Computational Optimization and Applications (2022) 83:651–691
https://doi.org/10.1007/s10589-022-00396-6

Some modified fast iterative shrinkage thresholding algorithms with a new adaptive non-monotone stepsize strategy for nonsmooth and convex minimization problems

Hongwei Liu[1] · Ting Wang[1] · Zexian Liu[2]

Received: 14 October 2020 / Accepted: 30 June 2022 / Published online: 26 July 2022

Abstract
The " fast iterative shrinkage-thresholding algorithm " (FISTA) is one of the most famous first order optimization schemes, and the stepsize, which plays an important role in theoretical analysis and numerical experiment, is always determined by a constant relating to the Lipschitz constant or by a backtracking strategy. In this paper, we design a new adaptive non-monotone stepsize strategy (NMS), which allows the stepsize to increase monotonically after finite iterations. It is remarkable that NMS can be successfully implemented without knowing the Lipschitz constant or without backtracking. And the additional cost of NMS is less than the cost of some existing backtracking strategies. For using NMS to the original FISTA (FISTA_NMS) and the modified FISTA (MFISTA_NMS), we show that the convergence results stay the same. Moreover, under the error bound condition, we show that FISTA_NMS achieves the rate of convergence to $o\left(\frac{1}{k^6}\right)$ and MFISTA_NMS enjoys the convergence rate related to the value of parameter of t_k, that is $o\left(\frac{1}{k^{2(a+1)}}\right)$; and the iterates generated by the above two algorithms are convergent. In addition, by taking advantage of the restart technique to accelerate the above two methods, we establish the linear convergences of the function values and iterates under the error bound condition. We conduct some numerical experiments to examine the effectiveness of the proposed algorithms.

Keywords FISTA · Proximal-based method · Adaptive non-monotone stepsize strategy · Inertial forward-backward algorithms · Convergence rate · Convex optimization

Mathematics Subject Classification 94A12 · 65K10 · 94A08 · 90C06 · 90C25

✉ Ting Wang
wangting_7640@163.com

Extended author information available on the last page of the article

1 Introduction

We consider the nonsmooth optimization problem:

(P) $\min_{x\in R^n} F(x) = f(x) + g(x)$.

The following assumptions are made throughout the paper:

(A) $g : R^n \to]-\infty, +\infty]$ is a proper, convex, " proximal-friendly " [11] and lower semi-continuous function.

(B) $f : R^n \to]-\infty, +\infty[$ is a smooth convex function and continuously differentiable with Lipschitz continuous gradient, i.e., there exists a Lipschitz constant L_f such that for every $x, y \in R^n$, $\|\nabla f(x) - \nabla f(y)\| \le L_f\|x - y\|$ and $\|\cdot\|$ denotes the standard Euclidean norm.

(C) Problem (P) is solvable, i.e., $X^* := \arg\min F \neq \emptyset$, and for $x^* \in X^*$ we set $F^* := F(x^*)$.

Problem (P) arises in many contemporary applications such as machine learning [24], compressed sensing [13], and image processing [8]. And due to the importance and the popularity of the problem (P), various attempts have been made to solve it efficiently, especially when the problem instances are of large scale. One popular class of methods for solving problem (P) are first-order methods due to their cheap iteration cost and good convergence properties. Among them, the proximal gradient (PG) method [14, 17, 22] is arguably the most fundamental one, in which the basic iteration is

$$x_{k+1} = \mathrm{prox}_{\lambda_k g}\big(x_k - \lambda_k \nabla f(x_k)\big),\ \ \lambda_k \in]0, 1/L_f], \tag{1}$$

where $\mathrm{prox}_{\lambda g}(\cdot) = \arg\min_x \left\{ g(x) + \frac{1}{2\lambda}\|x - \cdot\|^2 \right\}$ denotes the proximal operator of λg, and λ_k indicates the stepsize and has an upper bound relating to Lipschitz constant. The convergence of PG has been well studied in the literature under various contexts and frameworks (The detailed information can be referred to [7, 9, 18, 20]). However, PG can be slow in practice, see, for example, [23].

Various ways have thus been made to accelerate the proximal gradient algorithm. By performing the extrapolation technique, a prototypical algorithm takes the following form:

$$\begin{aligned} y_{k+1} &= x_k + \gamma_k\big(x_k - x_{k-1}\big), \\ x_{k+1} &= p_{\lambda_{k+1} g}\big(y_{k+1}\big), \end{aligned} \tag{2}$$

where γ_k is the extrapolation parameter satisfying $0 \le \gamma_k \le 1$, $\lambda_{k+1} \in]0, 1/L_f]$, and

$$p_{\lambda g}(y) = \arg\min_x \big\{ Q_\lambda(x, y) \big\} = \mathrm{prox}_{\lambda g}(y - \lambda \nabla f(y)). \tag{3}$$

Here $Q(x, y)$ is the approximation function of $F(x)$ at the given point y, where

$$Q_\lambda(x,y) = g(x) + f(y) + \langle \nabla f(y), x-y\rangle + \frac{1}{2\lambda}\|x-y\|^2, \quad \forall x \in R^n. \tag{4}$$

One representative algorithm that takes the form of (2) and with the extrapolation parameter

$$\gamma_k = \frac{t_k - 1}{t_{k+1}}, \text{ where } t_1 = 1, t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2} \tag{5}$$

is the fast iterative shrinkage-thresholding algorithm (FISTA), which was proposed by Beck and Teboulle [5] and was based on the idea that was introduced and developed by Nesterov [21] for minimizing a smooth convex function. The stepsize λ_k can be dynamically updated to estimate the Lipschitz constant L_f by a backtracking stepsize rule. FISTA is a very effective algorithm that keeps the simplicity of schemes like PG and improves the convergence rate of objective function value to $O\left(k^{-2}\right)$ for solving the problem (P), hence, it has become a standard algorithm [17] and motivates subsequent studies on the extrapolation scheme (2), see, for example, [6, 19, 22, 23, 29, 32]. Though FISTA is surprisingly efficient, the convergence of the whole iterative sequence generated by FISTA is still unclear [12, 30]. Chambolle and Dossal [12] established the convergence of the sequence generated by FISTA with the new parameter $\gamma_k = \frac{k-1}{k+a}$ for a fixed $a > 2$ and the assumption $\lambda_k \in]0, \frac{1}{L_f}]$, for problem (P). Furthermore, Attouch and Peypouquet [1] proved that the convergence rate of function values generated by the algorithm in [12] is $o\left(k^{-2}\right)$ and they considered the convergence of iterates and the rate of convergence of function values for the scheme of (2) with various options of extrapolation parameter γ_k in [2]. In addition, in [3], the authors showed that the algorithm in [12] with $a \in (0, 2)$ enjoys the convergence rates of function value as $O\left(k^{-\frac{2(a-1)}{3}}\right)$. Under the convex setting, Wen, Chen and Pong [30] established the R-linear convergence of the sequences generated by (2) with a parameter $\sup \gamma_k < 1$ based on the error bound condition. However, the stepsize in these algorithms such as [1–3, 12, 30] is directly related to the Lipschitz constant, which results in that the algorithm implementation as well as the theoretical analysis rely heavily on the Lipschitz constant.

Backtracking for estimating Lipschitz constant works well in practice but the principal drawback is that the stepsize λ_k generated by the backtracking strategy in FISTA is non-increasing. This non-increasing property can substantially limit the performance of FISTA when a small stepsize is encountered early in the algorithm since this causes the stepsize taken at that point, and at all subsequent iterates, to be very small. Scheinberg, Goldfarb and Bai [27] developed a new backtracking strategy, which allows stepsize to increase. This new backtracking strategy [27] starts with a new initial value at the beginning of each iteration, rather than the stepsize of last iteration like the backtracking in FISTA, and estimates the local Lipschitz constant L_k, which is often smaller than L_f. Hence, $\frac{1}{L_k}$ may be a better estimate for the stepsize than $\frac{1}{L_f}$. With this new backtracking strategy, they proposed a new version of accelerated FISTA (FISTA_BKTR), which reduces the number of iteration greatly

and the calculating cost is much less than the one in backtracking rule of original FISTA, and the convergence result is still $O\left(1/k^2\right)$.

It is natural that each time the backtracking step operates, the calculating cost of algorithm will increase. Although both of the mentioned backtracking strategies work well, we still pursue to develop a stepsize strategy, which does not use the backtracking procedure and can bring some numerical improvements and some new theoretical results. In this paper, we exploit a new adaptive non-monotone stepsize technique (NMS) to determine λ_k in (2), where the stepsize increases monotonically after finite iterations. We prove that FISTA with NMS keeps $O\left(1/k^2\right)$ convergence rate of the objective function values, which is similar to original FISTA and FISTA_BKTR. By using the new choice of t_k in [12] and the new adaptive non-monotone technique, we present a modified FISTA (MFISTA) with NMS which also achieves $o\left(\frac{1}{k^2}\right)$ convergence rate of the objective function values. Also, the convergence of the iterative sequence is established without depending on the Lipschitz constant L_f unlike the analysis in [12]. Meanwhile, we prove that both of those two algorithms with NMS enjoy $o\left(\frac{1}{k}\right)$ convergent rates of the norm of subdifferential of the objective function. Furthermore, under the error bound condition, we prove that FISTA and MFISTA with NMS can achieve some improved convergence rates for objective function values and the sequence of iterates is convergent; we also take advantage of the restart technique in [23] to accelerate the above FISTA methods with NMS, and establish the linear convergences of the function values and iterative sequence under the error bound condition.

The reminder of the paper is organized as follows. In Sect. 2, we provide a new adaptive non-monotone stepsize strategy. In Sect. 3, we propose an algorithm FISTA_NMS by combining FISTA with the new adaptive non-monotone technique, which ensures the similar convergence rate of the objective function values with FISTA, and a faster convergence rate of the norm of subdifferential of function value than FISTA (See Sect. 3.1 for details). Also, with a small modification, we present a MFISTA_NMS which has similar theoretical results like [1, 12, 26] (See Sect. 3.2 for details). In Sect. 3.3, under the error bound condtion, we show that FISTA_NMS and MFISTA_NMS enjoy improved convergence results. In Sect. 4, we use the restart technique in [23] to accelerate the above methods and establish the linear convergences of the function values and iterates under the error bound condition. Numerical results are reported in Sect. 5. In the last section, conclusions and discussions are presented.

2 Adaptive non-monotone stepsize strategy

In this section, we present a new adaptive non-monotone stepsize strategy.

We first state the algorithm of FISTA with backtracking [5] and the detailed algorithm of FISTA_BKTR [27] as follows.

Denote the computations of $t_{k+1} := \frac{1+\sqrt{1+4\theta_k t_k^2}}{2}$ and $y_{k+1} := x_k + \frac{t_k-1}{t_{k+1}}\left(x_k - x_{k-1}\right)$ by $(t_{k+1}, y_{k+1}) = FistaStep\left(x_k, x_{k-1}, t_k, \theta_k\right)$. And in following two algorithms,

$p_{\lambda_k g}(y_k)$ is defined in (3) with $\lambda := \lambda_k, y := y_k$ and $Q_{\lambda_k}\left(p_{\lambda_k g}(y_k), y_k\right)$ is defined in (4) with $\lambda := \lambda_k, y := y_k$ and $x := p_{\lambda_k g}(y_k)$.

Algorithm 1 FISTA with backtracking

Step 0. Set $t_1 = 1$ and $y_1 = x_0, \lambda_0 > 0$; $\eta < 1$.
Step k. (1) Finding the smallest nonnegative integers i_k such that with $\lambda_k = \eta^{i_k}\lambda_{k-1}$

$$F\left(p_{\lambda_k g}(y_k)\right) \le Q_{\lambda_k}\left(p_{\lambda_k g}(y_k), y_k\right) \tag{6}$$

(2) Compute $(t_{k+1}, y_{k+1}) = FistaStep(x_k, x_{k-1}, t_k, 1)$

Since that (6) holds if $\lambda_k \le \frac{1}{L_f}$, we have that $\lambda_k > \frac{\eta}{L_f}$, which means that the lower bound of stepsize is related to L_f, and λ_k in Algorithm 1 can be seen as an estimate for the global Lipschitz constant. It is easy to obtain that there are at most $\log_{\frac{1}{\eta}}\left(\lambda_0 L_f\right) + 1$ backtracking steps at each iteration [27]. Each time the backtracking performs, $p_{\lambda_k g}(y_k)$ and $f\left(p_{\lambda_k g}(y_k)\right)$ must be recomputed, that is the main cost of FISTA_backtracking.

To obtain larger stepsize than Algorithm 1, The following Algorithm 2 proposes a new backtracking step rule, which starts with a new initial value at the beginning of each iteration and can be reduced to Algorithm 1 if we set $\lambda_k^0 = \lambda_{k-1}$.

Algorithm 2 FISTA_BKTR

Step 0. Set $t_1 = 1, t_0 = 0, 0 < \beta < 1, \theta_0 = 1$ and $y_1 = x_0 = x^{-1}, \lambda_1^0 > 0$;
Step k. (1) Set $\lambda_k := \lambda_1^0$, and compute $\nabla f(y_k), p_{\lambda_k}(y_k)$
(2) If $F\left(p_{\lambda_k g}(y_k)\right) > Q_{\lambda_k}\left(p_{\lambda_k g}(y_k), y_k\right)$
set $\lambda_k := \beta\lambda_k, \theta_{k-1} := \theta_{k-1}/\beta$
$(t_k, y_k) = FistaStep(x_{k-1}, x_{k-2}, t_{k-1}, \theta_{k-1})$
return to (2)
(3) $x_k := p_{\lambda_k g}(y_k)$
choose $\lambda_{k+1}^0 > 0$ and set $\theta_k := \lambda_k/\lambda_{k+1}^0$
$(t_{k+1}, y_{k+1}) = FistaStep(x_k, x_{k-1}, t_k, \theta_k)$

We see that the updating rule of θ_k is equivalent to $\theta_k = \frac{\lambda_k}{\lambda_{k+1}}$ and λ_k in Algorithm 2 is an estimate for the local Lipschitz constant, while the λ_k in Algorithm 1 is an estimate of the global Lipschitz constant. Similar to the analysis of Algorithm 1, the lower bound of stepsize is related to the local Lipschitz constant L_k for $\nabla f(x)$ restricted to the interval $\left[p_{\lambda_k g}(y_k), y_k\right]$ for any $\lambda_k \le \frac{1}{L_k}$, which is less than or equal to L_f. If the backtracking step is performed, the values of $f(y_k)$, $\nabla f(y_k)$, $p_{\lambda_k g}(y_k)$ and $f\left(p_{\lambda_k g}(y_k)\right)$ must be recomputed. Here, we can see that computation of $f(y_k)$ and $\nabla f(y_k)$ will be additional costs over against Algorithm 1 for the case that ∇f is non-linear; otherwise, those computation can be negligible. Since the option of initial stepsize is related to the number of backtracking steps closely, based on the idea of Nesterov [22], the author chose $\lambda_k^0 = \frac{\lambda_{k-1}}{\sigma}(\sigma \ge \beta)$ to reduce the total number of backtracking steps to $[1 + \frac{\ln\sigma}{\ln\beta}](Iter + 1) + \frac{1}{\ln\beta}[\ln\frac{\sigma\lambda_0}{\beta/L_f}]_+$, where *Iter* means the total number of iterations of Algorithm 2.

Although Algorithm 2 greatly reduces the number of cycle of the internal loop, and generates better stepsize, it still may have additional costs per backtracking step, especially, when the function f is non-linear, the computations of $f(y_k)$, $\nabla f(y_k)$, $p_{\lambda_k g}(y_k)$ and $f(p_{\lambda_k g}(y_k))$ will occupy the CPU time. Hence, we design a stepsize strategy that directly gives the stepsize at each iteration, which avoids any extra computations due to line search. We present the adaptive non-monotone stepsize strategy as follows.

Algorithm 3 Adaptive non-monotone stepsize strategy

For the general sequences $\{x_k\}, \{y_k\}$, let $\sum_{k=1}^{\infty} E_k$ be a convergent nonnegative series. Set $0 < \mu_1 < \mu_0 < 1$.

