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

October 5, 2021

Abstract

In this paper, we propose a restart scheme for FISTA (Fast Iterative Shrinking-Threshold Algorithm) [12]. This method which is a generalization of Nesterov's accelerated gradient algorithm [29] 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 knowledge on the growth 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 the efficiency of this method.

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

Let $N \in \mathbb{N}^*$ and consider a composite optimization problem:

$$\min_{x \in \mathbb{R}^N} F(x). \tag{1}$$

The objective function $F : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ belongs to the class $\mathcal{H}_L$ of composite functions: $F = f + h$ where f is a convex, differentiable function having a L-Lipschitz gradient and h is a convex function whose proximal operator is known. The set of minimizers of F denoted by X^* is assumed to be non empty.

For the class $\mathcal{H}_L$ of composite functions, a classical minimization algorithm is the Forward-Backward algorithm (FB) that 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, FISTA (Fast Iterative Shrinkage-Thresholding Algorithm) [12], based on the ideas of acceleration of Nesterov [29], ensures a better asymptotic bound: $F(x_k) - F^* = \mathcal{O}\left(\frac{1}{k^2}\right)$.

In many optimization problems, as for instance in statistics or in image processing, the objective function F naturally satisfies some additional geometry assumptions such as a quadratic growth condition:

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

which allows to reach better decay rates. Note that in the convex setting, the quadratic growth condition (2) is equivalent to a Łojasiewicz property with an exponent $\frac{1}{2}$ [26, 27, 14]. This set of functions includes the set of strongly convex functions but it is much larger. For example, it contains functions associated to the mean square problem and the LASSO [15, Corollary 9]. Note that the uniqueness of the minimizer of F is not required.

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 (2) and the Lipschitz constant L of ∇f [19]. It turns out that this exponential decay cannot be observed in many numerical experiments because the condition number $\kappa = \frac{\mu}{L} \ll 1$ 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 the number of iterations required to reach a precision ε is proportional to $\frac{L}{\mu}$ for FB while it is proportional to $\sqrt{\frac{L}{\mu}}$ for FISTA.

Assuming that F has a unique minimizer, the variation of Polyak's Heavy Ball method [32] introduced in [8] reaches a better decay rate: $\mathcal{O}\left(e^{-(2-\sqrt{2})\sqrt{\kappa}k}\right)$. Unfortunately, this scheme requires an accurate a priori estimation of the growth parameter μ to get this fast decay. 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 introduce a new algorithm to minimize efficiently composite functions of the class $\mathcal{H}_L$ having additionally the quadratic growth property (2). More precisely this new algorithm is based on an original restart rule of FISTA ensuring:

$$F(x_k) - F^* = \mathcal{O}\left(e^{-\frac{1}{12}\sqrt{\kappa}k}\right) \tag{3}$$

without any a priori knowledge of μ, and 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]. Compared to [2], we improve the decay rate and reduce the number of estimations of $F(x_k)$ during the iterations (notice that these evaluations may heavily impact the numerical cost and calculation time).

The article is organized as follows. Section 2 provides a state of the art of this optimization problem. 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 introduce some notations. Given a differentiable function f, ∇f denotes the gradient of f. Given a convex lower semicontinuous function h, ∂h denotes the convex subgradient of h defined by:

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

Given a vector x, $\|x\|$ and $\|x\|_1$ respectively denote its Euclidean norm and its 1-norm.

We define the proximal operator of $h : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ a convex lower semicontinuous function as follows:

$$\forall x \in \mathbb{R}^N, \quad \mathrm{prox}_h(x) = \underset{y \in \mathbb{R}^N}{\mathrm{argmin}}\ h(y) + \frac{1}{2}\|y - x\|^2. \tag{5}$$

In this paper we focus on the class $\mathcal{H}_L$ of composite functions defined as: $F = f + h$ where f is a convex, differentiable function having a L-Lipschitz gradient and h is a convex function whose proximal operator is known. The set X^* of minimizers of F is assumed to be non empty.

Definition 1 (Quadratic growth condition $\mathcal{G}_\mu^2$)**.** *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 a quadratic growth condition* $\mathcal{G}_\mu^2$ *for some* $\mu > 0$ *if it satisfies:*

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

Classically the quadratic growth condition $\mathcal{G}_\mu^2$ can be seen as a relaxation of the strong convexity. In the convex setting, the condition $\mathcal{G}_\mu^2$ is equivalent to a global Łojasiewicz property with an exponent $\frac{1}{2}$ [14, 19]: 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 $F^* = \inf F$. If F satisfies a quadratic growth condition $\mathcal{G}_\mu^2$ for some $\mu > 0$, then F has a global Łojasiewicz property with an exponent $\frac{1}{2}$:

$$\forall x \in \mathbb{R}^N, \quad 2\mu \left(F(x) - F^*\right) \leqslant d(0, \partial F(x))^2. \tag{7}$$

To set some definitions we say that:

Definition 2. *An optimization algorithm provides a fast exponential decay on the class of functions* F *satisfying a growth condition* $\mathcal{G}_\mu^2$ *with parameter* μ *and having a* L*-Lipschitz gradient if there exists a constant* $K > 0$ *independent of* μ *and* L *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 3. *An optimization algorithm provides a low exponential decay on the class of functions* F *satisfying a growth condition* $\mathcal{G}_\mu^2$ *with parameter* μ *and having a* L*-Lipschitz gradient if there exists a constant* $K > 0$ *independent of* μ *and* L *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 $\frac{\mu}{L} \ll \sqrt{\frac{\mu}{L}}$.

2.2 Literature review

Let $L > 0$. Let $F : \mathbb{R}^n \to \mathbb{R}$ be a composite function of the class $\mathcal{H}_L$ satisfying additionally a growth condition for some parameter $\mu > 0$, namely μ-strong convexity or the quadratic growth condition $\mathcal{G}_\mu^2$. The Lipschitz constant L is assumed to be known.

A classical method to minimize such a function is the Forward-Backward algorithm which is an adaptation of Gradient Descent method. If F is μ-strongly

convex or satisfies $\mathcal{G}_\mu^2$ for some $\mu > 0$, this scheme provides a low exponential decay $F(x_k) - F^* = \mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$ [19].

