FISTA and Extensions – Review and New Insights

Weiwei Kong, Jefferson G. Melo, and Renato D.C. Monteiro

July 2021

1 Introduction

The purpose of this technical report is to review the main properties of an accelerated composite gradient (ACG) method commonly referred to as the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA). In addition, we state a version of FISTA for solving both convex and strongly convex composite minimization problems and derive its iteration complexities to generate iterates satisfying various stopping criteria, including one which arises in the course of solving other composite optimization problems via inexact proximal point schemes. This report also discusses different reformulations of the convex version of FISTA and how they relate to other formulations in the literature.

Organization. Section 2 contains three subsections. The first one describes a composite optimization problem and its main assumptions. The second subsection states and analyze a variant of FISTA, called S-FISTA, for solving the aforementioned problem. The third subsection establishes some iteration-complexity bounds for S-FISTA to obtain approximate stationary solution for the composite optimization problem we are interested in. Section 3 presents an alternative formulation for S-FISTA and shows that it becomes the well-known FISTA for solving (non strongly) convex composite optimization problems.

1.1 A Brief History of FISTA

An earlier prototype of FISTA was given in [4], which proposed an ACG method named the Fast Gradient Method (FGM) for solving smooth convex (non-composite) optimization problems. FISTA, which is an extension of [4] to smooth convex composite optimization problems, was then proposed in [2]. Its monotonically decreasing variant called M-FISTA was later proposed in [3].

2 A Strongly Convex Extension of FISTA

This section contains three subsections. The first one describes a composite optimization problem and its main assumptions. The second subsection states and analyze a variant of FISTA, called S-FISTA, for solving the aforementioned problem. The third subsection establishes some iteration-complexity bounds for S-FISTA to obtain approximate stationary solution for the composite optimization problem we are interested in.

2.1 Problem Description and Assumptions

We are interested in the following problem

$$\phi^* := \min\{\phi(x) := f(x) + h(x) : x \in \mathbb{R}^n\} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable $\bar{\mu}_f$-convex function and $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a possibly nonsmooth $\bar{\mu}_h$-convex function, and $\bar{\mu}_f, \bar{\mu}_h \geq 0$.

In addition, the following assumptions are made.

(A) Problem (1) has an optimal solution.

(B) There exists a scalar $\bar{L}_f \geq \bar{\mu}_f$ such that

$$f(\cdot) \leq l_f(\cdot, z) + \frac{\bar{L}_f}{2}\|\cdot - z\|^2, \tag{2}$$

where

$$l_f(\cdot, z) := f(z) + \langle \nabla f(z), \cdot - z \rangle. \tag{3}$$

Clearly the following inclusion holds for any solution x^* of (1):

$$0 \in \nabla f(x^*) + \partial h(x^*).$$

For a given tolerance $\rho > 0$, we say that a pair $(y, u) \in \mathbb{R}^n \times \mathbb{R}^n$ is a ρ-approximate stationary solution for problem (1) if the following relations hold

$$u \in \nabla f(y) + \partial h(y), \quad \|u\| \leq \rho. \tag{4}$$

Next, we introduce a scalar which measures the distance of the initial point x_0 to the solution set of (1).

$$d_0 := \min\{\|x_0 - x^*\| : x^* \text{ is a solution of } (1)\}. \tag{5}$$

Recall that if a function $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is ν-convex then, for every z^* that minimizes ψ, we have

$$\psi(z^*) + \frac{\nu}{2}\|\cdot - z^*\|^2 \leq \psi(\cdot). \tag{6}$$

Moreover, since f is $\bar{\mu}_f$-convex, the following inequality holds for any $z \in \mathbb{R}^n$:

$$l_f(\cdot, z) + \frac{\bar{\mu}_f}{2}\|\cdot - z\|^2 \leq f(\cdot), \tag{7}$$

where $l_f(\cdot, z)$ is as in (3).

Throughout this note we use the following notation $\log_1^+(\cdot) := \max\{\log(\cdot), 1\}$.

2.2 Statement and Properties of S-FISTA

Recall that FISTA is a popular ACG variant for solving (1) for the case where $\bar{\mu}_f = \bar{\mu}_h = 0$. This subsection describes an extension of FISTA for solving (1) for the general case where $\bar{\mu}_f, \bar{\mu}_h \geq 0$.

We start by stating a strongly convex variant of FISTA, referred to as S-FISTA, for solving (1).

Algorithm 1 (S-FISTA)

---

0. Let initial point $x_0 \in \operatorname{dom} h$ and scalars $L_f > \bar{L}_f$, $\mu_f \in [0, \bar{\mu}_f]$ and $\mu_h \in [0, \bar{\mu}_h]$ be given, and set $x_0 = y_0$, $A_0 = 0$, $\tau_0 = 1$, $\lambda = 1/(L_f - \mu_f)$, $\mu = \mu_f + \mu_h$, and $k = 0$;
1. Compute

$$a_k = \frac{\lambda\tau_k + \sqrt{(\lambda\tau_k)^2 + 4\lambda\tau_k A_k}}{2}, \quad A_{k+1} = A_k + a_k, \quad \tilde{x}_k = \frac{A_k y_k + a_k x_k}{A_{k+1}}; \tag{8}$$

2. Compute

$$y_{k+1} := \underset{x \in \operatorname{dom} h}{\operatorname{argmin}} \left\{ q_k^L(x; \tilde{x}_k) := \ell_f(x; \tilde{x}_k) + h(x) + \frac{L}{2}\|x - \tilde{x}_k\|^2 \right\}, \tag{9}$$

$$\tau_{k+1} = \tau_k + \mu a_k, \tag{10}$$

$$x_{k+1} = \frac{1}{\tau_{k+1}} \left[\frac{a_k}{\lambda}(y_{k+1} - \tilde{x}_k) + \mu a_k y_{k+1} + \tau_k x_k \right]; \tag{11}$$

3. Set $k \leftarrow k + 1$ and go to step 1.

We now make some comments about S-FISTA. First, if $\mu = 0$, then $\tau_k = 1$ for every $k \geq 0$. Second, the first and second relations in (8) imply that

$$\frac{\tau_k A_{k+1}}{a_k^2} = \frac{1}{\lambda} = L_f - \mu_f. \tag{12}$$

Third, it will be shown in Section 3 that when $\mu = 0$, S-FISTA is actually FISTA.