If $\langle \nabla f(x_k) - \nabla f(y_k), x_k - y_k \rangle > \frac{\mu_0}{\lambda_k} \|x_k - y_k\|^2$ holds, set

$$\lambda_{k+1} = \mu_1 \frac{\|x_k - y_k\|^2}{\langle \nabla f(x_k) - \nabla f(y_k), x_k - y_k \rangle} \tag{7}$$

Otherwise,

$$\lambda_{k+1} = \lambda_k (1 + E_k) \tag{8}$$

In Algorithm 3, we use the condition

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \frac{\mu_0}{\lambda} \|x - y\|^2 \text{ where } \mu_0 \in]0, 1[\tag{9}$$

to control the increase or decrease of the stepsize λ_k. When the condition (9) does not holds, the stepsize λ_k is determined by (7), which implies that $\lambda_{k+1} < \lambda_k$. Conversely, $\lambda_{k+1} \geq \lambda_k$. The $\sum_{k=1}^{\infty} E_k$ is called control series, which can be corrected adaptively for better control of stepsize growth. For the choice of E_k, we will discuss later in this section.

It is remarkable that it is not required to know the Lipschitz constant or use a line search procedure when one uses Algorithm 3 to determine the stepsize λ_k. Now we study some important properties of the stepsize $\{\lambda_k\}$ generated by Algorithm 3.

Lemma 2.1 *Let $\{\lambda_k\}$ be the sequence generated by Algorithm 3. We have that the sequence $\{\lambda_k\}$ is convergent, and*

$$\lambda_k \geq \lambda_{min} := \min\left\{\lambda_1, \frac{\mu_1}{L_f}\right\}, \quad \forall k \geq 1. \tag{10}$$

Proof First, we prove that $\forall k \geq 1$, $\lambda_k \geq \min\left\{\lambda_1, \frac{\mu_1}{L_f}\right\}$ holds by induction.

For $k = 1$, the conclusion is obvious. Suppose that the conclusion holds true for some $k = p \geq 1$. Then, for $k = p + 1$, there are two situations:

(1) λ_{p+1} is generated by (7). We obtain

$$\lambda_{p+1} = \mu_1 \frac{\left\|x_p - y_p\right\|^2}{\left\langle \nabla f(x_p) - \nabla f(y_p), x_p - y_p \right\rangle} \geq \frac{\mu_1}{L_f}, \tag{11}$$

the inequality follows from the fact that f is Lipschitz continuous gradient.

(2) λ_{p+1} is generated by (8). We obtain

$$\lambda_{p+1} \geq \lambda_p \geq \min\left(\lambda_1, \frac{\mu_1}{L_f}\right). \tag{12}$$

From (11) and (12), we conclude that $\forall k \geq 1$, $\lambda_k \geq \min\left\{\lambda_1, \frac{\mu_1}{L_f}\right\}$ holds for $\forall k \geq 1$.

Denote that

$$\ln \lambda_{i+1} - \ln \lambda_i = \left(\ln \lambda_{i+1} - \ln \lambda_i\right)^+ - \left(\ln \lambda_{i+1} - \ln \lambda_i\right)^-, \tag{13}$$

where $(\cdot)^+ = \max\{0, \cdot\}$, $(\cdot)^- = -\min\{0, \cdot\}$. Following the fact that

$$\ln \lambda_{i+1} - \ln \lambda_i \leq \ln\left(1 + E_i\right) \leq E_i, \forall i \geq 1, \tag{14}$$

we have

$$\left(\ln \lambda_{i+1} - \ln \lambda_i\right)^+ \leq E_i, \forall i = 1, 2, \cdots, \tag{15}$$

which implies that $\sum_{i=1}^{\infty}\left(\ln \lambda_{i+1} - \ln \lambda_i\right)^+$ is convergent from the fact that $\sum_{i=1}^{\infty} E_i$ is a convergent nonnegative series.

The convergence of $\sum_{i=1}^{\infty}\left(\ln \lambda_{i+1} - \ln \lambda_i\right)^-$ also can be proved as follows.

Assume by contradiction that $\sum_{i=1}^{\infty}\left(\ln \lambda_{i+1} - \ln \lambda_i\right)^- = +\infty$. Based on the convergence of $\sum_{i=1}^{\infty}\left(\ln \lambda_{i+1} - \ln \lambda_i\right)^+$ and the equality

$$\begin{aligned} \ln \lambda_{k+1} - \ln \lambda_1 &= \sum_{i=1}^{k}\left(\ln \lambda_{i+1} - \ln \lambda_i\right) \\ &= \sum_{i=1}^{k}\left(\ln \lambda_{i+1} - \ln \lambda_i\right)^+ - \sum_{i=1}^{k}\left(\ln \lambda_{i+1} - \ln \lambda_i\right)^- \end{aligned} \tag{16}$$

we can easily deduce $\lim_{k\to\infty} \ln \lambda_k = -\infty$, which is a contradiction with $\lambda_k \geq \min\left\{\lambda_1, \frac{\mu_1}{L_f}\right\} > 0$. As a result, $\sum_{i=1}^{\infty}\left(\ln \lambda_{i+1} - \ln \lambda_i\right)^-$ is a convergent series. Then, in view of (16), we obtain the sequence $\left\{\lambda_k\right\}$ is convergent. □

Lemma 2.2 *For the sequence $\left\{\lambda_k\right\}$ generated by Algorithm 3, there exists a positive integer $\hat{k} \geq 1$ such that condition (9) holds constantly for every $k > \hat{k}$.*

Proof Suppose the conclusion is not true, i.e. there exists a sequence $\left\{k_j\right\}$, where $k_j \to \infty$, such that

$$\begin{aligned}\left\|x_{k_j}-y_{k_j}\right\|^2 &< \frac{\lambda_{k_j}}{\mu_0}\left\langle \nabla f\left(x_{k_j}\right)-\nabla f\left(y_{k_j}\right), x_{k_j}-y_{k_j}\right\rangle \\ &= \frac{\lambda_{k_j}}{\lambda_{k_j+1}}\frac{1}{\mu_0}\lambda_{k_j+1}\left\langle \nabla f\left(x_{k_j}\right)-\nabla f\left(y_{k_j}\right), x_{k_j}-y_{k_j}\right\rangle \\ &= \frac{\lambda_{k_j}}{\lambda_{k_j+1}}\frac{\mu_1}{\mu_0}\left\|x_{k_j}-y_{k_j}\right\|^2.\end{aligned} \tag{17}$$

Combining this with the fact

$$\lim_{j\to\infty}\frac{\lambda_{k_j}}{\lambda_{k_j+1}}\frac{\mu_1}{\mu_0}=\frac{\mu_1}{\mu_0}<1, \tag{18}$$

which follows from Lemma 2.1, we obtain

$$\left\|x_{k_j}-y_{k_j}\right\|^2 < \left\|x_{k_j}-y_{k_j}\right\|^2, \quad \text{for } j \text{ is sufficiently large,} \tag{19}$$

which is a contradiction. Therefore, (9) will hold constantly after finite iterations $\hat{k}$. □

For the rest of this article, we always denote that $k_0=\hat{k}+1$ is the first positive integer such that λ_k satisfies the condition (9), which means that condition (9) holds for any $k\geq k_0$. It follows from Lemma 2.2 that the stepsize $\{\lambda_k\}$ generated by Algorithm 3 increases monotonically after $\hat{k}$ steps.

According to Lemmas 2.1 and 2.2, we can easily obtain the following conclusion.

Corollary 2.1 *For the sequence $\{\lambda_k\}$ generated by Algorithm 3, denote that $\lim_{k\to\infty}\lambda_k=\lambda^*$. Then, for any $k\geq 1$, we have $\lambda_k\leq\lambda_{\max}:=\max\{\lambda_1,\cdots,\lambda_{\hat{k}},\lambda^*\}$.*

Now, we discuss the choice of E_k. In Algorithm 3, we set $E_k:=\frac{w_k}{k^p}(p>1)$, where $\{w_i\}$ is a nonnegative bounded sequence. Generally, we set the value of p close to 1. For the choice of w_k, we can adjust the value of w_k based on the angle between the vectors x_k-x_{k-1} and $x_{k-1}-x_{k-2}$. If the value $\frac{\langle x_k-x_{k-1},x_{k-1}-x_{k-2}\rangle}{\|x_k-x_{k-1}\|\|x_{k-1}-x_{k-2}\|}$ is close to 1, it may be caused by a small stepsize, then, we may want to use a larger stepsize in the next iteration. Hence, we can set the value of w_k adaptively. In the following, we give the details for setting w_k.

Set $w_k=\eta_1$, if $\langle x_k-x_{k-1},x_{k-1}-x_{k-2}\rangle\leq 0.9\|x_k-x_{k-1}\|\|x_{k-1}-x_{k-2}\|$;

set $w_k=\eta_3$, if $\langle x_k-x_{k-1},x_{k-1}-x_{k-2}\rangle\geq 0.98\|x_k-x_{k-1}\|\|x_{k-1}-x_{k-2}\|$;

set $w_k=\eta_2$, otherwise, where $0<\eta_1<\eta_2<\eta_3$. In the numerical experiment, $\eta_1=1,\eta_2=2,\eta_3=10$.

3 FISTA-type algorithm with the adaptive non-monotone stepsize

Based on the adaptive non-monotone stepsize strategy, we present a class of FISTA-type algorithms with non-monotone stepsize (FISTA-type_NMS) and show its convergence results under different inertial terms. The algorithm scheme is as follows:

Algorithm 4 FISTA-type_NMS

Step 0. Take $y_1 = x_0 \in R^n, t_1 = 1, 0 < \mu_1 < \mu_0 < 1$ and $\lambda_1 > 0$

Step k. compute

$$x_k = p_{\lambda_k g}(y_k), \text{ where } p_{\lambda g}(\cdot) \text{ is defined in (3)}$$

Set λ_{k+1} via the adaptive non-monotone stepsize strategy (Algorithm 3)

Choose t_{k+1} such that $\gamma_k = \frac{t_k - 1}{t_{k+1}} \leq 1$ and compute

$$y_{k+1} = x_k + \gamma_k (x_k - x_{k-1})$$

We first show some key results and the theoretical analysis of the algorithms proposed in this paper relies heavily on it. For ease of description, we define the following sequences.

Notation 3.1 Let $\{x_k\}$ and $\{y_k\}$ be generated by the Algorithm 4 and x^* be a fixed minimizer of F. Then, for the convergence of objective function values holds, the sequence $\{v_k\}$ tends to zero when k goes to infinity

$$v_k := F(x_k) - F(x^*). \tag{20}$$

The sequence $\{\delta_k\}$ means the local variation of the sequence $\{x_k\}$

$$\delta_k := \frac{1}{2}\|x_k - x_{k-1}\|^2, \tag{21}$$

and the sequence $\{\Gamma_k\}$, denoting the distance between $\{y_k\}$ and $\{p_{\lambda_k g}(y_k)\}$, is

$$\Gamma_k := \frac{1}{2}\|x_k - y_k\|^2, \tag{22}$$

and we define Φ_k as the distance between $\{x_k\}$ and a fixed minimizer $\{x^*\}$

$$\Phi_k := \frac{1}{2}\|x_k - x^*\|^2. \tag{23}$$

Lemma 3.1 [5] *For any $y \in R^n$, one has $z = p_{\lambda g}(y)$ if and only if there exists $\sigma(y) \in \partial g(z)$ the subdifferential of $g(\cdot)$, such that*

$$\nabla f(y) + \frac{1}{\lambda}(z - y) + \sigma(y) = 0.$$

Lemma 3.2 *For any $y \in R^n, \mu_0 \in]0, 1]$, if y and $p_{\lambda g}(y)$ satisfy the condition (9), then, for any $x \in R^n$,*

$$F(x) - F(p_{\lambda g}(y)) \geq \frac{\bar{\mu}}{\lambda}\|p_{\lambda g}(y) - y\|^2 + \frac{1}{\lambda}\langle y - x, p_{\lambda g}(y) - y\rangle, \tag{24}$$

where $\bar{\mu} = 1 - \frac{\mu_0}{2}$, *if* f *is a quadratic function*; $\bar{\mu} = 1 - \mu_0$, *if* f *is a non-quadratic function*.

Proof Since f, g are convex, we have

$$\begin{aligned} f(x) &\geq f(y) + \langle x - y, \nabla f(y)\rangle, \\ g(x) &\geq g\big(p_{\lambda g}(y)\big) + \big\langle x - p_{\lambda g}(y), \gamma(y)\big\rangle, \end{aligned} \tag{25}$$

where $\gamma(y) = -\nabla f(y) - \frac{1}{\lambda}\big(p_{\lambda g}(y) - y\big) \in \partial g\big(p_{\lambda g}(y)\big)$, and $\partial g(\cdot)$ denotes the subdifferential of $g(\cdot)$. Then,

$$\begin{aligned} &F(x) - F\big(p_{\lambda g}(y)\big) \\ &= f(x) + g(x) - f\big(p_{\lambda g}(y)\big) - g\big(p_{\lambda g}(y)\big) \\ &\geq f(y) + \langle x - y, \nabla f(y)\rangle + \Big\langle p_{\lambda g}(y) - x, \nabla f(y) + \frac{1}{\lambda}\big(p_{\lambda g}(y) - y\big)\Big\rangle \\ &\quad - f\big(p_{\lambda g}(y)\big) \\ &= f(y) - f\big(p_{\lambda g}(y)\big) + \big\langle p_{\lambda g}(y) - y, \nabla f(y)\big\rangle + \frac{1}{\lambda}\big\langle p_{\lambda g}(y) - x, p_{\lambda g}(y) - y\big\rangle \\ &= f(y) - f\big(p_{\lambda g}(y)\big) + \big\langle p_{\lambda g}(y) - y, \nabla f(y)\big\rangle + \frac{1}{\lambda}\big\langle y - x, p_{\lambda g}(y) - y\big\rangle \\ &\quad + \frac{1}{\lambda}\big\|p_{\lambda g}(y) - y\big\|^2. \end{aligned} \tag{26}$$

Denote

$$\begin{aligned} \Delta = f(y) - f\big(p_{\lambda g}(y)\big) + \big\langle p_{\lambda g}(y) - y, \nabla f(y)\big\rangle + \frac{1}{\lambda}\big\langle y - x, p_{\lambda g}(y) - y\big\rangle \\ + \frac{1}{\lambda}\big\|p_{\lambda g}(y) - y\big\|^2. \end{aligned} \tag{27}$$

The proof is derived by dividing into two cases.

