Restart FISTA with Global Linear Convergence

Teodoro Alamo, Pablo Krupa and Daniel Limon

arXiv:1906.09126v2 [math.OC] 24 Dec 2019

Abstract—Fast Iterative Shrinking-Threshold Algorithm (FISTA) is a popular fast gradient descent method (FGM) in the field of large scale convex optimization problems. However, it can exhibit undesirable periodic oscillatory behaviour in some applications that slows its convergence. Restart schemes seek to improve the convergence of FGM algorithms by suppressing the oscillatory behaviour. Recently, a restart scheme for FGM has been proposed that provides linear convergence for non strongly convex optimization problems that satisfy a quadratic functional growth condition. However, the proposed algorithm requires prior knowledge of the optimal value of the objective function or of the quadratic functional growth parameter. In this paper we present a restart scheme for FISTA algorithm, with global linear convergence, for non strongly convex optimization problems that satisfy the quadratic growth condition without requiring the aforementioned values. We present some numerical simulations that suggest that the proposed approach outperforms other restart FISTA schemes.

Keywords: Fast gradient method, restart FISTA, convex optimization, linear convergence, quadratic functional growth condition.

I. INTRODUCTION

Fast gradient methods (FGM) were introduced by Yurii Nesterov in [3], [4], where it was shown that these methods provide a convergence rate $O(1/k^2)$ for smooth convex optimization problems with non strongly convex objective functions [4], where k is the iteration counter. These methods were generalized to composite non smooth convex optimization problems in [5], [6], [7]. The resulting algorithm is commonly known as FISTA algorithm [5]. Because of its complexity certification, it is often used in the context of embedded model predictive control [8], [9], [10]. Another possibility to address composite convex optimization problems is to use splitting methods like ADMM [11], [12], [13].

FISTA algorithms can be applied in a primal setting (as in the Lasso problem [5]), or in a dual one [14], [15]. They can be thought of as a momentum method, since the linearization point at each iteration depends on the previous iterations. Since the momentum grows with the iteration counter, the algorithm can exhibit undesirable periodic oscillating behavior for certain applications, which slows the convergence rate. To mitigate this, restart schemes have been proposed in the literature which stop the algorithm when a certain criteria is met. It is then restarted using the last value provided by the stopped algorithm as the new initial condition [16], [17], [18].

In [16] two heuristic restart schemes for FGM are proposed which exhibit improved convergence rates over nonrestart FGM schemes. These restart schemes reset the momentum of the FGM in order to eliminate the undesirable oscillations whenever the periodical behavior is detected. A restart scheme similar to the ones in [16] with $O(1/k^2)$ convergence rate for smooth convex optimization is presented in [18]. In [19], an algorithm is proposed that uses the restart schemes from [16]. Numerical results show improvements over previous restart schemes for FGM, but no theoretical results on convergence rates are provided.

Recently, linear convergence rate has been derived for several first order methods applied to convex optimization problems with non strongly convex objective functions that satisfy a relaxation of the strong convexity known as the quadratic functional growth [20].

In [20, Subsection 5.2.2] a restarting scheme of FGM is presented with global linear convergence rate for convex optimization problems that satisfy the functional growth condition with parameter μ . However, in order to implement this strategy, prior knowledge is needed of either the optimal value of the objective function or the value of μ , which can be challenging to compute.

In this paper we propose a novel restart scheme for FISTA algorithm applied to solving convex constrained problems. We show that the algorithm guarantees global linear convergence rate $O(1/\sqrt{\mu})$ for convex optimization problems with non strongly convex objective functions that satisfy the quadratic functional growth condition with parameter μ . The proposed algorithm does not require prior knowledge of the value of μ or of the optimal value of the objective function. We provide theoretical upper bounds on the number of iterations of the algorithm needed to achieve a given accuracy.

Additionally, we show numerical results comparing the proposed algorithm with the heuristic restart schemes from [16] and the restart scheme from [20] for Lasso problems.

In Section II we introduce the problem formulation. Section III presents FISTA algorithm and some restart schemes. The convergence rate of non restart FISTA algorithm under the satisfaction of the quadratic functional growth condition is presented in Section IV. In Section V we present the proposed restart scheme for FISTA and state its global linear convergence. Numerical results comparing the proposed algorithm with other restart schemes applied to FISTA are shown in Section VI. Finally, conclusions are presented in Section VII.

T. Alamo, P. Krupa and D. Limon are at the Systems Engineering and Automation Department, University of Seville, Spain. E-mail: {talamo,pkrupa,dlm}@us.es

The authors acknowledge MINERCO and FEDER funds for funding project DPI2016-76493-C3-1-R, and MCIU and FSE for the FPI-2017 grant.

This paper constitutes an extended and revised version of [1]. Some of the technical results presented in this paper are used in [2].

Notation: Given vectors x and y, we denote by $\langle x, y \rangle$ their scalar product, i.e. $\langle x, y \rangle \doteq x^{\top} y$. Given vector x, $||x||_2$ denotes its Euclidean norm $(||x||_2 \doteq \sqrt{x^{\top}x})$, and $\|\cdot\|_1$ denotes its l_1 -norm (sum of the absolute values of the components of x). Given $R \succ 0$ we denote by $\|\cdot\|_R$ the weighted Euclidean norm $||x||_R \doteq \sqrt{x^\top R x}$, and by $||x||_* \doteq ||x||_{R^{-1}}$ its dual norm. $\ln(\cdot)$ is the natural logarithm and e is Euler's number. $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x; [x] denotes the smallest integer greater than or equal to x. Given a set $\mathcal{X} \subseteq \mathbb{R}^n$ we denote by $I_{\mathcal{X}}$ its indicator function. That is, $I_{\mathcal{X}}(x) = 0$ if $x \in \mathcal{X}$, and $I_{\mathcal{X}}(x) = \infty$ if $x \notin \mathcal{X}$. The relative interior of set \mathcal{X} is denoted by $ri(\mathcal{X})$. Given the extended real valued function $f: \mathbb{R}^n \to (-\infty, \infty]$ we denote by $\operatorname{dom}(f)$ its effective domain. That is, dom $(f) \doteq \{ x \in \mathbb{R}^n : f(x) < \infty \}.$ We denote by epi(f) the epigraph of f. That is, $epi(f) \doteq \{ (x,t) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq t \}.$ We say that function $f: \mathbb{R}^n \to (-\infty, \infty]$ is closed if its epigraph is a closed set. We say that $f : \mathbb{R}^n \to (-\infty, \infty]$ is proper if its effective domain is not empty. That is, if f is not identically equal to ∞ . We say that a vector $d \in \mathbb{R}^n$ is a subgradient of f at a point $x \in \text{dom}(f)$ if $f(y) \ge f(x) + \langle d, y - x \rangle$, $\forall y \in \mathbb{R}^n$. The set of all subgradients of f at x is called the subdifferential of f at x and is denoted by $\partial f(x)$.

II. PROBLEM FORMULATION

We address the problem of solving the composite convex minimization problem

$$f^* = \min_{x \in \mathcal{X}} f(x) = \min_{x \in \mathcal{X}} \Psi(x) + h(x), \tag{1}$$

under the following assumption.

Assumption 1. We assume that

(i) $h : \mathbb{R}^n \to \mathbb{R}$ is a smooth differentiable convex function. That is, there is $R \succ 0$ such that the inequality

$$h(x) \le h(y) + \langle \nabla h(y), x - y \rangle + \frac{1}{2} ||x - y||_R^2,$$
 (2)

is satisfied for every $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$.

- (ii) $\Psi : \mathbb{R}^n \to (-\infty, \infty]$ is a closed convex function and $\mathcal{X} \subseteq \mathbb{R}^n$ is a closed convex set.
- (iii) Denote $f \doteq \Psi + h$. The minimization problem

$$\min_{x \in \mathcal{X}} f(x)$$

is solvable. That is, there is $x^* \in \mathcal{X} \cap \operatorname{dom}(\Psi)$ such that $f^* = f(x^*) = \inf_{x \in \mathcal{X}} f(x)$.

We notice that it is standard to write down the first point of Assumption 1 as

$$h(x) \le h(y) + \langle \nabla h(y), x - y \rangle + \frac{L}{2} ||x - y||_S^2,$$
 (3)

where parameter L serves to characterize the smoothness of h and S is a positive definite matrix. Constant L provides

a bound on the Lipschitz constant of the gradient $\nabla h(\cdot)$ [4, Subsection 2.1]. Since

$$\frac{L}{2}\|x-y\|_{S}^{2} = \frac{1}{2}\|x-y\|_{LS}^{2}$$

we have that (3) implies (2) if we take R = LS. This simplifies the algebraic expressions needed to analyze the convergence of the proposed algorithm.

We notice that Assumption 1 guarantees that the minimization problem (1) is solvable. The optimal set Ω is defined as

$$\Omega \doteq \{ x : x \in \mathcal{X}, f(x) = f^* \}.$$

This set is a singleton if f(x) is strictly convex. Given $x \in \mathbb{R}^n$ we will denote \bar{x} its closest element in the optimal set Ω (with respect to the norm $\|\cdot\|_R$). That is,

$$\bar{x} \doteq \arg\min_{z \in \Omega} \|x - z\|_R. \tag{4}$$

Given $y \in \mathbb{R}^n$, one could use the local information given by $\nabla h(y)$ to minimize the value of $f = \Psi + h$ around y. Under Assumption 1, this can be done obtaining the minimizer of the strictly convex optimization problem

$$\min_{x \in \mathcal{X}} \Psi(x) + \langle \nabla h(y), x - y \rangle + \frac{1}{2} \|x - y\|_R^2.$$

It is well known that this problem is solvable and has a unique solution if Assumption 1 holds (see, for example, Subsection 6.1 in [21] for an analogous result). For completeness we provide a proof of this statement in Appendix A (see Property 5).

