A FISTA-type accelerated gradient algorithm for solving smooth nonconvex composite optimization problems

Jiaming Liang ∗ Renato D.C. Monteiro ∗ Chee-Khian Sim †

May 20, 2019

Abstract

In this paper, we describe and establish iteration-complexity of two accelerated composite gradient (ACG) variants to solve a smooth nonconvex composite optimization problem whose objective function is the sum of a nonconvex differentiable function $f$ with a Lipschitz continuous gradient and a simple nonsmooth closed convex function $h$. When $f$ is convex, the first ACG variant reduces to the well-known FISTA for a specific choice of the input, and hence the first one can be viewed as a natural extension of the latter one to the nonconvex setting. The first variant requires as input a pair $(M,m)$, $M$ being a Lipschitz constant of $\nabla f$ and $m$ being a lower curvature of $f$ such that $m \leq M$ (possibly $m < M$), which is usually hard to obtain or poorly estimated. The second variant on the other hand can start from an arbitrary input pair $(M,m)$ of positive scalars and its complexity is shown to be not worse, and better in some cases, than that of the first variant for a large range of the input pairs. Finally, numerical results are provided to illustrate the efficiency of the two ACG variants.

1 Introduction

Accelerated gradient methods for solving convex noncomposite programs were originally developed by Nesterov in his celebrated work [18]. Subsequently, several variants of this method (see for example [1, 13, 17, 19, 20, 24]) were developed for solving convex simple-constrained or composite programs which we refer generically to as ACG variants. These variants have also been used as subroutines in several inexact-type proximal algorithms for solving convex-concave saddle point and monotone Nash equilibrium problems (see for example [3, 8, 9, 11, 20, 21]).

In this paper, we study ACG algorithms to solve the smooth nonconvex composite optimization (SNCO) problem

$$\phi_* := \min \{ \phi(z) := f(z) + h(z) : z \in \mathbb{R}^n \}$$

where $h : \mathbb{R}^n \to (-\infty, \infty]$ is a proper lower-semicontinuous convex function with bounded dom $h$ and $f$ is a real-valued differentiable (possibly nonconvex) function whose gradient is $M$-Lipschitz continuous on dom $h$, i.e., for every $z, z' \in \text{dom } h$,

$$\|\nabla f(z') - \nabla f(z)\| \leq M \|z' - z\|.$$
The first analysis of an ACG algorithm for solving (1) under the above assumption appears in [5] where essentially a well-known ACG variant which solves the convex version of (1) is also shown to solve its nonconvex version in the following sense: for a given tolerance $\hat{\rho} > 0$, it computes $(\hat{y}, \hat{v}) \in \text{dom} \ h \times \mathbb{R}^n$ such that $\hat{v} \in \nabla f(\hat{y}) + \partial h(\hat{y})$ and $\|\hat{v}\| \leq \hat{\rho}$ in

\[
O \left( \frac{M\bar{m}D_h^2}{\hat{\rho}^2} + \left( \frac{Md_0}{\hat{\rho}} \right)^{2/3} \right)
\]

iterations where $d_0$ is the distance of the initial point $x_0$ to the optimal solution set of (1), $D_h$ is the diameter of $\text{dom} \ h$ and $\bar{m}$ is the smallest scalar $m \geq 0$ satisfying

\[
-\frac{m}{2} \|z' - z\|^2 \leq f(z') - f(z) - \langle \nabla f(z), z' - z \rangle \quad \forall z, z' \in \text{dom} \ h.
\]

We refer to the ACG variant of [5] to as the AG method and note that each one of its iterations performs exactly two resolvent evaluations of $h$, i.e., an evaluation of the point-to-point operator $(I + \tau \partial h)^{-1}(\cdot)$ for some $\tau > 0$.

This paper describes and establishes the iteration-complexities of two ACG variants for solving the nonconvex version of (1). The first variant can be viewed as a direct extension of the FISTA in [1] for solving the convex version of (1). In contrast to an iteration of the AG method, every iteration of the first variant performs exactly one resolvent evaluation of $h$. One drawback of the first variant is that it requires as input a pair $(m, M)$, $M$ being a Lipschitz constant as in (2) and $m$ a nonnegative scalar satisfying (4), which is usually hard to obtain or poorly estimated. Letting $(\bar{m}, \bar{M})$ denote the smallest pair with the above properties, a second variant is proposed to remedy the aforementioned drawback in that it works regardless of the choice of input pair $(m, M)$ (i.e., not necessarily satisfying (2) and (4)), and its complexity is shown not to be worse than (3) when $M \geq \bar{M}$ and $m \in [\bar{m}, \bar{M}]$. Moreover, when $m \in [\bar{m}, \bar{M}]$, the complexity of the second variant is empirically argued to behave as (3) with $M = \bar{M}$ for a large range of scalars $M$ such that $M \leq \bar{M}$ (see the paragraph following Corollary 3.4) and our computational results demonstrate that taking $M$ relatively smaller than $\bar{M}$ can substantially improve its performance. It is also shown that all iterations of the second variant, with the exception of a few ones whose total number is log-bounded, perform exactly one resolvent evaluation of $h$.

Related works. Inspired by [5], other papers have proposed ACG variants for solving (1) under the assumption that $f$ is a nonconvex continuously differentiable function with Lipschitz continuous gradient, and $h$ is a simple lower semi-continuous convex (see e.g. [4, 6]) or nonconvex (see e.g. [14, 15, 25]) function. Similar to an iteration of the two ACG variants in our paper, the one of the algorithms in [15, 25] requires exactly one resolvent evaluation of $h$. However, while every iteration of the variants studied here are always accelerated, the ones of the latter algorithms can be a simple composite gradient (and unaccelerated) step whenever a certain descent property is not satisfied.

Another approach for solving (1) consists of using a descent unaccelerated inexact proximal-type method where each prox subproblem is constructed to be (possibly strongly) convex and hence solved by an ACG variant (see [2, 12, 22]). Moreover, the approach has the benefit of working with a larger prox stepsize and hence of having a better outer iteration-complexity than the approaches of the previous paragraph. However, each of its outer iteration still has to perform a uniformly bounded number of inner iterations to approximately solve a prox subproblem. Overall, it is shown that its inner-iteration complexity is better than the iteration-complexities of the methods in the previous paragraph, particularly when $\bar{m} \ll \bar{M}$. As in the papers [4, 6, 14, 15, 25] in the previous
paragraph, it is worth noting that the method in \cite{22} attempts to perform an accelerated step whenever a certain descent property holds and, in case of failure, it performs an unaccelerated prox step similar to the one used in the methods of\cite{2,12}.

Finally, a hybrid approach that borrows ideas from the above group of papers is presented in \cite{16}. More specifically, the latter work presents an accelerated inexact proximal point method reminiscent of those presented in \cite{7,17,23}, but in which only the convex version of \cite{11} are considered. Each (outer) iteration of the method requires that a prox subproblem be approximately solved by using an ACG variant in the same way as in the papers \cite{2,12}. Hence, similar to the methods of the previous paragraph, this method performs both outer and inner iterations with the major difference that every outer iteration is an accelerated step (as in the papers \cite{4,6,14,15,25}) with a large proximal stepsize (as in the papers \cite{2,12}).

Organization of the paper. Subsection 1.1 presents basic definitions and notations used throughout the paper. Section 2 presents assumptions made on the SNCO problem, describes the first ACG variant, which is an extension of FISTA to the SNCO problem and is referred to as NC-FISTA, and establishes its iteration-complexity for obtaining a stationary point of the SNCO problem. Section 3 presents an adaptive variant of NC-FISTA, namely, ADAP-NC-FISTA, and establishes its iteration-complexity. Section 4 presents computational results showing the efficiency of NC-FISTA and ADAP-NC-FISTA. Section 5 finishes the paper by presenting a few concluding remarks. Finally, supplementary technical results are provided in the appendix.

1.1 Basic definitions and notation

This subsection provides some basic definitions and notations used in this paper.

Let \( \Psi : \mathbb{R}^n \to (-\infty, +\infty] \) be given. The effective domain of \( \Psi \) is denoted by \( \text{dom} \, \Psi := \{ x \in \mathbb{R}^n : \psi(x) < \infty \} \) and \( \Psi \) is proper if \( \text{dom} \, \Psi \neq \emptyset \). Moreover, a proper function \( \Psi : \mathbb{R}^n \to (-\infty, +\infty] \) is \( \mu \)-strongly convex for some \( \mu \geq 0 \) if

\[
\Psi(\beta z + (1 - \beta)z') \leq \beta \Psi(z) + (1 - \beta)\Psi(z') - \frac{\beta(1 - \beta)\mu}{2} \| z - z' \|^2
\]

for every \( z, z' \in \text{dom} \, \Psi \) and \( \beta \in [0, 1] \). Let \( \partial \Psi(z) \) denote the subdifferential of \( \Psi \) at \( z \in \text{dom} \, \Psi \). If \( \Psi \) is differentiable at \( \tilde{z} \in \mathbb{R}^n \), then its affine approximation \( \ell_{\Psi}(\cdot; \tilde{z}) \) at \( z' \) is defined as

\[
\ell_{\Psi}(z; z') := \Psi(z') + \langle \nabla \Psi(z'), z - \tilde{z} \rangle \quad \forall z \in \mathbb{R}^n.
\]

The normal cone of \( X \) at \( z \) is denoted by \( N_X(z) \), i.e. \( N_X(z) = \{ u \in \mathbb{R}^n : \langle u, \tilde{z} - z \rangle \leq 0, \forall \tilde{z} \in X \} \). Let \( \text{Conv} (\mathbb{R}^n) \) denote the set of all proper lower semi-continuous convex functions \( \Psi : \mathbb{R}^n \to (-\infty, +\infty] \). Define \( \log^+(s) := \max\{\log s, 0\} \) and \( \log^+_1(s) := \max\{\log s, 1\} \) for \( s > 0 \).

