lambda n h_n \tag{106}$$

using the reduced notations (107):

$$w_n = 2s(F(x_n) - F(x^*)), \quad h_n = \|x_n - x^*\|^2,\ \delta_n = \|x_n - x_{n-1}\|,\ \alpha_n = \frac{n}{n+\alpha},\ \lambda = \frac{2\alpha}{3}, \tag{107}$$

A.1.1 Proof of Lemma 1.

First step: using the reduced notations (107), we prove that:

$$\begin{aligned} E_{n+1} - \left(1 - \frac{\frac{2\alpha}{3} - 2}{n}\right) E_n \quad &\leqslant \quad \frac{4\alpha K(\alpha)}{3}\frac{h_n}{n} + A_1(n,\alpha)\delta_n + B_1(n,\alpha)(h_{n-1} - h_n) \\ &\quad + B_3(n,\alpha)(h_{n+1} - h_n - \delta_{n+1}) \end{aligned} \tag{108}$$

with:

$$\begin{aligned} A_1(n,\alpha) &= \frac{17\alpha^2}{9} - \frac{8\alpha}{3} + 2 - \alpha\frac{(10\alpha^2 - 18\alpha + 9)n + 7\alpha^3 - 12\alpha^2 + 6\alpha}{3(n+\alpha)^2}, \\ B_1(n,\alpha) &= -\frac{2}{9}\alpha^2 + \frac{4}{3}\alpha - 1 + \frac{1}{3}\frac{3\alpha - 2\alpha^3}{n+\alpha} + \frac{1}{27}\frac{8\alpha^3 - 24\alpha^2}{n}, \\ B_3(n,\alpha) &= \frac{2}{3}\alpha - 1. \end{aligned}$$

Indeed:

$$\begin{aligned} E_{n+1} - \left(1 - \frac{\frac{2\alpha}{3} - 2}{n}\right) E_n \quad &= \quad (n+1)^2 w_{n+1} - \left(1 - \frac{\frac{2\alpha}{3} - 2}{n}\right) n^2 w_n \\ &\quad + \left((n+1)^2 - \lambda(n+1)\right)\delta_{n+1} - \left(1 - \frac{\frac{2\alpha}{3} - 2}{n}\right)\left(n^2 - \lambda n\right)\delta_n \\ &\quad + \left(\lambda^2 - \lambda(n+1) - \lambda n\left(1 - \frac{\frac{2\alpha}{3} - 2}{n}\right)\right) h_n + \lambda(n+1)h_{n+1} \\ &\quad - (\lambda^2 - \lambda n)\left(1 - \frac{\frac{2\alpha}{3} - 2}{n}\right) h_{n-1} \end{aligned} \tag{109}$$

Observe now that:

$$
\begin{aligned}
&(n+1)^2 w_{n+1} - \left(1 - \frac{\frac{2\alpha}{3}-2}{n}\right) n^2 w_n \\
&= \left(1 - \frac{\frac{2\alpha}{3}-2}{n}\right) n^2 (w_{n+1}-w_n) + \left((n+1)^2 - n^2\left(1 - \frac{\frac{2\alpha}{3}-2}{n}\right)\right) w_{n+1} \\
&= n\left(n - (\frac{2\alpha}{3}-2)\right)(w_{n+1}-w_n) + (\frac{2\alpha}{3}n+1) w_{n+1}
\end{aligned}
$$

Combining the two following inequalities

$$
w_{n+1} - w_n \leqslant \alpha_n^2 \delta_n - \delta_{n+1} \tag{110}
$$

from [12] and:

$$
w_{n+1} \leqslant \|x_n + \alpha_n(x_n - x_{n-1}) - x^*\|^2 - \|x_{n+1} - x^*\|^2
$$

from [2], or equivalently with our notations:

$$
w_{n+1} \leqslant (1+\alpha_n)h_n - \alpha_n h_{n-1} - h_{n+1} + (\alpha_n + \alpha_n^2)\delta_n \tag{111}
$$

we then deduce:

$$
\begin{aligned}
&(n+1)^2 w_{n+1} - \left(1 - \frac{\frac{2\alpha}{3}-2}{n}\right) n^2 w_n \\
&\leqslant n\left(n - \frac{2\alpha}{3} + 2\right)(\alpha_n^2\delta_n - \delta_{n+1}) + (\frac{2\alpha}{3}n+1)\left((1+\alpha_n)h_n - \alpha_n h_{n-1} - h_{n+1} + (\alpha_n+\alpha_n^2)\delta_n\right)
\end{aligned}
$$

It follows:

$$
E_{n+1} - \left(1 - \frac{\frac{2\alpha}{3}-2}{n}\right) E_n \leqslant A_1(n,\alpha)\delta_n + A_2(n,\alpha)\delta_{n+1} + B_1(n,\alpha)h_{n-1} + B_2(n,\alpha)h_n + B_3(n,\alpha)h_{n+1} \tag{112}
$$

where:

$$
A_1(n,\alpha) = \frac{17\alpha^2}{9} - \frac{8\alpha}{3} + 2 - \alpha\frac{(10\alpha^2 - 18\alpha + 9)n + 7\alpha^3 - 12\alpha^2 + 6\alpha}{3(n+\alpha)^2}, \quad A_2(n,\alpha) = 1 - \frac{2\alpha}{3}. \tag{113}
$$

and

$$
B_1(n,\alpha) = -\frac{2}{9}\alpha^2 + \frac{4}{3}\alpha - 1 + \frac{1}{3}\frac{3\alpha - 2\alpha^3}{n+\alpha} + \frac{1}{27}\frac{8\alpha^3 - 24\alpha^2}{n}, \quad B_2(n,\alpha) = \frac{2}{9}\alpha^2 - 2\alpha + 2 - \frac{1}{3}\frac{3\alpha - 2\alpha^3}{n+\alpha}, \tag{114}
$$

and

$$
B_3(n,\alpha) = \frac{2}{3}\alpha - 1. \tag{115}
$$

Observe now that: $A_2(n,\alpha) = -B_3(n,\alpha)$ and:

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

so that (112) becomes:

$$
\begin{aligned}
E_{n+1} - \left(1 - \frac{\frac{2\alpha}{3}-2}{n}\right) E_n \quad \leqslant \quad & \frac{4\alpha K(\alpha)}{3}\frac{h_n}{n} + A_1(n,\alpha)\delta_n + B_1(n,\alpha)(h_{n-1} - h_n) \\
& + B_3(n,\alpha)(h_{n+1} - h_n - \delta_{n+1})
\end{aligned} \tag{116}
$$

Step 2: First observe that combining the growth condition $\mathcal{G}_\mu^2$ with the control of the values by the energy (namely: $E_n \geqslant n^2 w_n$ for all n), we have:

$$\forall n \in \mathbb{N}^*,\ \frac{h_n}{n} \leqslant \frac{w_n}{\kappa n} \leqslant \frac{E_n}{\kappa n^3} \leqslant \frac{E_n}{\kappa n(n-\frac{2\alpha}{3})^2},$$

so that applying the following Lemma whose proof is detailed in Appendix A.1.3:

