A FISTA-Type First Order Algorithm on Composite Optimization Problems that is Adaptable to the Convex Situation

Chee-Khian Sim

May 25, 2020

Abstract

In this note, we propose a FISTA-type first order algorithm, VAR-FISTA, to solve a composite optimization problem. A distinctive feature of VAR-FISTA is its ability to exploit the convexity of the function in the problem, resulting in an improved iteration complexity when the function is convex compared to when it is nonconvex. The iteration complexity result for the convex and nonconvex case obtained in the note are compatible to the best known in the literature so far.

Keywords. Fast iterative shrinkage thresholding algorithm (FISTA); Composite optimization problem; Iteration complexity.

1 Introduction

Using first order methods is the preferred approach to solve large scale optimization problems that arise in application areas such as machine learning. Fast iterative shrinkage thresholding algorithm (FISTA), an efficient first order method, is proposed in [1] to solve composite optimization problems when the functions involved are convex; see also [11, 12, 13]. Recently, there are interests in the study of first order algorithms, such as FISTA variants, to solve composite optimization problems with nonconvex functions. These works include [2, 3, 4, 5, 6, 7, 8, 9, 10, 14, 15].

In this note, we propose a FISTA-type first order algorithm, VAR-FISTA, to solve composite optimization problems. This algorithm is inspired by the algorithm ADAP-NC-FISTA in [10]. The algorithm in this note is designed in such a way that when the functions involved in the composite optimization problem are convex, it is able to exploit the convexity of the problem leading to an iteration complexity of $\mathcal{O}((1/\hat{\rho})^{2/3})$, while an iteration complexity of $\mathcal{O}(1/\hat{\rho}^2)$ is achieved in the nonconvex case. These complexity results are the best known in the literature so far. The contribution of this note is twofold. First, the algorithm requires only one resolvent evaluation in an iteration, unlike the algorithms in [4, 6]. Second, other than information on function and gradient values, no other data information, such as Lipschitz constant or lower curvature, are required from the problem, as in [3, 15], for the algorithm to run. It should finally be noted that the algorithm, ADAP-NC-FISTA, in [10] also shares these same features as our algorithm, but in [10], the iteration complexity of $\mathcal{O}((1/\hat{\rho})^{2/3})$ when the functions in the composite optimization problem are convex cannot be directly established.

2 A Composite Optimization Problem

We consider the following composite optimization problem:

$$\min_{u\in\Re^n} \phi(u) := f(u) + h(u), \tag{1}$$

where f is continuously differentiable, can be nonconvex on Ω,

$$\|\nabla f(u_1) - \nabla f(u_2)\| \le M\|u_1 - u_2\|, \ \forall\, u_1, u_2 \in \Omega, \tag{2}$$

with $M > 0$ and Ω is a closed convex set in $\Re^n$, that is, the gradient of f is Lipschitz continuous on Ω, and h is a proper lower semi-continuous convex function, which can be nonsmooth, with dom $h \subset \Re^n$ being closed and bounded. We assume that dom $h \subseteq \Omega$. Let $\overline{M}(> 0)$ be the smallest M satisfying (2). There exists $m \ge 0$ such that

$$-\frac{m}{2}\|u_1 - u_2\|^2 \le f(u_1) - \ell_f(u_1; u_2), \forall\, u_1, u_2 \in \Omega, \tag{3}$$

where

$$\ell_f(u_1; u_2) := f(u_2) + \langle \nabla f(u_2), u_1 - u_2\rangle, \tag{4}$$

since by (2),

$$|f(u_1) - l_f(u_1; u_2)| \le \frac{M}{2}\|u_1 - u_2\|, \ \forall\, u_1, u_2 \in \Omega. \tag{5}$$

Hence, an m that satisfies (3) is $\overline{M}$. Let $\underline{m} \ge 0$ be the smallest $m \ge 0$ such that (3) holds. We have $0 \le \underline{m} \le \overline{M}$. Observe that if $\underline{m} = 0$, then by (3), f is convex on Ω, while if $\underline{m} > 0$, then f is nonconvex on Ω.

Let us denote $y^* \in$ dom h to be an optimal solution to Problem (1), which exists since dom h is closed and bounded.

A necessary condition for $u \in \text{dom}\, h$ to be a local minimum of Problem (1) is $0 \in \nabla f(u) + \partial h(u)$. Motivated by this condition, we have the notion of an $\hat{\rho}$-approximate solution of Problem (1), which is a pair $(\hat{u}, \hat{v})$ which satisfies

$$\hat{v} \in \nabla f(\hat{u}) + \partial h(\hat{u}), \quad \|\hat{v}\| \le \hat{\rho}, \tag{6}$$

where $\hat{\rho} > 0$ is a given tolerance.

3 A FISTA-Type Algorithm on Problem (1)

In the following, we propose a variant of FISTA, which we call VAR-FISTA, to find an $\hat{\rho}$-approximate solution to Problem (1), for a given tolerence $\hat{\rho} > 0$. This algorithm is inspired by the algorithm ADAP-FISTA in [10].

VAR-FISTA

Initialization: Let $\xi_0 = 0$, $\lambda_0 > 0$, $\theta > 1$, $0 < \gamma < 1$, tolerance $\hat{\rho} > 0$ and initial point $y_0 \in \text{dom}\, h$, and set $y_0^{\min} = x_0 = y_0$, $A_0 = 12$, $L_0 = 0$.

