manuscript No.
(will be inserted by the editor)

A Nonmonontone Accelerated Proximal Gradient Method with Variable Stepsize Strategy for Nonsmooth and Nonconvex Minimization Problems

Hongwei Liu[1] · Ting Wang[1] ·

Received: date / Accepted: date

Abstract We propose a new nonmonontone accelerated proximal gradient method with variable stepsize strategy for minimizing the sum of a nonsmooth function with a smooth one in the nonconvex setting. In this algorithm, the objective function value be allowed to increase discontinuously, but is decreasing from the overall point of view. The variable stepsize strategy don't need a line search or to know the Lipschitz constant, which makes the algorithm easier to implement. Every sequence of iterates generated by the algorithm converges to a critical point of the objective function. Further, under the assumption that the objective function satisfies the Kurdyka-Łojasiewicz inequality, we prove the convergence rates of the objective function value and the iterates. Moreover, numerical results on both convex and nonconvex problems are reported to demonstrate the effectiveness and superiority of the proposed methods and stepsize strategy.

Keywords Nonconvex, Nonsmooth, Accelerated proximal gradient method, variable stepsize strategy, Kurdyka-Łojasiewicz property, Convergence

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

1 Introduction

Triggered by practical problems in signal processing, image processing, and machine learning [24,28,47, 48], there has been an increased interest in so-called composite objective functions:

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

where f is a smooth (possibly nonconvex) function with Lipschitz continuous gradient and g is a proper

Ting Wang (✉)

E-mail: wangting_7640@163.com

Hongwei Liu

E-mail: hwliuxidian@163.com

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

lower semicontinuous (possibly 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 $\inf F > -\infty$.

In convex optimization, the property that one of the function is smooth and another is convex makes the proximal gradient (PG) method [46] well defined and be a benchmark approach for solving the problem (P). The concrete iterative scheme of this method can be read as:

$$x_{k+1} \in \operatorname{prox}_{\lambda_k g}\left(x_k - \lambda_k \nabla f\left(x_k\right)\right), \tag{1}$$

where $\lambda_k > 0$ denotes the stepsize and the proximal mapping of λg is defined by

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

Algorithm (1) is a descent scheme provided that $0 < \lambda_k \leq \frac{1}{L_f}$ and the sequence generated by it converges (weakly in an infinite-dimensional space) to the minimizer; and the convergence rate of objective function values is $o\left(\frac{1}{k}\right)$ [16,23,46]. In such convex setting, the well-known iterative shrinkage and soft-thresholding algorithm (ISTA) [14,27] and projected gradient method [39] are also special cases of this method with proximal operator derived by l_1-norm and indicator function of certain convex set, respectively.

Some accelerated proximal gradient (APG) algorithms be proposed in order to accelerate the convergence rate and enhance the numerical performance of the PG method by incorporating an inertial term, which is computed by the difference of the two preceding iterations, i.e.,

$$\begin{aligned} y_{k+1} &= x_k + \gamma_k\left(x_k - x_{k-1}\right) \quad \text{with} \quad \gamma_k \in [0,1) \\ x_{k+1} &\in \operatorname{prox}_{\lambda_{k+1} g}\left(y_{k+1} - \lambda_{k+1}\nabla f\left(y_{k+1}\right)\right). \end{aligned} \tag{3}$$

This seminal work proposed by Nestorev [41], who showed the $O\left(\frac{1}{k^2}\right)$ convergence rate with $\gamma_k = \frac{t_k - 1}{t_{k+1}}$, where $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ and $t_0 = 1$ for smooth setting and several extensions have been made in the nonsmooth and convex setting, for example, [4,6,8,9,12,13,26,29,33,35,36,49]. In the case of nonconvex, the inertial optimization algorithm has the ability to detect multiple critical points of nonconvex functions via an appropriate control of the ineritial parameter, while, non-inertial methods lack this property [20]. However, since that the proximal operator is not anymore single-valued [18], the convergence analysis for the inertial algorithm becomes more complex. A common method is to assume that the functions in the objective have the Kurdyka-Lojasiewicz property [34], which is almost always satisfied from a practical point of view of image processing, computer vision, or machine learning. Under the assumption that the nonsmooth part of the objective function is convex, while the smooth counterpart is allowed to be nonconvex, Ochs et al. [42] proposed an inertial proximal algorithm for nonconvex optimization (ipiano) and obtained convergence result based on the Kurdyka-Łojasiewicz inequality; Wenbo & Chen [53] proved the $R-$linear convergence for the APG with a extrapolation coefficient, which has an upper bound less than 1, under the error bound condition [50] (Bolte in [18] showed that there has a quantitative relationship between the error bound (EB) condition and the Kurdyka-Łojasiewicz property); Wu & Li [54] considered a general inertial proximal gradient method with two different extrapolation coefficients,

and the local linear convergence also can be established for the proposed method by using the EB condition.

With the growing use of non-convex objective function in applied fields like image processing and machine learning, the needs of numerical methods for the fully nonconvex setting increased significantly. However, it is difficult to extend the results directly from convex setting to nonconvex setting. A class of Bregman proximal gradient methods [57,43], which can be seen as a further developments on PG, be used for solving the fully nonconvex settting. The Bregman proximal gradient methods replaces the proximal mapping in (2) by

$$\text{prox}_{\lambda g}^{h}(u,v) := \underset{x\in R^n}{\arg\min}\left\{g(x) + \langle x, v\rangle + D_h(x,u)\right\}, \tag{4}$$

where $D_h(x,u)$, called Bregman distance, be defined as $D_h(x,u) := h(x) - (h(u) + \langle \nabla h(u), x-u\rangle)$ with a strong convex function h. Easy to observe that (4) can reduce to (2) if taking $h := \frac{1}{2}\|\cdot\|^2$. More on the property of the Bregman distance can refer to [10] and the detailed theoretical analysis for convex scheme can be found in [21,44,51]. Recently, several efforts on using Bregman proximal gradient methods incorporated inertial term to solve the nonconvex case of problem (P). Bolte [19] proposed a Bregman PG method for nonconvex setting, that is:

$$x_{k+1} \in \text{prox}_{\lambda_k g}^{h}(x_k, \lambda_k \nabla f(x_k) - \gamma_k(x_k - x_{k-1}))$$

and showed that every sequence of iterates generated by this algorithm converges to a critical point of the objective function provided an appropriate regularization of the objective satisfies the Kurdyka-Łojasiewicz inequality. This algoirthm can be reduced to the one proposed by Ochs [42] if taking $h := \frac{1}{2}\|\cdot\|^2$. Further, taking $\beta_k = 0$, the algorithm can be reduced to the one in [17]. Wu&Li [54] proposed an inertial Bregman proximal gradient method, that is

$$x_{k+1} \in \text{prox}_{\lambda_k g}^{h}(x_k, \lambda_k \nabla f(z_k) - \gamma_k(x_k - x_{k-1})),$$

where $z_k = x_k + \alpha_k(x_k - x_{k-1})$, and show that the sequence converges to the stationary point of objective function and linear convergence rate under Kurdyka-Łojasiewicz framework. More efforts on Bregman proximal gradient methods for solving nonconvex scheme can be found in [7,22,38,40,32,52].

Here, we foucus on the algorithm proposed in [37]. The author based on the idea of Beck and Teboulle's monotone FISTA [13], proposed a monotone APG for fully nonconvex setting. In this method, a proximal gradient step be using as the monitor to make the sufficient descent condition $F(x_{k+1}) \leq F(x_k) - \delta\|v_{k+1} - x_k\|^2$ satisfies, and the convergence rate of function value can be obtained under the assumption that objective function F has the Kurdyka-Łojasiewicz property. Meanwhile, the author extended it to a nonmonotone scheme by relaxing the sufficient descent condition as $F(x_{k+1}) \leq c_k - \delta\|v_{k+1} - x_k\|^2$, where c_k is a relaxation of $F(x_k)$, but can not obtain the corresponding convergence

result. The concrete iterative scheme is :

$$(nmAPG)\begin{cases} y_{k+1} = x_k + \frac{t_{k-1}}{t_k}\left(z_k - x_k\right) + \frac{t_{k-1}-1}{t_k}\left(x_k - x_{k-1}\right) \\ z_{k+1} = \mathrm{prox}_{\lambda g}\left(y_{k+1} - \lambda \nabla f\left(y_{k+1}\right)\right) \\ \text{if } F\left(z_{k+1}\right) \leq c_k - \rho \|z_{k+1} - y_k\|^2 \\ \qquad\qquad x_{k+1} = z_{k+1} \\ \text{else} \\ \qquad\qquad v_{k+1} = \mathrm{prox}_{\lambda g}\left(x_k - \lambda \nabla f\left(x_k\right)\right) \\ \qquad\qquad x_{k+1} = \begin{cases} z_{k+1}, & \text{if } F\left(z_{k+1}\right) \leq F\left(v_{k+1}\right), \\ v_{k+1}, & \text{otherwise.} \end{cases} \\ q_{k+1} = \eta q_k + 1,\ c_{k+1} = \left(\eta q_k c_k + F\left(x_{k+1}\right)\right)/q_{k+1} \end{cases}$$

Inspired by the algorithms in [37], we similarly combining the proximal gradient step to the inertial proximal gradient method to propose a new accelerated proximal gradient method with variable stepsize strategy (newAPG_vs) for solving the fully nonconvex and nonsmooth problem (P) in this paper. The newAPG_vs algorithm is nonmonotone, which allows function value increasing but the rising value no more than the drop-out value at the previous iteration such that the algorithm is declining on the whole. Although the constant stepsize is feasible, we still propose a variable stepsize strategy to speed up the convergence of algorithm from the numerical point of view. We can show that every accumulation point is a critical point. Then, under the assumption that objective function F have the KŁ property, we can obtain the convergence rates of function value and iterates.