Lemma 5 *for all $n \geqslant 1$ and any $(A,B) \in \mathbb{R}^2$*

$$A\delta_n + B(h_{n-1}-h_n) \leqslant \left(2|A+B| + \frac{\sqrt{2}|B|}{\sqrt{s\mu}} + \frac{8\alpha^2}{9s\mu n^2}\right)\frac{E_n}{(n-\frac{2\alpha}{3})^2}. \tag{117}$$

we can prove that:

$$\frac{4\alpha K(\alpha)}{3}\frac{h_n}{n} + A_1(n,\alpha)\delta_n + B_1(n,\alpha)(h_{n-1}-h_n) \leqslant \frac{\widetilde{C}_1(n,\alpha,\kappa)E_n}{(n-\frac{2\alpha}{3})^2} \tag{118}$$

and:

$$B_3(n,\alpha)(h_{n+1}-h_n-\delta_{n+1}) \leqslant \frac{\widetilde{C}_2(n,\alpha,\kappa)E_{n+1}}{(n+1-\frac{2\alpha}{3})^2} \tag{119}$$

where:

$$\widetilde{C}_1(n,\alpha,\kappa) = 2|\frac{5}{3}\alpha^2 - \frac{4\alpha}{3} + 1 + R(n,\alpha)| + \sqrt{2}\frac{|-\frac{2\alpha^2}{9}+\frac{4\alpha}{3}-1+Q(n,\alpha)|}{\sqrt{\kappa}} + \frac{8\alpha^2}{9\kappa n^2} + \frac{4\alpha K(\alpha)}{3\kappa n} \tag{120}$$

and:

$$\widetilde{C}_2(n,\alpha,\kappa) = \sqrt{2}\frac{\frac{2\alpha}{3}-1}{\sqrt{\kappa}} + \frac{8\alpha^2}{9\kappa(n+1)^2}. \tag{121}$$

with: $|R(\alpha,n)| \leqslant \frac{K\alpha^3}{n}$ and $|Q(\alpha,n)| \leqslant \frac{K\alpha^3}{n}$. Finally observe that for all $n \geqslant \frac{2\alpha}{3}(1+\frac{1}{\sqrt{\kappa}})$, we have:

$$\frac{1}{n-\frac{2\alpha}{3}} \leq \frac{1}{n}\left(1+\sqrt{\kappa}\right) \quad \text{and} \quad \frac{1}{n+1-\frac{2\alpha}{3}} \leq \frac{1}{n+1}\left(1+\sqrt{\kappa}\right) \tag{122}$$

hence:

$$E_{n+1} - \left(1 - \frac{\frac{2\alpha}{3}-2}{n}\right)E_n \leqslant (1+\sqrt{\kappa})^2\left(\widetilde{C}_1(n,\alpha,\kappa)\frac{E_n}{n^2} + \widetilde{C}_2(n,\alpha,\kappa)\frac{E_{n+1}}{(n+1)^2}\right). \tag{123}$$

Step 3: The last step is to uniformly bound the coefficients $\widetilde{C}_1(n,\alpha,\kappa)$ and $\widetilde{C}_2(n,\alpha,\kappa)$ with respect to n. For any $n \geqslant \frac{\alpha}{\sqrt{\kappa}}$ and $\alpha \geqslant 3$, we have:

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

with: $\tilde{c}_2 = \frac{8}{9\sqrt{2}}$. The calculations to bound the coefficient $\widetilde{C}_1(n,\alpha,\kappa)$ are similar but a little more painful. For all $n \geqslant \frac{\alpha}{\sqrt{\kappa}}$, we have:

$$4\alpha\frac{K(\alpha)}{3\kappa n} \leqslant \frac{8\alpha(\alpha-3)}{27\sqrt{\kappa}}$$

so that:

$$
\begin{aligned}
\widetilde{C}_1(n,\alpha,\kappa) &= 2|\frac{5}{3}\alpha^2-\frac{4\alpha}{3}+1+R(n,\alpha)|+\sqrt{2}\frac{|-\frac{2\alpha^2}{9}+\frac{4\alpha}{3}-1+Q(n,\alpha)|}{\sqrt{\kappa}}+\frac{8\alpha^2}{9\kappa n^2}+4\alpha\frac{K(\alpha)}{3\kappa n}\\
&\leqslant \sqrt{\frac{2}{\kappa}}\left(\frac{2\alpha^2}{9}-\frac{4\alpha}{3}+1+|Q(n,\alpha)|+(\frac{5}{3}\alpha^2-\frac{4\alpha}{3}+1+|R(n,\alpha)|)\sqrt{2\kappa}+\frac{4\sqrt{2}\alpha^2}{9n^2\sqrt{\kappa}}+\frac{4\sqrt{2}\alpha(\alpha-3)}{27}\right)\\
&\leqslant \sqrt{\frac{2}{\kappa}}\left(\frac{4\alpha^2}{9}-2\alpha+1+|Q(n,\alpha)|+(\frac{5}{3}\alpha^2-\frac{4\alpha}{3}+1+|R(n,\alpha)|)\sqrt{2\kappa}+\frac{4\sqrt{2}\alpha^2}{9n^2\sqrt{\kappa}}\right)
\end{aligned}
$$

Let $P(\alpha)=\frac{4\alpha^2}{9}-2\alpha+1$. The coefficient $\widetilde{C}_1(n,\alpha,\kappa)$ can be rewritten as:

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

Studying the variations of the function $\alpha\mapsto\frac{\alpha^2}{P(\alpha)}$, we can prove that for any real $\alpha_0>\frac{9+3\sqrt{5}}{4}$, we have for any $\alpha\geqslant\alpha_0$

$$
P(\alpha)\geqslant P(\alpha_0) \quad \text{and} \quad 0<\frac{\alpha^2}{P(\alpha)}\leqslant\frac{\alpha_0^2}{P(\alpha_0)}
$$

so that:

$$
\forall n\geqslant\frac{\alpha}{\sqrt{\kappa}},\ \frac{\alpha^3}{nP(\alpha)}\leqslant\frac{\alpha^2}{P(\alpha)}\sqrt{\kappa}\leqslant\frac{\alpha_0^2}{P(\alpha_0)}\sqrt{\kappa}
$$

and

$$
\forall n\geqslant\frac{\alpha}{\sqrt{\kappa}},\ \sqrt{\frac{2}{\kappa}}\frac{4\alpha^2}{9P(\alpha)n^2}\leqslant\frac{4\sqrt{2}}{9P(\alpha_0)}\sqrt{\kappa} \tag{124}
$$

It finally exists some real constant $\tilde{c}_1$ such that for any $n\geq\frac{\alpha}{\sqrt{\kappa}}$,

$$
\widetilde{C}_1(n,\alpha,\kappa)\leqslant\sqrt{\frac{2}{\kappa}}\left(\frac{4\alpha^2}{9}-2\alpha+1\right)\left(1+\tilde{c}_1\sqrt{\kappa}\right). \tag{125}
$$

Combining (122) and (125), there thus exists $\kappa_0>0$ such that for any $\kappa\leqslant\kappa_0$ and for any $\alpha\geqslant\alpha_0>\frac{9+3\sqrt{5}}{4}$, the inequality (81) holds as expected:

$$
E_{n+1}-\left(1-\frac{\frac{2\alpha}{3}-2}{n}\right)E_n\leqslant\frac{C_1(\alpha,\kappa)E_n}{n^2}+\frac{C_2(\alpha,\kappa)E_{n+1}}{(n+1)^2}
$$

with:

$$
C_1(\alpha,\kappa) = \sqrt{\frac{2}{\kappa}}(\frac{4\alpha^2}{9}-2\alpha+1)(1+\sqrt{\kappa})^2(1+\tilde{c}_1\sqrt{\kappa}) \tag{126}
$$

$$
C_2(\alpha,\kappa) = \sqrt{\frac{2}{\kappa}}(\frac{2\alpha}{3}-1)(1+\sqrt{\kappa})^2(1+\tilde{c}_2\sqrt{\kappa}) \tag{127}
$$

■

A.1.2 Proof of Lemma 2

Assume that the energy E_n satisfies:

$$
E_{n+1}-\left(1-\frac{\frac{2\alpha}{3}-2}{n}\right)E_n\leqslant\frac{C_1(\alpha,\kappa)E_n}{n^2}+\frac{C_2(\alpha,\kappa)E_{n+1}}{(n+1)^2}
$$

i.e.:

$$\left(1-\frac{C_2(\alpha,\kappa)}{(n+1)^2}\right)E_{n+1}-\left(1-\frac{\frac{2\alpha}{3}-2}{n}+\frac{C_1(\alpha,\kappa)}{n^2}\right)E_n\leqslant 0. \tag{128}$$

Let $n_0 \geqslant \frac{\alpha}{\sqrt{\kappa}}$. We then deduce:

$$\forall n\geqslant n_0,\ \log(E_{n+1})-\log(E_{n_0})\leqslant\sum_{k=n_0}^{n}\log\left(\frac{1-\frac{\frac{2\alpha}{3}-2}{k}+\frac{C_1(\alpha,\kappa)}{k^2}}{1-\frac{C_2(\alpha,\kappa)}{(k+1)^2}}\right). \tag{129}$$

