Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 19, No. 3, pp. 529–549, 2023

A CLASS OF FAST ITERATIVE SHRINKAGE THRESHOLDING ALGORITHM FOR NONSMOOTH OPTIMIZATION*

Hanlin Zhou, Wanyou Cheng, Jianhao Ye and Jiahao Zhang

Abstract: In this paper, we propose a class of fast iterative shrinkage thresholding algorithm which contains the original FISTA scheme, the FISTA-CD scheme and the FISTA-Mod. We prove that the objective value sequence achieves $O(\frac{1}{k^2})$ complexity. Moreover, we prove that any limit point of the sequence generated by the proposed algorithm is a minimizer of the objective function. At last, we propose a modified adaptive restart strategy which can dramatically improve the convergence rate of the proposed algorithms. Numerical results demonstrate that the proposed algorithm is competitive with several known methods.

Key words: *nonsmooth nonconvex penalty, sparse optimization, subspace optimization*

Mathematics Subject Classification: *90C06, 90C25, 65Y20, 94A08*

1 Introduction

In this paper, we focus on the following structured non-smooth optimization problem

$$\min_{x\in\mathcal{R}^n} \phi(x) := f(x) + r(x), \tag{1.1}$$

where f is convex and differentiable with gradient $\nabla f(x)$ being L_f-Lipschitz continuous and $r(x)$ is proper, convex and lower-semi-continuous. We assume that the set of minimizers of problem (1.1) is non-empty, i.e. $\operatorname{Argmin}\phi \neq \emptyset$. The problem (1.1) often appears in many applications, such as in compressive sensing, machine learning, high-dimensional variable selection, data fitting and image reconstruction [4, 6, 8].

A typical optimization strategy for solving large-scale problem (1.1) is a so-called proximal gradient algorithm (PG). The proximal gradient algorithm has the property that the objective value sequence achieves $O(\frac{1}{k})$ complexity and can be slow in practice in its original form [14]. Acceleration techniques for PG has firstly been considered by Nesterov [13] for projected gradient descent. Beck and Teboulle [2] proposed the fast iterative shrinkage/thresholding algorithm (FISTA) and proved that the objective value sequence of FISTA

*Supported by the Chinese NSF Grant (nos. 12271187, 11961011 and 11971106), by special projects in key fields of Guangdong Universities(no.2021ZDZX1054) and by the Natural Science Foundation of Guangdong province (no.2022A1515010567).

achieves $O(\frac{1}{k^2})$ complexity. FISTA performs an extrapolation technique which has the following form

$$\left\{ \begin{array}{l} y^k = x^k + \beta^k(x^k - x^{k-1}), \\ x^{k+1} = \arg\min_{x \in \mathcal{R}^n} \{\nabla f(y^k)^T x + \frac{1}{2\mu}\|x - y^k\|^2 + r(x)\}, \end{array} \right. \tag{1.2}$$

where $\beta^k = \frac{t^k - 1}{t^{k+1}}$, $t^{k+1} = \frac{1+\sqrt{1+4(t^k)^2}}{2}$ and $t^1 = 1$. Besides achieving optimal convergence rate $O(\frac{1}{k^2})$ for objective function value, Chambolle and Dossal [7] established the convergence of the whole convergence of (1.2) with the extrapolation sequence

$$\beta^k = \frac{T^k - 1}{T^{k+1}}, \;\; T^k = \frac{k + a - 1}{a}, \tag{1.3}$$

where $a > 2$. Tao, Boley and Zhang [16] established local linear convergence of ISTA and FISTA and showed that FISTA's convergence rate can slow down as it proceeds, eventually becoming slower than ISTA when they are applied to solve the LASSO problem. Johnstone and Moulin [9] established the convergence of the whole convergence generated by (1.2) with extrapolation sequence $0 \leq \beta^k \leq \bar{\beta}$ for some constant $0 \leq \bar{\beta} < 1$. They also showed that the proposed algorithm is locally linearly convergent for the LASSO problem. O'Donoghue and Candsè [14] proposed an adaptive restart scheme for β^k based on FISTA for solving problem (1.1) with $r(x) = 0$. The basic idea of restarting is that once the objective function value of $\phi(x^k)$ is about to increase, the algorithm resets the extrapolation parameter β^k. Ochs and Pock [15] proposed an adaptively FISTA algorithm and established that it is equivalent to a proximal variant of the SR1 quasi-Newton method. They also proved that every limit point of the sequence generated by the algorithm is a stationary point of problem (1.1) where f is differentiable with gradient $\nabla f(x)$ being L_f-Lipschitz continuous and $r(x)$ is lower-semicontinuous. However, the adaptively FISTA can not guarantee an optimal convergence rate of $O(\frac{1}{k^2})$ as for FISTA. Under the error bound condition [12], Wen et al.[17] showed that if the extrapolation coefficients $\{\beta^k\}$ are chosen below a given threshold, then the sequence generated converges R-linearly to a stationary point of problem (1.1) without convexity of f. Moreover, the corresponding sequence of objective values is also R-linearly convergent. Bareilles and Iutzeler [1] proposed two modified proximal gradient methods exhibiting a stable identification behavior while maintaining the convergence rate of FISTA both in theory and in practice. In their algorithm, efficient practical tests to determine whether or not to perform the FISTA iteration. We refer to papers [1, 5, 10, 11] for more advances in this area.

Recently, Liang et al. [11] proposed a modified FISTA scheme (1.2) with the extrapolation sequence

$$\beta^k = \frac{t^k - 1}{t^{k+1}}, \;\; t^{k+1} = \frac{p + \sqrt{q + 4(t^k)^2}}{2}, \tag{1.4}$$

where $p, q \in [0, 1]$ and $t^1 = 1$. The modified FISTA scheme has the property that the objective value sequence achieves $O(\frac{1}{k^2})$ complexity. Numerical results in [11] and [7] showed that different parameters seriously affect practical performance of their algorithm. To improve the efficiency of the modified FISTA scheme of [11] and the algorithm of [7] further, by the use of the convex combination of T^k of (1.3) and t^k of (1.4), we propose a class of fast iterative shrinkage thresholding algorithm to solve problem (1.1). The proposed algorithm includes FISTA, the algorithms of [7] and [11] as special cases. The main contributions of our paper are summarized as follows

(1) The sequence $\{\|x^k - x^*\|\}$ is convergent for any $x^* \in \operatorname{Arg}\min\phi$ and any limit point of x^k is a solution of problem (1.1).

(2) The objective value sequence of the proposed algorithm possesses $O(\frac{1}{k^2})$ complexity.

(3) We give the possibility reason why algorithm of Chambolle and Dossal's [7] with relatively large values of a performs better than FISTA and Liang's algorithm [11] do.

(4) We proposed a modified adaptive restart strategy which can be up to an order faster than the original scheme.

The rest of the paper is organized as follows. Some notations and preliminary results are given in Section 2. The proposed algorithm and its convergence results are presented in Section 3. In Section 4, we propose a modified adaptive restart strategy. We report some numerical results in Section 5 and make some conclusions in the last section.

Throughout the paper, $\|\cdot\|$ denotes the Euclidean norm of vectors. Let x^* be a global minimizer of the problem (1.1). Let x_T denote the collection of columns and entries of x, whose indices are in an index set $T \subseteq \{1,2,3,...,n\}$, respectively. Denote the subdifferential of $r(x)$ at x by $\partial r(x)$.

2 Preliminaries

In this section, we present some notations, lemmas and theorems which will be used in the rest of the paper.

Lemma 2.1 (Lemma 2.1 of [2])**.** *Suppose that f is continuously differentiable function with Lipschitz continuous gradient and Lipschitz constant L_f. Then, for any $\gamma \in (0, \frac{1}{L_f}]$,*

$$f(x) \leq f(y) + \nabla f(y)^T(x-y) + \frac{1}{2\gamma}\|x-y\|^2 \tag{2.1}$$

for every $x, y \in \mathcal{R}^n$.