The solution to this optimization problem leads to the notion of composite gradient mapping [6], which constitutes a generalization of the gradient mapping that can be found in [4, Subsection 2.2] for the particular case $\Psi(\cdot) = 0$. See also [5] for the particular case $\mathcal{X} = \mathbb{R}^n$.

Definition 1 (Composite Gradient Mapping g(y)). Under Assumption 1, and given $y \in \mathbb{R}^n$, we define

$$y^+ \doteq \arg\min_{x \in \mathcal{X}} \Psi(x) + \langle \nabla h(y), x - y \rangle + \frac{1}{2} \|x - y\|_R^2,$$

$$g(y) \doteq R(y - y^+).$$

We notice that the composite gradient mapping is closely related to the notion of proximal operator [22], [21, Chapter 6]. For example, one could state, after some manipulations, the computation of the composite gradient mapping as the computation of a proximal operator. In the context of optimal gradient methods, it is assumed that the computation of y^+ is cheap. This is the case when \mathcal{X} is a simple set (box, \mathbb{R}^n , etc.), R diagonal, and $\Psi(\cdot)$ a separable function. For example, in the well known Lasso optimization problem, the computation of y^+ resorts to the computation of the shrinkage operator [5]. See [23], Section 6 of [22], Chapter 28 in [24], or Chapter 6 in [21], for numerous examples in which the computation of the composite gradient mapping is simple.

The following property gathers well-known properties of the composite gradient mapping g(y) and its dual norm $||g(y)||_* = ||g(y)||_{R^{-1}}$ [5], [6]. For completeness, we include the proof in Appendix B.

Property 1. Suppose that Assumption 1 holds. Then,

(i) For every $y \in \mathbb{R}^n$ and $x \in \mathcal{X}$:

$$f(y^{+}) - f(x) \leq \langle g(y), y^{+} - x \rangle + \frac{1}{2} ||g(y)||_{*}^{2}$$
$$= \langle g(y), y - x \rangle - \frac{1}{2} ||g(y)||_{*}^{2}$$
$$= -\frac{1}{2} ||y^{+} - x||_{R}^{2} + \frac{1}{2} ||y - x||_{R}^{2}$$

(ii) For every $y \in \mathcal{X}$:

$$\frac{1}{2} \|g(y)\|_*^2 \le f(y) - f(y^+) \le f(y) - f^*.$$

The composite gradient serves to characterize optimality [6]. That is, under Assumption 1 we have the following equivalence

$$y \in \Omega \Leftrightarrow g(y) = 0.$$

This fact is proved in Appendix C.

III. RESTART FISTA SCHEMES

For a given initial condition $z \in \mathbb{R}^n$, a minimum number of iterations $k_{min} \ge 0$, and an exit condition E_c , the non restart FISTA algorithm [5] is shown in Algorithm 1. This algorithm solves $\min_{x \in \mathcal{X}} h(x) + \Psi(x)$ under Assumption 1.

Algorithm 1: FISTA

_					
	Require: $z \in \mathbb{R}^n$, $k_{min} \ge 0$, E_c				
1	$y_0 = x_0 = z^+, t_0 = 1, k = 0$				
2	repeat				
3	k = k + 1				
4	$x_k = y_{k-1}^+$				
5	$t_k = \frac{1}{2} \left(1 + \sqrt{1 + 4t_{k-1}^2} \right)$				
6	$y_k = x_k + \frac{t_{k-1} - 1}{t_k} (x_k - x_{k-1})$				
7	Compute exit condition E_c				
8	s until E_c and $k \ge k_{min}$				
	Output: $r = x_k, n = k$				

Since the optimality of x_k is equivalent to $g(x_k) = 0$ (see Property 7 in Appendix C), a typical choice for non restart FISTA schemes is to choose k_{min} equal to zero and codify the exit condition

$$\|g(x_k)\|_* \le \epsilon,$$

where $\epsilon > 0$ is an accuracy parameter. It is also common to use the exit condition $||g(y_{k-1})||_* \leq \epsilon$, since this exit condition requires y_{k-1}^+ , which has already been computed in step 4 of the algorithm.

It is well known that under Assumption 1, see also (3), the iterations of non-restart FISTA satisfy [5], [6],

$$f(x_k) - f^* \le \frac{2}{(k+1)^2} \|x_0 - \bar{x}_0\|_R^2, \ \forall k \ge 1,$$
 (5)

where \bar{x}_0 represents the point in the optimal set Ω closest to the initial condition x_0 of the algorithm (see (4)). For the sake of completeness, we present a detailed proof of this claim in Appendix D. We also prove in the same appendix that the sequence $\{y_k\}$ generated by Algorithm 1 (FISTA) satisfies

$$||g(y_k)||_* \le \frac{4||x_0 - \bar{x}_0||_R}{k+2}, \forall k \ge 0.$$

In restart schemes, one invokes several times FISTA algorithm with a relaxed exit condition. Typical choices are (see [16]),

(i) Function scheme:

$$E_c^f = \text{True} \Leftrightarrow f(x_k) \ge f(x_{k-1}).$$
 (6)

(ii) Gradient scheme:

$$E_c^g = \text{True} \Leftrightarrow \langle g(y_{k-1}), x_{k-1} - x_k \rangle \le 0.$$
 (7)

Given initial condition $r_0 \in \mathcal{X}$, a minimum number of iterations $k_{min} \geq 0$, an exit condition E_c , and an accuracy parameter $\epsilon > 0$, the standard restart FISTA algorithm is shown in Algorithm 2.

1	Algorithm 2: Restart FISTA				
	Require: $r_0 \in \mathcal{X}, k_{min} \geq 0, \epsilon > 0, E_c$				
1	j = 0				
2	repeat				
3	j = j + 1				
4	$r_j = \text{FISTA}(r_{j-1}, k_{min}, E_c)$				
5	until $ g(r_j) _* \le \epsilon$				
	Output: $x^* = r_j$				

The implementation of Algorithm 2 usually provides better performance than the original non restart version [16], [18].

IV. CONVERGENCE OF RESTART FISTA UNDER A QUADRATIC FUNCTIONAL GROWTH CONDITION

It has been recently shown in [20] that some relaxations of the strong convexity conditions of the objective function are sufficient for obtaining linear convergence for several first order methods. In particular, the following relaxation of strong convexity suffices to guarantee linear convergence of different gradient optimization schemes for smooth functions $(\Psi(\cdot) = 0)$. See [20, Subsection 5.2.2].

Assumption 2 (Quadratic Functional Growth). *We assume that the optimization problem*

$$f^* = \min_{x \in \mathcal{X}} f(x)$$

is solvable and satisfies the following quadratic functional growth condition with parameter $\mu > 0$:

$$f(x) - f^* \ge \frac{\mu}{2} \|x - \bar{x}\|_R^2, \ \forall x \in \mathcal{X},$$

where \bar{x} denotes the closest element to x in the optimal set Ω (see (4)).

As can be seen in [20, Subsection 3.4], strong convexity implies quadratic functional growth. This means that the quadratic functional growth setting encompasses a broad family of convex functions.

It is also shown in [20, Subsection 5.2.2] that if the value of f^* is known and $\Psi(\cdot) = 0$, then a restart FISTA based on the exit condition

$$E_c^* = \operatorname{True} \Leftrightarrow f(x_k) - f^* \le \frac{f(x_0) - f^*}{e^2}, \qquad (8)$$

exhibits global linear convergence. This exit condition is easily implementable if the optimal value f^* is known. This is the case, for example, in some formulations of feasibility optimization problems, in which the optimal value f^* is equal to zero for every feasible solution. This restart scheme corresponds to an optimal restart rate of $\frac{2e}{\sqrt{\mu}}$ [20, Subsection 5.2.2].

We present now a novel result that further characterizes the convergence properties of the non restart FISTA algorithm under Assumption 2.

Property 2. Under Assumptions 1 and 2, the iterations of FISTA algorithm satisfy

(i)
$$f(x_k) - f^* \leq \frac{4(f(x_0) - f^*)}{\mu(k+1)^2}$$
, for all $k \geq 1$.
(ii) $f(x_k) \leq f(x_0)$, for all $k \geq \left\lfloor \frac{2}{\sqrt{\mu}} \right\rfloor$.
(iii) $f(x_k) - f^* \leq \frac{f(x_0) - f(x_k)}{e}$, for all $k \geq \left\lfloor \frac{2\sqrt{e+1}}{\sqrt{\mu}} \right\rfloor$.

Proof. See Appendix F.

V. RESTART FISTA WITH GLOBAL LINEAR CONVERGENCE

In this section we propose a novel restart FISTA algorithm (Algorithm 3) that exhibits global linear convergence under the quadratic functional growth condition. The algorithm uses exit condition E_c^l , which is defined to be true if the following two conditions are satisfied,

$$E_c^l = \operatorname{True} \Leftrightarrow \begin{cases} f(x_m) - f(x_k) \le \frac{f(x_0) - f(x_m)}{e} & \text{(9a)} \\ f(x_k) \le f(x_0), & \text{(9b)} \end{cases}$$

$$\int f(x_k) \le f(x_0), \qquad ($$

with $m = \lfloor \frac{\kappa}{2} \rfloor + 1$.

Inequality (9b) guarantees that the output of the FISTA algorithm is no larger than the one corresponding to its initial condition.

