manuscript No.
(will be inserted by the editor)

On the Convergence Results of a class of Nonmonotone Accelerated Proximal Gradient Methods for Nonsmooth and Nonconvex Minimization Problems

Ting Wang[1] · Hongwei Liu[1] ·

Received: date / Accepted: date

Abstract In this paper, we consider a class of nonsmooth problem that is the sum of a Lipschitz differentiable function and a nonsmooth and proper lower semicontinuous function. We discuss here the convergence rate of the function values for a nonmonotone accelerated proximal gradient method, which proposed in [35] for the nonconvex case but with incomplete theoretical analysis. Further, we proposed a hybrid proximal gradient method for the nonconvex setting and show the corresponding theoretical analysis under the assumption that objective function has the Kurdyka-Łojasiewicz property. Numerical experiments on nonconvex models to demonstrate the advantage of the proposed method.

Keywords accelerated proximal gradient method, Kurdyka-Łojasiewicz property, Convergence

Mathematics Subject Classification (2000) 94A12 · 65K10 · 94A08 · 90C25 ·

1 Introduction

The proximal gradient method [2,41] is a benchmark approach for minimizing the composite optimization problem of the sum of two functions, one of which is smooth with Lipschitz continuous gradient and the other is nonsmooth and proper lower semicontinuous. For the setting that objective is convex, it is well known that this method globally convergent to an optimal solution with the complexity $o\left(k^{-1}\right)$ on the objective function values [14,20]. Many stepsize rules like the line search in [12,23] are also be proposed to speed up the performance and achieved the same complexity. Meanwhile, Bauschke, Bolte and Teboulle [17] introducing the so-called NoLips algorithm close to the PG method with the involvement of Bregman distance [10]. Due to the PG method's simplicity and efficiency, a lot of variants of PG method have been proposed in order to acquire better complexity and faster convergence speed and numerical performance. One of the popular strategies is to incorporate an inertial force,

Hongwei Liu (✉)
E-mail: hwliuxidian@163.com

Ting Wang
E-mail: wangting_7640@163.com

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

also called extrapolation, into the iterative scheme. It can be read as

$$\begin{aligned} y_{k+1} &= x_k + \alpha_k \left(x_k - x_{k-1}\right) \\ x_{k+1} &= \operatorname{prox}_{\lambda_{k+1} g}\left(y_{k+1} - \lambda_{k+1} \nabla f\left(y_{k+1}\right)\right), \end{aligned}$$

where λ_k denotes stepsize and α_k denotes the inertial parameter and

$$\operatorname{prox}_{\lambda g}(x) = \arg\min_{u \in R^n} \left\{ g(u) + \frac{1}{2\lambda} \|x - u\|^2 \right\}.$$

One of the most famous choices of inertial parameter is the FISTA [12] scheme presented as

$$\alpha_k = \frac{t_k - 1}{t_{k+1}}, \text{ where } t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2} \text{ and } t_1 = 1.$$

For the convex setting, FISTA improves the complexity for function values to $O\left(k^{-2}\right)$ but no convergence of iterates. However, under the error bound condition, it has been proved in [31] that the sequence of iterates generated by FISTA converges strongly to the minimizer with $O\left(k^{-2}\right)$ rate of convergence, meanwhile, the convergence rate of function value is $o\left(k^{-6}\right)$. Many efforts on the modification of FISTA can be found in [13, 22, 40], and the corresponding theoretical analyses have been well studied [5, 6, 8, 9, 24, 29, 32, 42].

Subsequently, scholars focus on nonconvex setting. While for the nonconvex setting, the convergence of the whole sequence is hard to establish. Some works considering the problem of " convex + nonconvex ", i.e., the nonsmooth part of the objective function is convex, while the smooth counterpart is allowed to be nonconvex, for example, Ochs [39] showed that the sequence generated by an Inertial proximal gradient method is globally convergent to a stationary point under the assumption that objective function satisfies the Kurdyka-Łojasiewicz property [30]. Wenbo, under the error bound condition [43] (Bolte2015 [16] showed that there has a quantitative relationship between the error bound condition and the Kurdyka-Łojasiewicz property), showed that the R-linearly convergence rates of the sequences of objective values and iterates if the extrapolation coefficients are chosen below a shreshold [44]. The methods based on the Bregman distance were also studied in [7, 18, 19, 28, 34, 37].

In this paper, we focus on the problem with fully nonconvex setting, which can be described as follows:

$$\text{(P)} \qquad \min_x F(x) = f(x) + g(x),$$

where f is a smooth nonconvex function with Lipschitz continuous gradient and g is a proper lower semicontinuous, nonconvex and nonsmooth function. Furthermore, we require F to be coercive, i.e., $\|x\|_2 \to \infty$ implies that $F(x) \to \infty$ and bounded from below by some value $\inf F > -\infty$. For this " nonconvex + nonconvex " setting considered in this paper, there have several works to solve it. Under the mild assumption that objective function F or an appropriate regularization of the objective function F satisfies the KŁ property, the convergence analyses of all these works be achieved partly, but, incomplete. For example, Lihuan [35] proposed an nonmonontone accelerated proximal gradient method for this fully nonconvex problem (P), however, the author just proved that depending on subsequence, the accumulation point of the generated sequence of iterates converges to stationary point of the objective function. The convergence of the whole sequence of iterates is unknown, and the convergence rate for the objective function values or sequence of iterates is still unknown. Bolte [16] based on the Bregman distance, proved that every sequence of iterates generated by an inertial forward-backward algorithm converges

to a critical point of the objective function. However, no the corresponding results of convergence rate. Wu [45] proposed an inertial Bregman proximal gradient method with two different inertial steps and showed the sequence of iterates is linearly convergent under the assumption that objective function satisfies the KŁ property with the exponent belongs to $\left[\frac{1}{2}, 1\right)$. They didn't prove the convergence rate for the exponent belongs to $\left[0, \frac{1}{2}\right)$. A complete convergence analysis can be founded in László [36], in which the author consider a special case of problem (P) with $f = 0$. Since the above analyses, the aim of this paper is to further exploit the convergence result of the proximal gradient method for solving the problem (P).

Our contributions We first prove a supplementary conclusion that is the convergence rate of objective function values of the nonmonotone accelerated proximal gradient method proposed in [35], while the convergence of iterates is still unknown. Hence, based on the idea of [35], we propose a hybrid proximal gradient methods for solving the nonconvex problem (P) and show that all the accumulation point of the iterates generated by our algorithm converges to critical point of F. Further, we obtain the convergence rates of function values and the iterates by assuming that the objective function F satisfies the KŁ property.

The reminder of the paper is organized as follows. Some preliminaries are summarized in Section 2. In Section 3, by assuming that objective function F satisfies the KŁ property, we bulid the convergence rate of function values for the nonmonotone accelerated proximal gradient method (nmAPG) proposed in [35]. In Section 4, we provide a hybrid proximal gradient algorithm, which is a modification of nmAPG, and show that any accumulation point of generated iterates converges to critical point. Then, based on the assumption that objective function F satisfies the KŁ property, we show the convergence rates of function values and iterates. Numerical results are reported in Section 5.

2 preliminaries

In this section we recall some notions and results which are needed throughout this paper. The domain of function $g : R^n \to (-\infty, +\infty]$ is defined by $\operatorname{dom} g = \{x \in R^n : g(x) < +\infty\}$. We say that g is property if $\operatorname{dom} g \neq \emptyset$. The Fermat rule reads in this nonsmooth setting as if $x \in R^n$ is a local minimizer of g, then, $0 \in \partial g(x)$. Notice that in case g is continuously differentiable around $x \in R^n$, we have $\partial g(x) = \{\nabla g(x)\}$. We denote by

$$\operatorname{crit} g = \left\{x \in R^n : 0 \in \partial g(x)\right\}$$

the set of (limiting)-critical points of g. Let us mention also the following subdifferential rule: if $g : R^n \to (-\infty, +\infty]$ is proper and lower semicontinuous and $f : R^n \to R$ is a continuously differentiable function, then $\partial(f + g)(x) = \partial g(x) + \nabla f(x)$. In addition, a simpler notation will be adopted in the following full text:

$$T_{\lambda g}(y) := \operatorname{prox}_{\lambda g}(y - \lambda \nabla f(y)).$$

Next, we give a well-known lemma for smooth functions.

Lemma 2.1 *[11, 38] Let $f : R^n \to R$ be a continuously differentiable function with gradient ∇f assumed L_f-Lipschitz continuous. Then, for any $u, v \in R^n$, we have*

$$|f(u) - f(v) - \langle u - v, \nabla f(v)\rangle| \leq \frac{L_f}{2}\|u - v\|^2. \tag{1}$$

We turn now our attention to functions satisfying the Kurdyka-Łojasiewicz property. This class of functions will play a crucial role when proving the convergence results of the proposed algorithm.

For $\eta \in (0, +\infty]$, we denote by Θ_η the class of concave and continuous functions $\varphi : [0, \eta) \to [0, +\infty)$ such that $\varphi(0) = 0$, φ is continuously differentiable on $(0, \eta)$, continuous at 0 and $\varphi'(s) > 0$ for all $s \in (0, \eta)$.

Definition 2.1 [15] (Kurdyka-Łojasiewicz property and KŁ function) Let $F : R^n \to R$ be a differentiable function. We say that F satisfies the Krudyka-Lojasiewica (KŁ) property at $\bar{x} \in R^n$ if there exists $\eta \in (0, +\infty]$, a neighborhood U of $\bar{x}$ and a function $\varphi \in \Theta_\eta$ such that for all x in the intersection

$$U \cap \left\{x \in R^m : F(\bar{x}) < F(x) < F(\bar{x}) + \eta\right\}$$

the following, so called KŁ inequality, holds

$$\varphi'(F(x) - F(\bar{x}))\, dist(0, \partial F(x)) \geq 1.$$

If F satisfies the KŁ property at each point in R^m, then F is called a KŁ function.

An important role in our convergence analysis will be played by the following uniformized KL property given in [15].

Lemma 2.2 *[15] Let $\Omega \subseteq R^n$ be a compact set and let $F : R^n \to (-\infty, +\infty]$ be a proper and lower semicontinuous function. Assume that F is constant on Ω and F satisfies the KŁ property at each point of Ω. Then, there exist $\varepsilon, \eta > 0$ and $\varepsilon, \eta > 0, \varphi \in \Theta_\eta$ such that for all $\bar{x} \in \Omega$ and for all x in the intersection*

$$\left\{x \in R^n : dist(x, \Omega) < \varepsilon\right\} \cap \left\{x \in R^n : F(\bar{x}) < F(x) < F(\bar{x}) + \eta\right\},$$

the following inequality holds

$$\varphi'(F(x) - F(\bar{x}))\, dist(0, \partial F(x)) \geq 1.$$

3 The convergence rate of function values for the nonmonotone APG with fixed stepsize

Inspired by the monotone FISTA (MFISTA) algorithm for convex case proposed by Beck [13], Lihuan in [35] proposed a monotone APG method which can be used for the nonconvex case, further, the author relaxed the monotone APG to a nonmonotone one by using a relaxed sufficient descent condition. We give the concrete scheme as follows:

Denote $\{1, 2, \cdots, k\} = \Omega_1 \cup \Omega_2$, where $\Omega_1 := \left\{k_1, k_2, \cdots, k_j, \cdots\right\}$ is the set such that (3) in Algorithm 1 be executed for all $k = k_j \in \Omega_1$ and $\Omega_2 := \left\{m_1, m_2, \cdots, m_j, \cdots\right\}$ such that (4) in Algorithm 1 be executed for all $k = m_j \in \Omega_2$. Note that $\Omega_1 \cap \Omega_2 = \emptyset$.

