FISTA restart using an automatic estimation of the growth parameter

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

FISTA restart using an automatic estimation of the growth parameter

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

Jean-Francois.Aujol@math.u-bordeaux.fr,

{Charles.Dossal,Hippolyte.Labarriere,Aude.Rondepierre}@insa-toulouse.fr

February 26, 2021

Abstract

In this paper, we propose a novel restart scheme for FISTA (Fast Iterative Shrinking-Threshold Algorithm) [12]. This method which is a generalization of Nesterov's accelerated gradient algorithm [22] is widely used in the field of large convex optimization problems and it provides fast convergence results under a strong convexity assumption. These convergence rates can be extended for weaker hypotheses such as the Łojasiewicz property but it requires prior knowledge on the function of interest. In particular, most of the schemes providing a fast convergence for non-strongly convex functions satisfying a quadratic growth condition involve the growth parameter which is generally not known. Recent works [2, 1] show that restarting FISTA could ensure a fast convergence for this class of functions without requiring any geometry parameter. We improve these restart schemes by providing a better asymptotical convergence rate and by requiring a lower computation cost. We present numerical results emphasizing that our method is efficient especially in terms of computation time.

Key-words FISTA, restart, convex optimization, Łojasiewicz property, convergence rate, growth parameter.

*Univ. Bordeaux, Bordeaux INP, CNRS, IMB, UMR 5251, F-33400 Talence, France
†IMT, Univ. Toulouse, INSA Toulouse, Toulouse, France
‡LAAS, Univ. Toulouse, CNRS, Toulouse, France

1 Introduction

This article introduces a new algorithm to minimize efficiently a large set of convex functions with a quadratic growth. It achieves the best decay rate on this class of functions, in the case when the quadratic growth constant is not known.

We are interested in the minimization of a function F defined as follows. Let $F = f + h$ a convex function with $F : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ with $N > 0$. The function f is assumed to be convex, differentiable and with L-Lipschitz gradient, while h is assumed to be a convex function whose proximal operator is known. The set of minimizers of F denoted X^* is non empty. The Forward-Backward algorithm (FB) produces a sequence $(x_k)_{k\geqslant 1}$ ensuring $F(x_k) - F^* = \mathcal{O}\left(\frac{1}{k}\right)$ [16] where $F^* = \inf F$. Under the same assumptions, the FISTA Algorithm [12], based on the ideas of acceleration of Nesterov [22], ensures a better asymptotical bound: $F(x_k) - F^* = \mathcal{O}\left(\frac{1}{k^2}\right)$.

In numerous optimization problems, as for instance in statistics or in image processing, the function F to minimize satisfies more hypotheses, which allows to reach better decay rates. This paper focuses on the set of functions satisfying a quadratic growth condition, i.e such that it exists $\mu > 0$ such that the following inequality holds:

$$\forall x \in \mathbb{R}^N,\ \frac{\mu}{2} d(x, X^*)^2 \leqslant F(x) - F^*. \tag{1}$$

This set of functions includes the set of strongly convex functions, but it is much larger. For example, it contains functions F associated to the mean square problem and the LASSO [15]. Note that the uniqueness of the minimizer of F is not required. Recall that under a convexity assumption this set is equal to the set of functions satisfying a global Łojasiewicz property of parameter $\frac{1}{2}$ [14].

On this set of functions, FB reaches an exponential decay rate: $\mathcal{O}\left(e^{-\kappa k}\right)$ where $\kappa := \frac{\mu}{L}$ is the ratio between the growth parameter μ defined in (1) and the Lipschitz constant L of ∇f [17]. It turns out that this exponential decay cannot be observed in many numerical experiments because the condition number $\kappa = \frac{\mu}{L} << 1$ can be very small, especially in large dimension problems. Most of the time, up to a large accuracy, the quadratic bound of FISTA is numerically better than the theoretical exponential decay of FB. Indeed FB needs at least several times $\frac{L}{\mu}$ iterations to provide good results while FISTA provides interesting ones as soon as k is proportional to $\sqrt{\frac{L}{\mu}}$.

Under additional hypotheses, some inertial algorithms such as the Heavy Ball of Polyak [25] and its numerous variants [18, 10, 9] reach a better decay rate: $\mathcal{O}\left(e^{-c\sqrt{\kappa}k}\right)$ where c depends on the regularity hypotheses of F and on the variant of the algorithm.

Unfortunately, these inertial algorithms need an accurate a priori estimation of the growth parameter μ to get these fast decays. Moreover, in many situations, μ

is small and unknown. An incorrect estimation of μ may significantly reduce the speed of the algorithm.

In this paper, we propose an algorithm based on an original restart rule of FISTA ensuring:

$$F(x_k) - F^* = \mathcal{O}\left(e^{-\frac{1}{12}\sqrt{\kappa}k}\right) \tag{2}$$

wihthout any a priori knowledge of μ. This algorithm relies on an iterative estimation of μ comparing some values $F(x_k)$. The restart rule is inspired by the one proposed by Alamo et al [2]. With respect to [2], we improve the decay rate and we reduce the number of estimations of $F(x_k)$ during the iterations (notice that these evaluations may heavily impact the numerical cost and the time of computation).

As a consequence, our algorithm provides the best decay rate on the set of convex functions satisfying the quadratic growth condition (1), in the case when the growth parameter μ is not known.

The article is structured as follows. In Section 2, a state of the art of this optimization problem is given. Section 3 is then devoted to the definition and the properties of our restart algorithm. We also propose some numerical experiments and comparisons. The proofs of the results of Section 3 are postponed to Section 4.

2 State of the art

2.1 Framework

We consider a composite function $F : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ defined as $F = f + h$ where f is a convex differentiable function having a L-Lipschitz gradient and h is a convex proper lower semicontinuous function. We suppose that F has a non empty set of minimizers. The focus of this paper lies in the efficient minimization problem:

$$\min_{x \in \mathbb{R}^N} F(x). \tag{3}$$

We introduce some notations that will be needed later on. The set of minimizers of F is denoted X^* and $F^* = \inf F$. The gradient of f is denoted by ∇f and the convex subdifferential of h is denoted by ∂h. We recall that:

$$\forall x \in \mathbb{R}^N, \quad \partial h(x) = \left\{ s \in \mathbb{R}^N \mid \forall y \in \mathbb{R}^N, h(y) \geqslant h(x) + \langle s, y - x \rangle \right\}. \tag{4}$$

This paper focuses on the set of functions which have a global Łojasiewicz property [19, 20, 13] with an exponent $\frac{1}{2}$. This assumption is equivalent to a global quadratic growth condition around the set of minimizers of F.

Definition 1. *Let $F : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ be a proper lower semicontinuous convex function with a non empty set of minimizers X^*. Let $F^* = \inf F$. The function F satisfies the growth condition $\mathcal{G}_\mu^2$ for some $\mu > 0$ if we have:*

$$\forall x \in \mathbb{R}^N, \quad \frac{\mu}{2} d(x, X^*)^2 \leqslant F(x) - F^*. \tag{5}$$

or equivalently F has a global Łojasiewicz property with an exponent $\frac{1}{2}$ if there exists $c > 0$ such that:

$$\forall x \in \mathbb{R}^N, \quad F(x) - F^* \leqslant c d(0, \partial F(x))^2. \tag{6}$$

This inequality is valid for $c = \frac{1}{2\mu}$.

This property is sufficiently weak to gather a large set of functions. The class of functions satisfying the growth condition $\mathcal{G}_\mu^2$ includes quadratic functions as well as μ-strongly convex functions and μ-quasi-strongly convex functions introduced by Necoara et al. in [21]. We give a special interest to this assumption as it is satisfied by many functions that are widely used in statistics and image processing. A well-known example of such a function is the LASSO function:

$$F(x) = \frac{1}{2}\|Ax - y\|^2 + \lambda \|x\|_1. \tag{7}$$

Bolte et al. prove in [15, Corollary 9] that the LASSO function (7) has a Łojasiewicz property of exponent $\frac{1}{2}$.

This paper focuses on the functions satisfying the hypothesis $\mathcal{H}$ defined as follows:

Definition 2. *Let $F : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$. The function F satisfies the hypothesis $\mathcal{H}(\mu)$ if:*

i. it can be written $F = f + h$ where f is a convex differentiable function having a L-Lipschitz gradient and h is a convex proper lower semicontinuous function.

ii. it has a non-empty set of minimizers X^*.

iii. it satisfies the growth condition $\mathcal{G}_\mu^2$.

2.2 Literature review

Studying the convergence rate of first order algorithms is a great point of interest in current research. A common way to get convergence results is to analyze an ODE associated to the algorithm of interest. For example, Su et al. bring out in

[31] that Nesterov's accelerated gradient method can be seen as a discretization of an ODE (8) modeling a dynamical system:

$$\forall t \geq t_0, \quad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla F(x(t)) = 0. \tag{8}$$

In many recent works, convergence results in a continuous setting are obtained by using a Lyapunov analysis, see e.g. [31, 7, 4, 5, 6, 8, 9, 10, 28, 29, 27].