(1) In the case that f is a quadratic function, without loss of generality, assume that

$$f(x) = \frac{1}{2}x^T Ax + b^T x, \nabla f(x) = Ax + b. \tag{28}$$

It is easy to obtain that

$$f(x) - f(y) = \frac{1}{2}\langle \nabla f(x) + \nabla f(y), x - y\rangle. \tag{29}$$

Then,

$$\begin{aligned}\Delta &= \tfrac{1}{2}\left\langle \nabla f(y) + \nabla f\left(p_{\lambda g}(y)\right), y - p_{\lambda g}(y)\right\rangle + \left\langle \nabla f(y), p_{\lambda g}(y) - y\right\rangle \\ &\quad + \tfrac{1}{\lambda}\left\langle y - x, p_{\lambda g}(y) - y\right\rangle + \tfrac{1}{\lambda}\left\|p_{\lambda g}(y) - y\right\|^2 \\ &= \tfrac{1}{2}\left\langle \nabla f(y) - \nabla f\left(p_{\lambda g}(y)\right), p_{\lambda g}(y) - y\right\rangle + \tfrac{1}{\lambda}\left\langle y - x, p_{\lambda g}(y) - y\right\rangle \\ &\quad + \tfrac{1}{\lambda}\left\|p_{\lambda g}(y) - y\right\|^2 \\ &\geq \tfrac{1}{\lambda}\left(1 - \tfrac{\mu_0}{2}\right)\left\|p_{\lambda g}(y) - y\right\|^2 + \tfrac{1}{\lambda}\left\langle y - x, p_{\lambda g}(y) - y\right\rangle.\end{aligned} \tag{30}$$

(2) In the case that f is a non-quadratic function,

$$\begin{aligned}\Delta &\geq \left\langle \nabla f\left(p_{\lambda g}(y)\right), y - p_{\lambda g}(y)\right\rangle + \left\langle \nabla f(y), p_{\lambda g}(y) - y\right\rangle \\ &\quad + \tfrac{1}{\lambda}\left\langle y - x, p_{\lambda g}(y) - y\right\rangle + \tfrac{1}{\lambda}\left\|p_{\lambda g}(y) - y\right\|^2 \\ &= \left\langle \nabla f(y) - \nabla f\left(p_{\lambda g}(y)\right), p_{\lambda g}(y) - y\right\rangle + \tfrac{1}{\lambda}\left\langle y - x, p_{\lambda g}(y) - y\right\rangle \\ &\quad + \tfrac{1}{\lambda}\left\|p_{\lambda g}(y) - y\right\|^2 \\ &\geq \tfrac{1}{\lambda}\left(1 - \mu_0\right)\left\|p_{\lambda g}(y) - y\right\|^2 + \tfrac{1}{\lambda}\left\langle y - x, p_{\lambda g}(y) - y\right\rangle.\end{aligned} \tag{31}$$

The last inequalities of (30) and (31) are from the condition (9).
By combining (26), (27), (30) and (31), we can easily obtain (24). □

Remark 3.1 when $\bar{\mu} \geq \frac{1}{2}$, the result (24) of Lemma 3.2 reduces to the Lemma 2.3 in [5]

$$F(x) - F\left(p_{\lambda g}(y)\right) \geq \frac{1}{2\lambda}\left\|p_{\lambda g}(y) - y\right\|^2 + \frac{1}{\lambda}\left\langle y - x, p_{\lambda g}(y) - y\right\rangle, \tag{32}$$

which plays a crucial role for the analysis of FISTA.

We always choose the value range of $\mu_0 \in]0, 1[$ for the quadratic function and $\mu_0 \in]0, \frac{1}{2}[$ for the non-quadratic function, i.e. $\bar{\mu} > \frac{1}{2}$, which means that Lemma 3.2 is a result stronger than Lemma 2.3 of [5].

Further, it follows from the identity

$$\langle a - b, a - c\rangle = \frac{1}{2}\|a - b\|^2 + \frac{1}{2}\|a - c\|^2 - \frac{1}{2}\|b - c\|^2 \tag{33}$$

and (24) that

$$\begin{aligned}F\left(p_{\lambda g}(y)\right) + \frac{\left\|p_{\lambda g}(y) - x\right\|^2}{2\lambda} &\leq F(x) + \frac{\|y - x\|^2}{2\lambda} - \left(\frac{2\bar{\mu} - 1}{2\lambda}\right)\left\|p_{\lambda g}(y) - y\right\|^2 \\ &\leq F(x) + \frac{\|x - y\|^2}{2\lambda},\ \forall x \in R^n, \bar{\mu} > \frac{1}{2}.\end{aligned} \tag{34}$$

Independent of the inertial term, we can obtain the following inequality:

Lemma 3.3 *Let* $\{x_k\}, \{y_k\}$ *be generated by the Algorithm 4. Then, for any* $k \geq k_0 := \hat{k} + 1$, *where* $\hat{k}$ *is defined in Lemma 2.2, we have*

$$\lambda_k t_k^2 v_k - \lambda_{k+1} t_{k+1}^2 v_{k+1} - \rho_k v_k \geq \frac{1}{2}\left(\|u_{k+1}\|^2 - \|u_k\|^2\right) + \left(\bar{\mu} - \frac{1}{2}\right)\left\|t_{k+1}\left(x_{k+1} - y_{k+1}\right)\right\|^2,$$

where $\rho_k = \lambda_k t_k^2 - \lambda_{k+1} t_{k+1}\left(t_{k+1} - 1\right)$, $u_k = t_k x_k - \left(t_k - 1\right)x_{k-1} - x^*$ *and* v_k *is defined in* (20).

Proof Invoking Lemmas 2.2 and 3.2, we obtain that (24) holds for every $k \geq k_0 := \hat{k} + 1$, where $\hat{k}$ is defined in Lemma 2.2.

Denote that $u_k = t_k x_k - \left(t_k - 1\right)x_{k-1} - x^*$. We apply the inequality (24) at the points $\left(x := x_k, y := y_{k+1}\right)$ with $\lambda := \lambda_{k+1}$, and likewise at the points $\left(x := x^*, y := y_{k+1}\right)$, to get

$$\begin{aligned} \lambda_{k+1}\left(v_k - v_{k+1}\right) &\geq \bar{\mu}\left\|x_{k+1} - y_{k+1}\right\|^2 + \left\langle x_{k+1} - y_{k+1}, y_{k+1} - x_k\right\rangle, \\ -\lambda_{k+1} v_{k+1} &\geq \bar{\mu}\left\|x_{k+1} - y_{k+1}\right\|^2 + \left\langle x_{k+1} - y_{k+1}, y_{k+1} - x^*\right\rangle, \end{aligned} \tag{35}$$

where $\left\{v_k\right\}$ is defined in (20). Multiplying the first inequality above by $\left(t_{k+1} - 1\right)$ and adding it to the second inequality, we have

$$\begin{aligned} &\lambda_{k+1}\left(\left(t_{k+1} - 1\right)v_k - t_{k+1} v_{k+1}\right) \\ &\quad \geq \bar{\mu} t_{k+1}\left\|x_{k+1} - y_{k+1}\right\|^2 + \left\langle x_{k+1} - y_{k+1}, t_{k+1} y_{k+1} - \left(t_{k+1} - 1\right)x_k - x^*\right\rangle. \end{aligned} \tag{36}$$

Further, multiplying (37) by t_{k+1}, we obtain

$$\begin{aligned} &\lambda_k t_k^2 v_k - \lambda_{k+1} t_{k+1}^2 v_{k+1} - \left(\lambda_k t_k^2 - \lambda_{k+1} t_{k+1}\left(t_{k+1} - 1\right)\right)v_k \\ &\geq \bar{\mu}\left\|t_{k+1}\left(x_{k+1} - y_{k+1}\right)\right\|^2 + \left\langle t_{k+1}\left(x_{k+1} - y_{k+1}\right), t_{k+1} y_{k+1} - \left(t_{k+1} - 1\right)x_k - x^*\right\rangle \\ &= \tfrac{1}{2}\left\|t_{k+1}\left(x_{k+1} - y_{k+1}\right)\right\|^2 + \left\langle t_{k+1}\left(x_{k+1} - y_{k+1}\right), t_{k+1} y_{k+1} - \left(t_{k+1} - 1\right)x_k - x^*\right\rangle \\ &\qquad + \left(\bar{\mu} - \tfrac{1}{2}\right)\left\|t_{k+1}\left(x_{k+1} - y_{k+1}\right)\right\|^2 \\ &= \frac{1}{2}\left(\|u_{k+1}\|^2 - \|u_k\|^2\right) + \left(\bar{\mu} - \frac{1}{2}\right)\left\|t_{k+1}\left(x_{k+1} - y_{k+1}\right)\right\|^2. \end{aligned} \tag{37}$$

□

3.1 FISTA algorithm with the adaptive non-monotone stepsize

In this subsection, we consider the parameter

$$t_1 = 1 \text{ and } t_{k+1} = \frac{1 + \sqrt{1 + 4\left(\lambda_k / \lambda_{k+1}\right)t_k^2}}{2} \tag{38}$$

in the setting of Algorithm 4, which will be called the FISTA algorithm with the adaptive non-monotone stepsize (FISTA_NMS).

Since

$$t_{k+1} - t_k = \frac{1 + \sqrt{1 + 4\frac{\lambda_k}{\lambda_{k+1}} t_k^2}}{2} - t_k$$
$$= \frac{-4\left(1 - \frac{\lambda_k}{\lambda_{k+1}}\right) t_k^2 + 4t_k}{2\left(\sqrt{1 + 4\frac{\lambda_k}{\lambda_{k+1}} t_k^2} + 2t_k - 1\right)} \geq \frac{-4t_k + 4t_k}{2\left(\sqrt{1 + 4\frac{\lambda_k}{\lambda_{k+1}} t_k^2} + 2t_k - 1\right)} = 0,$$

it's easy to show that

$$\gamma_k = \frac{t_k - 1}{t_{k+1}} \leq \frac{t_k - 1}{t_k} = 1 - \frac{1}{t_k} \leq 1. \tag{39}$$

Hence, (39) is an appropriate parameter option for Algorithm 4. In the following, we firstly prove a trivial fact about this $\{t_k\}$.

Lemma 3.4 *Let* $\{t_k\}$ *be generated by FISTA_NMS. Then, we obtain that* $1/t_k = \Theta(1/k)$.

Proof Rearranging the expression of t_{k+1}, we have $2\sqrt{\lambda_{k+1}} t_{k+1} = \sqrt{\lambda_{k+1}} + \sqrt{\lambda_{k+1} + 4\lambda_k t_k^2}$. Denote that $w_k = \sqrt{\lambda_k} t_k$, then

$$2w_{k+1} = \sqrt{\lambda_{k+1}} + \sqrt{\lambda_{k+1} + 4w_k^2}, \tag{40}$$

it is easy to get that $\{w_k\}$ increasing monotonically.

The following proves that $\lim_{k\to\infty} w_k = +\infty$. Suppose that $\lim_{k\to\infty} w_k = w < +\infty$. Using Lemmas 2.1 and 2.2, denoted $\lim_{k\to\infty} \lambda_k = \lambda^* > 0$, we have $2w = \sqrt{\lambda^*} + \sqrt{\lambda^* + 4w^2}$, which implies a contradiction that $4w^2 - 4w\sqrt{\lambda^*} = 4w^2$. Therefore, $\lim_{k\to\infty} w_k = +\infty$.

Using the Stolz theorem, we deduce

$$\begin{aligned}\lim_{k\to\infty}\frac{t_k}{k}&=\lim_{k\to\infty}\frac{w_k}{\sqrt{\lambda_k}k}=\frac{1}{\sqrt{\lambda^*}}\lim_{k\to\infty}\left(w_{k+1}-w_k\right)\\&\overset{(41)}{=}\frac{1}{\sqrt{\lambda^*}}\lim_{k\to\infty}\frac{1}{2}\left(\sqrt{\lambda_{k+1}}+\sqrt{\lambda_{k+1}+4w_k^2}\right)-w_k\\&=\frac{1}{\sqrt{\lambda^*}}\lim_{k\to\infty}\frac{1}{2}\left(\sqrt{\lambda_{k+1}+4w_k^2}-\left(2w_k-\sqrt{\lambda_{k+1}}\right)\right)\\&=\frac{1}{\sqrt{\lambda^*}}\lim_{k\to\infty}\frac{1}{2}\left(\frac{4w_k\sqrt{\lambda_{k+1}}}{\sqrt{\lambda_{k+1}+4w_k^2}+\left(2w_k-\sqrt{\lambda_{k+1}}\right)}\right)\\&=\frac{1}{\sqrt{\lambda^*}}\lim_{k\to\infty}\left(\frac{2\sqrt{\lambda_{k+1}}}{\sqrt{\lambda_{k+1}/w_k^2+4}+2-\sqrt{\lambda_{k+1}/w_k^2}}\right)=\frac{1}{2}.\end{aligned}\tag{41}$$

Hence, we have $1/t_k=\Theta\left(1/k\right)$. □

Next, we will show that FISTA_NMS enjoys the $O\left(1/k^2\right)$ convergence rate of the objective function values and $o\left(\frac{1}{k}\right)$ convergence rate of the norm of subdifferential of function value.

Theorem 3.1 (Convergence rate) *Let* $\{x_k\},\{y_k\}$ *be generated by FISTA_NMS. Then,*

(*a*) $$F\left(x_k\right)-F(x^*)\le O\left(1/k^2\right),\quad \forall x^*\in X^* \text{ and } \forall k\ge 1.\tag{42}$$

(*b*) *The series* $\sum_{k=1}^{\infty}k^2\left\|x_k-y_k\right\|^2$ *is convergent and* $\liminf_{k\to\infty}k^{1.5}\left\|x_k-y_k\right\|=0$.

Proof Define the quantities $a_k=\lambda_k t_k^2 v_k, b_k=\frac{1}{2}\left\|u_k\right\|^2$, where v_k was defined in (20) and u_k was defined in the statement of Lemma 3.3. Since $t_{k+1}=\frac{1+\sqrt{1+4\left(\lambda_k/\lambda_{k+1}\right)t_k^2}}{2}$, one have that $\rho_k=\lambda_k t_k^2-\lambda_{k+1}\left(t_{k+1}^2-t_{k+1}\right)=0$. Then, the inequality in Lemma 3.3 can be rewritten as

$$a_k-a_{k+1}\ge\left(b_{k+1}-b_k\right)+\left(\bar{\mu}-\frac{1}{2}\right)\left\|t_{k+1}\left(x_{k+1}-y_{k+1}\right)\right\|^2\ge b_{k+1}-b_k,\quad\forall k\ge k_0.\tag{43}$$

It is not difficult to show that there exists a constant $c>0$ such that

$$a_k+b_k\le a_{k_0}+b_{k_0}\le c,\tag{44}$$

which implies that $\lambda_k t_k^2 v_k\le c$. Applying (10), we have

$$v_k = F(x_k) - F(x^*) \le \frac{c}{\lambda_k t_k^2} \le \frac{c}{\lambda_{\min} t_k^2}, \tag{45}$$

then, Lemma 3.4 yields the result that $F(x_k) - F(x^*) \le O(1/k^2)$.

Rearranging (44) we see that

$$(a_k + b_k) - (a_{k+1} + b_{k+1}) \ge \left(\bar{\mu} - \frac{1}{2}\right) \left\| t_{k+1}(x_{k+1} - y_{k+1}) \right\|^2, \quad \forall k \ge k_0.$$

Summing the inequality from $k = k_0$ to $k = N$, we obtain that

$$(a_{k_0} + b_{k_0}) - (a_{N+1} + b_{N+1}) \ge \left(\bar{\mu} - \frac{1}{2}\right) \sum_{k=k_0}^{N} \left\| t_{k+1}(x_{k+1} - y_{k+1}) \right\|^2, \tag{46}$$

where $\bar{\mu} - \frac{1}{2} > 0$ based on the choice of μ_0. Then, from $a_k + b_k > 0$ and Lemma 3.4 we have $\sum_{k=1}^{\infty} k^2 \|x_k - y_k\|^2$ is convergent.

Further, we can obtain that $\liminf_{k\to\infty} k^{1.5} \|x_k - y_k\| = 0$. Suppose that there exists a constant c such that $\liminf_{k\to\infty} k^3 \|x_k - y_k\|^2 = c > 0$, then for k is sufficiently large, we have $k^3 \|x_k - y_k\|^2 > \frac{c}{2}$, which means that $k^2 \|x_k - y_k\|^2 > \frac{c}{2k}$. Summing the inequality from $k = 1$ to $k = \infty$, we obtain that $\sum_{k=1}^{+\infty} k^2 \|x_k - y_k\|^2 > \sum_{k=1}^{+\infty} \frac{c}{2k}$, which is a contradiction since the left side of the inequality is a convergent series, but the right side is a divergent series. □

Remark 3.2 Denote $\psi_k = \nabla f(x_k) - \frac{1}{\lambda_k}(x_k - y_k + \lambda_k \nabla f(y_k))$. Based on Lemma 3.1, we have $\psi_k \in \partial F(x_k)$. It follows from Lemmas 2.1, 3.4, the conclusion $\lim_{k\to\infty} k^2 \|x_k - y_k\|^2 = 0$ and the fact that $\|\psi_k\| \le (L_f + 1/\lambda_{\min}) \|x_k - y_k\|$ that $\lim_{k\to\infty} k \|\psi_k\| = 0$, which implies that $\|\psi_k\| = o\left(\frac{1}{k}\right)$. However, for the FISTA, we deduce that $t_k^2 \|x_k - y_k\|^2$ is bounded from the proof of Lemma 4.1 in [5], i.e. $\|\psi_k\| = O\left(\frac{1}{k}\right)$. Hence, it seems that the sequence $\{\|\psi_k\|\}$ generated by FISTA_NMS converges to zero faster than the one generated by FISTA. In Sect. 5, the numerical performances can verify this.