2 The NC-FISTA for solving the SNCO Problem

This section describes the assumptions made on our problem of interest, namely, problem \( (1) \). It also presents and establishes the iteration-complexity of the first ACG variant, namely NC-FISTA, for obtaining an approximate stationary point of \( (1) \).

Throughout this paper, we consider problem \( (1) \) and make the following assumptions on it:

(A1) \( h \in \text{Conv} (\mathbb{R}^n); \)

(A2) \( \text{dom} \, h \) is bounded;

3
(A3) \( f \) is differentiable on a closed convex set \( \Omega \supseteq \text{dom } h \) and there exists a scalar \( M > 0 \) such that (2) holds for every \( z, z' \in \Omega \);

(A4) \( f \) is nonconvex on \( \text{dom } h \) and there exists \( m > 0 \) satisfying (4).

Throughout this paper, we denote the diameter of \( \text{dom } h \) as
\[
D_h := \sup_{u, u' \in \text{dom } h} \| u' - u \| < \infty
\]
where its finiteness is due to (A2). Moreover, let \( \bar{M} \) denote the smallest scalar \( M \) satisfying (2). Also, let \( \bar{m} \) denote the smallest scalar \( m \geq 0 \) satisfying (4). Clearly, \( 0 \leq \bar{m} \leq \bar{M} \).

We now make a few remarks about the above assumptions. First, it is implied by (A1)-(A3) that the set \( X^* \) of optimal solutions to (1) is nonempty and compact. Second, (A4) implies that \( \bar{m} \geq 0 \). Third, our interest is in the case where \( \bar{m} \ll \bar{M} \) since this case naturally arises in the context of penalty methods for solving linearly constrained composite nonconvex optimization problems (e.g., see Section 4 of [12]).

For \( z \in \text{dom } h \) to be a local minimum of (1), a necessary condition is that \( z \) is a stationary point of (1), i.e., \( 0 \in \nabla f(z) + \partial h(z) \). Motivated by this remark, the following notion of an approximate solution to problem (1) is proposed: a pair \( (\hat{y}, \hat{v}) \) is said to be a \( \hat{\rho} \)-approximate solution to (1), for a given tolerance \( \hat{\rho} > 0 \), if
\[
\hat{v} \in \nabla f(\hat{y}) + \partial h(\hat{y}), \quad \| \hat{v} \| \leq \hat{\rho}.
\]

We are now ready to state the NC-FISTA for solving (1).

**NC-FISTA**

0. Let a pair \((m, M)\) of \( f \) over \( \Omega \) such that \( 2M > m \), scalar \( \xi \in [m, 2M] \), stepsize \( \lambda \in (0, 1/M) \), tolerance \( \hat{\rho} > 0 \) and initial point \( y_0 \in \text{dom } h \) be given and choose initial parameter \( A_0 > 0 \) satisfying
\[
(\xi - m) \left( 1 + \sqrt{1 + 4A_0} \right) \geq 4\xi; \quad (7)
\]
set \( x_0 = y_0 \) and \( k = 0 \);

1. compute
\[
a_k = \frac{1 + \sqrt{1 + 4A_k}}{2}, \quad A_{k+1} = A_k + a_k; \quad (8)
\]

2. compute
\[
x_k = \frac{A_k}{A_{k+1}} y_k + \frac{a_k}{A_{k+1}} x_k \quad (9)
\]
\[
y_{k+1} = \text{argmin}_{u} \left\{ l_f(u; \bar{x}_k) + h(u) + \frac{1}{2} \left( \frac{1}{\lambda} + \frac{2\xi}{a_k} \right) \| u - \bar{x}_k \|^2 \right\}, \quad (10)
\]
\[
\hat{x}_{k+1} = \frac{(a_k + 2\xi \lambda) y_{k+1} - (a_k - 1) y_k}{2\xi \lambda + 1}, \quad x_{k+1} = P_\Omega(\hat{x}_{k+1}); \quad (11)
\]
3. compute

\[ v_{k+1} = \left( \frac{1}{\lambda} + \frac{2\xi}{a_k} \right) (\bar{x}_k - y_{k+1}) + \nabla f(y_{k+1}) - \nabla f(\bar{x}_k); \quad (12) \]

if \( \|v_{k+1}\| \leq \hat{\rho} \) then output \((\hat{y}, \hat{v}) = (y_{k+1}, v_{k+1})\) and stop; otherwise, set \( k \leftarrow k + 1 \) and go to step 1.

We now make a few remarks about the NC-FISTA. First, it follows from (10) that \( \{y_k\} \subset \text{dom} \ h, \) and hence \( \{y_k\} \) is bounded in view of (A2). Second, the definition of \( \{x_k\} \) in (11) implies that \( \{x_k\} \subset \Omega, \) and hence that \( \{\bar{x}_k\} \subset \Omega \) in view of (9). Hence, if \( \Omega \) is chosen to be compact, then the latter two sequences will also be bounded but our analysis does not make such assumption on \( \Omega. \) Third, if \( \Omega = \mathbb{R}^n, \) then each iteration of the NC-FISTA requires one resolvent evaluation of \( \partial h \) in (10), i.e., an evaluation of \( (I + \tau \partial h)^{-1} \) for some \( \tau > 0. \) Otherwise, it requires an extra projection onto \( \Omega \) in (11) which, depending on the problem instance and the set \( \Omega, \) might be considerably cheaper than a resolvent evaluation of \( \partial h. \) Fourth, an example of scalars \( \xi \) and \( A_0 \) satisfying (7) is \( \xi = 2m \) and \( A_0 = 12. \) Many other choices are possible although none of them with \( A_0 = 0. \) Thus, \( A_0 \) can be chosen to be \( \mathcal{O}(1), \) i.e., independent of the parameters \( m \) and \( M, \) and hence is not included in the complexity bounds derived in this paper. Fifth, when \( f \) is convex, i.e. \( \bar{m} = 0, \) NC-FISTA reduces to FISTA with the choice of \( m = 0 \) and \( \xi = 0. \) Finally, (8) imply that

\[ A_{k+1} = a_k^2. \quad (13) \]

We first establish a number of technical results. The first one establishes an important inequality satisfied by \( \xi. \)

**Lemma 2.1** Assume that \( m \geq \bar{m} \) and that the pair \((\xi, A_0) \in [m, \infty) \times \mathbb{R}_+ \) satisfies (7). Then, we have

\[ \bar{m} + \frac{2\xi}{a_k} \leq \xi \quad \forall k \geq 0. \quad (14) \]

**Proof:** Noting that (7), relation (8) with \( k = 0, \) and the assumptions that \( \xi \geq m \) and \( m \geq \bar{m} \) imply that

\[ \xi - \bar{m} \geq \xi - m \geq \frac{2\xi}{a_0}, \]

and using the fact that \( \{a_k\} \) is increasing, we conclude that for every \( k \geq 0, \)

\[ \xi - \bar{m} \geq \frac{2\xi}{a_0} \geq \frac{2\xi}{a_k}. \]

The following results introduce two functions that play important roles in our analysis of NC-FISTA and establish some basic facts about them.

**Lemma 2.2** For every \( k \geq 0, \) if we define

\[ \tilde{\gamma}_k(u) := l_f(u; \bar{x}_k) + h(u) + \frac{\xi}{a_k} \|u - \bar{x}_k\|^2, \quad (15) \]

\[ \gamma_k(u) := \tilde{\gamma}_k(y_{k+1}) + \frac{1}{\lambda} (\bar{x}_k - y_{k+1}, u - y_{k+1}) + \frac{\xi}{a_k} \|u - y_{k+1}\|^2, \quad (16) \]

then the following statements hold:
The definition of $\gamma_k$ for every $u$ implies that
\begin{equation}
\min_u \left\{ \tilde{\gamma}_k(u) + \frac{1}{2\lambda} \|u - \tilde{x}_k\|^2 \right\} = \min_u \left\{ \gamma_k(u) + \frac{1}{2\lambda} \|u - \tilde{x}_k\|^2 \right\},
\end{equation}
and these minimization problems have $y_{k+1}$ as unique optimal solution;

(b) for every $u \in \text{dom } h$,
\[\tilde{\gamma}_k(u) - \phi(u) \leq \frac{1}{2} \left( \tilde{m} + \frac{2\xi}{a_k} \right) \|u - \tilde{x}_k\|^2;\]

(c) $x_{k+1} = \arg\min_{u \in \Omega} \left\{ a_k \tilde{\gamma}_k(u) + \|u - x_k\|^2/(2\lambda) \right\}$.