Using now the following classical inequalities:

$$\forall x>-1,\ \frac{x}{x+1}\leq\log(1+x)\leq x, \tag{130}$$

we get:

$$\log\left(1-\frac{\frac{2\alpha}{3}-2}{k}+\frac{C_1(\alpha,\kappa)}{k^2}\right)\leq-\frac{\frac{2\alpha}{3}-2}{k}+\frac{C_1(\alpha,\kappa)}{k^2} \tag{131}$$

and

$$-\log\left(1-\frac{C_2(\alpha,\kappa)}{(k+1)^2}\right)\leqslant\frac{C_2(\alpha,\kappa)}{(k+1)^2-C_2(\alpha,\kappa)} \tag{132}$$

We therefore get:

$$\log\left(\frac{1-\frac{\frac{2\alpha}{3}-2}{k}+\frac{C_1(\alpha,\kappa)}{k^2}}{1-\frac{C_2(\alpha,\kappa)}{(k+1)^2}}\right)\leq-\frac{\frac{2\alpha}{3}-2}{k}+\frac{C_1(\alpha,\kappa)}{k^2}+\frac{C_2(\alpha,\kappa)}{(k+1)^2-C_2(\alpha,\kappa)} \tag{133}$$

Hence:

$$\log(E_{n+1})-\log(E_{n_0})\leqslant\sum_{k=n_0}^{n}\left(-\frac{\frac{2\alpha}{3}-2}{k}+\frac{C_1(\alpha,\kappa)}{k^2}+\frac{C_2(\alpha,\kappa)}{(k+1)^2-C_2(\alpha,\kappa)}\right) \tag{134}$$

We are now going to make use of the fact that the functions $x\mapsto\frac{1}{x}$, $x\mapsto\frac{1}{x^2}$ and $x\mapsto\frac{C_2(\alpha,\kappa)}{x^2-C_2(\alpha,\kappa)}$ are decreasing functions on $(C_2,+\infty)$. Observe that all coefficients in the very last inequality are actually non negative since $\alpha\geqslant\alpha_0>3$. We then have:

$$\int_k^{k+1}\frac{dx}{x}\leq\frac{1}{k},\qquad\frac{1}{k^2}\leq\int_{k-1}^{k}\frac{dx}{x^2} \tag{135}$$

and:

$$\frac{C_2(\alpha,\kappa)}{(k+1)^2-C_2(\alpha,\kappa)}\leq\int_k^{k+1}\frac{C_2(\alpha,\kappa)}{x^2-C_2(\alpha,\kappa)}\,dx \tag{136}$$

so that:

$$\log(E_{n+1})-\log(E_{n_0})\leqslant-\left(\frac{2\alpha}{3}-2\right)\int_{n_0}^{n+1}\frac{dx}{x}+C_1(\alpha,\kappa)\int_{n_0-1}^{n}\frac{dx}{x^2}+C_2(\alpha,\kappa)\int_{n_0}^{n+1}\frac{dx}{x^2-C_2(\alpha,\kappa)}$$

Noticing that:

$$\frac{1}{x^2-C_2(\alpha,\kappa)}=\frac{1}{2\sqrt{C_2(\alpha,\kappa)}}\left(\frac{1}{x-\sqrt{C_2(\alpha,\kappa)}}-\frac{1}{x+\sqrt{C_2(\alpha,\kappa)}}\right),$$

we eventually get:

$$\begin{aligned}\log(E_{n+1})-\log(E_{n_0})\ &\leqslant\ -\left(\frac{2\alpha}{3}-2\right)\log\left(\frac{n+1}{n_0}\right)+C_1(\alpha,\kappa)\left(\frac{1}{n_0-1}-\frac{1}{n}\right)\\&\quad+\frac{\sqrt{C_2(\alpha,\kappa)}}{2}\log\left(\frac{(n+1-\sqrt{C_2(\alpha,\kappa)})(n_0+\sqrt{C_2(\alpha,\kappa)})}{(n+1+\sqrt{C_2(\alpha,\kappa)})(n_0-\sqrt{C_2(\alpha,\kappa)})}\right)\end{aligned} \tag{137}$$

i.e.:

$$\log(E_{n+1}) - \log(E_{n_0}) \leqslant -\left(\frac{2\alpha}{3}-2\right)\log\left(\frac{n+1}{n_0}\right) + C_1(\alpha,\kappa)\left(\frac{1}{n_0-1}-\frac{1}{n}\right) + \frac{\sqrt{C_2(\alpha,\kappa)}}{2}\left(\log\left(\frac{n+1-\sqrt{C_2(\alpha,\kappa)}}{n+1+\sqrt{C_2(\alpha,\kappa)}}\right) + \log\left(\frac{n_0+\sqrt{C_2(\alpha,\kappa)}}{n_0-\sqrt{C_2(\alpha,\kappa)}}\right)\right) \tag{138}$$

Taking the exponential, we get:

$$E_{n+1} \leq E_{n_0}\left(\frac{n+1}{n_0}\right)^{-\left(\frac{2\alpha}{3}-2\right)} \exp(\tilde{\Phi}(n_0) - \tilde{\Phi}(n+1)) \tag{139}$$

with:

$$\tilde{\Phi}(n) = \frac{C_1(\alpha,\kappa)}{n-1} + \frac{\sqrt{C_2(\alpha,\kappa)}}{2}\log\left(\frac{n+\sqrt{C_2(\alpha,\kappa)}}{n-\sqrt{C_2(\alpha,\kappa)}}\right).$$

Let us finally compute a more tractable bound on the function $\tilde{\Phi}(n)$: using the inequality $\log(1+x) \leq x$ for $x \leq 1$, we have:

$$0 \leq \log\left(\frac{n+\sqrt{C_2(\alpha,\kappa)}}{n-\sqrt{C_2(\alpha,\kappa)}}\right) = \log\left(1 + \frac{2\sqrt{C_2(\alpha,\kappa)}}{n-\sqrt{C_2(\alpha,\kappa)}}\right) \leqslant \frac{2\sqrt{C_2(\alpha,\kappa)}}{n-\sqrt{C_2(\alpha,\kappa)}} \tag{140}$$

Hence we deduce that:

$$0 \leq \frac{\sqrt{C_2(\alpha,\kappa)}}{2}\log\left(\frac{n+\sqrt{C_2(\alpha,\kappa))}}{n-\sqrt{C_2(\alpha,\kappa)}}\right) \leqslant \frac{C_2(\alpha,\kappa)}{n-\sqrt{C_2(\alpha,\kappa)}} \tag{141}$$

Now, using the definition of the coefficients $C_1(\alpha,\kappa)$ and $C_2(\alpha,\kappa)$ given in Lemma 1, we get:

$$\begin{aligned} 0 \leq \tilde{\Phi}(n) &\leq \frac{C_1(\alpha,\kappa)}{n-1} + \frac{C_2(\alpha,\kappa)}{n-\sqrt{C_2(\alpha,\kappa)}} \leqslant \frac{2C_1(\alpha,\kappa)}{n} + \frac{C_2(\alpha,\kappa)}{n-\sqrt{C_2(\alpha,\kappa)}} \\ &\leqslant \frac{1}{n}\sqrt{\frac{2}{\kappa}}(1+\sqrt{\kappa})^2\left[2\left(\frac{4\alpha^2}{9}-2\alpha+1\right)(1+\tilde{c}_1\sqrt{\kappa}) + \left(\frac{2\alpha}{3}-1\right)\frac{1+\tilde{c}_2\sqrt{\kappa}}{1-\frac{\sqrt{C_2(\alpha,\kappa)}}{n}}\right] \end{aligned} \tag{142}$$

Observe then that for any $n \geqslant \frac{\alpha}{\sqrt{\kappa}}$,

$$\frac{1}{1-\frac{\sqrt{C_2(\alpha,\kappa)}}{n}} \leqslant \frac{1}{1-\frac{\sqrt{C_2(\alpha,\kappa)}}{\alpha}\sqrt{\kappa}} \leqslant 1 + 2\frac{\sqrt{C_2(\alpha,\kappa)}}{\alpha}\sqrt{\kappa}$$

so that there exists a real constant $\tilde{c}_3$ such that for κ small enough and any $\alpha \geqslant \alpha_0$ we have:

$$\frac{1}{1-\frac{\sqrt{C_2(\alpha,\kappa)}}{n}} \leqslant 1 + \tilde{c}_3\kappa^{1/4}.$$