Next, we present some technical lemmas about S-FISTA.

Lemma 2.1. *For every $k \geq 0$ and $x \in \mathbb{R}^n$, define*

$$\tilde{\gamma}_k(x) := \ell_f(x; \tilde{x}_k) + h(x) + \frac{\mu_f}{2}\|x - \tilde{x}_k\|^2, \tag{13}$$

$$\gamma_k(x) := \tilde{\gamma}_k(y_{k+1}) + \frac{1}{\lambda}\langle \tilde{x}_k - y_{k+1}, x - y_{k+1} \rangle + \frac{\mu}{2}\|x - y_{k+1}\|^2. \tag{14}$$

Then, the following statements hold for every $k \geq 0$:

a) $\tilde{\gamma}_k(y_{k+1}) = \gamma_k(y_{k+1})$*;*

b) $\tilde{\gamma}_k \le \phi$ *and*

$$y_{k+1} = \operatorname*{argmin}_x \left\{ \tilde{\gamma}_k(x) + \frac{1}{2\lambda} \|x - \tilde{x}_k\|^2 \right\}; \tag{15}$$

c) $\gamma_k \le \tilde{\gamma}_k$ *and*

$$y_{k+1} = \min_x \left\{ \gamma_k(x) + \frac{1}{2\lambda} \|x - \tilde{x}_k\|^2 \right\}; \tag{16}$$

d) $x_{k+1} = \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ a_k \gamma_k(x) + \tau_k \|x - x_k\|^2 /2 \right\}$ *and*

$$x_{k+1} = x_k + \frac{a_k}{\tau_{k+1}} \left[\frac{1}{\lambda}(y_{k+1} - \tilde{x}_k) + \mu(y_{k+1} - x_k) \right];$$

e) $\tau_k = 1 + A_k \mu$.

Proof. a) It clearly follows from (14) that $\gamma_k(y_{k+1}) = \tilde{\gamma}_k(y_{k+1})$.

b) The inequality $\tilde{\gamma}_k \le \phi$ follows from the definition of ϕ in (1), (7), and (13). Moreover, (15) follows from (9), (13), and the fact that $\lambda = 1/(L_f - \mu_f)$.

c) Define $\hat{\gamma}_k := \tilde{\gamma}_k - \mu \| \cdot - y_{k+1} \|^2/2$. Since $\tilde{\gamma}_k$ is μ-convex, it follows that $\hat{\gamma}_k$ is convex, and hence that $\partial \tilde{\gamma}_k(y_{k+1}) = \partial \hat{\gamma}_k(y_{k+1})$, in view of the subgradient rule for the sum of two convex functions. Also, the optimality condition for (15) implies that

$$\frac{\tilde{x}_k - y_{k+1}}{\lambda_k} \in \partial \tilde{\gamma}_k(y_{k+1}) = \partial \hat{\gamma}_k(y_{k+1}),$$

which, in view of the definition of $\hat{\gamma}_k$ and its subgradient at y_{k+1}, is easily seen to be equivalent to $\gamma_k \le \tilde{\gamma}_k$. Now, since $\nabla \gamma_k(y_{k+1}) = (\tilde{x}_k - y_{k+1})/\lambda$ by (14), we easily see that y_{k+1} satisfies the optimality condition for (16), and hence (16) in follows.

d) It follows from (14), (10) and (11) that

$$\begin{aligned} a_k \nabla \gamma_k(x_{k+1}) + \tau_k(x_{k+1} - x_k) &= \left[\frac{a_k}{\lambda}(\tilde{x}_k - y_{k+1}) + a_k \mu (x_{k+1} - y_{k+1}) \right] + \tau_k(x_{k+1} - x_k) \\ &= \frac{a_k}{\lambda}(\tilde{x}_k - y_{k+1}) + \tau_{k+1} x_{k+1} - a_k \mu y_{k+1} - \tau_k x_k = 0, \end{aligned}$$

and hence that the first claim in d) follows. The second claim follows from (10) and (11).

e) This identity follows follows immediately from (10) and the second identity in (8). □

Lemma 2.2. *For every* $k \ge 0$ *and* $x \in \operatorname{dom} h$*, we have*

$$q_k^L(x; \tilde{x}_k) \ge \phi(x) + \frac{1}{2}\left(L_f - \bar{L}_f \right) \|x - \tilde{x}_k\|^2$$

where $q_k^L(\cdot;\cdot)$ *is defined in* (9).

Proof. Using the definitions of ϕ and $q_k^L(\cdot;\tilde{x}_k)$ in (1) and (9), respectively, and inequality (2), we have

$$\begin{aligned} q_k^L(x;\tilde{x}_k) &= \left(\ell_f(x;\tilde{x}_k) + \frac{\bar{L}_f}{2}\|x-\tilde{x}_k\|^2\right) + h(x) + \frac{1}{2}\left(L_f - \bar{L}_f\right)\|x-\tilde{x}_k\|^2 \\ &\geq \phi(x) + \frac{1}{2}\left(L_f - \bar{L}_f\right)\|x-\tilde{x}_k\|^2 \end{aligned}$$

for every $x \in \mathbb{R}^n$. □

Lemma 2.3. *For every $k \geq 0$ and $x \in \mathbb{R}^n$, we have*

$$A_k\gamma_k(y_k) + a_k\gamma_k(x) + \frac{1}{2}\|x_k - x\|^2 - \frac{1}{2}\|x_{k+1}-x\|^2 \geq A_{k+1}q_k^L(y_{k+1};\tilde{x}_k). \tag{17}$$

where $\gamma_k(\cdot)$ and $q_k^L(\cdot;\cdot)$ are defined in (14) and (9), respectively.