**Proof:** (a) By definition of $\tilde{\gamma}_k$ and $\gamma_k$ in (15) and (16) respectively, they are clearly $(2\xi/a_k)$-strongly convex. By (10) and the definition of $\tilde{\gamma}_k$ in (15), $y_{k+1}$ is the optimal solution to the first minimization problem in (17). Since the objective function of this minimization problem is $[(1/\lambda) + (2\xi/a_k)]$-strongly convex, it follows that for all $u \in \mathbb{R}^n$,
\[\tilde{\gamma}_k(y_{k+1}) + \frac{1}{2\lambda} \|y_{k+1} - \tilde{x}_k\|^2 + \frac{1}{2} \left( \frac{1}{\lambda} + \frac{2\xi}{a_k} \right) \|y_{k+1} - u\|^2 \leq \tilde{\gamma}_k(u) + \frac{1}{2\lambda} \|u - \tilde{x}_k\|^2 \quad (18)\]
On the other hand, the definition of $\gamma_k$ in (16) and the relation
\[\|y_{k+1} - \tilde{x}_k\|^2 + \|y_{k+1} - u\|^2 - \|u - \tilde{x}_k\|^2 = 2\langle \tilde{x}_k - y_{k+1}, u - y_{k+1} \rangle.\]

imply that
\[\tilde{\gamma}_k(y_{k+1}) + \frac{1}{2\lambda} \|y_{k+1} - \tilde{x}_k\|^2 + \frac{1}{2} \left( \frac{1}{\lambda} + \frac{2\xi}{a_k} \right) \|y_{k+1} - u\|^2 = \gamma_k(u) + \frac{1}{2\lambda} \|u - \tilde{x}_k\|^2 \quad (19)\]
Thus, it follows from (18) and (19) that $\gamma_k \leq \tilde{\gamma}_k$ and $\tilde{\gamma}_k(y_{k+1}) = \gamma_k(y_{k+1})$. The latter conclusion then implies that
\[\gamma_k(y_{k+1}) + \frac{1}{2\lambda} \|y_{k+1} - \tilde{x}_k\|^2 = \tilde{\gamma}_k(y_{k+1}) + \frac{1}{2\lambda} \|y_{k+1} - \tilde{x}_k\|^2 \]
\[\leq \gamma_k(u) + \frac{1}{2\lambda} \|u - \tilde{x}_k\|^2 \leq \tilde{\gamma}_k(u) + \frac{1}{2\lambda} \|u - \tilde{x}_k\|^2 \]
for every $u \in \mathbb{R}^n$, and hence that the remaining conclusions of (a) follows.

(b) This statement follows from assumption (A4) and the definition of $\tilde{\gamma}_k(u)$ in (15).

(c) Using the expressions for $\tilde{x}_k$ and $\tilde{x}_{k+1}$ in (9) and (11), respectively, it is easy to see that $\tilde{x}_{k+1}$ is the (unique) global minimizer of the function $a_k \gamma_k(u) + \|u - x_k\|^2/(2\lambda)$ over the whole space $\mathbb{R}^n$. The definition of $x_{k+1}$ and the previous observation then imply that the conclusion of (c) holds.

The following result states a recursive inequality which plays an important role in the convergence rate analysis of NC-FISTA.

**Lemma 2.3** For every $u \in \Omega$ and $k \geq 0$, we have
\[\lambda A_{k+1} \phi(y_{k+1}) + \left( \xi \lambda + \frac{1}{2} \right) \|u - x_{k+1}\|^2 + \frac{(1 - \lambda M_k) A_{k+1}}{2} \|y_{k+1} - \tilde{x}_k\|^2 \leq \lambda A_k \gamma_k(y_k) + \lambda a_k \gamma_k(u) + \frac{1}{2} \|u - x_k\|^2,\]
where
\[M_k := \frac{2 \left[ f(y_{k+1}) - \ell_f(y_{k+1}; \tilde{x}_k) \right]}{\|y_{k+1} - \tilde{x}_k\|^2}.\]
Proof: Using the definition of $\tilde{M}_k$ and (15), we conclude that
\[
\lambda \phi(y_{k+1}) + \frac{(1-\lambda \tilde{M}_k)}{2} \|y_{k+1} - \tilde{x}_k\|^2 = \lambda \gamma_k(y_{k+1}) + \left(\frac{1}{2} - \frac{\xi \lambda}{a_k}\right) \|y_{k+1} - \tilde{x}_k\|^2 \\
\leq \lambda \gamma_k(y_{k+1}) + \frac{1}{2} \|y_{k+1} - \tilde{x}_k\|^2.
\] (20)
Using the fact that, by Lemma 2.2(a), $\gamma_k$ is $(2\xi/a_k)$-strongly convex and $y_{k+1}$ is an optimal solution of (17), and relations (9) and (13), we conclude that for every $r \in \Omega$,
\[
A_{k+1} \left(\lambda \gamma_k(y_{k+1}) + \frac{1}{2} \|y_{k+1} - \tilde{x}_k\|^2\right) \\
\leq A_{k+1} \left(\lambda \gamma_k \left(\frac{A_k y_k + a_k x_{k+1}}{A_{k+1}}\right) + \frac{1}{2} \left|\frac{A_k y_k + a_k x_{k+1}}{A_{k+1}} - \tilde{x}_k\right|^2\right) \\
\leq \lambda A_k \gamma_k(y_k) + \frac{A_{k+1}}{2} \left|\frac{A_k y_k + a_k x_{k+1}}{A_{k+1}} - \tilde{x}_k\right|^2 \\
= \lambda A_k \gamma_k(y_k) + \lambda a_k \gamma_k(x_{k+1}) + \frac{1}{2} \|x_{k+1} - x_k\|^2 \\
\leq \lambda A_k \gamma_k(y_k) + \lambda a_k \gamma_k(u) + \frac{1}{2} \|u - x_k\|^2 - \frac{2\xi \lambda + 1}{2} \|u - x_{k+1}\|^2,
\] (21)
where the last inequality follows from Lemma 2.2(c) and the fact that $\lambda a_k \gamma_k(u) + \|u - x_k\|^2/2$ is $(2\xi + 1)$-strongly convex. The result now follows by combining (20) and (21).

Lemma 2.4 For every $k \geq 1$ and $u \in \text{dom } h$, we have
\[
\sum_{i=0}^{k-1} \left(\frac{1-\lambda \hat{M}_i}{2} A_{i+1} \|y_{i+1} - \tilde{x}_i\|^2\right) \\
\leq \left(\lambda A_0 \phi(y_0) - \phi(u)\right) + \frac{1}{2} \|u - x_0\|^2 - \left(\lambda A_k \phi(y_k) - \phi(u)\right) + \frac{1}{2} \|u - x_k\|^2 \\
+ \xi \lambda \|u - x_0\|^2 - \xi \lambda \|u - x_k\|^2 + 3\xi \lambda D_h^2 + \tilde{m} \lambda D_h^2 \sum_{i=0}^{k-1} a_i.
\] (22)
Proof: Let $i \geq 0$ and $u \in \text{dom } h$ be given. Using the definition of $\tilde{x}_i$ in (9), relations (5) and (13), the fact that $A_{i+1} = A_i + a_i \geq A_i$ due to (8), the inequality \(|a + b|^2 \leq 2\|a\|^2 + \|b\|^2\) for any $a, b \in \mathbb{R}^n$, we obtain
\[
A_i \|y_i - \tilde{x}_i\|^2 + a_i \|u - \tilde{x}_i\|^2 \\
= A_i a_i^2 \|x_i - y_i\|^2 + a_i \left|\frac{A_i}{A_{i+1}} (u - y_i) + \frac{a_i}{A_{i+1}} (u - x_i)\right|^2 \\
= A_i \left|\frac{(x_i - u) - (y_i - u)}{A_{i+1}}\right|^2 + a_i \left|\frac{A_i}{A_{i+1}} (u - y_i) + \frac{a_i}{A_{i+1}} (u - x_i)\right|^2 \\
\leq \frac{2A_i}{A_{i+1}} \left[\|u - x_i\|^2 + \|u - y_i\|^2\right] + 2a_i \left[\frac{A_i^2}{A_{i+1}^2} \|u - y_i\|^2 + \frac{a_i^2}{A_{i+1}^2} \|u - x_i\|^2\right] \\
\leq \frac{2A_i}{A_{i+1}} \left[\|u - x_i\|^2 + \|u - y_i\|^2\right] + 2a_i \|u - y_i\|^2 + \frac{2a_i}{A_{i+1}} \|u - x_i\|^2,
\[ \leq 2\|u - x_i\|^2 + 2(1 + a_i)D_h^2. \]  

(23)

Now, using Lemma 2.3 relation (8), some simple algebraic manipulations, statements (a) and (b) of Lemma 2.2 and the above inequality, we conclude that for every \( i \geq 0 \),

\[
\begin{align*}
1 - \lambda M_i & \frac{1}{2} A_{i+1} \|y_{i+1} - \tilde{x}_i\|^2 + \xi \lambda \|u - x_{i+1}\|^2 \\
- \left( \lambda A_i (\phi(y_i) - \phi(u)) + \frac{1}{2} \|u - x_i\|^2 \right) + \left( \lambda A_{i+1} (\phi(y_{i+1}) - \phi(u)) + \frac{1}{2} \|u - x_{i+1}\|^2 \right) \\
\leq \lambda A_i (\gamma_i(y_i) - \phi(y_i)) + \lambda a_i (\gamma_i(u) - \phi(u)) \\
\leq \frac{\lambda}{2} \left( \bar{m} + \frac{2\xi}{a_i} \right) (A_i \|y_i - \tilde{x}_i\|^2 + a_i \|u - \tilde{x}_i\|^2) \\
\leq \lambda \left( \bar{m} + \frac{2\xi}{a_i} \right) (\|u - x_i\|^2 + (1 + a_i) D_h^2).
\end{align*}
\]

It follows from the above inequality and Lemma 2.1 that

\[
\begin{align*}
\frac{1 - \lambda M_i}{2} A_{i+1} \|y_{i+1} - \tilde{x}_i\|^2 \\
- \left( \lambda A_i (\phi(y_i) - \phi(u)) + \frac{1}{2} \|u - x_i\|^2 \right) + \left( \lambda A_{i+1} (\phi(y_{i+1}) - \phi(u)) + \frac{1}{2} \|u - x_{i+1}\|^2 \right) \\
\leq \lambda \left( \bar{m} + \frac{2\xi}{a_i} \right) (\|u - x_i\|^2 + (1 + a_i) D_h^2) - \xi \lambda \|u - x_{i+1}\|^2 \\
\leq \xi \lambda (\|u - x_i\|^2 - \|u - x_{i+1}\|^2) + \left( \bar{m} + \frac{2\xi}{a_i} \right) (1 + a_i) \lambda D_h^2 \\
\leq \xi \lambda (\|u - x_i\|^2 - \|u - x_{i+1}\|^2) + (3\xi + \bar{m}a_i) \lambda D_h^2.
\end{align*}
\]

Inequality (22) now follows by summing the above inequality from \( i = 0 \) to \( i = k - 1 \).