To set some definitions we would say that:

Definition 3. *An optimization algorithm provides a fast exponential decay on the class of functions F satisfying a growth condition with parameter μ and having a L-Lipschitz gradient if there exists $K > 0$ such that the sequence $(x_k)_{k\in\mathbb{N}}$ provided by this algorithm satisfies*

$$F(x_k) - F^* = \mathcal{O}\left(e^{-K\sqrt{\frac{\mu}{L}}k}\right).$$

Definition 4. *An optimization algorithm provides a low exponential decay on the class of functions F satisfying a growth condition with parameter μ and having a L-Lipschitz gradient if there exists $K > 0$ such that the sequence $(x_k)_{k\in\mathbb{N}}$ provided by this algorithm satisfies*

$$F(x_k) - F^* = \mathcal{O}\left(e^{-K\frac{\mu}{L}k}\right).$$

In most practical cases the function F is ill-conditioned i.e. $\mu \ll L$. The difference between a fast and a low exponential decay is then significant since $\frac{\mu}{L} \ll \sqrt{\frac{\mu}{L}}$.

If F is supposed to have a Łojasiewisz property with an exponent $\frac{1}{2}$, several classical first-order methods provide a low exponential decay. Let F be differentiable with a L-Lipschitz gradient. The Gradient Descent is a classical algorithm defined by:

$$\forall k > 0, \quad x_{k+1} = x_k - s\nabla F(x_k), \quad s > 0, \tag{9}$$

which ensures that $F(x_k) - F^* = \mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$ in this setting. Polyak proves in [26] that the best rate is obtained for $s = \frac{2}{L+\mu}$.

In [26], Polyak introduces the Heavy-Ball method (10) defined by:

$$\forall k > 0, \quad \begin{cases} x_k = y_{k-1} - s\nabla F(x_{k-1}) \\ y_k = x_k + \alpha(x_k - x_{k-1}) \end{cases}, \quad s > 0, \tag{10}$$

which is a discretization of the Heavy-Ball ODE defined by:

$$\forall t \geq t_0, \quad \ddot{x}(t) + \alpha\dot{x}(t) + \nabla F(x(t)) = 0. \tag{11}$$

This algorithm is designed to reach a fast exponential decay for C^2 strongly-convex functions. In [18], Ghadimi et al. get a low exponential decay if F is C^1 with L-Lipschitz gradient and Siegel proves in [30] that a fast exponential decay can be found in the same setting. In [10], Aujol et al. show that this method gives a low exponential decay for functions satisfying the growth condition $\mathcal{G}_\mu^2$.

A variation of the Heavy-Ball method is proposed in [10] by using an other discretization of the ODE (11). This scheme provides a fast exponential decay in $\mathcal{O}\left(e^{-(2-\sqrt{2})\sqrt{\frac{\mu}{L}}k}\right)$ for the set $\mathcal{G}_\mu^2$ assuming the uniqueness of the minimizer.

Necoara et al. observe in [21] that the classical accelerated gradient algorithm introduced by Nesterov (12) in [22] can have a fast exponential decay by restarting it in a specific way.

$$\forall k > 0, \quad \begin{cases} x_k = y_{k-1} - s\nabla F(y_{k-1}) \\ y_k = x_k + \dfrac{k-1}{k+2}(x_k - x_{k-1}) \end{cases}, \quad s > 0. \tag{12}$$

The authors prove that if F is μ-strongly convex and the algorithm (12) is restarted every $k^* = \lfloor 2e\sqrt{\frac{L}{\mu}} \rfloor$ iterations, we get that $F(x_k) - F^* = \mathcal{O}\left(e^{-\frac{1}{e}\sqrt{\frac{\mu}{L}}k}\right)$. This result holds if F has a Łojasiewisz property of exponent $\frac{1}{2}$ by replacing the parameter of strong convexity by the growth parameter μ satisfying (5). This scheme has a fast convergence rate but it requires to know or to estimate the growth parameter μ beforehand. This may be restrictive in numerical experiments since the growth parameter is rarely known. Moreover, this convergence rate does not hold if μ is overestimated. A possibility is to underestimate μ which causes the algorithm to slow down.

O'Donoghue and Candès propose heuristic restart rules for accelerated gradient method in [24]. These rules are based on empirical observations and they give efficient numerical results. For example, the first scheme consists in restarting the algorithm each iterations where $F(x_{k+1}) > F(x_k)$ holds. This scheme is a way to avoid the oscillations of Nesterov's accelerated gradient method and is particularly efficient on the set $\mathcal{G}_\mu^2$ but it requires to compute $F(x_k)$ at each iteration and no convergence rate has been found. Beck and Teboulle use a similar strategy in [11] to build a monotone version of Nesterov's accelerated gradient method. This scheme provides the same convergence rate as the original version.

In [2], Alamo et al. propose an adaptative restart scheme of accelerated gradient method with a fast exponential decay. This method does not require any estimation of μ and relies on a restart condition and a doubling condition. The combination of these two conditions enables to estimate μ at each restart. This strategy allows this method to ensure that $F(x_k) - F^* = \mathcal{O}\left(e^{-\frac{1}{16}\sqrt{\frac{\mu}{L}}k}\right)$. However, $F(x_k)$ has to be evaluated at least at half of the iterations. Thus, this algorithm

may not be efficient in terms of calculation time if computing F is expensive which is mostly the case in practice.

These schemes can be adapted to a non differentiable setting and some convergence results hold for such functions. Let F be a convex proper lower semicontinuous function which has a non-empty set of minimizers X^* and which satisfies the growth condition $\mathcal{G}_\mu^2$. We define the proximal operator of $h : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ a convex semicontinuous function as follows:

$$\forall x \in \mathbb{R}^N, \quad \text{prox}_h(x) = \underset{y \in \mathbb{R}^N}{\text{argmin}}\ h(y) + \frac{1}{2}\|y - x\|^2. \tag{13}$$

The Forward-Backward algorithm can be seen as a Gradient Descent method in a non differentiable setting. If F is defined such that $F = f + h$ where f is a convex differentiable function having a L-Lipschitz gradient and h is a convex proper lower semicontinuous function, then this scheme is defined by

$$\forall k > 0, \quad x_k = \text{prox}_{sh}(x_{k-1} - s\nabla f(x_{k-1})), \quad s > 0. \tag{14}$$

The convergence results obtained for a differentiable function hold in this setting and this algorithm provides a low exponential decay. Similarly, the Heavy-Ball method can be adapted to non differentiable functions and Nesterov's accelerated gradient is generalized to this setting in [12] resulting in FISTA. The convergence rates stated previously are still valid, and they are summarized in Table 1.

Algorithm	**References**	**Convergence rate**	**Limitations**
Forward-Backward	Garrigos et al. [17]	$\mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$	-
Heavy-Ball method	Polyak [26]	$\mathcal{O}\left(e^{-4\frac{\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}k}\right)$	Requires μ-strong convexity, a C^2 assumption and an estimation of μ
Heavy-Ball variation	Aujol et al. [10]	$\mathcal{O}\left(e^{-(2-\sqrt{2})\sqrt{\frac{\mu}{L}}k}\right)$	Requires an estimation of μ and uniqueness of minimizer
Optimal FISTA restart	Necoara et al. [21]	$\mathcal{O}\left(e^{-\frac{1}{e}\sqrt{\frac{\mu}{L}}k}\right)$	Requires an estimation of μ
Monotone FISTA	Beck and Teboulle [11]	$\mathcal{O}\left(k^{-2}\right)$	Requires to compute F regularly
Empirical FISTA restart	O'Donoghue and Candès [24]	-	Numerically fast but no convergence rate
FISTA restart by Alamo et al.	Alamo et al. [2]	$\mathcal{O}\left(e^{-\frac{1}{16}\sqrt{\frac{\mu}{L}}k}\right)$	Requires to compute F regularly
Adaptative restart	Theorem 1.	$\mathcal{O}\left(e^{-\frac{1}{12}\sqrt{\frac{\mu}{L}}k}\right)$	-

Table 1: Convergence rates of classical algorithms for a possibly non differentiable function F satisfying $\mathcal{H}(\mu)$ for some $\mu > 0$.

We propose a novel restart scheme of accelerated gradient method and FISTA which provides a fast exponential decay in $\mathcal{O}\left(e^{-\frac{1}{12}\sqrt{\frac{\mu}{L}}k}\right)$. This algorithm does not require to estimate the growth parameter μ. Moreover, the value of $F(x_k)$ is computed for a reduced number of iterations. As a consequence, this method has a fast convergence rate in terms of iterations and computational time as well.