$\boldsymbol{k}^{\textbf{th}}$ **Iteration** ($k = 1, 2, \ldots$):

$k1$. Set $\lambda = \lambda_{k-1}$, $\xi = \xi_{k-1}$ and compute

$$a_{k-1} = \frac{1+\sqrt{1+4A_{k-1}}}{2}, \quad A_k = A_{k-1} + a_{k-1}, \quad \tilde{x}_k = \frac{A_{k-1}}{A_k} y_{k-1} + \frac{a_{k-1}}{A_k} x_{k-1}; \tag{7}$$

$k2$. compute

$$\tau = \frac{2\xi\lambda}{a_{k-1}},$$

$$y = \text{argmin}_u \left\{ l_f(u;\tilde{x}_k) + h(u) + \frac{1+\tau}{2\lambda}\|u-\tilde{x}_k\|^2 \right\}, \tag{8}$$

$$U = \frac{2\left[f(y) - \ell_f(y;\tilde{x}_k)\right]}{\|y-\tilde{x}_k\|^2}; \tag{9}$$

$$\tilde{y}^{\min} = \text{argmin}\left\{\phi(\tilde{y})\ ;\ \tilde{y} = y_{k-1}^{\min}, y\right\}, \tag{10}$$

$$L = \max\left\{ \frac{2[\ell_f(y_{k-1};\tilde{x}_k) - f(y_{k-1})]}{\|y_{k-1}-\tilde{x}_k\|^2}, \frac{2[\ell_f(\tilde{y}^{\min};\tilde{x}_i) - f(\tilde{y}^{\min})]}{\|\tilde{y}^{\min}-\tilde{x}_i\|^2}, L_{k-1}, 0\ ;\ i = 1,\ldots,k \right\}; \tag{11}$$

$k3$. If $U\lambda > \gamma$ or $\xi\lambda_{k-1} < L\lambda + \tau$ or $\xi\lambda_{i-1} < L\lambda_i + \tau_i$ for some $i = 1,\ldots,k-1$, go to step $k2$ with (ξ,λ) given by

$$(\xi,\lambda) = \text{UPDATE}(\xi,\lambda,\lambda_1,\ldots,\lambda_{k-1},\tau,\tau_1,\ldots,\tau_{k-1},L,U,\theta,\gamma);$$

else set $\tau_k = \tau$, $y_k^{\min} = \tilde{y}^{\min}$, $y_k = y$, $\lambda_k = \lambda$, $U_k = U$, $L_k = L$ and $\xi_k = \xi$;

$k4$. compute

$$x_k = P_\Omega\left(\frac{(1+\tau_k)A_k}{a_{k-1}(\tau_k a_{k-1}+1)} y_k - \frac{A_{k-1}}{a_{k-1}(\tau_k a_{k-1}+1)} y_{k-1} \right), \tag{12}$$

$$v_k = \frac{1+\tau_k}{\lambda_k}(\tilde{x}_k - y_k) + \nabla f(y_k) - \nabla f(\tilde{x}_k). \tag{13}$$

Termination: If at the end of the k^{th} iteration, $\|v_k\| \le \hat{\rho}$, then output $(\hat{y},\hat{v}) = (y_k, v_k)$, and **exit**.

We describe the subroutine in VAR-FISTA in the following:

```
$(\xi,\lambda) = \text{UPDATE}(\xi,\lambda,\lambda_1,\ldots,\lambda_{k-1},\tau,\tau_1,\ldots,\tau_{k-1},L,U,\theta,\gamma)$:
if $U\lambda > \gamma$ then set
    $\lambda \leftarrow \min\{\lambda/\theta, \gamma/U\};$    (14)
if
    $\xi\lambda_{k-1} < L\lambda + \tau$ or $\xi\lambda_{i-1} < L\lambda_i + \tau_i$ for some $i = 1,\ldots,k-1$    (15)
then
    if $\xi = 0$, set
        $\xi \leftarrow 1;$    (16)
    else set
        $\xi \leftarrow 2\xi;$    (17)
```

In the above algorithm, VAR-FISTA, steps $k2$ and $k3$ can be performed more than once in an iteration since if the conditions in step $k3$ are not satisfied, then step $k2$ needs to be performed again with an updated (ξ, λ) obtained from the subroutine in step $k3$. The conditions in step $k3$ are then checked again. This will continue until the conditions are satisfied. It should be noted however that the total number of times this occurs in an iteration is bounded. In fact, if N_0 is the number of iterations taken by the algorithm before termination, then the total number of executions of steps $k2$ and $k3$ is bounded by $N_0 + \max\left\{\frac{\log(\lambda_0/\lambda)}{\log\theta}, \frac{\log\bar{\xi}}{\log 2}, 0\right\}$.

There is no particular reason for setting A_0 to be 12 in the above algorithm. We do this for the sake of convenience. A_0 can be set to any positive number without affecting the results in this paper.

Note that in the above algorithm, for all $k \geq 1$, $y_k^{\min} \in \text{dom } h$ and $x_k, \tilde{x}_k \in \Omega$.

We remark that for every $k \geq 1$, we have $y_k \in \text{dom } h$ and $v_k \in \nabla f(y_k) + \partial h(y_k)$. Hence, upon termination of the algorithm, we obtain an $\hat{\rho}$-approximate solution, $(\hat{y}, \hat{v})$, of Problem (1).

Also, we remark that λ_k in the algorithm can be viewed as an estimation of the reciprocal of $\overline{M}$, while ξ_k is an estimation of $\underline{m}$. Hence, when f is convex, in which case $\underline{m} = 0$, $\xi_k = 0$, which also implies that $\tau_k = 0$ for all $k \geq 1$. The algorithm then reduces to FISTA with constant stepsize [1] on Problem (1) when we set $\lambda_k = 1/\overline{M}$ for all $k \geq 0$.

Note that $\{a_k\}$ and $\{A_k\}$ in (7) are related by

$$A_{k+1} = a_k^2. \tag{18}$$

Furthermore, we observe that by defining for $k \geq 1$

$$\tilde{\gamma}_k(u) := l_f(u; \tilde{x}_k) + h(u) + \frac{\tau_k}{2\lambda_k}\|u - \tilde{x}_k\|^2, \ u \in \text{dom } h, \tag{19}$$

$$\gamma_k(u) := \tilde{\gamma}_k(y_k) + \frac{1}{\lambda_k}\langle \tilde{x}_k - y_k, u - y_k\rangle + \frac{\tau_k}{2\lambda_k}\|u - y_k\|^2, \ u \in \Omega, \tag{20}$$

it is easy to check that x_k given in (12) is the unique optimal solution to the following optimization problem.

$$\min_{u\in\Omega} a_{k-1}\gamma_k(u) + \frac{1}{2\lambda_k}\|u - x_{k-1}\|^2. \tag{21}$$

In the definition of γ_k in (20), we note that aside from the quadratic term, γ_k is the "linearization" of $\tilde{\gamma}_k$ at $u = y_k$.