For any $\gamma > 0$, consider the following quadratic approximation of $\phi(x)$ at a given y:

$$Q_\gamma(x,y) := f(y) + \nabla f(y)^T(x-y) + \frac{1}{2\gamma}\|x-y\|^2 + r(x). \tag{2.2}$$

It is obvious that $Q_\gamma(x,y)$ is strongly convex with respect to x, hence it admits a unique minimizer

$$p_\gamma(y) := \arg\min_{x\in\mathcal{R}^n} Q_\gamma(x,y). \tag{2.3}$$

Lemma 2.2 (Lemma 2.3 of [2])**.** *Let $y \in \mathcal{R}^n$ and $\gamma \in (0, \frac{1}{L_f}]$ such that*

$$\phi(p_\gamma(y)) \leq Q_\gamma(p_\gamma(y), y). \tag{2.4}$$

Then for any $x \in \mathcal{R}^n$, we have

$$\phi(x) - \phi(p_\gamma(y)) \geq \frac{1}{2\gamma}\|p_\gamma(y) - y\|^2 + \frac{1}{\gamma}(p_\gamma(y) - y)^T(y - x). \tag{2.5}$$

Lemma 2.3 (Lemma 4.2 of [2])**.** *Let $\{a^k\}$ and $\{b^k\}$ be positive sequences of reals satisfying*

$$a^k - a^{k+1} \geq b^{k+1} - b^k, \quad \forall k \geq 1, \quad \text{with } a^1 + b^1 \leq c, \quad c > 0. \tag{2.6}$$

Then, $a^k \leq c$ for every $k \geq 1$.

Suppose $p \in (0, 1]$ and $q > 0$. Given a positive integer l, define the sum $S_l = \frac{q}{4p} \sum_{i=0}^{l} \frac{1}{1+i}$ and a new sequence $\bar{t}^k$ by

$$\bar{t}^k = 1 + S_l + (\frac{p}{2} + \frac{q}{4p(l+1)})k. \tag{2.7}$$

Denote $\lceil x \rceil$ the smallest integer that is larger than x, and define the following two constants

$$b = \lceil \frac{p+2}{p + \frac{q}{2p(l+1)}} \rceil, \;\; c = \lceil \frac{p+2+2S_l}{p + \frac{q}{2p(l+1)}} \rceil. \tag{2.8}$$

Lemma 2.4 (Lemma 3.6 of [11]). *For the t^k update rule (1.4) with $q > 0$ and $0 < p \leq 1$. Let $t^0 = 1$, then for all positive integer k, it holds that*

$$t^k \geq \frac{(k+1)p}{2}. \tag{2.9}$$

Lemma 2.5 (Lemma 3.11 of [11]). *For the t^k update rule (1.4) with $p, q \in (0, 1]$. it holds that*

$$t^k \leq \bar{t}^k \tag{2.10}$$

for all positive integer k.

Lemma 2.6 (Lemma 3.12 of [11]). *For all $j \geq 1$, define $\beta^{j,k} = \prod_{i=j}^{k} a_i = \prod_{i=j}^{k} \frac{t_{i-1}-1}{t_i}$ for all j, k, and $\beta^{j,k} = 1$ for all $k < j$. Let $l \geq \lceil \frac{q}{p(2-p)} \rceil$, then for all j, it holds that*

$$\sum_{k=j}^{\infty} \beta^{j,k} \leq j + c + 2b. \tag{2.11}$$

Lemma 2.7 (Lemma 4.1 of [7]). *For all $j \geq 1$, let us define $\beta^{j,k} = \prod_{l=j}^{k} \alpha^l = \prod_{i=j}^{k} \frac{l-1}{l+a}$ for all $k \geq j$, and $\beta^{j,k} = 1$ for all $k < j$. Then, we have for all j*

$$\sum_{k=j}^{\infty} \beta^{j,k} \leq \frac{j+5}{2}. \tag{2.12}$$

Lemma 2.8 (Lemma 3.1 of [7]). *Let $\gamma \in (0, \frac{1}{L_f}]$, where L_f is the Lipschitz constant of ∇f, and $\bar{x} = P_\gamma(y)$. Then*

$$\phi(\bar{x}) + \frac{\|\bar{x} - x\|^2}{2\gamma} \leq \phi(x) + \frac{\|x - y\|^2}{2\gamma}, \;\; \forall x \in \mathcal{R}^n. \tag{2.13}$$

Theorem 2.9 (Theorem 3.2 of [7]). *If the sequence $\{t^n\}$ satisfies $\rho^n = (t^{n-1})^2 - (t^n)^2 + t^n \geq 0$ and $t^1 = 1$, if $0 < \gamma \leq \frac{1}{L_f}$, then for any $N \geq 2$*

$$(t^{N+1})^2 w^{N+1} + \sum_{n=1}^{N} \rho^{n+1} w^n \leq \frac{\nu^0 - \nu^{N+1}}{\gamma} \tag{2.14}$$

where $w^n = \phi(x^n) - \phi(x^*)$, $\nu^n = \frac{\|u^n - x^*\|^2}{2}$ and $u^n = x^{n-1} + t^n(x^n - x^{n-1})$.

$\boxed{3}$ Algorithm and Its Convergence

As we all known, the modified FISTA scheme [11] and the algorithm [7] possess properties (1) and (2). Moreover, numerical results show that the modified FISTA scheme [11] and the algorithm [7] perform better than FISTA in practice. To improve the efficiency of the modified FISTA scheme [11] and the algorithm [7] further, we propose a class of fast iterative shrinkage thresholding algorithm which utilize the convex combination of T^k of (1.3) and t^k of (1.4). Specifically, we set

$$\beta^k = \frac{w^k - 1}{w^{k+1}}, \quad w^k = \beta T^k + (1-\beta)t^k, \;\; \forall \beta \in [0,1] \tag{3.1}$$

in (1.2). The new algorithm is described below.

Algorithm 3.1. (A class of FISTA schme)

Step 0. Given an initial point $x^0 = x^1 \in \mathcal{R}^n$, constants $p \in (0,1]$, $0 < q \leq (2-p)^2$, $t^1 = 1$, $a \geq 2$, $\beta \in [0,1]$ and $0 < \gamma \leq \frac{1}{L_f}$. Set $k := 1$.

Step 1. If the stopping condition is satisfied, then stop. Otherwise, go to Step 2.

Step 2. Compute β^k by (3.1).

Step 3. Update $y^k = x^k + \beta^k(x^k - x^{k-1})$ and $x^{k+1} = P_\gamma(y^k)$.

Step 4. Set $k := k+1$ and go to Step 1.

From Algorithm 3.1, it is easy to see that Algorithm 3.1 with $\beta = 0$ reduces to the modified FISTA [11], while Algorithm 3.1 with $\beta = 1$ reduces to the algorithm [7].

$\boxed{3.1}$ Global convergence of function sequence

In this section, we shall establish that the new algorithm achieves the worst-case $O(\frac{1}{k^2})$ optimal convergence rate in terms of objective function value. The proof of the following theorem heavily depends on the proof of Lemma 4.1 and Theorem 4.4 [2]. For completeness, we presented the details of the proofs. We first give a useful lemma.

Lemma 3.2. *Suppose that w^k is defined by (3.1) with $T^0 = 1$ and $t^0 = 1$.*

(i) *If $a \geq 2$, $0 < p \leq 1$ and $0 < q \leq (2-p)^2$, then $w^k \geq (\frac{\beta}{a} + \frac{1-\beta}{2}p)(k+1)$.*

(ii) *If $0 < q \leq (2-p)^2$, $p \in (0,1]$ and $p + \sqrt{q} \leq 1$ and $a \geq 2$, then we have*

$$(w^{k+1})^2 - w^{k+1} - (w^k)^2 \leq \beta(1-\beta)((\frac{k+a}{a})(p+\sqrt{q}-1) + (\frac{2}{a}-1)t^k) \leq 0. \tag{3.2}$$

Proof. By Lemma 2.4, we have

$$w^k = \beta T^k + (1-\beta)t^k \geq (\frac{\beta}{a} + \frac{1-\beta}{2}p)(k+1) + \frac{\beta(a-2)}{a} \geq (\frac{\beta}{a} + \frac{1-\beta}{2}p)(k+1). \tag{3.3}$$

Now, we begin to prove (ii). Since $a \geq 2$, we have

$$\begin{aligned}(T^{k+1})^2 - T^{k+1} - (T^k)^2 &= (T^{k+1}+T^k)(T^{k+1}-T^k) - T^{k+1} \\ &= \frac{2k+2a+1}{a^2} - \frac{k+a+1}{a} \\ &= \frac{k(2-a) - ((a-\frac{1}{2})^2 - \frac{5}{4})}{a^2} \\ &\leq 0. \end{aligned} \tag{3.4}$$