3 Contributions

In this section we introduce a restart scheme for FISTA in order to minimize a convex function F satisfying a Łojasiewisz property with an exponent $\frac{1}{2}$. We give a convergence rate and describe the underlying strategy.

3.1 Adaptative FISTA restart scheme

We give some notations in order to simplify the writing of the algorithm. Let y^+ be the vector given by a step of Forward-Backward algorithm (14) on y with $s = \frac{1}{L}$. Namely, y^+ is defined by:

$$y^+ = \text{prox}_{\frac{1}{L}h}(y - \frac{1}{L}\nabla f(y)), \tag{15}$$

We define the composite gradient mapping g as follows:

$$g(y) = y - y^+. \tag{16}$$

Given initial condition $z \in \mathbb{R}^N$ and a number of iterations n, FISTA [12] can be written as Algorithm 1.

Algorithm 1 : FISTA

Require: $z \in \mathbb{R}^N,\ n \in \mathbb{N}$
  $y_0 = x_0 = z^+,\ k = 0$
  repeat
    $k = k + 1$
    $x_k = y_{k-1}^+$
    $y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1})$
  until $k \geqslant n$
  return $r = x_k$

We introduce a restart scheme which aims to accelerate the convergence of FISTA. This algorithm involves several variables of interest:

- the sequence $(r_j)_{j\in\mathbb{N}}$ corresponds to the iterates of the algorithm. For all $j > 0$, r_j is obtained from $j-1$ consecutive restarts of FISTA.
- the sequence $(n_j)_{j\in\mathbb{N}}$ refers to the number of iterations of FISTA following the j-th restart. Namely, for all $j \geqslant 0$ we have:

$$r_{j+1} = \text{FISTA}(r_j, n_j),$$

- the sequence $(\tilde{\mu}_j)_{j\geqslant 2}$ estimates the growth parameter μ at each restart. This estimation is built from the known convergence results of FISTA and considering a comparison of the cost function F computed at three iterates.

For all $j \geqslant 2$, the number of iterations n_j is defined according to n_{j-1}, $\tilde{\mu}_j$ and the predefined parameter $C > 0$. The algorithm verifies if a condition called doubling condition is fulfilled. If the condition holds true, then n_{j-1} is considered too small and n_j is set to $2n_{j-1}$. In the other case, the number of iterations is not increased.

The exit condition $\|g(r_j)\| \leqslant \varepsilon$ is motivated by the following lemma which is proven in Section 4.5:

Lemma 1. *Let F be a function satisfying $\mathcal{H}(\mu)$ for some $\mu > 0$. Then for all $x \in \mathbb{R}^N$ we have:*

$$F(x^+) - F^* \leqslant \frac{2L^2}{\mu}\|g(x)\|^2. \tag{17}$$

The exit criteria combined to the inequality (17) enable the algorithm to ensure that the vector r_j satisfies:

$$F(r_j^+) - F^* \leqslant \frac{2L^2\varepsilon^2}{\mu}, \tag{18}$$

without computing F^*.

For a given initial condition $r_0 \in \mathbb{R}^N$, we propose the algorithm shown in Algorithm 2.

Algorithm 2 : Restart scheme

Require: $r_0 \in \mathbb{R}^N, j = 1$
  $n_0 = \lfloor 2C \rfloor$
  $r_1 = \text{FISTA}(r_0, n_0)$
  $n_1 = \lfloor 2C \rfloor$
  repeat
    $j = j + 1$
    $r_j = \text{FISTA}(r_{j-1}, n_{j-1})$
    $\tilde{\mu}_j = \min\limits_{\substack{i\in\mathbb{N}^* \\ i<j}} \dfrac{4L}{(n_{i-1}+1)^2} \dfrac{F(r_{i-1}) - F(r_j)}{F(r_i) - F(r_j)}$
    if $n_{j-1} \leqslant C\sqrt{\frac{L}{\tilde{\mu}_j}}$ **then**
      $n_j = 2n_{j-1}$
    end if
  until $\|g(r_j)\| \leqslant \varepsilon$
  return $r = r_j$

Theorem 1. *Let F be a function satisfying $\mathcal{H}(\mu)$ for some $\mu > 0$. Let $(r_j)_{j\in\mathbb{N}}$ and $(n_j)_{j\in\mathbb{N}}$ be the sequences provided by Algorithm 2 with parameters $C > 4$ and $\varepsilon > 0$. Then the number of iterations required to guarantee $\|g(r_j)\| \leqslant \varepsilon$ is bounded and satisfies*

$$\sum_{i=0}^{j} n_i \leqslant \frac{4C}{\log\left(\frac{C^2}{4}-1\right)}\sqrt{\frac{L}{\mu}}\left(2\log\left(\frac{C^2}{4}-1\right)+\log\left(1+\frac{16}{C^2-16}\frac{2(F(r_0)-F^*)}{L\varepsilon^2}\right)\right). \tag{19}$$

Corollary 1. *Let F be a function satisfying $\mathcal{H}(\mu)$ for some $\mu > 0$. If $C > 4$ and $\varepsilon > 0$, then the sequences $(r_j)_{j\in\mathbb{N}}$ and $(n_j)_{j\in\mathbb{N}}$ provided by Algorithm 2 satisfy*

$$F(r_j^+) - F^* = \mathcal{O}\left(e^{-\frac{\log\left(\frac{C^2}{4}-1\right)}{4C}\sqrt{\frac{\mu}{L}}\sum_{i=0}^{j} n_i}\right). \tag{20}$$

Specifically, if C is chosen to maximize $\frac{\log\left(\frac{C^2}{4}-1\right)}{4C}$, namely $C \approx 6.38$, then there exists $K > \frac{1}{12}$ such that the sequences $(r_j)_{j\in\mathbb{N}}$ and $(n_j)_{j\in\mathbb{N}}$ provided by Algorithm 2 satisfy

$$F(r_j^+) - F^* = \mathcal{O}\left(e^{-K\sqrt{\frac{\mu}{L}}\sum_{i=0}^{j} n_i}\right). \tag{21}$$

Corollary 1 states that Algorithm 2 provides asymptotically a fast exponential decay. We recall that under our assumptions, the Forward-Backward algorithm and the Gradient Descent method (if F is differentiable) ensure that $F(x_k) - F^* = \mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$. As $\frac{\mu}{L} \ll \sqrt{\frac{\mu}{L}}$ when $\frac{\mu}{L}$ is small, the convergence rate of Algorithm 2 is much better than the convergence rate of these two methods. Similarly, if we do not consider that F has a unique minimizer, Algorithm 2 has a better convergence rate than any Heavy-Ball method in the literature so far.

Elementary computations [21, 24] show that restarting FISTA every $k^* = \lfloor 2e\sqrt{\frac{L}{\mu}} \rfloor$ iterations ensures a faster convergence than Algorithm 2. However it requires a low estimation of the growth parameter μ. Applying this method to a function whose growth parameter can not be estimated may not be relevant. Algorithm 2 is slightly slower but it is applicable to a larger spectrum of problems.

The empiric restart schemes introduced by O'Donoghue and Candès in [24] can not be compared theoretically to Algorithm 2 since no convergence rate has been proven so far. However, the rate given in Corollary 1 is significantly faster than the rate of the monotone version of FISTA ($\mathcal{O}\left(k^{-2}\right)$) introduced by Beck and Teboulle in [11].

Algorithm 2 has been inspired by the adaptative restart scheme of Alamo et al. [2] and its convergence rate (21) provided by Corollary 1 is slightly faster. Under the same assumptions, the method introduced in [2] ensures that:

$$F(r_j^+) - F^* = \mathcal{O}\left(e^{-\frac{1}{16}\sqrt{\frac{\mu}{L}}\sum_{i=0}^{j} n_i}\right). \tag{22}$$

Moreover, both schemes rely on exit and doubling conditions which require to compute F. The restart scheme of Alamo et al. [2] requires a computation of F at least one iteration out of two. These computations are significantly less numerous in Algorithm 2 as it occurs every $\lfloor 2C \rfloor$ iterations in the worst case. This difference between the two schemes has a major impact on their performance in terms of computing time. Computing F can be really expensive in some numerical cases and reducing these computations is crucial.

3.2 Structure of the algorithm

Algorithm 2 is an adaptative restart scheme based on FISTA [12]. At each step j, FISTA is restarted for n_{j-1} iterations and it gives the vector r_j:

$$r_j = \text{FISTA}(r_{j-1}, n_{j-1}).$$

The key parameter of such an algorithm is the sequence $(n_j)_{j\in\mathbb{N}}$. The strategy behind Algorithm 2 is to estimate the growth parameter $\mu > 0$ satisfying (5) at each step j. The estimation of μ denoted $\tilde{\mu}_j$ is then used to define n_j.

The estimation of μ is based on well-known convergence results on FISTA stated in Proposition 1 and proven in Section 4.4.