Proof. Using Lemma 2.1(d), the facts that $\tau_{k+1} = \tau_k + \mu a_k$ (see step 3 of Algorithm 1) and $\psi_k := a_k\gamma_k(\cdot) + \tau_k\|\cdot - \tilde{x}_k\|^2/2$ is $(\tau_k + \mu a_k)$-convex, it follows from (6) with $\psi = \psi_k$ and $\nu = \tau_{k+1}$ that

$$a_k\gamma_k(x) + \frac{\tau_k}{2}\|x - x_k\|^2 - \frac{\tau_{k+1}}{2}\|x - x_{k+1}\|^2 \geq a_k\gamma_k(x_{k+1}) + \frac{\tau_k}{2}\|x_{k+1}-x_k\|^2 \quad \forall x \in \mathbb{R}^n.$$

Using the convexity of γ_k, the definitions of A_{k+1} and $\tilde{x}_k$ in (8), and relation (12), we have

$$\begin{aligned} &A_k\gamma_k(y_k) + a_k\gamma_k(x_{k+1}) + \frac{\tau_k}{2}\|x_{k+1}-x_k\|^2 \\ &\quad\geq A_{k+1}\gamma_k\left(\frac{A_ky_k + a_kx_{k+1}}{A_{k+1}}\right) + \frac{\tau_kA_{k+1}^2}{2a_k^2}\left\|\frac{A_ky_k + a_kx_{k+1}}{A_{k+1}} - \frac{A_ky_k + a_kx_k}{A_{k+1}}\right\|^2 \\ &\quad= A_{k+1}\left[\gamma_k\left(\frac{A_ky_k + a_kx_{k+1}}{A_{k+1}}\right) + \frac{1}{2\lambda}\left\|\frac{A_ky_k + a_kx_{k+1}}{A_{k+1}} - \tilde{x}_k\right\|^2\right] \\ &\quad\geq A_{k+1}\min_x\left\{\gamma_k(x) + \frac{1}{2\lambda}\|x-\tilde{x}_k\|^2\right\} \\ &\quad= A_{k+1}\left[\tilde{\gamma}_k(y_{k+1}) + \frac{L_f-\mu_f}{2}\|y_{k+1}-\tilde{x}_k\|^2\right] = A_{k+1}q_k^L(y_{k+1};\tilde{x}_k) \end{aligned}$$

where the second last equality is due to Lemma 2.1(b) and the fact that $\lambda^{-1} = L_f - \mu_f$, and the last one is due to (13) and the definition of $q_k^L(\cdot;\cdot)$ in (9). The lemma now follows by combining the above two conclusions. □

The next two results provide some important recursive formulas.

Lemma 2.4. *For every $k \geq 0$ and $x \in \mathbb{R}^n$, we have*

$$A_k\phi(y_k) + a_k\gamma_k(x) + \frac{\tau_k}{2}\|x_k - x\|^2 - \frac{\tau_{k+1}}{2}\|x_{k+1}-x\|^2 \tag{18}$$

$$\geq A_{k+1}\phi(y_{k+1}) + \frac{A_{k+1}}{2}\left(L_f - \bar{L}_f\right)\|y_{k+1}-\tilde{x}_k\|^2. \tag{19}$$

where $\gamma_k(\cdot)$ is defined in (14).

Proof. The conclusion of this result follows immediately from Lemma 2.2 with $x = y_{k+1}$ and Lemma 2.3. $\square$

Lemma 2.5. *For every $k \geq 0$ and $x \in \operatorname{dom} h$, we have*

$$\eta_k(x) - \eta_{k+1}(x) \geq \frac{A_{k+1}}{2} \left(L_f - \bar{L}_f\right) \|\tilde{y}_{k+1} - \tilde{x}_k\|^2$$

where

$$\eta_k(x) := A_k[\phi(y_k) - \phi(x)] + \frac{\tau_k}{2}\|x - x_k\|^2.$$

Proof. Using Lemma 2.3 and the fact that $\gamma_k \leq \phi$ by Lemma 2.1(b)-(c), we have

$$\begin{aligned} A_k\phi(y_k) + a_k\phi(x) + \frac{\tau_k}{2}\|x_k - x\|^2 - \frac{\tau_{k+1}}{2}\|x_{k+1} - x\|^2 \\ \geq A_{k+1}\phi(y_{k+1}) + \frac{A_{k+1}}{2}\left(L_f - \bar{L}_f\right)\|\tilde{y}_{k+1} - \tilde{x}_k\|^2. \end{aligned}$$

The conclusion of the lemma now follows by subtracting $A_{k+1}\phi(x)$ from both sides of the above inequality, and using the identity $A_{k+1} = A_k + a_k$ and the definition of $\eta_k(x)$. $\square$

Next, we state a basic result that will be useful in deriving complexity bounds for S-FISTA.

Lemma 2.6. *For every $k \geq 0$ and $x \in \operatorname{dom} h$, we have*

$$A_k[\phi(y_k) - \phi(x)] + \frac{\tau_k}{2}\|x - x_k\|^2 \leq \frac{1}{2}\|x - x_0\|^2 - \frac{1}{2}\left(L_f - \bar{L}_f\right)\sum_{i=0}^{k-1} A_{i+1}\|\tilde{y}_{i+1} - \tilde{x}_i\|^2.$$

Proof. This result follows by summing the inequality of Lemma 2.5 from $k = 0$ to $k = k-1$, and using the fact that $A_0 = 0$ and the definition of $\eta_k(\cdot)$ in Lemma 2.5. $\square$

The below result gives some estimates on the sequence $\{A_k\}$.