On the other hand, by (1.4) and $0 < q \leq (2-p)^2$, we have

$$\begin{aligned}(t^{k+1})^2 - pt^{k+1} + \frac{p^2-q}{4} = (t^k)^2 &\iff (t^{k+1})^2 - t^{k+1} + (1-p)t^{k+1} + \frac{p^2-q}{4} = (t^k)^2 \\ &\implies (t^{k+1})^2 - t^{k+1} + (1-p)t^0 + \frac{p^2-q}{4} \leq (t^k)^2 \\ &\implies (t^{k+1})^2 - t^{k+1} + \frac{(2-p)^2-q}{4} \leq (t^k)^2 \\ &\implies (t^{k+1})^2 - t^{k+1} \leq (t^k)^2. \end{aligned} \tag{3.5}$$

By the definition of w^k, we have

$$\begin{aligned} & (w^{k+1})^2 - w^{k+1} - (w^k)^2 \\ = \; & \beta^2(T^{k+1})^2 + (1-\beta)^2(t^{k+1})^2 + 2\beta(1-\beta)T^{k+1}t^{k+1} \\ & -\beta^2(T^k)^2 - (1-\beta)^2(t^k)^2 - 2\beta(1-\beta)T^kt^k - \beta T^{k+1} - (1-\beta)t^{k+1} \\ \leq \; & \beta^2T^{k+1} + (1-\beta)^2t^{k+1} + 2\beta(1-\beta)T^{k+1}t^{k+1} - 2\beta(1-\beta)T^kt^k - \beta T^{k+1} - (1-\beta)t^{k+1} \\ = \; & \beta(1-\beta)(2T^{k+1}t^{k+1} - 2T^kt^k - T^{k+1} - t^{k+1}) \\ \leq \; & \beta(1-\beta)(2(T^k + \frac{1}{a})(t^k + \frac{p+\sqrt{q}}{2}) - T^k - \frac{1}{a} - t^{k+1} - 2T^kt^k) \\ = \; & \beta(1-\beta)(2(T^kt^k + T^k\frac{p+\sqrt{q}}{2} + \frac{t^k}{a} + \frac{p+\sqrt{q}}{2a}) - T^k - \frac{1}{a} - t^{k+1} - 2T^kt^k) \\ \leq \; & \beta(1-\beta)(T^k(p+\sqrt{q}) + \frac{2t^k}{a} + \frac{p+\sqrt{q}}{a} - T^k - \frac{1}{a} - t^k) \\ = \; & \beta(1-\beta)((T^k + \frac{1}{a})(p+\sqrt{q}-1) + (\frac{2}{a}-1)t^k), \end{aligned}$$

where the first inequality uses (3.4) and (3.5), the second inequality uses $t^{k+1} > t^k + \frac{p+\sqrt{q}}{2}$ and the last inequality uses the monotonicity of t^k. □

Remark 3.3. As pointed out by [11], the practical performance of the modified FISTA [11] depends on the speed of β^k approaching 1. From Figure 1, we see that β^k approaches 1 much slower for the choice of $p = \frac{1}{20}, q = 0.1$. This fact also shows that the condition $p + \sqrt{q} \leq 1$ may be a good choice for controlling the speed of β^k approaching 1.

The following theorem shows that the proposed algorithm achieves the worst-case $O(\frac{1}{k^2})$ optimal convergence rate in terms of objective function value.

Theorem 3.4. *Suppose that $p \in (0,1]$, $0 < q \leq (2-p)^2$, $p + \sqrt{q} \leq 1$ and $a \geq 2$ for the proposed algorithm. Then*

$$\phi(x^k) - \phi(x^*) \leq \frac{1}{2\gamma c^2(k+1)^2}\|x^0 - x^*\|^2, \tag{3.6}$$

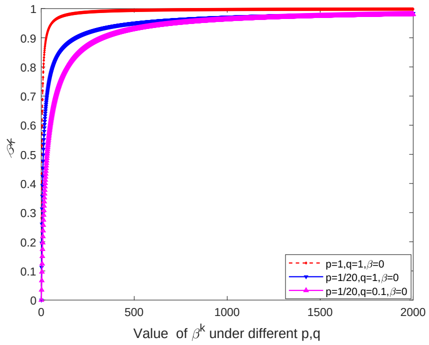

Figure 1: p, q control the speed of β^k approaching its limits in Algorithm 3.1

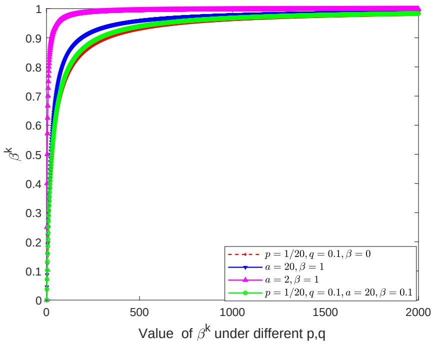

Figure 2: p, q control the speed of β^k approaching its limits in Algorithm 3.1

where $c = \frac{\beta}{a} + \frac{1-\beta}{2}p$ *and* $\gamma \in (0, \frac{1}{L_f}]$ *and* $x^* \in \operatorname{Arg\,min} \phi$.

Proof. Define $v^k = \phi(x^k) - \phi(x^*)$ and $u^k = w^k x^k - (w^k - 1)x^{k-1} - x^*$. Then we apply Lemma 2.2 at the points $(x = x^k, y = y^k)$ and at $(x = x^*, y = y^k)$ to get

$$\begin{aligned} 2\gamma(v^k - v^{k+1}) &\geq \|x^{k+1} - y^k\|^2 + 2(x^{k+1} - y^k)^T(y^k - x^k) \\ -2\gamma v^{k+1} &\geq \|x^{k+1} - y^k\|^2 + 2(x^{k+1} - y^k)^T(y^k - x^*). \end{aligned}$$

Multiplying the first inequality by $w^{k+1} - 1$ and then adding to the second inequality, we get

$$2\gamma((w^{k+1}-1)v^k - w^{k+1}v^{k+1}) \geq w^{k+1}\|x^{k+1} - y^k\|^2 + 2(x^{k+1} - y^k)^T(w^{k+1}y^k - (w^{k+1}-1)x^k - x^*). \tag{3.7}$$

Multiplying above inequality by w^{k+1} and using (i) of Lemma 3.2, we get

$$\begin{aligned} 2\gamma((w^k)^2 v^k - (w^{k+1})^2 v^{k+1}) \geq{}& \|w^{k+1}(x^{k+1} - y^k)\|^2 \\ &+ 2w^{k+1}(x^{k+1} - y^k)^T(w^{k+1}y^k - (w^{k+1} - 1)x^k - x^*). \end{aligned} \tag{3.8}$$

By the Pythagoras relation $\|b - a\|^2 + 2(b - a)^T(a - c) = \|b - c\|^2 - \|a - c\|^2$, we get

$$\begin{aligned} 2\gamma((w^k)^2 v^k - (w^{k+1})^2 v^{k+1}) \geq{}& \|w^{k+1}x^{k+1} - (w^{k+1} - 1)x^k - x^*\|^2 \\ &- \|w^{k+1}y^k - (w^{k+1} - 1)x^k - x^*\|^2. \end{aligned} \tag{3.9}$$

Since $y^k = x^k + \frac{w^k - 1}{w^{k+1}}(x^k - x^{k-1})$, we get

$$2\gamma(w^k)^2 v^k - 2\gamma(w^{k+1})^2 v^{k+1} \geq \|u^{k+1}\|^2 - \|u^k\|^2. \tag{3.10}$$

Define $a^k = 2\gamma(w^k)^2 v^k$, $b^k = \|u^k\|^2$ and $c = \|y^1 - x^*\|^2 = \|x^0 - x^*\|^2$. Proceeding the similar analysis as that of Theorem 4.4 of [2], we get $a^1 + b^1 \leq c$. By Lemma 2.3, we get

$$2\gamma(w^k)^2 v^k \leq \|x^0 - x^*\|^2. \tag{3.11}$$

Furthermore, by Lemma 3.2, we get the conclusion. □

3.2 Convergence analysis of $\{x^k\}$

In this section, we shall show that $\{\|x^k - x^*\|\}$ is convergent and any limit point of $\{x^k\}$ is a solution of problem (1.1). The proof of the following theorem heavily depends on the proof of Lemma 4.1 and Theorem 4.1 of [7] and Lemma 3.12 and Theorem 3.5 of [11]. Before presenting the result, we first present some supporting lemmas. By the definition of w^k and Lemma 2.5, we get the following the lemma.

