Yokohama Publishers
ISSN 1349-8169 ONLINE JOURNAL

Pacific Journal of Optimization
Vol. 19, No. 3, pp. 477–487, 2023

CONVERGENCE ANALYSIS OF PROXIMAL GRADIENT ALGORITHM WITH EXTRAPOLATION FOR A CLASS OF CONVEX NONSMOOTH MINIMIZATION PROBLEMS*

Mengxi Pan and Bo Wen†

Abstract: In this paper, we consider the proximal gradient algorithm with extrapolation ($\mathrm{PG_e}$) for solving a class of nonsmooth convex minimization problems, whose objective function is the sum of a continuously differentiable convex function with Lipschitz gradient and a proper closed convex function. We first establish the subsequential convergence of iterate generated by $\mathrm{PG_e}$, then we prove that the convergence rate of objective function is $O(1/k)$, which implies that the convergence rate of FISTA with fixed restart is also $O(1/k)$. Finally, we conduct some numerical experiments to illustrate the theoretical results.

Key words: *nonsmooth convex minimization, proximal gradient algorithm, extrapolation, convergence analysis*

Mathematics Subject Classification: *90C25, 65K05, 90C30*

1 Introduction

In this paper, we consider the following optimization problem:

$$\min_{x\in R^n} F(x) = f(x) + g(x), \tag{1.1}$$

where f is a smooth convex function with Lipschitz continuous gradient and g is a proper closed convex function.

In recent years, problem (1.1) draws many attention due to its wide range applications, such as matrix completion [3] and compressed sensing [4, 6]. From the assumptions of the objective function, we know that problem (1.1) is a nonsmooth convex optimization problem. Hence, many popular algorithms can be applied to solve this problem, including various bundle methods [9, 11, 17] and proximal gradient algorithm [10]. However, the original proximal gradient algorithm can be slow when it is applied to solve the practical problems. Then, the proximal gradient algorithm can be accelerated by performing various extrapolation techniques, which means adding momentum terms that involve previous iterates for updating the current iterate, see, for example, [1, 5, 8, 14]. One popular and commonly used extrapolation method is the accelerated gradient algorithm [15], which was proposed

*This work was supported in part by NSFC11801131, Natural Science Foundation of Hebei Province (Grant No. A2019202229)

†Corresponding author

by Nesterov in 1983 for solving a class of smooth convex optimization problems. A typical algorithm takes the following form:

$$\begin{cases} y^k = x^k + \beta_k \left(x^k - x^{k-1}\right), \\ x^{k+1} = y^k - s\nabla f\left(y^k\right), \end{cases} \tag{1.2}$$

where $s > 0$ is the step size, which depends on the Lipschitz continuity modulus of ∇f, and the extrapolation coefficients β_k satisfy $0 < \beta_k < 1$ for all k. It was shown in [15], by choosing specific extrapolation coefficients sequence $\{\beta_k\}$ with $\sup_k \beta_k = 1$, this algorithm has a faster convergence rate than the general gradient algorithm, which is

$$F\left(x^k\right) - \inf_{x\in\mathbb{R}^n} F = O\left(\frac{1}{k^2}\right),$$

where $\{x_k\}$ is generated by (1.2).

Recently Beck and Teboulle [1] proposed a fast iterative shrinkage-thresholding algorithm (FISTA) for solving a class of nonsmooth convex minimization problems, which extends Nesterov's original accelerated gradient algorithm to nonsmooth case. A similar algorithm was independently proposed by Nesterov [13]. Similar to (1.2), we just take the following form as a general case,

$$\begin{cases} y^k = x^k + \beta_k \left(x^k - x^{k-1}\right), \\ x^{k+1} = \arg\min\limits_{x\in\mathbb{R}^n} \left\{ \left\langle \nabla f\left(y^k\right), x\right\rangle + \dfrac{1}{2s}\left\|x - y^k\right\|^2 + g\left(x\right)\right\}, \end{cases} \tag{1.3}$$

where $s > 0$ is a constant, which depends on the Lipschitz continuity modulus of ∇f , and $0 < \beta_k < 1$ for all k. By choosing specific extrapolation coefficients $\{\beta_k\}$ with $\sup_k \beta_k = 1$, FISTA also exhibits a fast convergence rate, which is $O\left(1/k^2\right)$. Besides these works, many other accelerated algorithms based on Nesterov's technique [16, 12, 13] have been proposed, see [2, 18] and the references therein for an overview.

More recently, O'Donoghue and Candès [7] proposed an adaptive restart scheme for FISTA. Specifically, instead of following the recurrence relation of β_k in FISTA for all k, they reset $\beta_k = \beta_0$ every K iterations, where K is a positive number. Although numerically the algorithm obtained behaved well, no theoretical global convergence analysis was provided. In this paper, we mainly study the global convergence of a proximal gradient algorithm with extrapolation($\mathrm{PG_e}$), which can take the FISTA with both fixed and adaptive restart schemes as a special case.