In following theorem, we will show that the sequence $\{x_k\}$ has at least one accumulation point, and any accumulation point belongs to X^*.

Theorem 3.2 *For $\forall k \geq 1$, we have the sequence $\{x_k\}$ generated from FISTA_NMS is bounded, and all the accumulation points of $\{x_k\}$ belong to X^*.*

Proof From (45), we have that $b_k \leq c$ for any $k \geq k_0$.

With the definition of b_k and triangle inequality, we see that

$$\|x_k - x^*\| \leq \frac{\sqrt{2b_k}}{t_k} + \left(1 - \frac{1}{t_k}\right)\|x_{k-1} - x^*\| \leq \frac{\sqrt{2c}}{t_k} + \left(1 - \frac{1}{t_k}\right)\|x_{k-1} - x^*\|. \tag{47}$$

Let $M_0 = \max\left(2c, \left\|x_{k_0} - x^*\right\|\right)$. Then, we can easily prove that $\|x_k - x^*\| \leq M_0$ by induction, which implies $\{x_k\}$ is bounded. Assume that $\left\{x_{k_j}\right\}$ is a convergent subsequence of $\{x_k\}$ and $\lim_{j\to\infty} x_{k_j} = \bar{x}$.

In view of (43) and F is lower semi-continuous, we see that

$$F(\bar{x}) \leq \liminf_{j\to\infty} F\left(x_{k_j}\right) = \lim_{j\to\infty} F\left(x_{k_j}\right) = F(x^*). \tag{48}$$

Combining this with the fact that $F(\bar{x}) \geq F(x^*)$, we have $F(\bar{x}) = F(x^*)$, which means that $\bar{x} \in X^*$. □

3.2 Modified FISTA algorithm with the adaptive non-monotone stepsize

In the previous subsection, for the FISTA Algorithm with adaptive non-monontone stepsize strategy (FISTA_NMS), we proved that convergence rate of the function values remains $O\left(\frac{1}{k^2}\right)$, however, same as the original FISTA algorithm, the convergence of iterates generated by FISTA_NMS is still unknown. As mentioned in Sect. 1, Chambolle and Dossal [12] exploited a new $\gamma_k = \frac{t_k - 1}{t_{k+1}}$ with $t_k = \frac{k+a-1}{a}$, $a > 2$ for original FISTA, and established the convergence of the iterates generated by FISTA with this new parameter γ_k and a constant stepsize $\lambda_k \equiv \frac{1}{L_f}$ (MFISTA). Attouch and Peypouquet [1] proved that the convergence rate of function values of MFISTA is actually $o\left(\frac{1}{k^2}\right)$, better than $O\left(\frac{1}{k^2}\right)$ of FISTA. Moreover, MFISTA has a better numerical performance than FISTA.

In order to establish the convergence of iterates generated by FISTA_NMS, and accelerate its convergence speed, we consider the parameter

$$t_k = \frac{k + a - 1}{a},\ a > 2, \tag{49}$$

which satisfies that $\gamma_k = (k-1)/(k+a) \leq 1$. Hence, (50) is a good parameter option for Algorithm 4, which will be called the modified FISTA algorithm with the new adaptive non-monotone stepsize (MFISTA_NMS).

From Theorem 3.1, we can see that if the sequence $\{t_k\}$ satisfies

$$\rho_k = \lambda_k t_k^2 - \lambda_{k+1}\left(t_{k+1}^2 - t_{k+1}\right) \geq 0, \tag{50}$$

where $t_1 = 1$ and $\{\lambda_k\}$ is generated by the Algorithm 3, then the objective function values generated by Algorithm 4 has the $O\left(1/k^2\right)$ convergence rate. Particularly, $\rho_k = 0$ for $t_{k+1} = \frac{1+\sqrt{1+4\left(\lambda_k/\lambda_{k+1}\right)t_k^2}}{2}$. The following result is based on the analysis of ρ_k.

Lemma 3.5 *Let $\{x_k, y_k\}$ be the sequences generated via MFISTA_NMS. Assume that $\sum_{k=1}^{\infty} E_k$ is a convergent nonnegative series and $\{E_k\}$ is decreasing monotonically. We obtain the following conclusions.*

(a) *The series $\sum_{k=1}^{\infty} k\left(F\left(x_k\right) - F(x^*)\right)$ is convergent.*

(b) *The series $\sum_{k=1}^{\infty} k^2\left\|x_k - y_k\right\|^2$ is convergent and $\liminf_{k\to\infty} k^{1.5}\left\|x_k - y_k\right\| = 0$.*

Proof Based on the assumption that $\sum_{k=1}^{\infty} E_k$ is a convergent nonnegative series and $\{E_k\}$ is decreasing monotonically, we can easily obtain that $\lim_{k\to\infty} kE_k = 0$. Then, for any $k > \hat{k}$, the following equality

$$\begin{aligned}
\rho_{k-1} &= \frac{1}{a^2}\left(\lambda_{k-1}(k+a-2)^2 - \lambda_k(k-1)(k+a-1)\right)\\
&= \frac{1}{a^2}\left(\lambda_{k-1}(k+a-2)^2 - \lambda_k\left((k+a-2)^2 + (2-a)(k+a-2) + 1 - a\right)\right)\\
&= \frac{1}{a^2}\left(\left(\lambda_{k-1} - \lambda_k\right)(k+a-2)^2 + \lambda_k((a-2)(k+a-2) + a - 1)\right)\\
&= \frac{1}{a^2}\left(-\left|\lambda_{k-1} - \lambda_k\right|(k+a-2)^2 + \lambda_k((a-2)(k+a-2) + a - 1)\right)\\
&= \frac{1}{a^2}\left(-\lambda_{k-1}\cdot E_{k-1}\cdot(k+a-2)^2 + \lambda_k((a-2)(k+a-2) + a - 1)\right)
\end{aligned} \tag{51}$$

yields that $\lim_{k\to\infty} \frac{\rho_k}{k} = \frac{a-2}{a^2}\lambda^* \geq \omega_3$, where $\omega_3 = \frac{(a-2)}{a^2}\lambda_{\min}$. Note the fourth equality follows from the fact that the stepsize $\{\lambda_k\}$ increases monotonically after $\hat{k}$ step.

Invoking the inequality in Lemma 3.3 and combining $\lim_{k\to\infty} \frac{\rho_k}{k} \geq \omega_3$, then, for all k is sufficiently large, we have $\frac{\rho_k}{k} \geq \frac{\omega_3}{2}$ and

$$\begin{aligned}
&\lambda_k t_k^2 v_k - \lambda_{k+1} t_{k+1}^2 v_{k+1}\\
&\quad\geq \frac{1}{2}\left(\left\|u_{k+1}\right\|^2 - \left\|u_k\right\|^2\right) + \left(\bar{\mu} - \frac{1}{2}\right)\left\|t_{k+1}\left(x_{k+1} - y_{k+1}\right)\right\|^2 + \frac{\omega_3}{2}kv_k,
\end{aligned} \tag{52}$$

where $u_k = t_k x_k - \left(t_k - 1\right)x_{k-1} - x^*$ and v_k is defined in (20). We rearrange (52) into

$$\left(\lambda_k t_k^2 v_k + \frac{1}{2}\|u_k\|^2\right) - \left(\lambda_{k+1} t_{k+1}^2 v_{k+1} + \frac{1}{2}\|u_{k+1}\|^2\right) \geq \frac{\omega_3}{2} k v_k,$$

and

$$\left(\lambda_k t_k^2 v_k + \tfrac{1}{2}\|u_k\|^2\right) - \left(\lambda_{k+1} t_{k+1}^2 v_{k+1} + \tfrac{1}{2}\|u_{k+1}\|^2\right) \geq \left(\bar{\mu} - \tfrac{1}{2}\right)\left\|t_{k+1}\left(x_{k+1} - y_{k+1}\right)\right\|^2.$$

Summing the above two inequalities from $k = N_s$ to $k = N$, where N_s is sufficiently large, yields that $\sum_{k=1}^{\infty} k\left(F\left(x_k\right) - F(x^*)\right)$ and $\sum_{k=1}^{\infty} t_k^2\|x_k - y_k\|^2$ are convergent. With the definition of $t_k = \frac{k+a-1}{a}(a > 2)$, we can further obtain that $\sum_{k=1}^{\infty} k^2\|x_k - y_k\|^2$ is convergent. In addition, we can also show that $\liminf_{k\to\infty} k^{1.5}\|x_k - y_k\| = 0$. □

Remark 3.3 Recalling that $\partial F\left(x_k\right) \ni \psi_k = \nabla f\left(x_k\right) - \frac{1}{\lambda_k}\left(x_k - y_k + \lambda_k \nabla f\left(y_k\right)\right)$. It follows that $\sum_{k=1}^{\infty} k^2\left\|\psi_k\right\|^2$ is convergent from Lemma 3.5 (*b*), which is stronger than the fact in [1] that $\left\|\psi_k\right\| = o\left(\frac{1}{k}\right)$ for MFISTA.

Lemma 3.6 *Let $\{x_k\}$ be generated by MFISTA_NMS. Then, the series $\sum_{k=1}^{\infty} k\delta_k$ is convergent, where δ_k is defined in (21).*

Proof From (34) with $x := x_k, y := y_{k+1}$, we have that

$$\frac{\delta_{k+1}}{\lambda_{k+1}} - \gamma_k^2 \frac{\delta_k}{\lambda_k} \leq v_k - v_{k+1}, \forall k \geq k_0, \tag{53}$$

where $\gamma_k = (k-1)/(k+a)$, $v_k := F\left(x_k\right) - F(x^*)$ is defined in (20) and $\delta_k := \frac{1}{2}\|x_k - x_{k-1}\|^2$ is defined in (21).

Multiplying this inequality by $(k+a)^2$ and summing from $k = N_s$ to $k = N$, where N_s is sufficiently large, leads to

$$\begin{aligned}
&-\left(N_s - 1\right)^2 \frac{\delta_{N_s}}{\lambda_{N_s}} + \left[\left(N_s + a\right)^2 - \left(N_s + 1 - 1\right)^2\right] \frac{\delta_{N_s+1}}{\lambda_{N_s+1}} \\
&\quad + \left[\left(N_s + 1 + a\right)^2 - \left(N_s + 2 - 1\right)^2\right] \frac{\delta_{N_s+2}}{\lambda_{N_s+2}} \\
&\quad + \cdots + \left[(N - 1 + a)^2 - (N-1)^2\right] \frac{\delta_N}{\lambda_N} + (N+a)^2 \frac{\delta_{N+1}}{\lambda_{N+1}} \\
&\leq \left(N_s + a\right)^2 v_{N_s} + \left[\left(N_s + 1 + a\right)^2 - \left(N_s + a\right)^2\right] v_{N_s+1} \\
&\quad + \left[\left(N_s + 2 + a\right)^2 - \left(N_s + 1 + a\right)^2\right] v_{N_s+2} \\
&\quad + \cdots + \left[(N+a)^2 - (N - 1 + a)^2\right] v_N - (N+a)^2 v_{N+1},
\end{aligned} \tag{54}$$

i.e.,

$$\begin{aligned}(N+a)^2\frac{\delta_{N+1}}{\lambda_{N+1}}-\left(N_s-1\right)^2\frac{\delta_{N_s}}{\lambda_{N_s}}+\sum_{k=N_s+1}^{N}a(2k-2+a)\frac{\delta_k}{\lambda_k}\\ \leq\left(N_s+a\right)^2v_{N_s}-(N+a)^2v_{N+1}+\sum_{k=N_s+1}^{N}(2k+2a-1)v_k.\end{aligned}\tag{55}$$

From Lemma 3.5 (a) and $a>2$, the series $\sum_{k=1}^{\infty}k\frac{\delta_k}{\lambda_k}$ is convergent. Further, we obtain that $\sum_{k=1}^{\infty}k\delta_k$ is convergent by using $\lim_{k\to\infty}\lambda_k=\lambda^*>0$. □

Next, we construct the convergence rate of function values generated by MFISTA_NMS.

Theorem 3.3 *For the sequence* $\{x_k\}$ *generated by MFISTA_NMS, we have* $F(x_k)-F(x^*)=o\left(\frac{1}{k^2}\right)$ and $\|x_k-x_{k-1}\|=o\left(\frac{1}{k}\right)$.

Proof Recalling the definitions of $\{v_k\}$ in (20) and $\{\delta_k\}$ in (21). Denote $\phi_k=v_k+\frac{\delta_k}{\lambda_k}$.

From (53) and $\gamma_k\leq 1$, we can easily deduce that $\phi_{k+1}\leq\phi_k$ for $k\geq k_0$. Multiplying by $(k+1)^2$ we have

$$(k+1)^2\phi_{k+1}\leq(k+1)^2\phi_k=k^2\phi_k+(2k+1)\phi_k.\tag{56}$$

Then, we have

$$\left((k+1)^2\phi_{k+1}-k^2\phi_k\right)^+\leq(2k+1)\phi_k.\tag{57}$$

Since $\sum_{k=1}^{\infty}k\phi_k$ is convergent from Lemma 3.5 (a) and Lemma 3.6, we get that

$$\sum\nolimits_{k=1}^{+\infty}\left((k+1)^2\phi_{k+1}-k^2\phi_k\right)^+<+\infty.\tag{58}$$

We also can prove that $\sum_{k=1}^{+\infty}\left((k+1)^2\phi_{k+1}-k^2\phi_k\right)^-<+\infty$. Otherwise, based on (58) and the equality

$$\begin{aligned}&(k+1)^2\phi_{k+1}-\phi_1\\ &=\sum\nolimits_{i=1}^{k}\left((i+1)^2\phi_{i+1}-i^2\phi_i\right)^+-\sum\nolimits_{i=1}^{k}\left((i+1)^2\phi_{i+1}-i^2\phi_i\right)^-,\end{aligned}\tag{59}$$

we can get that $\lim_{k\to\infty}(k+1)^2\phi_{k+1}=-\infty$, which is a contradiction with the fact that $k^2\phi_k\geq 0$. Hence, in view of (59), we have that $\{k^2\phi_k\}$ is convergent. Further, we have $\liminf_{k\to\infty}k^2\phi_k=0$ by using the convergence of $\sum_{k=1}^{\infty}k\phi_k$. Hence, $\lim_{k\to\infty}k^2\phi_k=0$, which implies $F(x_k)-F(x^*)=o\left(\frac{1}{k^2}\right)$ and $\|x_k-x_{k-1}\|=o\left(\frac{1}{k}\right)$ by Lemma 2.1. □

Now, we give the proof of convergence of the sequence $\{x_k\}$ generated by MFISTA_NMS. Before that, we give some auxiliary results.

Lemma 3.7 *For $\forall k \geq 1$, we have the sequence $\{x_k\}$ generated from MFISTA_NMS is bounded, and all the accumulation points of $\{x_k\}$ belong to X^*.*

Proof The proof is similar to the Theorem 3.2.

Lemma 3.8 *For any $x^* \in X^*$, and the sequence $\{x_k\}$ generated by MFISTA_NMS, we have $\boldsymbol{\Phi}_k = \frac{1}{2}\|x_k - x^*\|^2$ is convergent.*

Proof Recalling (52) in the proof of Lemma 3.5, we have $a_k + b_k \geq a_{k+1} + b_{k+1}$ for all $k \geq k_0$, where $a_k := \lambda_k t_k^2 v_k$ and $b_k := \frac{1}{2}\left\|(t_k - 1)(x_k - x_{k-1}) + (x_k - x^*)\right\|^2$. Combining this and the fact that $a_k + b_k \geq 0$, we can easily deduce that the sequence $\{a_k + b_k\}$ is convergent.