Lemma 2.7. *For every $k \geq 1$, we have*

$$A_k \geq \frac{1}{L_f - \mu_f} \max\left\{\frac{k^2}{4}, \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L_f - \mu_f}}\right)^{2(k-1)}\right\}. \tag{20}$$

As a consequence, for a given $\bar{A} > 0$, we have $A_k \geq \bar{A}$ as long as

$$k \geq \min\left\{2\sqrt{(L_f - \mu_f)\bar{A}},\ \left[\frac{1}{2} + \sqrt{\frac{L_f - \mu_f}{\mu}}\right] \log_1^+((L_f - \mu_f)\bar{A}) + 1\right\}. \tag{21}$$

Proof. The first and second identities in (8) imply that

$$A_{k+1} = A_k + a_k \geq A_k + \left(\frac{\tau_k\lambda}{2} + \sqrt{\tau_k\lambda A_k}\right) \geq \left(\sqrt{A_k} + \frac{1}{2}\sqrt{\tau_k\lambda}\right)^2$$

which, together with the fact that $\tau_k = 1 + \mu A_k$, yields

$$\sqrt{A_{k+1}} \geq \sqrt{A_k} + \frac{1}{2}\sqrt{\tau_k \lambda} = \sqrt{A_k} + \frac{1}{2}\sqrt{(1+\mu A_k)\lambda}.$$

Clearly, the last inequality implies the two inequalities

$$\sqrt{A_{k+1}} \geq \sqrt{A_k} + \frac{1}{2}\sqrt{\lambda}, \quad \sqrt{A_{k+1}} \geq \sqrt{A_k}\left(1 + \frac{1}{2}\sqrt{\mu\lambda}\right).$$

The first bound in (20) follows by summing the first inequality from $k = 0$ to $k = k-1$, and using the fact that $A_0 = 0$ and $\lambda = 1/(L_f - \mu_f)$ (see step 0 of S-FISTA). The second bound in (20) follows by successively using the second inequality from $k = 1$ to $k = k-1$ and using the fact that $A_1 = \lambda$.

Now to prove the last statement of the lemma note that (20) implies that in order to have $A_k \geq \bar{A}$, it is sufficient to have

$$\frac{1}{L_f - \mu_f} \max\left\{\frac{k^2}{4}, \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L_f - \mu_f}}\right)^{2(k-1)}\right\} \geq \bar{A}.$$

Clearly, the above condition is satisfied if one of the following conditions holds

$$k \geq 2\sqrt{(L_f - \mu_f)\bar{A}}, \qquad \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L_f - \mu_f}}\right)^{2(k-1)} \geq (L_f - \mu_f)\bar{A}.$$

The latter inequality is equivalent to

$$2(k-1)\log\left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L_f - \mu_f}}\right) \geq \log\left((L_f - \mu_f)\bar{A}\right).$$

Since $\log(1+x) \geq 1/(1+x^{-1})$ for any $x > 0$, it follows by using $x = \sqrt{\mu}/[2\sqrt{L_f - \mu_f}]$ that the above condition holds if

$$2(k-1)\left(\frac{1}{1 + \frac{2\sqrt{L_f - \mu_f}}{\sqrt{\mu}}}\right) \geq \log_1^+\left((L_f - \mu_f)\bar{A}\right)$$

which immediately proves the last statement of the lemma. □

The below result establishes a convergence rate and iteration-complexity bounds for S-FISTA to obtain a approximate (function value) solution of (1).

Proposition 2.8. *For every $k \geq 1$, we have*

$$\phi(y_k) - \phi^* \leq \frac{(L_f - \mu_f)d_0^2}{2} \min\left\{\frac{4}{k^2}, \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L_f - \mu_f}}\right)^{2(1-k)}\right\} \tag{22}$$

where ϕ^* and d_0 are as in (1) and (5), respectively. As a consequence, for any given $\bar{\varepsilon} > 0$, S-FISTA finds a point $y := y_k$ satisfying $\phi(y) - \phi^* \le \bar{\varepsilon}$ in at most

$$\mathcal{O}\left(\min\left\{d_0\sqrt{\frac{L_f - \mu_f}{\bar{\varepsilon}}}, \sqrt{\frac{L_f - \mu_f}{\mu}}\log_1^+\left(\frac{(L_f - \mu_f)d_0^2}{\bar{\varepsilon}}\right)\right\}\right)$$

iterations.

Proof. It follows from Lemma 2.6 with x^* such that $d_0 = \|x_0 - x^*\|$ that, for every $k \ge 0$,

$$\phi(y_k) - \phi^* \le \frac{1}{2A_k}d_0^2.$$

Hence, (22) follows immediately from (20). The last statement of the proposition follows immediately from the latter inequality and the last statement of Lemma 2.7 with $\bar{A} = d_0^2/(2\bar{\varepsilon})$. $\square$

2.3 Stationarity Complexity Bounds

This subsection is devoted to the study of iteration-complexity bounds for S-FISTA to compute several different notions of an approximate stationary solution of (1).

We start by establishing an iteration-complexity bound for S-FISTA to obtain an approximate stationary solution of (1) based on the generalized subdifferential of ϕ.

Lemma 2.9. *Assume that ∇f is L-Lipschitz continuous and define*

$$u_k = \nabla f(y_k) - \nabla f(\tilde{x}_{k-1}) + L_f(\tilde{x}_{k-1} - y_k).$$

Then, the following statements hold:

a) for every $k \ge 1$,

$$u_k \in \nabla f(y_k) + \partial h(y_k), \quad \min_{1\le i\le k}\|u_i\|^2 \le \frac{8L_f^2 d_0^2}{(L_f - \bar{L}_f)\sum_{i=1}^k A_i}; \tag{23}$$

b) for any $\rho > 0$, S-FISTA generates a ρ-approximate stationary solution pair $(y, u) := (y_k, u_k)$ in at most

$$\left\lceil \min\left\{\left(\frac{12\zeta d_0^2}{\rho^2}\right)^{1/3}, \left(1 + \frac{2\sqrt{L_f - \mu_f}}{\sqrt{\mu}}\right)\log\left(1 + \frac{\zeta(c^2-1)d_0^2}{\rho^2}\right)\right\}\right\rceil$$

iterations, where

$$\zeta = \zeta(\mu_f, L_f, \bar{L}_f) := \frac{8L_f^2(L_f - \mu_f)}{L_f - \bar{L}_f}, \qquad c = c(\mu_f, \mu, L_f) = 1 + \frac{1}{2}\sqrt{\frac{\mu}{L_f - \mu_f}}. \tag{24}$$