In details, we first establish the global subsequential convergence of iterates generated by $\mathrm{PG_e}$, then we prove that the global convergence rate of $\mathrm{PG_e}$ is $O\left(1/k\right)$. Hence, we can conclude that the global convergence rate of FISTA with both fixed and adaptive restart schemes is $O\left(1/k\right)$. Finally, some numerical experiments have been performed to show the efficiency of the proposed algorithm.

The contents of this paper are as follows. The notation and preliminary materials are described in Section 2. We present the convergence analyses of the proximal gradient algorithm with fixed extrapolation in Section 3. Numerical experiments are introduced in Section 4.

2 Preliminaries

In the whole paper, the problem is considered in $\mathbb{R}^n$ space, and we denote its inner product by $\langle\cdot,\cdot\rangle$. We use $\|\cdot\|$, $\|\cdot\|_1$ and $\|\cdot\|_\infty$ to denote the Euclidean norm, the ℓ_1 norm and the ℓ_∞

norm, respectively. We define A^T as the transpose of a matrix $A \in \mathbb{R}^{m\times n}$. For a symmetric matrix $A \in \mathbb{R}^{n\times n}$, we denote $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ as its largest and smallest eigenvalues, respectively.

For an extended-real-valued function $h : \mathbb{R}^n \to [-\infty, \infty]$, we denote its domain as $h = \{x \in \mathbb{R}^n : h(x) < \infty\}$. The function h is said to be proper if it never equals $-\infty$ and $dom\ h \neq \varnothing$. A proper function h is said to be closed if it is lower semicontinuous. For a proper closed function h, we say it is level bounded if the lower level sets of h are bounded, which means the set $\{x \in \mathbb{R}^n : h(x) \leq r\}$ is bounded for any $r \in \mathbb{R}$. We denote the subdifferential of a proper closed convex function $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ at $x \in dom\ h$ by

$$\partial h(x) = \{\xi \in \mathbb{R}^n : h(u) - h(x) - \langle \xi, u - x\rangle \geq 0, \forall u \in \mathbb{R}^n\}.$$

If h is additionally continuously differentiable, then the subdifferential of h reduces to the gradient of h denoted by ∇h.

Given a proper closed convex function h, the proximal operator of h at any $v \in \mathbb{R}^n$ is defined by

$$\text{prox}_h(v) = \underset{x\in\mathbb{R}^n}{\arg\min} \left\{ h(x) + \frac{1}{2}\|x - v\|^2 \right\}. \tag{2.1}$$

Before ending this section, we give the definition of the stationary point of (1.1). For an optimal solution $\hat{x}$ of (1.1), the following first-order necessary condition always holds:

$$0 \in \nabla f(\hat{x}) + \partial g(\hat{x}), \tag{2.2}$$

where ∇f denotes the gradient of f, ∂g denotes the subdifferential of g. We say that $\tilde{x}$ is a stationary point of (1.1) if it satisfies (2.2). We use χ to denote the set of stationary points of F. Since (1.1) is a convex optimization problem, we immediately obtain from the above discussion that χ also denotes the set of global minimizers of problem (1.1).

3 Convergence Analysis

In this section, we present the proximal gradient algorithm with extrapolation for solving (1.1) and study the convergence properties of the sequence generated by the algorithm.

First, from the assumptions in our problem (1.1), we note that the function g is proper closed convex and f has a Lipschitz continuous gradient; moreover, the function $F = f + g$ is level bounded. From these, we obtain that a minimizer of (1.1) exists and consequently, $\inf F > -\infty$. Let L be a Lipschitz continuity modulus of ∇f . We are now ready to present our algorithm.

Algorithm 1 Proximal gradient algorithm with extrapolation(PG_e)

Require: $x^0 \in dom\ g, \{\beta_k\} \subseteq (0, 1)$. Set $x^{-1} = x^0$.

1: **for** $k = 0, 1, 2, \ldots$ **do**

2:
$$\begin{aligned} y^k &= x^k + \beta_k\left(x^k - x^{k-1}\right) \\ x^{k+1} &= \text{prox}_{\frac{1}{L}g}\left(y^k - \tfrac{1}{L}\nabla f\left(y^k\right)\right) \end{aligned} \tag{3.1}$$

3: **end for**

We shall discuss the convergence behavior of Algorithm 1. We note first that the x-update in (3.1) is equivalent to the following relation

$$x^{k+1} = \underset{x\in\mathbb{R}^n}{\arg\min} \left\{ \left\langle \nabla f\left(y^k\right), x\right\rangle + \frac{L}{2}\left\|x - y^k\right\|^2 + g(x) \right\}, \tag{3.2}$$

which is due to the definition of proximal operator in (2.1). This fact will be used repeatedly in our convergence analysis below. Our analysis also relies heavily on the following auxiliary sequence:

$$H_{k,\gamma} = F\left(x^k\right) + \gamma \cdot \frac{L}{2}\left\|x^k - x^{k-1}\right\|^2, \tag{3.3}$$

where $\gamma \in (0,1)$ is a constant, $\left\{x^k\right\}$ is generated by Algorithm 1. Next, we will give some important lemmas used below.