With a any small positive constant δ, the author extend the monontone APG proposed in [13] by making sure the following relaxed sufficient descent conditions hold rather that the descent condition in [13].

$$F\left(x_{k_j}\right) = F\left(z_{k_j}\right) \leq c_{k_j - 1} - \delta \left\|x_{k_j} - y_{k_j}\right\|^2, \forall k_j \in \Omega_1$$
$$F\left(x_{m_j}\right) \leq F\left(v_{m_j}\right) \leq c_{m_j - 1} - \delta \left\|v_{m_j} - x_{m_j - 1}\right\|^2, \forall m_j \in \Omega_2,$$

Algorithm 1 Nonmonotone APG with fixed stepsize (nmAPG)

Step 0. Initial $z_0 = x_0 = x^-, t_0 = 1, t^- = 0$, $\delta > 0, c_0 = F(x_0), q_0 = 1$, $\lambda_x < \frac{1}{L}$, and $\lambda_y < \frac{1}{L}$.

Step k. Compute that

$$y_k = x_{k-1} + \frac{t_{k-2}}{t_{k-1}}(z_{k-1} - x_{k-1}) + \frac{t_{k-2}-1}{t_{k-1}}(x_{k-1} - x_{k-2}) \text{ and } z_k = T_{\lambda_y g}(y_k) \tag{2}$$

If $F(z_k) \leq c_{k-1} - \delta\|z_k - y_k\|^2$

$$x_k = z_k \tag{3}$$

else

$$\begin{aligned} v_k &= T_{\lambda_x g}(x_{k-1}) \\ x_k &= \begin{cases} z_k, & \text{if } F(z_k) \leq F(v_k), \\ v_k, & \text{otherwise.} \end{cases} \end{aligned} \tag{4}$$

end

$$t_k = \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2} \tag{5}$$

$$c_k = \frac{\eta q_{k-1} c_{k-1} + F(x_k)}{q_k}, \text{ where } q_k = 1 + \eta q_{k-1}. \tag{6}$$

Then, using the definition of $\{c_k\}$, it can be easily obtained that

$$c_{k_j} \leq c_{k_j-1} - \frac{\delta}{q_{k_j}}\left\|x_{k_j} - y_{k_j}\right\|^2, \forall k_j \in \Omega_1 \tag{7}$$

$$c_{m_j} \leq c_{m_j-1} - \frac{\delta}{q_{m_j}}\left\|v_{m_j} - x_{m_j-1}\right\|^2, \forall m_j \in \Omega_2. \tag{8}$$

Here, we recall some results proved in [35].

Lemma 3.1 *[35] In Algorithm 1, we have* $F(x_k) \leq c_k \leq A_k, A_k = \frac{\sum_{i=1}^k F(x_i)}{k}$ *and* $c_{k+1} \leq c_k$.

Theorem 3.1 *[35] Let* $\{x_k\}, \{y_k\}$ *and* $\{v_k\}$ *be generated by Algorithm 1. Then,* $\left\{x_{k_j}\right\}, \left\{y_{k_j}\right\}$ *and* $\left\{v_{k_j}\right\}$ *where* $k_j \in \Omega_1$ *generated by Algorithm 1 are bounded, and*

1. If Ω_1 *or* Ω_2 *is finite, then for any accumulation point* x^* *of* $\{x_k\}$, *we have* $0 \in \partial F(x^*)$.

2. If Ω_1 *and* Ω_2 *are both infinite, then for any accumulation point* x^* *of* $\left\{x_{k_j}\right\}$, y^* *of* $\left\{y_{k_j}\right\}$, *where* $k_j \in \Omega_1$ *and any accumulation point* v^* *of* $\left\{v_{m_j}\right\}$, x^* *of* $\left\{x_{m_j}\right\}$ *where* $m_j \in \Omega_2$, *we have* $0 \in \partial F(x^*)$, $0 \in \partial F(y^*)$ *and* $0 \in \partial F(v^*)$.

We can see in Theorem 3.1 that the author not obtain the convergence rate of the whole sequence of iterates, but proved that the accumulation point of sequence generated by Algorithm 1 converges to the stable point depending on the set Ω_1 and Ω_2. In addition, the author not obtain the convergence rate of function values. Through our further analysis of this algorithm, we notice that the convergence rate of the function values can be achieved. In the following, we provide complementary results on the convergence rate of function values under the assumption that objective F satisfies the KŁ property.

Note that several algorithms have been shown to converge [4], and an abstract convergence theorem for descent methods with certain properties is proved [1,3,21,26]. Obviously, the algorithm nmAPG not a descent method,

therefore this abstract convergence theorem is not applicable to it. While, from Lemma 3.1, we know that the sequence $\{c_k\}$ has the descent property. Hence, we can prove the convergence rate of $\{c_k - F^*\}$ depending on the sets Ω_1 and Ω_2, further, the convergence rate of $\{|F(x_k) - F^*|\}$ can be obtained.

Now, we show this complementary theorem.

Lemma 3.2 *Let $\{x_k\}$ and $\{c_k\}$ generated by Algorithm 1. Then, we have that $\{F(x_k)\}$ is convergence. Further, taking $\lim_{k\to\infty} F(x_k) = F^*$, we have $c_k \to F^*$ and $c_k \geq F^*$.*

Proof From Lemma 3.1 and the assumption that F is coercive, we have $\{c_k\}$ is nonincreasing and lower bounded. Hence, $\{c_k\}$ convergence. Assume that $c_k \to F^*$ as $k \to \infty$. Then, we have $\{F(x_k)\}$ converges to F^* since the definition of c_k in (5). The result $c_k \geq F^*$ can be directly obtained since $\{c_k\}$ is decreasing.

Remark. In the rest of this section, we suppose that $c_k > F^*$ for all k. If there exists $\dot{k}$ such that $c_k = F^*$ for $k > \dot{k}$, which means that $c_k = c_{k+1} = \cdots = F^*$. Then, by (7) and (8), we have x_k is stationary point, i.e., the algorithm terminates in finite steps.

Theorem 3.2 *Let $\{x_k\} := \{x_{k_j}\} \cup \{x_{m_j}\}$ be generated by Algorithm 1 with $k_j \in \Omega_1$ and $m_j \in \Omega_2$. Denote that $X^*_{\Omega_1}$ is the set of all accumulation points of $\{x_{k_j}\}_{k_j\in\Omega_1}$; $X^*_{\Omega_2}$ is the set of all accumulation points of $\{x_{m_j}\}_{m_j\in\Omega_2}$. Then, we have $F(X^*_{\Omega_1}) = F^* = F(X^*_{\Omega_2})$.*

Proof. For any $\bar{x} \in X^*_{\Omega_1}$, there exists a $\{x_{k_{j_i}}\}$ such that $\lim_{i\to\infty} x_{k_{j_i}} = \bar{x}$. It follows that

$$F(\bar{x}) \leq \liminf_{i\to\infty} F\left(x_{k_{j_i}}\right) = \lim_{j\to\infty} F\left(x_{k_j}\right) = F^* \tag{9}$$

from the fact that F is lower semicontinuous. In addition, using the inequality (93) with $k := k_j$ and $z := \bar{x}$ that is

$$F\left(x_{k_{j_i}}\right) + \left(\frac{1}{2\lambda} - \frac{L_f}{2}\right)\left\|x_{k_{j_i}} - y_{k_{j_i}}\right\|^2 \leq F(\bar{x}) + \left(\frac{1}{2\lambda} + \frac{L_f}{2}\right)\left\|\bar{x} - y_{k_{j_i}}\right\|^2, \tag{10}$$

and the fact that $\lim_{i\to\infty}\left\|x_{k_{j_i}} - y_{k_{j_i}}\right\|^2 = 0$ and $\lim_{i\to\infty}\left\|\bar{x} - y_{k_{j_i}}\right\|^2 = 0$, we have

$$F^* = \lim_{j\to\infty} F\left(x_{k_j}\right) = \limsup_{i\to\infty} F\left(x_{k_{j_i}}\right) \leq F(\bar{x}). \tag{11}$$

Combining (9) and (11), we have

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

Hence, the conclusion follows from the arbitrariness of $\bar{x}$. The result for the set $X^*_{\Omega_2}$ can be obtained similarly.

Theorem 3.3 *Let $\{x_k\}$ be generated by Algorithm 1. Assume that F satisfies the KŁ property at each point of* crit$F$*, and the desingularising function has the form of $\varphi(t) = \frac{C}{\theta}t^\theta$ for some $C > 0$, $\theta \in (0,1)$. Then,*
(1) If $\theta \in \left[\frac{1}{2}, 1\right)$, there exists $\dot{K} \in (0,1)$ such that

$$\left|F(x_k) - F^*\right| = O\left(\dot{K}^{\frac{k}{2}}\right).$$

(2) If $\theta \in \left(0, \frac{1}{2}\right)$,

$$\left|F(x_k) - F^*\right| = O\left(k^{-\frac{1}{1-2\theta}}\right).$$

Proof **Step 1.** We first prove that if Ω_1 is infinite, then, for any $k_j \in \Omega_1$, there exists $\tilde{K} \in (0,1)$ such that

$$c_{k_j} - F^* \leq \tilde{K}^{j-j_0}\left(c_{k_{j_0}} - F^*\right), \quad \text{for all } \theta \in \left[\frac{1}{2}, 1\right), \tag{12}$$

and

$$c_{k_j} - F^* = r_{k_j} \leq \left(\frac{2}{K_0\left(1-2\theta\right)\left(j-j_0\right)}\right)^{\frac{1}{1-2\theta}}, \quad \text{for all } \theta \in \left(0, \frac{1}{2}\right). \tag{13}$$

In addition, if Ω_2 is infinite, then, for any $m_j \in \Omega_2$, we have similarly results for $\left\{c_{m_j} - F^*\right\}$.