As it is stated in the following property, one of the main features of the proposed algorithm is that the number of iterations n_j required at each FISTA iteration $[r_j, n_j] = FISTA(r_{j-1}, n_{j-1}, E_c^l)$ is upper bounded by $\frac{4\sqrt{e+1}}{\sqrt{\mu}} \approx \frac{7.72}{\sqrt{\mu}}$. Moreover, the number of iterations required by the proposed algorithm to attain a given accuracy ϵ is upper bounded by

$$\frac{16}{\sqrt{\mu}} \left[\ln \left(1 + \frac{2(f(r_0) - f^*)}{\epsilon^2} \right) \right].$$

Algorithm 3: Linearly Convergent Restart FISTA (LCR-FISTA)

$$\begin{array}{c|c} \mbox{Require: } r_0 \in \mathcal{X}, \ \epsilon > 0 \\ \mbox{$\mathbf{1}$} \ n_0 = 0, \ j = 1 \\ \mbox{$\mathbf{2}$} \ [r_1, n_1] = \mbox{FISTA}(r_0, n_0, E_c^l) \\ \mbox{$\mathbf{3}$ repeat} \\ \mbox{$\mathbf{4}$} & \left| \begin{array}{c} j = j + 1 \\ [r_j, n_j] = \mbox{FISTA}(r_{j-1}, n_{j-1}, E_c^l) \\ \mbox{\mathbf{i}} \ f(r_{j-1}) - f(r_j) > \frac{1}{e} \left(f(r_{j-2}) - f(r_{j-1}) \right) \ \mbox{then} \\ \mbox{$\mathbf{1}$} \ n_j = 2n_{j-1} \\ \mbox{$\mathbf{8}$} & \mbox{\mathbf{e}} \ \mbox{$\mathbf{0}$} \\ \mbox{$\mathbf{0}$} \ \mbox{\mathbf{util}} \ ||g(r_j)||_* \le \epsilon \\ \mbox{$\mathbf{0}$} \ \mbox{$\mathbf{0}$} \ \mbox{$\mathbf{0}$} \\ \mbox{$\mathbf{0}$} \ \mbox{$\mathbf{0}$} \ \mbox{$\mathbf{0}$} \\ \mbox{$\mathbf{0}$} \ \mbox{$\mathbf{0}$} \ \mbox{$\mathbf{0}$} \\ \mbox{$\mathbf{0}$} \ \mbox{$\mathbf{0}$} \ \mbox{$\mathbf{0}$} \ \mbox{$\mathbf{0}$} \ \mbox{$\mathbf{0}$} \\ \mbox{$\mathbf{0}$} \ \mbox{$\mathbf{0}$$

Property 3. Suppose that Assumptions 1 and 2 hold. Then, the sequences $\{r_j\}$, $\{n_j\}$ provided by Algorithm 3 satisfy

(i)
$$\frac{1}{2} \|g(r_{j-1})\|_{*}^{2} \leq f(r_{j-1}) - f(r_{j}), \forall j \geq 1.$$

(ii) $n_{j} \leq \frac{4\sqrt{e+1}}{\sqrt{\mu}}, \forall j \geq 0.$
(iii) The number of iterations $(\sum_{j=1}^{j} n_{j})$ required to

(iii) The number of iterations
$$(\sum_{i=0}^{n} n_i)$$
 required to guarantee
 $\|g(r_j)\|_* \le \epsilon$ is no larger than
 $\frac{16}{\sqrt{\mu}} \left[\ln \left(1 + \frac{2(f(r_0) - f^*)}{\epsilon^2} \right) \right].$

Proof. See Appendix G.

We notice that the factor 16 in the worst case complexity analysis is conservative. The authors claim that a better factor might be obtained at the expense of a more involved proof.

VI. NUMERICAL RESULTS

We consider a weighted Lasso problem of the form

$$\min_{x} \frac{1}{2N} \|Ax - b\|_{2}^{2} + \|Wx\|_{1}, \tag{10}$$

where $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{N \times n}$ is sparse with an average of 90% of its entries being zero (sparsity was generated by setting a 0.9 probability for each element of the matrix to be 0), n > N, and $b \in \mathbb{R}^N$. Each nonzero element in A and b is obtained from a Gaussian distribution with zero mean and covariance 1. $W \in \mathbb{R}^{n \times n}$ is a diagonal matrix with elements obtained from a uniform distribution on the interval $[0, \alpha]$.

We note that Lasso problems (10) can be reformulated in such a way that they satisfy the quadratic growth condition [20, Section 6.3]. For this problem, inequality (2) of Assumption 1 is satisfied, for instance, for a matrix R chosen as

$$R_{i,i} = \sum_{j=1}^{n} |H_{i,j}|,$$

with $H = \frac{1}{N}A^{\top}A$. This is due to the Gershgorin Circle Theorem [25, Subsection 7.2]. See also [6, Section 6].

We show the results of applying algorithms 2 and 3 with an accuracy parameter $\epsilon = 10^{-11}$ using different restart schemes and values of N, n and α . We take $r_0 = 0$.

The restart schemes shown are E_c^f (6) and E_c^g (7) from [16], restart condition E_c^* (8) [20], and the restart condition E_c^l (9) proposed in this paper (using Algorithm 3). Additionally, we show the results of applying FISTA algorithm without using a restart scheme. In order to provide a fair comparison between the performance of the restart schemes, the algorithms are exited as soon as a value of y_k that satisfies $||g(y_{k-1})||_* \leq \epsilon$ is found. We note that, in order to implement the restart scheme based on E_c^* , we had to previously compute the optimal value f^* , which was done by using Algorithm 3 with $\epsilon = 10^{-12}$.

Tables I to III show results of performing 100 tests with different randomized problems (10) that share common values of parameters N, n and α . Tables show the average, median, maximum and minimum number of iterations.

 TABLE I

 Test 1. Comparison between restart schemes

Exit Cond.	E_c^l	No restart	E_c^f	E_c^g	E_c^*
Avg. Iter.	670.6	8207.2	1648.7	687.5	1569.5
Median Iter.	676	8241	1608.5	666.5	1571
Max. Iter.	783	10109	2156	930	2053
Min. Iter.	570	6737	1192	567	917
P is (100×10^{-11}) in (100×10^{-11})					

Results of 100 tests with $N = 600, n = 800, \alpha = 0.01, \epsilon = 10^{-11}$

 TABLE II

 Test 2. Comparison between restart schemes

Exit Cond.	E_c^l	No restart	E_c^f	E_c^g	E_c^*
Avg. Iter.	1683.7	34116.4	7743.3	1606.7	4601.9
Median Iter.	1659	33127.5	7242	1594	4503
Max. Iter.	2162	51201	14080	2201	7266
Min. Iter.	1406	24539	3894	1306	2499
Results for 100 tests with $N = 600$, $n = 800$, $\alpha = 0.003$, $\epsilon = 10^{-11}$.					

TABLE III Test 3. Comparison between restart schemes

Exit Cond.	E_c^l	No restart	E_c^f	E_c^g	E_c^*
Avg. Iter.	705.9	8379.5	1786.3	686	1709.4
Median Iter.	704.5	8135.5	1773	680.5	1703
Max. Iter.	873	12055	3218	892	2512
Min. Iter.	547	5943	987	529	1042

Results for 100 tests with $N = 300, n = 400, \alpha = 0.01, \epsilon = 10^{-11}$.

Figures 1 to 3 show the value of $||g(x_k)||_*$ for a randomly selected problem out of the randomized problems used to compute the results shown in tables I to III, respectively.

Figure 4 shows the value of n_j at each iteration j of Algorithm 3 for the three examples whose results are shown in Figures 1 to 3. Note that the final value of n_j is lower than



Fig. 1. Value of $||g(y_k)||_*$ for a problem (10) of Test 1.



Fig. 2. Value of $||g(y_k)||_*$ for a problem (10) of Test 2.

the previous one in all three instances due to the algorithm exiting as soon as the condition $||g(y_{k-1})||_* \le \epsilon$ is satisfied.

VII. CONCLUSIONS

In this paper we have presented a novel restart scheme with guaranteed global linear convergence. The algorithm relies on a quadratic functional growth condition. One of the advantages of the proposed algorithm is that it does not require the knowledge of the parameter μ that characterizes the quadratic functional growth condition, or the optimal value of the minimization problem. We provide an upper bound of the required number of iterations equal to

$$\frac{16}{\sqrt{\mu}} \left\lceil \ln \left(1 + \frac{2(f(r_0) - f^*)}{\epsilon^2} \right) \right\rceil.$$

We have presented numerical evidence of the good performance of the algorithm when compared with other restarts schemes. It outperforms the restart scheme based on the knowledge of the optimal value f^* .

REFERENCES

- T. Alamo, P. Krupa, and D. Limon, "Restart FISTA with global linear convergence," in 2019 18th European Control Conference (ECC). IEEE, 2019, pp. 1969–1974.
- [2] —, "Gradient based restart FISTA," in *Proceedings of the 58th IEEE Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 3936–3941.
- [3] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," Sov. Math. Dokl., vol. 27, no. 2, pp. 372–376, 1983.
- [4] —, Introductory Lectures on Convex Optimization: A Basic Course. Springer, 2004.



Fig. 3. Value of $||g(y_k)||_*$ for a problem (10) of Test 3.



Fig. 4. Value of n_j obtained using Algorithm 3 for the problems (10) whose result are shown in Figures 1 to 3.