Polyak introduces in [32] the Heavy-Ball method which ensures a fast convergence for μ-strongly convex functions that are twice differentiable. There exist many variations of this scheme and [8] proposes a Heavy-Ball method adapted to composite functions satisfying $\mathcal{G}_\mu^2$ for some $\mu > 0$. This scheme provides a fast exponential decay $\mathcal{O}\left(e^{-(2-\sqrt{2})\sqrt{\frac{\mu}{L}}k}\right)$ for this set of functions under a uniqueness of minimizers assumption. This method requires a prior estimate of the growth parameter μ.

Restarting schemes of FISTA [12] are efficient methods in this setting. FISTA is an inertial first order algorithm adapted from Nesterov accelerated gradient introduced in [29] which ensures that $F(x_k) - F^* = \mathcal{O}(k^{-2})$. Inertia generated in this scheme allows a fast convergence but also produces oscillations. Restarting FISTA is equivalent to set inertia to zero which helps to reduce oscillating behavior. A classical strategy introduced in [29] is to restart the algorithm at regular intervals. [28] show that restarting Nesterov accelerated gradient every $\lfloor 2e\sqrt{\frac{L}{\mu}} \rfloor$ iterations ensures that $F(x_k) - F^* = \mathcal{O}\left(e^{-\frac{1}{e}\sqrt{\frac{\mu}{L}}k}\right)$ for μ-strongly convex functions. This restart scheme and its convergence rate can be extended to FISTA and composite functions satisfying $\mathcal{G}_\mu^2$ [31, 28]. Note that an estimate of the growth parameter μ is required. This may be restrictive in numerical experiments since the growth parameter is rarely known.

O'Donoghue and Candès propose heuristic restart rules for accelerated gradient method and FISTA in [31]. These rules are based on empirical observations and they give highly effective numerical results especially for functions satisfying $\mathcal{G}_\mu^2$. However, no improved convergence rates was theoretically found so far.

Adaptive restart schemes take advantage of each iteration to estimate the geometry of F. This approach enables to fit the parameters of the algorithm progressively. [30, 22, 23] propose non restart schemes based on this strategy to compute an estimate of the strong convexity parameter. [17] provides an adaptive restart scheme for FISTA for the set of functions satisfying $\mathcal{G}_\mu^2$ which builds a sequence of estimates of the growth parameter μ. It requires a prior estimate μ_0 and the convergence rate of the method is given by $\mathcal{O}\left(e^{-\frac{\sqrt{2}-1}{2\sqrt{e}(2-\sqrt{\frac{\mu}{\mu_0}})}\sqrt{\frac{\mu}{L}}k}\right)$ if $\mu_0 \geqslant \mu$. As this rate significantly depends on $\frac{\mu}{\mu_0}$ this method might be less effective if μ_0 is highly overestimated. This rate is faster than $\mathcal{O}\left(e^{-\frac{1}{12}\sqrt{\frac{L}{\mu}}}\right)$ as long as $\mu_0 < 4\mu$.

Alamo et al. introduce in [2] and [1] adaptive restart schemes for FISTA under a $\mathcal{G}_\mu^2$ assumption on F where no prior estimate of $\mu > 0$ is required. The method

proposed in [2] provides a fast exponential decay $\mathcal{O}\left(e^{-\frac{1}{16}\sqrt{\frac{\mu}{L}}k}\right)$ and it relies on computations of $F(x_k)$ at least at half of the iterations. There exist techniques that reduce the computational cost of these evaluations. In [21] the authors introduce FASTA which is an adaptation of FISTA optimized for composite functions satisfying $\mathcal{H}_L$ such that $f(x) = \tilde{f}(Ax)$ for some function $\tilde{f}$ and some operator A. In this setting, this scheme reduces the computational cost of f and therefore of F. However in some common cases computing F is expensive despite these strategies. In these cases additional calculations may significantly slow down the method. The scheme introduced in [1] is based on gradient information and therefore it does not require to compute F. It provides the convergence rate $\mathcal{O}\left(e^{-\frac{1}{4e(1+\sqrt{\mu+1})}\frac{\mu}{L}k}\right)$ which can be seen as a low exponential decay as $e^{-\frac{1}{4e(1+\sqrt{\mu+1})}\frac{\mu}{L}k} > e^{-\frac{1}{8e}\frac{\mu}{L}k}$ for all $k > 0$.

Table 1: Convergence rates of classical algorithms for a possibly non differentiable function F satisfying $\mathcal{H}_L$ and $\mathcal{G}_\mu^2$ for some $L > 0$ and $\mu > 0$.

Algorithm	**References**	**Convergence rate**	**Limitations**
Forward-Backward	Garrigos et al. [19]	$\mathcal{O}\left(e^{-\frac{\mu}{L}k}\right)$	-
Heavy-Ball variation	Aujol et al. [8]	$\mathcal{O}\left(e^{-(2-\sqrt{2})\sqrt{\frac{\mu}{L}}k}\right)$	Requires an estimate of μ and uniqueness of minimizer
Optimal FISTA restart	Necoara et al. [28]	$\mathcal{O}\left(e^{-\frac{1}{e}\sqrt{\frac{\mu}{L}}k}\right)$	Requires an estimate of μ
Empirical FISTA restart	O'Donoghue and Candès [31]	$\mathcal{O}(k^{-2})$	Numerically fast but no improved theoretical convergence rate
FISTA restart by Fercoq and Qu [17]	Fercoq and Qu [17]	$\mathcal{O}\left(e^{-\frac{\sqrt{2}-1}{2\sqrt{e}(2-\sqrt{\frac{\mu}{\mu_0}})}\sqrt{\frac{\mu}{L}}k}\right)$	Requires an estimate μ_0 of μ
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
Gradient based FISTA restart	Alamo et al. [1]	$\mathcal{O}\left(e^{-\frac{1}{4e(1+\sqrt{\mu+1})}\frac{\mu}{L}k}\right)$	-
Adaptive restart	Section 3.1	$\mathcal{O}\left(e^{-\frac{1}{12}\sqrt{\frac{\mu}{L}}k}\right)$	-

We propose a 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 composite function F satisfying $\mathcal{H}_L$ and the growth condition $\mathcal{G}_\mu^2$ for some $L > 0$

and $\mu > 0$. We give a convergence rate and we describe the underlying strategy.