Therefore we finally get for any $n \geqslant \frac{\alpha}{\sqrt{\kappa}}$:

$$\tilde{\Phi}(n) \leq \frac{1}{n}\sqrt{\frac{2}{\kappa}}\left(1+\sqrt{\kappa}\right)^2\left(\left(\frac{8\alpha^2}{9}-4\alpha+2\right)\left(1+\tilde{c}_1\sqrt{\kappa}\right) + \left(\frac{2\alpha}{3}-1\right)\left(1+\tilde{c}_2\sqrt{\kappa}\right)\left(1+\tilde{c}_3\kappa^{1/4}\right)\right) \tag{143}$$

Since $\frac{8\alpha^2}{9} - 4\alpha + 2 + \frac{2\alpha}{3} - 1 = \frac{2}{9}(\alpha-3)(4\alpha-3)$, we then deduce that there exists $C_3 > 0$ (independent to α) such that

$$\forall n \geqslant \frac{\alpha}{\sqrt{\kappa}},\ \tilde{\Phi}(n) \leqslant \frac{2}{9n}\sqrt{\frac{2}{\kappa}}(\alpha-3)(4\alpha-3)\left(1+C_3\kappa^{1/4}\right) \tag{144}$$

Let us introduce:

$$\Phi(n)=\frac{2}{9n}\sqrt{\frac{2}{\kappa}}\,(\alpha-3)\,(4\alpha-3)\left(1+C_3\kappa^{1/4}\right).$$

As expected we finally have:

$$\forall n\geqslant n_0,\ E_{n+1}\leq E_{n_0}\left(\frac{n+1}{n_0}\right)^{-\left(\frac{2\alpha}{3}-2\right)}e^{\Phi(n_0)} \tag{145}$$

■

A.1.3 Technical lemma

Lemma 6 *Let M_n the mechanical energy, that is:*

$$M_n=F(x_n)-F(x^*)+\frac{1}{2s}\|x_n-x_{n-1}\|^2 \tag{146}$$

Then we have

$$\frac{E_n}{2sn^2}\quad\leqslant\quad\left(1+\frac{4\alpha^2}{9\kappa n^2}+\frac{4\alpha}{3\sqrt{\kappa}n}\right)M_n=\left(1+\frac{2\alpha}{3\sqrt{\kappa}n}\right)^2M_n \tag{147}$$

Proof: Let us first remark that:

$$\begin{aligned}
b_n &= \left\|\frac{2\alpha}{3}(x_{n-1}-x^*)+n(x_n-x_{n-1})\right\|^2=\left\|\frac{2\alpha}{3}(x_n-x^*)+\left(n-\frac{2\alpha}{3}\right)(x_n-x_{n-1})\right\|^2\\
&= \frac{4\alpha^2}{9}\|x_n-x^*\|^2+(n-\frac{2\alpha}{3})^2\|x_n-x_{n-1}\|^2+\frac{4\alpha}{3}(n-\frac{2\alpha}{3})\langle x_n-x^*,x_n-x_{n-1}\rangle\\
&\leqslant \frac{4\alpha^2}{9}\|x_n-x^*\|^2+n^2\|x_n-x_{n-1}\|^2+\frac{4\alpha}{3}(n-\frac{2\alpha}{3})\langle x_n-x^*,x_n-x_{n-1}\rangle
\end{aligned}$$

Using a discrete version of the inequality (51), we have:

$$|\langle x_n-x^*,x_n-x_{n-1}|\rangle\leqslant\frac{\sqrt{\kappa}}{2}\|x_n-x^*\|^2+\frac{1}{2\sqrt{\kappa}}\|x_n-x_{n-1}\|^2 \tag{148}$$

so that:

$$b_n\leq\frac{4\alpha^2}{9}\|x_n-x^*\|^2+n^2\|x_n-x_{n-1}\|^2+\frac{2\alpha n}{3}\left(\sqrt{\kappa}\|x_n-x^*\|^2+\frac{1}{\sqrt{\kappa}}\|x_n-x_{n-1}\|^2\right) \tag{149}$$

Hence:

$$\begin{aligned}
\frac{E_n}{2sn^2} &= F(x_n)-F^*+\frac{1}{2sn^2}b_n\\
&= M_n+\frac{2\alpha^2}{9sn^2}\|x_n-x^*\|^2+\frac{\alpha}{3sn}\left(\sqrt{\kappa}\|x_n-x^*\|^2+\frac{1}{\sqrt{\kappa}}\|x_n-x_{n-1}\|^2\right)
\end{aligned}$$

Using now the quadratic growth condition $\mathcal{G}^2_\mu$ and remembering that: $s\mu=\kappa$, we get:

$$\frac{E_n}{2sn^2}\quad\leqslant\quad\left(1+\frac{4\alpha^2}{9\kappa n^2}+\frac{4\alpha}{3\sqrt{\kappa}n}\right)M_n=\left(1+\frac{2\alpha}{3\sqrt{\kappa}n}\right)^2M_n$$

■

Proof of Lemma 5 Let us prove that for all $n \geqslant 1$ and any $(A,B) \in \mathbb{R}^2$

$$A\delta_n + B(h_{n-1} - h_n) \leqslant \left(2|A+B| + \frac{\sqrt{2}|B|}{\sqrt{s\mu}} + \frac{8\alpha^2}{9s\mu n^2}\right)\frac{1}{(n-\frac{2\alpha}{3})^2}E_n. \tag{150}$$

Firstly notice that

$$A\delta_n + B(h_{n-1} - h_n) = (A+B)\delta_n + B(h_{n-1} - h_n - \delta_n) \tag{151}$$

and for any $\theta > 0$

$$|h_{n-1} - h_n - \delta_n| = 2|\langle x_n - x_{n-1}, x_n - x^*\rangle| \leqslant \frac{h_n}{\theta} + \theta\delta_n. \tag{152}$$

Combining the last two inequalities, it follows that for any $\theta > 0$:

$$A\delta_n + B(h_{n-1} - h_n) \leqslant (A + B + \theta|B|)\delta_n + \frac{|B|}{\theta}h_n \tag{153}$$

To bound the coefficient of δ_n we use a specific expression of b_n:

$$b_n = \left\|\frac{2\alpha}{3}(x_n - x^*) + (n - \frac{2\alpha}{3})(x_n - x_{n-1})\right\|^2 \tag{154}$$

Applying the inequality $\|u\|^2 \leqslant 2\|u+v\|^2 + 2\|v\|^2$ to $u = (n - \frac{2\alpha}{3})(x_n - x_{n-1})$ and $v = \frac{2\alpha}{3}(x_n - x^*)$, we get:

$$\|u\|^2 \leqslant 2\|u+v\|^2 + 2\|v\|^2 \tag{155}$$

can be written

$$(n - \frac{2\alpha}{3})^2\delta_n \leqslant 2b_n + \frac{8\alpha^2}{9}h_n. \tag{156}$$

It follows that

$$\delta_n \leqslant \frac{2}{(n-\frac{2\alpha}{3})^2}b_n + \frac{8\alpha^2}{9(n-\frac{2\alpha}{9})^2}h_n. \tag{157}$$

and thus

$$A\delta_n + B(h_{n-1} - h_n) \leqslant (|A+B| + \theta|B|)\frac{2}{(n-\frac{2\alpha}{3})^2}b_n + \left(\frac{|B|}{\theta} + \frac{8\alpha^2}{9(n-\frac{3\alpha}{4})^2}\right)h_n \tag{158}$$

Using now the growth condition $h_n \leqslant \frac{1}{s\mu}w_n$ for all $n \in \mathbb{N}$, we get:

$$A\delta_n + B(h_{n-1} - h_n) \leqslant (|A+B| + \theta|B|)\frac{2}{(n-\frac{2\alpha}{3})^2}b_n + \left(\frac{|B|}{s\mu\theta} + \frac{8\alpha^2}{9s\mu(n-\frac{2\alpha}{3})^2}\right)w_n \tag{159}$$

Choosing $\theta = \frac{1}{\sqrt{2s\mu}}$ we finally deduce:

$$A\delta_n + B(h_{n-1} - h_n) \leqslant (2|A+B| + \frac{\sqrt{2}|B|}{\sqrt{s\mu}})\frac{1}{(n-\frac{2\alpha}{3})^2}b_n + \left(\frac{\sqrt{2}|B|}{\sqrt{s\mu}} + \frac{8\alpha^2}{9s\mu(n-\frac{2\alpha}{3})^2}\right)w_n \tag{160}$$

and

$$A\delta_n + B(h_{n-1} - h_n) \leqslant \left(2|A+B| + \frac{\sqrt{2}|B|}{\sqrt{s\mu}} + \frac{8\alpha^2}{9s\mu n^2}\right)\frac{1}{(n-\frac{2\alpha}{3})^2}E_n. \tag{161}$$

which concludes the proof of the lemma.