With Lemma 2.1, the definition of t_k in (50) and $\lim_{k\to\infty} k^2\big(F(x_k) - F(x^*)\big) = 0$ from Theorem 3.3, we obtain that $\lim_{k\to\infty} a_k = 0$, which implies that $\{b_k\}$ is convergent. From the definition of $\{b_k\}$, we see that

$$\begin{aligned} b_k &= \frac{1}{2}\left\|(t_k - 1)(x_k - x_{k-1}) + (x_k - x^*)\right\|^2 \\ &= \frac{1}{2}(t_k - 1)^2\|x_k - x_{k-1}\|^2 + \left\langle (t_k - 1)(x_k - x_{k-1}), (x_k - x^*)\right\rangle + \frac{1}{2}\|x_k - x^*\|^2. \end{aligned} \tag{60}$$

It follows from (50) and Theorem 3.3 that the first item of (60) converges to zero, i.e.,

$$\lim_{k\to\infty} \frac{1}{2}(t_k - 1)^2\|x_k - x_{k-1}\|^2 = 0. \tag{61}$$

In addition, from (50), Theorem 3.3 and the fact from Lemma 3.7 that $\|x_k - x^*\|$ is bounded, we have

$$\lim_{k\to\infty} \left\langle (t_k - 1)(x_k - x_{k-1}), (x_k - x^*)\right\rangle = 0. \tag{62}$$

With (60), (61), (62) and the fact that $\{b_k\}$ is convergent, we can obtain that $\boldsymbol{\Phi}_k = \frac{1}{2}\|x_k - x^*\|^2$ is convergent.

Theorem 3.4 *The sequence $\{x_k\}$ generated by MFISTA_NMS converges to a minimizer of F.*

Proof From Lemma 3.7, we have $\lim_{j\to\infty} x_{k_j} = \bar{x} \in X^*$. And with the result of Lemma 3.8, we have $\|x_k - \bar{x}\|^2$ is convergent by using $\bar{x}$ to replace x^*. Then one can easily deduce that the sequence $\left\{\|x_k - \bar{x}\|^2\right\}$ converges to zero, which implies $\{x_k\}$ converges to a minimizer of F. □

It is worth mentioning that the biggest difference between MFISTA_NMS and MFISTA is that MFISTA_NMS is not required to do any assumption relating to the L_f, while the condition $\lambda \in]0, \frac{1}{L_f}]$ in MFISTA plays an important role in algorithm implementation and theoretical analysis.

Meanwhile, it is unclear that whether the iterative sequence generated by MFISTA with the backtracking stepsize converges. However, MFISTA_NMS keeps the similar theoretical results with MFISTA and guarantees in addition the convergence rate of objective function values and the convergence of iterative sequence.

3.3 The convergence results for FISTA_NMS and MFISTA_NMS under the error bound condition

In the above analysis, we proved that for the FISTA_NMS and MFISTA_NMS, function values keep similar convergence rates with FISTA and MFISTA, however, the convergence of the iterates generated by FISTA or FISTA_NMS has not been established so far. Note that the convergence of the iterates generated by FISTA is a puzzling question in the study of numerical optimization methods, meanwhile, the linear convergence rate of function values always be expected, but there is no effective way to prove it. Recently, in [19], under the error bound condition, the authors proposed a *comparison method* to prove the convergence of iterates generated by FISTA and $o\left(\frac{1}{k^6}\right)$ rate of convergence for the function values. Moreover, the convergence rate of function values generated by MFISTA can be improved to $o\left(\frac{1}{k^{2(1+a)}}\right)$.

In this subsection, based on the idea of the proof in [19], we will derive that FISTA_NMS and MFISTA_NMS enjoy similar results (See Corollary 3.1 and Corollary 3.2). It is worth emphasizing that the main difference between our setting and the setting of [19] is that we use the adaptive non-monotone stepsize strategy but a constant stepsize was used in [19]. Although there are large overlaps with Theorem 2.6, Corollary 3.5 and Corollary 3.6 in [19], there are some subtle differences. Here, we present the detailed proofs for self-containedness and the convenience of the readers.

Now, we recall the error bound condition which is a key ingredient in proving convergence of iterative methods.

Assumption 3.1 (*Error bound condition*) For any $\xi \geq F^* := \inf_{x \in R^n} F(x)$, there exist $\varepsilon > 0$ and $\bar{\tau} > 0$ such that

$$\operatorname{dist}(x, X^*) \leq \bar{\tau} \left\| p_{\frac{1}{L_f} g}(x) - x \right\| \tag{63}$$

whenever $\left\| p_{\frac{1}{L_f} g}(x) - x \right\| < \varepsilon$ and $F(x) \leq \xi$.

Major contributions on developing and using error bound condition to derive convergence rate of iterative descent algorithms have been developed in a series of papers [4, 15, 16, 28, 30].

Lemma 3.9 [22, Lemma 2] *For $\lambda_1 \geq \lambda_2 > 0$, we have*

$$\left\|p_{\lambda_1 g}(x) - x\right\| \geq \left\|p_{\lambda_2 g}(x) - x\right\| \quad \text{and} \quad \frac{\left\|p_{\lambda_1 g}(x) - x\right\|}{\lambda_1} \leq \frac{\left\|p_{\lambda_2 g}(x) - x\right\|}{\lambda_2}, \tag{64}$$

where $p_{\lambda g}(\cdot)$ is defined in (3).

Lemma 3.10 *Suppose that Assumption 3.1 holds. Let $\{x_k\}$ be generated by Algorithm 4 and $x^* \in X^*$. Then, for k is sufficiently large, there exists $\tau_1 > 0$ such that*

$$F(x_{k+1}) - F(x^*) \leq \tau_1 \|y_{k+1} - x_{k+1}\|^2.$$

Proof Setting $\xi_0 = F(x_{k_0+1}) + \frac{1}{2\lambda_{k_0+1}}\left\|x_{k_0+1} - x_{k_0}\right\|^2$. From Lemma 2.1, we obtain that $\frac{1}{L_f} \geq \frac{\lambda_{\min}}{\mu_1}, \frac{\lambda_{\min}}{\mu_1} \leq \frac{\lambda_k}{\mu_1}$ and $\frac{\lambda_k}{\mu_1} \geq \lambda_k$, then, combining with Lemma 3.9, the nonexpansiveness property of the proximal operator and the fact that ∇f is Lipschitz continuous, we obtain that

$$\begin{aligned}
\left\|p_{\frac{1}{L_f}g}(x_k) - x_k\right\| &\leq \frac{\mu_1}{L_f\lambda_{\min}}\left\|p_{\frac{\lambda_{\min}}{\mu_1}g}(x_k) - x_k\right\| \leq \frac{\mu_1}{L_f\lambda_{\min}}\left\|p_{\frac{\lambda_k}{\mu_1}g}(x_k) - x_k\right\| \\
&\leq \frac{1}{L_f\lambda_{\min}}\left\|p_{\lambda_k g}(x_k) - x_k\right\| = \frac{1}{L_f\lambda_{\min}}\left\|p_{\lambda_k g}(x_k) - p_{\lambda_k g}(y_k)\right\| \\
&\leq \frac{1}{L_f\lambda_{\min}}(1 + \lambda_k L_f)\|x_k - y_k\|.
\end{aligned} \tag{65}$$

Applying the inequality (34) at the point $x := x_k$, $y := y_{k+1}$ and $\lambda := \lambda_{k+1}$, we obtain that for $k \geq k_0$,

$$\begin{aligned}
&\frac{2\bar{\mu} - 1}{2\lambda_{k+1}}\|x_{k+1} - y_{k+1}\|^2 \\
&\leq \left(F(x_k) + \frac{\gamma_k^2}{2\lambda_{k+1}}\|x_k - x_{k-1}\|^2\right) - \left(F(x_{k+1}) + \frac{1}{2\lambda_{k+1}}\|x_{k+1} - x_k\|^2\right) \\
&\leq \left(F(x_k) + \frac{1}{2\lambda_k}\|x_k - x_{k-1}\|^2\right) - \left(F(x_{k+1}) + \frac{1}{2\lambda_{k+1}}\|x_{k+1} - x_k\|^2\right),
\end{aligned} \tag{66}$$

where the second inequality follows the facts that $\gamma_k \leq 1$ and $\lambda_k \leq \lambda_{k+1}$, the latter relation was proved in Lemma 2.2.

It follows from (66) that for $k \geq k_0$, $\left\{F(x_{k+1}) + \frac{1}{2\lambda}\|x_{k+1} - x_k\|^2\right\}$ is non-increasing, and $\sum_{k=k_0}^{+\infty}\|x_{k+1} - y_{k+1}\|^2 < +\infty$ due to the convergence of $\{\lambda_k\}$. Then, we get

that for k is sufficiently large, $F(x_k) \leq \xi_0$ and there exists a $\varepsilon_0 > 0$ such that $\left\| p_{\frac{1}{L_f}g}(x_k) - x_k \right\| < \varepsilon_0$ by (65). Hence, combining with Assumption 3.1, we have for $\xi_0 = F(x_{k_0+1}) + \frac{1}{2\lambda_{k_0+1}} \left\| x_{k_0+1} - x_{k_0} \right\|^2$, there exists $\tau_0 > 0$, for k is sufficiently large, such that

$$\operatorname{dist}(x_k, X^*) \leq \tau_0 \left\| p_{\frac{1}{L_f}g}(x_k) - x_k \right\|. \tag{67}$$

In addition, applying the inequality (34) at the point $y := y_{k+1}$, $\lambda := \lambda_{k+1}$ with $x^*_{k+1} \in X^*$ satisfying $\operatorname{dist}(x_{k+1}, X^*) = \left\| x_{k+1} - x^*_{k+1} \right\|$, we obtain

$$\begin{aligned} F(x_{k+1}) - F(x^*) &\leq \frac{1}{2\lambda_{k+1}} \left\| y_{k+1} - x^*_{k+1} \right\|^2 - \frac{1}{2\lambda_{k+1}} \left\| x_{k+1} - x^*_{k+1} \right\|^2 \\ &= \frac{1}{2\lambda_{k+1}} \left\| y_{k+1} - x_{k+1} \right\|^2 + \frac{1}{\lambda_{k+1}} \left\langle y_{k+1} - x_{k+1}, x_{k+1} - x^*_{k+1} \right\rangle \\ &\leq \frac{1}{2\lambda_{k+1}} \left\| y_{k+1} - x_{k+1} \right\|^2 + \frac{1}{\lambda_{k+1}} \left\| y_{k+1} - x_{k+1} \right\| \operatorname{dist}(x_{k+1}, X^*). \end{aligned}$$

Then, combining (67), (65) and Corollary 2.1, we have for k is sufficiently large,

$$F(x_{k+1}) - F(x^*) \leq \tau_1 \left\| y_{k+1} - x_{k+1} \right\|^2, \tag{68}$$

where $\tau_1 = \frac{1}{\lambda_{\min}} \left(\frac{1}{2} + \frac{\tau_0}{\lambda_{\min} L_f} \left(1 + \lambda^* L_f \right) \right)$. □

Lemma 3.11 *Suppose that there exists a nonnegative sequence* $\{s_k\}$ *such that for* k *is sufficiently large,* $\alpha_k = \frac{s_k - 1}{s_{k+1}} \geq \gamma_k$, where $\gamma_k = \frac{t_k - 1}{t_{k+1}}$ and $\lim_{k \to +\infty} \gamma_k = 1$. Then, we have $\lim_{k \to \infty} s_k = +\infty$ and $\limsup_{k \to \infty} \frac{s^2_{k+1} - s^2_k}{s^2_k} \leq 0$.

Proof The proof is similar to Lemma 2.5 in [19]. □

The following theorem is largely similar to [19, Theorem 2.6], here, we still give a complete proof since our setting is equipped with the adaptive non-monotone stepsize, which is different from constant stepsize be used in [19].

Theorem 3.5 *Suppose that Assumption 3.1 holds and there exists a nonnegative sequence* $\{s_k\}$ *such that for* k *is sufficiently large,* $\alpha_k = \frac{s_k - 1}{s_{k+1}} \geq \gamma_k$, where $\gamma_k = \frac{t_k - 1}{t_{k+1}}$ and $\lim_{k \to +\infty} \gamma_k = 1$. Then, we have that $F(x_{k+1}) - F(x^*) = o\left(\frac{1}{s^2_{k+1}} \right)$ and $\left\| x_{k+1} - x_k \right\| = O\left(\frac{1}{s_{k+1}} \right)$.

Proof Denote that $E_k = s_{k+1}^2\big(F(x_k) - F(x^*)\big) + \frac{s_k^2}{2\lambda_k}\|x_k - x_{k-1}\|^2$. Applying inequality (34) at the point $x := x_k$, $y := y_{k+1}$ and $\lambda := \lambda_{k+1}$, we have for $k \geq k_0$,

$$F(x_{k+1}) - F(x^*) + \frac{2\bar{\mu}-1}{2\lambda_{k+1}}\|x_{k+1} - y_{k+1}\|^2 + \frac{1}{2\lambda_{k+1}}\|x_{k+1} - x_k\|^2 \leq F(x_k) - F(x^*) + \frac{\gamma_k^2}{2\lambda_k}\|x_k - x_{k-1}\|^2. \tag{69}$$

By the supposed condition, we have $\gamma_k^2 \leq \alpha_k^2$ for any $k \geq k_1$, where k_1 is sufficiently large, then,

$$F(x_{k+1}) - F(x^*) + \frac{2\bar{\mu}-1}{2\lambda_{k+1}}\|x_{k+1} - y_{k+1}\|^2 + \frac{1}{2\lambda_{k+1}}\|x_{k+1} - x_k\|^2 \leq F(x_k) - F(x^*) + \frac{\alpha_k^2}{2\lambda_k}\|x_k - x_{k-1}\|^2. \tag{70}$$

Multiplying by s_{k+1}^2, we have

$$\begin{aligned}
&E_{k+1} + \big(s_{k+1}^2 - s_{k+2}^2\big)\big(F(x_{k+1}) - F(x^*)\big) + \frac{2\bar{\mu}-1}{2\lambda_{k+1}}s_{k+1}^2\|x_{k+1} - y_{k+1}\|^2 \\
&\leq s_{k+1}^2\big(F(x_k) - F(x^*)\big) + \frac{(s_k-1)^2}{2\lambda_k}\|x_k - x_{k-1}\|^2 \\
&= s_{k+1}^2\big(F(x_k) - F(x^*)\big) + \frac{s_k^2}{2\lambda_k}\|x_k - x_{k-1}\|^2 \\
&\quad - \frac{2s_k - 1}{2\lambda_k}\|x_k - x_{k-1}\|^2 \leq E_k,\ \forall k \geq k_1.
\end{aligned}$$

Then, combining with the inequality in Lemma 3.10 and Corollary 2.1, we have

$$E_{k+1} + \left(\frac{\big(s_{k+1}^2 - s_{k+2}^2\big)}{s_{k+1}^2} + \frac{2\bar{\mu}-1}{2\tau_1\lambda_{\max}}\right)s_{k+1}^2\big(F(x_{k+1}) - F(x^*)\big) \leq E_k. \tag{71}$$

Since $\limsup_{k\to\infty} \frac{s_{k+2}^2 - s_{k+1}^2}{s_{k+1}^2} \leq 0$ from Lemma 3.11, we have $\frac{(s_{k+1}^2 - s_{k+2}^2)}{s_{k+1}^2} \geq -\frac{2\bar{\mu}-1}{4\tau_1\lambda_{\max}}$ for k is sufficiently large, then, (71) implies that, for k is sufficiently large,

$$E_{k+1} + \frac{2\bar{\mu}-1}{4\tau_1\lambda_{\max}}s_{k+1}^2\big(F(x_{k+1}) - F(x^*)\big) \leq E_k, \tag{72}$$

which means that $\{E_k\}$ is nonincreasing and convergent and $\sum_{k=1}^{\infty} s_{k+1}^2\big(F(x_{k+1}) - F(x^*)\big)$ is convergent. Hence, $F(x_{k+1}) - F(x^*) = o\left(\frac{1}{s_{k+1}^2}\right)$ holds ture.