3.1 Adaptive FISTA restart scheme

Let us introduce some notations to simplify the writing of the algorithm. Let:

$$y^+ = \text{prox}_{\frac{1}{L}h}(y - \frac{1}{L}\nabla f(y)) \tag{8}$$

be the vector given by a step of the Forward-Backward algorithm on y with $s = \frac{1}{L}$. The composite gradient mapping g is then defined as follows:

$$g(y) = L(y - y^+). \tag{9}$$

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. Algorithm 2 relies on a few sequences:

- the sequence $(r_j)_{j\in\mathbb{N}}$ corresponds to the iterates of the algorithm. For all $j > 0$, r_j is the output of the jth execution 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:

Let F be a function satisfying $\mathcal{H}_L$ and $\mathcal{G}_\mu^2$ for some $L > 0$ and $\mu > 0$. Then for all $x \in \mathbb{R}^N$ we have:

$$F(x^+) - F^* \leqslant \frac{2}{\mu}\|g(x)\|^2. \tag{10}$$

The exit criteria combined to the inequality (10) enable the algorithm to ensure that the vector r_j satisfies:

$$F(r_j^+) - F^* \leqslant \frac{2\varepsilon^2}{\mu}, \tag{11}$$

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**
      $n_j = 2n_{j-1}$
    else
      $n_j = n_{j-1}$
    end if
  until $\|g(r_j)\| \leqslant \varepsilon$
  return $r = r_j$

Let F be a function satisfying $\mathcal{H}_L$ and $\mathcal{G}_\mu^2$ for some $L > 0$ and $\mu > 0$. 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 $1 + \sum_{i=0}^{j} n_i$ required to guarantee $\|g(r_j)\| \leqslant \varepsilon$ is bounded and satisfies

$$\sum_{i=0}^{j} n_i \leqslant \frac{4C}{\log\left(\frac{C^2}{4}-1\right)} \sqrt{\frac{L}{\mu}} \left(2\log\left(\frac{C^2}{4}-1\right) + \log\left(1 + \frac{16}{C^2-16}\frac{2L(F(r_0)-F^*)}{\varepsilon^2}\right)\right). \tag{12}$$

Let F be a function satisfying $\mathcal{H}_L$ and $\mathcal{G}_\mu^2$ for some $L > 0$ and $\mu > 0$. 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

$$F(r_j^+) - F^* = \mathcal{O}\left(e^{-\frac{\log\left(\frac{C^2}{4}-1\right)}{4C}\sqrt{\frac{\mu}{L}}\sum_{i=0}^{j} n_i} \right). \tag{13}$$

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

$$F(r_j^+) - F^* = \mathcal{O}\left(e^{-K\sqrt{\frac{\mu}{L}}\sum_{i=0}^{j} n_i} \right). \tag{14}$$

Section 3.1 states that Algorithm 2 provides asymptotically a fast exponential decay. This convergence rate is faster than any method when considering a function F satisfying $\mathcal{H}_L$ and $\mathcal{G}_\mu^2$ where the parameter μ cannot be estimated. In this setting, Forward-Backward algorithm provides a low exponential decay and FISTA has no improved theoretical convergence rate. The variation of Heavy-Ball method introduced in [8], the FISTA restart scheme introduced in [17] and fixed restart of FISTA require to estimate the growth parameter to ensure a fast exponential decay.

The restart scheme introduced by Alamo et al. in [2] ensures the following rate:

$$F(r_j^+) - F^* = \mathcal{O}\left(e^{-\frac{1}{16}\sqrt{\frac{\mu}{L}}\sum_{i=0}^{j} n_i} \right), \tag{15}$$

which is slightly slower than (14). The gradient based FISTA restart proposed in [1] guarantees the theoretical rate $\mathcal{O}\left(e^{-\frac{1}{4e(1+\sqrt{\mu+1})}\frac{\mu}{L}k}\right)$ which can be considered

as an low exponential decay. As a consequence, this decay may be significantly slower than the decay of Algorithm 2 when $\frac{\mu}{L} \ll \sqrt{\frac{\mu}{L}}$ which is the case in numerous practical examples.

We recall that convergence results such as (14) and (15) are worst-case bounds. Consequently a faster theoretical rate does not necessarily guarantee a faster algorithm in general.

3.2 Structure of the algorithm

Algorithm 2 is an adaptive 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 (6) 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. This proposition is adapted from [2, Property 2] and introduces slightly more general claims leading to the estimate $\tilde{\mu}_j$ (24).

Proposition 1. *Let F be a function satisfying $\mathcal{H}_L$ and $\mathcal{G}_\mu^2$ for some $L > 0$ and $\mu > 0$. Then the sequence $(x_k)_{k\in\mathbb{N}}$ provided by Algorithm 1 satisfies*

$$\textit{(i)}\ \forall k \in \mathbb{N}^*,\ F(x_k) - F^* \leqslant \frac{4L}{\mu(k+1)^2}\left(F(x_0) - F^*\right), \tag{16}$$

$$\textit{(ii)}\ \forall k \in \mathbb{N}^*,\ F(x_k) \leqslant F(x_0), \tag{17}$$

$$\textit{(iii)}\ \forall k \in \mathbb{N},\ F(x_k) - F^* > \gamma\left(F(x_0) - F(x_k)\right) \implies (k+1)^2 < \frac{4L}{\mu}\left(1 + \frac{1}{\gamma}\right). \tag{18}$$

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:

$$\forall j \in \mathbb{N}^*,\ F(r_j) - F^* \leqslant \frac{4L}{\mu(n_{j-1}+1)^2}\left(F(r_{j-1}) - F^*\right), \tag{19}$$

which is equivalent to:

$$\forall j \in \mathbb{N}^*,\ \mu \leqslant \frac{4L}{(n_{j-1}+1)^2}\frac{F(r_{j-1}) - F^*}{F(r_j) - F^*}. \tag{20}$$

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 (19) and by considering a third point r_{j+1}. For all $j \in \mathbb{N}^*$,

$$F(r_j) - F^* > F(r_j) - F(r_{j+1}) = \gamma(F(r_{j-1}) - F(r_j)), \tag{21}$$

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:

$$\mu \leqslant \frac{4L}{(n_{j-1}+1)^2}\left(1 + \frac{1}{\gamma}\right). \tag{22}$$