– The proof of (12)

Following from $\operatorname{dist}\left(x_{k_j}, X^*_{\Omega_1}\right) \to 0$, $F\left(x_{k_j}\right) \to F^*$, and the fact that $X^*_{\Omega_1}$ is compact and $X^*_{\Omega_1} \subset \operatorname{crit} F$, using the uniform KŁ property in Lemma 2.2 with $\Omega := X^*_{\Omega_1}$, then, there exist $\varepsilon, \eta > 0$ and j_0 such that for any $j > j_0$, if $F\left(x_{k_j}\right) - F^* > 0$, we have

$$x_{k_j} \in \left\{v : \operatorname{dist}\left(v, X^*_{\Omega_1}\right) < \varepsilon\right\} \cap \left\{v : F^* < F\left(v\right) < F^* + \eta\right\}$$

and

$$1 \leq \left(\varphi'\left(F\left(x_{k_j}\right) - F^*\right)\right)^2 dist\big(0, \partial F\left(x_{k_j}\right)\big)^2. \tag{14}$$

Case 1. If $F\left(x_{k_j}\right) - F^* > 0$, using (14), we have

$$\begin{aligned} 1 &\leq \left(\frac{1}{\lambda} + L_f\right)^2 \left(\varphi'\left(F\left(x_{k_j}\right) - F^*\right)\right)^2 \left\|x_{k_j} - y_{k_j}\right\|^2 \\ &\leq K\big(F\left(x_{k_j}\right) - F^*\big)^{2\theta-2}\left(c_{k_j-1} - c_{k_j}\right), \quad \forall j > j_0, \end{aligned} \tag{15}$$

where $K = C^2\big(\frac{1}{\lambda} + L_f\big)^2 \frac{1}{\delta(1-\eta)}$ follows from (7) and

$$\lim_{k\to\infty} q_k = \lim_{k\to\infty} \sum\nolimits_{i=0}^{k-1} \eta^i = \frac{1}{1-\eta} \quad \text{and} \quad 1 \leq q_k \leq \frac{1}{1-\eta}, \ \forall k. \tag{16}$$

Since $0 < 2 - 2\theta \leq 1$, we have $\big(F\left(x_{k_j}\right) - F^*\big)^{2-2\theta} \geq F\left(x_{k_j}\right) - F^*$. Then, (14) becomes

$$F\left(x_{k_j}\right) - F^* \leq K\left(\left(c_{k_j-1} - F^*\right) - \left(c_{k_j} - F^*\right)\right). \tag{17}$$

Using the definition of c_k, we obtain that

$$F\left(x_{k_j}\right) - F^* = q_{k_j}\left(c_{k_j} - F^*\right) - \eta q_{k_j-1}\left(c_{k_j-1} - F^*\right), \tag{18}$$

combine with (17), we can obtain that

$$c_{k_j} - F^* \leq \left(\frac{K + \eta q_{k_j-1}}{q_{k_j} + K}\right)\left(c_{k_j-1} - F^*\right). \tag{19}$$

Case 2. If $F\left(x_{k_j}\right) - F^* \leq 0$, from (18), we have

$$c_{k_j} - F^* \leq \frac{\eta q_{k_j-1}}{q_{k_j}}\left(c_{k_j-1} - F^*\right). \tag{20}$$

Hence, by (19), (20) and the fact that $\{c_k\}$ is nonincreasing, we have that

$$c_{k_j} - F^* \leq \left(\frac{K + \eta q_{k_j-1}}{q_{k_j} + K}\right)\left(c_{k_j-1} - F^*\right) \leq \left(\frac{K + \eta q_{k_j-1}}{q_{k_j} + K}\right)\left(c_{k_{j-1}} - F^*\right)$$

holds for any $k_j \in \Omega_1$. Then, we obtain by recursion that

$$c_{k_j} - F^* \leq \left(\frac{K + \eta q_{k_j-1}}{q_{k_j} + K}\right)^{j-j_0} \left(c_{k_{j_0}} - F^*\right). \tag{21}$$

Hence, by (16), (21) becomes that

$$c_{k_j} - F^* \leq \left(1 - \frac{1}{K + 1/(1-\eta)}\right)^{j-j_0} \left(c_{k_{j_0}} - F^*\right) = \tilde{K}^{j-j_0} \left(c_{k_{j_0}} - F^*\right), \tag{22}$$

where $\tilde{K} = 1 - \frac{1}{K+1/(1-\eta)}$.

– The proof of (13)

Note that $2\theta - 2 \in (-2, -1), 2\theta - 1 \in (-1, 0)$ and $c_{k_{j_0}}^{2\theta-1} \leq c_{k_{j_0+1}}^{2\theta-1} \leq \cdots \leq c_{k_{j-1}}^{2\theta-1} \leq c_{k_j}^{2\theta-1}$. Define $\phi(t) = \frac{1}{1-2\theta} t^{2\theta-1}$, then, $\phi'(t) = -t^{2\theta-2}$. Also define that $r_{k_j} = c_{k_j} - F^*$.

Case 1. Consider $F\left(x_{k_j}\right) - F^* \geq \frac{1}{2}\left(c_{k_j} - F^*\right)$, then,

$$\left(F\left(x_{k_j}\right) - F^*\right)^{2\theta-2} \leq \frac{1}{2^{2\theta-2}} \left(r_{k_j}\right)^{2\theta-2}.$$

By (14), we have

$$\frac{1}{K} \leq \frac{1}{2^{2\theta-2}} \left(r_{k_j}\right)^{2\theta-2} \left(r_{k_j-1} - r_{k_j}\right) \leq \frac{1}{2^{2\theta-2}} \left(r_{k_j}\right)^{2\theta-2} \left(r_{k_{j-1}} - r_{k_j}\right), \quad \forall j > j_0. \tag{23}$$

(1) If $\left(r_{k_j}\right)^{2\theta-2} \leq 2\left(r_{k_{j-1}}\right)^{2\theta-2}$, then,

$$\begin{aligned} \phi\left(r_{k_j}\right) - \phi\left(r_{k_{j-1}}\right) &= \int_{r_{k_j}}^{r_{k_{j-1}}} t^{2\theta-2} dt \geq \left(r_{k_{j-1}} - r_{k_j}\right) \left(r_{k_{j-1}}\right)^{2\theta-2} \\ &\geq \frac{1}{2} \left(r_{k_{j-1}} - r_{k_j}\right) \left(r_{k_j}\right)^{2\theta-2} \\ (23) &\geq \frac{1}{2^{3-2\theta} K}. \end{aligned} \tag{24}$$

(2) If $\left(r_{k_j}\right)^{2\theta-2} > 2\left(r_{k_{j-1}}\right)^{2\theta-2}$, we have $\left(r_{k_j}\right)^{2\theta-1} > 2^{\frac{2\theta-1}{2\theta-2}} \left(r_{k_{j-1}}\right)^{2\theta-1}$. Then,

$$\begin{aligned} \phi\left(r_{k_j}\right) - \phi\left(r_{k_{j-1}}\right) &= \frac{1}{1-2\theta} \left(\left(r_{k_j}\right)^{2\theta-1} - \left(r_{k_{j-1}}\right)^{2\theta-1}\right) \\ &> \frac{1}{1-2\theta} \left(2^{\frac{2\theta-1}{2\theta-2}} - 1\right) \left(r_{k_{j-1}}\right)^{2\theta-1} \\ &\geq \bar{K} \left(r_{k_{j_0}}\right)^{2\theta-1}, \end{aligned} \tag{25}$$

where $\bar{K} = \frac{1}{1-2\theta} \left(2^{\frac{2\theta-1}{2\theta-2}} - 1\right)$.

Case 2. Consider $F\left(x_{k_j}\right) - F^* < \frac{1}{2}\left(c_{k_j} - F^*\right)$. Following from the definition of c_k, we have

$$c_{k_j} - F^* < \frac{\eta q_{k_j-1}}{q_{k_j}} \left(c_{k_j-1} - F^*\right) + \frac{1}{2q_{k_j}} \left(c_{k_j} - F^*\right) \leq \frac{\eta q_{k_j-1}}{q_{k_j}} \left(c_{k_{j-1}} - F^*\right) + \frac{1}{2q_{k_j}} \left(c_{k_j} - F^*\right).$$

By (16), we have

$$c_{k_j} - F^* < \frac{2\eta}{1+\eta} \left(c_{k_{j-1}} - F^*\right). \tag{26}$$

Then,

$$\left(r_{k_j}\right)^{2\theta-1} > \left(\frac{1+\eta}{2\eta}\right)^{1-2\theta} \left(r_{k_{j-1}}\right)^{2\theta-1}$$

and

$$\begin{aligned}\phi\left(r_{k_j}\right)-\phi\left(r_{k_{j-1}}\right) &= \frac{1}{1-2\theta}\left(\left(r_{k_j}\right)^{2\theta-1}-\left(r_{k_{j-1}}\right)^{2\theta-1}\right) \\ &> \frac{1}{1-2\theta}\left(\left(\frac{1+\eta}{2\eta}\right)^{1-2\theta}-1\right)\left(r_{k_{j-1}}\right)^{2\theta-1} \\ &\geq \hat{K}\left(r_{k_{j_0}}\right)^{2\theta-1}\end{aligned} \tag{27}$$

where $\hat{K}=\frac{1}{1-2\theta}\left(\left(\frac{1+\eta}{2\eta}\right)^{1-2\theta}-1\right)$. Hence, from (24), (25) and (27), we obtain that

$$\phi\left(r_{k_j}\right)-\phi\left(r_{k_{j-1}}\right)\geq K_0,$$

where $K_0=\min\left\{\frac{1}{2^{2-2\theta}\bar{K}},\bar{K}\left(r_{k_{j_0}}\right)^{2\theta-1},\hat{K}\left(r_{k_{j_0}}\right)^{2\theta-1}\right\}$. Then,

$$\phi\left(r_{k_j}\right)\geq\sum\nolimits_{i=j_0}^{j-1}\phi\left(r_{k_{i+1}}\right)-\phi\left(r_{k_i}\right)\geq\left(j-j_0\right)K_0, \tag{28}$$

i.e.,

$$c_{k_j}-F^*=r_{k_j}\leq\left(\frac{2}{K_0\left(1-2\theta\right)\left(j-j_0\right)}\right)^{\frac{1}{1-2\theta}}. \tag{29}$$

Step 2. Now, we consider the convergence rate of function values.

Case 1. If Ω_2 is finite, then, there exists $\tilde{k}$ such that for any $k>\tilde{k}$, (3) in Algorithm 1 is executed. Then, following from (12), there exists $k_0>\tilde{k}$ such that

$$\begin{aligned}\left|F\left(x_k\right)-F^*\right| &\leq q_k\left(c_k-F^*\right)+\eta q_{k-1}\left(c_{k-1}-F^*\right) \\ &\leq \frac{2}{1-\eta}\left(c_{k-1}-F^*\right)\leq\frac{2}{1-\eta}\left(c_{k_0}-F^*\right)\tilde{K}^{k-k_0},\quad\text{for all }\theta\in\left[\frac{1}{2},1\right).\end{aligned} \tag{30}$$

Similarly, following from (13), we have

$$\left|F\left(x_k\right)-F^*\right|\leq\frac{2}{1-\eta}\left(\frac{2}{K_0\left(1-2\theta\right)\left(k-k_0\right)}\right)^{\frac{1}{1-2\theta}},\quad\text{for all }\theta\in\left(0,\frac{1}{2}\right).$$

Case 2. If Ω_1 is finite, then, similarly results can be obtained for $\left|F\left(x_k\right)-F^*\right|$.