4 Iteration Complexity Results for VAR-FISTA

In this section, we derive iteration complexity results as stated in Theorem 4.10 to find an $\hat{\rho}$-approximate solution to Problem (1) using VAR-FISTA.

First, we define

$$D_h := \sup_{u_1, u_2 \in \text{dom } h} \|u_1 - u_2\|. \tag{22}$$

Note that D_h is finite since dom h is bounded.

We need the following results on $\{a_k\}$ and $\{A_k\}$ in deriving these iteration complexity results. The proof of the lemma is given in the appendix.

Lemma 4.1 *For every $k \geq 1$, the sequences $\{a_k\}$ and $\{A_k\}$ given in* (7) *satisfy*

$$\frac{k}{2} \leq a_{k-1} \leq 4k, \quad \sum_{i=1}^{k} A_i \geq \frac{k^3}{12}, \quad \frac{\sum_{i=1}^{k} a_{i-1}}{\sum_{i=1}^{k} A_i} \leq \frac{4}{k}.$$

In the following lemma, we put together properties of $\tau_k, y_k^{\min}, \lambda_k, U_k, L_k$ and ξ_k in VAR-FISTA. These results are useful in our analysis later.

Lemma 4.2 *The following statements hold for VAR-FISTA:*

(a) $\{\lambda_k\}$ is positive, non-increasing; $\{\xi_k\}$ and $\{L_k\}$ are non-negative, non-decreasing;

(b) for every $k \geq 1$,

$$\begin{aligned} &U_k \leq \overline{M}, \quad L_k \leq \underline{m}, \quad \tau_k = \frac{2\xi_k \lambda_k}{a_{k-1}}, \quad U_k \lambda_k \leq \gamma, \\ &\xi_k \lambda_{i-1} \geq L_k \lambda_i + \tau_i \geq L_i \lambda_i + \tau_i \geq 0, \ i = 1, \ldots, k, \\ &y_k^{\min} = \text{argmin} \left\{ \phi(\tilde{y}) \ ; \ \tilde{y} \in \{y_0, \ldots, y_k\} \right\}; \end{aligned}$$

(c) for every $k \geq 0$, $\lambda_k \geq \underline{\lambda} := \min\{\gamma/(\theta \overline{M}), \lambda_0\}$, $\xi_k \leq \max\{4\underline{m}, 1\}$; furthermore, if f is convex, then $\xi_k = 0$ for every $k \geq 0$.

Proof: (a) The first statement follows from $\lambda_0 > 0$, the assumption that $\theta > 1$ and the fact that the update procedure for λ in step $k3$ of VAR-FISTA either leaves λ unchanged or strictly decreases λ according to the update formula (14). That $\{\xi_k\}$ is non-negative, non-decreasing is obvious in view of (16) and (17) in step $k3$ of the algorithm, while $\{L_k\}$ is non-negative, non-decreasing hold due to (11) in step $k2$ of the algorithm.

(b) Since $\overline{M} > 0$ and $\underline{m} \geq 0$ satisfies (5) and (3) respectively, it follows that every quantity U (resp., L) computed in step $k2$ of VAR-FISTA, and hence U_k (resp., L_k), is bounded above by $\overline{M}$ (resp., $\underline{m}$). The other conclusions follow immediately from (a) and the definitions of τ_k, $y_k^{\min}$, y_k, λ_k, U_k, L_k and ξ_k in step $k3$ of VAR-FISTA.

(c) For contradiction, assume that $\lambda_k < \underline{\lambda}$ for some $k \geq 0$. Then, since $\lambda_k < \lambda_0$, λ_k has been obtained from a pair (λ, U) through the update formula (14) and we also have $U > 0$. Since $\overline{M} \geq U > 0$ and $\lambda_k < \gamma/(\theta M)$, it follows that $\gamma/U \geq \gamma/\overline{M} > \lambda_k$. Hence, it follows from (14) that $\lambda_k = \lambda/\theta$. On the other hand, noting that step $k3$ in VAR-FISTA implies that λ is no longer reduced whenever $\lambda \leq \gamma/\overline{M}$, we then conclude that $\lambda > \gamma/\overline{M}$, and hence that $\lambda_k = \lambda/\theta > \gamma/(\theta\overline{M})$. Since the latter conclusion contradicts our initial assumption, the first result in statement (c) follows. To show the second result in statement (c), for contradiction, assume that $\xi_k > \max\{4\underline{m}, 1\}$ for some $k \geq 0$. Since $\xi_k > 1$, we have $k \geq 1$, and we also have $\xi_k = 2\xi$, where ξ satisfies $\xi\lambda_{k-1} < L\lambda + \tau$ or $\xi\lambda_{i-1} < L\lambda_i + \tau_i$ for some $i = 1, \ldots, k-1$, according to (15). By $L \leq \overline{m}$ from (3) and (11), definition of τ and τ_i, $a_i \geq a_0 = 4$, $\lambda \leq \lambda_{k-1}$, $\lambda_i \leq \lambda_{i-1}$ and $\xi \geq \xi_i, i = 1, \ldots, k-1$, we have $L\lambda + \tau \leq \underline{m}\lambda_{k-1} + (\lambda_{k-1}\xi)/2$ or $L\lambda_i + \tau_i \leq \overline{m}\lambda_{i-1} + (\lambda_{i-1}\xi)/2$. Hence, $\xi\lambda_i < \lambda_i(\underline{m} + \xi/2)$ for some $i = 1, \ldots, k-1$, which implies that $\xi < 2\underline{m}$. Therefore, $\xi_k = 2\xi < 4\underline{m}$, which contradicts our initial assumption. The second result in statement (c) then follows. Furthermore, if f is convex, then $\xi = \tau = \tau_i = L = 0$ and hence (15) is always false, which implies that (16) and (17) are never executed. Therefore, $\xi_k = \xi_0 = 0$ for every $k \geq 1$. ■

Lemma 4.2 is similar to Lemma 3.1 in [10].