■

A.2 Technical lemmas for Theorem 7

The proof of this theorem is based on the analysis of the following Lyapunov sequence:

$$E_n = 2s\,n^2(F(x_n) - F(x^*) + \left\|\frac{\alpha}{2}(x_{n-1} - x^*) + (n - \frac{\alpha}{4})(x_n - x_{n-1})\right\|^2. \tag{162}$$

Using the reduced notations $w_n = 2s(F(x_n) - F(x^*))$ and:

$$h_n = \|x_n - x^*\|^2,\ \delta_n = \|x_n - x_{n-1}\|,\ \alpha_n = \frac{n}{n+\alpha},\ \lambda = \frac{\alpha}{2}, \tag{163}$$

the energy E_n can be rewritten as:

$$E_n = n^2 w_n + \left(\lambda^2 - \lambda n + \frac{\alpha}{4}\lambda\right) h_{n-1} + \left[\left(n - \frac{\alpha}{4}\right)^2 - \lambda(n - \frac{\alpha}{4})\right]\delta_n + \lambda(n - \frac{\alpha}{4})h_n \tag{164}$$

which can be seen as a discretization of the Lyapunov energy in the continuous setting.

A.2.1 Proof of Lemma 3

First step: using the reduced notations (163), we prove that:

$$\begin{aligned} E_{n+1} - \left(1 - \frac{\alpha-2}{n}\right)E_n \ &\leqslant\ \alpha K(\alpha)\frac{h_n}{n} + A_1(n,\alpha)\delta_n + B_1(n,\alpha)(h_{n-1} - h_n) \\ &\quad + A_2(n,\alpha)\delta_{n+1} + B_3(n,\alpha)(h_{n+1} - h_n) \end{aligned} \tag{165}$$

with:

$$\begin{aligned} A_1(n,\alpha) &= 1 - 2\alpha + \frac{37}{16}\alpha^2 - \frac{\alpha}{16n(n+\alpha)^2}\left[n^2\left(77\alpha^2 - 90\alpha + 24\right) + 2n\alpha\left(25\alpha^2 - 26\alpha + 8\right) - 3\alpha^3\left(\alpha - 2\right)\right] \\ A_2(n,\alpha) &= 1 - \alpha + \frac{\alpha^2}{16} \\ B_1(n,\alpha) &= -\frac{1}{2}\left(1 - 2\alpha + \frac{3}{4}\alpha^2\right) - \frac{\alpha}{8n(n+\alpha)}\left((\alpha^2 + 6\alpha - 4)n - 3\alpha^2(\alpha - 2)\right) \\ B_3(n,\alpha) &= -\frac{1}{8}(\alpha - 2)^2 \end{aligned}$$

Indeed:

$$\begin{aligned} E_{n+1} - \left(1 - \frac{\alpha-2}{n}\right)E_n \ &=\ (n+1)^2 w_{n+1} - \left(1 - \frac{\alpha-2}{n}\right)n^2 w_n \\ &\quad + \left(\lambda^2 - \lambda(n+1) + \frac{\alpha}{4}\lambda - \lambda\left(1 - \frac{\alpha-2}{n}\right)(n - \frac{\alpha}{4})\right)h_n \\ &\quad + \left((n + 1 - \frac{\alpha}{4})^2 - \lambda(n + 1 - \frac{\alpha}{4})\right)\delta_{n+1} \\ &\quad - \left(1 - \frac{\alpha-2}{n}\right)\left((n - \frac{\alpha}{4})^2 - \lambda(n - \frac{\alpha}{4})\right)\delta_n + \lambda(n + 1 - \frac{\alpha}{4})h_{n+1} \\ &\quad - (\lambda^2 - \lambda n + \frac{\alpha}{4}\lambda)\left(1 - \frac{\alpha-2}{n}\right)h_{n-1} \end{aligned} \tag{166}$$

Observe now that, combining the two following inequalities

$$w_{n+1} - w_n \leqslant \alpha_n^2\delta_n - \delta_{n+1} \tag{167}$$

from [12] and:

$$w_{n+1} \leqslant \frac{1}{2}\left(\|x_n + \alpha_n(x_n - x_{n-1}) - x^*\|^2 - \|x_{n+1} - x^*\|^2\right)$$

from [2], or equivalently with our notations:

$$w_{n+1} \leqslant \frac{1}{2}\left((1+\alpha_n)h_n - \alpha_n h_{n-1} - h_{n+1} + (\alpha_n + \alpha_n^2)\delta_n\right) \tag{168}$$

we deduce:

$$\begin{aligned}(n+1)^2 w_{n+1} - \left(1 - \frac{\alpha-2}{n}\right) n^2 w_n &= n\left(n - (\alpha-2)\right)(w_{n+1} - w_n) + (\alpha n + 1) w_{n+1} \\ &\leqslant n\left(n - \alpha + 2\right)(\alpha_n^2 \delta_n - \delta_{n+1}) \\ &\quad + \frac{1}{2}(\alpha n + 1)\left((1+\alpha_n)h_n - \alpha_n h_{n-1} - h_{n+1} + (\alpha_n + \alpha_n^2)\delta_n\right)\end{aligned}$$

Noticing that:

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

we get:

$$\begin{aligned}E_{n+1} - \left(1 - \frac{\alpha-2}{n}\right) E_n &\leqslant \alpha K(\alpha)\frac{h_n}{n} + A_1(n,\alpha)\delta_n + B_1(n,\alpha)(h_{n-1} - h_n) \\ &\quad + A_2(n,\alpha)\delta_{n+1} + B_3(n,\alpha)(h_{n+1} - h_n)\end{aligned} \tag{169}$$

as expected.

Step 2: combining the growth condition $\mathcal{G}_\mu^2$ with the control of the values by the energy (namely: $E_n \geqslant n^2 w_n$ for all n), we have:

$$\forall n \in \mathbb{N}^*, \ \frac{h_n}{n} \leqslant \frac{w_n}{\kappa n} \leqslant \frac{E_n}{\kappa n^3} \leqslant \frac{E_n}{\kappa n (n - \frac{3\alpha}{4})^2},$$

so that applying the following Lemma whose proof is detailed in Appendix A.2.3:

Lemma 7 *For any $n \geqslant 1$ and $(A,B) \in \mathbb{R}^2$, we have:*

$$A\delta_n + B(h_{n-1} - h_n) \leqslant \left(2|A+B| + \frac{\sqrt{2}|B|}{\sqrt{s\mu}} + \frac{\alpha^2}{4s\mu n^2}\right)\frac{1}{(n - \frac{3\alpha}{4})^2} E_n. \tag{170}$$

we can prove that:

$$\alpha K(\alpha)\frac{h_n}{n} + A_1(n,\alpha)\delta_n + B_1(n,\alpha)(h_{n-1} - h_n) \leqslant \frac{\widetilde{C}_1(n,\alpha,\kappa)E_n}{(n - \frac{3\alpha}{4})^2} \tag{171}$$

and:

$$A_2(n,\alpha)\delta_{n+1} + B_3(n,\alpha)(h_{n+1} - h_n) \leqslant \frac{\widetilde{C}_2(n,\alpha,\kappa)E_{n+1}}{(n + 1 - \frac{3\alpha}{4})^2} \tag{172}$$

where:

$$\widetilde{C}_1(n,\alpha,\kappa) = |\frac{31}{8}\alpha^2 - 2\alpha + 1 + R(n,\alpha)| + \frac{|\frac{3\alpha^2}{4} - 2\alpha + 1 + Q(n,\alpha)|}{\sqrt{2\kappa}} + \frac{\alpha^2}{4\kappa n^2} + \frac{\alpha K(\alpha)}{\kappa n} \tag{173}$$