Thus,

$$\forall j \in \mathbb{N}^*,\ \mu \leqslant \frac{4L}{(n_{j-1}+1)^2}\frac{F(r_{j-1}) - F(r_{j+1})}{F(r_j) - F(r_{j+1})}. \tag{23}$$

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 $(\tilde{\mu}_j)_{j\geqslant 2}$:

$$\forall j \geqslant 2, \quad \tilde{\mu}_j = \min_{\substack{i\in\mathbb{N}^* \\ i<j}} \frac{4L}{(n_{i-1}+1)^2}\frac{F(r_{i-1}) - F(r_j)}{F(r_i) - F(r_j)}. \tag{24}$$

This sequence satisfies the following property proven in Section 4.6: Let F be a function satisfying $\mathcal{H}_L$ and $\mathcal{G}^2_\mu$ for some $L > 0$ and $\mu > 0$. Then the sequence $(\tilde{\mu}_j)_{j\geqslant 2}$ provided by Algorithm 2 satisfies

$$\forall j \geqslant 2, \quad \tilde{\mu}_j \geqslant \tilde{\mu}_{j+1} > \mu. \tag{25}$$

It is worth noticing that this strategy could be adapted to other schemes than FISTA as long as theoretical convergence bounds are known. The estimate of the growth parameter indeed relies only on (16).

Once the growth parameter μ is estimated it is possible to set the number of iterations accordingly. It is known [29, 28, 31] 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. Its proof can be found in Section 4.7. Let F be a function satisfying $\mathcal{H}_L$ and $\mathcal{G}_\mu^2$ for some $L > 0$ and $\mu > 0$. Then the sequence $(n_j)_{j\in\mathbb{N}}$ provided by Algorithm 2 satisfies

$$\forall j \in \mathbb{N}, \quad n_j \leqslant 2C\sqrt{\frac{L}{\mu}}. \tag{26}$$

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

$$\forall j \in \mathbb{N}, \quad n_j \leqslant \gamma C\sqrt{\frac{L}{\mu}}. \tag{27}$$

The fastest asymptotical convergence rate is obtained for the choice $\gamma = 2$ (Algorithm 2).

3.3 Numerical experiments

In this section we illustrate the convergence results stated in previous sections with numerical experiments. We compare Algorithm 2 to the following set of methods:

1. The Forward-backward algorithm,
2. FISTA [12],
3. Empirical restart scheme of FISTA by O'Donoghue and Candès [31] with the restart condition:
$$F(x_k) > F(x_{k-1}), \tag{28}$$
4. Empirical restart scheme of FISTA by O'Donoghue and Candès [31] with the restart condition:
$$\langle g(y_{k-1}), x_k - x_{k-1}\rangle > 0, \tag{29}$$
5. Adaptive restart scheme introduced by Alamo et al. in [2],
6. Gradient based restart scheme introduced by Alamo et al. in [1].

Note that methods 3, 4, 5 and 6 are respectively refered as *Empirical restart scheme 1*, *Empirical restart scheme 2*, *Restart scheme 1 by Alamo et al.* and *Restart scheme 2 by Alamo et al.* in the following figures.

3.3.1 L^1-regularized least squares problem

The problem considered reads as follows:

$$\min_{x\in\mathbb{R}^N} F(x) = \frac{1}{2}\|Ax-b\|^2 + \rho\|x\|_1, \tag{30}$$

where $A \in \mathcal{M}_{N,N}(\mathbb{R})$, $b \in \mathbb{R}^N$, $\rho > 0$ and $N \geqslant 1$. The function F is a composite function satisfying $\mathcal{H}_L$ where $L > 0$ can easily be computed. The L^1-regularization prevents F to be strongly convex but it satisfies the assumption $\mathcal{G}^2_\mu$ for some $\mu > 0$. There exists no explicit formula for the growth parameter μ. Experiments are carried out by setting A as a random matrix of size $N \times N$. Results are provided for $N = 1000$ which ensures that the problem is ill-conditioned. The computational cost of F is negligible as optimization techniques such as FASTA [21] or speculative estimations [18] can be applied.

Figure 1 shows the decrease of the error according to the number of iterations. One can observe that Algorithm 2 and the restart scheme introduced by Alamo et al. in [2] provide similar precision with equal number of iterations. In this example additional computations of F are not expensive and consequently both methods perform similarly. This figure also illustrates the low exponential decay provided by Gradient based restart [1] as this scheme is significantly slower. Note that both empirical restart schemes provide significantly fast decay.

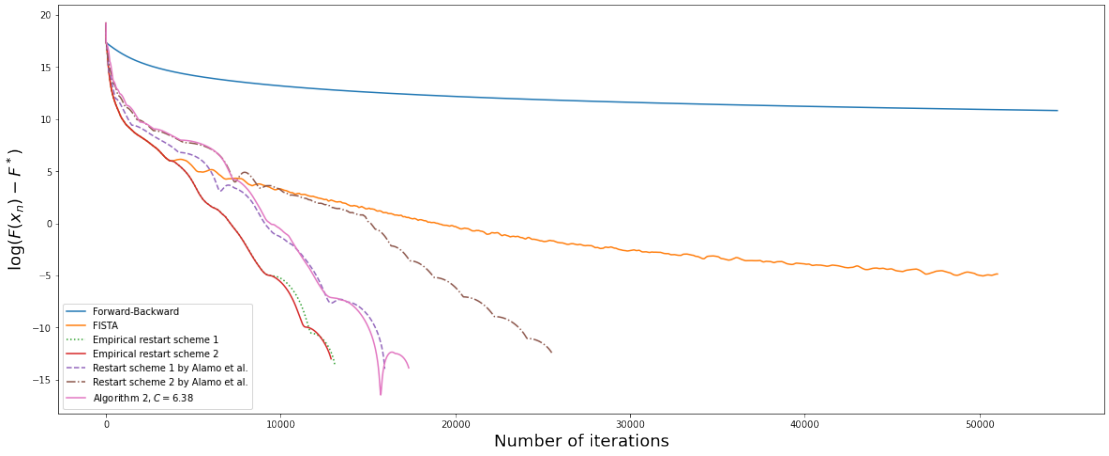

Figure 1: Comparison of several algorithms in terms of iterations for a L^1-regularized least squares problem with $N = 1000$.