- [5] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.
- [6] Y. Nesterov, "Gradient methods for minimizing composite functions," *Mathematical Programming*, vol. 140, pp. 125–161, 2013.
- [7] P. Tseng, "On accelerated proximal gradient methods for convexconcave optimization," Dept. Math., Univ. Washington, Seattle, WA, USA, Tech. Rep., 2008.
- [8] M. Kögel and R. Findeisen, "A fast gradient method for embedded linear predictive control," in *18th IFAC World Congress*, 2011, pp. 1362–1367.
- [9] S. Richter, C. Jones, and M. Morari, "Computational complexity certification for real-time MPC with input constraints based on the fast gradient method," *IEEE Transactions on Automatic Control*, vol. 57, no. 6, pp. 1391–1403, 2012.
- [10] P. Krupa, D. Limon, and T. Alamo, "Implementation of model predictive controllers in programmable logic controllers using IEC 61131-3 standard," in 2018 European Control Conference (ECC). IEEE, 2018, pp. 1–6.
- [11] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends* (*in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.
- [12] P. Giselsson and S.Boyd, "Linear convergence and metric selection in Douglas-Rachford splitting and ADMM," *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 532–544, 2017.
- [13] G. Banjac and P. J. Goulart, "Tight global linear convergence rate bounds for operator splitting methods," *IEEE Transactions on Automatic Control*, vol. 63, no. 12, pp. 4126–4139, 2018.
- [14] S. Richter, C. Jones, and M. Morari, "Certification aspects of the fast gradient method for solving the dual of parametric convex programs," *Mathematical Methods Operational Research*, vol. 77, pp. 305–321, 2013.
- [15] A. Beck and M. Teboulle, "A fast dual proximal gradient algorithm for convex minimization and applications," *Operations Research Letters*, vol. 42, no. 1, pp. 1–6, 2014.
- [16] B. O'Donoghue and E. Candes, "Adaptive restart for accelerated gradient schemes," *Foundations of Computational Mathematics*, pp. 1–18, 2013.

- [17] M. Alamir, "Monitoring control updating period in fast gradient based NMPC," in *Proceedings of the European Control Conference (ECC)*, 2013, pp. 3621–3626.
- [18] P. Giselsson and S. Boyd, "Monotonicity and restart in fast gradient methods," in *Proceedings of the 2014 IEEE 53rd Annual Conference* on Decision and Control, 2014, pp. 5058–5063.
- [19] D. Kim and J. A. Fessler, "Adaptive restart of the optimized gradient method for convex optimization," *Journal of Optimization Theory and Applications*, vol. 178, no. 1, pp. 240–263, 2018.
- [20] I. Necoara, Y. Nesterov, and F. Glineur, "Linear convergence of first order methods for non-strongly convex optimization," *Mathematical Programming*, pp. 1–39, 2018.
- [21] A. Beck, First-Order Methods in Optimization. SIAM, 2017.
- [22] N. Parikh and S. Boyd, "Proximal algorithms," Foundations and Trends in Optimization, vol. 1, no. 3, pp. 123–231, 2013.
- [23] P. Combettes and J. Pesquet, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer Optimization and its Applications, 2011, vol. 49, ch. Proximal Splitting Methods in Signal Processing, pp. 185–212.
- [24] H. Bauschke and P. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.
- [25] G. H. Golub and C. F. Van Loan, *Matrix computations*, 3rd ed. Johns Hopkins University Press, 1996.
- [26] D. Bertsekas, Convex Optimization Theory. Athena Scientific, 2009.

APPENDIX

A. Existence and Uniqueness of Composite Gradient

We present in this appendix some well known facts about convex analysis that are required to analyze the properties of the composite gradient.

Property 4. Suppose that

- (i) $\Psi : \mathbb{R}^n \to (-\infty, \infty]$ is a closed convex function.
- (ii) $\mathcal{X} \subseteq \mathbb{R}^n$ is a closed convex set.
- (iii) The set dom(Ψ) $\bigcap \mathcal{X}$ is non empty.
- (iv) $I_{\mathcal{X}} : \mathbb{R}^n \to \{0, \infty\}$ is the indicator function of \mathcal{X} . That is,

$$I_{\mathcal{X}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{X} \\ \infty & \text{otherwise.} \end{cases}$$

(v) The function $\Psi_{\mathcal{X}} : \mathbb{R}^n \to (-\infty, \infty]$ is defined as

$$\Psi_{\mathcal{X}}(x) \doteq \Psi(x) + I_{\mathcal{X}}(x), \ \forall x \in \mathbb{R}^n.$$

Then

- (i) The function $\Psi_{\mathcal{X}}$ is proper, closed, and convex.
- (ii) The relative interior of dom($\Psi_{\mathcal{X}}$) is non empty.
- (iii) There is $z \in \mathcal{X}$ and $d \in \mathbb{R}^n$ such that $\Psi_{\mathcal{X}}(z) < \infty$ and

$$\Psi_{\mathcal{X}}(x) \ge \Psi_{\mathcal{X}}(z) + \langle d, x - z \rangle, \ \forall x \in \mathbb{R}^n.$$

Proof. From dom(Ψ) $\bigcap \mathcal{X} \neq \emptyset$ we have that both dom(Ψ) and \mathcal{X} are non empty. The epigraph of the indicator function $I_{\mathcal{X}}$ is, by definition,

$$epi(I_{\mathcal{X}}) = \{ (x,t) \in \mathbb{R}^n \times \mathbb{R} : I_{\mathcal{X}}(x) \le t \}$$

= $\{ (x,t) \in \mathbb{R}^n \times \mathbb{R} : x \in \mathcal{X}, 0 \le t \}.$

Since \mathcal{X} and $\mathcal{T} \doteq \{ t \in \mathbb{R} : t \geq 0 \}$ are non empty closed sets, $\operatorname{epi}(I_{\mathcal{X}}) = \mathcal{X} \times \mathcal{T}$ is also a non empty closed convex set. Thus, by definition, $I_{\mathcal{X}} : \mathbb{R}^n \to \{0, \infty\}$ is a closed convex function. Since both Ψ and $I_{\mathcal{X}}$ are closed convex functions, $\Psi_{\mathcal{X}} \doteq \Psi + I_{\mathcal{X}}$ is also a closed convex function (the sum of closed convex functions provides closed convex functions [26, Proposition 1.1.5]). Since dom $(\Psi_{\mathcal{X}}) = \text{dom}(\Psi) \bigcap \mathcal{X} \neq \emptyset$, we infer that the domain of $\Psi_{\mathcal{X}}$ is non empty. This implies that $\Psi_{\mathcal{X}}$ is not identically equal to ∞ . Moreover, since $\Psi : \mathbb{R}^n \to (-\infty, \infty]$ we have that $\Psi_{\mathcal{X}} : \mathbb{R}^n \to (-\infty, \infty]$. We conclude that $\Psi_{\mathcal{X}}(x) > -\infty$ for every $x \in \mathbb{R}^n$. From this and the fact that $\Psi_{\mathcal{X}}$ is not identically equal to ∞ we have that $\Psi_{\mathcal{X}}$ is proper.

Since $\operatorname{dom}(\Psi_{\mathcal{X}})$ is a non empty convex set, it has a non empty relative interior $\operatorname{ri}(\operatorname{dom}(\Psi_{\mathcal{X}}))$ (see [26, Proposition 1.3.2]).

It is a well know fact from convex analysis that the subdifferential of a proper convex function at a point in the relative interior of its domain is non empty [26, Proposition 5.4.1]. Suppose now that $z \in ri(dom(\Psi_{\mathcal{X}}))$. Since $\Psi_{\mathcal{X}}$ is a proper convex function we have that the subdifferential of $\Psi_{\mathcal{X}}$ at z is non empty. This means, by definition, that there is $d \in \mathbb{R}^n$ such that

$$\Psi_{\mathcal{X}}(x) \ge \Psi_{\mathcal{X}}(z) + \langle d, x - z \rangle, \ \forall x \in \mathbb{R}^n$$

Property 5. Suppose that Assumption 1 holds. Given any $y \in \mathbb{R}^n$, consider the quadratic function $h_y : \mathbb{R}^n \to \mathbb{R}$ defined as

$$h_y(x) \doteq \langle \nabla h(y), x - y \rangle + \frac{1}{2} \|x - y\|_R^2.$$

Then, the minimization problem

$$\min_{x \in \mathcal{X}} \Psi(x) + h_y(x) \tag{11}$$

is solvable and has a unique solution. That is, there exists a unique point $y^+ \in \mathcal{X}$ such that

$$\Psi(y^+) + h_y(y^+) = \inf_{x \in \mathcal{X}} \Psi(x) + h_y(x) < \infty$$

Proof. Notice that the minimization problem (11) is equivalent to

$$\min_{x \in \mathbb{R}^n} \Psi(x) + I_{\mathcal{X}}(x) + h_y(x),$$

where $I_{\mathcal{X}}$ is the indicator function of \mathcal{X} . If we define $\Psi_{\mathcal{X}} \doteq \Psi + I_{\mathcal{X}}$ we can rewrite the original problem (11) as

$$\min_{x \in \mathbb{R}^n} \Psi_{\mathcal{X}}(x) + h_y(x)$$

We notice that the assumptions of Property 4 are satisfied if Assumption 1 holds. Thus, we infer from Property 4 that $\Psi_{\mathcal{X}} : \mathbb{R}^n \to (-\infty, \infty]$ is a proper closed convex function. We also have that the quadratic function $h_y : \mathbb{R}^n \to \mathbb{R}$ is also proper and closed because it is a real valued continuous function (see [26, Proposition 1.1.3]). Since the sum of closed functions is closed (see [26, Proposition 1.1.5]), we infer that $F_y \doteq \Psi_{\mathcal{X}} + h_y$ is a closed function. Moreover, from Property 4 we also have that there is $z \in \mathcal{X}$ and $d \in \mathbb{R}^n$ such that

(i) $\Psi_{\mathcal{X}}(z) < \infty$. (ii) $\Psi_{\mathcal{X}}(x) \ge \Psi_{\mathcal{X}}(z) + \langle d, x - z \rangle, \forall x \in \mathbb{R}^n$. Therefore,

$$F_{y}(z) = \Psi_{\mathcal{X}}(z) + h_{y}(z) = \gamma_{z} < \infty,$$

$$F_{y}(x) = \Psi_{\mathcal{X}}(x) + h_{y}(x) \qquad (12)$$

$$\geq \Psi_{\mathcal{X}}(z) + \langle d, x - z \rangle + h_{y}(x), \ \forall x \in \mathbb{R}^{n}.$$