3.1 Auxiliary lemmas.

Lemma 3.1. *Suppose that there exists $\gamma \in (0,1)$ such that $\gamma \geq \sup_k \beta_k$. Let $\left\{x^k\right\}$ be a sequence generated by Algorithm 1. Then the following statements hold:*

(i) *The sequence $\{H_{k,\gamma}\}$ is nonincreasing.*

(ii) *The sequence $\left\{x^k\right\}$ is bounded.*

(iii) *The sequence $\{H_{k,\gamma}\}$ is convergent.*

(iv) $\sum_{k=0}^{\infty}\left\|x^{k+1} - x^k\right\|^2 < \infty$.

Proof. We first prove (i). Fix any $z \in dom\ g$. Using the definition of x^{k+1} in (3.2) and the strong convexity of the objective in the minimization problem (3.2), we obtain

$$\begin{aligned} g\left(x^{k+1}\right) \leq g(z) + \left\langle -\nabla f\left(y^k\right), x^{k+1} - z\right\rangle + \frac{L}{2}\left\|z - y^k\right\|^2 \\ -\frac{L}{2}\left\|x^{k+1} - y^k\right\|^2 - \frac{L}{2}\left\|x^{k+1} - z\right\|^2. \end{aligned} \tag{3.4}$$

On the other hand, from the fact that ∇f is Lipschitz continuous with a Lipschitz continuity modulus L, we have

$$f\left(x^{k+1}\right) \leq f\left(y^k\right) + \left\langle \nabla f\left(y^k\right), x^{k+1} - y^k\right\rangle + \frac{L}{2}\left\|x^{k+1} - y^k\right\|^2. \tag{3.5}$$

Summing (3.4) and (3.5), we see further that

$$f\left(x^{k+1}\right) + g\left(x^{k+1}\right) \leq f\left(y^k\right) + g(z) + \left\langle \nabla f\left(y^k\right), z - y^k\right\rangle + \frac{L}{2}\left\|z - y^k\right\|^2 - \frac{L}{2}\left\|x^{k+1} - z\right\|^2. \tag{3.6}$$

Then using the convexity of f, we have

$$f\left(y^k\right) - f(z) \leq \left\langle \nabla f\left(y^k\right), y^k - z\right\rangle. \tag{3.7}$$

Combining (3.6) with (3.7) and the fact that $F = f + g$, we obtain that

$$F\left(x^{k+1}\right) \leq F(z) + \frac{L}{2}\left\|z - y^k\right\|^2 - \frac{L}{2}\left\|x^{k+1} - z\right\|^2. \tag{3.8}$$

Then using the definition of the y-update in (3.1) that $y^k - x^k = \beta_k\left(x^k - x^{k-1}\right)$ and (3.8) with $z = x^k$, we obtain that

$$F\left(x^{k+1}\right) - F\left(x^k\right) \leq \frac{L{\beta_k}^2}{2}\left\|x^k - x^{k-1}\right\|^2 - \frac{L}{2}\left\|x^{k+1} - x^k\right\|^2. \tag{3.9}$$

From (3.9) and the definition of $\{H_{k,\gamma}\}$, we see further that

$$\begin{aligned} H_{k+1,\gamma} - H_{k,\gamma} &= F\left(x^{k+1}\right) + \gamma \cdot \frac{L}{2}\left\|x^{k+1} - x^k\right\|^2 - F\left(x^k\right) - \gamma \cdot \frac{L}{2}\left\|x^k - x^{k-1}\right\|^2 \\ &\leqslant (\gamma - 1) \cdot \frac{L}{2}\left\|x^{k+1} - x^k\right\|^2 + \left({\beta_k}^2 - \gamma\right) \cdot \frac{L}{2}\left\|x^k - x^{k-1}\right\|^2. \end{aligned} \tag{3.10}$$

Since $\gamma \in (0,1)$ and $\gamma \geq \sup_k \beta_k$ by our assumption, we have

$$\gamma - 1 \leqslant 0 \text{ and } {\beta_k}^2 - \gamma \leqslant 0.$$

Consequently, $H_{k+1,\gamma} - H_{k,\gamma} \leqslant 0$; i.e., $\{H_{k,\gamma}\}$ is nonincreasing. This proves (i).