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:

$$\min_x F(x) = \frac{1}{2}\|Mx - y\|^2 + \lambda\|Tx\|_1, \tag{31}$$

where T is an orthogonal transformation ensuring that Tx^0 is sparse. In this example, x^0 is piecewise smooth so T is set as an orthogonal wavelet transform. In this setting, the objective function F satisfies $\mathcal{H}_L$ and $\mathcal{G}_\mu^2$ where $L = 1$ and $\mu > 0$ cannot be computed directly.

Figure 2: Example of image inpainting: the damaged image y is on the left and an approximation of the solution of (31) is on the right.

This inpainting problem (31) is an example of a highly expensive function F to compute. As shown in Figure 3 and Figure 4, this high cost significantly penalizes restart schemes that rely on regular computations of F. It appears that Algorithm 2 is still efficient in this situation as its additional computations of the cost function are reduced.

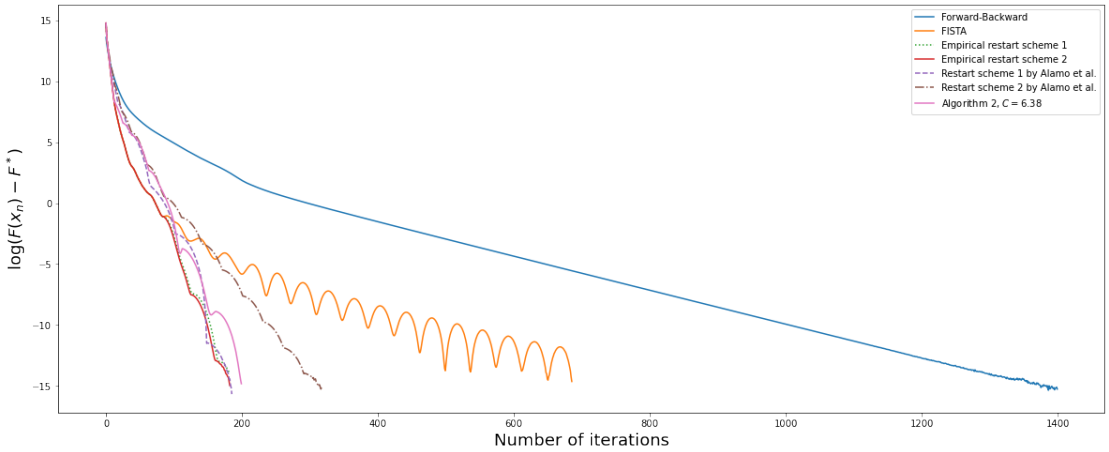

Figure 3: Comparison of several algorithms in terms of iterations for an inpainting problem.

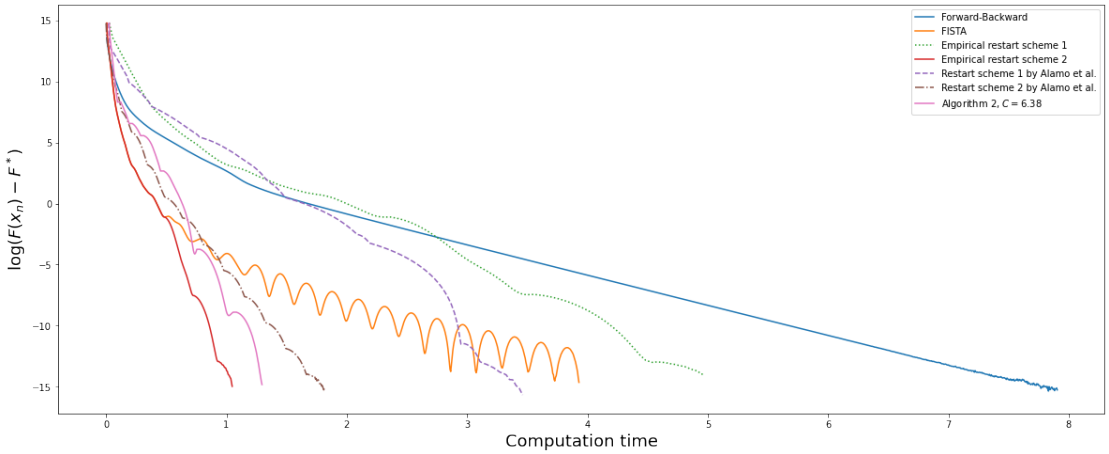

Figure 4: Comparison of several algorithms in terms of computation time for an inpainting problem.

4 Proofs

4.1 Sketch of proof

As the proof of Section 3.1 is technical, it is split into the following parts:

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 Section 4.2 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 Section 3.1 comes from Section 3.2 which gives an upper bound of n_j.

4.2 Proof of Section 3.1

Let $C > 4$ and $\varepsilon > 0$. We define $T = 1 + \left\lceil \frac{\log\left(1+\frac{16}{C^2-16}\frac{2(F(r_0)-F^*)}{L\varepsilon^2}\right)}{\log\left(\frac{C^2}{4}-1\right)} \right\rceil$ 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 $s \geqslant 1$:

$$\forall j \in [\![s+1, s+T]\!], \quad n_{j-1} > C\sqrt{\frac{L}{\tilde{\mu}_j}}, \tag{32}$$

and consequently:

$$\forall j \in [\![s+1, s+T]\!], \quad n_j = n_s. \tag{33}$$

Hence:

$$\begin{aligned}
\forall j \in [\![s+2, s+T]\!],\ \tilde{\mu}_j &= \min_{\substack{i\in\mathbb{N}^* \\ i<j}} \frac{4L}{(n_{i-1}+1)^2}\frac{F(r_{i-1})-F(r_j)}{F(r_i)-F(r_j)} \\
&\leqslant \min_{\substack{i\in\mathbb{N}^* \\ s<i<j}} \frac{4L}{(n_{i-1}+1)^2}\frac{F(r_{i-1})-F(r_j)}{F(r_i)-F(r_j)} \\
&\leqslant \min_{\substack{i\in\mathbb{N}^* \\ s<i<j}} \frac{4L}{{n_{i-1}}^2}\frac{F(r_{i-1})-F(r_j)}{F(r_i)-F(r_j)} \\
&\leqslant \min_{\substack{i\in\mathbb{N}^* \\ s<i<j}} \frac{4L}{{n_s}^2}\frac{F(r_{i-1})-F(r_j)}{F(r_i)-F(r_j)} \\
&\leqslant \frac{4L}{{n_s}^2} \min_{\substack{i\in\mathbb{N}^* \\ s<i<j}} \frac{F(r_{i-1})-F(r_j)}{F(r_i)-F(r_j)},
\end{aligned} \tag{34}$$