Remark 4.3 *VAR-FISTA is able to "detect" when f is convex, in which case, ξ_k is always equal to 0 for all k, unlike when f is nonconvex. This leads to better iteration complexity for VAR-FISTA as shown in Theorem 4.10 below. Although the algorithm performs differently in terms of update from ξ_k to ξ_{k+1} depending on whether f is convex or nonconvex, we carry out the analysis to find the iteration complexity for VAR-FISTA in an unified manner. We do this by defining $\bar{\xi}$ to be such that*

$$\bar{\xi} := \begin{cases} \max\{4\underline{m}, 1\}, & \textit{if } \underline{m} > 0, \\ 0, & \textit{if } \underline{m} = 0. \end{cases} \tag{23}$$

It is easy to see from the above definition of $\bar{\xi}$ that its value is zero only when f is convex. Observe also from (23) and Lemma 4.2(c) that $\forall\ k \geq 0$, $\xi_k \leq \bar{\xi}$.

The following lemma is crucial for us to arrive at the iteration complexity results for VAR-FISTA in Theorem 4.10.

Lemma 4.4 *The total number of times, n_0, the value of ξ_k changes as k increases is of the order* $\max\{\log \underline{m}, 1\}$*.*

Proof: We observe that if $\xi \geq 2\underline{m}$, the inequalities in (15) do not hold, which follows from Lemma 4.2(a), $a_k \geq a_0 = 4$ and that L in (11) is always less than or equal to $\underline{m}$. This, together with the update formula (17), leads to the result in the lemma. ■

The following lemma provides a bound on $\|x_k - x_0\|$.

Lemma 4.5 *We have for $k \geq 0$, $\|x_k - x_0\| \leq Ck$, where*

$$C := 2(2 + \bar{\xi}\lambda_0)D_h.$$

Proof: We have for $k \geq 1$, by (7), (12), (18), Lemma 4.2, (22), the last sentence in Remark 4.3 and Lemma 4.1, that

$$\begin{aligned}
\|x_k - x_0\| &\leq \left\| \frac{(1+\tau_k)A_k}{a_{k-1}(\tau_k a_{k-1}+1)} y_k - \frac{A_{k-1}}{a_{k-1}(\tau_k a_{k-1}+1)} y_{k-1} - x_0 \right\| \\
&= \frac{1}{(2\xi_k\lambda_k + 1)a_{k-1}} \|(1+\tau_k)A_k y_k - A_{k-1}y_{k-1} - (\tau_k a_{k-1}+1)a_{k-1}x_0\| \\
&\leq \frac{1}{a_{k-1}} \|(1+\tau_k)A_k y_k - A_{k-1}y_{k-1} - (\tau_k a_{k-1}+1)a_{k-1}x_0\| \\
&= \frac{1}{a_{k-1}} \|A_{k-1}(y_k - y_{k-1}) + (\tau_k a_{k-1}+1)a_{k-1}(y_k - x_0)\| \\
&\leq \frac{D_h}{a_{k-1}}(A_{k-1} + (\tau_k a_{k-1}+1)a_{k-1}) \\
&= D_h\left(\frac{A_{k-1}}{a_{k-1}} + 2\xi_k\lambda_k + 1\right) \leq D_h\left(\frac{A_{k-1}}{a_{k-1}} + 2\bar{\xi}\lambda_0 + 1\right)
\end{aligned}$$

$$\leq \; D_h(a_{k-1} + 2\bar{\xi}\lambda_0) \leq D_h(4k + 2\bar{\xi}\lambda_0) \leq 2(2 + \bar{\xi}\lambda_0)D_h k.$$

The conclusion of the lemma then follows. ■

Below are two technical results that are needed in the analysis to arrive at Theorem 4.10. Proposition 4.6 is used to prove Lemma 4.9.

Proposition 4.6 *For $u \in$ dom h, for every $k \geq 1$, we have*

$$A_{k-1}\|y_{k-1} - \tilde{x}_k\|^2 \leq 2\|u - x_{k-1}\|^2 + 2D_h^2, \tag{24}$$

$$a_{k-1}\|u - \tilde{x}_k\|^2 \leq \frac{2}{a_{k-1}}\|u - x_{k-1}\|^2 + 2a_{k-1}D_h^2. \tag{25}$$

Proof: By the definition of $\tilde{x}_k$ in (7), relations (22) and (18), the fact that $A_k = A_{k-1}+a_{k-1} \geq A_{k-1}$ due to (7), the inequality $\|a + b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$ for any $a, b \in \mathbb{R}^n$, we obtain for $u \in$ dom h,

$$\begin{aligned} A_{k-1}\|y_{k-1} - \tilde{x}_k\|^2 &= \frac{A_{k-1}a_{k-1}^2}{A_k^2}\|x_{k-1} - y_{k-1}\|^2 = \frac{A_{k-1}}{A_k}\|(x_{k-1} - u) - (y_{k-1} - u)\|^2 \\ &\leq \frac{2A_{k-1}}{A_k}\left[\|u - x_{k-1}\|^2 + \|u - y_{k-1}\|^2\right] \leq 2\|u - x_{k-1}\|^2 + 2D_h^2. \end{aligned}$$

Hence, (24) holds. Arguing in a similar manner, (25) holds as well. ■

The following technical result allows us to arrive at Lemma 4.8, which through Lemma 4.9, then leads to Theorem 4.10, the main result of this section. This proposition is also needed in the proof of Lemma 4.9. The proof of this proposition and that of Lemma 4.8 are similar to that of Lemma 2.2 and Lemma 2.3 in [10] respectively, and are provided in the appendix of this note for the sake of completeness.