From the sequence $\{H_{k,\gamma}\}$ is nonincreasing and the definition of $H_{k,\gamma}$, we see that

$$F\left(x^k\right) \leq H_{k,\gamma} \leq H_{0,\gamma} < \infty.$$

Since F is level bounded, we conclude that $\left\{x^k\right\}$ is bounded, which proves (ii).

Next, recall that $\inf F > -\infty$. Hence, $H_{k,\gamma} = F\left(x^k\right) + \gamma \cdot \frac{L}{2}\left\|x^k - x^{k-1}\right\|^2$ is bounded from below. This together with the fact that $\{H_{k,\gamma}\}$ is nonincreasing, implies that $\{H_{k,\gamma}\}$ is convergent. This proves (iii).

Finally, since $\gamma \in (0,1)$, we have from (3.10) that

$$H_{k+1,\gamma} - H_{k,\gamma} \leqslant -\left(\gamma - {\beta_k}^2\right) \cdot \frac{L}{2}\left\|x^k - x^{k-1}\right\|^2. \tag{3.11}$$

Summing both sides of (3.11) from 1 to N, we see further that

$$0 \leq \sum_{k=1}^{N}\left(\gamma - {\beta_k}^2\right) \cdot \frac{L}{2}\left\|x^k - x^{k-1}\right\|^2 \leq \sum_{k=1}^{N}\left(H_{k,\gamma} - H_{k+1,\gamma}\right) = H_{1,\gamma} - H_{N+1,\gamma}, \tag{3.12}$$

where the nonnegativity follows from the fact that $\gamma \in (0,1)$. Since $\{H_{k,\gamma}\}$ is convergent, letting $N \to \infty$ in (3.12), we conclude that the infinite sum exists and is finite, i.e.,

$$\sum_{k=1}^{\infty}\left(\gamma - {\beta_k}^2\right) \cdot \frac{L}{2}\left\|x^k - x^{k-1}\right\|^2 < \infty.$$

Since $\gamma \geq \sup_k \beta_k$, the conclusion in (iv) follows immediately. This completes the proof. □

Lemma 3.2. *Let $w \in \mathbb{R}^n$ and define an auxiliary sequence $h_k = \frac{1}{2}\left\|x^k - w\right\|^2$, then*

$$h_k - h_{k-1} = \left\langle x^k - w, x^k - x^{k-1}\right\rangle - \frac{1}{2}\left\|x^k - x^{k-1}\right\|^2. \tag{3.13}$$

Proof. By calculating, we obtain that

$$\begin{aligned} h_{k-1} &= \frac{1}{2}\left\|x^{k-1} - w\right\|^2 = \frac{1}{2}\left\|x^{k-1} - x^k + x^k - w\right\|^2 \\ &= \frac{1}{2}\left\langle\left(x^{k-1} - x^k\right) + \left(x^k - w\right), \left(x^{k-1} - x^k\right) + \left(x^k - w\right)\right\rangle \\ &= \frac{1}{2}\left\|x^{k-1} - x^k\right\|^2 + \frac{1}{2}\left\|x^k - w\right\|^2 + \left\langle x^k - w, x^{k-1} - x^k\right\rangle \\ &= \frac{1}{2}\left\|x^k - x^{k-1}\right\|^2 + h_k + \left\langle x^k - w, x^{k-1} - x^k\right\rangle, \end{aligned}$$

which implies that

$$h_k - h_{k-1} = \left\langle x^k - w, x^k - x^{k-1}\right\rangle - \frac{1}{2}\left\|x^k - x^{k-1}\right\|^2.$$

□

3.2 Convergence analysis of $\{x^k\}$ and $\{F(x^k)\}$.

Now we are ready to analyze the convergence behavior of the proposed algorithm $\mathrm{PG_e}$.

Theorem 3.3. *Suppose that there exists $\gamma \in (0,1)$ such that $\gamma \geq \sup_k \beta_k$. Let $\{x^k\}$ be a sequence generated by Algorithm 1. Then any accumulation point of $\{x^k\}$ is a minimizer of F.*

Proof. Let $\bar{x}$ be an accumulation point. Then there exists a subsequence $\{x^{k_i}\}$ such that $\lim_{i\to\infty} x^{k_i} = \bar{x}$. Using the first-order optimality condition of the minimization problem (3.2) at point x^{k_i+1}, we obtain

$$-L\left(x^{k_i+1} - y^{k_i}\right) \in \nabla f\left(y^{k_i}\right) + \partial g\left(x^{k_i+1}\right).$$

Combining this with the definition of y^{k_i} , which is $y^{k_i} = x^{k_i} + \beta_{k_i}\left(x^{k_i} - x^{k_i-1}\right)$, we see further that

$$-L\left[\left(x^{k_i+1} - x^{k_i}\right) - \beta_{k_i}\left(x^{k_i} - x^{k_i-1}\right)\right] \in \nabla f\left(y^{k_i}\right) + \partial g\left(x^{k_i+1}\right). \tag{3.14}$$

On the other hand, by the triangle inequality, we have

$$\left\|x^{k_i+1} - \bar{x}\right\| \leq \left\|x^{k_i+1} - x^{k_i}\right\| + \left\|x^{k_i} - \bar{x}\right\|,$$

from the above relation and the fact that $\left\|x^{k_i+1} - x^{k_i}\right\| \to 0$ from Lemma 3.1(iv) and $\lim_{i\to\infty} x^{k_i} = \bar{x}$, we immediately deduce that $\lim_{i\to\infty} x^{k_i+1} = \bar{x}$.