The reminder of this paper is organized as follows. In Section 2, we provide our algorithm and show that any accumulation point of generated iterates converges to critical point. In Section 3, we suppose that objective function satisfies the KŁ inequality and show the convergence rates of function values and iterates. Numerical results are reported in Section 4.

2 A New Nonmonontone Accelerated Proximal Gradient Method with Variable Stepsize Strategy

In the following Algorithm 1, we give the concrete scheme of the new nonmonotone accelerated proximal gradient method with variable stepsize strategy (newAPG_vs). Easy to see that if the last iteration satisfies the sufficient descent condition, we introduce a trial step $\hat{x}$ generated by inertial proximal gradient method, which be accepted if function value at this trial point nonincreasing or increasing but the rising function value is less than δ times of the drop value of the previous step, where $\delta \in (0, 1)$; Otherwise, we use the proximal gradient method. If the sufficient descent condition can not be satisfied at the last iteration, we directly use the proximal gradient method to generate the iterates. Hence, the iterates generated by Algorithm 1 can be divided into two cases. Note that

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

Case 1. The trial step be accepted, i.e., $x_{k+1} = T_{\lambda_{k+1}}(y_{k+1})$ where $y_{k+1} = x_k + \gamma_k (x_k - x_{k-1})$, which be called InertialStep and satisfies the sufficient descent condition $\|x_k - x_{k-1}\|^2 \leq c(F(x_k) - F(x_{k-1}))$ and $F(x_{k+1}) \leq F(x_k) + \min(Q_k, \delta(F(x_{k-1}) - F(x_k)))$, which means that we allow the function value at present iteration to increase appropriately, but the increasing value cannot exceed δ times of the decrease of previous iteration. Meanwhile, we can deduce that

$$\begin{aligned} F(x_{k-1}) - F(x_{k+1}) &= F(x_{k-1}) - F(x_k) + F(x_k) - F(x_{k+1}) \\ &\geq (1-\delta)(F(x_{k-1}) - F(x_k)) \\ &\geq \left(\frac{1-\delta}{c}\right)\|x_k - x_{k-1}\|^2. \end{aligned} \tag{5}$$

Case 2. The trial step not be accepted, i.e., $x_{k+1} = T_{\lambda_{k+1}}(y_{k+1})$ where $y_{k+1} = x_k$, which be called ZeroStep since the inertial term equals to 0. Lemma 2.3 will show that the function value is decreasing if using the ZeroStep.

Algorithm 1 A New Nonmonontone Accelerated Proximal Gradient Method with Variable Stepsize Strategy (newAPG_vs)

Step 0. Take $x_0 \in R^n, \lambda_1 > 0, x_1 = p_{\lambda_1 g}(x_0), 0 < \mu_1 < \mu_0 < 1$. $\delta \in (0,1)$ and c is a large sufficiently positive constant.

Let $\sum_{k=1}^{\infty} Q_k$ and $\sum_{k=1}^{\infty} \mathrm{E}(k)$ are two convergent positive series. Set $0 < \mu_1 < \mu_0 < 1$ and $\gamma_k \in [0,1)$.

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

$$\lambda_{k+1} = \frac{\mu_1 \cdot \|x_k - y_k\|^2}{2|f(x_k) - f(y_k) - \langle \nabla f(x_k), x_k - y_k\rangle|} \tag{6}$$

otherwise, set

$$\lambda_{k+1} = \lambda_k + \min\{1, \lambda_k\}\mathrm{E}(k). \tag{7}$$

end

If $\|x_k - x_{k-1}\|^2 \leq c(F(x_{k-1}) - F(x_k))$

compute $\hat{y} = x_k + \gamma_k(x_k - x_{k-1})$ and $\hat{x} = T_{\lambda_{k+1}}(\hat{y})$

If $F(\hat{x}) \leq F(x_k) + \min(Q_k, \delta(F(x_{k-1}) - F(x_k)))$

$$y_{k+1} = \hat{y} \quad \text{and} \quad x_{k+1} = \hat{x} \tag{8}$$

else

$$x_{k+1} = T_{\lambda_{k+1}}(y_{k+1}) \quad \text{where} \quad y_{k+1} = x_k \tag{9}$$

end

else

$$x_{k+1} = T_{\lambda_{k+1}}(y_{k+1}) \quad \text{where} \quad y_{k+1} = x_k \tag{10}$$

end

The variable stepsize strategy in Algorithm 1 is also nonmonotonic. It uses the condition

$$2|f(x_k) - f(y_k) - \langle \nabla f(x_k), x_k - y_k\rangle| \leq \frac{\mu_0}{\lambda_k}\|x_k - y_k\|^2 \tag{11}$$

to control the increase or decrease of the stepsize. When the condition (11) does not holds, the stepsize λ_{k+1} is determined by (6), which implies that $\lambda_{k+1} < \lambda_k$. Conversely, $\lambda_{k+1} \geq \lambda_k$. And $\sum_{k=1}^{\infty} \mathrm{E}(k)$, which is called control series, is used for controlling the growth rate of stepsize. To analyze the convergence of Algorithm 1, we start from some significant properties of the stepsize $\{\lambda_k\}$ generated by Algorithm 2.

Lemma 2.1 *Let $\{\lambda_k\}$ be the sequence generated by Algorithm 2. We have that the sequence $\{\lambda_k\}$ is convergent. And for all k,*

$$\lambda_k \geq \lambda_{\min} = \min\left\{\lambda_1, \frac{\mu_1}{L_f}\right\}. \tag{12}$$

Proof See the detailed proof in the Appendix A.

Lemma 2.2 *For the sequence $\{\lambda_k\}$ generated by Algorithm 2, there exists a $\hat{k} \geq 1$, for every $k > \hat{k}$, condition (11) holds constantly.*

Proof The proof of this Lemma is developed in the Appendix B.

The Lemma 2.2 proved that the stepsize $\{\lambda_k\}$ generated by the variable stepsize strategy is non-monotone at previous finite steps $\hat{k}$, and after $\hat{k}$ step, it will increase monotonically.

Corollary 2.1 *For the sequence $\{\lambda_k\}$ generated by the variable stepsize strategy in Algorithm 1, denote $\lim_{k\to\infty} \lambda_k = \lambda^*$. Then, for any $k > \hat{k}$, we have $\lambda_k \leq \lambda^*$. And, there exists $\lambda_{\max} = \max\left(\lambda^*, \lambda_0, \cdots, \lambda_{\hat{k}}\right)$ such that $\lambda_k \leq \lambda_{\max}$ for all k.*

Remark 1 The Algorithm 1 with a constant stepsize can still be well defined if we set $\lambda < \frac{1}{L_f}$.

Now we begin to analyze the convergence of Algorithm 1 by proving some important properties in the following lemmas.

Lemma 2.3 *For $\mu_0 \in]0, 1]$, if $\{x_{k+1}\}$ and $\{y_{k+1}\}$ satisfy the condition (11), then, for any $z \in R^n$,*

$$F(x_{k+1}) + \left(\frac{1-\mu_0}{2\lambda_{k+1}}\right)\|x_{k+1} - y_{k+1}\|^2 \leq F(z) + \left(\frac{1}{2\lambda_{k+1}} + \frac{L_f}{2}\right)\|z - y_{k+1}\|^2, \quad \forall k > \hat{k}. \tag{13}$$

Proof. By the iterative scheme

$$x_{k+1} = T_{\lambda_{k+1}}(y_{k+1}) = \arg\min_z \left\{\langle \nabla f(y_{k+1}), z - y_{k+1}\rangle + \frac{1}{2\lambda_{k+1}}\|z - y_{k+1}\|^2 + g(z)\right\}, \tag{14}$$

we can deduce that

$$g(x_{k+1}) + \langle \nabla f(y_{k+1}), x_{k+1} - y_{k+1}\rangle + \frac{1}{2\lambda_{k+1}}\|x_{k+1} - y_{k+1}\|^2 \leq g(z) + \langle \nabla f(y_{k+1}), z - y_{k+1}\rangle + \frac{1}{2\lambda_{k+1}}\|z - y_{k+1}\|^2. \tag{15}$$

Using the fact that $-\nabla f$ is Lipschitz continuous, we have for any $y, z \in R^n$

$$f(z) \geq f(y) + \langle \nabla f(y), z - y\rangle - \frac{L_f}{2}\|z - y\|^2 \tag{16}$$

and recall the condition (11) that

$$f(x_{k+1}) - f(y_{k+1}) - \langle \nabla f(x_{k+1}), x_{k+1} - y_{k+1}\rangle \leq \frac{\mu_0}{2\lambda_{k+1}}\|x_{k+1} - y_{k+1}\|^2, \quad \forall k > \hat{k}. \tag{17}$$

Adding both side of (15) by $f(x_{k+1})$, the conclusion (13) follows from (16) with $y := y_{k+1}$ and (17).