We infer from (12) that the closed function $F_y: \mathbb{R}^n \to (-\infty, \infty]$ is not identically equal to ∞ and therefore, proper. We conclude that F_y is a proper closed convex function. From Weiertrasss' Theorem (see Proposition 3.2.1 in [26]) we have that the set of minima of F_y over \mathbb{R}^n is nonempty and compact if there is a scalar $\bar{\gamma}$ such that the level set $\Phi(\bar{\gamma}) = \{x : F_y(x) \leq \bar{\gamma}\}$ is nonempty and bounded. From (12) we have that $\Phi(\gamma_z)$ is nonempty. Moreover, we also infer from (12) that $\Phi(\gamma_z)$ is a bounded set because F_y is lower bounded by a strictly convex quadratic function of x. We conclude that

$$\min_{x \in \mathcal{X}} \Psi(x) + h_y(x) = \min_{x \in \mathbb{R}^n} \Psi_{\mathcal{X}}(x) + h_y(x)$$
$$= \min_{x \in \mathbb{R}^n} F_y(x) \le \gamma_z < \infty$$

is a solvable optimization problem. That is, there is $y^+ \in \mathcal{X}$ such that

$$\Psi(y^+) + h_y(y^+) = \inf_{x \in \mathcal{X}} \Psi(x) + h_y(x) < \infty.$$

The set of minimizers consists of a single element y^+ because of the strictly convex nature of F_y (h_y is a strictly convex function).

B. Proof of Property 1.

We prove in this appendix Property 1, which is rewritten here for the reader's convenience.

Property 6. Suppose that Assumption 1 holds. Then,

(i) For every $y \in \mathbb{R}^n$ and $x \in \mathcal{X}$:

$$f(y^+) - f(x) \le \langle g(y), y^+ - x \rangle + \frac{1}{2} ||g(y)||_*^2$$
 (13a)

$$= \langle g(y), y - x \rangle - \frac{1}{2} ||g(y)||_{*}^{2}$$
(13b)

$$= -\frac{1}{2} \|y^{+} - x\|_{R}^{2} + \frac{1}{2} \|y - x\|_{R}^{2}.$$
 (13c)

(ii) For every $y \in \mathcal{X}$:

$$\frac{1}{2} \|g(y)\|_*^2 \le f(y) - f(y^+) \le f(y) - f^*.$$

Proof. From Property 5 we have that there is a (unique) $y^+ \in \mathcal{X}$ such that

$$\Psi(y^+) + h_y(y^+) \le \Psi(x) + h_y(x), \ \forall x \in \mathcal{X},$$
(14)

where $h_y(x) \doteq \langle \nabla h(y), x - y \rangle + \frac{1}{2} ||x - y||_R^2$. Denote now $\Psi_{\mathcal{X}} = \Psi + I_{\mathcal{X}}$, where $I_{\mathcal{X}} : \mathbb{R}^n \to \{0, \infty\}$ is the indicator function of \mathcal{X} . Since $y^+ \in \mathcal{X}$ we have $I_{\mathcal{X}}(y^+) = 0$. Therefore, inequality (14) implies

$$\Psi_{\mathcal{X}}(y^+) + h_y(y^+) \le \Psi_{\mathcal{X}}(x) + h_y(x), \ \forall x \in \mathbb{R}^n.$$

Denote now $F_y = \Psi_{\mathcal{X}} + h_y$. From last inequality we have

$$F_y(y^+) \le F_y(x), \ \forall x \in \mathbb{R}^n$$

By definition of subdifferential at a point, we have that the previous inequality implies

$$0 \in \partial F_y(y^+). \tag{15}$$

We have that $\Psi_{\mathcal{X}}$ is a proper closed function and $\operatorname{ri}(\operatorname{dom}(\Psi_{\mathcal{X}})) \neq \emptyset$ (see the first two claims of Property 4). The domain of the quadratic function $h_y : \mathbb{R}^n \to \mathbb{R}$ is \mathbb{R}^n . Since h_y is a continuous real value function in \mathbb{R}^n , it is also closed (see Proposition 1.1.3 in [26]). We have that

$$\operatorname{ri}(\operatorname{dom}(\Psi_{\mathcal{X}})) \bigcap \operatorname{ri}(\operatorname{dom}(h_y)) = \operatorname{ri}(\operatorname{dom}(\Psi_{\mathcal{X}})) \bigcap \mathbb{R}^n$$
$$= \operatorname{ri}(\operatorname{dom}(\Psi_{\mathcal{X}})) \neq \emptyset.$$

Since $F_y = \Psi_{\mathcal{X}} + h_y$ is equal to the sum of two closed convex functions and

$$\operatorname{ri}(\operatorname{dom}(\Psi_{\mathcal{X}})) \bigcap \operatorname{ri}(\operatorname{dom}(h_y)) \neq \emptyset,$$

we have $\partial F_y(y^+) = \partial \Psi_{\mathcal{X}}(y^+) + \partial h_y(y^+)$ (see Proposition 5.4.6 in [26]). The subdifferential of the differentiable function h_y at y^+ is $\nabla h_y(y^+) = \nabla h(y) + R(y^+ - y)$. Thus, we obtain from (15)

$$0 \in \partial F_y(y^+) = \partial \Psi_{\mathcal{X}}(y^+) + \partial h_y(y^+)$$

= $\partial \Psi_{\mathcal{X}}(y^+) + \nabla h(y) + R(y^+ - y).$

Since g(y) is defined as $R(y - y^+)$ we obtain

$$g(y) - \nabla h(y) \in \partial \Psi_{\mathcal{X}}(y^+).$$

By definition of $\partial \Psi_{\mathcal{X}}(\cdot)$ we have

$$\Psi_{\mathcal{X}}(x) \ge \Psi_{\mathcal{X}}(y^+) + \langle g(y) - \nabla h(y), x - y^+ \rangle, \ \forall x \in \mathbb{R}^n.$$

Obviously, since $\mathcal{X} \subseteq \mathbb{R}^n$, this implies

$$\Psi_{\mathcal{X}}(x) \ge \Psi_{\mathcal{X}}(y^+) + \langle g(y) - \nabla h(y), x - y^+ \rangle, \ \forall x \in \mathcal{X}.$$

Since $y^+ \in \mathcal{X}$ and $\Psi_{\mathcal{X}} = \Psi$ for every $x \in \mathcal{X}$, we obtain

$$\Psi(x) \ge \Psi(y^+) + \langle g(y) - \nabla h(y), x - y^+ \rangle, \ \forall x \in \mathcal{X}.$$
 (16)

The convexity of $h(\cdot)$ implies

$$h(x) \ge h(y) + \langle \nabla h(y), x - y \rangle, \quad \forall x \in \mathcal{X}.$$

Adding this inequality to (16) yields

$$f(x) = \Psi(x) + h(x)$$

$$\geq \Psi(y^{+}) + \langle g(y) - \nabla h(y), x - y^{+} \rangle$$

$$+ h(y) + \langle \nabla h(y), x - y \rangle$$

$$= \Psi(y^{+}) + \langle g(y), x - y^{+} \rangle$$

$$+ h(y) + \langle \nabla h(y), y^{+} - y \rangle, \ \forall x \in \mathcal{X}.$$
(17)

From Assumption 1 we have

$$\begin{split} h(y) &\geq h(y^{+}) - \langle \nabla h(y), y^{+} - y \rangle - \frac{1}{2} \|y^{+} - y\|_{R}^{2} \\ &= h(y^{+}) - \langle \nabla h(y), y^{+} - y \rangle - \frac{1}{2} \|R^{-1}g(y)\|_{R}^{2} \\ &= h(y^{+}) - \langle \nabla h(y), y^{+} - y \rangle - \frac{1}{2} \|g(y)\|_{*}^{2}. \end{split}$$

Adding this inequality to (17) yields

$$f(x) \ge \Psi(y^{+}) + h(y^{+}) + \langle g(y), x - y^{+} \rangle - \frac{1}{2} ||g(y)||_{*}^{2}$$

= $f(y^{+}) + \langle g(y), x - y^{+} \rangle - \frac{1}{2} ||g(y)||_{*}^{2}, \forall x \in \mathcal{X}.$

From this inequality we have

$$f(y^+) - f(x) \le \langle g(y), y^+ - x \rangle + \frac{1}{2} ||g(y)||_*^2, \ \forall x \in \mathcal{X}.$$

This proves (13a). We now prove (13b) and (13c) by means of simple algebraic manipulations.

$$f(y^{+}) - f(x) \leq \langle g(y), y^{+} - x \rangle + \frac{1}{2} ||g(y)||_{*}^{2}$$

$$= \langle g(y), y - x + y^{+} - y \rangle + \frac{1}{2} ||g(y)||_{*}^{2}$$

$$= \langle g(y), y - x \rangle + \langle g(y), y^{+} - y \rangle + \frac{1}{2} ||g(y)||_{*}^{2}$$

$$= \langle g(y), y - x \rangle + \langle g(y), -R^{-1}g(y) \rangle + \frac{1}{2} ||g(y)||_{*}^{2}$$

$$= \langle g(y), y - x \rangle - ||g(y)||_{*}^{2} + \frac{1}{2} ||g(y)||_{*}^{2}$$

$$= \langle g(y), y - x \rangle - \frac{1}{2} ||g(y)||_{*}^{2}, \forall x \in \mathcal{X}.$$
(18)

This proves (13b). From this inequality, and the definition of g(y), we obtain

$$f(y^{+}) - f(x) \leq \langle R(y - y^{+}), y - x \rangle - \frac{1}{2} ||R(y - y^{+})||_{*}^{2}$$

$$= -\langle R(y - y^{+}), x - y \rangle - \frac{1}{2} ||y - y^{+}||_{R}^{2}$$

$$= -\frac{1}{2} ||y - y^{+} + x - y||_{R}^{2} + \frac{1}{2} ||x - y||_{R}^{2}$$

$$= -\frac{1}{2} ||y^{+} - x||_{R}^{2} + \frac{1}{2} ||y - x||_{R}^{2}, \forall x \in \mathcal{X}.$$

This proves (13c). Suppose now that $y \in \mathcal{X}$. Particularizing inequality (18) to x = y yields

$$\frac{1}{2} \|g(y)\|_*^2 \le f(y) - f(y^+), \ \forall y \in \mathcal{X}.$$

The inequality $f(y) - f(y^+) \le f(y) - f^*$ trivially follows from $f^* \le f(y^+)$.