Proposition 1. *Let F be a function satisfying $\mathcal{H}(\mu)$ for some $\mu > 0$. Then the*

sequence $(x_k)_{k\in\mathbb{N}}$ *provided by Algorithm 1 satisfies*

$$\textit{(i)}\ \forall k \in \mathbb{N}^*,\ F(x_k) - F^* \leqslant \frac{4L}{\mu(k+1)^2}\left(F(x_0) - F^*\right), \tag{23}$$

$$\textit{(ii)}\ \forall k \in \mathbb{N}^*,\ F(x_k) \leqslant F(x_0), \tag{24}$$

$$\textit{(iii)}\ \forall k \in \mathbb{N},\ F(x_k) - F^* > \gamma\left(F(x_0) - F(x_k)\right) \implies k < 2\sqrt{\frac{L}{\mu}}\sqrt{1+\frac{1}{\gamma}} - 1. \tag{25}$$

Given the first claim of Proposition 1, the most direct strategy to estimate μ is to compare $F(r_j) - F^*$ and $F(r_{j-1}) - F^*$ at each step j as we have:

$$\forall j \in \mathbb{N}^*,\ F(r_j) - F^* \leqslant \frac{4L}{\mu(n_{j-1}+1)^2}\left(F(r_{j-1}) - F^*\right), \tag{26}$$

which is equivalent to:

$$\forall j \in \mathbb{N}^*,\ \mu \leqslant \frac{4L}{(n_{j-1}+1)^2}\frac{F(r_{j-1}) - F^*}{F(r_j) - F^*}. \tag{27}$$

In many cases the optimal value F^* is not known and μ cannot be estimated in this way. This issue can be avoided by rewriting (26) and by considering a third point r_{j+1}. For all $j \in \mathbb{N}^*$,

$$F(r_j) - F^* > F(r_j) - F(r_{j+1}) = \gamma(F(r_{j-1}) - F(r_j)), \tag{28}$$

where $\gamma = \dfrac{F(r_j) - F(r_{j+1})}{F(r_{j-1}) - F(r_j)}$. The third claim of Proposition 1 gives us that:

$$\mu \leqslant \frac{4L}{(n_{j-1}+1)^2}\left(1 + \frac{1}{\gamma}\right). \tag{29}$$

Thus,

$$\forall j \in \mathbb{N}^*,\ \mu \leqslant \frac{4L}{(n_{j-1}+1)^2}\frac{F(r_{j-1}) - F(r_{j+1})}{F(r_j) - F(r_{j+1})}. \tag{30}$$

In this way, it is possible to get an estimation of μ at each restart for $j \geqslant 2$ by comparing $F(r_{j-1}) - F(r_j)$ and $F(r_{j-2}) - F(r_j)$. This strategy is implemented in Algorithm 2 to build the sequence $(\tilde{\mu}_j)_{j\geqslant 2}$:

$$\forall j \geqslant 2, \quad \tilde{\mu}_j = \min_{\substack{i\in\mathbb{N}^* \\ i<j}} \frac{4L}{(n_{i-1}+1)^2}\frac{F(r_{i-1}) - F(r_j)}{F(r_i) - F(r_j)}. \tag{31}$$

This sequence satisfies the following property proven in Section 4.6:

Lemma 2. *Let F be a function satisfying $\mathcal{H}(\mu)$ for some $\mu > 0$. Then the sequence $(\tilde{\mu}_j)_{j\geqslant 2}$ provided by Algorithm 2 satisfies*

$$\forall j \geqslant 2, \quad \tilde{\mu}_j \geqslant \tilde{\mu}_{j+1} > \mu. \tag{32}$$

It is worth noticing that this strategy is not inherent to FISTA or Nesterov's accelerated gradient scheme. It relies on the fact that an upper bound of $F(x_k)-F^*$ is known for all $k > 0$. Note that a similar approach could be applied for the Heavy-ball method (10) as the growth parameter μ has to be known to get an optimal decay rate.

Once the growth parameter μ is estimated it is possible to set the number of iterations accordingly. It is known that the optimal number of iterations before a restart of FISTA is $k^* = \lfloor 2e\sqrt{\frac{L}{\mu}} \rfloor$. However, considering $\tilde{\mu}$ an upper estimation of μ, the growth condition $\mathcal{G}^2_{\tilde{\mu}}$ is not satisfied by F. Consequently, setting $k^* = \lfloor 2e\sqrt{\frac{L}{\tilde{\mu}}} \rfloor$ does not ensure a fast exponential decay. Since $\tilde{\mu}_j \geqslant \mu$ for all $j \geqslant 2$, setting $n_j = \lfloor 2e\sqrt{\frac{L}{\tilde{\mu}_j}} \rfloor$ may not be efficient.

The strategy of Algorithm 2 is to reset n_j by using a doubling condition that depends on the parameter $C > 0$. At each step the doubling condition $n_j \leqslant C\sqrt{\frac{L}{\tilde{\mu}_j}}$ is evaluated. If the condition is satisfied then the number of iterations n_j is considered too small and it is doubled before the next restart. This process ensures that $(n_j)_{j\in\mathbb{N}}$ is an increasing and bounded sequence which satisfies the following lemma.

Lemma 3. *Let F be a function satisfying $\mathcal{H}(\mu)$ for some $\mu > 0$. Then the sequence $(n_j)_{j\in\mathbb{N}}$ provided by Algorithm 2 satisfies*

$$\forall j \in \mathbb{N}, \quad n_j \leqslant 2C\sqrt{\frac{L}{\mu}}. \tag{33}$$

Remark 1. *An alternative approach would be to multiply the number of iterations n_j by $\gamma > 1$ instead of 2 when $n_j \leqslant C\sqrt{\frac{L}{\tilde{\mu}_j}}$. In that case the upper bound of n_j becomes*

$$\forall j \in \mathbb{N}, \quad n_j \leqslant \gamma C\sqrt{\frac{L}{\mu}}. \tag{34}$$

However, calculations show that $\gamma = 2$ is the optimal choice in order to get faster asymptotical convergence rates.

3.3 Numerical experiments

In this section we illustrate the convergence results stated beforewise with numerical experiments. We compare Algorithm 2 to the state of the art for some classical minimization problems.

3.3.1 Least squares problem

We first solve a classical least squares problem. This minimization problem reads as follows:

$$\min_{x\in\mathbb{R}^N} F(x) = \frac{1}{2}\|Ax-b\|^2, \tag{35}$$

where $A \in \mathcal{M}_{N,N}(\mathbb{R})$, $b \in \mathbb{R}^N$ and $N \geq 1$. F is a convex, differentiable function with a continuous L-Lipschitz gradient. Moreover, F has the Łojasiewicz property with an exponent $\frac{1}{2}$ [15, Corollary 9] and the growth parameter μ can easily be computed. However, we compare algorithms which do not require to know this parameter. Computations are indeed done for the following methods:

1. Gradient descent (9),
2. Nesterov's accelerated gradient algorithm without restart (12),
3. Empirical restart scheme of Nesterov's accelerated gradient algorithm by O'Donoghue and Candès [24] with the following restart condition:

$$F(x_k) > F(x_{k-1}), \tag{36}$$

4. Empirical restart scheme of Nesterov's accelerated gradient algorithm by O'Donoghue and Candès [24] with the following restart condition:

$$\langle \nabla F(y_{k-1}), x_k - x_{k-1}\rangle > 0, \tag{37}$$

5. Adaptative restart scheme introduced by Alamo et al. in [2],
6. Algorithm 2 with $C = 6.38$.

Figure 1 shows the evolution of $\log(F(x_n) - F^*)$ according to n. It highlights the fast convergence of the restart schemes in comparison to Gradient descent and Nesterov's accelerated gradient method without restart. The restart schemes proposed by O'Donoghue and Candès provide significantly fast decay despite not having any theoretical convergence rate. In this example, the adaptative restart scheme proposed by Alamo et al. is slightly faster than Algorithm 2 in terms of iterations.

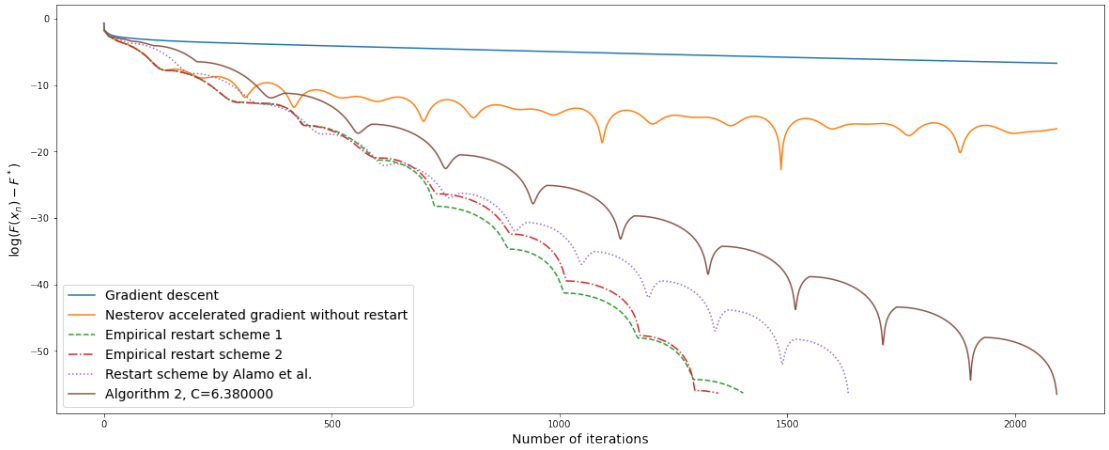

Figure 1: Comparison of several algorithms in terms of iterations for a least squares problem with $N = 10$.