Case 3. If Ω_1 and Ω_2 are both infinite, define that $\bar{\Omega}_1:=\Omega_1\cap\{1,2,\cdots,k\}$ and $\bar{\Omega}_2:=\Omega_2\cap\{1,2,\cdots,k\}$. Without losing generality, we suppose that $\left|\bar{\Omega}_2\right|\geq\left|\bar{\Omega}_1\right|$, which means that $\left|\bar{\Omega}_2\right|\geq\frac{k}{2}$. Then, following from (12) and (13), we have

$$\left|F\left(x_k\right)-F^*\right|\leq\frac{2}{1-\eta}\left(c_{k_0}-F^*\right)\tilde{K}^{\frac{k}{2}-k_0}$$

and

$$\left|F\left(x_k\right)-F^*\right|\leq\frac{2}{1-\eta}\left(\frac{2}{K_0\left(1-2\theta\right)\left(2/k-k_0\right)}\right)^{\frac{1}{1-2\theta}},\quad\text{for all }\theta\in\left(0,\frac{1}{2}\right).$$

Note that the convergence result for the whole sequence generated by Algorithm 1 is still unknown. According to this, it is presented in the following a variant of Algorithm 1, that is Hybrid_PG, which can be proved that all the accumulation point of the generated sequence belongs to citical point of F, further, the convergence rates of function values and iterates can also be obtained.

4 A Hybrid proximal gradient method

In the following, based on the idea of Algorithm 1, we propose a new nonmonotone proximal gradient algorithm, which has a more simpler frame but a faster convergence results from the numerical point of view. Also, the theoretical analysis is more complete.

Algorithm 2 Hybrid proximal gradient algorithm (Hybrid_PG)

Step 0. Take $x_0 = x^- \in R^n, \lambda_0 > 0, c_0 = F(x_0), q_0 = 1$. $\{\alpha_k\} \subset [0,1)$ be the inertial term.

Compute

$$y_k = x_{k-1} + \alpha_{k-1}(x_{k-1} - x_{k-2}) \tag{31}$$

$$z_k = T_\lambda(y_k) \tag{32}$$

If $F(z_k) \leq c_{k-1} - \delta\|z_k - x_{k-1}\|^2$

$$x_k = z_k \tag{33}$$

else

$$x_k = T_\lambda(y_k), \text{ where } y_k = x_{k-1} \tag{34}$$

end

Compute

$$c_k = \frac{\eta q_{k-1} c_{k-1} + F(x_k)}{q_k}, \text{ where } q_k = 1 + \eta q_{k-1}. \tag{35}$$

Remark. There are two differences between Algorithm 1 and Algorithm 2. (1) In Algorithm 1, the relaxed sufficient descent condition reads as $F(z_k) \leq c_{k-1} - \delta\|z_k - y_k\|^2$ and Algirithm 2 using the term that $\|z_k - x_{k-1}\|^2$. (2) If the relaxed sufficient descent condition not holds, Algorithm 1 choose the smaller function values between z_k and $T_\lambda(x_{k-1})$ as the iteration point, however, Algorithm 2 using the point $T_\lambda(x_{k-1})$ directly.

Here, we introduce two options of the stepsize strategies: *The Constant stepsize.* Let $\lambda < \frac{1}{L_f}$. Then, following from Appendix A that :

$$F(x_k) + \left(\frac{1}{2\lambda} - \frac{L_f}{2}\right)\|x_k - y_k\|^2 \leq F(z) + \left(\frac{1}{2\lambda} + \frac{L_f}{2}\right)\|z - y_k\|^2. \tag{36}$$

Note that the constant stepsize strategy requires to know the Lipschitz constant L_f, which greatly limits the application of Algorithm 2. Hence, we also show a *variable stepsize strategy*, which not require to know L_f when the algorithm be executed: The similar strategy is also used in [32] and [33], hence, the following lemma show the corresponding property of the generated stepsize but omit the detailed proof.

Lemma 4.1 *Let $\{\lambda_k\}$ be the sequence generated by Algorithm 3. We have that the sequence $\{\lambda_k\}$ is convergent. And for all k,*

$$\lambda_k \geq \underline{\lambda} = \min\left\{\lambda_1, \frac{\mu_1}{L_f}\right\}. \tag{40}$$

Algorithm 3 nonmonotone variable stepsize strategy (nm_VS)

If the condition

$$2\left|f\left(x_{k-1}\right)-f\left(y_{k-1}\right)-\left\langle\nabla f\left(x_{k-1}\right),x_{k-1}-y_{k-1}\right\rangle\right|>\frac{\mu_0}{\lambda_{k-1}}\left\|x_{k-1}-y_{k-1}\right\|^2 \tag{37}$$

holds, set

$$\lambda_k=\frac{\mu_1\cdot\left\|x_{k-1}-y_{k-1}\right\|^2}{2\left|f\left(x_{k-1}\right)-f\left(y_{k-1}\right)-\left\langle\nabla f\left(x_{k-1}\right),x_{k-1}-y_{k-1}\right\rangle\right|} \tag{38}$$

otherwise, set

$$\lambda_k=\lambda_{k-1}+\min\left\{1,\lambda_{k-1}\right\}\mathrm{E}\left(k\right), \tag{39}$$

where $\sum_{k=1}^{\infty}E\left(k\right)$ is a positive convergence series.

end

Also, there exists a $\hat{k}\geq 1$, *for every* $k>\hat{k}$, *condition (37) holds constantly. Denote* $\lim\limits_{k\to\infty}\lambda_k=\lambda^*$. *Then, there exists* $\bar{\lambda}=\max\left(\lambda^*,\lambda_0,\cdots,\lambda_{\hat{k}}\right)$ *such that* $\lambda_k\leq\bar{\lambda}$ *for all* k. *In addition, for any* $k>\hat{k}$,

$$F\left(x_k\right)+\left(\frac{1-\mu_0}{2\lambda_k}\right)\left\|x_k-y_k\right\|^2\leq F\left(z\right)+\left(\frac{1}{2\lambda_k}+\frac{L_f}{2}\right)\left\|z-y_k\right\|^2,\ \ \forall z\in R^n. \tag{41}$$

Hence, denote $\lambda_{\min}=\lambda_{\max}=\lambda$ if using the constant stepsize strategy; $\lambda_{\min}=\underline{\lambda}$ and $\lambda_{\max}=\bar{\lambda}$ if using the variable stepsize strategy. Combine (36) and (41), we can obtain than there exists k_1 such that for any $k>\hat{k}$,

$$M_0\|x_k-y_k\|^2\leq F\left(z\right)-F\left(x_k\right)+\left(\frac{1}{2\lambda_{\min}}+\frac{L_f}{2}\right)\left\|z-y_k\right\|^2,\ \ \forall z\in R^n, \tag{42}$$

where $M_0=\min\left\{\frac{1-\mu_0}{2\lambda_{\max}},\frac{1}{2}\left(\frac{1}{\lambda_{\max}}+L_f\right)\right\}$.

Similar with the Lemma 3.1, we can also obtain the following lemma for Algorithm 2. See Appendix B for the concrete proof.

Lemma 4.2 *Let* $\{x_k\}$ *be generated by Algorithm 2. Then, we have* $F\left(x_k\right)\leq c_k$ *and* $c_{k+1}\leq c_k$.

Now, we analyze the convergence of Algorithm 2 using the following Lemma.

Lemma 4.3 *[15] Let* $\left(x_k,u_k\right)$ *be a sequence such that* $x_k\to x$, $u_k\to u$, $F\left(x_k\right)\to F\left(x\right)$ *and* $u_k\in\partial F\left(x_k\right)$, *then* $u\in\partial F\left(x\right)$.

Theorem 4.1 *Let* $\{x_k\}$ *generated by Algorithm 2. Then, all the accumulation point of the* $\{x_k\}$ *belongs to* crit F.

Proof. We can easy to show that $\{x_k\}$ is bounded by the fact that $\{F\left(x_k\right)\}$ is level bounded. Suppose that $\left\{x_{k_j}\right\}$ is a convergent subsequence of $\{x_k\}$ and $\lim\limits_{j\to\infty}x_{k_j}=\hat{x}$.

Following from $x_k=T_\lambda\left(y_k\right)$, we have

$$\nabla f\left(x_k\right)-\nabla f\left(y_k\right)-\frac{1}{\lambda_k}\left(x_k-y_k\right)\in\partial F\left(x_k\right). \tag{43}$$

Since the fact that ∇f is Lipschitz continuous gradient, we obtain that

$$\left\|\nabla f\left(x_k\right)-\nabla f\left(y_k\right)-\frac{1}{\lambda_k}\left(x_k-y_k\right)\right\|\leq\left(L_f+\frac{1}{\lambda_k}\right)\left\|x_k-y_k\right\| \tag{44}$$

(i) If x_k generated by (33), then, $y_k=x_{k-1}+\alpha_k\left(x_{k-1}-x_{k-2}\right)$ and

$$\left\|x_k-x_{k-1}\right\|^2\leq\frac{1}{\delta}\left(c_{k-1}-F\left(x_k\right)\right), \tag{45}$$

$$\|x_k - y_k\| \leq \|x_k - x_{k-1}\| + \|x_{k-1} - x_{k-2}\|. \tag{46}$$

(ii) If x_k generated by (34), then, $y_k = x_{k-1}$ and (42) becomes that

$$\|x_k - x_{k-1}\|^2 \leq \frac{1}{M_0}\left(F\left(x_{k-1}\right) - F\left(x_k\right)\right), \tag{47}$$

$$\|x_k - y_k\| = \|x_k - x_{k-1}\|. \tag{48}$$

Combine (45) and (47), we get that for any k,

$$\begin{aligned}\|x_k - x_{k-1}\|^2 &\leq \max\left(\frac{1}{\delta}, \frac{1}{M_0}\right)\left(c_{k-1} - F\left(x_k\right)\right) \\ (35) &= \max\left(\frac{1}{\delta}, \frac{1}{M_0}\right) q_{k-1}\left(c_{k-1} - c_k\right) \\ &\leq \max\left(\frac{1}{\delta}, \frac{1}{M_0}\right)\frac{1}{1-\eta}\left(c_{k-1} - c_k\right).\end{aligned} \tag{49}$$

where the last inequality follows from (16). Summing (49) over $k = 1, 2, \cdots, \infty$, we can obtain that

$$\sum\nolimits_{k=1}^{\infty} \|x_k - x_{k-1}\|^2 < \infty$$

and $\|x_k - x_{k-1}\| \to 0$. Hence, combine with (46), (48) and (44), we have

$$\left\|\nabla f\left(x_k\right) - \nabla f\left(y_k\right) - \frac{1}{\lambda_k}\left(x_k - y_k\right)\right\| \to 0. \tag{50}$$

In addition, we have

$$\left\langle \nabla f\left(y_k\right), x_k - y_k\right\rangle + \frac{1}{2\lambda_k}\|x_k - y_k\|^2 + g\left(x_k\right) \leq \left\langle \nabla f\left(y_k\right), \bar{x} - y_k\right\rangle + \frac{1}{2\lambda_k}\|\bar{x} - y_k\|^2 + g\left(\bar{x}\right), \tag{51}$$

which means that $\limsup_{k\to\infty} g\left(x_k\right) \leq g\left(\bar{x}\right)$. Combining with $\liminf_{k\to\infty} g\left(x_k\right) \geq g\left(\bar{x}\right)$ from the definition of lower semicontinuous of g, we have $\lim_{k\to\infty} g\left(x_k\right) = g\left(\bar{x}\right)$. Moreover, since f is continuously differentiable, we have $\lim_{k\to\infty} f\left(x_k\right) = f\left(\bar{x}\right)$. Hence,

$$\lim_{k\to\infty} F\left(x_k\right) = F\left(\bar{x}\right). \tag{52}$$

Combining $\lim_{j\to\infty} x_{k_j} = \hat{x}$, (43), (50), (52), and using Lemma 4.3, we have $0 \in \partial F\left(\hat{x}\right)$.