Proposition 4.7 *$\tilde{\gamma}_k$ defined in (19) and γ_k defined in (20) are (τ_k/λ_k)-strongly convex functions, $\gamma_k(u) \leq \tilde{\gamma}_k(u)$ $\forall$ $u \in$ dom h, $\tilde{\gamma}_k(y_k) = \gamma_k(y_k)$,*

$$\min_u \left\{\tilde{\gamma}_k(u) + \frac{1}{2\lambda_k}\|u - \tilde{x}_k\|^2\right\} = \min_u \left\{\gamma_k(u) + \frac{1}{2\lambda_k}\|u - \tilde{x}_k\|^2\right\}, \tag{26}$$

and these minimization problems have y_k as their unique optimal solution;

Lemma 4.8 *We have for $k \geq 1$, for every $u \in \Omega$,*

$$\begin{aligned} &\lambda_k A_k\phi(y_k) + \frac{\tau_k a_{k-1} + 1}{2}\|u - x_k\|^2 + \frac{(1 - \gamma)A_k}{2}\|y_k - \tilde{x}_k\|^2 \\ &\leq \lambda_k A_{k-1}\gamma_k(y_{k-1}) + \lambda_k a_{k-1}\gamma_k(u) + \frac{1}{2}\|u - x_{k-1}\|^2, \end{aligned} \tag{27}$$

The inequality (27) in the above lemma is the basic inequality fundamental in proving the iteration complexity results for VAR-FISTA, and is the key result needed to show that the following lemma holds.

Lemma 4.9 *For every $k \geq 1$,*

$$\begin{aligned} \frac{1 - \gamma}{2}\left(\sum_{i=1}^{k} A_i\|y_i - \tilde{x}_i\|^2\right) &\leq \lambda_0 A_0(\phi(y_0) - \phi(y_k^{\min})) + \left(\frac{1}{2} + 2\bar{\xi}\lambda_0 n_0\right) D_h^2 \\ &+ 2\bar{\xi}\lambda_0 C^2\left(\sum_{i=1}^{k}\frac{i^2}{a_{i-1}} + n_0 k^2\right) + \bar{\xi}\lambda_0 D_h^2\sum_{i=1}^{k}(3 + a_{i-1}). \end{aligned} \tag{28}$$

Proof: For $k \geq 1$, let $i_j \leq k$, $j \geq 1$, be such that $\xi_i = \xi_{i_j}$ for $i = i_{j-1}+1, \ldots, i_j$, where $i_0 = 0$ and $i_{n_1} = k$. Note that $n_1 \leq n_0$, by Lemma 4.4. From (27) in Lemma 4.8, where we let $u = y_k^{\min}$, for $i_{j-1}+1 \leq i \leq i_j$, we have

$$\begin{aligned}
&\frac{1-\gamma}{2} A_i \|y_i - \tilde{x}_i\|^2 \\
&- \left(\lambda_{i-1} A_{i-1} (\phi(y_{i-1}) - \phi(y_k^{\min})) + \frac{1}{2} \|y_k^{\min} - x_{i-1}\|^2 \right) + \left(\lambda_i A_i (\phi(y_i) - \phi(y_k^{\min})) + \frac{1}{2} \|y_k^{\min} - x_i\|^2 \right) \\
&\leq \lambda_i A_{i-1} (\gamma_i(y_{i-1}) - \phi(y_{i-1})) + \lambda_i a_{i-1} (\gamma_i(y_k^{\min}) - \phi(y_k^{\min})) - \frac{\tau_i a_{i-1}}{2} \|y_k^{\min} - x_i\|^2 \\
&+ (\lambda_i - \lambda_{i-1}) A_{i-1} \left(\phi(y_{i-1}) - \phi(y_k^{\min}) \right). \qquad (29)
\end{aligned}$$

Observe that by Proposition 4.7, the definition of $\tilde{\gamma}_i$ in (19), and L_i in view of (11) that for $i = i_{j-1}+1, \ldots, i_j$,

$$\begin{aligned}
\gamma_i(y_{i-1}) - \phi(y_{i-1}) &\leq \tilde{\gamma}_i(y_{i-1}) - \phi(y_{i-1}) = \ell_f(y_{i-1}; \tilde{x}_i) - f(y_{i-1}) + \frac{\tau_i}{2\lambda_i} \|y_{i-1} - \tilde{x}_i\|^2 \\
&\leq \left(\frac{L_i}{2} + \frac{\tau_i}{2\lambda_i} \right) \|y_{i-1} - \tilde{x}_i\|^2, \qquad (30)
\end{aligned}$$

and

$$\begin{aligned}
\gamma_i(y_k^{\min}) - \phi(y_k^{\min}) &\leq \tilde{\gamma}_i(y_k^{\min}) - \phi(y_k^{\min}) = \ell_f(y_k^{\min}; \tilde{x}_i) - f(y_k^{\min}) + \frac{\tau_i}{2\lambda_i} \|y_k^{\min} - \tilde{x}_i\|^2 \\
&\leq \left(\frac{L_k}{2} + \frac{\tau_i}{2\lambda_i} \right) \|y_k^{\min} - \tilde{x}_i\|^2. \qquad (31)
\end{aligned}$$

From (29), for $i_{j-1}+1 \leq i \leq i_j$, using (30), (31), $\{\lambda_i\}$ is non-increasing in view of Lemma 4.2(a), $\phi(y_k^{\min}) \leq \phi(y_{i-1})$, Proposition 4.6 where $u = y_k^{\min}$, $0 \leq L_i\lambda_i + \tau_i \leq \xi_{i_j}\lambda_{i-1}$ and $0 \leq L_k\lambda_i + \tau_i \leq \xi_k\lambda_{i-1}$ in view of Lemma 4.2(b), $\tau_i = 2\xi_i\lambda_i/a_{i-1}$, $\xi_i = \xi_{i_j}$, $\lambda_{j-1} \leq \lambda_0$ and the last statement in Remark 4.3, we conclude that