Lemma 3.5. *Let $p \in (0,1]$ and $0 < q \le (2-p)^2$. Then we have*

$$w^k \le (1-\beta)\bar{t}^k + \beta T^k. \tag{3.12}$$

Lemma 3.6. *Let $p \in (0,1]$ and $0 < q \le (2-p)^2$. If $2 < ap < 2a-2$, then there exists $k^0 = \frac{2a-2-ap}{ap-2}$ such that for all $k > k^0$*

$$t^k > T^k. \tag{3.13}$$

Proof. By (2.9) and $2 < ap < 2a-2$, we have

$$\begin{aligned} t^k - T^k &\ge \frac{k+1}{2}p - \frac{k+a-1}{a} \\ &= k(\frac{p}{2} - \frac{1}{a}) + \frac{p}{2} - \frac{a-1}{a} \\ &= k(\frac{ap-2}{2a}) + \frac{ap-2a+2}{2a} \\ &> 0, \end{aligned}$$

when $k > k^0 = \frac{2a-2-ap}{ap-2}$. □

The following lemma shows that β^k of (3.1) is a convex combination of $\frac{T^k-1}{T^{k+1}}$ and $\frac{t^k-1}{t^{k+1}}$. Moreover, the sequence $\{\beta^k\}$ of (3.1) is bounded above.

Lemma 3.7. *Let $p \in (0,1]$, $0 < q \le (2-p)^2$ and $a \ge 2$. Then we have*

$$\beta^k = \frac{w^k-1}{w^{k+1}} = \alpha\frac{T^k-1}{T^{k+1}} + (1-\alpha)\frac{t^k-1}{t^{k+1}} \le \alpha\frac{k-1}{k+a} + (1-\alpha)(1 - \frac{b}{k+c}), \tag{3.14}$$

where $\alpha \in [0,1]$. If $2 < ap < 2a-2$, and $k > k^0 = \frac{2a-2-ap}{ap-2}$, then $\alpha < \beta$. If $a \ge b-1$, then $\beta^k \le 1 - \frac{b}{k+c}$. If $2 \le a < b-1$, then there exists a positive integer k^1 such that

$$\beta^k \le \frac{k-1}{k+a}, \ \ \forall k > k^1. \tag{3.15}$$

Proof.

$$\begin{aligned} \frac{w^k-1}{w^{k+1}} &= \frac{\beta(T^k-1) + (1-\beta)(t^k-1)}{\beta T^{k+1} + (1-\beta)t^{k+1}} \\ &= \frac{\beta T^{k+1}\frac{T^k-1}{T^{k+1}} + (1-\beta)t^{k+1}\frac{t^k-1}{t^{k+1}}}{\beta T^{k+1} + (1-\beta)t^{k+1}} \\ &= \alpha\frac{T^k-1}{T^{k+1}} + (1-\alpha)\frac{t^k-1}{t^{k+1}}, \end{aligned}$$

where $\alpha = \frac{\beta T^{k+1}}{\beta T^{k+1}+(1-\beta)t^{k+1}}$. By Lemma 3.6, we have $t^k > T^k$ when $k > k^0 = \frac{a(2-p)}{ap-2}$. Thus we have $\alpha < \beta$ when $k > k^0$. By the definition of T^k, we get $\frac{T^k-1}{T^{k+1}} = \frac{k-1}{k+a}$. On the other hand, by Lemma 2.5, we get

$$\begin{aligned}
\frac{t^k-1}{t^{k+1}} &= \frac{2t^k-2}{p+\sqrt{q+4(t^k)^2}} \leq \frac{p+2t^k-2-p}{p+2t^k} \\
&= 1-\frac{2+p}{p+2t^k} \\
&\leq 1-\frac{2+p}{p+2+2S_l+(p+\frac{q}{2p(l+1)})k} \\
&= 1-\frac{b}{k+c},
\end{aligned}$$

where b and c appears in (2.8). Now, we consider

$$\begin{aligned}
1-\frac{b}{k+c} &\geq \frac{k-1}{k+a} = 1-\frac{a+1}{k+a} \\
&\Leftrightarrow \frac{b}{k+c} \leq \frac{a+1}{k+a} \\
&\Leftrightarrow b(k+a) \leq (k+c)(a+1) \\
&\Leftrightarrow k(b-a-1) \leq a(c-b)+c \\
&\Leftarrow a \geq b-1.
\end{aligned}$$

Thus, we get $\beta^k \leq (1-\frac{b}{k+c})$ as $a \geq b-1$. In addition, we have

$$\begin{aligned}
1-\frac{b}{k+c} &\leq \frac{k-1}{k+a} = 1-\frac{a+1}{k+a} \\
&\Leftrightarrow k(b-a-1) \geq a(c-b)+c.
\end{aligned}$$

Thus, we get

$$\beta^k \leq \frac{k-1}{k+a} \tag{3.16}$$

when $k > [\frac{a(c-b)+c}{b-a-1}]$ and $2 \leq a < b-1$. □

Remark 3.8. As we mentioned in **Remark 3.1**, the practical performance of the modified FISTA [11] and the algorithm [7] depend on the speed of β^k approaching 1. For the algorithm [7], $\beta^k = \frac{T^{k-1}-1}{T^k} = \frac{k-1}{k+a}$ is monotonically decreasing on a for fixed k. In addition, from the proof of above lemma, we see that the speed of β^k approaching 1 of the modified FISTA [11] is slower than one of the algorithm [7] when $2 \leq a < b-1$ and k is sufficiently large enough. Thus, the algorithm [7] with a smaller a maybe lead to a weak performance. Meanwhile, above lemma shows that the proposed algorithm with a larger a possibly leads to a faster practical performance. Figure 2 shows graphically the behaviour of β^k under different a and β, which verifies above argument.

The following lemma shows that $\sum_{k=j}^{\infty} \beta^{j,k}$ is bounded above where $\beta^{j,k} = \prod_{l=j}^{k} \beta^l$ for all j,k and $\beta^{j,k} = 1$ for all $k < j$.

Lemma 3.9. *For all $j \geq 1$, define $\beta^{j,k} = \prod_{l=j}^{k} \beta^l$ for all j,k, and $\beta^{j,k} = 1$ for all $k < j$.*

(i) *Let $l \geq \lceil \frac{q}{p(2-p)} \rceil$ and $a \geq b-1$, then for all $j \geq 1$, it holds that*

$$\sum_{k=j}^{\infty} \beta^{j,k} \leq j + c + 2b, \tag{3.17}$$

where b and c appears in (2.8).

(ii) *If $2 < a < b-1$, then for all $j \geq 1$, it holds that*

$$\sum_{k=j}^{\infty} \beta^{j,k} \leq \max(\frac{j+5}{2}, \frac{k^1+5}{2} + k^1 - j). \tag{3.18}$$

Proof. Item (i) follows from Lemma 2.6 and Lemma 3.7. If $2 < a < b-1$, we consider two different cases. If $j \geq k^1$, then the conclusion follows from Lemma 2.7. Suppose $j < k^1$. Then we have

$$\beta^{j,k} = \prod_{l=j}^{k} \beta^l \leq \prod_{l=k^1}^{k} \beta^l = \frac{k^1-1}{k^1+2}\frac{k^1}{k^1+3}\frac{k^1+1}{k^1+4}\cdots\frac{k-3}{k}\frac{k-2}{k+1}\frac{k-1}{k+2} \leq (\frac{k^1+1}{k})^3 \tag{3.19}$$

for all $k - k^1 \geq 2$. Observe that $\beta^1 = 0$, then $\forall k > 1$, $\beta^{1,k} = 0$. It follows that for all $2 \leq j \leq k^1$

$$\begin{aligned}
\sum_{k=j}^{\infty} \beta^{j,k} &= \sum_{k=j}^{k^1+1} \beta^{j,k} + \sum_{k=k^1+2}^{\infty} \beta^{j,k} \\
&\leq k^1 - j + 2 + (k^1+1)^3 \sum_{k=k^1+2}^{\infty} \frac{1}{k^3} \\
&\leq k^1 - j + 2 + (k^1+1)^3 \int_{t=k^1+1}^{\infty} \frac{dt}{t^3} \\
&\leq k^1 - j + 2 + (k^1+1)^3 \frac{1}{2(k^1+1)^2} \\
&= k^1 - j + \frac{k^1+5}{2}.
\end{aligned}$$

The proof is completed. □

Define $\delta^k = \frac{\|x^k - x^{k-1}\|^2}{2}$. The following theorem shows that the series $\sum_{k=1}^{\infty} k\delta^k$ is convergent.