Lemma 4.4 *Let $\{x_k\}$ and $\{c_k\}$ generated by Algorithm 2. Then, we have that $\{F\left(x_k\right)\}$ is convergence. Further, taking $\lim_{k\to\infty} F\left(x_k\right) = F^*$, we have $c_k \to F^*$ and $c_k \geq F^*$.*

Theorem 4.2 *Denote X^* is the set of all accumulation points of $\{x_k\}$ generated by Algorithm 2. For $\xi = \lim_{k\to\infty} F\left(x_k\right)$, we have $F\left(X^*\right) \equiv \xi$.*

Proof. For any $\bar{x} \in X^*$, there exists a $\left\{x_{k_j}\right\}$ such that $\lim_{j\to\infty} x_{k_j} = \bar{x}$. It follows that

$$F\left(\bar{x}\right) \leq \liminf_{j\to\infty} F\left(x_{k_j}\right) = \lim_{k\to\infty} F\left(x_k\right) = \xi \tag{53}$$

from the fact that F is lower semicontinuous. In addition, recalling (42) and set $k := k_j$ and $x = \bar{x}$, we have

$$F\left(x_{k_j}\right) + M_0\left\|x_{k_j} - y_{k_j}\right\|^2 \leq F\left(\bar{x}\right) + \left(\frac{1}{2\lambda_{\min}} + \frac{L_f}{2}\right)\left\|\bar{x} - y_{k_j}\right\|^2. \tag{54}$$

Following from $\lim_{j\to\infty}\left\|x_{k_j}-y_{k_j}\right\|^2=0$ and $\lim_{j\to\infty}\left\|\bar{x}-y_{k_j}\right\|^2=0$, we have

$$\xi=\lim_{j\to\infty}F\left(x_k\right)=\limsup_{j\to\infty}F\left(x_{k_j}\right)\leq F\left(\bar{x}\right). \tag{55}$$

Combining (53) and (55), we have

$$\xi=\lim_{k\to\infty}F\left(x_k\right)=F\left(\bar{x}\right).$$

Hence, the conclusion follows from the arbitrariness of $\bar{x}$.

The following two theorems show the convergence rates of the objective function values and iterates generated by Algorithm 2. We suppose that $c_k>F^*$ for all k. Otherwise, by (56) and (43), this algorithm terminates in finite steps.

Theorem 4.3 *Let $\{x_k\}$ be generated by Algorithm 2. Assume that F satisfy the KŁ property at each point of* crit F*, and the desingularising function has the form of $\varphi(t)=\frac{C}{\theta}t^\theta$ for some $C>0$, $\theta\in[0,1)$. Then,*
(1) If $\theta\in\left[\frac{1}{2},1\right)$, there exists $\mathcal{D}\in(0,1)$ such that

$$\left|F\left(x_k\right)-F^*\right|=O\left(\mathcal{D}^{\frac{k}{2}}\right).$$

(2) If $\theta\in\left(0,\frac{1}{2}\right)$,

$$\left|F\left(x_k\right)-F^*\right|=O\left(k^{-\frac{1}{1-2\theta}}\right).$$

Proof **Step 1.** We first prove that for any $k>\hat{k}$, there exists $M>0$ such that

$$\left\|x_k-y_k\right\|^2\leq M\left(c_{k-2}-c_k\right). \tag{56}$$

(i) Consider that x_k generated by (33), i.e.,

$$y_k=x_{k-1}+\alpha_{k-1}\left(x_{k-1}-x_{k-2}\right) \tag{57}$$

and

$$\left\|x_k-x_{k-1}\right\|^2\leq\frac{1}{\delta}\left(c_{k-1}-F\left(x_k\right)\right) \tag{58}$$

Then,

$$\left\|x_k-y_k\right\|^2\leq 2\left(\left\|x_k-x_{k-1}\right\|^2+\left\|x_{k-1}-x_{k-2}\right\|^2\right)\leq 2\left(\frac{1}{\delta}(c_{k-1}-F\left(x_k\right))+\left\|x_{k-1}-x_{k-2}\right\|^2\right). \tag{59}$$

If x_{k-1} generated by (33), then,

$$\left\|x_{k-1}-x_{k-2}\right\|^2\leq\frac{1}{\delta}\left(c_{k-2}-F\left(x_{k-1}\right)\right). \tag{60}$$

If x_{k-1} generated by (34), using (42) with $k:=k-1$ and $z:=x_{k-2}$, we have

$$\left\|x_{k-1}-x_{k-2}\right\|^2\leq\frac{1}{M_0}\left(F\left(x_{k-2}\right)-F\left(x_{k-1}\right)\right)\leq\frac{1}{M_0}\left(c_{k-2}-F\left(x_{k-1}\right)\right),\ \ \forall k>\hat{k}. \tag{61}$$

From (60) and (61), (59) becomes that

$$\begin{aligned}\left\|x_k-y_k\right\|^2&\leq 2\max\left(\frac{1}{M_0},\frac{1}{\delta}\right)\left(c_{k-1}-F\left(x_k\right)+c_{k-2}-F\left(x_{k-1}\right)\right)\\&\leq\max\left(\frac{1}{M_0},\frac{1}{\delta}\right)\left(\frac{2}{1-\eta}\right)\left(c_{k-2}-c_k\right),\ \ \forall k>\hat{k},\end{aligned} \tag{62}$$

where the last inequality follows from (16) and

$$c_{k-1} - F(x_k) = q_k (c_{k-1} - c_k). \tag{63}$$

(ii) Consider that x_k generated by (33), similar with (61) we have

$$\|x_k - y_k\|^2 = \|x_k - x_{k-1}\|^2 \leq \frac{1}{M_0} (c_{k-1} - F(x_k)) \leq \frac{1}{M_0} (c_{k-2} - F(x_k)). \tag{64}$$

Hence, set $M = \max\left(\frac{1}{M_0}, \frac{1}{\delta}\right)\left(\frac{2}{1-\eta}\right)$, we get (56) from (62) and (64).

Step 2. We show that $|F(x_k) - F^*|$ converges linearly to 0 for $\theta \in \left[\frac{1}{2}, 1\right)$.

Since $dist(x_k, X^*) \to 0$, $F(x_k) \to F^*$, $X^* \subset \operatorname{crit} F$ and X^* is compact set, using Lemma 2.2 with $\Omega := X^*$, there exists $k_0 > \hat{k}$ ($\hat{k}$ be defined in Lemma 4.1) such that for any $k > k_0$, if $F(x_k) - F^* > 0$, we have

$$x_k \in \left\{ v | \, dist\left(v, X^*\right) < \varepsilon \cap F^* < F(v) < F^* + \eta \right\}$$

and

$$1 \leq \left(\varphi'\left(F(x_k) - F^*\right)\right)^2 dist(0, \partial F(x_k))^2. \tag{65}$$

Case 1. If $F(x_k) - F^* > 0$, by (65), (43) and (44), we have

$$1 \leq \left(\frac{1}{\lambda_{\min}} + L_f\right)^2 \left(\varphi'\left(F(x_k) - F^*\right)\right)^2 \|x_k - y_k\|^2 \tag{66}$$

$$(62) \leq D\left(F(x_k) - F^*\right)^{2\theta-2} (c_{k-2} - c_k),$$

where $D = MC^2\left(\frac{1}{\lambda_{\min}} + L_f\right)^2$.

Consider that $\theta \in \left[\frac{1}{2}, 1\right)$. Since $0 < 2 - 2\theta \leq 1$, we have $\left(F(x_k) - F^*\right)^{2-2\theta} \geq F(x_k) - F^*$. Then, (65) becomes

$$F(x_k) - F^* \leq D\left(\left(c_{k-2} - F^*\right) - \left(c_k - F^*\right)\right). \tag{67}$$

By (35) and the fact that $\{c_k\}$ is nonincreasing, we obtain that

$$F(x_k) - F^* = q_k\left(c_k - F^*\right) - \eta q_{k-1}\left(c_{k-1} - F^*\right) \geq q_k\left(c_k - F^*\right) - \eta q_{k-1}\left(c_{k-2} - F^*\right), \tag{68}$$

then, (67) becomes that

$$c_k - F^* \leq \left(\frac{\eta q_{k-1} + D}{q_k + D}\right)\left(c_{k-2} - F^*\right). \tag{69}$$

Case 2. If $F(x_k) - F^* \leq 0$, follows from (68) that

$$q_k\left(c_k - F^*\right) \leq \eta q_{k-1}\left(c_{k-2} - F^*\right),$$

we obtain

$$c_k - F^* \leq \left(\frac{\eta q_{k-1}}{q_k}\right)\left(c_{k-2} - F^*\right). \tag{70}$$

Hence, combine (69) and (70), we have

$$c_k - F^* \leq \left(\frac{\eta q_{k-1} + D}{q_k + D}\right)\left(c_{k-2} - F^*\right) = \left(1 - \frac{1}{q_k + D}\right)\left(c_{k-2} - F^*\right) \leq \tilde{D}\left(c_{k-2} - F^*\right), \tag{71}$$

where $\tilde{D}=\left(1-\frac{1}{1/(1-\eta)+D}\right)\in(0,1)$ since (16). Hence, (71) can be deduced by recursion that for any $k>k_0$,

$$c_k-F^*\leq\tilde{D}^{\frac{k-k_0}{2}}\left(c_{k_0}-F^*\right),\ \text{ for }\ \theta\in\left[\frac{1}{2},1\right). \tag{72}$$

Further, by (35), we obtain

$$\left|F\left(x_k\right)-F^*\right|\leq q_k\left|c_k-F^*\right|+\eta q_{k-1}\left|c_{k-1}-F^*\right|\leq\frac{2}{1-\eta}\left|c_{k-1}-F^*\right|. \tag{73}$$

Hence, the result (ii) can be completed follows from (72).