$$\begin{aligned}
&\frac{1-\gamma}{2} A_i \|y_i - \tilde{x}_i\|^2 \\
&- \left(\lambda_{i-1} A_{i-1} (\phi(y_{i-1}) - \phi(y_k^{\min})) + \frac{1}{2} \|y_k^{\min} - x_{i-1}\|^2 \right) + \left(\lambda_i A_i (\phi(y_i) - \phi(y_k^{\min})) + \frac{1}{2} \|y_k^{\min} - x_i\|^2 \right) \\
&\leq \frac{1}{2} (L_i\lambda_i + \tau_i) A_{i-1} \|y_{i-1} - \tilde{x}_i\|^2 + \frac{1}{2} (L_k\lambda_i + \tau_i) a_{i-1} \|y_k^{\min} - \tilde{x}_i\|^2 - \frac{\tau_i a_{i-1}}{2} \|y_k^{\min} - x_i\|^2 \\
&\leq (L_i\lambda_i + \tau_i)(\|y_k^{\min} - x_{i-1}\|^2 + D_h^2) + (L_k\lambda_i + \tau_i) \left(\frac{1}{a_{i-1}} \|y_k^{\min} - x_{i-1}\|^2 + a_{i-1} D_h^2 \right) \\
&- \frac{\tau_i a_{i-1}}{2} \|y_k^{\min} - x_i\|^2 \\
&\leq \xi_{i_j} \lambda_{i-1} (\|y_k^{\min} - x_{i-1}\|^2 + D_h^2) + \xi_k \lambda_{i-1} \left(\frac{1}{a_{i-1}} \|y_k^{\min} - x_{i-1}\|^2 + a_{i-1} D_h^2 \right) - \xi_i \lambda_i \|y_k^{\min} - x_i\|^2 \\
&\leq \xi_{i_j} (\lambda_{i-1} \|y_k^{\min} - x_{i-1}\|^2 - \lambda_i \|y_k^{\min} - x_i\|^2) + \bar{\xi} \lambda_0 \left(\frac{1}{a_{i-1}} \|y_k^{\min} - x_{i-1}\|^2 + (1 + a_{i-1}) D_h^2 \right). \\
&\qquad (32)
\end{aligned}$$

Summing the inequality in (32) from $i = 1$ to k, we obtain

$$\frac{1-\gamma}{2} \left(\sum_{i=1}^{k} A_i \|y_i - \tilde{x}_i\|^2 \right) \leq \lambda_0 A_0 (\phi(y_0) - \phi(y_k^{\min})) + \frac{1}{2} \|y_k^{\min} - x_0\|^2$$

$$+\,\bar{\xi}\lambda_0 \sum_{i=1}^{k} \left(\frac{1}{a_{i-1}} \|y_k^{\min} - x_{i-1}\|^2 + (1+a_{i-1})D_h^2 \right) + \sum_{j=1}^{n_1} \xi_{i_j}\lambda_{i_j-1}\|y_k^{\min} - x_{i_j-1}\|^2. \tag{33}$$

Now, for $0 \le i \le k$, by Lemma 4.5,

$$\|y_k^{\min} - x_i\|^2 \le 2(\|y_k^{\min} - x_0\|^2 + \|x_0 - x_i\|^2) \le 2\|y_k^{\min} - x_0\|^2 + 2C^2 i^2.$$

Therefore, by the above and that $\xi_{i_j}\lambda_{i_{j-1}} \le \bar{\xi}\lambda_0$, we have from (33)

$$\frac{1-\gamma}{2}\left(\sum_{i=1}^{k} A_i \|y_i - \tilde{x}_i\|^2 \right) \le \lambda_0 A_0(\phi(y_0) - \phi(y_k^{\min})) + \left(\frac{1}{2} + 2\bar{\xi}\lambda_0 \left(n_1 + \sum_{i=1}^{k} \frac{1}{a_{i-1}} \right) \right) \|y_k^{\min} - x_0\|^2$$

$$+\,2\bar{\xi}\lambda_0 C^2 \left(\sum_{i=1}^{k} \frac{i^2}{a_{i-1}} + n_1 k^2 \right) + \bar{\xi}\lambda_0 D_h^2 \sum_{i=1}^{k} (1 + a_{i-1}).$$

The conclusion of the lemma then follows by noting that $n_1 \le n_0$, $a_{i-1} \ge 1$ and the definition of D_h in (22). ■

We are now ready to state the iteration complexity results of VAR-FISTA to solve Problem (1).

Theorem 4.10 *VAR-FISTA terminates to obtain an $\hat{\rho}$-approximate solution $(\hat{y}, \hat{v})$ to Problem (1) in at most*

$$\left(\frac{3C_1 L_1}{\hat{\rho}^2} \right)^{1/3} + \left(\frac{3C_1\bar{\xi}\lambda_0(2C^2 + 3D_h^2)}{\hat{\rho}^2} \right)^{1/2} + \frac{C_1\bar{\xi}\lambda_0(6C^2(1+n_0) + D_h^2)}{\hat{\rho}^2} + 1 \tag{34}$$

iterations, where

$$C_1 = \left(\frac{8}{1-\gamma} \right) \left(\frac{1}{\underline{\lambda}} + \frac{1}{2}\bar{\xi} + \overline{M} \right)^2,$$
$$L_1 = \lambda_0 A_0(\phi(y_0) - \phi(y^*)) + \left(\frac{1}{2} + 2\bar{\xi}\lambda_0 n_0 \right) D_h^2,$$

and recall that $C = 2(2 + \bar{\xi}\lambda_0)D_h$ and $n_0 = \mathcal{O}(\max\{\log \underline{m}, 1\})$. Furthermore, if f in Problem (1) is convex, then the iteration complexity of VAR-FISTA to solve the problem is improved, and it finds an $\hat{\rho}$-approximate solution $(\hat{y}, \hat{v})$ to Problem (1) in at most

$$\left(\frac{3C_2 L_2}{\hat{\rho}^2} \right)^{1/3} + 1 \tag{35}$$

iterations, where

$$C_2 = \left(\frac{8}{1-\gamma} \right) \left(\frac{1}{\underline{\lambda}} + \overline{M} \right)^2,$$
$$L_2 = \lambda_0 A_0(\phi(y_0) - \phi(y^*)) + \frac{1}{2} D_h^2.$$