C. Characterization of optimality

The following property serves to characterize the optimality of a given point $y \in \mathbb{R}^n$.

Property 7. Suppose that Assumption 1 holds. Then $y \in \mathbb{R}^n$ belongs to the optimal set

$$\Omega = \{ x : x \in \mathcal{X}, f(x) = f^* \}$$

if and only if g(y) = 0.

Proof. We first show that g(y) = 0 implies $y \in \Omega$. Since $R \succ 0$, we infer from equality $g(y) = R(y - y^+)$ that g(y) = 0 is equivalent to $y = y^+$. Suppose that $x^* \in \Omega \subseteq \mathcal{X}$. Then, we obtain from g(y) = 0, $y = y^+ \in \mathcal{X}$, and the first claim of Property 1, the following inequality

$$f(x^*) \ge f(y^+) - \langle g(y), y^+ - x^* \rangle - \frac{1}{2} ||g(y)||_*^2$$

= $f(y^+) = f(y).$

That is, $f^* = f(x^*) \ge f(y)$. Since $y = y^+ \in \mathcal{X}$, this is possible only if y is also optimal $(f(y) = f^*)$. This proves that g(y) = 0 implies $y \in \Omega$. We now prove that $y \in \Omega$ implies g(y) = 0. Suppose that $y \in \Omega$. Then, $f(y) = f^*$ and we obtain from the second claim of Property 1

$$\frac{1}{2} \|g(y)\|_*^2 \le f(y) - f^* = 0.$$

This implies g(y) = 0.

D. Convergence of non restart FISTA

Property 8. Suppose that Assumption 1 holds. Then, the sequences $\{x_k\}$ and $\{y_k\}$ generated by Algorithm 1 (FISTA) satisfy

(i)
$$f(x_k) - f^* \leq \frac{2\|x_0 - \bar{x}_0\|_R^2}{(k+1)^2}$$
, for all $k \geq 1$
(ii) $\|g(y_k)\|_* \leq \frac{4\|x_0 - \bar{x}_0\|_R}{k+2}$, for all $k \geq 0$,

where \bar{x}_0 represents the point in the optimal set Ω closest to the initial condition x_0 of the algorithm.

Proof. First claim:

We denote $g_k \doteq g(y_k)$, $\forall k \ge 0$. Additionally, we recall that $\|\cdot\|_* \doteq \|\cdot\|_{R^{-1}}$.

From step 4 of FISTA algorithm we have

$$x_k = y_{k-1}^+, \ \forall k \ge 1.$$
 (19)

This implies that

$$g_k = R(y_k - y_k^+) = R(y_k - x_{k+1}), \ \forall k \ge 0.$$

Particularizing inequality (13c) of the first claim of Property 6 to $y = y_0 \in \mathbb{R}^n$, and $x = \bar{x}_0 \in \Omega \subseteq \mathcal{X}$, we obtain

$$f(y_0^+) - f(\bar{x}_0) \le -\frac{1}{2} \|y_0^+ - \bar{x}_0\|_R^2 + \frac{1}{2} \|y_0 - \bar{x}_0\|_R^2$$

By construction we have that $x_0 = y_0$ and $x_1 = y_0^+$. Furthermore, by definition of \bar{x}_0 , we have $f(\bar{x}_0) = f^*$. Therefore we can rewrite previous inequality as

$$f(x_1) - f^* \le -\frac{1}{2} \|x_1 - \bar{x}_0\|_R^2 + \frac{1}{2} \|x_0 - \bar{x}_0\|_R^2 \qquad (20)$$
$$\le \frac{1}{2} \|x_0 - \bar{x}_0\|_R^2.$$

This proves the claim of the property for k = 1. We now proceed to prove the claim for $k \ge 2$. From equality (19) we have

$$x_{k+1} = y_k^+, \ \forall k \ge 1.$$

Therefore, from inequality (13b) of Property 6 we obtain that for every $x \in \mathcal{X}$ and every $k \ge 1$

$$f(x) \geq f(x_{k+1}) + \frac{1}{2} ||g_k||_*^2 - \langle g_k, y_k - x \rangle.$$

We notice that, by construction, $x_k \in \mathcal{X}$, $k \ge 1$. Particularizing at x_k and \bar{x}_0 , we obtain from last inequality

$$f(x_k) \ge f(x_{k+1}) + \frac{1}{2} ||g_k||_*^2 - \langle g_k, y_k - x_k \rangle, \ \forall k \ge 1, (21a)$$

$$f(\bar{x}_0) \ge f(x_{k+1}) + \frac{1}{2} ||g_k||_*^2 - \langle g_k, y_k - \bar{x}_0 \rangle, \ \forall k \ge 1. (21b)$$

In order to write down the proof in a compact way, we introduce the following incremental notation, valid for all $k \ge 0$,

$$\delta f_k \doteq f(x_k) - f^*,$$

$$\delta x_k \doteq x_k - \bar{x}_0,$$

$$\delta y_k \doteq y_k - \bar{x}_0,$$

Inequalities (21a) and (21b) in an incremental notation, are

$$\delta f_k - \delta f_{k+1} \ge \frac{1}{2} \|g_k\|_*^2 - \langle g_k, \delta y_k - \delta x_k \rangle, \ \forall k \ge 1, \ (22a)$$
$$-\delta f_{k+1} \ge \frac{1}{2} \|g_k\|_*^2 - \langle g_k, \delta y_k \rangle, \ \forall k \ge 1.$$
(22b)

We introduce now the auxiliary variable Γ_k , defined as

$$\Gamma_k \doteq t_{k-1}^2 \delta f_k - t_k^2 \delta f_{k+1}, \ \forall k \ge 1.$$

From Property 9 in appendix E we have

$$t_{k-1}^2 = t_k^2 - t_k, \ \forall k \ge 1.$$

We now use this identity to obtain

$$\Gamma_{k} = (t_{k}^{2} - t_{k})\delta f_{k} - t_{k}^{2}\delta f_{k+1}
= (t_{k}^{2} - t_{k})(\delta f_{k} - \delta f_{k+1}) - t_{k}\delta f_{k+1}, \ \forall k \ge 1.$$
(23)

In view of Property 9, $t_k \ge 1$, $\forall k \ge 0$. This implies that we can replace, in inequality (23), $\delta f_k - \delta f_{k+1}$ and $-\delta f_{k+1}$ by the lower bounds given by inequalities (22a) and (22b). In this way we obtain

$$\Gamma_k \ge (t_k^2 - t_k) \left(\frac{1}{2} \|g_k\|_*^2 - \langle g_k, \delta y_k - \delta x_k \rangle \right)$$
$$+ t_k \left(\frac{1}{2} \|g_k\|_*^2 - \langle g_k, \delta y_k \rangle \right)$$
$$= \frac{t_k^2}{2} \|g_k\|_*^2 - \langle g_k, t_k^2 (\delta y_k - \delta x_k) + t_k \delta x_k \rangle, \forall k \ge 1.(24)$$

From step 6 of the algorithm we have for all $k \ge 1$ that $y_k = x_k + \frac{t_{k-1} - 1}{t_k}(x_k - x_{k-1})$. This can be rewritten in incremental notation as

$$\delta y_k - \delta x_k = \frac{t_{k-1} - 1}{t_k} (\delta x_k - \delta x_{k-1}), \ \forall k \ge 1.$$
 (25)

We now define, for every $k \ge 1$

$$s_k \doteq \delta x_{k-1} + t_{k-1} (\delta x_k - \delta x_{k-1}).$$
 (26)

From the definition of s_k and (25) we obtain

$$s_{k} - \delta x_{k} = \delta x_{k-1} + t_{k-1} (\delta x_{k} - \delta x_{k-1}) - \delta x_{k}$$

= $(t_{k-1} - 1) (\delta x_{k} - \delta x_{k-1})$
= $t_{k} (\delta y_{k} - \delta x_{k}), \forall k \ge 1.$ (27)

From (24) and (27) we obtain

$$\Gamma_{k} \geq \frac{1}{2} \| t_{k} g_{k} \|_{*}^{2} - \langle g_{k}, t_{k} (s_{k} - \delta x_{k}) + t_{k} \delta x_{k} \rangle$$

= $\frac{1}{2} \| t_{k} g_{k} \|_{*}^{2} - \langle t_{k} g_{k}, s_{k} \rangle, \ \forall k \geq 1.$ (28)

Using (26) and (27) we now show that g_k can be written in terms of s_k and s_{k+1} .

$$t_{k}g_{k} = t_{k}R(y_{k} - x_{k+1}) = t_{k}R(\delta y_{k} - \delta x_{k+1}) = t_{k}R(\delta y_{k} - \delta x_{k} + \delta x_{k} - \delta x_{k+1}) = R(s_{k} - \delta x_{k} + t_{k}(\delta x_{k} - \delta x_{k+1})) = R(s_{k} - s_{k+1}), \forall k \ge 1.$$
(29)