Further, since the convergence of $\{E_k\}$, we have $\left\{s_{k+1}^2\|x_{k+1}-x_k\|^2\right\}$ is bounded, which means that $\|x_{k+1}-x_k\| \le O\left(\frac{1}{s_{k+1}}\right)$, i.e., there exists a constant $c_1 > 0$ such that for k is sufficiently large, $\|x_{k+1}-x_k\| \le \frac{c_1}{s_{k+1}}$. □

Corollary 3.1 *Suppose that Assumption 3.1 holds. Let $\{x_k\}$ be generated by FISTA_NMS and $x^* \in X^*$. Then,*

(1) $F(x_k) - F(x^*) = o\left(\frac{1}{k^6}\right)$ and $\|x_k - x_{k-1}\| = O\left(\frac{1}{k^3}\right)$.

(2) $\{x_k\}$ *converges to* $\bar{x} \in X^*$ at the $O\left(\frac{1}{k^2}\right)$ *rate of convergence.*

Proof Denote that $s_1 = s_2 = 1$ and $s_k = \frac{(k-1)^3}{\left(\int_1^{k-1}\frac{\ln x}{x^2}dx\right)^2}$, $\forall k \ge 3$. Based on the proof of Corollary 3.5 in [19], we have $\alpha_k \ge \gamma_k$ for k is sufficiently large and $s_k \sim k^3$. Then, By Theorem 3.5, we can get that $F(x_k) - F(x^*) = o\left(\frac{1}{k^6}\right)$ and there exists a $c' > 0$ such that for sufficiently large k, $\|x_{k+1}-x_k\| \le \frac{c'}{k^3}$. Then, we can deduce that

$$\begin{aligned}\forall p > 1, \quad \|x_{k+p}-x_k\| &\le \sum_{i=k+1}^{k+p}\|x_i - x_{i-1}\| \\ &\le c'\sum_{i=k+1}^{k+p}\frac{1}{i^3} \le c'\int_k^{k+p}\frac{1}{x^3}dx.\end{aligned}$$

Then,

$$\|x_k - \bar{x}\| \le \frac{c'}{2k^2}.$$

□

Corollary 3.2 *Suppose that Assumption 3.1 holds. Let $\{x_k\}$ be generated by MFISTA_NMS and $x^* \in X^*$. Then,*

(1) $F(x_k) - F(x^*) = o\left(\frac{1}{k^{2(a+1)}}\right)$ and $\|x_k - x_{k-1}\| = O\left(\frac{1}{k^{a+1}}\right)$.

(2) $\{x_k\}$ *converges to* $\bar{x} \in X^*$ *at the* $O\left(\frac{1}{k^a}\right)$ *rate of convergence.*

Proof For $a \ge 1$, denote $s_k = (k+a-1)^{a+1}$; otherwise, denote $s_1 = s_2 = 1$, and $s_k = \frac{(k+a-1)^{a+1}}{\int_1^{k-1}\frac{\ln x}{x^{1+a}}dx}$, $\forall k \ge 3$. Based on the proof of Corollary 3.6 in [19], we have $\gamma_k \le \alpha_k$ for k is sufficiently large and $s_k = O\left(k^{a+1}\right)$. Then, By Theorem 3.5, we can get that $F(x_k) - F(x^*) = o\left(\frac{1}{k^{2(a+1)}}\right)$ and there exists a $c'' > 0$ such that for sufficiently large k, $\|x_{k+1}-x_k\| \le \frac{c''}{k^{a+1}}$. Then, we can deduce that

$$\forall p > 1, \quad \left\|x_{k+p} - x_k\right\| \leq \sum_{i=k+1}^{k+p} \left\|x_i - x_{i-1}\right\| \leq c'' \sum_{i=k+1}^{k+p} \frac{1}{i^{a+1}} \leq c'' \int_k^{k+p} \frac{1}{x^{a+1}} dx.$$

Then,

$$\left\|x_k - \bar{x}\right\| \leq \frac{c''}{ak^a}.$$

□

It is noted that the stepsize λ_k generated by Algorithm 3 increases monotonically after finite iterations, while the stepsize λ_k generated by backtracking of FISTA_BKTR may increases or decreases in the backtracking process. Meanwhile, we cannot obtain a similar inequality with (34) based on FISTA_BKTR. Hence, using the same idea of proof in [19], backtracking of FISTA_BKTR can not obtain the results in Corollaries 3.1 and 3.2, and to our knowledge, there aren't similar results in the literature. From this point of view, FISTA with λ_k generated by the new stepsize strategy (Algorithm 3) enjoys better theoretical properties than Algorithm 2 (FISTA_BKTR).

To further illustrate this point, we consider a restart technique, which is crucially important in improving the theoretical results and accelerating the numerical performance of the algorithm, to improve our algorithms in next section.

4 Restart FISTA algorithm with the adaptive non-monotone stepsize strategy

O'Donoghue and Candès [23] introduced two simple heuristic adaptive restart techniques that can improve the convergence rate of accelerated gradient schemes. One restart technique is fixed restarting, that restarts the algorithm every K iterations and takes the last point generated by the algorithm as the starting point. Another is the adaptive restart, which restarts the algorithm based on the following schemes: 1) function scheme: $F(x_k) > F(x_{k-1})$; 2)gradient scheme: $(y_k - x_k)^T (x_k - x_{k-1}) > 0$.

O'Donoghue and Candès pointed out that both of the two adaptive restart schemes perform similarly well. But when the iteration point is close to the minimum, the algorithm with the gradient restart technique is more numerically stable. Therefore, we combine the fixed restarting with the gradient restart technique to improve the performance of FISTA_NMS and MFISTA_NMS in this section.

We present algorithms as follows.

Algorithm 5 FISTA_NMS_restart

Step 0. Given $K \in R$ and take $y_1 = x_0 \in R^n, t_1 = 1, 0 < \mu_1 < \mu_0 \leq 1$, $\lambda_1 > 0$ and $\tilde{k} = 1$
Step k. Compute

$$x_k = p_{\lambda_k g}(y_k), \text{ where } p_{\lambda g}(\cdot) \text{ is defined in (3)}$$

Set $\tilde{k} := \tilde{k} + 1$ and compute λ_{k+1} via Algorithm 3.
If $(y_k - x_k)^T (x_k - x_{k-1}) > 0$ or $\tilde{k} = K$ holds, set $t_k = 1, \tilde{k} = 1$

$$t_{k+1} = \left(1 + \sqrt{1 + 4(\lambda_k/\lambda_{k+1}) t_k^2}\right) \tag{72}$$

$$y_{k+1} = x_k + ((t_k - 1)/t_{k+1})(x_k - x_{k-1}).$$

Algorithm 6 MFISTA_NMS_restart

Step 0. Given $K \in R$ and take $y_1 = x_0 \in R^n, 0 < \mu_1 < \mu_0 \leq 1, a > 2, \lambda_1 > 0$ and $\tilde{k} = 1$.
Step k. Compute

$$x_k = p_{\lambda_k g}(y_k), \text{ where } p_{\lambda g}(\cdot) \text{ is defined in (3)}$$

Set $\tilde{k} := \tilde{k} + 1$
If $(y_k - x_k)^T (x_k - x_{k-1}) > 0$ or $\tilde{k} = K$ holds, set $\tilde{k} = 1$

$$y_{k+1} = x_k + \left(\frac{\tilde{k} - 1}{\tilde{k} + a}\right)(x_k - x_{k-1}) \tag{73}$$

Compute λ_{k+1} via Algorithm 3.

The schemes of FISTA_BKTR and MFISTA_BKTR combining the restart strategy separately, namely FISTA_BKTR_restart and MFISTA_BKTR_restart, are similar to the above two algorithms, here we omit the unnecessary details. In the following, we prove that under the error bound condition, the sequences generated by Algorithm 6 and Algorithm 7 are *R*-linearly convergent; Moreover, the corresponding sequences of objective function values are also *R*-linearly convergent. Note that whether the FISTA_BKTR with restart strategy enjoys similar convergence results is unknown.

Before proceeding with the convergence results, we give some auxiliary conclusions as follows.

Definition 4.1 [25] Let the iterative sequence $\{x_k\}$ generated by an algorithm converges to x^* in some norm. If there is a positive constant $\beta \in (0, 1)$ which is independent of the iterative number k, such that

$$\lim_{k \to \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = \beta,$$

then the sequence $\{x_k\}$ is said to be Q-linear convergence. If there is a positive constant $\alpha \in (0, 1)$ such that

$$\limsup_{k\to\infty} \left\|x_k - x^*\right\|^{\frac{1}{k}} = \alpha,$$

then the sequence $\{x_k\}$ is said to be R-linear convergence.

Lemma 4.1 [25] *The sequence $\{x_k\}$ is said to be R-linear convergence if there is a sequence of nonnegative scalars $\{q_k\}$* such that

$$\|x_k - x^*\| \le q_k \text{ for all } k, \text{ and } \{q_k\} \text{ converges } Q\text{–linearly to zero.}$$

Lemma 4.2 *Let $\{A_k\}, \{B_k\}$ and $\{C_k\}$ be three nonnegative sequences. Suppose that there exist $0 < \tau < 1, l > 0$* and $k_l > 0$ *such that $A_{k+1} + B_{k+1} + C_{k+1} \le A_k + \tau B_k$* and $A_k \le lC_k$ *hold for any $k > k_l$, we have $\{A_{k+1} + \alpha B_{k+1}\}$ converges Q-linearly to zero, where $\alpha = \min\left(\frac{1}{1+\frac{1}{l}}, \tau\right)$. And both of $\{A_k\}$* and $\{B_k\}$ *converges Q-linearly to zero.*

Proof We can easy to deduce that for any $k > k_l$,

$$\left(1 + \frac{1}{l}\right)A_{k+1} + B_{k+1} \le A_k + \tau B_k. \tag{74}$$

Denote $\alpha = \min\left(\frac{1}{1+\frac{1}{l}}, \tau\right)$ and $\beta = \max\left(\frac{1}{1+\frac{1}{l}}, \tau\right)$. Using the definitions of α and β and (74), we obtain

$$\begin{aligned} A_{k+1} + \alpha B_{k+1} &\le A_{k+1} + \left(\frac{1}{1+\frac{1}{l}}\right)B_{k+1} \le \left(\frac{1}{1+\frac{1}{l}}\right)A_k \\ &+ \left(\frac{\tau}{1+\frac{1}{l}}\right)B_k \le \beta\left(A_k + \alpha B_k\right), \end{aligned} \tag{75}$$

which means that $\{A_k + \alpha B_k\}$ converges Q-linearly to zero.

Further, we can deduce that $\{A_k\}$ and $\{B_k\}$ converge R-linearly to zeros using Lemma 4.1. □ □

Notation 4.1 Since Algorithm 5 and Algorithm 6 are special instances of Algorithm 4, then, similar to the Notation 3.1 in Sect. 3, we define the similar sequences $\{v_k\}, \{\delta_k\}$ and $\{\Gamma_k\}$, which are defined in (20),(21) and (22) respectively, for $\{x_k\}$ and $\{y_k\}$ generated by Algorithm 5 and Algorithm 6.

Theorem 4.1 *Suppose that Assumption 3.1 holds. Then, both of the sequences $\{x_k\}$ generated by the Algorithm 6 and Algorithm 7 are convergent and converges R-linearly to their limits. Also, $\{F(x_k)\}$ converges R-linearly to $F(x^*)$.*

Proof For the t_k generated by Algorithm 6, it follows from

$$t_{k+1}-1=\frac{\sqrt{1+4\left(\frac{\lambda_k}{\lambda_{k+1}}\right)t_k^2}-1}{2}\leq\frac{\sqrt{1+4t_k^2}-1}{2}<t_k,\ \forall\ k\ \text{is sufficiently large} \tag{76}$$

that there exists a $\hat{M}$ such that $t_k\leq\hat{M}$. Based on Lemma 2.1 and Corollary 2.1, we have

$$0\leq 1-\frac{\lambda_k}{\lambda_{k+1}}\leq\frac{1}{\hat{M}}\leq\frac{1}{t_k},$$

holds for k is sufficiently large. Then, similar with (40), it's easy to show that

$$\gamma_k=\frac{t_k-1}{t_{k+1}}\leq\frac{t_k-1}{t_k}=1-\frac{1}{t_k}\leq\frac{\hat{M}-1}{\hat{M}}<1.$$

From Algorithm 7, it is obvious that $\gamma_k=\frac{k-1}{k+a}\leq\frac{K-1}{K+a}<1$. Thus, for Algorithm 6 or Algorithm 7, there exists a $\bar{\gamma}$ such that $\gamma_k\leq\bar{\gamma}<1$.

Denote N_s is a sufficiently large positive integer. Recalling the definitions of $\{v_k\}$ and $\{\delta_k\}$ in (20) and (21). Let $\xi=v_{N_s}+\frac{\delta_{N_s}}{\lambda_{N_s}}+F(x^*)$. From the Assumption 3.1, we can deduce that for this ξ, there exist $\varepsilon>0$ and $\bar{\tau}>0$ such that $\text{dist}(x,X^*)\leq\bar{\tau}\left\|p_{\frac{1}{L_f}g}(x)-x\right\|$ holds for $\left\|p_{\frac{1}{L_f}g}(x)-x\right\|<\varepsilon$ and $F(x)\leq\xi$.

Recalling the definition of $\{\Gamma_{k+1}\}$ in (22). From (34), we obtain that

$$v_{k+1}+\frac{\delta_{k+1}}{\lambda_{k+1}}+\frac{2\bar{\mu}-1}{\lambda_{k+1}}\Gamma_{k+1}\leq v_k+\bar{\gamma}^2\frac{\delta_k}{\lambda_k}, \tag{77}$$

it is easy to get

$$v_{k+1}+\frac{\delta_{k+1}}{\lambda_{k+1}}\leq v_k+\frac{\delta_k}{\lambda_k}, \tag{78}$$

which means that for k is sufficiently large, $\left\{v_k+\frac{\delta_k}{\lambda_k}\right\}$ is nonincreasing. This together with the fact that $\left\{v_k+\frac{\delta_k}{\lambda_k}\right\}$ is bounded below deduce that $\left\{v_k+\frac{\delta_k}{\lambda_k}\right\}$ is convergent. Moreover, it follows from (78) that

$$v_k+\frac{\delta_k}{\lambda_k}\leq v_{N_s}+\frac{\delta_{N_s}}{\lambda_{N_s}}$$

which implies that for $k\geq N_s$

$$F(x_k)\leq\xi. \tag{79}$$

Based on Lemma 3.9, the nonexpansiveness property of the proximal operator [9], ∇f is Lipschitz continuous and $\lambda_{\min} \le \lambda_k \le \lambda^*$ for k is sufficiently large, we deduce that

$$\left\| p_{\frac{1}{L_f} g}(x_k) - x_k \right\| \le \frac{\mu_1}{\lambda_{\min} L_f} \left(1 + \lambda^* L_f\right) \|x_k - y_k\|. \tag{80}$$

See the detailed proof in (65).