Lemma 2.4 *For $\{x_k\}$ generated by the Algorithm 1, we have both of $F(x_k)$ and $\sum_{k=1}^{\infty} |F(x_k) - F(x_{k-1})|$ are convergent.*

Proof. Based on the scheme of Algorithm 1, we know that for the InertialStep, i.e., $x_{k+1} = T_{\lambda_{k+1}}(y_{k+1})$ where $y_{k+1} = x_k + \gamma_k(x_k - x_{k-1})$, it satisfied that $F(x_{k+1}) - F(x_k) \leq Q_k$, which means that

$$(F(x_{k+1}) - F(x_k))^+ \leq Q_k. \tag{18}$$

For the ZeroStep, i.e., $x_{k+1} = T_{\lambda_{k+1}}(y_{k+1})$ where $y_{k+1} = x_k$, using (13) with $z := x_k$, we have

$$(F(x_{k+1}) - F(x_k))^+ = 0 \leq Q_k. \tag{19}$$

Then, combining (18) and (19), we have $\sum_{k=1}^{\infty} (F(x_{k+1}) - F(x_k))^+$ is convergent since that $\sum_{k=1}^{\infty} Q_k$ is a convergent positive series. Next, we show that $\sum_{k=1}^{\infty} (F(x_{k+1}) - F(x_k))^-$ is convergent. We know that

$$F(x_{k+1}) - F(x_k) = (F(x_{k+1}) - F(x_k))^+ - (F(x_{k+1}) - F(x_k))^-$$

and

$$\sum_{k=1}^{\infty} F(x_{k+1}) - F(x_k) = \sum_{k=1}^{\infty} (F(x_{k+1}) - F(x_k))^+ - \sum_{k=1}^{\infty} (F(x_{k+1}) - F(x_k))^-. \tag{20}$$

Assume to the contrary that $\sum_{k=1}^{\infty} (F(x_{k+1}) - F(x_k))^- = +\infty$, then following from (20) that $F(x_k) \to -\infty$, which contradicts the fact that $\{F(x_k)\}$ is bounded below. Hence, we have $\sum_{k=1}^{\infty} (F(x_{k+1}) - F(x_k))^-$ is convergent. Then $\{F(x_k)\}$ is convergent following from (20). Further, $\sum_{k=1}^{\infty} |F(x_{k+1}) - F(x_k)|$ is convergent since that $\sum_{k=1}^{\infty} |F(x_{k+1}) - F(x_k)| = \sum_{k=1}^{\infty} (F(x_{k+1}) - F(x_k))^+ + \sum_{k=1}^{\infty} (F(x_{k+1}) - F(x_k))^-$.

Lemma 2.5 *For $\{x_k\}, \{y_k\}$ generated by the Algorithm 1. Then $\sum_{k=1}^{\infty} \|x_k - y_k\|^2$ is convergent.*

Proof. Using (13) with $z := x_k$, we have

$$\begin{aligned} \left(\frac{1-\mu_0}{2\lambda_{k+1}}\right)\|x_{k+1} - y_{k+1}\|^2 &\leq F(x_k) - F(x_{k+1}) + \left(\frac{1}{2\lambda_{k+1}} + \frac{L_f}{2}\right)\|x_k - y_{k+1}\|^2 \\ &\leq |F(x_{k+1}) - F(x_k)| + \left(\frac{1}{2\lambda_{k+1}} + \frac{L_f}{2}\right)\|x_k - y_{k+1}\|^2. \end{aligned} \tag{21}$$

For the InertialStep, we have

$$\begin{aligned} &\left(\frac{1-\mu_0}{2\lambda_{k+1}}\right)\|x_{k+1} - y_{k+1}\|^2 \\ &\leq |F(x_{k+1}) - F(x_k)| + \left(\frac{1}{2\lambda_{k+1}} + \frac{L_f}{2}\right)\gamma_k^2\|x_k - x_{k-1}\|^2 \\ &\leq |F(x_{k+1}) - F(x_k)| + \left(\frac{1}{2\lambda_{k+1}} + \frac{L_f}{2}\right)c\,(F(x_{k-1}) - F(x_k)) \\ &\leq |F(x_{k+1}) - F(x_k)| + \left(\frac{1}{2\lambda_{k+1}} + \frac{L_f}{2}\right)c\,|F(x_{k-1}) - F(x_k)|. \end{aligned} \tag{22}$$

For the ZeroStep, i.e. $y_{k+1} = x_k$, we can deduce (21) to

$$\left(\frac{1-\mu_0}{2\lambda_{k+1}}\right)\|x_{k+1}-y_{k+1}\|^2 \le |F(x_{k+1})-F(x_k)| \tag{23}$$
$$\le |F(x_{k+1})-F(x_k)| + \left(\frac{1}{2\lambda_{k+1}}+\frac{L_f}{2}\right)c\,|F(x_{k-1})-F(x_k)|.$$

Combining (22), (23) with Lemma 2.4 and Lemma 2.1, we can deduce that $\sum_{k=1}^{\infty}\|x_{k+1}-y_{k+1}\|^2$ is convergent.

Lemma 2.6 *For $\{x_k\}$ generated by the Algorithm 1. We have $\sum_{k=1}^{\infty}\|x_{k+1}-x_k\|^2$ is convergent.*

Proof. For the InertialStep, we have

$$\|x_{k+1}-x_k\|^2 \le 2\|x_{k+1}-y_{k+1}\|^2 + 2\|y_{k+1}-x_k\|^2 \tag{24}$$
$$\le 2\|x_{k+1}-y_{k+1}\|^2 + 2\|x_k-x_{k-1}\|^2$$
$$\le 2\|x_{k+1}-y_{k+1}\|^2 + 2c\,(F(x_{k-1})-F(x_k))$$
$$\le 2\|x_{k+1}-y_{k+1}\|^2 + 2c|F(x_{k-1})-F(x_k)|.$$

For the ZeroStep, obviously,

$$\|x_{k+1}-x_k\|^2 = \|x_{k+1}-y_{k+1}\|^2 \le 2\|x_{k+1}-y_{k+1}\|^2 + 2c|F(x_{k-1})-F(x_k)|. \tag{25}$$

Combining (24), (25), the conclusion that $\sum_{k=1}^{\infty}\|x_{k+1}-x_k\|^2$ is convergent follows from Lemma 2.4 and Lemma 2.5.

Lemma 2.7 *[17] Let (x_k,u_k) be a sequence such that $x_k\to x$, $u_k\to u$, $F(x_k)\to F(x)$ and $u_k\in\partial F(x_k)$, then $u\in\partial F(x)$.*

Theorem 2.1 *Let $\{x_k\}$ generated by Algorithm 1. Then, all the accumulation point of the $\{x_k\}$ belongs to* $\operatorname{crit} F := \{x\in R^n : 0\in\partial F(x)\}$.

Proof. We can easy to show that $\{x_k\}$ is bounded by the fact that $\{F(x_k)\}$ is coercive. Suppose that $\{x_{k_j}\}$ is a convergent subsequence of $\{x_k\}$ and $\lim_{j\to\infty} x_{k_j} = \hat{x}$. Following from (14), we have

$$\nabla f(x_{k+1}) - \nabla f(y_{k+1}) - \frac{1}{\lambda_{k+1}}(x_{k+1}-y_{k+1}) \in \partial F(x_{k+1}). \tag{26}$$

Since the fact that ∇f is Lipschitz continuous gradient and Lemma 2.5, we obtain that

$$\left\|\nabla f(x_{k+1}) - \nabla f(y_{k+1}) - \frac{1}{\lambda_{k+1}}(x_{k+1}-y_{k+1})\right\| \le \left(L_f+\frac{1}{\lambda_{k+1}}\right)\|x_{k+1}-y_{k+1}\| \to 0. \tag{27}$$

In addition, from (15) with $z := \hat{x}$, $k+1 := k_j+1$, we have

$$\left\langle \nabla f(y_{k_j+1}), x_{k_j+1}-y_{k_j+1}\right\rangle + \frac{1}{2\lambda_{k_j+1}}\left\|x_{k_j+1}-y_{k_j+1}\right\|^2 + g(x_{k_j+1})$$
$$\le \left\langle \nabla f(y_{k_j+1}), \hat{x}-y_{k_j+1}\right\rangle + \frac{1}{2\lambda_{k_j+1}}\left\|\hat{x}-y_{k_j+1}\right\|^2 + g(\hat{x}), \tag{28}$$

which means that $\limsup_{j\to\infty} g\left(x_{k_j+1}\right) \leq g\left(\hat{x}\right)$. Combining with $\liminf_{j\to\infty} g\left(x_{k_j+1}\right) \geq g\left(\hat{x}\right)$ from the definition of lower semicontinuous of g, we have $\lim_{j\to\infty} g\left(x_{k_j+1}\right) = g\left(\hat{x}\right)$. Moreover, since f is continuously differentiable, we have $\lim_{j\to\infty} f\left(x_{k_j+1}\right) = f\left(\hat{x}\right)$. Hence,

$$\lim_{j\to\infty} F\left(x_{k_j+1}\right) = F\left(\hat{x}\right). \tag{29}$$

Combining $\lim_{j\to\infty} x_{k_j} = \hat{x}$, (26), (27) and (29), using Lemma 2.7, we have $0 \in \partial F\left(\hat{x}\right)$.