With this expression for $t_k g_k$ we obtain from (28)

$$\begin{split} \Gamma_k &\geq \frac{1}{2} \|R(s_k - s_{k+1})\|_*^2 - \langle R(s_k - s_{k+1}), s_k \rangle \\ &= \frac{1}{2} \|s_{k+1} - s_k\|_R^2 + \langle R(s_{k+1} - s_k), s_k \rangle \\ &= \frac{1}{2} \|(s_{k+1} - s_k) + s_k\|_R^2 - \frac{1}{2} \|s_k\|_R^2 \\ &= \frac{1}{2} \|s_{k+1}\|_R^2 - \frac{1}{2} \|s_k\|_R^2, \ \forall k \geq 1. \end{split}$$

Thus, for every $k \ge 1$,

$$\Gamma_k = t_{k-1}^2 \delta f_k - t_k^2 \delta f_{k+1} \ge \frac{1}{2} \|s_{k+1}\|_R^2 - \frac{1}{2} \|s_k\|_R^2.$$

Equivalently

$$t_k^2 \delta f_{k+1} + \frac{1}{2} \|s_{k+1}\|_R^2 \le t_{k-1}^2 \delta f_k + \frac{1}{2} \|s_k\|_R^2, \ \forall k \ge 1.$$

Since this inequality holds for every $k \ge 1$ we can apply it in a recursive way to obtain

$$t_k^2 \delta f_{k+1} + \frac{1}{2} \|s_{k+1}\|_R^2 \le t_0^2 \delta f_1 + \frac{1}{2} \|s_1\|_R^2$$

= $\delta f_1 + \frac{1}{2} \|\delta x_0 + t_0 (\delta x_1 - \delta x_0)\|_R^2$
= $\delta f_1 + \frac{1}{2} \|x_1 - \bar{x}_0\|_R^2, \ \forall k \ge 1.$

From (20) we have

$$f(x_1) - f^* + \frac{1}{2} \|x_1 - \bar{x}_0\|_R^2 \le \frac{1}{2} \|x_0 - \bar{x}_0\|_R^2.$$

Thus,

$$t_k^2 \delta f_{k+1} + \frac{1}{2} \|s_{k+1}\|_R^2 \le \frac{1}{2} \|x_0 - \bar{x}_0\|_R^2, \ \forall k \ge 1.$$
(30)

Therefore,

 $t_k^2(f(x_{k+1}) - f^*) + \frac{1}{2} \|s_{k+1}\|_R^2 \le \frac{1}{2} \|x_0 - \bar{x}_0\|_R^2, \ \forall k \ge 1.$

From this inequality, and taking now into account that $t_k \ge \frac{k+2}{2}$ for all $k \ge 0$ (second claim of Property 9), we conclude

$$f(x_{k+1}) - f^* \le \frac{\|x_0 - \bar{x}_0\|_R^2}{2t_k^2} \le \frac{2\|x_0 - \bar{x}_0\|_R^2}{(k+2)^2}, \ \forall k \ge 1.$$

That is,

$$f(x_k) - f^* \le \frac{2\|x_0 - \bar{x}_0\|_R^2}{(k+1)^2}, \ \forall k \ge 2$$

Second claim:

We first prove the claim for k = 0.

$$\begin{aligned} \|g(y_0)\|_* &= \|R(y_0 - y_0^+)\|_* = \|y_0 - y_0^+\|_R \\ &= \|x_0 - x_1\|_R = \|x_0 - \bar{x}_0 + \bar{x}_0 - x_1\|_R \\ &\leq \|x_0 - \bar{x}_0\|_R + \|x_1 - \bar{x}_0\|_R. \end{aligned}$$

From (20) we derive

$$\|x_1 - \bar{x}_0\|_R \le \|x_0 - \bar{x}_0\|_R.$$
(31)

Thus,

$$||g(y_0)||_* \le ||x_0 - \bar{x}_0||_R + ||x_1 - \bar{x}_0||_R \le 2||x_0 - \bar{x}_0||_R.$$

We now prove the claim for k > 0. From (30) we also have

$$\|s_{k+1}\|_{R} \le \|x_0 - \bar{x}_0\|_{R}, \ \forall k \ge 1.$$
(32)

We also have that

$$s_1 = \delta x_0 + t_0 (\delta x_1 - \delta x_0) = x_1 - \bar{x}_0.$$
(33)

From (31) we derive $||s_1||_R = ||x_1 - \bar{x}_0||_R \le ||x_0 - \bar{x}_0||_R$. From this and (32) we obtain

$$\|s_k\|_R \le \|x_0 - \bar{x}_0\|_R, \ \forall k \ge 1.$$
(34)

1.

From here we derive, for every $k \ge 1$,

$$\|s_{k+1} - s_k\|_R \le \|s_{k+1}\|_R + \|s_k\|_R$$

 $\leq \|x_0 - \bar{x}_0\|_R + \|x_0 - \bar{x}_0\|_R = 2\|x_0 - \bar{x}_0\|_R.$ From (29) we have

$$g_k = \frac{1}{t_k} R(s_k - s_{k+1}), \forall k \ge$$

Therefore, for every $k \ge 1$

$$||g_k||_* = \frac{1}{t_k} ||s_k - s_{k+1}||_R$$

$$\leq \frac{2}{t_k} ||x_0 - \bar{x}_0||_R$$

$$\leq \frac{4}{k+2} ||x_0 - \bar{x}_0||_R.$$

We notice that the last inequality is due to the second claim of Property 9. This proves the second claim of the property.

E. Properties of the sequence $\{t_k\}$

Property 9. Let us suppose that $t_0 = 1$ and that

$$t_k \doteq \frac{1}{2} \left(1 + \sqrt{1 + 4t_{k-1}^2} \right), \ \forall k \ge 1.$$

Then

(i)
$$t_{k-1}^2 = t_k^2 - t_k$$
, for all $k \ge 1$.
(ii) $t_k \ge \frac{k+2}{2} \ge 1$, for all $k \ge 0$.

Proof.

(i) For every $k \ge 1$, t_k is defined as one of the roots of

$$t_k^2 - t_k - t_{k-1}^2 = 0.$$

Therefore we obtain $t_{k-1}^2 = t_k^2 - t_k$.

(ii) The claim is trivially satisfied for k equal to 0. We now show that if the claim is satisfied for k - 1 then it is also satisfied for k.

$$t_k = \frac{1}{2} \left(1 + \sqrt{1 + 4t_{k-1}^2} \right)$$

$$\geq \frac{1}{2} \left(1 + \sqrt{4t_{k-1}^2} \right) = \frac{1}{2} + t_{k-1}.$$

Since the claim is assumed to be satisfied for k-1 we have $t_{k-1} \ge \frac{k+1}{2}$ and consequently

$$t_k \ge \frac{1}{2} + \frac{k+1}{2} = \frac{k+2}{2}.$$

F. Proof of Property 2

From equation (5) we have

$$f(x_k) - f^* \le \frac{2}{(k+1)^2} ||x_0 - \bar{x}_0||_R^2, \ \forall k \ge 1.$$

Due to Assumption 2 we also have

$$\frac{\mu}{2} \|x_0 - \bar{x}_0\|_R^2 \le f(x_0) - f^*.$$

Therefore,

$$f(x_k) - f^* \le \frac{4}{\mu(k+1)^2} (f(x_0) - f^*), \ \forall k \ge 1.$$
 (35)

This proves the first claim. Denote

$$\alpha_k \doteq \frac{4}{\mu(k+1)^2}, \quad \forall k \ge 1$$

With this notation we rewrite (35) as

$$f(x_k) - f^* \le \alpha_k (f(x_0) - f^*), \ \forall k \ge 1.$$
 (36)

Suppose now that $k \ge \left\lfloor \frac{2}{\sqrt{\mu}} \right\rfloor$. Then,

$$\alpha_k = \frac{4}{\mu(k+1)^2} \le \frac{4}{\mu\left(\left\lfloor\frac{2}{\sqrt{\mu}}\right\rfloor + 1\right)^2} < \frac{4}{\mu\left(\left\lfloor\frac{2}{\sqrt{\mu}}\right\rfloor^2 = 1.$$

Therefore,

$$\alpha_k \in (0,1), \ \forall k \ge \left\lfloor \frac{2}{\sqrt{\mu}} \right\rfloor.$$
 (37)

This, along with inequality (36), yields

$$f(x_k) - f^* \le f(x_0) - f^*, \ \forall k \ge \left\lfloor \frac{2}{\sqrt{\mu}} \right\rfloor.$$

Equivalently,

$$f(x_k) \le f(x_0), \ \forall k \ge \left\lfloor \frac{2}{\sqrt{\mu}} \right\rfloor.$$

This proves the second claim of the property. In view of inequality (36) we have

$$f(x_k) - f^* \le \alpha_k (f(x_0) - f^*)$$

= $\alpha_k (f(x_0) - f(x_k) + f(x_k) - f^*)$
= $\alpha_k (f(x_0) - f(x_k)) + \alpha_k (f(x_k) - f^*).$

Therefore,

$$(1 - \alpha_k)(f(x_k) - f^*) \le \alpha_k(f(x_0) - f(x_k)).$$
(38)