The following result develops a convergence rate bound for the quantity \( \min_{1 \leq i \leq k} \|v_i\|^2 \) which, due to the stopping criterion in step 3 of NC-FISTA, is crucial for proving Theorem 2.6.

**Lemma 2.5** For every \( k \geq 1 \),

\[
\min_{1 \leq i \leq k} \|v_i\|^2 \leq \left( \frac{3}{\lambda} + M \right)^2 \left( \frac{4}{1 - \lambda M} \right) \left( \frac{2\lambda \bar{m} D_h^2}{k} + \frac{18\xi \lambda D_h^2}{k^2} + \frac{3 \left[ 2\lambda A_0 (\phi(y_0) - \phi_0) + 5d_0^2 \right]}{k^3} \right) 
\]  

(25)

where

\( d_0 := \inf_{x^* \in X^*} \|x^* - y_0\| = \inf_{x^* \in X^*} \|x^* - x_0\|. \)

(26)

**Proof:** First note that (17) and relation (8) with \( k = 0 \) imply that \( a_0 > 2 \). Hence, it follows from the assumptions that \( \lambda < 1/M \) and \( \xi \leq 2M \) that \( \xi < 2/\lambda \). This observation, the assumptions that \( \nabla f \) is \( M \)-Lipschitz continuous (see (A3)) and \( M < 1/\lambda \) (see step 0 of NC-FISTA), relation (12) and the fact that \( \{a_k\} \) is increasing then imply that

\[
\min_{1 \leq i \leq k} \|v_i\|^2 \leq \left( \frac{1}{\lambda} + \frac{2\xi}{a_0} + M \right)^2 \min_{0 \leq i \leq k-1} \|y_{i+1} - \tilde{x}_i\|^2 \leq \left( \frac{3}{\lambda} + M \right)^2 \min_{0 \leq i \leq k-1} \|y_{i+1} - \tilde{x}_i\|^2. 
\]  

(27)
Moreover, due to the first remark after assumptions (A1)-(A3), there exists $x^* \in X^*$ such that $\|x^* - x_0\| = d_0$. Noting that $x^* \in \text{dom } h$, and using Lemma $2.4$ with $u = x^*$, the fact that $M_k \leq \tilde{M}$ for $k \geq 0$ and the observation that $\xi \lambda < 2$, we conclude that

$$1 - \lambda \tilde{M} \left( \sum_{i=0}^{k-1} A_{i+1} \right) \min_{0 \leq i \leq k-1} \|y_{i+1} - \tilde{x}_i\|^2 \leq \sum_{i=0}^{k-1} \left( \frac{1 - \lambda \tilde{M}}{2} A_{i+1} \|y_{i+1} - \tilde{x}_i\|^2 \right)$$

$$\leq \lambda A_0 (\phi(y_0) - \phi_*) + \left( \frac{1}{2} + \xi \lambda \right) d_0^2 + 3 \xi \lambda D_h^2 k + \bar{m} \lambda D_h^2 \sum_{i=0}^{k-1} a_i$$

$$\leq \lambda A_0 (\phi(y_0) - \phi_*) + \frac{5}{2} d_0^2 + 3 \xi \lambda D_h^2 k + \bar{m} \lambda D_h^2 \sum_{i=0}^{k-1} a_i.$$
3 An Adaptive Variant of the NC-FISTA

This section describes the second ACG variant studied in this paper, namely ADAP-NC-FISTA which, in contrast to NC-FISTA, does not require knowledge of a pair \((m, M)\) as input. Instead of choosing constant stepsize \(\lambda\) and parameter \(\xi\) dependent on the pair \((m, M)\), it chooses them adaptively (see (31), (32) and (33) below).

We begin by describing ADAP-NC-FISTA which, in contrast to NC-FISTA, generates sequences \(\{\lambda_k\}\) and \(\{\xi_k\}\) in place of constant stepsize \(\lambda\) and parameter \(\xi\). Note that it requires as input an initial arbitrary pair \((\lambda_0, \xi_0)\) of positive scalars.

ADAP-NC-FISTA

0. Let \(\xi_0 > 0, \lambda_0 > 0\), tolerance \(\hat{\rho} > 0\), and initial point \(y_0 \in \text{dom} h\) and set \(x_0 = y_0, A_0 = 12\) and \(k = 0\);

1. compute \(a_k\) and \(A_{k+1}\) as in (5), \(\tilde{x}_k\) as in (9), and

\[
\bar{m}_{k+1} = \max \left\{ \frac{2[\ell_f(y_k; \tilde{x}_k) - f(y_k)]}{\|y_k - \tilde{x}_k\|^2}, \frac{2[\ell_f(y_0; \tilde{x}_k) - f(y_0)]}{\|y_0 - \tilde{x}_k\|^2}, 0 \right\}; \tag{28}
\]

2. let

\[
y_k(\lambda, \xi) := \arg\min_u \left\{ \ell_f(u; \tilde{x}_k) + h(u) + \frac{1}{2} \left( \frac{1}{\lambda} + \frac{2\xi}{a_k} \right) \|u - \tilde{x}_k\|^2 \right\}, \tag{29}
\]

\[
M_k(\lambda, \xi) := \frac{2[f(y_k(\lambda, \xi)) - \ell_f(y_k(\lambda, \xi); \tilde{x}_k)]}{\|y_k(\lambda, \xi) - \tilde{x}_k\|^2}; \tag{30}
\]

call subroutine SUB stated below to compute \((\lambda_{k+1}, \xi_{k+1}) = (\lambda, \xi)\) satisfying

\[
\xi \geq \xi_k, \quad \lambda \leq \lambda_k, \tag{31}
\]

\[
\lambda M_k(\lambda, \xi) \leq 0.9, \tag{32}
\]

\[
\xi \left( \lambda - \frac{2\lambda}{a_k} \right) \geq \bar{m}_{k+1} \lambda, \tag{33}
\]

and go to step 3;

3. compute

\[
y_{k+1} = y_k(\lambda_{k+1}, \xi_{k+1}), \quad M_{k+1} = M_k(\lambda_{k+1}, \xi_{k+1}), \tag{34}
\]

\[
x_{k+1} = P_\Omega \left( \frac{(a_k + 2\xi_{k+1}\lambda_{k+1})y_{k+1} - (a_k - 1)y_k}{2\xi_{k+1}\lambda_{k+1} + 1} \right), \tag{35}
\]

\[
v_{k+1} = \left( \frac{1}{\lambda_{k+1}} + \frac{2\xi_{k+1}}{a_k} \right) (\tilde{x}_k - y_{k+1}) + \nabla f(y_{k+1}) - \nabla f(\tilde{x}_k);
\]

if \(\|v_{k+1}\| \leq \hat{\rho}\) then output \((\hat{y}, \hat{v}) = (y_{k+1}, v_{k+1})\) and stop; otherwise, set \(k \leftarrow k + 1\) and go to step 1.
We will now describe the subroutine SUB used in step 2 of ADAP-NC-FISTA to compute \((\lambda, \xi)\) satisfying conditions (31)-(33).

**SUB**

0. Let \(\theta > 1\) and \((\lambda_k, \xi_k) \in \mathbb{R}^2_+\) be given and set \((\lambda, \xi) = (\lambda_k, \xi_k)\);

1. compute \(y_k(\lambda, \xi)\) and \(M_k(\lambda, \xi)\) according to (29) and (30), respectively, if \((\lambda, \xi)\) satisfy (32) and (33), then output \((\lambda, \xi)\) and stop; else, if (32) is not satisfied then set
   \[\lambda^+ \leftarrow \min\{\lambda/\theta, 0.9/M_k(\lambda, \xi)\};\]  
   if (33) is not satisfied then set \(\xi^+ \leftarrow 2\xi;\)

2. set \((\lambda, \xi) = (\lambda^+, \xi^+)\) and go to step 1.

A few remarks are made about ADAP-NC-FISTA and the subroutine SUB. First, the quantities \(y_k(\lambda_k+1, \xi_k+1)\) and \(M_k(\lambda_k+1, \xi_k+1)\) in (34) are actually computed inside the subroutine SUB. Second, ADAP-NC-FISTA consists of two types of iterations, namely, the ones indexed by \(k\) which we refer to as the outer iterations and the ones performed inside SUB which we refer to as the inner iterations. Third, each inner iteration performs exactly one resolvent evaluation of \(h\) while computing \(y_k(\lambda, \xi)\).