Lemma 3.10. *Suppose $p \in (0,1]$, $q \leq (2-p)^2$, $p + \sqrt{q} \leq 1$ and $a \geq 2$. Then $\sum_{k=1}^{\infty} k\delta^k < \infty$. In particular, there exists $C > 0$ such that for all $k \geq 1$, $\delta^k \leq \frac{C}{k^2}$.*

Proof. Applying Lemma 2.8 with $y = y^k = x^k + \beta^k(x^k - x^{k-1})$ and $x = x^k$ leads to

$$\phi(x^{k+1}) + \frac{\|x^{k+1} - x^k\|^2}{2\gamma} \leq \phi(x^k) + \frac{(\beta^k)^2\|x^k - x^{k-1}\|^2}{2\gamma}. \tag{3.20}$$

Then we have

$$\delta^{k+1} - (\beta^k)^2\delta^k \leq \gamma(v^k - v^{k+1}) \tag{3.21}$$

where $v^k = \phi(x^k) - \phi(x^*)$. Define $a^k = \frac{t^k-1}{t^{k+1}}$ and $b^k = \frac{T^k-1}{T^{k+1}} = \frac{k-1}{k+a}$. By Lemma 3.7, we have

$$\begin{aligned}(\beta^k)^2 &= (\alpha b^k + (1-\alpha)a^k)^2 \\ &= \alpha^2(b^k)^2 + (1-\alpha)^2(a^k)^2 + 2\alpha(1-\alpha)a^k b^k \\ &\leq \alpha^2(b^k)^2 + (1-\alpha)^2(a^k)^2 + \alpha(1-\alpha)((a^k)^2 + (b^k)^2) \\ &= \alpha(b^k)^2 + (1-\alpha)(a^k)^2. \qquad (3.22)\end{aligned}$$

We consider two different cases. If $a \geq b-1$, by Lemma 3.7, (3.21) and (3.22), we get

$$(\beta^k)^2 \leq (\bar{a}^k)^2 \qquad (3.23)$$

and

$$\delta^{k+1} - (\bar{a}^k)^2\delta^k \leq \gamma(v^k - v^{k+1}), \qquad (3.24)$$

where

$$\bar{a}^k = \begin{cases} 1 - \frac{b}{k+c}, & \text{if } k \geq 2; \\ 0, & \text{if } k = 1. \end{cases} \qquad (3.25)$$

Multiplying (3.24) by $(k+c)^2$ and summing from $k=1$ to N, we get

$$\sum_{k=1}^{N}(k+c)^2(\delta^{k+1} - (\bar{a}^k)^2\delta^k) \leq \gamma\sum_{k=1}^{N}(k+c)^2(v^k - v^{k+1}). \qquad (3.26)$$

Since $\bar{a}^1 = 0$, we get

$$\begin{aligned}\sum_{k=1}^{N}(k+c)^2(\delta^{k+1} - (\bar{a}^k)^2\delta^k) &= \sum_{k=1}^{N}(k+c)^2\delta^{k+1} - \sum_{k=2}^{N}(k+c-b)^2\delta^k \\ &= (N+c)^2\delta^{N+1} + \sum_{k=1}^{N-1}(k+c)^2\delta^{k+1} - \sum_{k=2}^{N}(k+c-b)^2\delta^k \\ &= (N+c)^2\delta^{N+1} + \sum_{k=2}^{N}((k+c-1)^2 - (k+c-b)^2)\delta^k \\ &= (N+c)^2\delta^{N+1} + \sum_{k=2}^{N}(2k+2c-b-1)(b-1)\delta^k \\ &> (N+c)^2\delta^{N+1} + \sum_{k=2}^{N}(2k)(b-1)\delta^k.\end{aligned}$$

By (3.26) and the last inequality, we get

$$\begin{aligned}(N+c)^2\delta^{N+1}+\sum_{k=2}^{N}2k(b-1)\delta^k &\leq \gamma\sum_{k=1}^{N}(k+c)^2(v^k-v^{k+1})\\ &= \gamma(1+c)^2v^1-\gamma(N+c)^2v^{N+1}+\gamma\sum_{k=2}^{N}(k+c)^2v^k\\ &\quad -\gamma\sum_{k=1}^{N-1}(k+c)^2v^{k+1}\\ &\leq \gamma(1+c)^2v^1+\gamma\sum_{k=2}^{N}((k+c)^2-(k+c-1)^2)v^k\\ &= \gamma(1+c)^2v^1+\gamma\sum_{k=2}^{N}(2k+2c-1)v^k\\ &\leq \gamma(1+c)^2v^1+\gamma\sum_{k=2}^{N}(2k+2c)v^k.\end{aligned}$$

From the proof of Theorem 3.3 of [11] and the last inequality, we get that $\sum_{k=1}^{\infty}k\delta^k<\infty$ and there exists $C>0$ such that $\delta^k\leq\frac{C}{k^2}$ for all $k\geq 1$. If $2\leq a<b-1$, by Lemma 3.7, then there exists a positive integer k^1 such that $a^k<b^k$. Further by (3.21) and (3.22) we have

$$\delta^{k+1}-(b^k)^2\delta^k\leq\gamma(v^k-v^{k+1}) \tag{3.27}$$

for all $k\geq k^1$. Multiplying (3.27) by $(k+a)^2$ and summing from $k=k^1$ to N, we get

$$\sum_{k=k^1}^{N}(k+a)^2(\delta^{k+1}-(b^k)^2\delta^k)\leq\gamma\sum_{k=k^1}^{N}(k+a)^2(v^k-v^{k+1}). \tag{3.28}$$

Similarly, we get

$$\begin{aligned}&(N+a)^2\delta^{N+1}+\sum_{k=k^1+1}^{N-1}((k+a-1)^2-(k-1)^2)\delta^k-(k^1-1)\delta^{k^1}\\ &\leq\gamma((a+k^1)^2v^1-(N+a)^2v^{N+1}+\sum_{k=k^1+1}^{N}((k+a)^2-(k+a-1)^2)v^k).\end{aligned}$$

That is,

$$\begin{aligned}&(N+a)^2\delta^{N+1}+\sum_{k=k^1+1}^{N}a(2k+a-2)\delta^k-(k^1-1)\delta^{k^1}\\ &\leq\gamma((a+k^1)^2v^1-(N+a)^2v^{N+1}+\sum_{k=k^1+1}^{N}(2k+2a-1)v^k).\end{aligned}$$

By Theorem 2.9 and the last inequality, we get the conclusion. □

The following theorem shows that $\{\|x^k-x^*\|\}$ is convergent and any limit point of $\{x^k\}$ is a solution of problem (1.1).

Theorem 3.11. *Suppose $p \in (0,1]$, $q \le (2-p)^2$, $p + \sqrt{q} \le 1$ and $a \ge 2$. Then $\{\|x^k - x^*\|\}$ is convergent and any limit point of $\{x^k\}$ is a solution of problem (1.1).*

Proof. Define $v^k = \frac{\|x^k - x^*\|^2}{2}$ and $\mathcal{T}^k = \frac{\|x^{k+1} - y^k\|^2}{2}$. By the definition of y^k, we get

$$\begin{aligned} v^k - v^{k+1} &= \frac{\|x^k - x^{k+1} + x^{k+1} - x^*\|^2}{2} - \frac{\|x^{k+1} - x^*\|^2}{2} \\ &= \delta^{k+1} + (x^k - x^{k+1})^T (x^{k+1} - x^*) \qquad (3.29) \\ &= \delta^{k+1} + (y^k - x^{k+1})^T (x^{k+1} - x^*) - \beta^k (x^k - x^{k-1})^T (x^{k+1} - x^*). \qquad (3.30) \end{aligned}$$

By using the monotonicity of ∂r, the definition of x^* and $y^k - x^{k+1} - \gamma \nabla f(y^k) \in \gamma \partial r(x^{k+1})$, we get

$$(y^k - x^{k+1} - \gamma \nabla f(y^k) + \gamma \nabla f(x^*))^T (x^{k+1} - x^*) \ge 0 \qquad (3.31)$$

which yields

$$(y^k - x^{k+1})^T (x^{k+1} - x^*) + \gamma (\nabla f(x^*) - \nabla f(y^k))^T (x^{k+1} - x^*) \ge 0. \qquad (3.32)$$