Proof. a) It follows from (9) with $k = k-1$ and its associated optimality condition that

$$0 \in \nabla f(\tilde{x}_{k-1}) + \partial h(y_k) + L_f(y_k - \tilde{x}_{k-1})$$

which, in view of the definition of u_k, immediately implies the inclusion of the lemma. Using the definition of u_k, assumption that ∇f is L_f-Lipschitz continuous on $\mathbb{R}^n$, and the triangle inequality for norms, it follows that

$$\|u_k\| \leq \|\nabla f(y_k) - \nabla f(\tilde{x}_{k-1})\| + L_f\|y_k - \tilde{x}_{k-1}\| \leq 2L_f\|y_k - \tilde{x}_{k-1}\|.$$

Now using Lemma 2.6 with $x = x^*$ where $d_0 = \|x_0 - x^*\|$, we conclude that

$$d_0^2 \geq \frac{1}{2}\left(L_f - \bar{L}_f\right)\sum_{i=1}^{k} A_i\|\tilde{y}_i - \tilde{x}_{i-1}\|^2 \geq \frac{L_f - \bar{L}_f}{8L_f^2}\sum_{i=1}^{k} A_i\|u_i\|^2,$$

and hence that the statement in (a) holds.

b) First note that in view of a) the inclusion in (4) holds with for any $(y, u) := (y_k, u_k)$. Now, recall that

$$\sum_{i=1}^{k} i^2 = \frac{k(k+1)(2k+1)}{6} \geq \frac{k^3}{3}, \qquad \sum_{i=1}^{k} c^{2(i-1)} = \frac{c^{2k}-1}{c^2-1}$$

for any nonzero scalar $c \neq \pm 1$. Hence, considering c as in (24), it follows from the above relations and (20) that

$$\sum_{i=1}^{k} A_i \geq \sum_{i=1}^{k} \frac{1}{L_f - \mu_f}\max\left\{\frac{i^2}{4}, c^{2(i-1)}\right\} \geq \frac{1}{L_f - \mu_f}\max\left\{\frac{k^3}{12}, \frac{c^{2k}-1}{c^2-1}\right\}$$

which combined with (23) and (24) implies that

$$\min_{1\leq i\leq k}\|u_i\|^2 \leq \frac{8L_f^2(L_f - \mu_f)d_0^2}{L_f - \bar{L}_f}\min\left\{\frac{12}{k^3}, \frac{c^2-1}{c^{2k}-1}\right\} = \zeta d_0^2 \min\left\{\frac{12}{k^3}, \frac{c^2-1}{c^{2k}-1}\right\}.$$

Hence, in order to obtain $\min_{1\leq i\leq k}\|u_i\| \leq \rho$, it is sufficient to have

$$\zeta d_0^2 \min\left\{\frac{12}{k^3}, \frac{c^2-1}{c^{2k}-1}\right\} \leq \rho^2,$$

or equivalently, one of the following inequalities should hold

$$k \geq \left(\frac{12\zeta d_0^2}{\rho^2}\right)^{1/3}, \quad \frac{\zeta d_0^2\,(c^2-1)}{\rho^2} \leq c^{2k} - 1. \tag{25}$$

Note that the latter inequality is equivalent to

$$\frac{\log\left(1 + \frac{\zeta d_0^2(c^2-1)}{\rho^2}\right)}{2\log c} \leq k.$$

Hence, since $\log(1+x) \geq 1/(1+x^{-1})$ for any $x > 0$, it follows by using $x = \sqrt{\mu}/(2\sqrt{L_f - \mu_f})$ and the definition of c that the above condition holds if

$$\left(1 + \frac{2\sqrt{L_f - \mu_f}}{\sqrt{\mu}}\right) \log\left(1 + \zeta d_0^2 (c^2 - 1)\rho^{-2}\right) \leq k.$$

Hence, the last statement of the lemma follows from the above conclusion, the first inequality in (25), and the definition of ζ in (24). □

Before discussing some more exotic notions of approximate solutions, we first establish some properties regarding Γ_k and its relation to ϕ.

Lemma 2.10. *Define* $\Gamma_k : \mathbb{R}^n \to \mathbb{R}$ *as*

$$\Gamma_k(x) := \frac{1}{A_k} \sum_{i=0}^{k-1} a_i \gamma_i(x) \quad \forall x \in \mathbb{R}^n. \tag{26}$$

Then, for every $k \geq 1$*, the following statements hold:*

a) $\Gamma_k \leq \phi$ *and* Γ_k *is a* μ*-convex quadratic function with Hessian equal to* μI*;*

b) for every $x \in \mathbb{R}^n$*, we have*

$$\Gamma_k(x) \geq \phi(y_k) + \frac{1}{2A_k}\left(\tau_k \|x_k - x\|^2 - \|x_0 - x\|^2\right) \tag{27}$$

where $\gamma_k(\cdot)$ *is defined in* (14)*;*

Proof. a) In view of (14), the above definition of Γ_k, (b) and (c) of Lemma 2.1, and the second relation in (8), it follows that Γ_k is a convex combination of a μ-convex quadratic functions minorizing ϕ and whose Hessian are all equal to μI. Hence, a) follows.

b) This statement follows by summing the inequality in Lemma 2.4 from $k = 0$ to $k = k - 1$, using the definition of Γ_k, and the fact that $A_0 = 0$ and $\tau_0 = 1$ (see step 0 of S-FISTA). □

The next result shows some important relations on the pair (v_k, η_k) defined below in (28). This pair of elements can be incorporated in S-FISTA in order to apply it to inexactly solve some proximal subproblems.