Step 3. We now show that $\left|F\left(x_k\right)-F^*\right|$ sublinearly converges to 0 at the $O\left(k^{-\frac{1}{1-2\theta}}\right)$ rate of convergence for $\theta\in\left(0,\frac{1}{2}\right)$.

Since that $\theta\in\left(0,\frac{1}{2}\right)$, then, $2\theta-2\in(-2,-1),2\theta-1\in(-1,0)$. And $c_{k_0}^{2\theta-1}\leq c_{k_0+2}^{2\theta-1}\leq\cdots\leq c_{k-2}^{2\theta-1}\leq c_k^{2\theta-1}$. Define $\phi\left(t\right)=\frac{1}{1-2\theta}t^{2\theta-1}$, then, $\phi'\left(t\right)=-t^{2\theta-2}$. Also define that $r_k=c_k-F^*$.

Case 1. Consider $F\left(x_k\right)-F^*\geq\frac{1}{2}\left(c_k-F^*\right)$. Then, $\left(F\left(x_k\right)-F^*\right)^{2\theta-2}\leq\frac{1}{2^{2\theta-2}}\left(c_k-F^*\right)^{2\theta-2}$, and, (65) becomes

$$\frac{1}{D}\leq\left(F\left(x_k\right)-F^*\right)^{2\theta-2}\left(c_{k-2}-c_k\right)\leq\frac{1}{2^{2\theta-2}}\left(r_k\right)^{2\theta-2}\left(r_{k-2}-r_k\right). \tag{74}$$

If $\left(r_k\right)^{2\theta-2}\leq2\left(r_{k-2}\right)^{2\theta-2}$, then,

$$\phi\left(r_k\right)-\phi\left(r_{k-2}\right)=\int_{r_k}^{r_{k-2}}t^{2\theta-2}dt\geq\left(r_{k-2}-r_k\right)\left(r_{k-2}\right)^{2\theta-2}\geq\frac{1}{2}\left(r_{k-2}-r_k\right)\left(r_k\right)^{2\theta-2}\geq\frac{1}{2^{3-2\theta}D}. \tag{75}$$

If $\left(r_k\right)^{2\theta-2}>2\left(r_{k-2}\right)^{2\theta-2}$, we have $\left(r_k\right)^{2\theta-1}>2^{\frac{2\theta-1}{2\theta-2}}\left(r_{k-2}\right)^{2\theta-1}$. Then,

$$\phi\left(r_k\right)-\phi\left(r_{k-2}\right)=\frac{1}{1-2\theta}\left(\left(r_k\right)^{2\theta-1}-\left(r_{k-2}\right)^{2\theta-1}\right)\geq\frac{1}{1-2\theta}\left(2^{\frac{2\theta-1}{2\theta-2}}-1\right)\left(r_{k-2}\right)^{2\theta-1}\geq\hat{D}(r_{k_0})^{2\theta-1}, \tag{76}$$

where $\hat{D}=\frac{1}{1-2\theta}\left(2^{\frac{2\theta-1}{2\theta-2}}-1\right)$.

Case 2. Consider $F\left(x_k\right)-F^*<\frac{1}{2}\left(c_k-F^*\right)$. From (35), we have

$$c_k-F^*=\left(\frac{\eta q_{k-1}}{q_k}\right)\left(c_{k-1}-F^*\right)+\left(\frac{1}{q_k}\right)\left(F\left(x_k\right)-F^*\right)<\left(\frac{\eta q_{k-1}}{q_k}\right)\left(c_{k-1}-F^*\right)+\left(\frac{1}{2q_k}\right)\left(c_k-F^*\right),$$

i.e.,

$$c_k-F^*<\left(1-\frac{1}{2q_k-1}\right)\left(c_{k-1}-F^*\right)\leq\frac{2\eta}{1+\eta}\left(c_{k-1}-F^*\right).$$

Then,

$$\left(r_k\right)^{2\theta-1}>\left(\frac{1+\eta}{2\eta}\right)^{1-2\theta}\left(r_{k-1}\right)^{2\theta-1}\geq\left(\frac{1+\eta}{2\eta}\right)^{1-2\theta}\left(r_{k-2}\right)^{2\theta-1}$$

and

$$\begin{aligned}\phi\left(r_k\right)-\phi\left(r_{k-2}\right)&=\frac{1}{1-2\theta}\left(\left(r_k\right)^{2\theta-1}-\left(r_{k-2}\right)^{2\theta-1}\right)\\&>\frac{1}{1-2\theta}\left(\left(\frac{1+\eta}{2\eta}\right)^{1-2\theta}-1\right)\left(r_{k-2}\right)^{2\theta-1}\\&\geq\bar{D}(r_{k_0})^{2\theta-1},\end{aligned} \tag{77}$$

where $\bar{D}=\frac{1}{1-2\theta}\left(\left(\frac{1+\eta}{2\eta}\right)^{1-2\theta}-1\right)$. Hence, from (75), (76) and (77), we obtain that

$$\phi\left(r_k\right)-\phi\left(r_{k-2}\right)\geq D_0,$$

where $D_0 = \min\left(\frac{1}{2^{3-2\theta}D}, \hat{D}(r_{k_0})^{2\theta-1}, \bar{D}(r_{k_0})^{2\theta-1}\right)$. Then, for $k > k_0$,

$$\phi(r_k) \geq (\phi(r_k) - \phi(r_{k-2})) + (\phi(r_{k-2}) - \phi(r_{k-4})) + \cdots + (\phi(r_{k_0+2}) - \phi(r_{k_0})) \geq \left(\frac{k-k_0}{2}\right) D_0, \tag{78}$$

i.e.,

$$(r_k)^{2\theta-1} \geq (1-2\theta)\left(\frac{k-k_0}{2}\right) D_0.$$

Hence, for any $k > k_0$,

$$c_k - F^* = r_k \leq \left(\frac{2}{D_0(1-2\theta)(k-k_0)}\right)^{\frac{1}{1-2\theta}}, \quad \text{for } \theta \in \left(0, \frac{1}{2}\right). \tag{79}$$

Similar with the proof of result (ii), the proof of result (iii) can be completed follows from (73) and (79).

Theorem 4.4 *Let $\{x_k\}$ be generated by Algorithm 2. Assume that F satisfy the KŁ property at each point of* crit F*, and the desingularising function has the form of $\varphi(t) = \frac{C}{\theta}t^\theta$ for some $C > 0$, $\theta \in \left(\frac{1}{4}, 1\right)$.*

(1) If $\theta \in \left[\frac{1}{2}, 1\right)$, then, $\{x_k\}$ R-linearly converges to its limit point.

(2) If $\theta \in \left(\frac{1}{4}, \frac{1}{2}\right)$, we have $\{x_k\}$ sublinearly converges to its limit point with $O\left(k^{-\frac{4\theta-1}{2(1-2\theta)}}\right)$ convergence rate.

Proof For x_k generated by (33),

$$\begin{aligned} \|x_k - x_{k-1}\|^2 = \|z_k - x_{k-1}\|^2 &\leq \frac{1}{\delta}(c_{k-1} - F(x_k)) \\ &= \frac{q_k}{\delta}\left(\left(c_{k-1} - F^*\right) - \left(c_k - F^*\right)\right) \\ (16) &\leq \frac{1}{\delta(1-\eta)}\left(\left(c_{k-1} - F^*\right) - \left(c_k - F^*\right)\right). \end{aligned} \tag{80}$$

For x_k generated by (34), using (42) with $z := x_{k-1}$, we have

$$\begin{aligned} \|x_k - x_{k-1}\|^2 &\leq \frac{1}{M_0}(F(x_{k-1}) - F(x_k)) \leq \frac{1}{M_0}(c_{k-1} - F(x_k)) \\ &= \frac{q_k}{M_0}\left(\left(c_{k-1} - F^*\right) - \left(c_k - F^*\right)\right) \\ &\leq \frac{1}{M_0(1-\eta)}\left(\left(c_{k-1} - F^*\right) - \left(c_k - F^*\right)\right). \end{aligned} \tag{81}$$

Hence, for all k, we have

$$\|x_k - x_{k-1}\|^2 \leq \frac{1}{1-\eta}\max\left(\frac{1}{\delta}, \frac{1}{M_0}\right)\left(\left|c_{k-1} - F^*\right| + \left|c_k - F^*\right|\right) \leq \bar{M}\left(\left|c_{k-1} - F^*\right|\right),$$

where $\bar{M} = \frac{2}{1-\eta}\max\left(\frac{1}{\delta}, \frac{1}{M_0}\right)$. Following from (72) and (79), we can obtain that

$$\|x_k - x_{k-1}\| \leq \sqrt{\bar{M}r_{k_0}}\tilde{D}^{\frac{k-(k_0+1)}{2}}, \quad \text{for } \theta \in \left[\frac{1}{2}, 1\right), \tag{82}$$

and

$$\|x_k - x_{k-1}\| \leq \left(\frac{2\bar{M}}{D_0(1-2\theta)(k-(k_0+1))}\right)^{\frac{1}{2(1-2\theta)}}, \quad \text{for } \theta \in \left(0, \frac{1}{2}\right). \tag{83}$$

Hence, for any $p > 0$, by (82), we have for $\theta \in \left[\frac{1}{2}, 1\right)$,

$$\begin{aligned} \left\|x_{k+p} - x_k\right\| \leq \sum\nolimits_{i=k+1}^{k+p}\|x_i - x_{i-1}\| &\leq \sqrt{\bar{M}r_{k_0}}\int_k^{k+p}\tilde{D}^{\frac{x-(k_0+1)}{2}}dx \\ = -\sqrt{\bar{M}r_{k_0}}\frac{1}{\left|\ln\tilde{D}\right|}\left(\tilde{D}^{\frac{x-(k_0+1)}{2}}\right)\Big|_k^{k+p} &\leq \sqrt{\bar{M}r_{k_0}}\frac{1}{\left|\ln\tilde{D}\right|}\left(\tilde{D}^{\frac{k-(k_0+1)}{2}}\right)\Big|, \end{aligned} \tag{84}$$

i.e., $\{x_k\}$ is Cauchy sequence. Taking $\lim_{k\to\infty} x_k = \bar{x}$, then, as $p \to \infty$,

$$\|x_k - \bar{x}\| \leq \sqrt{\bar{M} r_{k_0}} \frac{1}{\left|\ln \tilde{D}\right|} \tilde{D}^{\frac{k-(k_0+1)}{2}}, \quad \text{for } \theta \in \left[\frac{1}{2}, 1\right). \tag{85}$$

Also, by (83), for any $p > 0$, we have

$$\begin{aligned} \left\|x_{k+p} - x_k\right\| &\leq \sum\nolimits_{i=k+1}^{k+p} \|x_i - x_{i-1}\| \\ &\leq \sqrt[2(1-2\theta)]{\frac{2\bar{M}}{D_0(1-2\theta)}} \int_k^{k+p} (x - (k_0+1))^{-\frac{1}{2(1-2\theta)}} dx \\ &= -\sqrt[2(1-2\theta)]{\frac{2\bar{M}}{D_0(1-2\theta)}} \cdot \frac{2(1-2\theta)}{4\theta - 1} (x - (k_0+1))^{\frac{1-4\theta}{2(1-2\theta)}} \Big|_k^{k+p} \\ &\leq -\sqrt[2(1-2\theta)]{\frac{2\bar{M}}{D_0(1-2\theta)}} \cdot \frac{2(1-2\theta)}{4\theta - 1} (k - (k_0+1))^{\frac{1-4\theta}{2(1-2\theta)}}, \end{aligned} \tag{86}$$

i.e., $\{x_k\}$ is Cauchy sequence. Taking $\lim_{k\to\infty} x_k = \bar{x}$, then, as $p \to \infty$,

$$\|x_k - \bar{x}\| \leq \sqrt[2(1-2\theta)]{\frac{2\bar{M}}{D_0(1-2\theta)}} \frac{2(1-2\theta)}{4\theta - 1} (k - (k_0+1))^{-\frac{4\theta-1}{2(1-2\theta)}}, \quad \text{for } \theta \in \left(\frac{1}{4}, \frac{1}{2}\right). \tag{87}$$

Hence, the proof be completed from (85) and (87).