Recalling (77), for $k \ge N_s$ we have

$$\left(\frac{2\bar{\mu} - 1}{\lambda_{k+1}}\right) \Gamma_{k+1} \le \left(v_k + \frac{\delta_k}{\lambda_k}\right) - \left(v_{k+1} + \frac{\delta_{k+1}}{\lambda_{k+1}}\right).$$

Summing from $k = N_s$ to $k = N$ and letting $N \to \infty$, we obtain that $\sum_{k=1}^{\infty} \Gamma_k$, i.e. $\sum_{k=1}^{\infty} \|x_k - y_k\|^2$ is convergent from Lemma 2.1. Then, combining with (80), we have

$$\lim_{k \to \infty} \left\| p_{\frac{1}{L_f} g}(x_k) - x_k \right\| = 0. \tag{81}$$

Following (79) and (81), we have $F(x_k) \le \xi$ and $\left\| p_{\frac{1}{L_f} g}(x_k) - x_k \right\| < \varepsilon$ hold for k is sufficiently large. Then, combining with (63), there exists $\tau_2 > 0$ for k is sufficiently large,

$$\operatorname{dist}(x_k, X^*) \le \tau_2 \|x_k - y_k\|. \tag{82}$$

From (34) with $y := y_{k+1}$ and $\lambda := \lambda_{k+1}$, we have

$$\begin{aligned} F(x_{k+1}) &\le F(x) + \frac{\|x - y_{k+1}\|^2}{2\lambda_{k+1}} = F(x) + \frac{\|x - x_{k+1} + x_{k+1} - y_{k+1}\|^2}{2\lambda_{k+1}} \\ &\le F(x) + \frac{1}{\lambda_{k+1}} \left(\|x - x_{k+1}\|^2 + \|x_{k+1} - y_{k+1}\|^2 \right). \end{aligned} \tag{83}$$

Choose x to be an $x^*_{k+1} \in X^*$ satisfying $\left\| x^*_{k+1} - x_{k+1} \right\| = \operatorname{dist}(x_{k+1}, X^*)$, then,

$$\begin{aligned} F(x_{k+1}) - F(x^*) &\le \frac{1}{\lambda_{k+1}} \left(\left\| x^*_{k+1} - x_{k+1} \right\|^2 + \|x_{k+1} - y_{k+1}\|^2 \right) \\ &= \frac{1}{\lambda_{k+1}} \left(\operatorname{dist}^2(x_{k+1}, X^*) + \|x_{k+1} - y_{k+1}\|^2 \right) \\ &\le \tau_3 \|x_{k+1} - y_{k+1}\|^2, \end{aligned} \tag{84}$$

where $\tau_3 = \frac{1 + (\tau_2)^2}{\lambda_{\min}}$ and the last inequality is from (82) and Lemma 2.1, i.e.,

$$v_{k+1} \le 2\tau_3 \Gamma_{k+1} \tag{85}$$

holds.

It follows from (77) and (85) and Lemma 4.2 that $\left\{ v_k + \alpha \frac{\delta_k}{\lambda_k} \right\}$ converges Q-linearly to zero. And $F(x_k)$ converges R-linearly to $F(x^*)$, $\left\{ \|x_{k+1} - x_k\|^2 \right\}$ converges R-linearly to zero.

With the R-linear convergence of $\left\{ \|x_{k+1} - x_k\|^2 \right\}$, we obtain that there exist $0 < \bar{c} < 1$ and $M_1 > 0$, such that

$$\|x_k - x_{k-1}\| \leq M_1 \bar{c}^k.$$

Consequently, for any $m_2 > m_1 > 0$, we have

$$\|x_{m_2} - x_{m_1}\| \leq \sum_{k=m_1+1}^{m_2} \|x_k - x_{k-1}\| \leq M_1 \cdot \frac{\bar{c}^{m_1}}{1-\bar{c}}$$

showing that $\{x_k\}$ is a Cauchy sequence and hence convergent. Denoting its limit by $x^*$and passing to the limit as $m_2 \to \infty$ in the above relation, we see further that

$$\|x_{m_1} - x^*\| \leq M_1 \cdot \frac{\bar{c}^{m_1}}{1-\bar{c}}$$

that means that the sequence $\{x_k\}$ converges R-linearly to its limit. □

Remark 4.1 Under the error bound condition, Wen et al. [30] proved that for FISTA equipping with the restart scheme and the constant stepsize $\frac{1}{L_f}$, the sequences $\{x_k - x^*\}$ and $\{F(x_k) - F(x^*)\}$ converge R-linearly to zero. In Theorem 4.1, we show similar results hold for FISTA and MFISTA with stepsize generated by Algorithm 3 based on the error bound condition and restart scheme. The proposed algorithm implementations are independent of L_f. In the proof of Theorem 4.1, the main contribution of Algorithm 3 is that it generates a stepsize sequence which is convergent and increases monotonically after finite iterations. We see that backtracking strategy in FISTA_BKTR does not have this property, hence, it is not clear whether FISTA_BKTR can obtain the linear convergence.

5 Numerical experiments

5.1 Performance comparison of FISTA algorithms based on different stepsize strategies

We conduct numerical experiments to demonstrate effectiveness of our algorithms by testing the following six algorithms:

- FISTA_NMS
- MFISTA_NMS($a = 4$)
- FISTA_BKTR
- MFISTA_BKTR($a = 4$)

– FISTA_backtracking
– SpaRSA: This algorithm is a non-monotone proximal gradient method, whose description can be found in [31].

Algorithm 7 SpaRSA

Step 0. Set $\lambda_0 > 0$, $\eta < 1$, $\bar{M} = 5, \beta = 0.25, \bar{\sigma} = 10^{-5}$, and $\lambda_{\max} = 1/\lambda_{\min} = 10^{30}$.
Step k. (1) Set $\lambda_k = \max\left\{\lambda_{\min}, \min\left(\lambda_{\max}, \lambda^{BB}\right)\right\}$,

where $\lambda^{BB} = \left|\frac{\|s_k\|^2}{s_k^T v_k}\right|$ or $\left|\frac{s_k^T v_k}{\|v_k\|^2}\right|$, $s_k = x_{k+1} - x_k, v_k = \nabla f(x_{k+1}) - \nabla f(x_k)$

(2) repeat

$$x_{k+1} = p_{\lambda_k g}(x_k)$$
$$\lambda_k = \eta \lambda_k$$

until $F(x_{k+1}) \leq \max\limits_{i=\max(k-\bar{M},0),\cdots,k} F(x_i) - \frac{\bar{\sigma}}{2}\lambda_k \|x_{k+1} - x_k\|^2$

Termination condition: The inequality $\|\psi_k\| \leq \varepsilon$ is often used to be the termination condition for all comparison algorithms, where $\psi_k = \nabla f(x_k) - \frac{1}{\lambda_k}(x_k - y_k + \lambda_k \nabla f(y_k)) \in \partial F(x_k)$. However, we notice that if F is flat, the distance between two iterates will be very far but the value of $\|\psi_k\|$ is close to 0; and vice versa. Hence, we terminate the test algorithms when $\min\left(\|\psi_k\|, \|x_k - x_{k-1}\|\right) \leq \varepsilon$.

Test Function: The numerical experiments are conducted on the following two types of test functions: (1) the l_1-regularized least squares problem; (2) the l_1- regularized logistic regression. It's obvious that the first problem is the case that f is a quadratic function, thus we need to restrict the parameter $\mu_1 < \mu_0 < 1$; for the latter that f is a non-quadratic function, $\mu_1 < \mu_0 < 1/2$. In the numerical experiment, we set $E_k = \frac{w_k}{k^{1.1}}, \forall k \geq 1$ be the control series for the new adaptive non-monotone stepsize strategy, parameter w_k same as the setting we introduced in Sect. 2; $\mu_0 = 0.99, \mu_1 = 0.95$ for the test function (1), $\mu_0 = 0.49, \mu_1 = 0.45$ for the test function (2); $\varepsilon = 1e-5$. For the backtracking scheme, we set $\eta = 0.5$; For the BKTR scheme, we set $\beta = 0.5$ and $\lambda_k^0 = \frac{\lambda_{k-1}}{0.8}$.

Table 1 Comparison of algorithms for solving (86) with $n = 800, m = 8000, s = 80, \sigma = 1$

		Iter	Mult	Time
$\sigma = 1, s = 80$	FISTA_NMS	3287	6576	24.5598
	FISTA_BKTR	3101	15394	56.1681
	FISTA_backtracking	9113	36489	132.9190
	SpaRSA	10001	20005	74.8957
	MFISTA_NMS	2850	5702	21.3633
	MFISTA_BKTR	2495	12385	45.2483

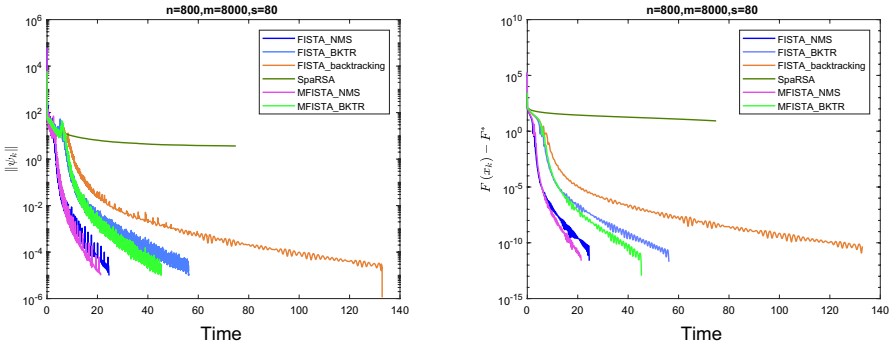

Fig. 1 Performance profile for the convergences of $\|\psi_k\|$ and $F(x_k) - F(x^*)$ with $\sigma = 1$

Table 2 Comparison of algorithms for solving (86) with $n = 800, m = 8000, s = 80, \sigma = 0.1$

		Iter	Mult	Time
$\sigma = 0.1, s = 80$	FISTA_NMS	10028	20058	77.7090
	FISTA_BKTR	9718	48252	175.6361
	FISTA_backtracking	34635	138580	502.9829
	SpaRSA	50001	100005	373.1670
	MFISTA_NMS	7631	15264	56.8800
	MFISTA_BKTR	7534	37407	136.6205

5.1.1 l_1-regularized least squares problem

l_1-regularized least squares problem is described as follows:

$$\min_x F(x) = \frac{1}{2}\|Ax - b\|^2 + \sigma\|x\|_1, \tag{86}$$

where the linear operator A and observation b is generated by the following scheme:

A = randn(n, m);
xstar = ones(m, 1);
Set s : The number of non − zero elements of xstar
I = randperm(m); xstar(I(1 : m − s)) = 0;
b = A ∗ xstar+0.1 ∗ randn(n, 1);

In the numerical experiments, we take $n = 800, m = 8000$.

Note that in this linear inverse problem, $\nabla f(x) = A^T(Ax - b)$, which is linear, hence, we can directly compute $\nabla f(y_k)$ by linear relationship between $\nabla f(x_{k-1})$ and $\nabla f(x_{k-2})$; since $Ay_k - b$ can be computed by linear relationship between $Ax_{k-1} - b$ and $Ax_{k-2} - b$, so the computation of $f(y_k) = \frac{1}{2}\|Ay_k - b\|^2$ can be almost negligible.

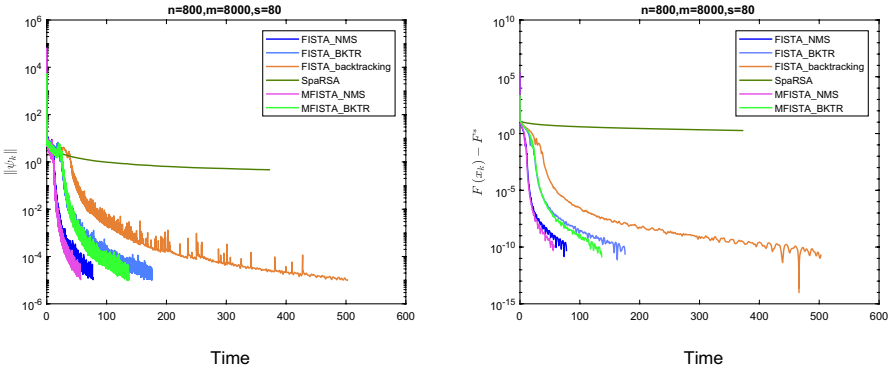

Fig. 2 Performance profile for the convergences of $\|\psi_k\|$ and $F(x_k) - F(x^*)$ with $\sigma = 0.1$

Table 3 Comparison of algorithms for solving (86) with $n = 800, m = 8000, s = 200, \sigma = 1$

		Iter	Mult	Time
$\sigma = 1, s = 200$	FISTA_NMS	29125	58252	220.8459
	FISTA_BKTR	29728	147714	547.6624
	FISTA_backtracking	47226	188938	705.4287
	SpaRSA	100001	200005	765.5658
	MFISTA_NMS	16894	33790	128.2840
	MFISTA_BKTR	14213	70687	264.4558

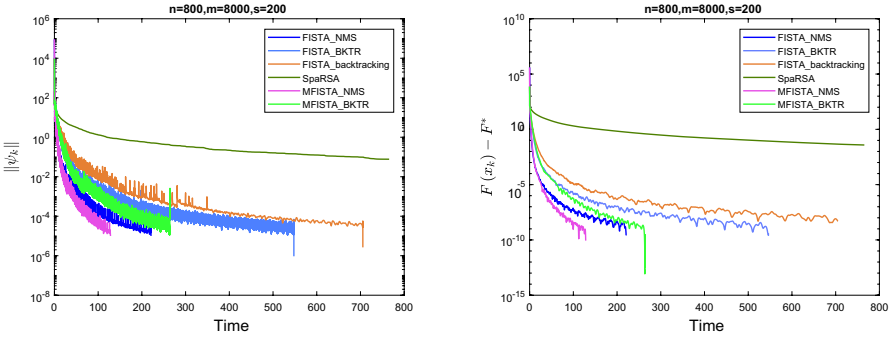

Fig. 3 Performance profile for the convergences of $\|\psi_k\|$ and $F(x_k) - F(x^*)$ with $\sigma = 1$

Through numerical experiments, we find that for FISTA_backtracking and FISTA_BKTR, the condition $F(p_{\lambda_k g}(y_k)) \leq Q_{\lambda_k}(p_{\lambda_k g}(y_k), y_k)$ is difficult to distinguish if we set ε too small, which means that these two backtracking schemes are not suitable for applications with high precision requirements like medical imaging. We consider the influence of such factors like sparsity $\left(\frac{s}{m}\right)$ and regularization parameter σ on the algorithms. The selection of regularization parameter is separately $\sigma = 1$ and

$\sigma = 0.1$. *Iter* denotes the total number of iterations and *Mult* denotes the number of matrix–vector product for compute $Ax - b$ and *Time* denotes the CPU time.

From Table 1, 2, 3, we see that under the setting of different parameters and different sparsity, our algorithms FISTA_NMS and MFISTA_NMS hava significant improvment over FISTA_backtracking and SpaRSA, and comparing with FISTA_BKTR and MFISTA_BKTR, we see that FISTA_BKTR is a little better than FISTA_NMS for the total number of iterations, but much more than FISTA_NMS for the number of matrix–vector product, the comparison with other two algorithms MFISTA_NMS and MFISTA_BKTR shows similar results. In order to more intuitively show the effectiveness of our algorithms, we plot how $\|\psi_k\|$ and $F(x_k) - F(x^*)$ change during the progress of the six algorithms, where F^* be the smallest terminating $F(x_k)$ among all methods.