Figure 2 shows that the calculations of F have a great influence on the computational time of each algorithm. The restart condition (36) proposed in [24] ensures a fast convergence in terms of iterations but it is computationnaly expensive. Similarly, the method introduced by Alamo et al. is slowed down by its additional computations. As Algorithm 2 does not require many evaluations of F, its convergence rate is still fast according to computation time.

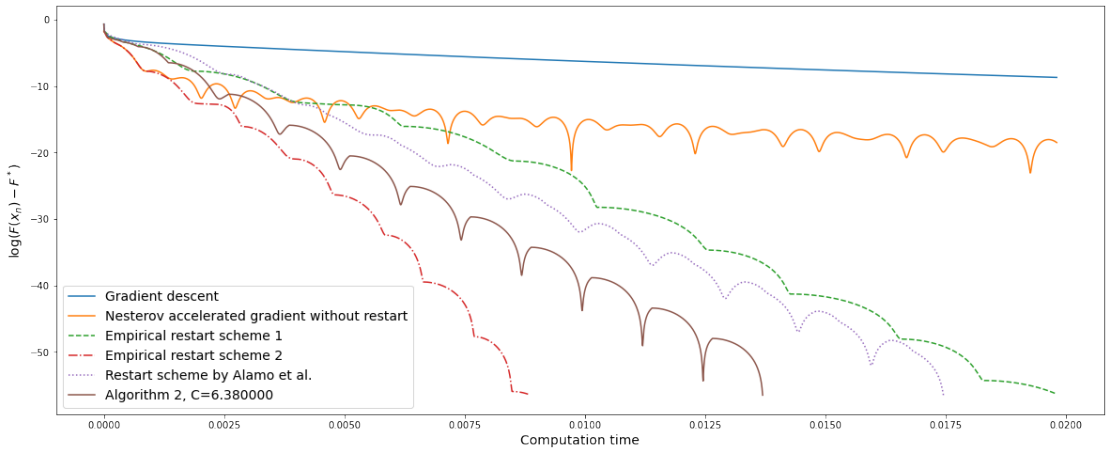

Figure 2: Comparison of several algorithms in terms of computation time for a least squares problem with $N = 10$.

3.3.2 Inpainting

Consider an image x^0 and a masking operator M. Let $y = Mx^0$ be the damaged version of x^0. The objective is to get an approximation of x^0 knowing y and M. This problem can be written as follows:

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

where T is an orthogonal transformation ensuring that Tx^0 is sparse. In this example, x^0 is piecewise smooth so we chose T as an orthogonal wavelet transform.

Figure 3: Example of image inpainting: the damaged image y is on the left and an approximation of the solution of (38) is on the right.

The objective function F can be rewritten as $F = f + h$ where f is a convex differentiable function having a L-Lipschitz gradient and h is a convex proper lower semicontinuous function. We compare the following methods:

1. Forward-backward (14),
2. FISTA without restart,
3. Empirical restart scheme of FISTA by O'Donoghue and Candès [24] with the restart condition (36),
4. Empirical restart scheme of FISTA by O'Donoghue and Candès [24] with the adapted restart condition:

$$\langle g(y_{k-1}), x_k - x_{k-1}\rangle > 0, \tag{39}$$

5. Adaptative restart scheme introduced by Alamo et al. in [2],

6. Algorithm 2 with $C = 6.38$.

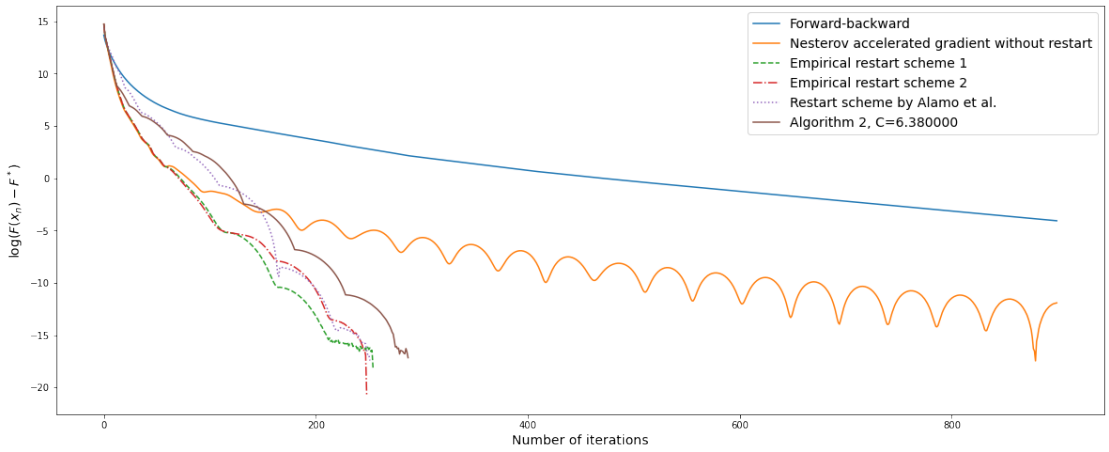

Figure 4: Comparison of several algorithms in terms of iterations for an inpainting problem.

Figure 4 and Figure 5 show that the restart schemes provide a faster convergence than classical algorithms but the computations of F can significantly slow down these methods. Moreover, the calculation cost of the function F defined in (38) is particularly high. As a consequence the restart scheme with (36) as a restart condition and the method of Alamo et al. are not as efficient as Figure 4 suggests. It appears that Algorithm 2 is a satisfying scheme as it provides a fast exponential decay which is proved theoretically and it is not affected by the expensive calculation cost of F.

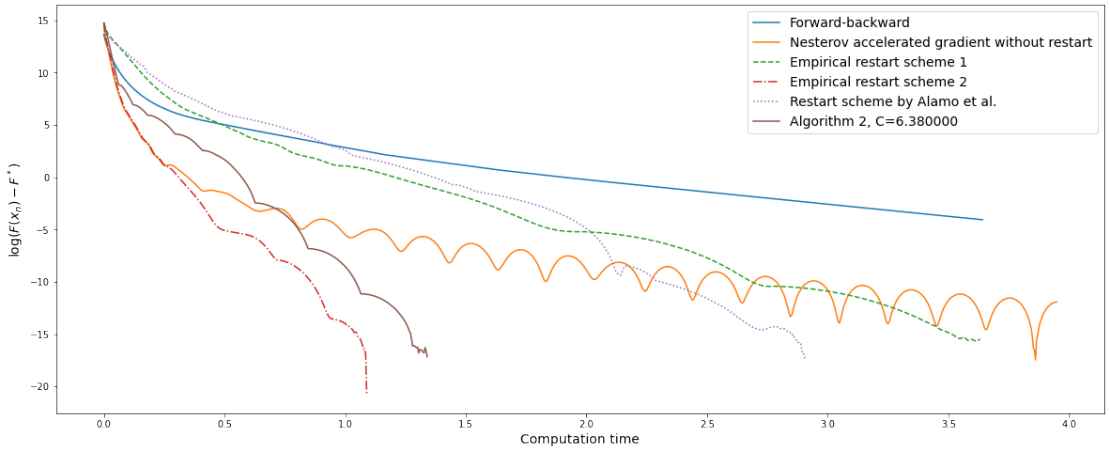

Figure 5: Comparison of several algorithms in terms of computation time for an inpainting problem.

4 Proofs

4.1 Sketch of proof

The proof of Theorem 1 is technical. Therefore we give a structure of it:

1. We show that there is at least one doubling step every T iterations for a well-chosen T.
   (a) We suppose that there is no doubling step from $j = s+1$ to $j = s+T$.
   (b) We exhibit the geometrical decrease of $(F(r_{j-1})-F(r_j))_{j\in[\![s+1,s+T]\!]}$ which represents the gain of the j-th execution of FISTA.
   (c) We apply Lemma 4 to show that we can find an upper bound of $\|g(r_{j-1})\|$ which depends on $F(r_{j-1}) - F(r_j)$ for all $j \in [\![s+1, s+T]\!]$.
   (d) We exploit the geometrical decrease $(F(r_{j-1})-F(r_j))_{j\in[\![s+1,s+T]\!]}$ to show that the exit condition is satisfied for $j = s+T$.
2. We use the first point to show that the number of iterations $\sum_{i=0}^{j} n_i$ is necessarily bounded by $2Tn_j$. The conclusion of Theorem 1 comes from Lemma 3 which gives an upper bound of n_j.