5 Numercial results

In this section, we apply the hybrid proximal gradient method (Hybrid_PG) to solve the following two types of nonconvex models:

Case 1. Nonconvex penalty model

$$\min_{x \in R^n} \frac{1}{2}\|Ax - b\|^2 + \mu\psi(x), \tag{88}$$

This model, arised in signal processing [27], belongs to "convex + nonconvex" case. Here, taking $A \in R^{n\times m}$ with random Gaussian entries and $b := A\hat{x} + 0.01\epsilon$, where $\hat{x}$ is a random sparse vector in R^n with the density $\frac{s}{m}$ and ϵ is a noise vector generated randomly. The parameter μ is used for balancing those two objective terms. $\psi : R^n \to R$ represents the regularizer to characterize the sparsity of solution. Here, we test (88) with two regularizers ψ :

– $\psi(x) = \|x\|_{1/2}^{1/2}$

The problem of finding a sparse solution $x \in R^m$ to a least-squares problem $Ax = b$ can be modeled as (88) with L_0 norm $\|x\|_0$, namely the number of nonzero elements of x. Since solving this model generally is intractable, many researchers have suggested to relax the L_0 regularization and, instead, to consider the following $L_{1/2}$ regularization [46]:

$$\min_{x \in R^m} \frac{1}{2}\|Ax - b\|^2 + \|x\|_{1/2}^{1/2}, \tag{89}$$

where $\|x\|_{1/2}^{1/2} := \sum_{i=1}^n |x_i|^{1/2}$ is separable, and its proximal mapping can be founded in [46].

– $\psi(x) = \sum_{i=1}^n g_\kappa(|x_i|)$

The smoothly clipped absolute deviation (SCAD) penalty problem [25], which arising in statistical learning, can be read as:

$$\min_{x\in R^m} \frac{1}{2}\|Ax-b\|^2+\mu\sum\nolimits_{i=1}^{n} g_\kappa\left(|x_i|\right), \tag{90}$$

where

$$g_\kappa\left(|x_i|\right):=\begin{cases}\kappa\,|x_i|\,, & |x_i|\le\kappa\\ \frac{-|x_i|^2+2c\kappa|x_i|-\kappa^2}{2(c-1)}, & \kappa<|x_i|\le c\kappa\\ \frac{(c+1)\kappa^2}{2}, & |x_i|>c\kappa,\end{cases}\qquad (c>2,\kappa>0)$$

As analyzed in [25,48], empirically the parameter c and κ could be better chosen through the cross-validation or generalized cross-validation technique by some given samples. Here, for the sole purpose of showing the numerical efficiency, we simply fix them as $c=3.7$ and $\kappa=0.1\sqrt{2\log(m)}$.

Case 2. Nonconvex constraint model

$$\min_{x\in\Delta_r^u} \frac{1}{2}x^TAx-b^Tx, \tag{91}$$

where $\Delta_r^u:=\left\{x\in R^m:\sum_{i=1}^m x_i=l,\ \|x\|_0\le r,\ 0\le x_i\le u,\ i=1,\cdots,m\right\}$. Notice that one can rewrite (91) in the form of problem (P) by defining $f(x)=\frac{1}{2}x^TAx-b^Tx$ and $g(x)=\delta_S(x)$, where $S=\Delta_r^u$. It is clear that f has a Lipschitz continuous gradient and g is nonconvex. The projection on S we refer the reader to [47]. For each $m=500,1000,2000$, we generate matrix $A:=B^T+B$ to make f is nonconvex, where $B\in R^{m\times m}$ be generated with i.i.d. standard Gaussian entries. Taking $b=\mathrm{randn}(m,1)$, $l=\max\{1,10t\}$ where t is chosen uniformly at random from $[0,1]$, $r=\lfloor\frac{m}{100}\rfloor$ and $u=\max\{10,l\}$.

Comparison algorithms We consider following different algorithms for each class of problems: (1) FISTA with fixed stepsize proposed in [12]. (2) nmAPG with fixed stepsize proposed in [35] (3) Hybrid_PG with fixed stepsize. Setting the fixed stepsize $\lambda_k\equiv\frac{0.98}{L_f}$. (4) Hybrid_PG with variable stepsize strategy nm_VS with $\mu_0=0.99$, $\mu_1=0.95$ and $E_k=\frac{1}{k^{1.1}}$. Note that FISTA is not necessarily convergent for nonconvex optimization theoretically. We initialize these four algorithm at the initial point $x_0=[0,0,\cdots,0]^T$ for Case 1 and $x_0=[l,0,\cdots,0]^T$ for Case 2. And set the parameters $\delta:=10^{-6}$ and $\eta:=0.8$. Besides, we use the stopping criteria

$$\|\psi_k\|\le TOL$$

where $\partial F(x_k)\ni\psi_k=\nabla f(x_k)-\nabla f(y_k)-\frac{1}{\lambda_k}(x_k-y_k)$ and $TOL:=10^{-5}$.

Table 1: Numerical comparisons of tested algorithms for solving the nonconvex penalty model

	n=100,m=1000,s=20				n=300,m=3000,s=30				n=500,m=5000,s=50			
	$L_{\frac{1}{2}}$		SCAD		$L_{\frac{1}{2}}$		SCAD		$L_{\frac{1}{2}}$		SCAD	
	Iter	CPUs	Iter	CPUs	Iter	CPUs	Iter	CPUs	Iter	CPUs	Iter	CPUs
FISTA	2944	0.4458	3572	0.6496	1432	1.5245	5775	5.9708	1417	10.0800	7912	59.2045
nmAPG	2246	0.4016	2615	0.5316	1234	1.2871	4664	5.0479	1286	9.2879	6787	53.4717
Hybrid_PG	743	0.1168	1517	0.2870	714	0.7799	1320	1.3947	780	5.7880	1695	12.6378
Hybrid_PG_vs	276	0.0566	990	0.2151	409	0.5099	936	1.1416	386	3.4644	1183	10.3891

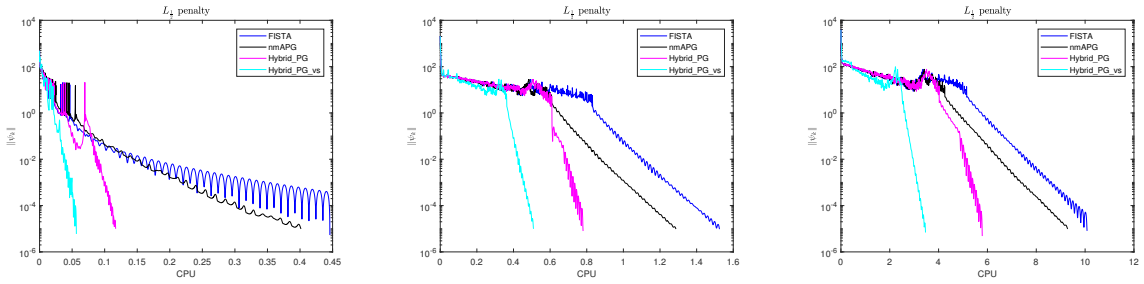

Fig. 1: Performance profile for the convergence of $\|\psi_k\|$ for solving the nonconvex penalty model with $L_{\frac{1}{2}}$ penalty. Left: Example with $n = 100, m = 1000, s = 20$ with $\mu = 1$. Middle: Example with $n = 300, m = 3000, s = 30$ with $\mu = 0.1$. Right: Example with $n = 500, m = 5000, s = 50$ with $\mu = 0.25$.

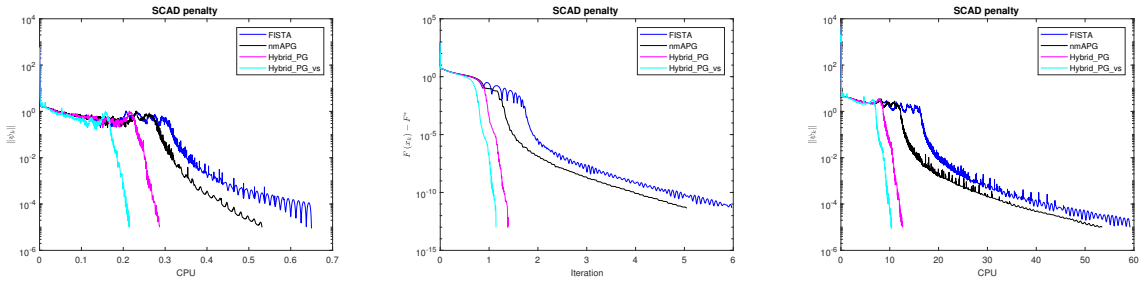

Fig. 2: Performance profile for the convergence of $\|\psi_k\|$ for solving the nonconvex penalty model with SCAD penalty. Set $\mu = 0.25$. Left: Example with $n = 100, m = 1000, s = 20$. Middle: Example with $n = 300, m = 3000, s = 30$. Right: Example with $n = 500, m = 5000, s = 50$.

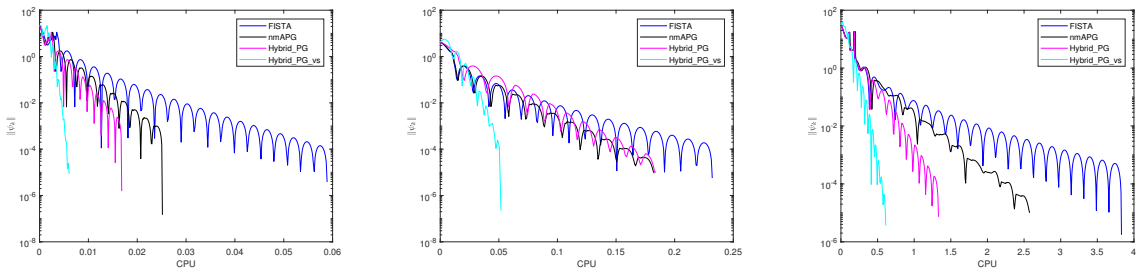

Fig. 3: Performance profile for the convergence of $\|\psi_k\|$ for solving the nonconvex constraint model.Left: Example with $m = 500$. Middle: Example with $m = 1000$. Right: Example with $m = 2000$.