Suppose now that $k \geq \left\lfloor \frac{2\sqrt{e+1}}{\sqrt{\mu}} \right\rfloor$. This implies $k \geq \left\lfloor \frac{2}{\sqrt{\mu}} \right\rfloor$ and consequently $1 - \alpha_k > 0$ (see (37)). Dividing both terms of inequality (38) by $1 - \alpha_k$, we get

$$f(x_k) - f^* \le \frac{\alpha_k}{1 - \alpha_k} (f(x_0) - f(x_k))$$

$$= \frac{\frac{4}{\mu(k+1)^2}}{1 - \frac{4}{\mu(k+1)^2}} (f(x_0) - f(x_k))$$

$$= \frac{4(f(x_0) - f(x_k))}{\mu(k+1)^2 - 4}$$

$$\le \frac{4(f(x_0) - f(x_k))}{\mu(\left\lfloor \frac{2\sqrt{e+1}}{\sqrt{\mu}} \right\rfloor + 1)^2 - 4}$$

$$\le \frac{4(f(x_0) - f(x_k))}{\mu(\frac{2\sqrt{e+1}}{\sqrt{\mu}})^2 - 4}$$

$$= \frac{4(f(x_0) - f(x_k))}{4(e+1) - 4} = \frac{f(x_0) - f(x_k)}{e}$$

G. Proof of Property 3

By construction, $r_{j-1} \in \mathcal{X}$, for all $j \ge 1$. Therefore, we have from the second claim of Property 1, that

$$\frac{1}{2} \|g(r_{j-1})\|_*^2 \le f(r_{j-1}) - f(r_{j-1}^+), \ \forall j \ge 1.$$
(39)

We also notice that r_j is computed invoking FISTA algorithm using r_{j-1} as initial condition ($z = r_{j-1}$). That is,

$$[r_j, n_j] = FISTA(r_{j-1}, n_{j-1}, E_c^l).$$

Since the output value $f(r_j)$ is forced to be no larger than the one corresponding to $x_0 = z^+ = r_{j-1}^+$, we have $f(r_j) \le f(r_{j-1}^+)$. Therefore, we obtain from inequality (39) that

$$\frac{1}{2} \|g(r_{j-1})\|_*^2 \le f(r_{j-1}) - f(r_{j-1}^+) \\ \le f(r_{j-1}) - f(r_j).$$

This proves the first claim of the property. We now show that if $n_{j-1} \leq \frac{4\sqrt{e+1}}{\sqrt{\mu}}$, then the value n_j obtained from

$$[r_j, n_j] = FISTA(r_{j-1}, n_{j-1}, E_c^l),$$

also satisfies

$$n_j \le \frac{4\sqrt{e+1}}{\sqrt{\mu}}.\tag{40}$$

Denote

$$\bar{m} = \left\lfloor \frac{2\sqrt{e+1}}{\sqrt{\mu}} \right\rfloor$$

Since $\bar{m} \geq \left\lfloor \frac{2\sqrt{e+1}}{\sqrt{\mu}} \right\rfloor$, we infer, from the third claim of Property 2, that

$$f(x_{\bar{m}}) - f^* \le \frac{f(x_0) - f(x_{\bar{m}})}{e}$$

From this inequality, we obtain

$$f(x_{\bar{m}}) - f(x_k) \le f(x_{\bar{m}}) - f^* \le \frac{f(x_0) - f(x_{\bar{m}})}{e}$$

Therefore, the first exit condition is satisfied for $m = \bar{m}$. Since $m = \lfloor \frac{k}{2} \rfloor + 1$ we have $m \ge \frac{k}{2}$. This means that for $m = \bar{m}$, the corresponding value for k is no larger than

$$2\bar{m} = 2\left\lfloor \frac{2\sqrt{e+1}}{\sqrt{\mu}} \right\rfloor \le \frac{4(\sqrt{e+1})}{\sqrt{\mu}}.$$

We also notice that, in view of the second claim of Property 2, the additional exit condition $f(x_k) \leq f(x_0)$ is satisfied for every

$$k \ge \left\lfloor \frac{2}{\sqrt{\mu}} \right\rfloor.$$

Therefore, $n_{j-1} \leq \frac{4\sqrt{e+1}}{\sqrt{\mu}}$ implies that n_j , obtained from $[r_j, n_j] = FISTA(r_{j-1}, n_{j-1}, E_c^l)$, also satisfies (40). We now prove, by reduction to the absurd, that n_j cannot be larger than $\frac{4\sqrt{e+1}}{\sqrt{\mu}}$. Suppose that

$$n_j > \frac{4\sqrt{e+1}}{\sqrt{\mu}}.\tag{41}$$

Because of the previous discussion, the previous inequality could be forced only by the doubling step $n_j = 2n_{j-1}$ of the algorithm. That is, inequality (41) is possible only if there is s such that $n_{s-1} > \frac{2\sqrt{e+1}}{\sqrt{\mu}}$ and

$$f(r_{s-1}) - f(r_s) > \frac{f(r_{s-2}) - f(r_{s-1})}{e}$$

Since

$$[r_{s-1}, n_{s-1}] = FISTA(r_{s-2}, n_{s-2}, E_c^l)$$

we have that r_{s-1} is obtained from r_{s-2} applying

$$n_{s-1} > \frac{2\sqrt{e+1}}{\sqrt{\mu}}$$

iterations of FISTA algorithm. However, we have from the third claim of Property 2 that this number of iterations implies

$$f(r_{s-1}) - f(r_s) \le f(r_{s-1}) - f^* \le \frac{f(r_{s-2}^+) - f(r_{s-1})}{e}.$$

From the second claim of Property 1 we also have $f(r_{s-2}^+) \leq f(r_{s-2})$. Thus,

$$f(r_{s-1}) - f(r_s) \le \frac{f(r_{s-2}) - f(r_{s-1})}{e}.$$

That is, there is no doubling step if $n_{s-1} \geq \frac{2\sqrt{e+1}}{\sqrt{\mu}}$. This proves the second claim of the property.

We now show that there is a doubling step at least every

$$T \doteq \left\lceil \ln \left(1 + \frac{2(f(r_0) - f^*)}{\epsilon^2} \right) \right\rceil$$

steps of the algorithm. Suppose that there is no doubling step from iteration j = s + 1 to j = s + T, where $s \ge 1$. That is,

$$f(r_{j-1}) - f(r_j) \le \frac{f(r_{j-2}) - f(r_{j-1})}{e}, \ \forall j \in [s+1, s+T].$$

From this, and the first claim of the property, we obtain the following sequence of inequalities

$$\frac{1}{2} \|g(r_{s+T-1})\|_{*}^{2} \leq f(r_{s+T-1}) - f(r_{s+T}) \\
\leq \frac{f(r_{s+T-2}) - f(r_{s+T-1})}{e} \leq \left(\frac{1}{e}\right)^{T} (f(r_{s-1}) - f(r_{s})) \\
\leq \left(\frac{1}{e}\right)^{T} (f(r_{s-1}) - f^{*}) \leq \left(\frac{1}{e}\right)^{T} (f(r_{0}) - f^{*}) \\
= \left(\frac{1}{e}\right)^{\left\lceil \ln\left(1 + \frac{2(f(r_{0}) - f^{*})}{e^{2}}\right)\right\rceil} (f(r_{0}) - f^{*}) \\
\leq \left(\frac{1}{e}\right)^{\ln\left(1 + \frac{2(f(r_{0}) - f^{*})}{e^{2}}\right)} (f(r_{0}) - f^{*}) \\
= \left(\frac{1}{1 + \frac{2(f(r_{0}) - f^{*})}{e^{2}}\right) (f(r_{0}) - f^{*}) \leq \frac{\epsilon^{2}}{2}.$$

We conclude that T consecutive iterations without doubling step implies that the exit condition is satisfied $(||g(r_{s+T-1})||_* \le \epsilon)$. We conclude that there must be at least one doubling step every T iterations. This implies that there exist $j \in [s+1, s+T]$ such that

$$f(r_{j-1}) - f(r_j) > \frac{f(r_{j-2}) - f(r_{j-1})}{e}.$$

Therefore, $n_j = 2n_{j-1}$. Moreover, since $\{n_j\}$ is a non decreasing sequence, we get $n_{s+T} \ge n_j = 2n_{j-1} \ge 2n_s$, $\forall s \ge 1$. That is,

$$n_s \le \frac{n_{s+T}}{2}, \ \forall s \ge 1.$$
(42)

Suppose that j is rewritten as j = m + nT, where $0 \le m < T$ and $n \ge 0$. From the non decreasing nature of $\{n_j\}$,

$$\sum_{i=0}^{j} n_{i} = \sum_{i=0}^{m+nT} n_{i} = \sum_{i=0}^{m} n_{i} + \sum_{\ell=0}^{n-1} \sum_{i=1}^{T} n_{m+i+\ell T}$$

$$\leq Tn_{m} + T \sum_{\ell=1}^{n} n_{m+\ell T} = T \sum_{\ell=0}^{n} n_{m+\ell T} = T \sum_{\ell=0}^{n} n_{j-\ell T}.$$
(43)

Also, from inequality (42), we have $n_{j-T} \leq \frac{n_j}{2}$. Using this inequality in a recursive manner we obtain

$$n_{j-\ell T} \leq \left(\frac{1}{2}\right)^{\ell} n_j, \ \ell = 0, \dots, n.$$

This, allows us to infer from (43) that

$$\sum_{i=0}^{j} n_i \le T \sum_{\ell=0}^{n} \left(\frac{1}{2}\right)^{\ell} n_j \le T \sum_{\ell=0}^{\infty} \left(\frac{1}{2}\right)^{\ell} n_j = 2Tn_j.$$

The last claim of the property follows directly from this one and the bound $n_j \leq \frac{4\sqrt{e+1}}{\sqrt{\mu}}$ of the second claim. That is, if j denotes the first index for which $||g(r_j)||_* \leq \epsilon$, we get that the number of total iterations is bounded by

$$\sum_{i=0}^{j} n_i \le 2Tn_j \le \frac{8T\sqrt{e+1}}{\sqrt{\mu}} \le \frac{16T}{\sqrt{\mu}}$$
$$= \frac{16}{\sqrt{\mu}} \left[\ln\left(1 + \frac{2(f(r_0) - f^*)}{\epsilon^2}\right) \right]$$