The last inequality and (3.30) lead to

$$v^k - v^{k+1} \ge \delta^{k+1} + \gamma (\nabla f(y^k) - \nabla f(x^*))^T (x^{k+1} - x^*) - \beta^k (x^k - x^{k-1})^T (x^{k+1} - x^*). \quad (3.33)$$

Note that

$$\begin{aligned} &(\nabla f(y^k) - \nabla f(x^*))^T (x^{k+1} - x^*) \\ &\quad = (\nabla f(y^k) - \nabla f(x^*))^T (x^{k+1} - y^k + y^k - x^*) \\ &\quad \ge \frac{1}{L_f} \|\nabla f(y^k) - \nabla f(x^*)\|^2 + (\nabla f(y^k) - \nabla f(x^*))^T (x^{k+1} - y^k) \\ &\quad \ge \frac{1}{L_f} \|\nabla f(y^k) - \nabla f(x^*)\|^2 - \frac{1}{L_f} \|\nabla f(y^k) - \nabla f(x^*)\|^2 - \frac{L_f}{2} \mathcal{T}^k \\ &\quad = -\frac{L_f}{2} \mathcal{T}^k. \end{aligned}$$

By the last inequality and (3.33), we get

$$v^k - v^{k+1} \ge \delta^{k+1} - \frac{\gamma L_f}{2} \mathcal{T}^k - \beta^k (x^k - x^{k-1})^T (x^{k+1} - x^*). \qquad (3.34)$$

Again using (3.29), we get

$$v^{k-1} - v^k = \delta^k + (x^{k-1} - x^k)^T (x^k - x^*). \qquad (3.35)$$

The last inequality and (3.35) lead to

$$\begin{aligned} v^{k+1} - v^k + \beta^k (v^{k-1} - v^k) &\le -\delta^{k+1} + \beta^k \delta^k + \frac{\gamma L_f}{2} \mathcal{T}^k + \beta^k (x^k - x^{k-1})^T (x^{k+1} - x^k) \\ &= -\mathcal{T}^k + \frac{\gamma L_f}{2} \mathcal{T}^k + (\beta^k + (\beta^k)^2) \delta^k, \end{aligned}$$

where we used the fact

$$\delta^{k+1} - \beta^k (x^k - x^{k-1})^T (x^{k+1} - x^k) = -(\beta^k)^2 \frac{\|x^k - x^{k-1}\|^2}{2} + \frac{\|x^{k+1} - y^k\|^2}{2}. \qquad (3.36)$$

Since $0 < \beta^k \leq 1$ and $1 - \frac{\gamma L_f}{2} > 0$, we get

$$v^{k+1} - v^k - \beta^k(v^k - v^{k-1}) \leq -(1 - \frac{\gamma L_f}{2})\mathcal{T}^k + 2\beta^k\delta^k \leq 2\beta^k\delta^k. \tag{3.37}$$

Letting $\theta^k = \max(0, v^k - v^{k-1})$, we get

$$\theta^{k+1} \leq \beta^k(\theta^k + 2\delta^k). \tag{3.38}$$

Since $\beta^1 = 0$, by applying the last inequality recursively, we obtain

$$\theta^{k+1} \leq 2\sum_{j=2}^{k}(\prod_{l=j}^{k}\beta^l)\delta^j = 2\sum_{j=2}^{k}\beta^{j,k}\delta^j. \tag{3.39}$$

By Lemma 3.9, we get

$$\begin{aligned}
\sum_{k=2}^{\infty}\theta^k &\leq 2\sum_{j=1}^{\infty}\sum_{j=2}^{k}\beta^{j,k}\delta^j \\
&\leq 2\sum_{j=2}^{\infty}\delta^j(\sum_{k=j}^{\infty}\beta^{j,k}) \\
&\leq 2\sum_{j=1}^{\infty}\delta^j(\max(j + c + 2b, \max(\frac{j+5}{2}, k^1 - j + \frac{k^1+5}{2}))).
\end{aligned}$$

By Lemma 3.10, we obtain $\sum_{k=1}^{\infty}\theta^k < \infty$. Define $s^k = v^k - \sum_{i=1}^{k}\theta^i$. Since $v^k \geq 0$, $\theta^i \geq 0$ and $\sum_{k=1}^{\infty}\theta^k < \infty$, we get $s^k \geq -\sum_{i=1}^{k}\theta^i$, which shows s^k is bounded below. Furthermore, by the definition of s^k, we get

$$s^{k+1} = v^{k+1} - \theta^{k+1} - \sum_{i=1}^{k}\theta^i \leq v^{k+1} - v^{k+1} + v^k - \sum_{i=1}^{k}\theta^i = s^k, \tag{3.40}$$

which implies that $\{s^k\}$ is a nondecreasing sequence. Hence $\{s^k\}$ is convergent. It follows that $\{v^k\}$ is convergent, meaning that $\lim\limits_{k\infty}\|x^k - x^*\|$ exists for any x^* such that $0 \in \nabla f(x^*) + \partial r(x^*)$.

Let $\bar{x}$ be any limit point of $\{x^k\}$. Then there exists an infinite set $K \subset \{1, 2, \cdots, n\}$ such that $x^k \to \bar{x}$ for $k \in K$. By the optimal condition of $P_\gamma(y^k)$, we get

$$y^k - x^{k+1} - \gamma\nabla f(y^k) \in \gamma\partial r(x^{k+1}) \tag{3.41}$$

for any $k \in K$. Note that Lemma 3.7 shows that $0 \leq \beta^k < 1$ and Lemma 3.10 shows that $x^k - x^{k-1} \to 0$ as $k \to \infty$. Since $\nabla f(.)$ is Lipschitz continuous, $y^k = x^k + \beta^k(x^k - x^{k-1})$ and $\partial r(x)$ is closed, we get $0 \in \nabla f(\bar{x}) + \partial r(\bar{x})$, i.e. $\bar{x}$ is a solution of (1.1). The proof is completed. □

4 A Modified Restarting Strategy

As pointed out by Tao, Boley and Zhang [16] that ISTA and FISTA have the local linear convergence, but FISTA's convergence rate can slow down as it proceeds, eventually becoming slower than ISTA when they are applied to solve the LASSO problem. This is mainly caused by the oscillatory behavior of scheme [10]. An efficient way to deal way with oscillation is the restart technique developed in [14]. The basic idea of restart is that if the condition

$$(y^k - x^{k+1})^T(x^{k+1} - x^k) > 0 \tag{4.1}$$

is satisfied, then the algorithm resets t^k and y^k in (1.2). Many numerical experiments [11, 10, 14] show that this restart strategy is very efficient. Moreover, the restart technique (4.1) has also been adopted in the popular software TFOCS [3]. Note that if $y^k = x^k$ in (1.2), then (4.1) reduces to

$$(y^k - x^{k+1})^T(x^{k+1} - x^k) = -\|y^k - x^{k+1}\|^2 > 0. \tag{4.2}$$

It seems reasonable to use the following restart condition

$$(y^k - x^{k+1})^T(x^{k+1} - x^k) > -c\|y^k - x^{k+1}\|^2, \tag{4.3}$$

where $0 \leq c \leq 1$. Clearly, the condition (4.3) is weaker than the condition (4.1) and (4.3) with $c = 0$ reduces to the condition (4.1). Moreover, the condition (4.3) is equivalent to

$$(y^k - x^{k+1})^T((1-c)x^{k+1} - x^k + cy^k) > 0. \tag{4.4}$$

which implies that the condition (4.1) does not add extra computation cost when it was compared with the condition (4.1).

5 Numerical Experiments

In this section, we state some numerical experiments to test the performance of the restart strategy (4.3) and the proposed algorithm. We now briefly describe the implementation details of Algorithm 3.1. We implement Algorithm 3.1 with the following parameters $\beta = 0.5$, $p = 0.98$, $q = 10^{-4}$ and $a = 2.1$. For convenience, we abbreviate Algorithm 3.1 as **FISTA-c**, Algorithm 3.1 with the restart technique (4.3) as **FISTA-c-new-ar**, **FISTA** with the restart technique (4.1) as **FISTA-ar** and **FISTA** with the restart technique (4.3) as **FISTA-new-ar** respectively. All codes are written in MATLAB 7.0 and all tests described in this section are performed on a PC with Intel I5-3230 2.6GHZ CPU processor and 16G RAM memory with a Windows operating system.