Theorem 2.2 *Denote $\omega\left(x_k\right)$ is the set of all accumulation points of $\{x_k\}$ generated by Algorithm 2. For $F^* = \lim_{k\to\infty} F\left(x_k\right)$, we have $F\left(\omega\left(x_k\right)\right) \equiv F^*$.*

Proof. For any $\hat{x} \in \omega\left(x_k\right)$, there exists a $\left\{x_{k_j}\right\}$ such that $\lim_{j\to\infty} x_{k_j} = \hat{x}$. It follows that

$$F\left(\hat{x}\right) \leq \liminf_{j\to\infty} F\left(x_{k_j}\right) = \lim_{k\to\infty} F\left(x_k\right) = F^* \tag{30}$$

from the fact that F is lower semicontinuous. In addition, recalling (13) and set $x = \hat{x}$, we have

$$F\left(x_{k_j+1}\right) + \left(\frac{1-\mu_0}{2\lambda_{k_j+1}}\right)\left\|x_{k_j+1} - y_{k_j+1}\right\|^2 \leq F\left(\hat{x}\right) + \left(\frac{1}{2\lambda_{k_j+1}} + \frac{L_f}{2}\right)\left\|\hat{x} - y_{k_j+1}\right\|^2. \tag{31}$$

Following from $\lim_{j\to\infty}\left\|x_{k_j+1} - y_{k_j+1}\right\|^2 = 0$, $\lim_{j\to\infty}\left\|\hat{x} - y_{k_j+1}\right\|^2 = 0$ and $\lim_{j\to\infty} \lambda_{k_j+1} = \lambda^*$, we have

$$F^* = \lim_{k\to\infty} F\left(x_k\right) = \limsup_{j\to\infty} F\left(x_{k_j+1}\right) \leq F\left(\hat{x}\right). \tag{32}$$

Combining (30) and (32), we have

$$F^* = \lim_{k\to\infty} F\left(x_k\right) = F\left(\hat{x}\right).$$

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

3 convergence rate of the function values.

In order to continue our analysis for the convergence rates of the function values and iterates, a slightly more assumption to the objective, namely that it satisfies the Kurdyka-Łojasiewicz inequality be in common use.

We state the definition of the Kurdyka-Łojasiewicz property: 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 3.1 (Kurdyka-Łojasiewicz property) [17] Let $F : R^m \to R$ be a differentiable function. We say that F satisfies the Krudyka-Lojasiewica (KL) property at $\bar{x} \in R^m$ 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

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

the following, so called KŁ inequality, holds

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

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

Following uniformized KŁ property given in [17] plays an important role in our convergence analysis.

Lemma 3.1 *[17] Let* $\mathrm{X} \subseteq R^n$ *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* X *and* F *satisfies the KŁ property at each point of* X. *Then, there exist* $\varepsilon, \eta > 0$ *and* $\varepsilon, \eta > 0, \varphi \in \Theta_\eta$ *such that for all* $\bar{x} \in \mathrm{X}$ *and for all* x *in the intersection*

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

the following inequality holds

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

A remarkable aspect of KŁ functions is that they are ubiquitous in applications, for example, semi-algebraic, subanalytic and log-exp. To the class of KŁ functions belong real sub-analytic, semi-convex,uniformly convex and convex functions satisfying a growing condition, we refer the reader to [1,2,3,5,11,15] and the references therein for more details regarding all the classed mentioned above and illustrating examples. Further, based on the KŁ property, an abstract convergence theorem for descent methods with certain properties is proved in [1,2,3,25,31]. Obviously, our algorithm not a descent method, therefore this abstract convergence theorem is not applicable to ours. To obtain the convergence rate of function value and iterates of our algorithm, we define two sets as follows:

$$\bar{\Omega} = \{i | F\left(x_{i-1}\right) > F\left(x_i\right) \text{ and } F\left(x_{i+1}\right) \geq F\left(x_i\right)\}$$

and

$$\Omega = \{1, 2, \cdots\} \setminus \bar{\Omega}.$$

The following two lemmas show the properties of the set Ω, which is crucial for the later proofs.

Lemma 3.2 *For any* $i \in \Omega$, *we have* $i + 1 \in \Omega$ *or* $i + 2 \in \Omega$.

Proof It is obviously if $i + 1 \in \Omega$. Otherwise, we have $i + 1 \in \bar{\Omega}$, i.e., $F\left(x_i\right) > F\left(x_{i+1}\right)$ and $F\left(x_{i+2}\right) \geq F\left(x_{i+1}\right)$, which means that the point x_{i+2} is produced by the InertialStep and the function value from x_{i+1} to x_{i+2} is nondecreasing, hence, the sufficient descent condition not holds. Then, the next step must be the ZeroStep, which means that $F\left(x_{i+3}\right) < F\left(x_{i+2}\right)$, i.e., $i + 2 \in \Omega$.

Lemma 3.3 *For any* $i_j, i_{j+1} \in \Omega$, *we have* $F\left(x_{i_{j+1}}\right) < F\left(x_{i_j}\right)$.

Proof From Lemma 3.2, there must be $i_{j+1} = i_j + 1$ or $i_{j+1} = i_j + 2$. Assume to the contrary that there exists a subscript $\bar{i}_j \in \Omega$ such that $F\left(x_{\bar{i}_{j+1}}\right) \geq F\left(x_{\bar{i}_j}\right)$.

(I) Considering the case that $i_{j+1} = i_j + 1$. Since that $F\left(x_{\bar{i}_{j+1}}\right) = F\left(x_{\bar{i}_j+1}\right) \geq F\left(x_{\bar{i}_j}\right)$, the function value from $x_{\bar{i}_j}$ to $x_{\bar{i}_j+1}$ is nondecreasing, the point $x_{\bar{i}_j+1}$ must be generated by the InertialStep and the previous iteration $x_{\bar{i}_j-1}$ must satisfies the sufficient descent condition, then, $F\left(x_{\bar{i}_j-1}\right) > F\left(x_{\bar{i}_j}\right)$ and $F\left(x_{\bar{i}_j+1}\right) \geq F\left(x_{\bar{i}_j}\right)$, i.e., $\bar{i}_j \in \bar{\Omega}$, which contradicts the fact that $\bar{i}_j \in \Omega$.

(II) Considering the case that $\bar{i}_{j+1} = \bar{i}_j + 2$, which implies that $\bar{i}_j + 1 \in \bar{\Omega}$, i.e., $F\left(x_{\bar{i}_j}\right) > F\left(x_{\bar{i}_j+1}\right)$ and $F\left(x_{\bar{i}_j+2}\right) \geq F\left(x_{\bar{i}_j+1}\right)$, which means that the iteration $x_{\bar{i}_j+2}$ generated by the InertialStep, then, from (5), we can obtain that

$$F\left(x_{\bar{i}_j}\right) - F\left(x_{\bar{i}_j+2}\right) \geq \frac{1-\delta}{c_k}\left\|x_{\bar{i}_j+1} - x_{\bar{i}_j}\right\|^2 > 0,$$

which contradicts the fact that $F\left(x_{\bar{i}_j+2}\right) = F\left(x_{\bar{i}_{j+1}}\right) \geq F\left(x_{\bar{i}_j}\right)$.

In the following theorems we provide convergence rates for objective function value and iterates generated by Algorithm 1 by assuming that the objective function F satisfies the KŁ property with a desingularizing function $\varphi(t) := \frac{C}{\theta}t^\theta$.

Theorem 3.1 *(Convergence rate of objective function values) 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 $\theta = 1$, $F(x_k)$ converges in finite steps.

(2) If $\theta \in \left[\frac{1}{2}, 1\right)$, there exists $Q \in (0,1)$ such that

$$\left|F(x_k) - F^*\right| = O\left(Q^k\right).$$

(3) 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 Define $r_k = F(x_k) - F^*$.