Proof: Using the facts that $\{a_k\}$ is increasing, $a_0 = 4$, Lemma 4.2(b), (c), and the last statement in Remark 4.3, we have for $k \ge 1$,

$$\frac{1+\tau_k}{\lambda_k} = \frac{1}{\lambda_k} + \frac{2\xi_k}{a_{k-1}} \le \frac{1}{\underline{\lambda}} + \frac{2\bar{\xi}}{a_0} = \frac{1}{\underline{\lambda}} + \frac{1}{2}\bar{\xi},$$

and hence, together with (2), by (13), we obtain

$$\min_{1\le i\le k}\|v_i\| \le \min_{1\le i\le k}\left(\frac{1+\tau_i}{\lambda_i}+\overline{M}\right)\|y_i-\tilde{x}_i\| \le \tilde{C}\min_{1\le i\le k}\|y_i-\tilde{x}_i\|,$$

where

$$\tilde{C}=\frac{1}{\underline{\lambda}}+\frac{1}{2}\bar{\xi}+\overline{M}.$$

Using the above inequality, (28) in Lemma 4.9, definition of y^* and the first inequality in Lemma 4.1, we obtain for $k\ge 1$,

$$\begin{aligned}
&\left(\frac{1-\gamma}{2}\sum_{i=1}^{k}A_i\right)\min_{1\le i\le k}\|v_i\|^2\\
&\le \tilde{C}^2\left[\lambda_0A_0(\phi(y_0)-\phi(y_k^{\min}))+\left(\frac{1}{2}+2\bar{\xi}\lambda_0n_0\right)D_h^2+2\bar{\xi}\lambda_0C^2\left(\sum_{i=1}^{k}\frac{i^2}{a_{i-1}}+n_0k^2\right)\right.\\
&\left.+\bar{\xi}\lambda_0D_h^2\sum_{i=1}^{k}(3+a_{i-1})\right]\\
&\le \tilde{C}^2\left[\lambda_0A_0(\phi(y_0)-\phi(y^*))+\left(\frac{1}{2}+2\bar{\xi}\lambda_0n_0\right)D_h^2+2\bar{\xi}\lambda_0C^2\left(\sum_{i=1}^{k}2i+n_0k^2\right)\right.\\
&\left.+\bar{\xi}\lambda_0D_h^2\left(3k+\sum_{i=1}^{k}a_{i-1}\right)\right].
\end{aligned}$$

Hence the complexity result (34) follows from the above inequality, the third and fourth inequality in Lemma 4.1. The result (35) follows from (34) and $\bar{\xi}=0$ by (23) where $\underline{m}=0$ since f is convex. ■

5 Conclusion

In this note, we propose a first order algorithm, VAR-FISTA, to solve composite optimization problems, and establish iteration complexity result for the convex and nonconvex case in Theorem 4.10 that are best known in the literature so far. We remark that even though the iteration complexity for the convex case is better than that for the nonconvex case as shown in Theorem 4.10, implementation[1] of the algorithm shows that the number of iterations to obtain an $\hat{\rho}$-approximate solution for instances of the quadratic programming problem as found in [10] is worse when the instance is convex than when it is nonconvex. This phenomenon appears to occur as well in [10] for other first order methods tested in the paper. We do not have a reasonable explanation for this unusual phenomenon.

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.

[1] We do not provide numerical results that we obtained in this note.

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

[4] S. Ghadimi, G. Lan, and H. Zhang. Generalized uniformly optimal methods for nonlinear programming. *Journal of Scientific Computing*, 79:1854–1881, 2019.

[5] 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 programs. *SIAM Journal on Optimization*, 29(4):2566–2593, 2019.

[6] 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.

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

[8] 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.

[9] J. Liang and R. D. C. Monteiro. An average curvature accelerated composite gradient method for nonconvex smooth composite optimization problems. *Available on arXiv:1909.04248*, 2019.

[10] J. Liang, R. D. C. Monteiro, and C.-K. Sim. A FISTA-type accelerated gradient algorithm for solving smooth nonconvex composite optimization problems. *Available on arXiv:1905.07010v2*, 2019.

[11] 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.

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

[13] Y. E. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103:127–152, 2005.

[14] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. `http://www.mit.edu/~dimitrib/PTseng/papers.html`, 2008.

[15] Q. Yao, J. T. Kwok, F. Gao, W. Chen, and T.-Y. Liu. Efficient inexact proximal gradient algorithm for nonconvex problems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3308–3314. IJCAI, 2017.

A Appendix

Proof of Lemma 4.1: For $k \geq 1$, observe that

$$\frac{1}{2} + \sqrt{A_{k-1}} \leq a_{k-1} = \frac{1 + \sqrt{1 + 4A_{k-1}}}{2} \leq 2\sqrt{A_{k-1}}.$$

It follows that

$$\left(\sqrt{A_{k-1}}+\frac{1}{2}\right)^2 \le A_{k-1}+\sqrt{A_{k-1}}+\frac{1}{2} \le A_k = A_{k-1}+a_{k-1}$$
$$\le A_{k-1}+2\sqrt{A_{k-1}} \le (\sqrt{A_{k-1}}+1)^2.$$

Hence,

$$\sqrt{A_0}+\frac{k}{2} \le \sqrt{A_{k-1}}+\frac{1}{2} \le \sqrt{A_k} \le \sqrt{A_{k-1}}+1 \le \sqrt{A_0}+k.$$

Since $A_k = a_{k-1}^2$ and $A_0 = 12$, we conclude from the above that

$$\frac{k}{2} \le a_{k-1} \le 4k. \tag{36}$$

Now, by $A_i = a_{i-1}^2$ and (36), we have

$$\sum_{i=1}^k A_i = \sum_{i=1}^k a_{i-1}^2 \ge \frac{1}{4}\sum_{i=1}^k i^2 = \frac{1}{24}k(k+1)(2k+1) \ge \frac{k^3}{12}.$$