From Figs. 1, 2, 3, we can see that even if regularization parameter selection and sparsity are different, FISTA_NMS has a significant improvement over the FISTA_BKTR, FISTA_backtracking and SpaRSA for the given test problems. In particular, at the maximum iteration point, SpaRSA is far from the optimal value,

Table 4 Comparison of algorithms for solving "heart_test"

	Iter	Fval	Gval	Time
FISTA_NMS	81255	81255	162510	8.0257
FISTA_BKTR	75631	199956	175608	11.8186
MFISTA_NMS	23598	23598	47196	2.3079
MFISTA_BKTR	20463	54100	47512	3.1921

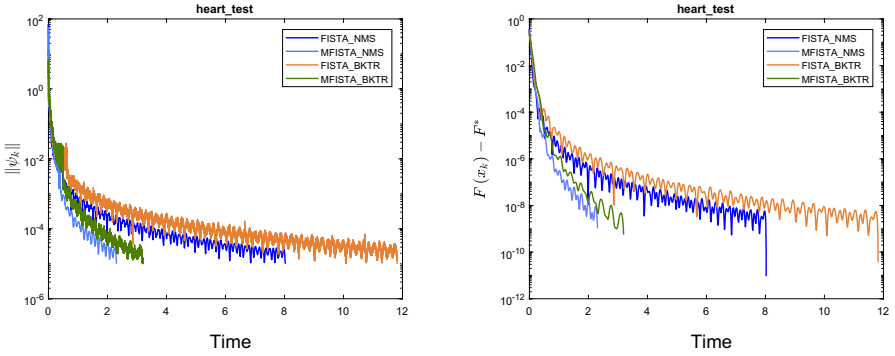

Fig. 4 Performance profile for solving "heart_test"

Table 5 Comparison of algorithms for solving "sonar_test"

	Iter	Fval	Gval	Time
FISTA_NMS	1038	1038	2076	0.1278
FISTA_BKTR	911	2406	2114	0.1696
MFISTA_NMS	730	730	1460	0.0878
MFISTA_BKTR	540	1422	1251	0.0994

it needs more iterations to meet the termination condition, and the corresponding *Mult* and *Time* will increase. Moreover, we can see that MFISTA_NMS is more efficient than MFISTA_BKTR, which means that our stepsize strategy is also effective for the modified algorithm MFISTA. Numerical experiments show that the new adaptive nonmonotone stepsize strategy is very useful for improving algorithm performances and our algorithms are very suitable for practical application problems such as sparse signal processing.

Since numerical performance of BKTR is better than backtracking and BB stepsize, in the following computational experiments, we just compare NMS and BKTR strategies with same algorithm schemes like FISTA and MFISTA for solving the sparse logistic regression problem to better illustrate the efficiency of NMS.

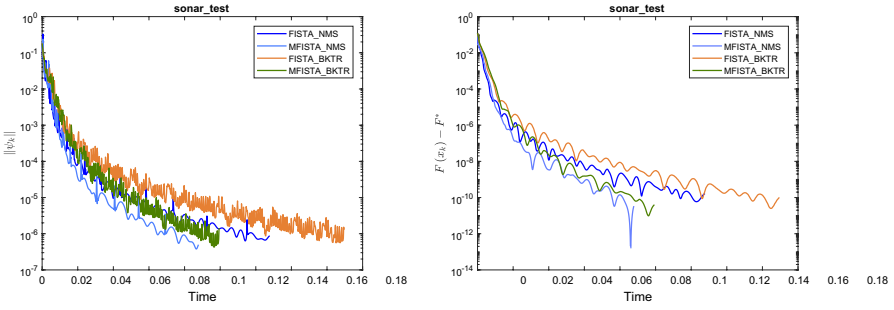

Fig. 5 Performance profile for solving "sonar_test"

Table 6 Comparison of algorithms for solving "mushroom"

	Iter	Fval	Gval	Time
FISTA_NMS	121	121	242	0.0361
FISTA_BKTR	106	250	231	0.0461
MFISTA_NMS	99	99	198	0.0259
MFISTA_BKTR	126	300	276	0.0484

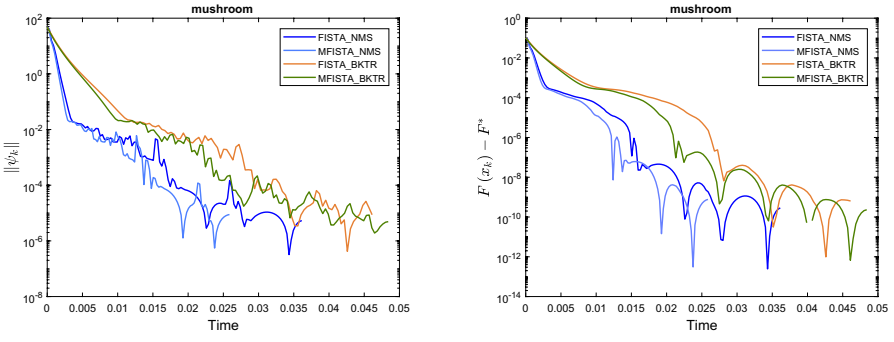

Fig. 6 Performance profile for solving "mushroom"

5.1.2 Sparse logistic regression

Consider the question

$$\min_x F(x) := \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp\left(-l_i\langle h_i, x\rangle\right)\right) + \sigma\|x\|_1, \tag{87}$$

where $x \in R^m, h_i \in R^m, l_i \in \{-1,1\}, i = 1,\cdots,n,$ and $\sigma = 1e-2$. The problem sparse logistic regression is a popular problem in machine learning applications, where $f(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp\left(-l_i\langle h_i, x\rangle\right)\right)$ is non-linear. Define $K_{ij} = -l_i h_{ij}$, and set $\tilde{f}(y) = \sum_{i=1}^{m} \log\left(1+\exp\left(y_i\right)\right)$. Then $f(x) = \tilde{f}(Kx)$, and $L_f = \frac{4}{n}\left\|K^T K\right\|$. Initial point x_0= zeros(m, 1). We take three datasets 'heart_test', 'sonar_test' and 'mushroom' from LIBSVM [10]. We report the number of iterations (*Iter*), calculation of function value (*Fval*), calculation of gradient value (*Gval*) and CPU time (*Time*) (Tables 4, 5, 6).

The algorithms for solving the Sparse logistic regression problem obtain similar results, i.e., though the number of iterations of algorithms with NMS is slightly worse than algorithms with BKTR, we can see that the algorithms with NMS are obviously better from the calculation times of function and gradient values and CPU time. Hence, FISTA_NMS outperforms the FISTA_BKTR, and meanwhile,

Table 7 Comparison of algorithms with restart scheme and without restart scheme for solving (86) with $n = 800, m = 8000$

		Iter	Mult	Time
$\sigma = 1, s = 80$	FISTA_NMS	3700	7402	28.4569
	FISTA_NMS_restart	678	1358	5.2647
	MFISTA_NMS	2967	5936	22.8726
	MFISTA_NMS_restart	776	1554	5.9951
	FISTA_BKTR_restart	632	3133	11.7834
	MFISTA_BKTR_restart	683	3388	12.7432
$\sigma = 0.1, s = 80$	FISTA_NMS	10461	20924	81.0398
	FISTA_NMS_restart	2512	5026	19.4928
	MFISTA_NMS	8000	16002	62.0796
	MFISTA_NMS_restart	2711	5424	21.1646
	FISTA_BKTR_restart	2270	11269	43.3095
	MFISTA_BKTR_restart	2356	11694	44.4426
$\sigma = 1, s = 200$	FISTA_NMS	32581	65164	262.3714
	FISTA_NMS_restart	13689	27380	110.0113
	MFISTA_NMS	16832	33666	132.1907
	MFISTA_NMS_restart	13342	26686	106.2211
	FISTA_BKTR_restart	8984	44701	179.8750
	MFISTA_BKTR_restart	10623	52859	218.2645

Table 8 Comparison of algorithms with restart scheme and without restart scheme for solving (87)

		Iter	Fval	Gval	Time
heart_test	FISTA_NMS	81305	81305	162610	7.8653
	FISTA_NMS_restart	7876	7876	15752	0.7712
	MFISTA_NMS	23586	23586	47172	2.2660
	MFISTA_NMS_restart	9536	9536	19072	0.9202
	FISTA_BKTR_restart	7046	18626	16358	1.0999
	MFISTA_BKTR_restart	7689	20326	17851	1.1962
sonar_test	FISTA_NMS	1081	1081	2162	0.1567
	FISTA_NMS_restart	237	237	474	0.0383
	MFISTA_NMS	660	660	1320	0.0941
	MFISTA_NMS_restart	235	235	470	0.0364
	FISTA_BKTR_restart	211	552	487	0.0495
	MFISTA_BKTR_restart	170	446	393	0.0391
mushroom	FISTA_NMS	114	114	228	0.0366
	FISTA_NMS_restart	72	72	144	0.0251
	MFISTA_NMS	115	115	230	0.0361
	MFISTA_NMS_restart	58	58	116	0.0224
	FISTA_BKTR_restart	87	199	186	0.0418
	MFISTA_BKTR_restart	86	195	183	0.0361

MFISTA_NMS is more efficient than the MFISTA_BKTR. Observe that sometimes FISTA_NMS is faster than MFISTA_BKTR for some test problems, like the Sparse logistic regression with "mushroom" dataset (Figs. 4, 5, 6).

5.2 Performance comparison of FISTA algorithms with restart based on different stepsize strategies

The main goal of our experiments in this subsection is to test that our algorithms combining with the *Restart* scheme are still effective. The test functions and the related parameter settings are same as Sect. 5.1.

First, we compare the following four algorithms: FISTA_NMS; FISTA_NMS_restart; MFISTA_NMS and MFISTA_NMS_restart. We can see that using the restart strategy, both of our algorithms' performances can be greatly improved, which shows from Table 7 that *Iter*, *Mult* and *Time* for solving the l_1-regularized least squares problem be greatly reduced; and from Table 8 that *Iter*, *Fval*, *Gval* and *Time* for solving the Sparse logistic regression be greatly reduced. Further, by comparing FISTA_BKTR_restart, MFISTA_BKTR_restart, FISTA_NMS_restart and MFISTA_NMS_restart, the numerical results elaborate that: after incorporating restart strategy into all the comparison algorithms, our algorithms are still superior to the other two comparison algorithms, which shows the stability of our algorithms.

6 Conclusion

In this paper, we introduce a new adaptive non-monotone stepsize strategy (NMS), which does not execute line search and is independent of the Lipschitz constant. Based on NMS, we propose FISTA_NMS that has $O\left(\frac{1}{k^2}\right)$ convergence rate of the objective function values, which is similar to FISTA. We construct the convergence of iterates generated by MFISTA_NMS based on the new adaptive non-monotone stepsize without depending on the Lipschitz constant. Also, the convergence rate of objective function values shares $o\left(\frac{1}{k^2}\right)$. Further, our algorithms FISTA_NMS and MFISTA_NMS achieve similar convergence rates in the norm of subdifferential of objective function. Under error bound condition, we prove that FISTA_NMS and MFISTA_NMS have improved convergence results, i.e., for FISTA_NMS, convergence rates of function values and iterates can be achieved to $o\left(\frac{1}{k^6}\right)$ and $O\left(\frac{1}{k^2}\right)$; for MFISTA_NMS, that are $o\left(\frac{1}{k^{2(a+1)}}\right)$ and $o\left(\frac{1}{k^a}\right)$. In addition, we improve our algorithms and give the proof of the linear convergence of function values and iterates by combining our algorithms with the restart strategy. Note that FISTA and MFISTA with backtracking schemes can not achieve the same results, which means that NMS has theoretical advantages. We demonstrate the performances of our schemes on some numerical examples to show that our stepsize strategy outperforms the backtracking.

Acknowledgements The work was supported by the National Natural Science Foundation of China (No. 11901561), the Natural Science Foundation of Guangxi (No. 2018GXNSFBA281180) and the Postdoctoral Fund Project of China (Grant No. 2019M660833), Guizhou Provincial Science and Technology Projects (No. QKHJC-ZK[2022]YB084).

Data availability Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

1. Attouch, H., Peypouquet, J.: The rate of convergence of Nesterov's accelerated forwardbackward method is actually faster than $\frac{1}{k^2}$. SIAM J. Optim. **26**, 1824–1834 (2016)
2. Attouch, H., Cabot, A.: Convergence rates of inertial forward-backward algorithms. SIAM J. Optim. **28**, 849–874 (2018)
3. Apidopoulos, V., Aujol, J., Dossal, C.: Convergence rate of inertial Forward–Backward algorithm beyond Nesterov's rule. Math. Program. **180**, 137–156 (2020)
4. Beck, A., Teboulle, M.: A linearly convergent dual-based gradient projection algorithm for quadratically constrained convex minimization. Math. Oper. Res. **31**, 398–417 (2006)
5. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**, 183–202 (2009)
6. Becker, S.R., Candès, E.J., Grant, M.C.: Templates for convex cone problems with applications to sparse signal recovery. Math. Prog. Comp. **3**, 165–218 (2011)
7. Bello, C., Jose, Y., Nghia, T.T.A.: On the convergence of the forward-backward splitting method with linesearches. Optim. Method Softw. **31**, 1209–1238 (2016)
8. Chambolle, A.: An algorithm for total variation minimization and applications. J. Math. Imaging Vis. **20**, 89–97 (2004)
9. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. **4**, 1168–1200 (2005)

10. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM. Trans. Intell. Syst. Technol. **2**, 1–27 (2011)
11. Combettes, P.L., Pesquet, J.C.: Proximal splitting methods in signal processing. In: Fixed-Point Algorithms for Inverse Problems in Science and Engineering. Springer, New York (2011)
12. Chambolle, A., Dossal, C.: On the convergence of the iterates of the fast iterative shrinkage-thresholding algorithm. J. Optim. Theory Appl. **166**, 968–982 (2015)
13. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**, 1289–1306 (2006)
14. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**, 964–979 (1979)
15. Luo, Z.Q., Tseng, P.: Error bound and the convergence analysis of matrix splitting algorithms for the affine variational inequality problem. SIAM J. Optim. **2**, 43–54 (1992)
16. Luo, Z.Q.: New error bounds and their applications to convergence analysis of iterative algorithms. Math. Program. **88**, 341–355 (2000)
17. Lorenz, D.A., Pock, T.: An inertial forward-backward algorithm for monotone inclusions. J. Math. Imaging Vis. **51**, 311–325 (2015)
18. Liang, J., Fadili, J., Peyré, G.: Convergence rates with inexact non-expansive operators. Math. Program. **159**, 403–434 (2016)
19. Liu, H.W., Wang, T., Liu, Z.X.: Convergence rate of inertial forward-backward algorithms based on the local error bound condition. arXiv:2007.07432
20. Molinari, C., Liang, J., Fadili, J.: Convergence rates of forward-douglas-rachford splitting Method. J. Optim. Theory Appl. **182**, 606–639 (2019)
21. Nesterov, Y.: A method for solving the convex programming problem with convergence rate $O\left(\frac{1}{k^2}\right)$. Dokl. Akad. Nauk SSSR **269**, 543–547 (1983)
22. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**, 125–161 (2013)
23. O'Donoghue, B., Candès, E.: Adaptive restart for accelerated gradient schemes. Found. Comput. Math. **15**, 715–732 (2015)
24. Sra, S., Nowozin, S., Wright, S.J.: Optimization for Machine Learning. MIT Press, Cambridge (2012)
25. Sun, W.Y., Yuan, Y.X.: Optimization Theory and Methods: Nonlinear Programming. Springer, Berlin (2010)
26. Su, W., Boyd, S., Candès, E.J.: A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. J. Mach. Learn. Res. **17**, 1–43 (2016)
27. Scheinberg, K., Goldfarb, D., Bai, X.: Fast first-order methods for composite convex optimization with backtracking. Found. Comput. Math. **14**, 389–417 (2014)
28. Tseng, P.: Approximation accuracy, gradient methods, and error bound for structured convex optimization. Math. Program. **125**, 263–295 (2010)
29. Tao, S., Boley, D., Zhang, S.: Local linear convergence of ISTA and FISTA on the LASSO problem. SIAM J. Optim. **26**, 313–336 (2016)
30. Wen, B., Chen, X.J., Pong, T.K.: Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. SIAM J. Optim. **27**, 124–145 (2017)
31. Wright, S.J., Nowak, R.D., Figueiredo, M.A.T.: Sparse reconstruction by separable approximation. IEEE T. Signal Proces. **57**, 2479–2493 (2009)
32. Wang, T., Liu, H.: Convergence results of a new monotone Inertial Forward-Backward Splitting algorithm under the local Hölder error bound condition. Appl. Math. Optim. (2022). https://doi.org/10.1007/s00245-022-09859-y

Authors and Affiliations

Hongwei Liu[1] · Ting Wang[1] · Zexian Liu[2]

Hongwei Liu
hwliu@mail.xidian.edu.cn

Zexian Liu
liuzexian2008@163.com

[1] School of Mathematics and Statistics, Xidian University, Xi'an 710126, China

[2] School of Mathematics and Statistics, Guizhou University, Guiyang 550025, China