Step 1. Considering the subsequence $\{k_j\} \subseteq \Omega$. From the Lemma 3.3, we can obtain that for any $k_j \in \Omega$, $\{F(x_{k_j})\}$ is monotonically decreasing. Then, $F(x_{k_j}) > F^*$ and $F(x_{k_j}) \to F^*$ as $j \to +\infty$, i.e., $r_{k_j} > 0$ and $r_{k_j} \to 0$ as $j \to +\infty$. In addition, from Theorem 2.1, we know that $w(x_{k_j}) \subset w(x_k) \subset \text{crit } F$. Since the assumption that F is coercive, we have $w(x_{k_j})$ is bounded. Also, it is compact. Hence, following from Lemma 3.1 with $\mathrm{X} := w(x_{k_j})$, there exist $\varepsilon > 0$, $\eta \in (0, +\infty]$, a concave function $\varphi(t) := \frac{C}{\theta}t^\theta$ and j_1 such that for all $j > j_1$,

$$x_{k_j} \in \left\{x | dist(x, w(x)) \leq \varepsilon \cap F^* < F(x) < F^* + \eta\right\}$$

such that

$$\varphi'\left(F(x_{k_j}) - F^*\right) dist\left(0, \partial F(x_{k_j})\right) \geq 1. \tag{33}$$

Recalling (13) and set $z := x_{k_j-1}$, and $k+1 := k_j$, we have that

$$\begin{aligned}\left\|x_{k_j} - y_{k_j}\right\|^2 &\leq \left(\frac{2\lambda_{k_j}}{1-\mu_0}\right)\left(F(x_{k_j-1}) - F(x_{k_j}) + \left(\frac{1}{2\lambda_{k_j}} + \frac{L_f}{2}\right)\left\|x_{k_j-1} - y_{k_j}\right\|^2\right) \\ &\leq M_1\left(F(x_{k_j-1}) - F(x_{k_j}) + M_2\left\|x_{k_j-1} - y_{k_j}\right\|^2\right),\ \forall j > \hat{j},\end{aligned} \tag{34}$$

where $\hat{j}$ such that $k_{\hat{j}} = \hat{k}$, $M_1 = \frac{2\lambda_{\max}}{1-\mu_0}$ and $M_2 = \frac{1}{2\lambda_{\min}} + \frac{L_f}{2}$.

(I) For the case that x_{k_j} be generated by ZeroStep, we have $F\left(x_{k_j-1}\right) > F\left(x_{k_j}\right)$, which means that $k_j - 1 \in \Omega$, i.e., $k_j - 1 = k_{j-1}$. Then, (34) becomes

$$\left\|x_{k_j} - y_{k_j}\right\|^2 \leq M_1\left(F\left(x_{k_j-1}\right) - F\left(x_{k_j}\right)\right) = M_1\left(F\left(x_{k_{j-1}}\right) - F\left(x_{k_j}\right)\right) \leq M_1\left(F\left(x_{k_{j-2}}\right) - F\left(x_{k_j}\right)\right) \quad (35)$$

where the last inequality follows from Lemma 3.3.

(II) For the case that x_{k_j} be generated by InertialStep, by the scheme of Algorithm 1, we know that x_{k_j-1} must satisfies the sufficient descent condition, then $F\left(x_{k_j-2}\right) > F\left(x_{k_j-1}\right)$, then, (34) becomes

$$\begin{aligned}\left\|x_{k_j} - y_{k_j}\right\|^2 &\leq M_1\left(F\left(x_{k_j-1}\right) - F\left(x_{k_j}\right) + M_2\left\|x_{k_j-1} - x_{k_j-2}\right\|^2\right) \quad (36)\\ (5) &\leq M_1\left(F\left(x_{k_j-1}\right) - F\left(x_{k_j}\right) + M_2\left(\frac{c}{1-\delta}\right)\left(F\left(x_{k_j-2}\right) - F\left(x_{k_j}\right)\right)\right)\\ &< M_3\left(F\left(x_{k_j-2}\right) - F\left(x_{k_j}\right)\right).\end{aligned}$$

where $M_3 = M_1\left(1 + M_2\left(\frac{c}{1-\delta}\right)\right)$.

(i) If $F\left(x_{k_j-1}\right) > F\left(x_{k_j}\right)$, we have $F\left(x_{k_j-2}\right) > F\left(x_{k_j-1}\right) > F\left(x_{k_j}\right)$, which means that $k_j - 1 = k_{j-1} \in \Omega$ and $k_j - 2 = k_{j-2} \in \Omega$. Hence, by Lemma 3.3, (36) becomes that

$$\left\|x_{k_j} - y_{k_j}\right\|^2 < M_3\left(F\left(x_{k_{j-2}}\right) - F\left(x_{k_j}\right)\right), \quad (37)$$

(ii) If $F\left(x_{k_j-1}\right) \leq F\left(x_{k_j}\right)$, combining with $F\left(x_{k_j-2}\right) > F\left(x_{k_j-1}\right)$, we have $k_j - 1 \in \bar{\Omega}$ and $k_j - 2 = k_{j-1} \in \Omega$. Hence, (36) becomes that

$$\left\|x_{k_j} - y_{k_j}\right\|^2 < M_3\left(F\left(x_{k_{j-1}}\right) - F\left(x_{k_j}\right)\right) \leq M_3\left(F\left(x_{k_{j-2}}\right) - F\left(x_{k_j}\right)\right). \quad (38)$$

From (35), (37) and (38), we obtain that for any $k_j \in \Omega$,

$$\left\|x_{k_j} - y_{k_j}\right\|^2 \leq M_3\left(F\left(x_{k_{j-2}}\right) - F\left(x_{k_j}\right)\right) = M_3\left(r_{k_{j-2}} - r_{k_j}\right) \quad \forall j > \hat{j}. \quad (39)$$

Since (33) with $\varphi(t) = \frac{C}{\theta}t^\theta$, $\varphi'(t) = Ct^{\theta-1}$, we have that for any $j > j_0 = \max\left(j_1, \hat{j}\right)$ such that

$$\begin{aligned}1 &\leq \left(\varphi'\left(F\left(x_{k_j}\right) - F^*\right) dist\left(0, \partial F\left(x_{k_j}\right)\right)\right)^2 \quad (40)\\ (27) &\leq \left(\varphi'\left(r_{k_j}\right)\right)^2\left(\frac{1}{\lambda_{k_j}} + L_f\right)^2\left\|x_{k_j} - y_{k_j}\right\|^2\\ (39) &\leq M_3\left(\frac{1}{\lambda_{k_j}} + L_f\right)^2\left(\varphi'\left(r_{k_j}\right)\right)^2\left\|r_{k_{j-2}} - r_{k_j}\right\|^2\\ &= \tilde{M}\left(r_{k_j}\right)^{2\theta-2}\left(r_{k_{j-2}} - r_{k_j}\right)\end{aligned}$$

where $\tilde{M} = C^2 M_3\left(\frac{1}{\lambda_{\min}} + L_f\right)^2$.

Case 1. $\theta = 1$. Then, (40) becomes that $1 \leq \tilde{M}\left(r_{k_{j-2}} - r_{k_j}\right)$, which against the fact that $r_{k_j} \to 0$. Hence, there exists $\bar{j}$ such that for any $j > \bar{j}$, $r_{k_j} = 0$, i.e., there exists $\bar{k} \in \Omega$ such that

$$r_k = 0, \quad \forall k > \bar{k} \text{ and } k \in \Omega. \quad (41)$$

Case 2. $\theta \in \left[\frac{1}{2}, 1\right)$. Since that $r_{k_j} \to 0$ and $0 < 2-2\theta \leq 1$, there exists j_2 such that $\left(r_{k_j}\right)^{2-2\theta} \geq r_{k_j}$ for all $j > j_2$. Hence, there exists $\tilde{j} > \max\left(j_0, j_2\right)$ such that for all $j > \tilde{j}$, (40) becomes

$$F\left(x_{k_j}\right) - F^* = r_{k_j} \leq \frac{\tilde{M}}{1+\tilde{M}} r_{k_{j-2}} \leq \cdots \leq \left(\frac{\tilde{M}}{1+\tilde{M}}\right)^{\frac{j-j_0}{2}} r_{k_{j_0}}. \tag{42}$$

Case 3. $\theta \in \left(0, \frac{1}{2}\right)$. We can easily obtain that $2\theta - 2 \in (-2, -1)$ and $2\theta - 1 \in (-1, 0)$. Then, since $r_{k_{j-2}} > r_{k_j}$, we have $\left(r_{k_{j-2}}\right)^{2\theta-2} < \left(r_{k_j}\right)^{2\theta-2}$ and $\left(r_{k_0}\right)^{2\theta-1} < \cdots < \left(r_{k_{j-2}}\right)^{2\theta-1} < \left(r_{k_j}\right)^{2\theta-1}$.

Define $\phi(t) = \frac{1}{1-2\theta} t^{2\theta-1}$, then, $\phi'(t) = -t^{2\theta-2}$.