Lemma 2.11. *Define*

$$v_k := \mu(y_k - x_k) + \frac{x_0 - x_k}{A_k}, \quad \eta_k := \frac{1}{2A_k}\left(\|x_0 - y_k\|^2 - \tau_k \|x_k - y_k\|^2\right). \tag{28}$$

Then, the following statements hold for every $k \geq 1$*:*

a) for every $x \in \mathbb{R}^n$*, we have*

$$\Gamma_k(x) - \frac{\mu}{2}\|x - y_k\|^2 \geq \phi(y_k) + \langle v_k, x - y_k \rangle - \eta_k, \tag{29}$$

where Γ_k *is as in* (26)*.*

b) we have

$$\eta_k \geq 0, \quad v_k \in \partial_{\eta_k}\left(\phi - \frac{\mu}{2}\|\cdot - y_k\|^2\right)(y_k),$$
$$\frac{1}{\tau_k}\|A_k v_k + y_k - x_0\|^2 + 2A_k\eta_k = \|y_k - x_0\|^2;$$

c) we have

$$\|v_k\| \leq \frac{1+\sqrt{\tau_k}}{A_k}\|y_k - x_0\|, \quad \eta_k \leq \frac{\|y_k - x_0\|^2}{2A_k}.$$

Proof. a) This statement follows from Lemma 2.10(b) and the fact that the definitions of v_k and η_k combined with the relation in Lemma 2.1(e) imply that

$$\frac{1}{2A_k}\left(\tau_k\|x_k - x\|^2 - \|x_0 - x\|^2\right) = \langle v_k, x - y_k\rangle - \eta_k + \frac{\tau_k - 1}{2A_k}\|x - y_k\|^2.$$

b) In view of Lemma 2.10(a), the inequality and the inclusion in (b) follow from the inequality in (a), first with $x = y_k$ and then with arbitrary $x \in \mathbb{R}^n$. The last relation in (b) follows from the definitions of v_k and η_k combined with Lemma 2.1(e).

c) These inequalities follow immediately from the last relation in (b) together with the triangle inequality for norms. □

The next result shows how the sequence $\{y_k\}$ together with the residuals pair sequence (v_k, η_k) defined in (28) can be used to generate an approximate solution based on a relative error criterion. An iteration complexity bound is also given for convenience.

Lemma 2.12. *Let $\{y_k\}$ be generated by S-FISTA and let $\{(v_k, \eta_k)\}$ defined as in (28). Then, for any $\tilde{\sigma} > 0$ and $k \geq 1$, the triple $(y, v, \eta) := (y_k, v_k, \eta_k)$ satisfies*

$$v \in \partial_\eta\left(\phi - \frac{\mu}{2}\|\cdot - y\|^2\right)(y), \quad \eta \geq 0, \quad \|v\|^2 + 2\eta \leq \tilde{\sigma}\|y - x_0\|^2, \tag{30}$$

as long as A_k satisfies

$$A_k \geq \bar{A} = \bar{A}(\mu, \tilde{\sigma}) := \frac{2\mu + 1 + \sqrt{(2\mu+1)^2 + 16\tilde{\sigma}}}{2\tilde{\sigma}}, \tag{31}$$

which in turn is satisfied in at most

$$\left\lceil \min\left\{2\sqrt{(L_f - \mu_f)\bar{A}},\ \left(\frac{1}{2} + \sqrt{\frac{L_f - \mu_f}{\mu}}\right)\log_1^+\left([L_f - \mu_f]\bar{A}\right) + 1\right\}\right\rceil$$

iterations of S-FISTA.

Proof. The first two relations in (30) follow immediately from Lemma 2.11. Now, it follows from Lemma 2.11(c) and Lemma 2.1(e) that

$$\|v_k\|^2 + 2\eta_k \leq \left[\frac{2}{A_k^2}(1+\tau_k) + \frac{1}{A_k}\right]\|y_k - x_0\|^2 = \frac{1}{A_k}\left(\frac{4}{A_k} + 2\mu + 1\right)\|y_k - x_0\|^2. \tag{32}$$

Since

$$\frac{1}{A_k}\left(\frac{4}{A_k}+2\mu+1\right) \le \tilde\sigma \Longleftrightarrow \tilde\sigma A_k^2-(2\mu+1)A_k-4 \ge 0,$$

we then conclude that the last inequality in (30) follows from (32) and the fact that the right hand side of (31) corresponds to the largest root of the above quadratic equation. The last statement of the lemma follows immediately from the last statement of Lemma 2.7. $\square$

The next result gives the complexity bound for a (slightly) different relative error criterion.

Lemma 2.13. *Let $\{(y_k, v_k, \eta_k)\}$ and $\bar A(\cdot,\cdot)$ be as in Lemma 2.12. Then for any $\sigma > 0$ and $k \ge 1$, the triple $(y, v, \eta) = (y_k, v_k, \eta_k)$ satisfies*

$$v \in \partial_\eta\left(\phi - \frac{\mu}{2}\|\cdot - y\|^2\right)(y), \quad \eta \ge 0, \quad \|v\|^2+2\eta \le \sigma\|v+y-y_0\|^2, \tag{33}$$

as long as $A_k \ge \bar A(\mu, \sigma/(1+\sqrt{\sigma})^2)$, which in turn is satisfied in at most

$$\left\lceil \min\left\{2\sqrt{(L_f-\mu_f)\mathcal{A}_{\mu,\sigma}}, \left(\frac{1}{2}+\sqrt{\frac{L_f-\mu_f}{\mu}}\right)\log_1^+\left([L_f-\mu_f]\mathcal{A}_{\mu,\sigma}\right)+1\right\}\right\rceil \tag{34}$$

iterations of S-FISTA, where $\mathcal{A}_{\mu,\sigma} := (2\mu+3)(1+\sqrt{\sigma})^2/\sigma$.

Proof. Let $\tilde\sigma = \sigma/(1+\sqrt{\sigma})^2 \in (0,1)$. Using Lemma 2.12 and the fact that $y_0 = x_0$, it follows that first two relations in (33) hold and $\|v\|^2+2\eta \le \tilde\sigma\|y-y_0\|^2$. Using the previous inequality, the definition of $\tilde\sigma$, and the relation $(a+b)^2 \le (1+\sqrt{\sigma})a^2+(1+1/\sqrt{\sigma})b^2$ for any $a, b \in \mathbb{R}$, we have

$$\|v\|^2+2\eta \le \frac{\sigma}{1+\sqrt{\sigma}}\|v+y-y_0\|^2+\frac{\sqrt{\sigma}}{1+\sqrt{\sigma}}\|v\|^2$$

which easily implies the last relation in (33). Finally, to obtain the bound in (34), we first use the definitions of $\tilde\sigma$ and $\bar A_{\mu,\sigma}$ with the fact $\tilde\sigma \in (0,1)$ to bound

$$\bar A(\mu,\tilde\sigma) \le \frac{2\mu+1}{\tilde\sigma}+\frac{2}{\sqrt{\tilde\sigma}} \le \frac{2\mu+3}{\tilde\sigma} = \frac{(2\mu+3)(1+\sqrt{\sigma})^2}{\sigma} = \mathcal{A}_{\mu,\sigma}.$$

The iteration complexity now follow from the last statement of Lemma 2.7 and the above bound. $\square$

The next result shows some bounds on the sequences $\{x_k\}$ and $\{y_k\}$ generated by S-FISTA.