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

$$\forall j \in [\![s+2, s+T]\!],\ \tilde{\mu}_j \leqslant \frac{4L}{{n_s}^2} \frac{F(r_{j-2}) - F(r_j)}{F(r_{j-1}) - F(r_j)}. \tag{35}$$

Combining (32) with (33) and (35) we get that:

$$n_s > C \sqrt{\frac{L}{\frac{4L}{{n_s}^2} \frac{F(r_{j-2}) - F(r_j)}{F(r_{j-1}) - F(r_j)}}} = n_s \frac{C}{2} \sqrt{\frac{F(r_{j-1}) - F(r_j)}{F(r_{j-2}) - F(r_j)}}. \tag{36}$$

This leads to the following inequality

$$F(r_{j-2}) - F(r_j) > \frac{C^2}{4}(F(r_{j-1}) - F(r_j)), \tag{37}$$

and then,

$$F(r_{j-2}) - F(r_{j-1}) > \left(\frac{C^2}{4} - 1\right)(F(r_{j-1}) - F(r_j)). \tag{38}$$

Since $C > 2$ we get that

$$F(r_{j-1}) - F(r_j) < \frac{4}{C^2 - 4}(F(r_{j-2}) - F(r_{j-1})). \tag{39}$$

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

$$\begin{aligned}
\tilde{\mu}_{s+1} &= \min_{\substack{i \in \mathbb{N}^* \\ i < s+1}} \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_{s+1})}{F(r_i) - F(r_{s+1})} \\
&\leqslant \frac{4L}{(n_{s-1}+1)^2} \frac{F(r_{s-1}) - F(r_{s+1})}{F(r_s) - F(r_{s+1})} \\
&\leqslant \frac{4L}{(\frac{n_s}{2}+1)^2} \frac{F(r_{s-1}) - F(r_{s+1})}{F(r_s) - F(r_{s+1})} \\
&\leqslant \frac{16L}{{n_s}^2} \frac{F(r_{s-1}) - F(r_{s+1})}{F(r_s) - F(r_{s+1})},
\end{aligned} \tag{40}$$

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

$$F(r_s) - F(r_{s+1}) < \frac{16}{C^2 - 16}(F(r_{s-1}) - F(r_s)). \tag{41}$$

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

$$F(r_s) - F(r_{s+1}) < \frac{16}{C^2 - 16}(F(r_{s-1}) - F(r_s)). \tag{42}$$

$$\forall j \in [\![s+2, s+T]\!], \quad F(r_{j-1}) - F(r_j) < \frac{4}{C^2-4}(F(r_{j-2}) - F(r_{j-1})). \tag{43}$$

We introduce Section 4.2 which links the composite gradient mapping g to the function F. This lemma is proven in Section 4.8: Let F satisfy the assumption $\mathcal{H}_L$ for some $L > 0$. Then the sequence $(r_j)_{j\in\mathbb{N}}$ provided by Algorithm 2 satisfies

$$\forall j \geqslant 1, \quad \frac{1}{2L}\|g(r_{j-1})\|^2 \leqslant F(r_{j-1}) - F(r_j). \tag{44}$$

From Section 4.2 and inequalities (42) and (43) we obtain the following sequence of inequalities

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

As a consequence, if there are T consecutive steps of Algorithm 2 without doubling the number of iterations, then the exit condition $\|g(r_j)\| \leqslant \varepsilon$ 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

$$n_{j-1} < C\sqrt{\frac{L}{\tilde{\mu}_j}}. \tag{45}$$

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

$$n_s \leqslant \frac{n_{s+T}}{2}, \ \forall s \geqslant 1. \tag{46}$$

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

$$\sum_{i=0}^{j} n_i = \sum_{i=0}^{m+nT} n_i = \sum_{i=0}^{m} n_i + \sum_{l=0}^{n-1}\sum_{i=1}^{T} n_{m+i+lT} \tag{47}$$

$$\leqslant T n_m + T\sum_{l=1}^{n} n_{m+lT} = T\sum_{l=0}^{n} n_{m+lT} = T\sum_{l=0}^{n} n_{j-lT}. \tag{48}$$

According to equation (46) we have $n_{j-T} \leqslant \frac{n_j}{2}$ and therefore

$$n_{j-lT} \leqslant \left(\frac{1}{2}\right)^l n_j, \ \forall l \in [\![0, n]\!]. \tag{49}$$

We obtain the following inequalities

$$\sum_{i=0}^{j} n_i \leqslant T\sum_{l=0}^{n} n_{j-lT} \leqslant T\sum_{l=0}^{n}\left(\frac{1}{2}\right)^l n_j \leqslant T\sum_{l=0}^{\infty}\left(\frac{1}{2}\right)^l n_j = 2Tn_j. \tag{50}$$

From (50) and Section 3.2 we get that for $j > 0$

$$\sum_{i=0}^{j} n_i \leqslant 2Tn_j \leqslant 4C\sqrt{\frac{L}{\mu}}T \leqslant 4C\sqrt{\frac{L}{\mu}}\left(1 + \left\lceil \frac{\log\left(1 + \frac{16}{C^2-16}\frac{2L(F(r_0)-F^*)}{\varepsilon^2}\right)}{\log\left(\frac{C^2}{4}-1\right)} \right\rceil\right) \tag{51}$$

$$\leqslant \frac{4C}{\log\left(\frac{C^2}{4}-1\right)}\sqrt{\frac{L}{\mu}}\left(2\log\left(\frac{C^2}{4}-1\right) + \log\left(1 + \frac{16}{C^2-16}\frac{2L(F(r_0)-F^*)}{\varepsilon^2}\right)\right). \tag{52}$$

□

4.3 Proof of Section 3.1

Let F satisfy $\mathcal{H}_L$ and $\mathcal{G}_\mu^2$ for some $L > 0$ and $\mu > 0$. 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:

$$\psi_\mu : \gamma \mapsto \frac{4C}{\log\left(\frac{C^2}{4}-1\right)}\sqrt{\frac{L}{\mu}}\left(2\log\left(\frac{C^2}{4}-1\right) + \log\left(1 + \frac{16}{C^2-16}\frac{2L(F(r_0)-F^*)}{\gamma}\right)\right). \tag{53}$$