The following lemma states some properties of ADAP-NC-FISTA.

**Lemma 3.1** The following statements hold for ADAP-NC-FISTA:

(a) for every \(k \geq 1\), we have \(\bar{M}_k \leq \bar{M}\) and \(\bar{m}_k \leq \bar{m}\);

(b) for every \(k \geq 1\),
   \[\bar{M}_k \lambda_k \leq 0.9, \quad \xi_k \lambda_{k-1} \geq \bar{m}_k \lambda_k + \frac{2\xi_k \lambda_k}{a_{k-1}};\]

(c) \(\{\lambda_k\}\) is non-increasing and \(\{\xi_k\}\) is non-decreasing.

(d) for every \(k \geq 0\),
   \[\lambda_k \geq \underline{\lambda} := \min\left\{\frac{0.9}{\theta \bar{M}}, \lambda_0\right\}, \quad \xi_k \leq \bar{\xi} := \max\{4\bar{m}, \xi_0\};\]  
   \(\lambda_k \leq \bar{M}_k(\lambda, \xi) = \min\left\{\lambda/\theta, 0.9/M_k(\lambda, \xi)\right\}\) for some \(\lambda\) and \(\xi\), and we have \(M_{k-1}(\lambda, \xi) > 0\). Since \(\bar{M} \geq M_{k-1}(\lambda, \xi) > 0\) and \(\lambda_k < 0.9/(\theta \bar{M})\), it follows
that $0.9/M_{k-1}(\lambda, \xi) \geq 0.9/M > \lambda_k$. Hence, it follows from (36) that $\lambda_k = \lambda/\theta$. On the other hand, noting that (32) implies that $\lambda$ is no longer reduced whenever $\lambda \leq 0.9/M$, we then conclude that $\lambda > 0.9/M$, and hence that $\lambda_k = \lambda/\theta > 0.9/(\theta M)$. Since the latter conclusion contradicts our initial assumption, the first result in statement (d) follows. To show the second result in statement (d), for contradiction, assume that $\xi_k > \xi$ for some $k \geq 0$. Since $\xi_k > \xi_0$, we have $k \geq 1$, and $\xi_k = 2\xi$ in view of (37), where $\xi$ satisfies $\xi\lambda_{k-1} < \bar{m}_k\lambda + 2\xi\lambda_{k-1} - \xi$ according to (33). By $\bar{m}_k < \bar{m}$ from Lemma 3.1(a), $a_{k-1} \geq a_0 = 4$ and $\lambda \leq \lambda_{k-1}$, we have $\bar{m}_k\lambda + 2\xi\lambda_{k-1} - \bar{m}\lambda_{k-1} + (\lambda_{k-1}\xi)/2$. Hence, $\xi\lambda_{k-1} < \lambda_{k-1}(\bar{m} + \xi/2)$, which implies that $\xi < 2\bar{m}$. Therefore, $\xi_k = 2\xi < 4\bar{m}$, which contradicts our initial assumption. The second result in statement (d) then follows.

We have the following lemma which allows us to provide the oracle complexity result for ADAP-NC-FISTA in Theorem 3.3.

**Lemma 3.2** For every $k \geq 1$, we have

$$\frac{1}{20} \left( \sum_{i=0}^{k-1} \frac{A_{i+1}}{\xi_{i+1}} \right) \min_{0 \leq i \leq k-1} \| y_{i+1} - \tilde{x}_i \|^2 \leq \lambda_0 D_h^2 \left( 3k + \bar{m} \sum_{i=0}^{k-1} \frac{a_i}{\xi_{i+1}} \right) + 2\lambda_0 \frac{\bar{A}}{\xi_0} (\phi(y_0) - \phi_*).$$

(39)

**Proof:** Using similar arguments as in the proof of Lemma 2.3, we conclude that for every $k \geq 0$ and $u \in \Omega$,

$$\lambda_{k+1} A_{k+1} \phi(y_{k+1}) + \left( \xi_{k+1} \lambda_{k+1} + \frac{1}{2} \right) \| u - x_{k+1} \|^2 + \frac{(1 - \lambda_{k+1} M_{k+1}) A_{k+1}}{2} \| y_{k+1} - \tilde{x}_k \|^2$$

$$\leq \lambda_{k+1} A_{k} \gamma_k(y_k) + \lambda_{k+1} a_k \gamma_k(u) + \frac{1}{2} \| u - x_k \|^2,$$

where

$$\gamma_k(u) := \tilde{\gamma}_k(y_k + 1) + \frac{\xi_{k+1}}{\lambda_{k+1}} \| x_k - y_{k+1} - u - y_{k+1} \|^2$$

and

$$\tilde{\gamma}_k(u) := \ell_f(u; \tilde{x}_k) + \| u - \tilde{x}_k \|^2.$$  

(40)

As in Lemma 2.2(a), we have $\gamma_k(u) \leq \tilde{\gamma}_k(u)$ for every $u \in \text{dom}\ h$. Hence, it follows from (41) and (28) that for every $k \geq 0$ and $u \in \{y_k, y_0\}$, we have

$$\gamma_k(u) - \phi(u) \leq \tilde{\gamma}_k(u) - \phi(u) = \ell_f(u; \tilde{x}_k) - f(u) + \frac{\xi_{k+1}}{a_k} \| u - \tilde{x}_k \|^2$$

$$\leq \frac{1}{2} \left( \frac{\bar{m}_{k+1} + 2\xi_{k+1}}{a_k} \right) \| u - \tilde{x}_k \|^2.$$

(42)

Using (40) and (23) both with $u = x_0$, and using (42), (a) and (b) of Lemma 3.1, and the facts that $x_0 = y_0$ and $\lambda_{i+1} \leq \lambda_i$, we conclude that for every $0 \leq i \leq k - 1$,

$$\frac{0.1}{2} A_{i+1} \| y_{i+1} - \tilde{x}_i \|^2 + \xi_{i+1} \lambda_{i+1} \| x_0 - x_{i+1} \|^2 + \left[ \lambda_{i+1} A_{i+1} (\phi(y_{i+1}) - \phi(y_0)) + \frac{1}{2} \| x_0 - x_{i+1} \|^2 \right]$$

$$- \left[ \lambda_{i+1} A_i (\phi(y_i) - \phi(y_0)) + \frac{1}{2} \| x_0 - x_{i} \|^2 \right]$$

$$\leq \lambda_{i+1} A_i (\gamma_i(y_i) - \phi(y_i)) + \lambda_{i+1} a_i (\gamma_i(y_0) - \phi(y_0))$$

$$\leq \frac{\lambda_{i+1}}{2} \left( \bar{m}_{i+1} + 2\xi_{i+1} \right) \left( A_i \| y_i - \tilde{x}_i \|^2 + a_i \| x_0 - \tilde{x}_i \|^2 \right)$$

12
Theorem 3.3
The following statements hold:

(a) every iterate \((y_k, v_k)\) generated by ADAP-NC-FISTA satisfies

\[ v_k \in \nabla f(y_k) + \partial h(y_k); \]

moreover, ADAP-NC-FISTA outputs a \(\hat{\rho}\)-approximate solution \((\hat{y}, \hat{v})\) in a finite number of outer iterations \(T\) which is bounded by

\[
T = O\left( \left( \frac{C_1|\phi(y_0) - \phi_*|}{\hat{\rho}^2} \right)^{1/3} + \left( \frac{C_1\xi_0D_h^2}{\hat{\rho}^2} \right)^{1/2} + \frac{C_1[mD_h^2 + \phi(y_0) - \phi_*]}{\hat{\rho}^2} + 1 \right) \tag{43}
\]
where $D_h$ is defined in (5), $\bar{m}$ and $\bar{M}$ are defined in the paragraph following assumptions (A1)-(A4), and

$$C_1 := \left( \frac{1}{\lambda_0} + \xi_0 + \bar{M} \right)^2 \max\{\bar{m}/\xi_0, 1\} \lambda_0;$$

(b) if $\xi_0 \geq 2 \bar{m}$, then an alternative bound on $T$ is

$$T = \mathcal{O} \left( \left( \frac{C_1 \phi(y_0) - \phi_\ast + (d_0^2/\lambda_0)}{\rho^2} \right)^{1/3} + \left( \frac{C_1 \xi_0 D_h^2}{\rho^2} \right)^{1/2} + \frac{C_1 \bar{m} D_h^2}{\rho^2} + 1 \right); \quad (44)$$

(c) the total number of inner iterations, and hence resolvent evaluations of $h$, performed by ADAP-NC-FISTA is bounded by

$$T + \mathcal{O} \left( \log_{1/2}^{+} \left( \max \left\{ \lambda_0 \bar{M}, \frac{\bar{m}}{\xi_0} \right\} \right) \right) \quad (45)$$

where $\log_{1/2}^{+} (\cdot)$ is defined in Subsection 1.1.