(i) If $\left(r_{k_j}\right)^{2\theta-2} \leq 2\left(r_{k_{j-2}}\right)^{2\theta-2}$, then, for any $j > j_0$,

$$\begin{aligned} \phi\left(r_{k_j}\right) - \phi\left(r_{k_{j-2}}\right) &= \int_{r_{k_{j-2}}}^{r_{k_j}} \phi'(t) dt = \int_{r_{k_j}}^{r_{k_{j-2}}} t^{2\theta-2} dt \\ &\geq \left(r_{k_{j-2}} - r_{k_j}\right)\left(r_{k_{j-2}}\right)^{2\theta-2} \\ &\geq \frac{1}{2}\left(r_{k_j} - r_{k_{j-2}}\right)\left(r_{k_j}\right)^{2\theta-2} \\ (40) &\geq \frac{1}{2\tilde{M}}. \end{aligned} \tag{43}$$

(ii) If $\left(r_{k_j}\right)^{2\theta-2} > 2\left(r_{k_{j-2}}\right)^{2\theta-2}$, then, $\left(r_{k_j}\right)^{2\theta-1} \geq 2^{\left(\frac{2\theta-1}{2\theta-2}\right)}\left(r_{k_{j-2}}\right)^{2\theta-1}$.

$$\begin{aligned} \phi\left(r_{k_j}\right) - \phi\left(r_{k_{j-2}}\right) &= \frac{1}{1-2\theta}\left(\left(r_{k_j}\right)^{2\theta-1} - \left(r_{k_{j-2}}\right)^{2\theta-1}\right) \\ &\geq \frac{1}{1-2\theta}\left(2^{\left(\frac{2\theta-1}{2\theta-2}\right)} - 1\right)\left(r_{k_{j-2}}\right)^{2\theta-1} \\ &\geq \frac{1}{1-2\theta}\left(2^{\left(\frac{2\theta-1}{2\theta-2}\right)} - 1\right)\left(r_{k_0}\right)^{2\theta-1}. \end{aligned} \tag{44}$$

Hence, by (43) and (44), we obtain that for any $j > j_0$,

$$\phi\left(r_{k_j}\right) - \phi\left(r_{k_{j-2}}\right) \geq D,$$

where $D = \min\left(\frac{1}{2\tilde{M}}, \frac{1}{1-2\theta}\left(2^{\left(\frac{2\theta-1}{2\theta-2}\right)} - 1\right)\left(r_{k_0}\right)^{2\theta-1}\right)$. Then, for $j > j_0$,

$$\begin{aligned} \phi\left(r_{k_j}\right) &\geq \left(\phi\left(r_{k_j}\right) - \phi\left(r_{k_{j-2}}\right)\right) + \left(\phi\left(r_{k_{j-2}}\right) - \phi\left(r_{k_{j-4}}\right)\right) + \cdots + \left(\phi\left(r_{k_{j_0+2}}\right) - \phi\left(r_{k_{j_0}}\right)\right) \\ &\geq \left(\frac{j-j_0}{2}\right) D, \end{aligned} \tag{45}$$

i.e.,

$$\left(r_{k_j}\right)^{2\theta-1} \geq (1-2\theta)\left(\frac{j-j_0}{2}\right) D,$$

and

$$F\left(x_{k_j}\right) - F^* = r_{k_j} \leq \left(\frac{2}{D(1-2\theta)(j-j_0)}\right)^{\frac{1}{1-2\theta}}.$$

Hence, for any $k \in \Omega$, there exists $\tilde{k} = k_{\tilde{j}} \in \Omega$ such that for any $k > \tilde{k}$,

$$\left|F(x_k) - F^*\right| = r_k \leq \left(\frac{\tilde{M}}{1+\tilde{M}}\right)^{\frac{k-\tilde{k}}{2}} r_{\tilde{k}} \quad \text{for} \quad \theta \in \left[\frac{1}{2}, 1\right), \tag{46}$$

and

$$\left|F(x_k) - F^*\right| = r_k \leq \left(\frac{2}{D(1-2\theta)\left(k-\tilde{k}\right)}\right)^{\frac{1}{1-2\theta}}, \quad \text{for} \quad \theta \in \left(0, \frac{1}{2}\right). \tag{47}$$

Step 2. Consider the case that $k \in \bar{\Omega}$.

In this case, $k-1, k+1 \in \Omega$, $F(x_{k+1}) \geq F(x_k)$, then, the iteration x_{k+1} must be generated by the InertialStep and $F^* < F(x_{k+1}) < F(x_{k-1})$. If $F(x_k) > F^*$, then,

$$\left|F(x_k) - F^*\right| = F(x_k) - F^* \leq F(x_{k+1}) - F^*. \tag{48}$$

Otherwise, $F(x_k) \leq F^*$, then,

$$\left|F(x_k) - F^*\right| = F^* - F(x_k) \leq F(x_{k+1}) - F(x_k) \leq \delta\left(F(x_{k-1}) - F(x_k)\right).$$

Since $F(x_{k-1}) - F(x_{k+1}) = F(x_{k-1}) - F(x_k) + F(x_k) - F(x_{k+1}) \geq (1-\delta)\left(F(x_{k-1}) - F(x_k)\right)$, we have

$$\begin{aligned} \left|F(x_k) - F^*\right| &\leq \left(\frac{\delta}{1-\delta}\right)\left(F(x_{k-1}) - F(x_{k+1})\right) \\ &\leq \left(\frac{\delta}{1-\delta}\right)\left(F(x_{k-1}) - F^*\right). \end{aligned} \tag{49}$$

Since that $k-1, k+1 \in \Omega$, and $\delta \in (0,1)$, we can deduce by (48) and (49) that

$$\left|F(x_k) - F^*\right| \leq \max\left(1, \frac{\delta}{1-\delta}\right)\left(F(x_{k-1}) - F^*\right).$$

Combining with (41), (46) and (47), we have that for any $k \in \bar{\Omega}$,

$$\left|F(x_k) - F^*\right| = 0, \quad \text{for} \quad \theta = 1, \tag{50}$$

$$\left|F(x_k) - F^*\right| \leq \left(\frac{\tilde{M}}{1+\tilde{M}}\right)^{\frac{k-1-\tilde{k}}{2}} r_{k_0}, \quad \text{for} \quad \theta \in \left[\frac{1}{2}, 1\right), \tag{51}$$

and

$$\left|F(x_k) - F^*\right| \leq \left(\frac{2M}{D(1-2\theta)\left(k-1-\tilde{k}\right)}\right)^{\frac{1}{1-2\theta}}, \quad \text{for} \quad \theta \in \left(0, \frac{1}{2}\right). \tag{52}$$

Hence, the proof be completed by (50), (51) and (52).

Theorem 3.2 *(Convergence rate of iterates) 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 $C > 0$, $\theta \in \left(\frac{1}{4}, 1\right]$. Then,*

(1) If $\theta = 1$, $\{x_k\}$ converges in finite steps.

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

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

Proof From (13) with $z := x_{k-1}$, $k+1 := k$, we have

$$\left(\frac{1-\mu_0}{2\lambda_k}\right)\|x_k - y_k\|^2 \leq F(x_{k-1}) - F(x_k) + \left(\frac{1}{2\lambda_k} + \frac{L_f}{2}\right)\|x_{k-1} - y_k\|^2, \quad \forall k > \hat{k}. \tag{53}$$

If the iteration x_k be generated by the ZeroStep, then, $y_k = x_{k-1}$, which means that

$$\|x_k - x_{k-1}\|^2 \leq \left(\frac{2\lambda_k}{1-\mu_0}\right)\left(F(x_{k-1}) - F(x_k)\right) \leq \left(\frac{2\lambda_{\max}}{1-\mu_0}\right)\left(\left|F(x_{k-1}) - F^*\right| + \left|F(x_k) - F^*\right|\right). \tag{54}$$

Otherwise, the iteration x_k be generated by the InertialStep, then,

$$\begin{aligned}\|x_k - x_{k-1}\|^2 &\leq 2\|x_k - y_k\|^2 + 2\|y_k - x_{k-1}\|^2 \\ &\leq 2\|x_k - y_k\|^2 + 2\|x_{k-1} - x_{k-2}\|^2 \\ &\leq 2\|x_k - y_k\|^2 + 2c\left(F\left(x_{k-2}\right) - F\left(x_{k-1}\right)\right).\end{aligned} \tag{55}$$

By (53), we have

$$\|x_k - y_k\|^2 \leq \left(\frac{2\lambda_k}{1-\mu_0}\right)\left(F\left(x_{k-1}\right) - F\left(x_k\right) + \left(\frac{1}{2\lambda_k} + \frac{L_f}{2}\right)c\left(F\left(x_{k-2}\right) - F\left(x_{k-1}\right)\right)\right), \tag{56}$$

we have

$$\begin{aligned}&\|x_k - x_{k-1}\|^2 \\ &\leq \left(\frac{4\lambda_k}{1-\mu_0}\right)\left(F\left(x_{k-1}\right) - F\left(x_k\right) + \left(\frac{1}{2\lambda_k} + \frac{L_f}{2}\right)c\left(F\left(x_{k-2}\right) - F\left(x_{k-1}\right)\right)\right) + 2c\left(F\left(x_{k-2}\right) - F\left(x_{k-1}\right)\right) \\ &\leq M_4\left(F\left(x_{k-1}\right) - F\left(x_k\right) + F\left(x_{k-2}\right) - F\left(x_{k-1}\right)\right) = M_4\left(F\left(x_{k-2}\right) - F\left(x_k\right)\right) \\ &\leq M_4\left(\left|F\left(x_{k-2}\right) - F^*\right| + \left|F\left(x_k\right) - F^*\right|\right)\end{aligned} \tag{57}$$

where $M_4 = \max\left(\left(\frac{4\lambda_{\max}}{1-\mu_0}\right), \left(\frac{4c\lambda_{\max}}{1-\mu_0}\right)\left(\frac{1}{2\lambda_k} + \frac{L_f}{2}\right) + 2c\right)$. Combining with (54) and (57), we have

$$\|x_k - x_{k-1}\|^2 \leq 2M_4 \cdot \max\left(\left|F\left(x_{k-2}\right) - F^*\right|, \left|F\left(x_{k-1}\right) - F^*\right|, \left|F\left(x_k\right) - F^*\right|\right). \tag{58}$$

Then, by the results of Theorem 3.1, we can obtain that

(1) for $\theta = 1$, $\{x_k\}$ converges in finite steps.