4.2 Proof of Theorem 1.

Let $C > 4$, $\varepsilon > 0$ and $T = 1 + \left\lceil \frac{\log\left(1+\frac{16}{C^2-16}\frac{2(F(r_0)-F^*)}{L\varepsilon^2}\right)}{\log\left(\frac{C^2}{4}-1\right)} \right\rceil$ where $\lceil\cdot\rceil$ is the ceiling function. We show that there is a doubling step at least every T iterations.

Assume that there is no doubling step from $j = s+1$ to $j = s+T$ where $s \geqslant 1$ which means that for a given $s \geqslant 1$:

$$\forall j \in [\![s+1, s+T]\!], \quad n_{j-1} > C\sqrt{\frac{L}{\tilde{\mu}_j}}, \tag{40}$$

and consequently:

$$\forall j \in [\![s+1, s+T]\!], \quad n_j = n_s. \tag{41}$$

Hence:

$$\begin{aligned}
\forall j \in [\![s+2, s+T]\!],\ \tilde{\mu}_j &= \min_{\substack{i \in \mathbb{N}^* \\ i<j}} \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_j)}{F(r_i) - F(r_j)} \\
&\leqslant \min_{\substack{i \in \mathbb{N}^* \\ s<i<j}} \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_j)}{F(r_i) - F(r_j)} \\
&\leqslant \min_{\substack{i \in \mathbb{N}^* \\ s<i<j}} \frac{4L}{{n_{i-1}}^2} \frac{F(r_{i-1}) - F(r_j)}{F(r_i) - F(r_j)} \\
&\leqslant \min_{\substack{i \in \mathbb{N}^* \\ s<i<j}} \frac{4L}{{n_s}^2} \frac{F(r_{i-1}) - F(r_j)}{F(r_i) - F(r_j)} \\
&\leqslant \frac{4L}{{n_s}^2} \min_{\substack{i \in \mathbb{N}^* \\ s<i<j}} \frac{F(r_{i-1}) - F(r_j)}{F(r_i) - F(r_j)},
\end{aligned} \tag{42}$$

as there is no doubling step for $j \geqslant s$. This then implies

$$\forall j \in [\![s+2, s+T]\!],\ \tilde{\mu}_j \leqslant \frac{4L}{{n_s}^2} \frac{F(r_{j-2}) - F(r_j)}{F(r_{j-1}) - F(r_j)}. \tag{43}$$

Combining (40) with (41) and (43) we get that:

$$n_s > C\sqrt{\frac{L}{\frac{4L}{{n_s}^2}\frac{F(r_{j-2})-F(r_j)}{F(r_{j-1})-F(r_j)}}} = n_s \frac{C}{2}\sqrt{\frac{F(r_{j-1}) - F(r_j)}{F(r_{j-2}) - F(r_j)}}. \tag{44}$$

This leads to the following inequality

$$F(r_{j-2}) - F(r_j) > \frac{C^2}{4}(F(r_{j-1}) - F(r_j)), \tag{45}$$

and then,

$$F(r_{j-2}) - F(r_{j-1}) > \left(\frac{C^2}{4} - 1\right)(F(r_{j-1}) - F(r_j)). \tag{46}$$

Since $C > 2$ we get that

$$F(r_{j-1}) - F(r_j) < \frac{4}{C^2 - 4}(F(r_{j-2}) - F(r_{j-1})). \tag{47}$$

We consider the case $j = s + 1$:

$$\begin{aligned}
\tilde{\mu}_{s+1} &= \min_{\substack{i \in \mathbb{N}^* \\ i < s+1}} \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_{s+1})}{F(r_i) - F(r_{s+1})} \\
&\leqslant \frac{4L}{(n_{s-1}+1)^2} \frac{F(r_{s-1}) - F(r_{s+1})}{F(r_s) - F(r_{s+1})} \\
&\leqslant \frac{4L}{(\frac{n_s}{2}+1)^2} \frac{F(r_{s-1}) - F(r_{s+1})}{F(r_s) - F(r_{s+1})} \\
&\leqslant \frac{16L}{{n_s}^2} \frac{F(r_{s-1}) - F(r_{s+1})}{F(r_s) - F(r_{s+1})},
\end{aligned} \tag{48}$$

as $n_s \leqslant 2n_{s-1}$.
By similar computations we get

$$F(r_s) - F(r_{s+1}) < \frac{16}{C^2 - 16}(F(r_{s-1}) - F(r_s)). \tag{49}$$

Since $C > 4$ we finally obtain the following inequalities

$$F(r_s) - F(r_{s+1}) < \frac{16}{C^2 - 16}(F(r_{s-1}) - F(r_s)). \tag{50}$$

$$\forall j \in [\![s+2, s+T]\!], \quad F(r_{j-1}) - F(r_j) < \frac{4}{C^2 - 4}(F(r_{j-2}) - F(r_{j-1})). \tag{51}$$

We introduce Lemma 4 which links the composite gradient mapping g to the function F. This lemma is proven in Section 4.8:

Lemma 4. *Let F be a function that can be written $F = f + h$ where f is a convex differentiable function having a L-Lipschitz gradient and h is a convex proper lower semicontinuous function. Then the sequence $(r_j)_{j \in \mathbb{N}}$ provided by Algorithm 2 satisfies*

$$\forall j \geqslant 1, \quad \frac{L}{2}\|g(r_{j-1})\|^2 \leqslant F(r_{j-1}) - F(r_j). \tag{52}$$

From Lemma 4 and inequalities (50) and (51) we obtain the following sequence of inequalities

$$\begin{aligned}
\frac{L}{2}\|g(r_{s+T-1})\|^2 &\leqslant F(r_{s+T-1}) - F(r_{s+T}) \\
&\leqslant \frac{4}{C^2-4}(F(r_{s+T-2}) - F(r_{s+T-1})) \\
&\leqslant \left(\frac{4}{C^2-4}\right)^{T-1}\left(\frac{16}{C^2-16}\right)(F(r_{s-1}) - F(r_s)) \\
&\leqslant \left(\frac{4}{C^2-4}\right)^{T-1}\left(\frac{16}{C^2-16}\right)(F(r_0) - F^*) \\
&\leqslant \left(\frac{4}{C^2-4}\right)^{\left\lceil \frac{\log\left(1+\frac{16}{C^2-16}\frac{2(F(r_0)-F^*)}{L\varepsilon^2}\right)}{\log\left(\frac{C^2}{4}-1\right)}\right\rceil}\left(\frac{16}{C^2-16}\right)(F(r_0) - F^*) \\
&\leqslant \left(\frac{4}{C^2-4}\right)^{\frac{\log\left(1+\frac{16}{C^2-16}\frac{2(F(r_0)-F^*)}{L\varepsilon^2}\right)}{\log\left(\frac{C^2}{4}-1\right)}}\left(\frac{16}{C^2-16}\right)(F(r_0) - F^*) \\
&\leqslant \frac{1}{1+\frac{16}{C^2-16}\frac{2(F(r_0)-F^*)}{L\varepsilon^2}}\left(\frac{16}{C^2-16}\right)(F(r_0) - F^*) \\
&\leqslant \frac{L\varepsilon^2}{2}.
\end{aligned}$$

As a consequence, if there are T consecutive steps of Algorithm 2 without doubling the number of iterations, then the exit condition $\|g(r_j)\|^2 \leqslant \varepsilon$ is eventually satisfied. This means that there is a doubling step at least every T steps and for all $s \geqslant 1$ there exists $j \in [\![s+1, s+T]\!]$ such that

$$n_{j-1} < C\sqrt{\frac{L}{\tilde{\mu}_j}}. \tag{53}$$

This implies that $n_j = 2n_{j-1}$. As $(n_j)_{j\in\mathbb{N}}$ is an increasing sequence, we get that $n_{s+T} \geqslant n_j = 2n_{j-1} \geqslant 2n_s$. And thus

$$n_s \leqslant \frac{n_{s+T}}{2}, \ \forall s \geqslant 1. \tag{54}$$

Let us rewrite j as $j = m + nT$ where $0 \leqslant m < T$ and $n \geqslant 0$. The increasing

nature of $(n_j)_{j\in\mathbb{N}}$ gives us that

$$\sum_{i=0}^{j} n_i = \sum_{i=0}^{m+nT} n_i = \sum_{i=0}^{m} n_i + \sum_{l=0}^{n-1}\sum_{i=1}^{T} n_{m+i+lT} \tag{55}$$

$$\leqslant Tn_m + T\sum_{l=1}^{n} n_{m+lT} = T\sum_{l=0}^{n} n_{m+lT} = T\sum_{l=0}^{n} n_{j-lT}. \tag{56}$$