From $A_i = a_{i-1}^2$, $A_i = A_{i-1}+a_{i-1}$ and (36), we have

$$\frac{\sum_{i=1}^k a_{i-1}}{\sum_{i=1}^k A_i} = \frac{\sum_{i=1}^k a_{i-1}}{\sum_{i=1}^k a_{i-1}^2} \le \frac{k\sum_{i=1}^k a_{i-1}}{\left(\sum_{i=1}^k a_{i-1}\right)^2} = \frac{k}{\sum_{i=1}^k a_{i-1}} = \frac{k}{A_k - A_0} \le \frac{k}{a_{k-1}^2} \le \frac{4}{k}.$$

■

Proof of Proposition 4.7: It is clear from the definition of $\tilde{\gamma}_k$ and γ_k that they are (τ_k/λ_k)-strongly convex. By (8), the way y_k is defined in step k3 of VAR-FISTA and the definition of $\tilde{\gamma}_k$ in (19), we see that y_k is the optimal solution to the first minimization problem in (26). Since the objective function of this minimization problem is $((1+\tau_k)/\lambda_k)$-strongly convex, it follows that $\forall\ u \in \Re^n$,

$$\tilde{\gamma}_k(y_k)+\frac{1}{2\lambda_k}\|y_k-\tilde{x}_k\|^2+\frac{1+\tau_k}{2\lambda_k}\|y_k-u\|^2 \le \tilde{\gamma}_k(u)+\frac{1}{2\lambda_k}\|u-\tilde{x}_k\|^2. \tag{37}$$

On the other hand, the definition of γ_k in (20) and the relation

$$\|y_k-\tilde{x}_k\|^2+\|y_k-u\|^2 = 2\langle\tilde{x}_k-y_k, u-y_k\rangle+\|u-\tilde{x}_k\|^2$$

imply that

$$\tilde{\gamma}_k(y_k)+\frac{1}{2\lambda_k}\|y_k-\tilde{x}_k\|^2+\frac{1+\tau_k}{2\lambda_k}\|y_k-u\|^2 = \gamma_k(u)+\frac{1}{2\lambda_k}\|u-\tilde{x}_k\|^2. \tag{38}$$

Hence, comparing (37) with (38), we have $\gamma_k(u) \le \tilde{\gamma}_k(u)\ \forall\ u \in \text{dom}\ h$, and from (38), we have $\tilde{\gamma}_k(y_k) = \gamma_k(y_k)$. Furthermore, $\gamma_k(y_k) = \tilde{\gamma}_k(y_k)$, (38) and $\gamma_k \le \tilde{\gamma}_k$ imply that

$$\begin{aligned}\gamma_k(y_k)+\frac{1}{2\lambda_k}\|y_k-\tilde{x}_k\|^2 &= \tilde{\gamma}_k(y_k)+\frac{1}{2\lambda_k}\|y_k-\tilde{x}_k\|^2\\ &\le \tilde{\gamma}_k(y_k)+\frac{1}{2\lambda_k}\|y_k-\tilde{x}_k\|^2+\frac{1+\tau_k}{2\lambda_k}\|y_k-u\|^2\end{aligned}$$

$$= \quad \gamma_k(u) + \frac{1}{2\lambda_k}\|u - \tilde{x}_k\|^2 \leq \tilde{\gamma}_k(u) + \frac{1}{2\lambda_k}\|u - \tilde{x}_k\|^2$$

for all $u \in \Re^n$, and hence the remaining conclusions of (a) follow. ■

Proof of Lemma 4.8: By Lemma 4.2(b), (9), the definition of $\tilde{\gamma}_k$ in (19) and Proposition 4.7, we have

$$\begin{aligned}
\lambda_k\phi(y_k) + \frac{1-\gamma}{2}\|y_k - \tilde{x}_k\|^2 &\leq \lambda_k\phi(y_k) + \frac{1 - U_k\lambda_k}{2}\|y_k - \tilde{x}_k\|^2 \\
&= \lambda_k\tilde{\gamma}_k(y_k) + \frac{1}{2}(1 - \tau_k)\|y_k - \tilde{x}_k\|^2 \\
&\leq \lambda_k\gamma_k(y_k) + \frac{1}{2}\|y_k - \tilde{x}_k\|^2. \qquad (39)
\end{aligned}$$

Since y_k is the optimal solution to the second minimization problem in (26), by convexity of γ_k, (7) and (18), the following holds for every $u \in \Omega$:

$$\begin{aligned}
& A_k\left(\lambda_k\gamma_k(y_k) + \frac{1}{2}\|y_k - \tilde{x}_k\|^2\right) \\
\leq\ & A_k\left(\lambda_k\gamma_k\left(\frac{A_{k-1}y_{k-1} + a_{k-1}x_k}{A_k}\right) + \frac{1}{2}\left\|\frac{A_{k-1}y_{k-1} + a_{k-1}x_k}{A_k} - \tilde{x}_k\right\|^2\right) \\
\leq\ & \lambda_k A_{k-1}\gamma_k(y_{k-1}) + \lambda_k a_{k-1}\gamma_k(x_k) + \frac{A_k}{2}\left\|\frac{A_{k-1}y_{k-1} + a_{k-1}x_k}{A_k} - \tilde{x}_k\right\|^2 \\
=\ & \lambda_k A_{k-1}\gamma_k(y_{k-1}) + \lambda_k a_{k-1}\gamma_k(x_k) + \frac{1}{2}\|x_k - x_{k-1}\|^2 \\
\leq\ & \lambda_k A_{k-1}\gamma_k(y_{k-1}) + \lambda_k a_{k-1}\gamma_k(u) + \frac{1}{2}\|u - x_{k-1}\|^2 - \frac{a_{k-1}\tau_k + 1}{2}\|u - x_k\|^2, \qquad (40)
\end{aligned}$$

where the last inequality holds since x_k is the optimal solution to the minimization problem (21), and its objective function is $((a_{k-1}\tau_k + 1)/\lambda_k)$-strongly convex. The result now follows by combining (39) and (40). ■