Lemma 2.14. *For every $k \ge 1$, the following estimates hold:*

$$\|x_k - x_0\| \le \left(\frac{1}{\sqrt{\tau_k}}+1\right)d_0, \quad \|y_k-x_0\| \le 2\left(1+\frac{2}{A_k\mu}\right)d_0,$$

Proof. Let x^* be a solution of (1) such that $\|x_0 - x^*\| = d_0$. It follows from Lemma 2.6 with $x = y_k$ and $x = x^*$ that

$$\tau_k \|x_k - y_k\|^2 \le \|x_0 - y_k\|^2, \quad \tau_k \|x_k - x^*\|^2 \le \|x_0 - x^*\|^2. \tag{35}$$

Hence, using the triangle inequality for norms, we have

$$\|x_0 - x_k\| \le \|x_0 - x^*\| + \|x_k - x^*\| \le \left(1 + \frac{1}{\sqrt{\tau_k}}\right) \|x_0 - x^*\| = \left(1 + \frac{1}{\sqrt{\tau_k}}\right) d_0,$$

which proves the first inequality of the lemma. Moreover, using the triangle inequality for norms and the first inequality in (35), we have

$$\|y_k - x_0\| \le \|x_0 - x_k\| + \|x_k - y_k\| \le \|x_0 - x_k\| + \frac{1}{\sqrt{\tau_k}} \|x_0 - y_k\|.$$

Rewriting the above inequality and using the first inequality of the lemma, we have

$$\left(1 - \frac{1}{\sqrt{\tau_k}}\right) \|x_0 - y_k\| \le \|x_0 - x_k\| \le \left(1 + \frac{1}{\sqrt{\tau_k}}\right) d_0.$$

Thus,

$$\|x_0 - y_k\| \le \frac{\sqrt{\tau_k} + 1}{\sqrt{\tau_k} - 1} d_0 = \frac{(\sqrt{\tau_k} + 1)^2}{\tau_k - 1} d_0 \le \frac{2(\tau_k + 1)}{\tau_k - 1} d_0 = 2\left(1 + \frac{2}{\tau_k - 1}\right) d_0.$$

The second inequality of the lemma now follows from the fact that $\tau_k = 1 + A_k \mu$ in view of Lemma 2.1(e). $\square$

The below result establishes some alternative iteration complexity bounds for the residuals pair (v_k, η_k) defined in (28).

Lemma 2.15. *The following inequalities hold*

$$\|v_k\| \le \frac{2}{A_k} \left(2 + \sqrt{\mu A_k}\right) \left(1 + \frac{2}{A_k \mu}\right) d_0, \quad \eta_k \le \frac{2}{A_k} \left(1 + \frac{2}{A_k \mu}\right)^2 d_0^2. \tag{36}$$

As a consequence, for given a given tolerance pair $(\varepsilon, \eta) \in \mathbb{R}^2_{++}$, *we have*

$$\|v_k\| \le \varepsilon, \qquad \eta_k \le \eta \tag{37}$$

in at most

$$k := \left\lceil \min \left\{ 8 \left(\frac{1}{\sqrt{\varepsilon}} + \frac{\sqrt{\mu d_0}}{\varepsilon} + \frac{\sqrt{d_0}}{\sqrt{\eta}} \right) \sqrt{\mathcal{M} d_0}, \right.\right.$$
$$\left.\left. \left[\frac{1}{2} + \sqrt{\frac{L_f - \mu_f}{\mu}} \right] \log_1^+ \left(16 \left[\frac{1}{\varepsilon} + \frac{\mu d_0}{\varepsilon^2} + \frac{d_0}{\eta} \right] \mathcal{M} d_0 \right) + 1 \right\} \right\rceil$$

iterations, where

$$\mathcal{M} = \mathcal{M}(\mu_f, \mu, L_f) := \left(1 + \frac{8(L_f - \mu_f)}{\mu}\right)^2 (L_f - \mu_f)$$

Proof. The inequalities in (36) follows by combining Lemma 2.1(e), Lemma 2.11(c), and Lemma 2.14.

Now, in view of (20), we have $A_k \geq A_1 \geq 1/[4(L_f - \mu_f)]$ for every $k \geq 1$. Hence, it follows from (36) that

$$\|v_k\| \leq \frac{2}{A_k}\left(2+\sqrt{\mu A_k}\right)\left(1+\frac{8(L_f-\mu_f)}{\mu}\right)d_0, \quad \eta_k \leq \frac{2}{A_k}\left(1+\frac{8(L_f-\mu_f)}{\mu}\right)^2 d_0^2$$

which implies that in order to (v_k, ε_k) to satisfy (37), it is sufficient to have

$$\frac{4}{A_k}\left(1+\frac{8(L_f-\mu_f)}{\mu}\right)d_0 \leq \frac{\varepsilon}{2}, \qquad \frac{2\sqrt{\mu}}{\sqrt{A_k}}\left(1+\frac{8(L_f-\mu_f)}{\mu}\right)d_0 \leq \frac{\varepsilon}{2},$$
$$\frac{2}{A_k}\left(1+\frac{8(L_f-\mu_f)}{\mu}\right)^2 d_0^2 \leq \eta.$$

Note that the above inequalities are satisfied if

$$A_k \geq \frac{8}{\varepsilon}\left(1+\frac{8(L_f-\mu_f)}{\mu}\right)d_0 + \left(\frac{16\mu}{\varepsilon^2}+\frac{2}{\eta}\right)\left(1+\frac{8(L_f-\mu_f)}{\mu}\right)^2 d_0^2.$$

Hence, the last statement of the lemma follows from the above inequalities, the last statement of Lemma 2.7, and the definition of $\mathcal{M}$. □

3 Alternate Formulations of S-FISTA

This section presents alternate formulations of S-FISTA for the case of $\mu = 0$. Although, we assume that $\mu = 0$, it is worth mentioning that similar results as the ones obtained in this section can be extended for the general case where $\mu \geq 0$.