In first numerical experiment, we test the efficiency of the restart condition (4.3). We consider the simple least square problem:

$$\min_{x \in \mathcal{R}^n} \phi(x) = \frac{1}{2}\|Ax\|^2, \tag{5.1}$$

where A is of the form

$$A = \begin{bmatrix} 2 & -2 & & & \\ -2 & 2 & -2 & & \\ & \ddots & \ddots & \ddots & \\ & & & -2 & 2 \end{bmatrix}. \tag{5.2}$$

This problem has a unique minimizer $x^* = 0$. Choose $L_f = \|A\|^2 = 16$. We set $c = \gamma = \frac{1}{16}$ and the initial point $x^0 = 10^4$ in all tested four algorithms. We stop all tested algorithm if $\phi(x^k) \leq 10^{-7}$ or the number of iteration exceeds 15000. Figures 5-6 plot the evolution of the objective function value $\phi(x^k)$ in log2 scale versus the iteration number and CPU time and

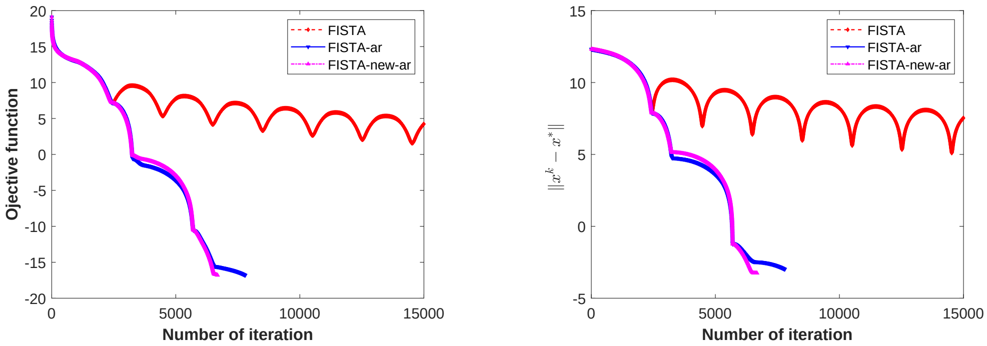

Figure 3: Performance profiles $m = n = 501$ for two different adaptive restart strategy

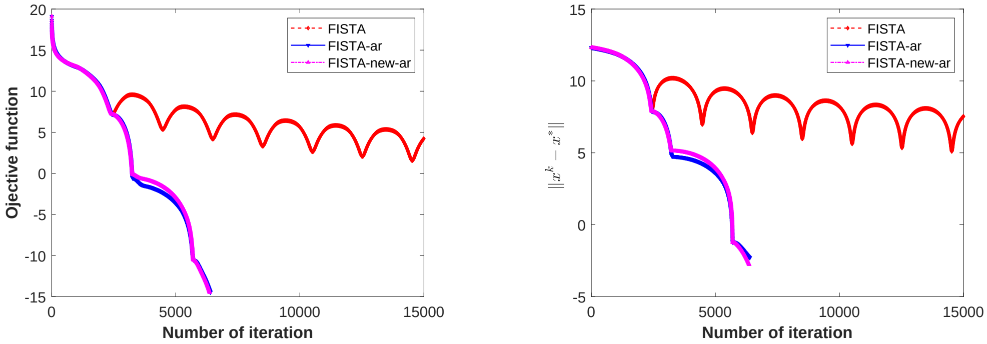

Figure 4: Performance profiles $m = n = 1001$ for two different adaptive restart strategy

the value $\|x^k - x^*\|$ in log2 scale versus the number of iteration. Tables 1-2 list the number of iteration (iter), the final function value (fun), the CPU time (time) and the number of restarting (restart). From Figures 3-4 and Tables 1-2, we see that FISTA-new-ar requires least the number of iteration and the CPU time. This shows the new restart strategy (4.3) is competitive with the restart strategy (4.1).

In second numerical experiment, we consider the following regularized logistic regression problem:

$$\min_{x \in \mathcal{R}^n} \phi(x) = \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-b_i x^T w_i) + \lambda \|x\|_1, \tag{5.3}$$

where $w_i \in \mathcal{R}^n$, $i = 1, 2, \cdots, m$, are the training samples and $b_i \in \{-1, 1\}$ are the corresponding labels. Problem (5.3) is used for training a linear classifier. We first test some simulated data. Specifically, $b_i \in \{-1, 1\}$, $i = 1, 2, \cdots, m$, $\lambda = 0.01$ and $A \in \mathcal{R}^{m \times n}$ are

Table 1: Data for $m = n = 501$

	iter	fun	time	restart
FISTA	15000	69.9267	43.71	0
FISTA-ar	7799	4.99e-08	19.20	4
FISTA-new-ar	6688	4.99e-08	17.89	4

Table 2: Data for $m = n = 701$

	iter	fun	time	restart
FISTA	15000	116.2188	55.71	0
FISTA-ar	6397	4.97e-08	19.32	3
FISTA-new-ar	6367	4.97e-08	18.31	3

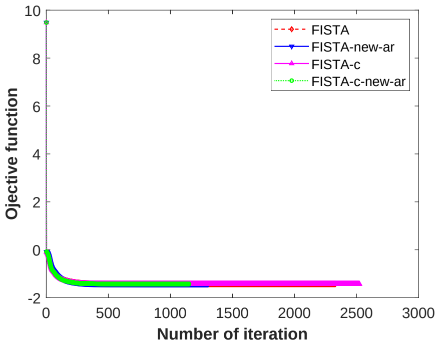

Figure 5: Performance profiles based on iteration number.

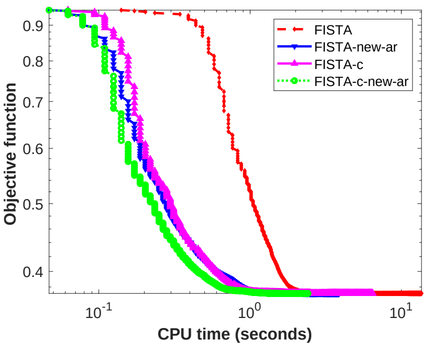

Figure 6: Performance profiles based on CPU time.

constructed randomly as follows:

$$\begin{aligned} &\mathrm{A = randn(m,n); x\_hat = zeros(n,1);} \\ &\mathrm{y = randperm(n); x\_hat(y(1:s)) = randn(s,1);} \\ &\mathrm{c = rand(1,1); b = sgn(Ax\_hat + c * ones(m,1));} \end{aligned} \tag{5.4}$$

where s is a given positive number. Choose $L_f = \|A\|^2$ and $w_i = A(i,:)$. At last, we test two problems which are downloaded from the website: https://www.csie.ntu.edu.tw/cjlin/libsvmtools
/datasets/. The input data w_i in each example has been normalized, that is $\|w_i\| = 1$ for all $i = 1, 2, \cdots m$, which leads to the Lipschitz constant $L_f = 1$. We set $c = \gamma = \frac{1}{L_f}$ and the initial point $x^0 = 0$ in all tested algorithms. In this test, we stop all tested algorithms if $\frac{|\phi(x^k) - \phi(x^{k-1})|}{\max\{|\phi(x^k)|, |\phi(x^{k-1})|\}} \leq 10^{-10}$ or the number of iteration exceeds 15000. Figures 5-12 plot the evolution of the objective function value $\phi(x^k)$ in log2 scale versus the iteration number and the evolution of the objective function value $\phi(x^k)$ versus the CPU time in log2 scale. Tables 3-6 list the number of iteration (iter), the final function value (fun), the CPU time (time) and the number of restart (restart). As far as the number of iteration and the CPU time are concerned, it is easy to see from Figures 5-12 and Tables 3-6 that FISTA-c-new-ar is the best in all tested four algorithms, followed by the FISTA-new-ar.

6 Conclusion

In this paper, we proposed a class of fast iterative shrinkage thresholding algorithm which contains the original FISTA scheme, the FISTA-CD scheme and the FISTA-Mod. We proved that the objective value sequence possesses $O(\frac{1}{k^2})$ complexity. Moreover, we showed that $\{\|x^k - x^*\|\}$ is convergent for any $x^* \in \operatorname{Arg\,min} \phi$ and any limit point of $\{x^k\}$ is a solution of

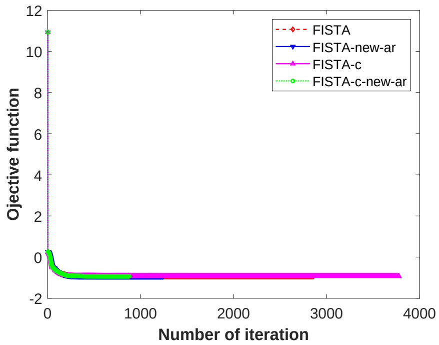

Figure 7: Performance profiles based on iteration number.