Taking limit in two sides of (3.14), and recalling that $\lim_{i\to\infty} x^{k_i+1} = \bar{x}$, which together with the continuity of ∇f and the closedness of ∂g imply that

$$0 \in \nabla f(\bar{x}) + \partial g(\bar{x}),$$

meaning that $\bar{x}$ is a minimizer of F. This completes the proof. □

Theorem 3.4. *Suppose that there exists $\gamma \in (0,1)$ such that $\gamma \geq \sup_k \beta_k$. Let $\{x^k\}$ be a sequence generated by Algorithm 1. Then*

$$F\left(x^{k+1}\right) - F(x^*) \leqslant \frac{1}{k}\left\{\frac{L}{2}\left\|x^1 - x^*\right\|^2 + \frac{L\gamma^2}{2(1-\gamma)}\left\|x^1 - x^0\right\|^2 + \frac{\gamma}{1-\gamma}\left(F\left(x^1\right) - F(x^*)\right) + C\right\}.$$

Proof. From (3.10) and $\gamma \geq \sup_k \beta_k$, we have

$$H_{k+1,\gamma} - H_{k,\gamma} \leqslant (\gamma - 1)\cdot\frac{L}{2}\left\|x^{k+1} - x^k\right\|^2 + \left(\gamma^2 - \gamma\right)\cdot\frac{L}{2}\left\|x^k - x^{k-1}\right\|^2,$$

then we obtain upon rearranging terms that

$$\left\|x^{k+1} - x^k\right\|^2 + \gamma\left\|x^k - x^{k-1}\right\|^2 \leqslant \frac{2}{L(1-\gamma)}\left(H_{k,\gamma} - H_{k+1,\gamma}\right). \tag{3.15}$$

Using $y^k - x^k = \beta_k\left(x^k - x^{k-1}\right)$ and (3.8) with $z = x^*$, we have

$$\begin{aligned} &F\left(x^{k+1}\right) - F\left(x^*\right) \\ \leqslant &\frac{L}{2}\left\|y^k - x^*\right\|^2 - \frac{L}{2}\left\|x^{k+1} - x^*\right\|^2 \\ = &\frac{L}{2}\left\|x^k - x^*\right\|^2 + \frac{L{\beta_k}^2}{2}\left\|x^k - x^{k-1}\right\|^2 + L\beta_k\left\langle x^k - x^*, x^k - x^{k-1}\right\rangle - \frac{L}{2}\left\|x^{k+1} - x^*\right\|^2. \end{aligned} \tag{3.16}$$

In the following, we show the sequence $\sum_{i=1}^{k}\left(H_{i+1,\gamma}-F\left(x^{*}\right)\right)$ is bounded. First, by the definition of $\{H_{k,\gamma}\}$, we have

$$
\begin{aligned}
&H_{k+1,\gamma}-F\left(x^{*}\right)=F\left(x^{k+1}\right)-F\left(x^{*}\right)+\gamma\cdot\frac{L}{2}\left\|x^{k+1}-x^{k}\right\|^{2}\\
\leqslant&\frac{L}{2}\left\|x^{k}-x^{*}\right\|^{2}+\frac{L{\beta_k}^{2}}{2}\left\|x^{k}-x^{k-1}\right\|^{2}+L\beta_{k}\left\langle x^{k}-x^{*},x^{k}-x^{k-1}\right\rangle-\frac{L}{2}\left\|x^{k+1}-x^{*}\right\|^{2}\\
&+\gamma\cdot\frac{L}{2}\left\|x^{k+1}-x^{k}\right\|^{2}\\
=&\frac{L}{2}\left\|x^{k}-x^{*}\right\|^{2}-\frac{L}{2}\left\|x^{k+1}-x^{*}\right\|^{2}+\frac{L{\beta_k}^{2}}{2}\left\|x^{k}-x^{k-1}\right\|^{2}+\gamma\cdot\frac{L}{2}\left\|x^{k+1}-x^{k}\right\|^{2}\\
&+L\beta_{k}\left(h_{k}-h_{k-1}+\frac{1}{2}\left\|x^{k}-x^{k-1}\right\|^{2}\right)\\
\leqslant&\frac{L}{2}\left\|x^{k}-x^{*}\right\|^{2}-\frac{L}{2}\left\|x^{k+1}-x^{*}\right\|^{2}+L\beta_{k}\left(h_{k}-h_{k-1}\right)+\frac{L\gamma}{2}\left\|x^{k}-x^{k-1}\right\|^{2}\\
&+\frac{L\gamma}{2}\left(\gamma\left\|x^{k}-x^{k-1}\right\|^{2}+\left\|x^{k+1}-x^{k}\right\|^{2}\right)\\
\leqslant&\frac{L}{2}\left\|x^{k}-x^{*}\right\|^{2}-\frac{L}{2}\left\|x^{k+1}-x^{*}\right\|^{2}+L\beta_{k}\left(h_{k}-h_{k-1}\right)+\frac{L\gamma}{2}\left\|x^{k}-x^{k-1}\right\|^{2}\\
&+\frac{\gamma}{1-\gamma}\left(H_{k,\gamma}-H_{k+1,\gamma}\right),
\end{aligned}
\tag{3.17}
$$