According to Section 3.1, the number of iterations required to ensure that $\|g(r_j)\| \leqslant \varepsilon$ satisfies:

$$\sum_{i=0}^{j} n_i \leqslant \psi_\mu(\varepsilon^2). \tag{54}$$

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

$$\psi_\mu^{-1}\left(\sum_{i=0}^{j} n_i\right) \geqslant \varepsilon^2, \tag{55}$$

where ψ_μ^{-1} is the inverse function of ψ_μ. By applying Section 3.1 we get that:

$$\begin{aligned} F(r_j^+) - F^* &\leqslant \frac{2}{\mu}\|g(r_j)\|^2 && (56)\\ &\leqslant \frac{2\varepsilon^2}{\mu} && (57)\\ &\leqslant \frac{2}{\mu}\psi_\mu^{-1}\left(\sum_{i=0}^{j} n_i\right). && (58) \end{aligned}$$

Elementary computations give us that:

$$\psi_\mu^{-1} : n \mapsto 2L\frac{16}{C^2-16}\frac{1}{e^{-2\log(\frac{C^2}{4}-1)}e^{\frac{\log(\frac{C^2}{4}-1)}{4C}\sqrt{\frac{\mu}{L}}n} - 1}(F(r_0) - F^*), \tag{59}$$

and thus we get:

$$F(r_j^+) - F^* \leqslant \frac{4L}{\mu}\frac{16}{C^2-16}\frac{1}{e^{-2\log(\frac{C^2}{4}-1)}e^{\frac{\log(\frac{C^2}{4}-1)}{4C}\sqrt{\frac{\mu}{L}}\sum_{i=0}^{j} n_i} - 1}(F(r_0) - F^*). \tag{60}$$

We can then conclude that

$$F(r_j^+) - F^* = \mathcal{O}\left(e^{-\frac{\log(\frac{C^2}{4}-1)}{4C}\sqrt{\frac{\mu}{L}}\sum_{i=0}^{j} n_i}\right). \tag{61}$$

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:

$$F(r_j^+) - F^* = \mathcal{O}\left(e^{-\frac{1}{12}\sqrt{\frac{\mu}{L}}\sum_{i=0}^{j} n_i}\right). \tag{62}$$

□

4.4 Proof of Proposition 1

(i) It is well known (see [12, 30, 2]) that as F satisfies $\mathcal{H}_L$ for some $L > 0$, the sequence $(x_k)_{k\in\mathbb{N}}$ provided by FISTA (Algorithm 1) satisfies

$$\forall k \in \mathbb{N}^*, \quad F(x_k) - F^* \leqslant \frac{2L}{(k+1)^2}\|x_0 - x^*\|^2, \tag{63}$$

where x^* is any minimizer of F. This inequality is true for all $x^* \in X^*$ so (63) can be rewritten

$$\forall k \in \mathbb{N}^*, \quad F(x_k) - F^* \leqslant \frac{2L}{(k+1)^2}d(x_0, X^*)^2. \tag{64}$$

Furthermore, F satisfies the growth condition $\mathcal{G}^2_\mu$ so we can conclude by combining (6) and (64).

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

$$\forall x \in \mathbb{R}^N, \quad F(y^+) + \frac{L}{2}\|y^+ - x\|^2 \leqslant F(x) + \frac{L}{2}\|x - y\|^2. \tag{65}$$

By definition of the proximal operator (5), y^+ is the unique minimizer of the function defined by

$$x \mapsto h(x) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x - y\|^2. \tag{66}$$

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

$$h(y^+)+\langle y^+-y, \nabla f(y)\rangle+\frac{L}{2}\|y^+-y\|^2+\frac{L}{2}\|y^+-x\|^2 \leqslant h(x)+\langle x-y, \nabla f(y)\rangle+\frac{L}{2}\|x-y\|^2. \tag{67}$$

f has a L-Lipschitz gradient which implies that

$$\forall (x, y) \in \mathbb{R}^N \times \mathbb{R}^N, \quad f(x) \leqslant f(y) + \langle x - y, \nabla f(y)\rangle + \frac{L}{2}\|x - y\|^2. \tag{68}$$

Thus we get that

$$h(y^+) + f(y^+) - f(y) + \frac{L}{2}\|y^+ - x\|^2 \leqslant h(x) + \langle x - y, \nabla f(y)\rangle + \frac{L}{2}\|x - y\|^2. \tag{69}$$

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

$$F(y^+) + \frac{L}{2}\|y^+ - x\|^2 \leqslant F(x) + \frac{L}{2}\|x - y\|^2. \tag{70}$$

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

$$F(x_{k+1}) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \leqslant F(x_k) + \frac{L}{2}\left(\frac{k-1}{k+2}\right)^2 \|x_k - x_{k-1}\|^2 \tag{71}$$

$$\leqslant F(x_k) + \frac{L}{2}\|x_k - x_{k-1}\|^2. \tag{72}$$

Moreover, if we apply (65) to $y = x = x_0$ and $y^+ = x_1$ we get that

$$F(x_1) + \frac{L}{2}\|x_1 - x_0\|^2 \leqslant F(x_0). \tag{73}$$

This implies that

$$\forall k \geq 1, \quad F(x_k) + \|x_k - x_{k-1}\|^2 \leqslant F(x_0), \tag{74}$$

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

$$\forall k \in \mathbb{N}^*, \quad F(x_k) - F^* \leqslant \frac{4L}{\mu(k+1)^2 - 4L}\left(F(x_0) - F(x_k)\right). \tag{75}$$

As a consequence, we have that for all $k > 0$ such that $(k+1)^2 \geqslant \frac{4L}{\mu}\left(1 + \frac{1}{\gamma}\right)$,

$$F(x_k) - F^* \leqslant \gamma\left(F(x_0) - F(x_k)\right). \tag{76}$$

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

□

4.5 Proof of Section 3.1

Suppose that F satisfies $\mathcal{H}_L$ and $\mathcal{G}_\mu^2$ for some $L > 0$ and $\mu > 0$. Then Section 2.1 states that there exists $c > 0$ such that

$$\forall x \in \mathbb{R}^N, \quad F(x) - F^* \leqslant c d(0, \partial F(x))^2. \tag{77}$$