Proof: (a) The first conclusion follows from the same argument as in the proof of Theorem 2.6. Using the facts that $\{a_k\}$ is increasing, $a_0 = 4$, and Lemma 3.1(d), we have

$$\frac{1}{\lambda_{k+1}} + \frac{2 \xi_{k+1}}{a_k} \leq \frac{1}{\Delta} + \frac{2 \xi_{k+1}}{a_0} \leq \frac{1}{\Delta} + \frac{1}{2} \xi$$

for every $k \geq 0$. This conclusion together with the definition of $\bar{M}$, assumption (A3) and (35) then imply that

$$\min_{1 \leq i \leq k} \|v_i\| \leq \min_{0 \leq i \leq k-1} \left( \frac{1}{\lambda_{i+1}} + \frac{2 \xi_{i+1}}{a_i} + \bar{M} \right) \|y_{i+1} - \tilde{x}_i\| \leq \left( \frac{1}{\lambda} + \frac{1}{2} \xi + \bar{M} \right) \min_{0 \leq i \leq k-1} \|y_{i+1} - \tilde{x}_i\|. \quad (46)$$

Moreover, by definitions of $\Delta$ and $\bar{\xi}$ in (38), and the fact that $\bar{m} \leq \bar{M}$, we have

$$\frac{1}{\Delta} + \frac{1}{2} \bar{\xi} + \bar{M} \leq \left( \frac{\theta}{0.9} + 3 \right) \left( \frac{1}{\lambda_0} + \xi_0 + \bar{M} \right) \leq (3\theta + 6) \sqrt{\frac{C_1}{\xi_0} \frac{\xi_0}{\lambda_0}}.$$

Using the above two inequalities, Lemma 3.2 and Lemma 3.1(d), and rearranging terms, we obtain

$$\left( \frac{1}{20} \sum_{i=0}^{k-1} A_{i+1} \right) \min_{1 \leq i \leq k} \|v_i\|^2 \leq (3\theta+6)^2 C_1 \left[ 2A_0(\phi(y_0) - \phi_\ast) + 3\xi_0 D_h^2 k + [\bar{m}D_h^2 + 2(\phi(y_0) - \phi_\ast)] \sum_{i=0}^{k-1} a_i \right].$$

The complexity bound (43) now follows immediately from the above inequality and Lemma A.1 in [16].

(b) The proof of this statement is similar to the proof of (a) except that Lemma A.1 is used in place of Lemma 3.2.

(c) It suffices to argue that the total number of times that the pair $(\lambda, \xi)$ is updated inside all calls to the subroutine SUB is bounded by the second term in (45). Indeed, this assertion follows from the following facts: the initial value of $(\lambda, \xi)$ is $(\lambda_0, \xi_0)$ (see step 1 of ADAP-NC-FISTA); in view of (42) and (43), the pair $(\lambda, \xi)$ is no longer updated whenever $\lambda \leq 0.9/\bar{M}$ and $\xi \geq 2 \bar{m}$, and;
due to (36) and (37), $\lambda$ is reduced by a factor less than or equal to $\theta > 1$ and $\xi$ is increased by a factor of 2 each time either one of them is updated.

A few remarks are made about Theorem 3.3. First, even though we have assumed throughout the paper that $\bar{m} > 0$ (see the second remark after assumptions (A1)-(A4)), both bounds (43) and (44) still hold when $\bar{m} = 0$, i.e., when $f$ is convex on $\text{dom} h$. Second, (43) and (44) are quite similar for the case in which $\bar{m} > 0$. Third, for the case in which $\bar{m} = 0$, though, in contrast to (43), the bound (44) with $\bar{m} = \xi_0 = 0$ yields an $O((1/\hat{\rho})^{2/3})$ iteration-complexity for finding $(\bar{y}, \bar{v})$ as in statement (a). It is worth mentioning that this bound is the same as the one obtained for a FISTA-type ACG variant studied in [17] (see Proposition 5.2) under the assumption that $\bar{m} = 0$.

Corollary 3.4 If $\Omega(\bar{m}) \leq \xi_0 \leq O\left(\max\{1/\lambda_0, \bar{M}\}\right)$, then the total number of inner iterations performed by ADAP-NC-FISTA is

$$
O\left(\left(C_2 \left[\frac{\phi(y_0) - \phi_* + (d_0^2/\lambda_0)}{\bar{\rho}^2}\right]\right)^{1/3} + \left(\frac{C_2 \xi_0 D_h^2 \bar{\rho}^2}{\lambda_0^2}\right)^{1/2} + \frac{C_2 \bar{m} D_h^2}{\bar{\rho}^2} + \log^+ (\lambda_0 \bar{M})\right),
$$

where

$$
C_2 := \left(\frac{1}{\lambda_0} + \bar{M}\right)^2 \lambda_0.
$$

Proof: The conclusion of the corollary follows immediately from Theorem 3.3(b) and the assumption on $\xi_0$.

We end this section by making a remark about the complexity derived in Corollary 3.4. The best choice of $\lambda_0$ which minimizes the constant $C_2$ is $\lambda_0 = \Theta(1/\bar{M})$. However, computational experiments indicate that taking larger values for $\lambda_0$ improves the performance of the method. One reason that may explain this phenomenon is that the constant $\bar{M}$ that appears in (46), and as a consequence in either $C_1$ or $C_2$, is very conservative since it can actually be replaced by the potentially smaller quantity

$$
\bar{L}_k := \frac{\|\nabla f(y_{k+1}) - \nabla f(\bar{x}_k)\|}{\|y_{k+1} - \bar{x}_k\|},
$$

where $\bar{k} = \arg\min_i \{\|y_i - \bar{x}_{i-1}\| : 1 \leq i \leq k\}$.

4 Computational Results

This section reports the experimental results of NC-FISTA and ADAP-NC-FISTA on three problems: nonconvex quadratic programming problem in both vector and matrix cases and nonnegative matrix factorization.

4.1 Nonconvex quadratic programming problem

This subsection discusses the performance of NC-FISTA and its adaptive variant to solve the same quadratic programming problem as in [12, 16], namely:

$$
\min \left\{ f(z) := -\frac{\alpha_1}{2}\|DBz\|^2 + \frac{\alpha_2}{2}\|A z - b\|^2 : z \in \Delta_n \right\},
$$

where $(\alpha_1, \alpha_2) \in \mathbb{R}_{++}^2$, $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix with diagonal entries sampled from the discrete uniform distribution $\mathcal{U}\{1, 1000\}$, matrices $A \in \mathbb{R}^{l \times n}$, $B \in \mathbb{R}^{n \times n}$ and vector $b \in \mathbb{R}^n$. The experimental results show that ADAP-NC-FISTA significantly outperforms NC-FISTA in terms of iteration-complexity and CPU time.
$\mathbb{R}^l$ are such that their entries are generated from the uniform distribution $U[0, 1]$, and $\Delta_n := \{z \in \mathbb{R}^n : \sum_{i=1}^n z_i = 1, \ z_i \geq 0 \}$ is the $(n-1)$-dimensional standard simplex. The dimensions are set to be $(l, n) = (20, 300)$. For some chosen curvature pairs $(m, M) \in \mathbb{R}_+^2$, the scalars $\alpha_1$ and $\alpha_2$ were chosen so that $M = \lambda_{\max}(\nabla^2 f)$ and $-m = \lambda_{\min}(\nabla^2 f)$ where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and smallest eigenvalues functions, respectively.

Our computational results presented below compare three methods: NC-FISTA, ADAP-NC-FISTA and the AG method proposed in [5], only since both are natural extensions of accelerated gradient variants for solving convex programs to the context of nonconvex optimization problems. We remark that their major difference lies in the fact that the AG method requires two resolvent evaluations of $\partial h$ per iteration while NC-FISTA and its adaptive variant require only one (see the third remark after NC-FISTA). In our implementation, all methods use the centroid of $\Delta_n$ as the initial point $z_0$ and terminate with a pair $(z, v)$ satisfying

$$v \in \nabla f(z) + N_{\Delta_n}(z), \quad \frac{\|v\|}{\|\nabla f(z_0)\| + 1} \leq 10^{-7}.$$  \hspace{1cm} (48)

For the AG method, we choose parameters $(\alpha_k, \beta_k, \lambda_k) = \left(\frac{2}{(k+1)}, 0.99/M, k\beta_k/2\right)$ for $k \geq 1$. The parameter pair $(\lambda, \xi) = (0.99/M, 1.05m)$ is chosen for NC-FISTA while the input triple $(\lambda_0, \xi_0, \theta) = (1, 1, 1.25)$ is used for ADAP-NC-FISTA. We implement all methods in MATLAB 2017b scripts and run them on a MacBook Pro with a 4-core Intel Core i7 processor and 16 GB of memory.

Table 1 and Table 2 present computational results which show the performance of all methods for different choices of parameter pairs $(m, M)$. The number of iterations are reported in the last three columns. The bold numbers highlight the method which has the best performance for each case. In the column "Function Value", we also present the objective function value of (47) in the last iteration of each case. Only one value is reported for each case, since the results are approximately the same for all methods.

| Size    | Function Value | Iteration Count |
|---------|----------------|-----------------|
| $M$     | $m$            | AG  | NC  | ADAP |
| 16777216| 16777216       | -2.24E+05 | 638  | 3732 | 18   |
| 16777216| 1048576        | -3.83E+04 | 2466 | 928  | 111  |
| 16777216| 65536          | -4.46E+02 | 12365| 2053 | 326  |
| 16777216| 4096           | 4.07E+03  | 16718| 9431 | 401  |
| 16777216| 256            | 4.38E+03  | 14467| 17093| 360  |
| 16777216| 16             | 4.40E+03  | 14457| 15860| 386  |

Table 1: Numerical results for AG, NC-FISTA and ADAP-NC-FISTA

| Size    | Function Value | Iteration Count |
|---------|----------------|-----------------|
| $M$     | $m$            | AG  | NC  | ADAP |
| 4000    | 1              | 9.68E-01 | 15830| 9431 | 176  |
| 16000   | 1              | 4.11E-00 | 16790| 14717| 282  |
| 64000   | 1              | 1.67E-01 | 14468| 17094| 361  |
| 256000  | 1              | 6.71E-01 | 13879| 18512| 360  |
| 1024000 | 1              | 2.68E-02 | 14457| 15860| 386  |
| 4096000 | 1              | 1.07E+03 | 14457| 15857| 386  |

Table 2: Numerical results for AG, NC-FISTA and ADAP-NC-FISTA
From Table 1 and 2, we conclude that ADAP-NC-FISTA performs fewer number of iterations than both AG and NC-FISTA. Note that ADAP-NC-FISTA exceeds the stepsize limitation \( \lambda < 1/M \) in AG and NC-FISTA.