where the first inequality follows from (3.16), the second equality is due to Lemma 3.2 with $w=x^{*}$, the last inequality holds from (3.15). And then summing both sides of (3.17) from 1 to k, we see further that

$$
\begin{aligned}
&\sum_{i=1}^{k}\left(H_{i+1,\gamma}-F\left(x^{*}\right)\right)\\
&\leqslant\frac{L}{2}\left\|x^{1}-x^{*}\right\|^{2}-\frac{L}{2}\left\|x^{k+1}-x^{*}\right\|^{2}+\frac{\gamma}{1-\gamma}\left(H_{1,\gamma}-H_{k+1,\gamma}\right)+\sum_{i=1}^{k}L\beta_{i}\left(h_{i}-h_{i-1}\right)\\
&\quad+\sum_{i=1}^{k}\frac{L\gamma}{2}\left\|x^{i}-x^{i-1}\right\|^{2}\\
&\leqslant\frac{L}{2}\left\|x^{1}-x^{*}\right\|^{2}-\frac{L}{2}\left\|x^{k+1}-x^{*}\right\|^{2}+\frac{\gamma}{1-\gamma}\left(H_{1,\gamma}-H_{k+1,\gamma}\right)+C\\
&=\frac{L}{2}\left\|x^{1}-x^{*}\right\|^{2}+\frac{\gamma}{1-\gamma}\left(F\left(x^{1}\right)-F\left(x^{k+1}\right)\right)+\frac{L\gamma^{2}}{2\left(1-\gamma\right)}\left\|x^{1}-x^{0}\right\|^{2}+C\\
&\leqslant\frac{L}{2}\left\|x^{1}-x^{*}\right\|^{2}+\frac{\gamma}{1-\gamma}\left(F\left(x^{1}\right)-F\left(x^{*}\right)\right)+\frac{L\gamma^{2}}{2\left(1-\gamma\right)}\left\|x^{1}-x^{0}\right\|^{2}+C,
\end{aligned}
$$

where the second inequality holds from Lemma 3.1 (ii) and (iv), there exists a constant C such that $\sum_{i=1}^{k}L\beta_{i}\left(h_{i}-h_{i-1}\right)+\sum_{i=1}^{k}\frac{L\gamma}{2}\left\|x^{i}-x^{i-1}\right\|^{2}<C$, the equality is due to the definition

of $H_{k+1,\gamma}$ and $H_{1,\gamma}$, respectively. Then,

$$\begin{aligned}
&F\left(x^{k+1}\right) - F\left(x^*\right) \\
&= H_{k+1,\gamma} - F\left(x^*\right) - \gamma \cdot \frac{L}{2}\left\|x^{k+1} - x^k\right\|^2 \\
&\leqslant H_{k+1,\gamma} - F\left(x^*\right) \leqslant \frac{1}{k}\sum_{i=1}^{k}\left(H_{i+1,\gamma} - F\left(x^*\right)\right) \\
&\leqslant \frac{1}{k}\left\{\frac{L}{2}\left\|x^1 - x^*\right\|^2 + \frac{\gamma}{1-\gamma}\left(F\left(x^1\right) - F\left(x^*\right)\right) + \frac{L\gamma^2}{2\left(1-\gamma\right)}\left\|x^1 - x^0\right\|^2 + C\right\}.
\end{aligned}$$