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

Let $x \in \mathbb{R}^N$. By definition of the proximal operator (5), x^+ is the unique minimizer of the function defined by

$$z \mapsto h(z) + \frac{L}{2}\|z - x + \frac{1}{L}\nabla f(x)\|^2 \tag{78}$$

and thus x^+ satisfies

$$0 \in \partial h(x^+) + \left\{ L(x^+ - x) + \nabla f(x) \right\}. \tag{79}$$

As a consequence we get that

$$g(x) - \nabla f(x) + \nabla f(x^+) \in \partial F(x^+). \tag{80}$$

Moreover as f has a L-Lipschitz gradient we have

$$\begin{aligned} \|g(x) - \nabla f(x) + \nabla f(x^+)\| &\leqslant \|g(x)\| + \|\nabla f(x^+) - \nabla f(x)\| && (81)\\ &\leqslant 2\|g(x)\|. && (82) \end{aligned}$$

By combining these inequalities we conclude that

$$\begin{aligned} F(x^+) - F^* &\leqslant \frac{1}{2\mu} d(0, \partial F(x^+))^2 && (83)\\ &\leqslant \frac{1}{2\mu} \|g(x) - \nabla f(x) + \nabla f(x^*)\|^2 && (84)\\ &\leqslant \frac{2}{\mu} \|g(x)\|^2. && (85) \end{aligned}$$

□

4.6 Proof of Section 3.2

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

$$\forall j \geqslant 2, \quad \tilde{\mu}_j = \min_{\substack{i \in \mathbb{N}^* \\ i<j}} \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_j)}{F(r_i) - F(r_j)}. \tag{86}$$

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

$$F(r_i) - F^* > F(r_i) - F(r_{i+k}) = \gamma(F(r_{i-1}) - F(r_i)), \tag{87}$$

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:

$$n_{i-1} < 2\sqrt{\frac{L}{\mu}} \sqrt{1 + \frac{1}{\gamma}} - 1. \tag{88}$$

Thus:

$$\mu < \frac{4L\left(1 + \frac{F(r_{i-1}) - F(r_i)}{F(r_i) - F(r_{i+k})}\right)}{(n_{i-1}+1)^2} = \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_{i+k})}{F(r_i) - F(r_{i+k})}. \tag{89}$$

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

$$\forall j \geqslant 2, \quad \tilde{\mu}_{j+1} = \min_{\substack{i \in \mathbb{N}^* \\ i<j+1}} \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_{j+1})}{F(r_i) - F(r_{j+1})} \tag{90}$$

$$\leqslant \min_{\substack{i \in \mathbb{N}^* \\ i<j}} \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_{j+1})}{F(r_i) - F(r_{j+1})}. \tag{91}$$

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

$$\forall j \geqslant 2, \quad \tilde{\mu}_{j+1} \leqslant \min_{\substack{i \in \mathbb{N}^* \\ i<j}} \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_{j+1})}{F(r_i) - F(r_{j+1})} \tag{92}$$

$$\leqslant \min_{\substack{i \in \mathbb{N}^* \\ i<j}} \frac{4L}{(n_{i-1}+1)^2} \frac{F(r_{i-1}) - F(r_j)}{F(r_i) - F(r_j)} \tag{93}$$

$$\leqslant \tilde{\mu}_j. \tag{94}$$

□

4.7 Proof of Section 3.2

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:

$$n_{j-1} \leqslant C \sqrt{\frac{L}{\tilde{\mu}_j}}. \tag{95}$$

The second claim of Proposition 1 implies that

$$n_{j-1} \leqslant C \sqrt{\frac{L}{\mu}}. \tag{96}$$

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 Section 4.2

Let $j \geqslant 1$. We can rewrite the inequality (65) for $x = y \in \mathbb{R}^N$:

$$\frac{L}{2}\|y^+ - y\|^2 \leqslant F(y) - F(y^+). \tag{97}$$

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

$$\frac{1}{2L}\|g(r_{j-1})\|^2 \leqslant F(r_{j-1}) - F(r_j). \tag{98}$$

□

5 Conclusions

We introduced an adaptive FISTA restart scheme for convex composite functions satisfying a quadratic growth condition around their minimizers. This method relies on an automatic estimation of the growth parameter which is generally not known in practice. The theoretical convergence rate provided is the fastest in the literature so far if the growth parameter cannot be estimated. Numerical experiments emphasize the efficiency of this method especially in the case when computing the function to minimize is expensive.

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. Aujol, C. Dossal, and A. Rondepierre. Convergence rates of the Heavy-Ball method with Lojasiewicz property. 2020.

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

[10] J.-F. Aujol, C. Dossal, and A. Rondepierre. Convergence rates of the Heavy-Ball method for quasi-strongly convex optimization. 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] O. Fercoq and Z. Qu. Adaptive restart of accelerated gradient methods under local quadratic growth condition. *IMA Journal of Numerical Analysis*, 39(4):2069–2095, 2019.

[18] M. I. Florea and S. A. Vorobyov. A generalized accelerated composite gradient method: Uniting nesterov's fast gradient method and fista. *IEEE Transactions on Signal Processing*, 68:3033–3048, 2020.

[19] 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.

[20] 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.

[21] T. Goldstein, C. Studer, and R. Baraniuk. A field guide to forward-backward splitting with a fasta implementation. *arXiv preprint arXiv:1411.3406*, 2014.

[22] C. C. Gonzaga and E. W. Karas. Fine tuning nesterov's steepest descent algorithm for differentiable convex programming. *Mathematical Programming*, 138(1):141–166, 2013.

[23] C. C. Gonzaga, E. W. Karas, and D. R. Rossetto. An optimal algorithm for constrained differentiable convex optimization. *SIAM Journal on Optimization*, 23(4):1939–1955, 2013.

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

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

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

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

[28] 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.

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

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

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

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

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

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

[35] V. Roulet and A. d'Aspremont. Sharpness, restart, and acceleration. *SIAM Journal on Optimization*, 30(1):262–289, 2020.

[36] O. Sebbouh, C. Dossal, and A. Rondepierre. Nesterov's acceleration and Polyak's Heavy-Ball method in continuous time: convergence rate analysis under geometric conditions and perturbations. 2019.

[37] 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.

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

[39] 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.

[40] 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.