### 4.2 Matrix problem

In this subsection, we test our methods on a matrix version of the nonconvex quadratic programming problem

\[
\min \left\{ f(Z) : = -\frac{\alpha_1}{2} \|DB(Z)\|^2 + \frac{\alpha_2}{2} \|A(Z) - b\|^2 : Z \in P_n \right\},
\]

where \( A : S^n_+ \rightarrow \mathbb{R}^l \) and \( B : S^n_+ \rightarrow \mathbb{R}^n \) are linear operators defined by

\[
[A(Z)]_i = \langle A_i, Z \rangle_F \text{ for } A_i \in \mathbb{R}^{n \times n} \text{ and } 1 \leq i \leq l,
\]

\[
[B(Z)]_j = \langle B_j, Z \rangle_F \text{ for } B_j \in \mathbb{R}^{n \times n} \text{ and } 1 \leq j \leq n,
\]

with entries of \( A_i, B_j \) sampled from the uniform distribution \( U[0, 1] \), and \( P_n \) denotes the spectraplex \( \{ z \in S^n_+ : \text{tr}(z) = 1 \} \).

Both number of iterations and running time are compared among AG, NC-FISTA and ADAP-NC-FISTA, since the resolvent evaluation performs eigenvalue decomposition and running time is affected by the expensive resolvent evaluation. All methods used the centroid of \( P_n \) as the initial point \( z_0 \), i.e. \( z_0 = I_n/n \), where \( I_n \) is the identity matrix of size \( n \times n \). Termination criterion is the same as \( (48) \) except that \( \Delta_n \) is replaced by \( P_n \). The parameters for the AG and NC-FISTA are chosen in the same way as described in Subsection 4.1, and input triple \((\lambda_0, \xi_0, \theta) = (1, 1000, 1.25)\) is chosen for ADAP-NC-FISTA.

| Size          | Function Value | Iteration Count | Running time(s) |
|---------------|----------------|-----------------|-----------------|
| \( M \)       | \( m \)        | AG | NC | ADAP | AG | NC | ADAP |
| 1000000       | 1000000        | -2.06E+05       | 46 | 38 | 44 | 1.63 | 0.96 | 1.69 |
| 1000000       | 1000000        | -3.65E+03       | 3809 | 7280 | 2209 | 137.73 | 187.10 | 88.61 |
| 1000000       | 1000000        | -1.74E+02       | 5400 | 2052 | 2595 | 197.46 | 52.73 | 103.85 |
| 1000000       | 1000000        | 2.05E+01        | 4621 | 3136 | 2641 | 163.07 | 48.88 | 84.33 |
| 1000000       | 1000000        | 3.67E+01        | 4476 | 8835 | 2643 | 157.14 | 235.21 | 109.31 |

Table 3: Numerical results for AG, NC-FISTA and ADAP-NC-FISTA

In Table 3 the dimensions are set to be \((l, n) = (50, 200)\) and 2.5% of entries in \( A_i, B_j \) are nonzero.

| Size          | Function Value | Iteration Count | Running time(s) |
|---------------|----------------|-----------------|-----------------|
| \( M \)       | \( m \)        | AG | NC | ADAP | AG | NC | ADAP |
| 1000000       | 1000000        | -1.78E+05       | 44 | 35 | 42 | 4.36 | 2.42 | 4.06 |
| 1000000       | 1000000        | -4.41E+03       | 1411 | 1174 | 534 | 134.39 | 79.81 | 54.32 |
| 1000000       | 1000000        | 2.12E+03        | 1963 | 701 | 872 | 195.25 | 48.88 | 84.33 |
| 1000000       | 1000000        | 2.54E+03        | 1935 | 3023 | 904 | 192.94 | 207.36 | 90.21 |
| 1000000       | 1000000        | 2.58E+03        | 1934 | 5767 | 907 | 189.45 | 497.81 | 91.06 |

Table 4: Numerical results for AG, NC-FISTA and ADAP-NC-FISTA
In Table 4, the dimensions are set to be \((l, n) = (50, 400)\) and \(0.5\%\) of entries in \(A_i, B_j\) are nonzero.

| Size | Function Value | Iteration Count | Running time(s) |
|------|----------------|-----------------|-----------------|
| \(M\) | \(m\) | AG | NC | ADAP | AG | NC | ADAP |
| 1000000 | 1000000 | -7.55E+04 | 69 | 62 | **31** | 21.99 | 18.87 | **10.57** |
| 1000000 | 100000 | 1.02E+03 | 277 | 108 | **26** | 118.97 | 22.68 | **9.31** |
| 1000000 | 10000 | 8.21E+03 | 491 | 523 | **61** | 173.33 | 110.00 | **20.71** |
| 1000000 | 1000 | 8.86E+03 | 531 | 1292 | **69** | 168.91 | 273.24 | **21.73** |
| 1000000 | 100 | 8.93E+03 | 535 | 1580 | **70** | 171.76 | 333.30 | **21.75** |

Table 5: Numerical results for AG, NC-FISTA and ADAP-NC-FISTA

In Table 5, the dimensions are set to be \((l, n) = (50, 800)\) and \(0.1\%\) of entries in \(A_i, B_j\) are nonzero.

From Tables 3, 4 and 5, we conclude that ADAP-NC-FISTA is superior to NC-FISTA and AG in both number of iterations and running time.

### 4.3 Nonnegative matrix factorization

In this subsection, we further test ADAP-NC-FISTA on a real life application rather than artificially generated problems and data. Nonnegative matrix factorization (NMF) is a popular dimension reduction method in which a data matrix \(X\) is factored into two matrices \(V\) and \(W\), with constraints that each entry in \(V\) and \(W\) is nonnegative.

\[
\min \left\{ f(V, W) := \frac{1}{2}\|X - VW\|_F^2 : V \geq 0, W \geq 0 \right\},
\]

where \(X \in \mathbb{R}^{n \times m}\), \(V \in \mathbb{R}^{n \times k}\) and \(W \in \mathbb{R}^{k \times m}\). Intuitively, the data matrix \(X\) is a collection of \(m\) data points in \(\mathbb{R}^n\), the columns of \(V\) can be viewed as the basis of all data points, and hence each data point is a linear combination of the basis, with weights in the corresponding column in \(W\). Because of its ability of extracting easily interpretable factors and automatically performing clustering, NMF finds a wide range of applications in practice, from text mining to image processing. Most of the NMF algorithms solve (53) in a two-block coordinate descent manner, by alternatively minimizing with respect to one of the two blocks, \(V\) or \(W\), while keeping the other one fixed. Alternating minimization is a natural idea for NMF, since the subproblem in one block is convex.

In this subsection, we apply ADAP-NC-FISTA to solve the nonconvex problem (53) directly by minimizing in \((V, W)\) jointly.

For a preliminary computational test, we apply ADAP-NC-FISTA to facial feature extraction. The problem is as described in (53), to factor out a data matrix into two matrices. The facial image dataset is provided by AT&T Laboratories Cambridge.\(^1\) There are ten different images of each of 40 distinct subjects, and each image is \(92 \times 112\) pixels, with 256 gray levels per pixel. It results in a matrix of size 10,304 \(\times\) 400, where each column of the data matrix is the vectorization of an image.

ADAP-NC-FISTA is benchmarked against the ANLS (Alternating Nonnegative Least Squares) method.\(^2\) ANLS alternatively solves minimization subproblems in \(V\) and \(W\) with nonnegative constraints and the other variable being fixed. We use the implementation of ANLS provided

---

1. https://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html
2. https://www.cc.gatech.edu/ hpark/nmfsoftware.html
by the authors of [10] as a benchmark for comparison. The ANLS code is modified with stopping criterion (54).

Both methods use the initial point \((V_0, W_0) = (1^{n \times k}/(nk), 1^{k \times m}/(km))\), where \(1^{n \times k}\) and \(1^{k \times m}\) are all one matrices of size \(n \times k\) and \(k \times m\). \(k\) is set to be 20. ADAP-NC-FISTA terminates with a pair \(((V, W), (S_V, S_W))\) satisfying

\[
(S_V, S_W) \in \nabla f(V, W) + N_F(V, W), \quad \frac{\| (S_V, S_W) \|_F}{\| \nabla f(V_0, W_0) \|_F + 1} \leq 10^{-7},
\]

where \(F = \{(V, W) : V \geq 0, W \geq 0\}\). The input triple \((\lambda_0, \xi_0, \theta) = (1, 100, 1.25)\) is chosen for the ADAP-NC-FISTA.

| Method       | Function Value | Iteration Count | Running time(s) |
|--------------|----------------|-----------------|-----------------|
| ADAP-NC-FISTA| 2.80E+09       | 47              | 6.94            |
| ANLS         | 1.20E+09       | 1000            | 137.58          |

Table 6: Numerical results for AG, NC-FISTA and ADAP-NC-FISTA

ADAP-NC-FISTA takes fewer number of iterations and time to find a stationary point than ANLS, but ANLS finds one with a smaller objective function value.