According to equation (54) we have $n_{j-T} \leqslant \frac{n_j}{2}$ and therefore

$$n_{j-lT} \leqslant \left(\frac{1}{2}\right)^l n_j,\ \forall l \in [\![0, n]\!]. \tag{57}$$

We obtain the following inequalities

$$\sum_{i=0}^{j} n_i \leqslant T\sum_{l=0}^{n} n_{j-lT} \leqslant T\sum_{l=0}^{n}\left(\frac{1}{2}\right)^l n_j \leqslant T\sum_{l=0}^{\infty}\left(\frac{1}{2}\right)^l n_j = 2Tn_j. \tag{58}$$

From (58) and Lemma 3 we get that for $j > 0$

$$\sum_{i=0}^{j} n_i \leqslant 2Tn_j \leqslant 4C\sqrt{\frac{L}{\mu}}T \leqslant 4C\sqrt{\frac{L}{\mu}}\left(1 + \left\lceil \frac{\log\left(1 + \frac{16}{C^2-16}\frac{2(F(r_0)-F^*)}{L\varepsilon^2}\right)}{\log\left(\frac{C^2}{4}-1\right)}\right\rceil\right) \tag{59}$$

$$\leqslant \frac{4C}{\log\left(\frac{C^2}{4}-1\right)}\sqrt{\frac{L}{\mu}}\left(2\log\left(\frac{C^2}{4}-1\right) + \log\left(1 + \frac{16}{C^2-16}\frac{2(F(r_0)-F^*)}{L\varepsilon^2}\right)\right). \tag{60}$$

□

4.3 Proof of Corollary 1.

Let $F = f + h$ be a function with a non empty set of minimizers X^* where f is a convex differentiable function with L-Lipschitz gradient and h is a convex function. We suppose that F has a Łojasiewicz property with an exponent $\frac{1}{2}$. Let $(r_j)_{j\in\mathbb{N}}$ and $(n_j)_{j\in\mathbb{N}}$ be the sequences provided by Algorithm 2 with $C > 4$ and $\varepsilon > 0$.

We consider the case in which the exit condition $\|g(r_j)\| \leqslant \varepsilon$ is satisfied at first for at least $8C\sqrt{\frac{L}{\mu}}$ iterations. We define the function $\psi_\mu : \mathbb{R}_+^* \to \left(8C\sqrt{\frac{L}{\mu}}, +\infty\right)$ such that:

$$\psi_\mu : \gamma \mapsto \frac{4C}{\log\left(\frac{C^2}{4}-1\right)}\sqrt{\frac{L}{\mu}}\left(2\log\left(\frac{C^2}{4}-1\right) + \log\left(1 + \frac{16}{C^2-16}\frac{2(F(r_0)-F^*)}{L\gamma}\right)\right). \tag{61}$$

According to Theorem 1, the number of iterations required to ensure that $\|g(r_j)\| \leqslant \varepsilon$ satisfies:

$$\sum_{i=0}^{j} n_i \leqslant \psi_\mu(\varepsilon^2). \tag{62}$$

As ψ_μ is a strictly decreasing function and $\sum_{i=0}^{j} n_i > 8C\sqrt{\frac{L}{\mu}}$ we can write that:

$$\psi_\mu^{-1}\left(\sum_{i=0}^{j} n_i\right) \geqslant \varepsilon^2, \tag{63}$$

where ψ_μ^{-1} is the inverse function of ψ_μ. By applying Lemma 1 we get that:

$$\begin{aligned} F(r_j^+) - F^* &\leqslant \frac{2L^2}{\mu}\|g(r_j)\|^2 && (64)\\ &\leqslant \frac{2L^2\varepsilon^2}{\mu} && (65)\\ &\leqslant \frac{2L^2}{\mu}\psi_\mu^{-1}\left(\sum_{i=0}^{j} n_i\right). && (66) \end{aligned}$$

Elementary computations give us that:

$$\psi_\mu^{-1} : n \mapsto \frac{2}{L}\frac{16}{C^2-16}\frac{1}{e^{-2\log(\frac{C^2}{4}-1)}e^{\frac{\log(\frac{C^2}{4}-1)}{4C}\sqrt{\frac{\mu}{L}}n} - 1}(F(r_0) - F^*), \tag{67}$$

and thus we get:

$$F(r_j^+) - F^* \leqslant \frac{4L}{\mu}\frac{16}{C^2-16}\frac{1}{e^{-2\log(\frac{C^2}{4}-1)}e^{\frac{\log(\frac{C^2}{4}-1)}{4C}\sqrt{\frac{\mu}{L}}\sum_{i=0}^{j} n_i} - 1}(F(r_0) - F^*). \tag{68}$$

We can then conclude that

$$F(r_j^+) - F^* = \mathcal{O}\left(e^{-\frac{\log(\frac{C^2}{4}-1)}{4C}\sqrt{\frac{\mu}{L}}\sum_{i=0}^{j} n_i}\right). \tag{69}$$

Minimizing the function $C \mapsto \frac{\log(\frac{C^2}{4}-1)}{4C}$ gives us the optimal value $\hat{C} \approx 6.38$. This choice leads to the following rate:

$$F(r_j^+) - F^* = \mathcal{O}\left(e^{-\frac{1}{12}\sqrt{\frac{\mu}{L}}\sum_{i=0}^{j} n_i}\right). \tag{70}$$

□

4.4 Proof of Proposition 1

(i) It is well known (see [12, 23, 2]) that as f is a convex differentiable function having a L-Lipschitz gradient and h is a convex proper lower semicontinuous function, the sequence $(x_k)_{k\in\mathbb{N}}$ provided by FISTA algorithm (1) satisfies

$$\forall k \in \mathbb{N}^*, \quad F(x_k) - F^* \leqslant \frac{2L}{(k+1)^2}\|x_0 - x^*\|^2, \tag{71}$$

where x^* is any minimizer of F. This inequality is true for all $x^* \in X^*$ so (71) can be rewritten

$$\forall k \in \mathbb{N}^*, \quad F(x_k) - F^* \leqslant \frac{2L}{(k+1)^2}d(x_0, X^*)^2. \tag{72}$$

Furthermore, F satisfies the growth condition $\mathcal{G}^2_\mu$ so we can conclude by combining (5) and (72).

(ii) We first prove the following claim. Let $y \in \mathbb{R}^N$. Then we have

$$\forall x \in \mathbb{R}^N, \quad F(y^+) + \frac{L}{2}\|y^+ - x\|^2 \leqslant F(x) + \frac{L}{2}\|x - y\|^2. \tag{73}$$

By definition of the proximal operator (13), y^+ is the unique minimizer of the function defined by

$$x \mapsto h(x) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x - y\|^2. \tag{74}$$

As this function is L-strongly convex we get that for all $x \in \mathbb{R}^N$,

$$h(y^+)+\langle y^+-y, \nabla f(y)\rangle+\frac{L}{2}\|y^+-y\|^2+\frac{L}{2}\|y^+-x\|^2 \leqslant h(x)+\langle x-y, \nabla f(y)\rangle+\frac{L}{2}\|x-y\|^2. \tag{75}$$

f has a L-Lipschitz gradient which implies that

$$\forall (x, y) \in \mathbb{R}^N \times \mathbb{R}^N, \quad f(x) \leqslant f(y) + \langle x - y, \nabla f(y)\rangle + \frac{L}{2}\|x - y\|^2. \tag{76}$$

Thus we get that

$$h(y^+) + f(y^+) - f(y) + \frac{L}{2}\|y^+ - x\|^2 \leqslant h(x) + \langle x - y, \nabla f(y)\rangle + \frac{L}{2}\|x - y\|^2. \tag{77}$$

The convexity of f gives us that $f(x) \geqslant f(y) + \langle x - y, \nabla f(y)\rangle$ and then

$$F(y^+) + \frac{L}{2}\|y^+ - x\|^2 \leqslant F(x) + \frac{L}{2}\|x - y\|^2. \tag{78}$$

By applying this inequality to $y = y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1})$, $y^+ = x_{k+1}$ and $x = x_k$ for $k \geqslant 1$ we have

$$F(x_{k+1}) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \leqslant F(x_k) + \frac{L}{2}\left(\frac{k-1}{k+2}\right)^2 \|x_k - x_{k-1}\|^2 \tag{79}$$

$$\leqslant F(x_k) + \frac{L}{2}\|x_k - x_{k-1}\|^2. \tag{80}$$

Moreover, if we apply (73) to $y = x = x_0$ and $y^+ = x_1$ we get that

$$F(x_1) + \frac{L}{2}\|x_1 - x_0\|^2 \leqslant F(x_0). \tag{81}$$

This implies that

$$\forall k \geq 1, \quad F(x_k) + \|x_k - x_{k-1}\|^2 \leqslant F(x_0), \tag{82}$$

and thus we can conclude.
(iii) By rewriting the first claim of Proposition 1 we get that

$$\forall k \in \mathbb{N}^*, \quad F(x_k) - F^* \leqslant \frac{4L}{\mu(k+1)^2 - 4L}\left(F(x_0) - F(x_k)\right). \tag{83}$$

As a consequence, we have that for all $k \geqslant 2\sqrt{\frac{L}{\mu}}\sqrt{1 + \frac{1}{\gamma}} - 1$,

$$F(x_k) - F^* \leqslant \gamma\left(F(x_0) - F(x_k)\right). \tag{84}$$

The contraposite of this proposition leads us to the expected conclusion.