(2) for $\theta \in \left[\frac{1}{2}, 1\right)$, there exists constant $C_1 > 0$ such that

$$\|x_k - x_{k-1}\| \leq \sqrt{2C_1 M_4 r_{k_0}}\, Q^{\frac{k}{2}}. \tag{59}$$

Hence, for any $p > 0$,

$$\begin{aligned}\left\|x_{k+p} - x_k\right\| &\leq \sum\nolimits_{i=k+1}^{k+p} \|x_i - x_{i-1}\| \leq \sqrt{2C_1 M_4 r_{k_0}} \int_k^{k+p} Q^{\frac{x}{2}}\, dx \\ &= -\frac{\sqrt{2C_1 M_5 r_{k_0}}}{|\ln Q|}\left(Q^{\frac{x}{2}}\right)\Big|_k^{k+p} \leq \frac{\sqrt{2C_1 M_5 r_{k_0}}}{|\ln Q|}\left(\sqrt{Q}\right)^k,\end{aligned} \tag{60}$$

i.e., $\{x_k\}$ is Cauchy sequence. Let $\lim\limits_{k\to\infty} x_k = \bar{x}$. As $p \to \infty$, we have

$$\|x_k - \bar{x}\| \leq \frac{\sqrt{2C_1 M_4 r_{k_0}}}{|\ln Q|}\left(\sqrt{Q}\right)^k.$$

(3) For $\theta \in \left(\frac{1}{4}, \frac{1}{2}\right)$, there exists constant $C_2 > 0$ such that

$$\|x_k - x_{k-1}\| \leq \sqrt{2C_2 M_4}\, k^{-\frac{1}{2(1-2\theta)}}.$$

Hence, for any $p > 0$,

$$\begin{aligned}\left\|x_{k+p} - x_k\right\| &\leq \sum\nolimits_{i=k+1}^{k+p} \|x_i - x_{i-1}\| \leq \sqrt{2C_2 M_4} \int_k^{k+p} x^{-\frac{1}{2(1-2\theta)}}\, dx \\ &= -\sqrt{2C_2 M_4}\frac{2\left(1-2\theta\right)}{4\theta - 1}\, x^{\frac{1-4\theta}{2(1-2\theta)}}\Big|_k^{k+p} \leq \sqrt{2C_2 M_4}\frac{2\left(1-2\theta\right)}{4\theta-1} k^{\frac{1-4\theta}{2(1-2\theta)}},\end{aligned} \tag{61}$$

i.e., $\{x_k\}$ is Cauchy sequence. Let $\lim\limits_{k\to\infty} x_k = \bar{x}$. As $p \to \infty$, we have

$$\|x_k - \bar{x}\| \leq \sqrt{2C_2 M_4}\frac{2\left(1-2\theta\right)}{4\theta-1} k^{-\frac{4\theta-1}{2(1-2\theta)}}.$$

The proof is completed.

4 Numerical Results

In this section, we conduct numerical experiments to illustrate the effectiveness of Algorithm 1 by considering three different types of problems: "convex + convex"; "convex + nonconvex"; "nonconvex + nonconvex". We consider four different algorithms for each class of problems: newAPG (Algorithm 1 with fixed stepsize); FISTA with fixed stepsize [12]; nmAPG with fixed stepsize (See Section 1) and newAPG_vs (Algorithm 1). Note that FISTA is not necessarily convergent for nonconvex optimization theoretically. We take $\lambda \equiv \frac{0.98}{L_f}$ for the first three algorithms and for the newAPG_vs, we set the initial stepsize λ_0 as local Lipschitz constant between initial point x_0 and $x_0 + 10^{-5}$. In the experiment, all algorithms use the same inertia term: $\gamma_k = \frac{t_k - 1}{t_{k+1}}$, where $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ and $t_1 = 1$. And for nmAPG, taking $\eta = 0.8$ for c_k and $\rho = 10^{-4}$. For the Algorithm 1, taking $\delta = 0.8$, $c = 10^4$, $Q_k = 0.99^k$, $E_k = 1/k^{1.1}$, $\mu_0 = 0.99$ and $\mu_1 = 0.95$.

The computational results are presented in following figures and tables. In each figure, we plot $\|\psi_k\|$ against the CPU time, where $\partial F(x_k) \ni \psi_k = \nabla f(x_k) - \nabla f(y_k) - \frac{1}{\lambda_k}(x_k - y_k)$. We also use the $\|\psi_k\| \leq TOL$ with $TOL = 10^{-5}$ to terminate algorithms. The number of iterations and CPU time for different settings of test problem be listed in tables.

4.1 "Convex + Convex"

In this subsection, we consider the LASSO:

$$\min_{x \in R^n} \frac{1}{2}\|Ax - b\|^2 + \mu\|x\|_1. \tag{62}$$

We observe that (62) is in the form of probelm (P) with$f(x) = \frac{1}{2}\|Ax - b\|^2$ and $g(x) = \mu\|x\|_1$. It is clear that f has a Lipschitz continuous gradient with $L_f = \lambda_{\max}\left(A^T A\right)$. In order to investigate the stability and efficiency of the algorithms, we test 3 scenarios with different n and m. We generated an $n \times m$ matrix A with i.i.d. standard Gaussian entries. Taking $x_0 = [0, 0, \cdots, 0]^T$. The vector $b \in R^n$ then generated as $b = A\hat{x} + 0.01\epsilon$, where $\hat{x}$ is an s-sparse random vector and ϵ has standard i.i.d. Gaussian entries. The computational results are presented in Fig. 1 and Table 1.

Table 1: Numerical comparisons of different algorithms for solving LASSO

	n=300,m=3000 s=30,$\mu = 0.25$		n=500,m=5000 s=50,$\mu = 0.01$		n=800,m=8000 s=80,$\mu = 0.1$	
	Iter	CPUs	Iter	CPUs	Iter	CPUs
FISTA	3038	2.3488	16496	107.7547	11819	213.0590
nmAPG	2739	2.1701	14260	97.1115	10838	201.1279
newAPG	799	0.6488	6072	39.9542	2389	43.6238
newAPG_vs	485	0.4780	4642	36.3045	1625	36.7511

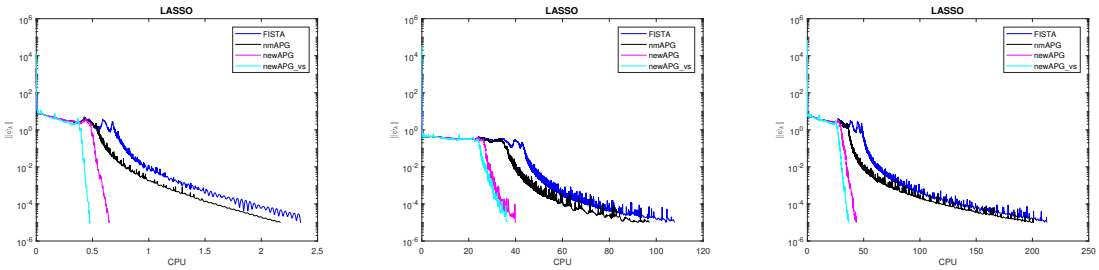

Fig. 1: Evolutions of $\|\psi_k\|$ with respect to the CPU time for solving LASSO. Left: Example with (n, m, s)=(300,3000,30) and $\mu = 0.25$; Middle: Example with (n, m, s)=(500,5000,50) and $\mu = 0.01$; Right: Example with (n, m, s)=(800,8000,80) and $\mu = 0.1$.

We can observe that the newAPG (Algorithm 1 with fixed stepsize) better than FISTA and nmAPG. In addition, newAPG_vs faster than newAPG, which means that the variable stepsize strategy can speed up the convergence of algorithm further.

4.2 "Convex + Nonconvex"

In this section, we provide a series of simulations to demonstrate the high performance of our algorithm. The numerical experiments are conducted by applying algorithm nmAPG_vs to nonconvex penalty model with $L_{1/2}$ and SCAD penalties. The concrete problems can be read as:

$$\min_{x\in R^m} \frac{1}{2}\|Ax-b\|^2 + \mu \|x\|_{1/2}^{1/2}, \tag{63}$$

where $\|x\|_{1/2}^{1/2} = \sum_{i=1}^n |x_i|^{1/2}$; and

$$\min_{x\in R^m} \frac{1}{2}\|Ax-b\|^2 + \mu \sum\nolimits_{i=1}^n g_\kappa(|x_i|), \tag{64}$$

where

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

The proximal mapping of $L_{1/2}$ and SCAD penalties can be found in [55] and [30,58] separately.