5 Concluding remarks

This paper presents two ACG variants and establishes their iteration-complexities for obtaining approximate stationary points of the SNCO problem. Numerical results are also given showing that they are both efficient in practice.

We have not assumed in our analysis that the set \(\Omega\) as in assumption (A3) is bounded. However, we remark that if \(\Omega\) is bounded then it can be shown using a simpler analysis than the one given in this paper that the version of the NC-FISTA with \(\xi = 0\) and \(\lambda = 1/(2M)\) has an

\[
O \left( \frac{M^2 \tilde{d}_0^2}{\rho^2} \right)^{1/3} + \left( \frac{M \tilde{m} D_\Omega^2}{\rho^2} \right)^{1/2} \left[ \frac{M \tilde{m} D_\Omega^2}{\rho^2} + 1 \right]
\]

iteration-complexity where \(D_\Omega := \sup_{u, u' \in \Omega} \| u' - u \| < \infty\). Moreover, it can be shown that a version of the ADAP-NC-FISTA in which \(\lambda_k\) is updated is a similar way and \(\xi_k = 0\) for every \(k\) has a guaranteed iteration-complexity which lies in between the one above and the one in (43).

Finally, we have implemented the two versions mentioned in the previous paragraph and tested them on problems for which \(\Omega\) is bounded but have observed that they are not as efficient as the corresponding ones studied in this paper.

References

[1] 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.

[2] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

[3] Y. Chen, G. Lan, and Y. Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.
[4] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, Jul 2018.

[5] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156:59–99, 2016.

[6] S. Ghadimi, G. Lan, and H. Zhang. Generalized uniformly optimal methods for nonlinear programming. *Available on arXiv:1508.07384*, 2015.

[7] O. Güler. New proximal point algorithms for convex minimization. *SIAM Journal on Optimization*, 2(4):649–664, 1992.

[8] Y. He and R. D. C. Monteiro. Accelerating block-decomposition first-order methods for solving composite saddle-point and two-player Nash equilibrium problems. *SIAM Journal on Optimization*, 25:2182–2211, 2015.

[9] Y. He and R. D. C. Monteiro. An accelerated HPE-type algorithm for a class of composite convex-concave saddle-point problems. *SIAM Journal on Optimization*, 26:29–56, 2016.

[10] J. Kim and H. Park. Toward faster nonnegative matrix factorization: A new algorithm and comparisons. In *2008 Eighth IEEE International Conference on Data Mining*, pages 353–362. IEEE, 2008.

[11] O. Kolossoski and R. D. C. Monteiro. An accelerated non-Euclidean hybrid proximal extragradient-type algorithm for convex-concave saddle-point problems. *Optimization Methods and Software*, 32:1244–1272, 2017.

[12] W. Kong, J. G. Melo, and R. D. C. Monteiro. Complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained nonconvex composite problem. *Preprint*, 2018.

[13] G. Lan, Z. Lu, and R. D. C. Monteiro. Primal-dual first-order methods with $O(1/\varepsilon)$ iteration-complexity for cone programming. *Math. Programming*, 126(1):1–29, 2011.

[14] H. Li and Z. Lin. Accelerated proximal gradient methods for nonconvex programming. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 379–387, December 2015.

[15] Q. Li, Y. Zhou, Y. Liang, and P. K. Varshney. Convergence analysis of proximal gradient with momentum for nonconvex optimisation. *Available on arXiv:1705.04925*, 2017.

[16] J. Liang and R. D. C. Monteiro. A doubly accelerated inexact proximal point method for nonconvex composite optimization problems. *Available on arXiv:1811.11378, submitted to SIAM Journal on Optimization*, 2018.

[17] R. D. C. Monteiro and B. F. Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM J. Optim.*, 23(2):1092–1125, 2013.

[18] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983.
A Supplementary Results

This section provides a bound on the quantity \( \min_{0 \leq i \leq k-1} \| y_{i+1} - \tilde{x}_i \|^2 \) for the case in which the parameter \( \xi_0 \) of the ADAP-NC-FISTA satisfies \( \xi_0 \geq 2\bar{m} \). Note that an alternative bound on this quantity has already been developed in Lemma 3.2 for any \( \xi_0 > 0 \).

Lemma A.1 For every \( k \geq 1 \), for \( \xi_0 \geq 2\bar{m} \), we have

\[
\frac{1}{20} \left( \sum_{i=0}^{k-1} A_{i+1} \right) \min_{0 \leq i \leq k-1} \| y_{i+1} - \tilde{x}_i \|^2 \leq \lambda_0 A_0 (\phi(y_0) - \phi^*) + \frac{\| x_0 - x^* \|^2}{2} + \lambda_0 D_h^2 \left( \xi_0 + 3\xi_0k + \bar{m} \sum_{i=0}^{k-1} a_i \right). \tag{55}
\]

Proof: Since \( \xi_0 \geq 2\bar{m} \), we have

\[
\left( \bar{m} + \frac{2\xi_0}{a_k} \right) \lambda_{k+1} \leq \xi_0 \lambda_k.
\]

The above inequality indicates that (55) is always satisfied with \( \xi = \xi_0 \) and \( \lambda = \lambda_{k+1} \), and hence \( \xi_k = \xi_0 \), for \( k \geq 0 \). Using similar arguments as in the proof of Lemma 3.2, we conclude that for every \( k \geq 0 \) and \( u \in \Omega \),

\[
\lambda_{k+1} A_{k+1} \phi(y_{k+1}) + \left( \xi_0 \lambda_{k+1} + \frac{1}{2} \right) \| u - x_{k+1} \|^2 + \frac{1}{2} \left( 1 - \lambda_{k+1} M_{k+1} \right) A_{k+1} \| y_{k+1} - \bar{x}_k \|^2 \\
\leq \lambda_{k+1} A_k \gamma_k(y_k) + \lambda_{k+1} a_k \gamma_k(u) + \frac{1}{2} \| u - x_k \|^2, \tag{56}
\]

where

\[
\gamma_k(u) := \bar{\gamma}_k(y_{k+1}) + \frac{1}{\lambda_{k+1}} \langle \bar{x}_k - y_{k+1}, u - y_{k+1} \rangle + \frac{\xi_0}{a_k} \| u - y_{k+1} \|^2.
\]
and
\[ \tilde{\gamma}_k(u) := \ell_f(u; \tilde{x}_k) + h(u) + \frac{\xi_0}{a_k} \| u - \tilde{x}_k \|^2. \]  
(57)

As in Lemma 2.2(a), we have \( \gamma_k(u) \leq \tilde{\gamma}_k(u) \) for every \( u \in \text{dom} \ h \). Hence, it follows from (41) and (4) that for every \( k \geq 0 \) and \( u \in \text{dom} \ h \), we have
\[ \gamma_k(u) - \phi(u) \leq \tilde{\gamma}_k(u) - \phi(u) = \ell_f(u; \tilde{x}_k) - f(u) + \frac{\xi_0}{a_k} \| u - \tilde{x}_k \|^2 \leq \frac{1}{2} \left( \bar{m} + \frac{2\xi_0}{a_k} \right) \| u - \tilde{x}_k \|^2. \]  
(58)

Taking \( u = x^* \), and using (56), (23), (58), (55), Lemma 3.1(b), and the facts that \( x_0 = y_0 \) and \( \lambda_{i+1} \leq \lambda_i \), we conclude that for every \( 0 \leq i \leq k - 1 \),
\[ 0.1 \frac{1}{2} A_{i+1} \| y_{i+1} - \tilde{x}_i \|^2 \]
\[ + \left[ \lambda_{i+1} A_{i+1} (\phi(y_{i+1}) - \phi_s) + \frac{1}{2} \| x^* - x_{i+1} \|^2 \right] - \left[ \lambda_i A_i (\phi(y_i) - \phi_s) + \frac{1}{2} \| x^* - x_i \|^2 \right] \]
\[ \leq \lambda_{i+1} A_i (\gamma_i(y_i) - \phi(y_i)) + \lambda_{i+1} a_i (\gamma_i(x^*) - \phi_s) + (\lambda_{i+1} - \lambda_i) A_i (\phi(y_i) - \phi_s) - \xi_0 \lambda_{i+1} |x_{i+1} - x^*|^2 \]
\[ \leq \lambda_{i+1} \left( \bar{m} + \frac{2\xi_0}{a_i} \right) \left( A_i \| y_i - \tilde{x}_i \|^2 + a_i \| x^* - \tilde{x}_i \|^2 \right) - \xi_0 \lambda_{i+1} |x_{i+1} - x^*|^2 \]
\[ \leq \lambda_{i+1} \left( \bar{m} + \frac{2\xi_0}{a_i} \right) \left( \| x^* - x_i \|^2 + (1 + a_i) D_h^2 \right) - \xi_0 \lambda_{i+1} |x_{i+1} - x^*|^2 \]
\[ \leq \xi_0 (\lambda_i \| x_i - x^* \|^2 - \lambda_{i+1} |x_{i+1} - x^*|^2) + (3\xi_0 + \bar{m} a_i) \lambda_i D_h^2. \]

The conclusion is obtained by summing the above inequality from \( i = 0 \) to \( k - 1 \).