and:

$$\widetilde{C}_2(n,\alpha,\kappa) = \left|\frac{\alpha^2}{8} + \alpha - 1\right| + \sqrt{2}\frac{(\alpha-2)^2}{8\sqrt{\kappa}} + \frac{\alpha^2}{4\kappa(n+1)^2}. \tag{174}$$

with: $|R(\alpha,n)| \leqslant \frac{K\alpha^3}{n}$ and $|Q(\alpha,n)| \leqslant \frac{K\alpha^3}{n}$. Finally observe that for all $n \geqslant \frac{3\alpha}{4}(1 + \frac{1}{\sqrt{\kappa}})$, we have:

$$\frac{1}{n - \frac{3\alpha}{4}} \leq \frac{1}{n}\left(1 + \sqrt{\kappa}\right) \quad \text{and} \quad \frac{1}{n + 1 - \frac{3\alpha}{4}} \leq \frac{1}{n+1}\left(1 + \sqrt{\kappa}\right) \tag{175}$$

hence:

$$E_{n+1} - \left(1 - \frac{\alpha-2}{n}\right) E_n \leqslant (1 + \sqrt{\kappa})^2 \left(\widetilde{C}_1(n,\alpha,\kappa)\frac{E_n}{n^2} + \widetilde{C}_2(n,\alpha,\kappa)\frac{E_{n+1}}{(n+1)^2}\right). \tag{176}$$

Step 3: The last step is to uniformly bound the coefficients $\widetilde{C}_1(n,\alpha,\kappa)$ and $\widetilde{C}_2(n,\alpha,\kappa)$ with respect to n. Let $\alpha_0 > 2$. For any $n \geqslant \frac{\alpha}{\sqrt{\kappa}}$ and $\alpha \geqslant \alpha_0$, we have:

$$\begin{aligned}\widetilde{C}_2(n,\alpha,\kappa) &= \frac{\alpha^2}{8}+\alpha-1+\sqrt{2}\frac{(\alpha-2)^2}{8\sqrt{\kappa}}+\frac{\alpha^2}{4\kappa(n+1)^2} \leqslant \frac{\alpha^2}{8}+\alpha-\frac{3}{4}+\sqrt{2}\frac{(\alpha-2)^2}{8\sqrt{\kappa}}\\ &\leqslant \frac{(\alpha-2)^2}{4\sqrt{2}\sqrt{\kappa}}\left(1+\frac{\alpha^2+8\alpha-6}{\sqrt{2}(\alpha-2)^2}\sqrt{\kappa}\right) \leqslant \frac{(\alpha-2)^2}{4\sqrt{2}\sqrt{\kappa}}\left(1+\widetilde{c}_2\sqrt{\kappa}\right)\end{aligned}$$

with $\widetilde{c}_2 = \frac{\alpha_0^2+8\alpha_0-6}{\sqrt{2}(\alpha_0-2)^2}$. The calculations to bound the coefficient $\widetilde{C}_1(n,\alpha,\kappa)$ are similar but a little more painful. Noticing that

$$\forall n \geqslant \frac{\alpha}{\sqrt{\kappa}},\ \frac{\alpha K(\alpha)}{\kappa n} \leqslant \frac{\alpha(\alpha-2)}{4\sqrt{\kappa}} \leqslant \frac{\alpha(\alpha-2)}{2\sqrt{2\kappa}}, \tag{177}$$

and that: $\frac{3\alpha^2}{4}-2\alpha+1>0$ for any $\alpha>2$, the coefficient $\widetilde{C}_1(n,\alpha,\kappa)$ can be rewritten as:

$$\begin{aligned}\widetilde{C}_1(n,\alpha,\kappa) &\leqslant \left|\frac{31}{8}\alpha^2-2\alpha+1\right|+|R(n,\alpha)|+\frac{\frac{3\alpha^2}{4}-2\alpha+1+|Q(n,\alpha)|}{\sqrt{2\kappa}}+\frac{\alpha^2}{4\kappa n^2}+\frac{\alpha K(\alpha)}{\kappa n}\\ &\leqslant \frac{(\alpha-2)(5\alpha-2)}{4\sqrt{2\kappa}}\left[1+\frac{4|Q(n,\alpha)|}{(\alpha-2)(5\alpha-2)}+\frac{\alpha^2\sqrt{2}}{(\alpha-2)(5\alpha-2)\sqrt{\kappa}n^2}\right.\\ &\quad\left.+\left(\left|\frac{31\alpha^2-16\alpha+8}{2(\alpha-2)(5\alpha-2)}\right|+\frac{4\,|R(n,\alpha)|}{(\alpha-2)(5\alpha-2)}\right)\sqrt{2\kappa}\right]\end{aligned}$$

Observe now that for any $\alpha \geqslant \alpha_0 > 2$, we can prove that:

$$\frac{1}{(\alpha-2)(5\alpha-2)} \leqslant \frac{1}{(\alpha_0-2)(5\alpha_0-2)} \quad \text{and} \quad \frac{\alpha^2}{(\alpha-2)(5\alpha-2)} \leqslant \frac{\alpha_0^2}{(\alpha_0-2)(5\alpha_0-2)} \tag{178}$$

so that there exists some real constant $\tilde{c}_1 > 0$ such that:

$$\forall n \geq \frac{\alpha}{\sqrt{\kappa}},\ \widetilde{C}_1(n,\alpha,\kappa) \leqslant \frac{(\alpha-2)(5\alpha-2)}{4\sqrt{2\kappa}}\left(1+\tilde{c}_1\sqrt{\kappa}\right) \tag{179}$$

Combining (123) and (125), there thus exists $\kappa_0 > 0$ such that for any $\kappa \leqslant \kappa_0$ and for any $\alpha \geqslant \alpha_0 > 2$, the inequality (93) holds as expected:

$$E_{n+1}-\left(1-\frac{\alpha-2}{n}\right)E_n \leqslant \frac{C_1(\alpha,\kappa)E_n}{n^2}+\frac{C_2(\alpha,\kappa)E_{n+1}}{(n+1)^2}$$

with:

$$C_1(\alpha,\kappa) = \frac{(\alpha-2)(5\alpha-2)}{4\sqrt{2\kappa}}\left(1+\tilde{c}_1\sqrt{\kappa}\right)\left(1+\sqrt{\kappa}\right)^2 \tag{180}$$

$$C_2(\alpha,\kappa) = \frac{(\alpha-2)^2}{4\sqrt{2}\sqrt{\kappa}}\left(1+\widetilde{c}_2\sqrt{\kappa}\right)\left(1+\sqrt{\kappa}\right)^2 \tag{181}$$

■

A.2.2 Proof of Lemma 4

Assume that the energy E_n satisfies:

$$E_{n+1}-\left(1-\frac{\alpha-2}{n}\right)E_n \leqslant \frac{C_1(\alpha,\kappa)E_n}{n^2}+\frac{C_2(\alpha,\kappa)E_{n+1}}{(n+1)^2} \tag{182}$$

i.e.:

$$\left(1-\frac{C_2(\alpha,\kappa)}{(n+1)^2}\right)E_{n+1}-\left(1-\frac{\alpha-2}{n}+\frac{C_1(\alpha,\kappa)}{n^2}\right)E_n\leqslant 0. \tag{183}$$

Let $n_0\geqslant\frac{\alpha}{\sqrt{\kappa}}$. We then deduce:

$$\forall n\geqslant n_0,\ \log(E_{n+1})-\log(E_{n_0})\leqslant\sum_{k=n_0}^{n}\log\left(\frac{1-\frac{\alpha-2}{k}+\frac{C_1(\alpha,\kappa)}{k^2}}{1-\frac{C_2(\alpha,\kappa)}{(k+1)^2}}\right). \tag{184}$$

Using now the following classical inequalities (if $-1<x$):

$$\frac{x}{x+1}\leq\log(1+x)\leq x \tag{185}$$

we get:

$$\log\left(1-\frac{\alpha-2}{k}+\frac{C_1(\alpha,\kappa)}{k^2}\right)\leq-\frac{\alpha-2}{k}+\frac{C_1(\alpha,\kappa)}{k^2} \tag{186}$$

and

$$-\log\left(1-\frac{C_2(\alpha,\kappa)}{(k+1)^2}\right)\leqslant\frac{C_2(\alpha,\kappa)}{(k+1)^2-C_2(\alpha,\kappa)} \tag{187}$$