We begin by deriving an alternate expression for y_{k+1}.

Lemma 3.1. *Assume that $\mu = 0$. Then, for every $k \geq 0$, we have*

$$y_{k+1} = \frac{A_k y_k + a_k x_{k+1}}{A_{k+1}}.$$

Proof. It follows from (11) and (12) that

$$\frac{a_k}{A_{k+1}}(x_{k+1} - x_k) = \frac{\lambda}{a_k}(x_{k+1} - x_k) = y_{k+1} - \tilde{x}_k.$$

On the other hand, it follows from the last identity in (8) that

$$\frac{a_k}{A_{k+1}}(x_{k+1} - x_k) = \frac{A_k y_k + a_k x_{k+1}}{A_{k+1}} - \tilde{x}_k.$$

The result now follows by combining the above two identities. □

The next result shows that the auxiliary sequence $\{\tilde{x}_k\}$ generated by S-FISTA can be expressed in terms of the sequence $\{y_k\}$ and a scalar sequence that can be easily generated by solving a quadratic equation.

Lemma 3.2. *Assume that $\mu = 0$ and, for every $k \geq 0$, define*

$$t_k := \frac{A_{k+1}}{a_k} = \frac{a_k}{\lambda}. \tag{38}$$

Then, for every $k \geq 0$, we have

$$\tilde{x}_{k+1} = y_{k+1} + \frac{t_k - 1}{t_{k+1}}(y_{k+1} - y_k) \tag{39}$$

and

$$t_{k+1}^2 - t_{k+1} - t_k^2 = 0. \tag{40}$$

Proof. First, note the that the second equality of (38) follows from (8). It follows from (38) with $k = k+1$ and the two last identities in (8) both with $k = k+1$ that

$$\tilde{x}_{k+1} - y_{k+1} = \frac{A_{k+1}y_{k+1} + a_{k+1}x_{k+1}}{A_{k+2}} - y_{k+1} = \frac{a_{k+1}}{A_{k+2}}(x_{k+1} - y_{k+1}) = \frac{1}{t_{k+1}}(x_{k+1} - y_{k+1}).$$

On the other hand, it follows from (38), the second identity in (8), and Lemma 3.1, that

$$\begin{aligned}(t_k - 1)(y_{k+1} - y_k) &= \left(\frac{A_{k+1}}{a_k} - 1\right)(y_{k+1} - y_k) = \frac{A_k}{a_k}(y_{k+1} - y_k) \\ &= \frac{1}{a_k}\left[A_k y_{k+1} - (A_{k+1}y_{k+1} - a_k x_{k+1})\right] = x_{k+1} - y_{k+1}.\end{aligned}$$

The first identity of the lemma now follows by combining the above two identities. Now, it follows from (38) that

$$t_k^2 = \frac{A_{k+1}}{\lambda}$$

for every $k \geq 0$. The last identity, together with (38) and the second identity in (8) with $k = k+1$, then imply that

$$t_{k+1}^2 - t_{k+1} = \frac{A_{k+2}}{\lambda} - \frac{a_{k+1}}{\lambda} = \frac{A_{k+1}}{\lambda} = t_k^2,$$

and hence that the second identity of the lemma also holds. □

We now make a few remarks about the relations above and how they relate to the ones given in FISTA. First, (40) implies that the iterates $\{t_k\}$ have the recursive form

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}.$$

Second, in view of the first remark, (39), and the fact that $t_0 = 1$, we conclude that the iterates $\{(y_k, \tilde{x}_k, t_k)\}$ generated by S-FISTA are the same as the ones generated by FISTA (see, for example, the definitions in [1, 2]).

The next result presents an alternative way of expressing the relations in (39) and (40).

Lemma 3.3. *Assume $\mu = 0$, let $\{t_k\}$ be as in* (38)*, and define $\alpha_k = 1/t_k$ for every $k \geq 0$. Then, the following relation holds*

$$\alpha_{k+1}^2 = (1 - \alpha_{k+1})\alpha_k^2, \tag{41}$$

$$\tilde{x}_{k+1} = y_{k+1} + \frac{\alpha_k (1 - \alpha_k)}{\alpha_k^2 + \alpha_{k+1}}(y_{k+1} - y_k). \tag{42}$$

Proof. It follows from (40) and the definition of α_k that

$$\frac{1}{\alpha_{k+1}^2} - \frac{1}{\alpha_{k+1}} - \frac{1}{\alpha_k^2} = 0.$$

Multiplying both sides by $\alpha_k^2 \alpha_{k+1}^2$, we arrive at

$$\alpha_k^2 - \alpha_k^2 \alpha_{k+1} - \alpha_{k+1}^2 = 0$$

which immediately implies (41). Now, note that (39) together with the definition of α_k imply that

$$\begin{aligned} \tilde{x}_{k+1} &= y_{k+1} + \frac{t_k - 1}{t_{k+1}}(y_{k+1} - y_k) = y_{k+1} + \alpha_{k+1}\left(\frac{1}{\alpha_k} - 1\right)(y_{k+1} - y_k) \\ &= y_{k+1} + \frac{\alpha_{k+1}}{\alpha_k}(1 - \alpha_k)(y_{k+1} - y_k), \end{aligned}$$

which in view of (41) proves (42). □

Similar to the remarks after Lemma 3.2, the above result shows that when $\mu = 0$ and h is the characteristic function of a simple set, the iterates $\{(y_k, \tilde{x}_k, t_k)\}$ generated by S-FISTA are the same as the ones generated by Nesterov's FGM in [6, Eq (2.2.63)] with $\alpha_0 = 1$ (see also [5, Eq (2.2.17)]).

References

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

[2] 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.

[3] 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.

[4] Y. Nesterov. A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. In *Doklady Akademii Nauk SSSR*, volume 269, pages 543–547, 1983.

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

[6] Y. Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.