This completes the proof. □

Remark 3.5. Since FISTA is a special extrapolation algorithm, whose extrapolation parameter is choose as $\beta_k = \frac{\theta_{k-1}-1}{\theta_k}$, $\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$. By simply calculation, one can deduce that $0 \leq \beta_k < 1$ for all k, and the sequence $\{\beta_k\}$ is nondecreasing. But the supreme of $\{\beta_k\}$ in FISTA is 1. FISTA with restart [7] is based on FISTA. The restart schemes include fixed restart and adaptive restart. In the fixed restart scheme, one can choose a positive integer K and restart every K iterations, while in the adaptive restart (gradient scheme), the algorithm restarts whenever $\left\langle y^k - x^{k+1}, x^{k+1} - x^k\right\rangle > 0$. As K is a fixed number, using this and the fact that the supreme of $\{\beta_k\}$ in FISTA is 1, we can immediately deduce that $\sup_k \beta_k < 1$ if the fixed restart scheme is invoked in FISTA. Hence there must exist a $\gamma > 0$ such that the assumption $\sup_k \beta_k \leq \gamma$ holds in FISTA with fixed restart scheme. Hence, by Theorem 3.4, one can immediately obtain that the global convergence rate of FISTA with fixed restart scheme or FISTA with both fixed restart and adaptive restart schemes is $O(1/k)$.

4 Numerical Experiments

In this section, we perform numerical experiments to study the behaviors of $\mathrm{PG_e}$. Since FISTA with both fixed restart scheme and adaptive restart scheme is a special case of $\mathrm{PG_e}$, we use FISTA with both fixed restart scheme and adaptive restart scheme to represent $\mathrm{PG_e}$. All codes are written in Matlab, and the experiments are performed in Matlab 2016a on a 64-bit PC with an Intel(R) Core(TM) i5-8250U CPU (1.80GHz) and 8GB of RAM.

we consider the l_1 regularized logistic regression problem:

$$v_{\log} := \min_{\tilde{x}\in\mathbb{R}^n, x_0\in\mathbb{R}} \sum_{i=1}^{m} \log\left(1 + \exp\left(-b_i\left(a_i^T\tilde{x} + x_0\right)\right)\right) + \lambda\|\tilde{x}\|_1, \tag{4.1}$$

where $a_i \in \mathbb{R}^n, b_i \in \{-1, 1\}, i = 1, 2, ..., m$, with b_i not all the same, $m < n$, and $\lambda > 0$ is the regularization parameter. It is easy to see that (4.1) is in the form of (1.1) with

$$f\left(x\right) = \sum_{i=1}^{m} \log\left(1 + \exp\left(-b_i(Dx)_i\right)\right), g\left(x\right) = \lambda\|\tilde{x}\|_1, \tag{4.2}$$

where $x := (\tilde{x}, x_0) \in \mathbb{R}^{n+1}$, and D is the matrix whose ith row is given by $\begin{pmatrix} a_i^T & 1\end{pmatrix}$. Moreover, one can show that ∇f is Lipschitz continuous with modulus $0.25\lambda_{\max}\left(D^TD\right)$. Thus, in our algorithms below we take $L = 0.25\lambda_{\max}\left(D^TD\right)$.

We now compare the numerical performance of $\mathrm{PG_e}$ with proximal gradient algorithm (PG) and FISTA. And we consider random instances for our experiments. For each (m, n, s) = (500, 5000, 50) and (800, 8000, 80), we generate an $m \times n$ matrix A with i.i.d. standard Gaussian entries. We then choose a support set T of size s uniformly at random and generate an s-sparse vector $\hat{x}$ supported on T with i.i.d. standard Gaussian entries. The vector b is then generated as $b = sign\,(A\hat{x} + ce)$, where c is chosen uniformly at random from $[0, 1]$.

Our computational results are presented in Figures 1 and 2. In part (a) of each figure, we plot $\left\|x^k - x^*\right\|$ against the number of iterations, where x^* denotes the approximate solution obtained at termination of the respective algorithm, while in part (b) of each figure, we plot $\left|F\left(x^k\right) - F_{\min}\right|$ against the number of iterations, where $F_{\min}$ denotes the minimum of the three objective values obtained from the above three algorithms. Moreover, compared with FISTA and the proximal gradient algorithm, $\mathrm{PG_e}$ performs better.