We therefore get:

$$\log\left(\frac{1-\frac{\alpha-2}{k}+\frac{C_1(\alpha,\kappa)}{k^2}}{1-\frac{C_2(\alpha,\kappa)}{(k+1)^2}}\right)\leqslant-\frac{\alpha-2}{k}+\frac{C_1(\alpha,\kappa)}{k^2}+\frac{C_2(\alpha,\kappa)}{(k+1)^2-C_2(\alpha,\kappa)} \tag{188}$$

Hence:

$$\log(E_{n+1})-\log(E_{n_0})\leqslant\sum_{k=n_0}^{n}\left(-\frac{\alpha-2}{k}+\frac{C_1(\alpha,\kappa)}{k^2}+\frac{C_2(\alpha,\kappa)}{(k+1)^2-C_2(\alpha,\kappa)}\right) \tag{189}$$

We are now going to make use of the fact that the functions $x\mapsto\frac{1}{x}$, $x\mapsto\frac{1}{x^2}$, and $x\mapsto\frac{C_2}{x^2-C_2}$ are decreasing functions on $(C_2,+\infty)$. Observe that all coefficients in the very last inequality are actually non negative for any $\alpha\geqslant\alpha_0>2$. We then have:

$$\int_k^{k+1}\frac{dx}{x}\leq\frac{1}{k},\quad\frac{1}{(k+1)^2}\leq\int_k^{k+1}\frac{dx}{x^2} \tag{190}$$

and

$$\frac{C_2(\alpha,\kappa)}{(k+1)^2-C_2(\alpha,\kappa)}\leqslant\int_k^{k+1}\frac{C_2(\alpha,\kappa)}{x^2-C_2(\alpha,\kappa)}\,dx \tag{191}$$

so that:

$$\log(E_{n+1})-\log(E_{n_0})\leqslant-(\alpha-2)\int_{n_0}^{n+1}\frac{dx}{x}+C_1(\alpha,\kappa)\int_{n_0-1}^{n}\frac{dx}{x^2}+C_2(\alpha,\kappa)\int_{n_0}^{n+1}\frac{dx}{x^2-C_2(\alpha,\kappa)} \tag{192}$$

Noticing that:

$$\frac{1}{x^2-C_2(\alpha,\kappa)}=\frac{1}{2\sqrt{C_2(\alpha,\kappa)}}\left(\frac{1}{x-\sqrt{C_2(\alpha,\kappa)}}-\frac{1}{x+\sqrt{C_2(\alpha,\kappa)}}\right),$$

we eventually get:

$$\begin{aligned}\log(E_{n+1})-\log(E_{n_0})\ &\leqslant\ -(\alpha-2)\log\left(\frac{n+1}{n_0}\right)+C_1(\alpha,\kappa)\left(\frac{1}{n_0-1}-\frac{1}{n}\right)\\&\quad+\frac{\sqrt{C_2(\alpha,\kappa)}}{2}\log\left(\frac{(n+1-\sqrt{C_2(\alpha,\kappa)})(n_0+\sqrt{C_2(\alpha,\kappa)})}{(n+1+\sqrt{C_2(\alpha,\kappa)})(n_0-\sqrt{C_2(\alpha,\kappa)})}\right)\end{aligned} \tag{193}$$

i.e.:

$$\log(E_{n+1}) - \log(E_{n_0}) \leqslant -(\alpha-2)\log\left(\frac{n+1}{n_0}\right) + C_1(\alpha,\kappa)\left(\frac{1}{n_0-1} - \frac{1}{n}\right) + \frac{\sqrt{C_2(\alpha,\kappa)}}{2}\log\left(\frac{n+1-\sqrt{C_2(\alpha,\kappa)}}{n+1+\sqrt{C_2(\alpha,\kappa)}}\right) + \frac{\sqrt{C_2(\alpha,\kappa)}}{2}\log\left(\frac{n_0+\sqrt{C_2(\alpha,\kappa)}}{n_0-\sqrt{C_2(\alpha,\kappa)}}\right) \tag{194}$$

Taking the exponential, we get:

$$E_{n+1} \le E_{n_0}\left(\frac{n+1}{n_0}\right)^{-(\alpha-2)} \exp(\tilde{\Phi}(n_0) - \tilde{\Phi}(n+1)) \tag{195}$$

with

$$\tilde{\Phi}(n) = \frac{C_1(\alpha,\kappa)}{n-1} + \frac{\sqrt{C_2(\alpha,\kappa)}}{2}\log\left(\frac{n+\sqrt{C_2(\alpha,\kappa))}}{n-\sqrt{C_2(\alpha,\kappa)}}\right) \tag{196}$$

Following the same calculations as in the case $\gamma = 1$, we have:

$$0 \le \frac{\sqrt{C_2(\alpha,\kappa)}}{2}\log\left(\frac{n+\sqrt{C_2(\alpha,\kappa))}}{n-\sqrt{C_2(\alpha,\kappa)}}\right) \leqslant \frac{C_2(\alpha,\kappa)}{n-\sqrt{C_2(\alpha,\kappa)}} \tag{197}$$

hence:

$$0 \le \tilde{\Phi}(n) \le \frac{C_1(\alpha,\kappa)}{n-1} + \frac{C_2(\alpha,\kappa)}{n-\sqrt{C_2(\alpha,\kappa)}} \leqslant \frac{2C_1(\alpha,\kappa)}{n} + \frac{C_2(\alpha,\kappa)}{n-\sqrt{C_2(\alpha,\kappa)}} \tag{198}$$

Now, using the definition of the coefficients $C_1(\alpha,\kappa)$ and $C_2(\alpha,\kappa)$ given in Lemma 3, we get:

$$0 \le \tilde{\Phi}(n) \le \frac{\alpha-2}{4n\sqrt{2\kappa}}(1+\sqrt{\kappa})^2\left[2(5\alpha-2)\left(1+\tilde{c}_1\sqrt{\kappa}\right) + (\alpha-2)\frac{1+\tilde{c}_2\sqrt{\kappa}}{1-\frac{\sqrt{C_2(\alpha,\kappa)}}{n}}\right]$$

Note now that for any $n \geqslant \frac{\alpha}{\sqrt{\kappa}}$,

$$\begin{aligned}\frac{1}{1-\frac{\sqrt{C_2(\alpha,\kappa)}}{n}} &\leqslant \frac{1}{1-\frac{\sqrt{C_2(\alpha,\kappa)}}{\alpha}\sqrt{\kappa}} \leqslant 1+2\frac{\sqrt{C_2(\alpha,\kappa)}}{\alpha}\sqrt{\kappa} \\ &\leqslant 1+2^{-1/4}\frac{\alpha-2}{\alpha}(1+\sqrt{\kappa})\kappa^{1/4}\sqrt{1+\tilde{c}_2\sqrt{\kappa}}\end{aligned}$$

so that there exists a real constant $\tilde{c}_3 > 0$ such that for κ small enough:

$$\frac{1}{1-\frac{\sqrt{C_2(\alpha,\kappa)}}{n}} \leqslant 1+\tilde{c}_3\kappa^{1/4}.$$

Therefore we finally get:

$$0 \le \tilde{\Phi}(n) \leqslant \frac{\alpha-2}{4n\sqrt{\kappa}}(1+\sqrt{\kappa})^2\left[2(5\alpha-2)\left(1+\tilde{c}_1\sqrt{\kappa}\right) + (\alpha-2)(1+\tilde{c}_2\sqrt{\kappa})(1+\tilde{c}_3\kappa^{1/4})\right] \tag{199}$$

so that there exists a real constant $C_3 > 0$ (independent of α) such that:

$$0 \le \tilde{\Phi}(n) \leqslant \frac{(\alpha-2)(11\alpha-6)}{4n\sqrt{2\kappa}}\left(1+C_3\kappa^{1/4}\right) \tag{200}$$

Introducing:

$$\Phi(n) = \frac{(\alpha-2)(11\alpha-6)}{4n\sqrt{2\kappa}}\left(1+C_3\kappa^{1/4}\right)$$

we finally have:

$$\forall n \geqslant n_0,\ E_n \le E_{n_0}\left(\frac{n}{n_0}\right)^{-(\alpha-2)} e^{\Phi(n_0)} \tag{201}$$