□

4.5 Proof of Lemma 1

Suppose that $F = f + h$ is a function with a non empty set of minimizers X^* where f is a convex differentiable function with L-Lipschitz gradient and h is a convex function. We consider that F satisfies $\mathcal{G}^2_\mu$ which is equivalent to say that F has a Łojasiewicz property with an exponent $\frac{1}{2}$. Therefore there exists $c > 0$ such that

$$\forall x \in \mathbb{R}^N, \quad F(x) - F^* \leqslant c d(0, \partial F(x))^2. \tag{85}$$

In particular, this assertion is true for $c = \frac{1}{2\mu}$.

Let $x \in \mathbb{R}^N$. By definition of the proximal operator (13), x^+ is the unique minimizer of the function defined by

$$z \mapsto h(z) + \frac{L}{2}\|z - x + \frac{1}{L}\nabla F(x)\|^2 \tag{86}$$

and thus x^+ satisfies

$$0 \in \partial h(x^+) + \left\{ L(x^+ - x) + \nabla f(x) \right\}. \tag{87}$$

As a consequence we get that

$$Lg(x) - \nabla f(x) + \nabla f(x^+) \in \partial F(x^+). \tag{88}$$

Moreover as f has a L-Lipschitz gradient we have

$$\begin{aligned} \|Lg(x) - \nabla f(x) + \nabla f(x^+)\| &\leqslant L\|g(x)\| + \|\nabla f(x^+) - \nabla f(x)\| && (89)\\ &\leqslant 2L\|g(x)\|. && (90) \end{aligned}$$

By combining these inequalities we conclude that

$$\begin{aligned} F(x^+) - F^* &\leqslant \frac{1}{2\mu} d(0, \partial F(x^+))^2 && (91)\\ &\leqslant \frac{1}{2\mu} \|Lg(x) - \nabla f(x) + \nabla f(x^*)\|^2 && (92)\\ &\leqslant \frac{2L^2}{\mu} \|g(x)\|^2. && (93) \end{aligned}$$

□

4.6 Proof of Lemma 2.

The sequence $(\tilde{\mu}_j)_{j \geqslant 2}$ is defined such that

$$\forall j \geqslant 2, \quad \tilde{\mu}_j = \min_{\substack{i \in \mathbb{N}^* \\ i<j}} \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_j)}{F(r_i) - F(r_j)}. \tag{94}$$

On the other hand, for all $i \in \mathbb{N}^*$ and $k \in \mathbb{N}^*$, we have:

$$F(r_i) - F^* > F(r_i) - F(r_{i+k}) = \gamma(F(r_{i-1}) - F(r_i)), \tag{95}$$

where $\gamma = \frac{F(r_i) - F(r_{i+k})}{F(r_{i-1}) - F(r_i)}$. Moreover, the third claim of Proposition 1 gives us that:

$$n_{i-1} < 2\sqrt{\frac{L}{\mu}} \sqrt{1 + \frac{1}{\gamma}} - 1. \tag{96}$$

Thus:

$$\mu < \frac{4L\left(1 + \frac{F(r_{i-1}) - F(r_i)}{F(r_i) - F(r_{i+k})}\right)}{(n_{i-1}+1)^2} = \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_{i+k})}{F(r_i) - F(r_{i+k})}. \tag{97}$$

As a consequence, $\mu < \tilde{\mu}_j$. Furthermore, we have

$$\forall j \geqslant 2, \quad \tilde{\mu}_{j+1} = \min_{\substack{i \in \mathbb{N}^* \\ i<j+1}} \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_{j+1})}{F(r_i) - F(r_{j+1})} \tag{98}$$

$$\leqslant \min_{\substack{i \in \mathbb{N}^* \\ i<j}} \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_{j+1})}{F(r_i) - F(r_{j+1})}. \tag{99}$$

The second claim of Proposition 1 implies that $F(r_{j+1}) \leqslant F(r_j)$. As a consequence the function defined by $y \mapsto \frac{F(r_{i-1})-y}{F(r_i)-y}$ is an increasing homographic function and we get that

$$\forall j \geqslant 2, \quad \tilde{\mu}_{j+1} \leqslant \min_{\substack{i \in \mathbb{N}^* \\ i<j}} \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_{j+1})}{F(r_i) - F(r_{j+1})} \tag{100}$$

$$\leqslant \min_{\substack{i \in \mathbb{N}^* \\ i<j}} \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_j)}{F(r_i) - F(r_j)} \tag{101}$$

$$\leqslant \tilde{\mu}_j. \tag{102}$$

□

4.7 Proof of Lemma 3.

The sequence $(n_j)_{j \in \mathbb{N}}$ is defined such that for all $j \geqslant 2$, $n_j = 2n_{j-1}$ if the following condition is satisfied:

$$n_{j-1} \leqslant C\sqrt{\frac{L}{\tilde{\mu}_j}}. \tag{103}$$

The second claim of Proposition 1 implies that

$$n_{j-1} \leqslant C\sqrt{\frac{L}{\mu}}. \tag{104}$$

This inequality ensures $n_j \leqslant 2C\sqrt{\frac{L}{\mu}}$ if $j \geqslant 2$. For $j = 0$ and $j = 1$, $n_j \leqslant 2C \leqslant 2C\sqrt{\frac{L}{\mu}}$ and we get the final conclusion.

4.8 Proof of Lemma 4

Let $j \geqslant 1$. We can rewrite the inequality (73) for $x = y \in \mathbb{R}^N$:

$$\frac{L}{2}\|y^+ - y\|^2 \leqslant F(y) - F(y^+). \tag{105}$$

By setting $y = r_{j-1}$ and as $F(r_{j-1}^+) \leqslant F(r_j)$ we conclude that

$$\frac{L}{2}\|g(r_{j-1})\|^2 \leqslant F(r_{j-1}) - F(r_j). \tag{106}$$

□

Acknowledgements

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

References

[1] T. Alamo, P. Krupa, and D. Limon. Gradient based restart FISTA. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3936–3941. IEEE, 2019.

[2] T. Alamo, D. Limon, and P. Krupa. Restart FISTA with global linear convergence. pages 1969–1974, 2019.

[3] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1-2):5–16, 2009.

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

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

[6] H. Attouch, Z. Chbani, and H. Riahi. Fast convex optimization via time scaling of damped inertial gradient dynamics. 2019.

[7] H. Attouch, Z. Chbani, and H. Riahi. Rate of convergence of the Nesterov accelerated gradient method in the subcritical case $\alpha \leqslant 3$. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:2, 2019.

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

[9] J.-F. Aujol, C. Dossal, and A. Rondepierre. Convergence rates of the Heavy-Ball method for quasi-strongly convex optimization. working paper or preprint, Apr. 2020.

[10] J.-F. Aujol, C. Dossal, and A. Rondepierre. Convergence rates of the Heavy-Ball method with Lojasiewicz property. Research report, IMB - Institut de Mathématiques de Bordeaux ; INSA Toulouse ; UPS Toulouse, Sept. 2020.

[11] A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE transactions on image processing*, 18(11):2419–2434, 2009.

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

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

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

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

[16] G. H. Chen and R. T. Rockafellar. Convergence rates in Forward-Backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997.

[17] 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.

[18] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson. Global convergence of the Heavy-Ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE, 2015.

[19] S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

[20] S. Łojasiewicz. Sur la géométrie semi-et sous-analytique. In *Annales de l'institut Fourier*, volume 43, pages 1575–1595, 1993.

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

[22] Y. Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. In *Sov. Math. Dokl*, volume 27.

[23] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

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

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

[26] B. T. Polyak. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1, 1987.

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

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

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

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

[31] W. Su, S. Boyd, and E. J. Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *The Journal of Machine Learning Research*, 17(1):5312–5354, 2016.

[32] M. V. Zibetti, E. S. Helou, R. R. Regatte, and G. T. Herman. Monotone FISTA with variable acceleration for compressed sensing magnetic resonance imaging. *IEEE transactions on computational imaging*, 5(1):109–119, 2018.