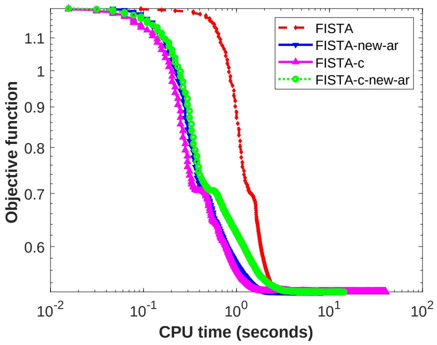

Figure 8: Performance profiles based on error in log scale.

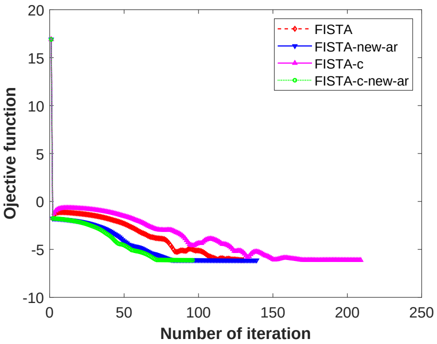

Figure 9: Performance profiles based on iteration number for abalone.

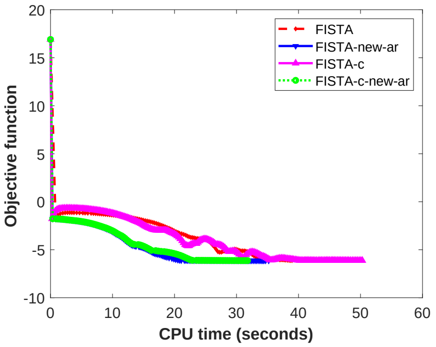

Figure 10: Performance profiles based on CPU time for abalone.

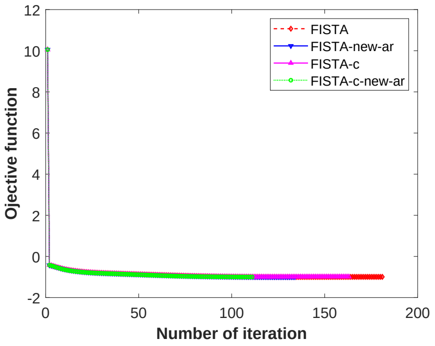

Figure 11: Performance profiles based on iteration number for space_ga.

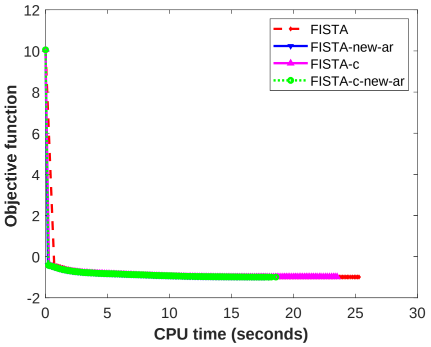

Figure 12: Performance profiles based on CPU time for space_ga.

Table 3: Data for $m = 200, n = 1000, s = 10$

	iter	fun	time	restart
FISTA	1610	0.1596	8.34	0
FISTA-new-ar	1248	0.1595	5.37	4
FISTA-c	2778	0.1589	7.12	0
FISTA-c-new-ar	1121	0.1595	4.39	7

Table 4: Data for $m = 500, n = 1000, s = 20$

	iter	fun	time	restart
FISTA	2858	0.3203	29.65	0
FISTA-new-ar	1243	0.3195	18.57	3
FISTA-c	3776	0.3171	40.21	0
FISTA-c-new-ar	876	0.3195	14.35	7

Table 5: Problem abalone with $m = 4177, n = 8$

	iter	fun	time	restart
FISTA	130	1.44e-02	39.06	0
FISTA-new-ar	139	1.43e-02	35.19	4
FISTA-c	209	1.43e-02	50.34	0
FISTA-c-new-ar	96	1.43e-02	31.78	4

Table 6: Problem space_ga with $m = 3107, n = 6$

	iter	fun	time	restart
FISTA	181	5.03e-01	25.41	0
FISTA-new-ar	134	5.02e-01	18.02	1
FISTA-c	163	5.05e-01	23.52	0
FISTA-c-new-ar	111	5.02e-01	18.58	1

problem (1.1). At last, we proposed a modified adaptive restart strategy which can dramatically improve the convergence rate of the proposed algorithm. The numerical comparisons with several state-of-art methods demonstrated the efficiency of the proposed algorithm.

7 Conflict of Interest

The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

References

[1] G. Bareilles and F. Iutzeler, On the interplay between acceleration and identification for the proximal gradient algorithm, *Comput. Optim. Appl.* 77 (2020) 351–378.

[2] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imaging Sci.* 2 (2009) 183–202.

[3] S. Becker, E.J. Candès, and M.C. Grant, Templates for convex cone problems with applications to sparse signal recovery, *Math. Program. Comput.* 3 (2011) 165–218.

[4] A.M. Bruckstein, D.L. Donoho and M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, *SIAM Rev.* 51 (2009) 34–81.

[5] L. Calatroni and A. Chambolle, Backtracking strategies for accelerated descent methods with smooth composite objectives, *SIAM J. Optim.* 29 (2019) 1772–1798.

[6] E.J. Candes, M.B. Wakin, and S. P. Boyd, Enhancing sparsity by reweighted ℓ_1 minimization, *J. Fourier Anal. Appl.* 14 (2008) 877–905.

[7] A. Chambolle and C. Dossal, On the convergence of the iterates of the "fast iterative shrinkage/thresholding-algorithm", *J Optimiz Theory App.* 166 (2015) 968–982.

[8] W.Y. Cheng and Y.H. Dai, Gradient-based method with active set strategy for ℓ_1 optimization, *Math. Comp.* 87 (2018) 1283–1305.

[9] P.R. Johnstone and P. Moulin, Local and global convergence of an inertial version of forward-backward splitting, arXiv preprint arXiv: 1502.02281v4 (2015).

[10] J. Liang, J. Fadili, and G. Peyré, Activity identification and local linear convergence of Forward-Backward-type methods, *SIAM J. Optim.* 27 (2017) 408–437.

[11] J. Liang, C.B. Schönlieb, C.B, Improving "fast iterative shrinkage-thresholding algorithm": Faster, smarter and greedier, arXiv preprint arXiv:1811.01430 (2021).

[12] Z.Q. Luo and P. Tseng, Error bounds and convergence analysis of feasible descent methods: general approach, *Ann. Oper. Res.*, 46 (1993) 157-178.

[13] Y.E. Nesterov, A method for solving the convex programming problem with convergence rate $O(\frac{1}{k^2})$, *Dokl. Akad. Nauk SSSR.* 269 (1983) 543–547.

[14] B. O'Donoghue and E. Candès, Adaptive restart for accelerated gradient schemes, *Found. Comput. Math.*, 15 (2015) 715–732.

[15] P. Ochs and T. Pock, Adaptive fista for nonconvex optimization, *SIAM J. Optim.* 29 (2019) 2482–2503.

[16] S. Tao, D. Boley, and S. Zhang, Local linear convergence of ISTA and FISTA on the LASSO problem, *SIAM J. Optim.* 26 (2016) 313–336.

[17] B.Wen, X.J. Chen and T.K. Pong, Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems, *SIAM J. Optim.* 27 (2017) 124–145.

Manuscript received 12 October 2021
revised 14 June 2022
accepted for publication 18 October 2022

Hanlin Zhou
College of Computer and Science Technology
Dongguan University of Technology
Dongguan 523000, China
E-mail address: 136607418@qq.com

Wanyou Cheng
College of Computer and Science Technology
Dongguan University of Technology
Dongguan 523000, China
E-mail address: chengwanyou@sina.com

Jianhao Ye
College of Computer and Science Technology
Dongguan University of Technology
Dongguan 523000, China
E-mail address: 458523944@qq.com

Jiahao Zhang
College of Computer and Science Technology
Dongguan University of Technology
Dongguan 523000, China
E-mail address: 1095199414@qq.com