■

A.2.3 Technical lemmas

Lemma 8 *Let M_n the mechanical energy, that is:*

$$M_n = F(x_n) - F(x^*) + \frac{1}{2s}\|x_n - x_{n-1}\|^2 \tag{202}$$

Then we have

$$\frac{E_n}{2sn^2} \leqslant \left(1 + \frac{\alpha}{2n\sqrt{\kappa}}\right)^2 M_n. \tag{203}$$

Proof: Let us first remark that:

$$\begin{aligned} b_n &= \left\|\frac{\alpha}{2}(x_{n-1} - x^*) + (n - \frac{\alpha}{4})(x_n - x_{n-1})\right\|^2 = \left\|\frac{\alpha}{2}(x_n - x^*) + \left(n - \frac{3\alpha}{4}\right)(x_n - x_{n-1})\right\|^2 \\ &= \frac{\alpha^2}{4}\|x_n - x^*\|^2 + \left(n - \frac{3\alpha}{4}\right)^2\|x_n - x_{n-1}\|^2 + \alpha\left(n - \frac{3\alpha}{4}\right)\langle x_n - x_{n-1}, x_n - x^*\rangle \end{aligned}$$

Using a discrete version of the inequality (51), we have:

$$|\langle x_n - x^*, x_n - x_{n-1}|\rangle \leqslant \frac{\sqrt{\kappa}}{2}\|x_n - x^*\|^2 + \frac{1}{2\sqrt{\kappa}}\|x_n - x_{n-1}\|^2 \tag{204}$$

so that:

$$b_n \leqslant \frac{\alpha^2}{4}\|x_n - x^*\|^2 + \left(n - \frac{3\alpha}{4}\right)^2\|x_n - x_{n-1}\|^2 + \frac{\alpha n}{2}\left(\sqrt{\kappa}\|x_n - x^*\|^2 + \frac{1}{\sqrt{\kappa}}\|x_n - x_{n-1}\|^2\right)$$

Hence:

$$\begin{aligned} \frac{E_n}{2sn^2} &= F(x_n) - F^* + \frac{1}{2sn^2}b_n \\ &\leqslant M_n + \frac{\alpha^2}{8hn^2}\|x_n - x^*\|^2 + \frac{\alpha n}{4sn^2}\left(\sqrt{\kappa}\|x_n - x^*\|^2 + \frac{1}{\sqrt{\kappa}}\|x_n - x_{n-1}\|^2\right) \end{aligned}$$

Using now the quadratic growth condition $\mathcal{G}^2_\mu$, we get:

$$\frac{E_n}{2sn^2} \leqslant \left(1 + \frac{\alpha}{2n\sqrt{\kappa}}\right)^2 M_n.$$

■

Proof of Lemma 7 Let us prove that for any $n \geqslant 1$ and $(A, B) \in \mathbb{R}^2$, we have:

$$A\delta_n + B(h_{n-1} - h_n) \leqslant \left(2|A + B| + \frac{\sqrt{2}|B|}{\sqrt{s\mu}} + \frac{\alpha^2}{4s\mu n^2}\right)\frac{1}{(n - \frac{3\alpha}{4})^2}E_n. \tag{205}$$

First, observe that:

$$\forall n \in \mathbb{N},\ A\delta_n + B(h_{n-1} - h_n) = (A + B)\delta_n + B(h_{n-1} - h_n - \delta_n) \tag{206}$$

and that for any $\theta > 0$,

$$|h_{n-1} - h_n - \delta_n| = 2|\langle x_n - x_{n-1}, x_n - x^*\rangle| \leqslant \frac{h_n}{\theta} + \theta\delta_n. \tag{207}$$

Combining the last two inequalities, it follows that:

$$\begin{aligned}\forall n\in\mathbb{N},\ A\delta_n+B(h_{n-1}-h_n) &\leqslant (A+B+\theta|B|)\delta_n+\frac{|B|}{\theta}h_n\\ &\leqslant (A+B+\theta|B|)\delta_n+\frac{|B|}{s\mu\theta}w_n\end{aligned}$$

using the growth condition: $h_n\leqslant\frac{1}{s\mu}w_n$ for all $n\in\mathbb{N}$. Let us now focus on the term b_n rewritten as:

$$b_n=\left\|\frac{\alpha}{2}(x_n-x^*)+(n-\frac{3\alpha}{4})(x_n-x_{n-1})\right\|^2 \tag{208}$$

Applying the inequality $\|u\|^2\leqslant 2\|u+v\|^2+\|v\|^2$ to: $u=(n-\frac{3\alpha}{4})(x_n-x_{n-1})$ and $v=\frac{\alpha}{2}(x_n-x^*)$, we get:

$$\forall n\in\mathbb{N},\ (n-\frac{3\alpha}{4})^2\delta_n\leqslant 2b_n+\frac{\alpha^2}{4}h_n. \tag{209}$$

so that:

$$\forall n\in\mathbb{N},\ \delta_n\leqslant\frac{2}{(n-\frac{3\alpha}{4})^2}b_n+\frac{\alpha^2}{4(n-\frac{3\alpha}{4})^2}h_n. \tag{210}$$

We thus deduce:

$$\forall n\in\mathbb{N},\ A\delta_n+B(h_{n-1}-h_n)\leqslant(|A+B|+\theta|B|)\frac{2}{(n-\frac{3\alpha}{4})^2}b_n+\left(\frac{|B|}{s\mu\theta}+\frac{\alpha^2}{4s\mu(n-\frac{3\alpha}{4})^2}\right)w_n \tag{211}$$

Choosing $\theta=\frac{1}{\sqrt{2s\mu}}$, we deduce:

$$A\delta_n+B(h_{n-1}-h_n)\leqslant(2|A+B|+\frac{\sqrt{2}|B|}{\sqrt{s\mu}})\frac{1}{(n-\frac{3\alpha}{4})^2}b_n+\left(\frac{\sqrt{2}|B|}{\sqrt{s\mu}}+\frac{\alpha^2}{4s\mu(n-\frac{3\alpha}{4})^2}\right)w_n \tag{212}$$

Remembering that: $E_n=n^2w_n+b_n$, we finally get:

$$A\delta_n+B(h_{n-1}-h_n)\leqslant\left(2|A+B|+\frac{\sqrt{2}|B|}{\sqrt{s\mu}}+\frac{\alpha^2}{4s\mu n^2}\right)\frac{1}{(n-\frac{3\alpha}{4})^2}E_n. \tag{213}$$

as expected.

■

References

[1] Teodoro Alamo, Daniel Limon, and Pablo Krupa. Restart fista with global linear convergence. In *2019 18th European Control Conference (ECC)*, pages 1969–1974. IEEE, 2019.

[2] V. Apidopoulos, J.-F. Aujol, Ch. Dossal, and A. Rondepierre. Convergence rates of an inertial gradient descent algorithm under growth and flatness conditions. *Mathematical Programming*, 2020.

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

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

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

[6] Jean-Francois Aujol, Ch Dossal, and Aude Rondepierre. Convergence rates of the heavy-ball method with lojasiewicz property. 2020.

[7] Jean-François Aujol, Charles Dossal, Hippolyte Labarrière, and Aude Rondepierre. Fista restart using an automatic estimation of the growth parameter. 2021.

[8] Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Convergence rates of the heavy-ball method for quasi-strongly convex optimization. 2021.

[9] 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.

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

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

[12] A. Chambolle and Ch. 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.

[13] Olivier Fercoq and Zheng Qu. Adaptive restart of accelerated gradient methods under local quadratic growth condition. *IMA Journal of Numerical Analysis*, 39(4):2069–2095, 2019.

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

[15] 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.

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

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

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

[19] 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.

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

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

[22] Chanwoo Park, Jisun Park, and Ernest K Ryu. Factor-$\sqrt{2}$ acceleration of accelerated gradient methods. *arXiv preprint arXiv:2102.07366*, 2021.

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

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

[25] Adrien Taylor and Yoel Drori. An optimal gradient method for smooth strongly convex minimization. *arXiv preprint arXiv:2101.09741*, 2021.