Comparison results Evolutions of the value $\|\psi_k\|$ with respect to the CPU times are presented in the following Fig.1, 2, 3. Also, we report the specific performance values including the iterations and CPU time in Tabel 1, 2. From the presented results, we can see that our algorithm Hybrid_PG are faster and more stable than FISTA and nmAPG with the same fixed stepsize strategy. Further, Hybrid_PG_vs has a significant improvement over other three algorithms, which means that the variable stepsize nm_VS is effective for speed up the convergence of algorithm.

Table 2: Numerical comparisons of tested algorithms for solving the nonconvex constraint model

	m=500		m=1000		m=2000	
	Iter	CPUs	Iter	CPUs	Iter	CPUs
FISTA	336	0.0589	303	0.2321	582	3.8301
nmAPG	188	0.0252	229	0.1827	361	2.5763
Hybrid_PG	141	0.0168	171	0.1839	198	1.3315
Hybrid_PG_vs	36	0.0061	37	0.0518	69	0.6125

A Proof of (10)

Since that

$$x_k = T_{\lambda_k}(y_k) = \arg\min_{z\in R^n}\left\{g(z) + f(y_k) + \langle \nabla f(y_k), z - y_k\rangle + \frac{1}{2\lambda_k}\|z - y_k\|^2\right\},$$

we have

$$g(x_k) + f(y_k) + \langle \nabla f(y_k), x_k - y_k\rangle + \frac{1}{2\lambda}\|x_k - y_k\|^2 \le g(z) + f(y_k) + \langle \nabla f(y_k), z - y_k\rangle + \frac{1}{2\lambda}\|z - y_k\|^2. \quad (92)$$

Since that ∇f is L_f-Lipschitz continuous, we obtain that

$$F(x_k) + \left(\frac{1}{2\lambda_k} - \frac{L_f}{2}\right)\|x_k - y_k\|^2 \le F(z) + \left(\frac{1}{2\lambda_k} + \frac{L_f}{2}\right)\|z - y_k\|^2. \quad (93)$$

by using Lemma .

B the proof of Lemma 4.2

Proof We prove result $F(x_k) \le c_k$ by induction. Obviously, the result holds for $k = 0$. Suppose that for all $k = 1, 2, \cdots, j-1$, the result holds, then, we consider $k = j$.

Define $D_j(t) = \frac{tc_{j-1} + F(x_j)}{t+1}$.

If x_j generated by (33), we have

$$F(x_j) = F(z_j) \le c_{j-1} - \delta\|z_j - x_{j-1}\|^2.$$

Otherwise, x_j generated by (34), we have

$$F(x_j) \le F(x_{j-1}) \le c_{j-1}, \quad (94)$$

where the last inequality follows from the induction step. Hence, for any x_j,

$$F(x_j) \le c_{j-1} \quad (95)$$

and

$$\frac{d}{dt}D_j(t) = \frac{c_{j-1} - F(x_j)}{(t+1)^2} \ge 0,$$

which means that $D_j(t)$ is nondecreasing. Then,

$$F(x_j) = D_j(0) \le D_j(\eta q_{j-1}) = c_j,$$

which completed the proof. Further, the result $c_{k+1} \le c_k$ is trivially since (95).

References

1. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features, Math. Program. 116, 5-16 (2008)

2. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. Math. Oper. Res. 35(2), 438-457 (2010)
3. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality, Math. Oper. Res. 35, 438-457 (2010)
4. Attouch, H., Bolte, J., Svaier, B.F.:Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. Math. Program. 137, 91-129 (2013)
5. Attouch, H., Peypouquet, J.: The rate of convergence of Nesterov's accelerated forward_backward method is actually faster than $\frac{1}{k^2}$. SIAM J. Optim. 26, 1824-1834 (2016)
6. Attouch, H., Cabot, A.: Convergence rates of inertial forward-backward algorithms. SIAM J. Optim. 28, 849-874 (2018)
7. Ahookhosh, M., Themelis, A., Patrinos, P.: A Bregman forward-backward linesearch algorithm for nonconvex composite optimization: superlinear convergence to nonisolated local minima (2019). arXiv:1905.11904
8. Apidopoulos, V., Aujol, J., Dossal, C.: Convergence rate of inertial Forward-Backward algorithm beyond Nesterov's rule. Math. Program. 180, 137-156 (2020)
9. Apidopoulos, V., Aujol, J., Dossal, C. et al.: Convergence rates of an inertial gradient descent algorithm under growth and flatness conditions. Math. Program. (2020). https://doi.org/10.1007/s10107-020-01476-3
10. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. U.S.S.R. Computational Mathematics and Mathematical Physics. 7, 200-217 (1967)
11. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and Distributed Computation: Numerical Methods. PrenticeHall, New Jersey (1989)
12. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. 2(1), 183-202 (2009)
13. Beck, A., Teboulle, M.: Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems.in IEEE Transactions on Image Processing. 18, 2419-2434 (2009)
14. Bauschke, H.H., Combettes, P. L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, Berlin (2011)
15. Bolte, J., Sabach, S., Teboulle. M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. 146(1-2), 459-494 (2014)
16. Bolte, J., Nguyen, T.P., Peypouquet, J., Suter, B.W.: From error bounds to the complexity of first-order descent methods for convex functions. Math. Program. 165, 1-37 (2015)
17. Bauschke, H.H., Bolte, J., Teboulle, M.: A Descent Lemma Beyond Lipschitz Gradient Continuity: First-Order Methods Revisited and Applications. Math. Oper. Res. 42(2), 330-348 (2016)
18. Bolte, J., Sabach, S., Teboulle, M., Vaisbourd, Y.: First-order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. SIAM J. Optim. 28, 2131-2151 (2018)
19. Bauschke, H.H., Bolte, J., Chen, J., Teboulle, M., Wang, X.: On linear convergence of non-Euclidean gradient methods without strong convexity and Lipschitz gradient continuity. J. Optim. Theory Appl. 182, 1068-1087 (2019)
20. Chen, G.H.G., Rockafellar, R.T.: Convergence rates in forward?backward splitting. SIAM J. Optim. 7(2), 421-444 (1997)
21. Chouzenoux, E., Pesquet, J.C., Repetti, A.: Variable metric forward-backward algorithmfor minimizing the sum of a differentiable function and a convex function. J. Optim. Theory Appl. (2013)
22. 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)
23. Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun. Pure Appl. Math. 57(11), 1413-1457 (2004)
24. Donghwan, K., Jeffrey, A.F.: Another look at the fast iterative shrinkage/thresholding algorithm (FISTA). SIAM J. Optim. 28, 223-250 (2018)

25. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96, 1348-1360 (2001)
26. Frankel, P., Garrigos, G., Peypouquet. J.: Splitting methods with variable metric for Kurdyka-Łojasiewicz functions and general convergence rates. J. Optim. Theory Appl. 165, 874-900, 2014
27. Ghayem, F., Sadeghi, M., Babaie-Zadeh, M., Chatterjee, S., Skoglund, M., Jutten, C.: Sparse signal recovery using iterative proximal projection. IEEE Trans. Signal Process. 66, 879-894 (2018)
28. Hien, L.T.K., Gillis, N., Patrinos, P.: Inertial block mirror descent method for non-convex non-smooth optimization (2019). arXiv:1903.01818
29. Johnstone, P.R., Moulin, P.: Local and global convergence of a general inertial proximal splitting scheme for minimizing composite functions. Comput. Optim. Appl. 67, 259-292 (2017)
30. Kurdyka, K.: On gradients of functions definable in o-minimal structures. Ann. I. Fourier. 48(3), 769-783 (1998)
31. Liu, H.W., Wang, T., Liu, Z.X.:Convergence rate of inertial forward-backward algorithms based on the local error bound condition. http://arxiv.org/pdf/2007.07432
32. Liu, H.W., Wang, T., Liu, Z.X.: Some modified fast iteration shrinkage thresholding algorithms with a new adaptive non-monotone stepsize strategy for nonsmooth and convex minimization problems. Optimization online. http://www.optimization-online.org/DB_HTML/2020/12/8169.html
33. Liu, H.W., Wang, T.: A Nonmonontone Accelerated Proximal Gradient Method with Variable Stepsize Strategy for Nonsmooth and Nonconvex Minimization Problems. Optimization online. http://www.optimization-online.org/DB_HTML/2021/04/8365.html
34. Lu, H., Freund, R.M., Nesterov, Y.: Relatively smooth convex optimization by first-order methods, and applications. SIAM J. Optim. 28, 333-354 (2018)
35. Li, H., Lin, Z.: Accelerated proximal gradient methods for nonconvex programming. In: Proceedings of NeurIPS, pp. 379-387 (2015)
36. László, S.C. Convergence rates for an inertial algorithm of gradient type associated to a smooth non-convex minimization. Math. Program. (2020). https://doi.org/10.1007/s10107-020-01534-w
37. Mukkamala, M.C., Ochs, P., Pock, T., Sabach, S.: Convex-concave backtracking for inertial Bregman proximal gradient algorithms in non-convex optimization (2019). arXiv:1904.03537
38. Ortega, J.M., Rheinboldt, W.C.: Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, New-York (1970)
39. Ochs, P., Chen, Y., Brox, T., Pock, T.: Inertial proximal algorithm for nonconvex optimization. SIAM Journal on Imaging Sciences. 7(2), 1388-1419 (2014)
40. O'Donoghue, B., Candès, E.: Adaptive restart for accelerated gradient schemes. Found Comput Math. 15, 715-732 (2015)
41. Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Optim. 1(3), 127-239 (2014)
42. Su, W., Boyd, S., Candes, E.J.: A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. J. Mach. Learn. Res. 17, 1-43 (2016)
43. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Math. Program. 117, 387-423 (2009)
44. 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)
45. Wu, Z.M., Li, C.S., Li, M., Lim, A.: Inertial proximal gradient methods with Bregman regularization for a class of nonconvex optimization problems. J Global Optim. https://doi.org/10.1007/s10898-020-00943-7
46. Xu, Z., Chang, X.Y., Xu, F.M., Zhang, H.: L1/2 Regularization: A Thresholding Representation Theory and a Fast Solver. IEEE Trans. Neural Netw. Learn. Syst. 23(7), 1013-1027 (2012)
47. Xu, F.M., Lu, Z.S., Xu, Z.B.: An efficient optimization approach for a cardinality-constrained index tracking problem. Optimization Methods and Software. 31(2), 258-271 (2016)
48. Zeng L.M., Xie. J.: Group variable selection via SCAD-l2. Statistics. 48, 49-66 (2014)