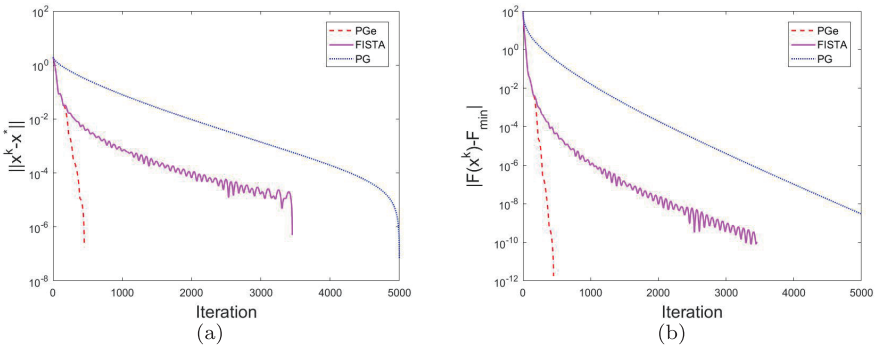

Figure 1: $n = 5000, m = 500, s = 50$

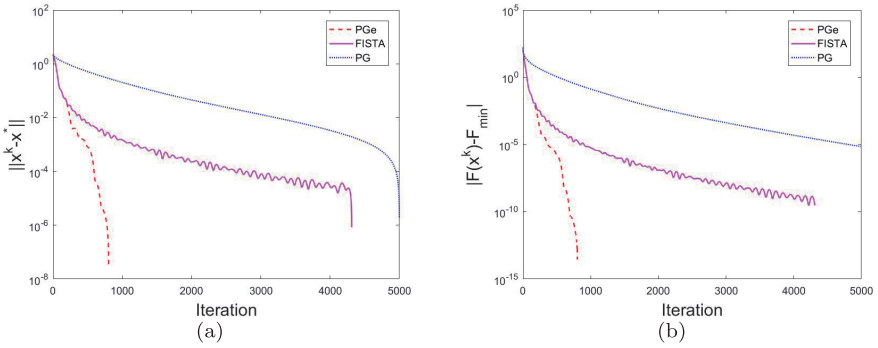

Figure 2: $n = 8000, m = 800, s = 80$

5 Conclusions

In this paper, we mainly study the convergence behavior of proximal gradient algorithm with extrapolation ($\mathrm{PG_e}$) for solving problem (1.1). We first prove the subsequential convergence of the iterate sequence $\{x^k\}$ generated by $\mathrm{PG_e}$, then under the assumption $\sup_k \beta_k \leq \gamma$, we establish the global convergence rate of $\mathrm{PG_e}$, which is $O\,(1/k)$. Hence, we can obtain the

convergence rate of FISTA with fixed restart scheme is $O(1/k)$ as a corollary. At last, some numerical experiments have been performed to illustrate the theoretical results.

Acknowledgements

The authors would like to thank the editor and anonymous reviewers for their insight and helpful comments and suggestions which improve the quality of the paper.

References

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

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

[3] E.J. Candès and B. Recht, Exact matrix completion via convex optimization, *Found. Comput. Math.* 9 (2009) 717–772.

[4] E.J. Candès and T. Tao, Decoding by linear programming, *IEEE Trans. Inform. Theory* 51 (2005) 4203–4215.

[5] X. Chen, W. Pan, J.K. Kwok and J.G. Carbonell, Accelerated gradient method for multi-task sparse learning problem, in: *2009 Ninth IEEE International Conference on Data Mining*, 2009, pp. 746–751.

[6] D.L. Donoho, Compressed sensing, *IEEE Trans. Inform. Theory* 52 (2006) 1289–1306.

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

[8] S. Ji and J. Ye, An accelerated gradient method for trace norm minimization, in: *Proceedings of the 26th International Conference on Machine Learning*, Canada Montreal, 2009.

[9] K.C. Kiwiel, A bundle bregman proximal method for convex nondifferentiable minimization, *Math. Program.* 85 (1999) 241–258.

[10] P.L. Lions and B. Mercier, Splitting algorithms for the sum of two nonlinear operators, *SIAM J. Numer. Anal.* 16 (1979) 964–979.

[11] M. Mäkelä, Survey of bundle methods for nonsmooth optimization, *Optim. Methods Softw.* 17 (2002) 1–29.

[12] Y. Nesterov, Smooth minimization of non-smooth functions, *Math. Program.* 103 (2005) 127–152.

[13] Y. Nesterov, Dual extrapolation and its applications to solving variational inequalities and related problems, *Math. Program.* 109 (2007) 319–344.

[14] Y. Nesterov, *Gradient Methods for Minimizing Composite Objective Function*, CORE Discussion Paper, 2007.

[15] Y. Nesterov, A method of solving a convex programming problem with convergence rate $O\left(1/k^2\right)$, *Soviet Math. Dokl.* 27 (1983) 372–376.

[16] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers, Boston, 2004.

[17] H. Schramm and J. Zowe, A version of the bundle idea for minimizing a nonsmooth functions: conceptual idea, convergence analysis, numercial results, *SIAM J. Optim.* 2 (1992) 121–152.

[18] P. Tseng, Approximation accuracy, gradient methods, and error bound for structured convex optimization, *Math. Program.* 125 (2010) 263–295.

Manuscript received 11 November 2021
revised 2 February 2022
accepted for publication 2 March 2022

Mengxi Pan
School of Science
Hebei University of Technology, Tianjin, China
E-mail address: panmengxi21@163.com

Bo Wen
Institute of Mathematics
Hebei University of Technology, Tianjin, China
E-mail address: wenbohit@163.com