Note that in [30] the values of the parameters c and κ were suggested to be chosen pairwise over a two-dimensional grids using some criteria such as the cross-validation; and $c = 3.7$ was suggested therein. And We set $\kappa = 0.1\sqrt{2\log(m)}$ inspired by [30]. Similar with the Subsection 4.1, we generate the matrix $A \in R^{n\times m}$ for (n, m, s)=(100,1000,20), (300,3000,30) and (500,5000,50) and vector $b \in R^n$. Taking $x_0 = [0, 0, \cdots, 0]^T$.

In Fig. 2, 3 and Table 2, we can see that, as in the previous subsection, the algorithm newAPG better than FISTA and nmAPG; and newAPG_vs is always the fastest algorithm.

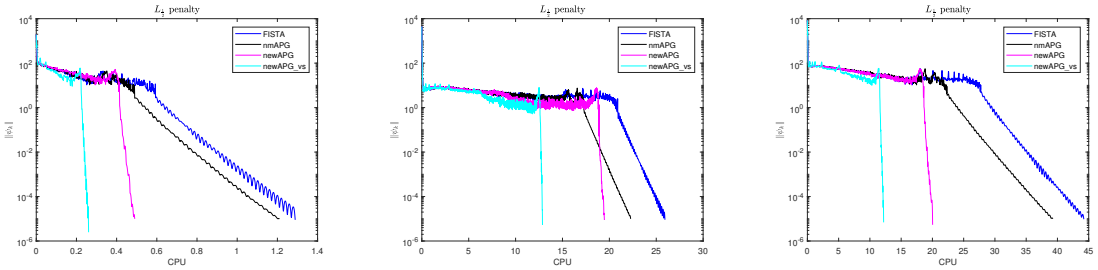

Fig. 2: Evolutions of $\|\psi_k\|$ with respect to the CPU time for $L_{\frac{1}{2}}$ penalty problem. Left: Example with (n, m, s)=(100,1000,20) and $\mu = 1$; Middle: Example with (n, m, s)=(300,3000,30) and $\mu = 0.1$; Right: Example with (n, m, s)=(500,5000,50) and $\mu = 0.25$.

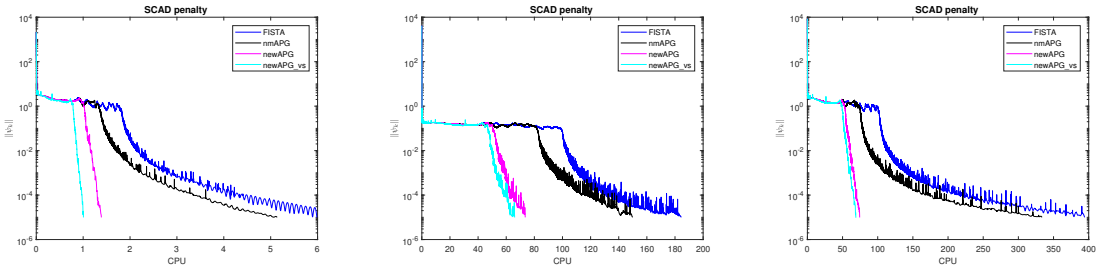

Fig. 3: Evolutions of $\|\psi_k\|$ with respect to the CPU time for SCAD penalty. Left: Example with (n, m, s)=(100,1000,20) and $\mu = 0.25$; Middle: Example with (n, m, s)=(300,3000,30) and $\mu = 0.25$; Right: Example with (n, m, s)=(500,5000,50) and $\mu = 0.25$.

Table 2: Numerical comparisons of different algorithms for solving the $L_{\frac{1}{2}}$ and SCAD penalty problems

	n=300,m=3000,s=30				n=500,m=5000,s=50				n=800,m=8000,s=80			
	$L_{\frac{1}{2}}$		SCAD		$L_{\frac{1}{2}}$		SCAD		$L_{\frac{1}{2}}$		SCAD	
	Iter	CPUs	Iter	CPUs	Iter	CPUs	Iter	CPUs	Iter	CPUs	Iter	CPUs
FISTA	1160	1.2887	5365	5.9869	3364	25.9288	23752	184.4883	1969	44.2255	18726	394.2484
nmAPG	1031	1.2101	4415	5.1376	2835	22.2703	17737	149.8109	1678	39.2621	15380	333.6720
newAPG	379	0.4901	1192	1.3888	2540	19.4692	9148	73.5371	899	20.0315	3520	74.3744
newAPG_vs	186	0.2605	786	1.0065	1435	12.8670	6947	66.0125	483	12.1949	2659	68.8939

4.3 "Nonconvex + Nonconvex"

In this subsection, we look at problems of the following form:

$$\min_{x \in \Delta_r^u} \frac{1}{2} x^T A x - b^T x, \tag{65}$$

where $\Delta_r^u := \left\{ x \in R^m : \sum_{i=1}^n x_i = s,\ \|x\|_0 \leq r,\ 0 \leq x_i \leq u,\ i = 1, \cdots, m \right\}$. Notice that one can rewrite (65) in the form of problem (P) by defining $f(x) = \frac{1}{2} x^T A x - b^T x$ 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 [56]. 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 $b = \text{randn}\,(m,1)$, $s = \max\{1, 10t\}$ where t is chosen uniformly at random from $[0,1]$, $r = \lfloor \frac{m}{100} \rfloor$ and $u = \max\{10, s\}$. Taking $x_0 = [s, 0, \cdots, 0]^T$.

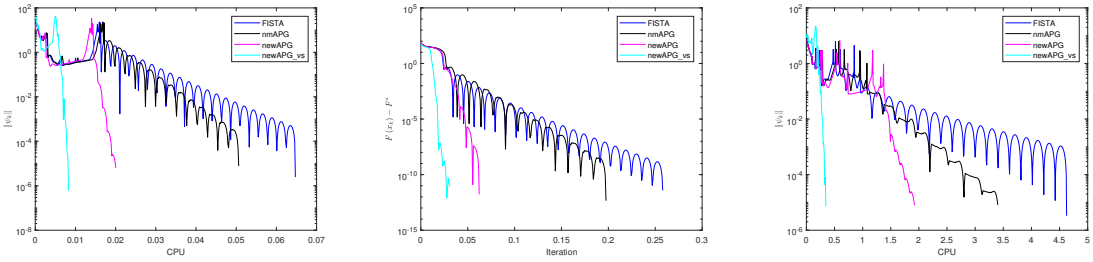

Fig. 4: Evolutions of $\|\psi_k\|$ with respect to the CPU time for nonconvex constraint problem. Left: Example with $m = 500$; Middle: Example with $m = 1000$; Right: Example with $m = 2000$.

Table 3: Numerical comparisons of different algorithms for solving the nonconvex constraint problem

	m=500		m=1000		m=2000	
	Iter	CPUs	Iter	CPUs	Iter	CPUs
FISTA	407	0.0647	346	0.2580	708	4.6253
nmAPG	302	0.0506	235	0.1975	466	3.4013
newAPG	141	0.0200	82	0.0623	270	1.9216
newAPG_vs	46	0.0083	23	0.0321	37	0.3456

The computational results are presented in Fig. 4 and Table 3. From the numerical results, we see that same as the previous subsections, our algorithm newAPG better that FISTA and nmAPG based on same fixed stepsize strategy; and newAPG with the variable stepsize strategy can speed up the convergence of algorithm further. Moreover, these three types of test problems show that our algorithm is effective for both convex and nonconvex problems.

A Proof of Lemma 2.1

Proof By the adaptive non-monotone stepsize strategy, we have for any $i \geq 1$

$$\lambda_{i+1} - \lambda_i \leq \mathrm{E}\,(i)\,. \tag{66}$$

Denote that

$$\lambda_{i+1} - \lambda_i = (\lambda_{i+1} - \lambda_i)^+ - (\lambda_{i+1} - \lambda_i)^-, \text{where}(\cdot)^+ = \max\{0, \cdot\}, (\cdot)^- = -\min\{0, \cdot\}, \tag{67}$$

we have

$$(\lambda_{i+1} - \lambda_i)^+ \leq \mathrm{E}\,(i)\,, \forall i = 1, 2, \cdots, \tag{68}$$

which implies that $\sum_{i=1}^{\infty}(\lambda_{i+1}-\lambda_i)^+$ is convergent from the fact that $\sum_{i=1}^{\infty}\mathrm{E}(i)$ is a convergent positive series.

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

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

$$\lambda_{k+1}-\lambda_1=\sum_{i=1}^{k}(\lambda_{i+1}-\lambda_i)=\sum_{i=1}^{k}(\lambda_{i+1}-\lambda_i)^+-\sum_{i=1}^{k}(\lambda_{i+1}-\lambda_i)^-. \tag{69}$$

We can easily deduce $\lim_{k\to\infty}\lambda_k=-\infty$, which is a contradiction with $\lambda_k>0$, $\forall k\geq 1$. Therefore, $\sum_{i=1}^{\infty}(\lambda_{i+1}-\lambda_i)^-$ is a convergent series. Then, in view of (69), we obtain the sequence $\{\lambda_k\}$ is convergent.

We can easily to prove that $\forall k\geq 1$, $\lambda_k\geq\min\left\{\lambda_1,\frac{\mu_1}{L_f}\right\}$ holds by induction.

B Proof of Lemma 2.2

Proof Suppose that the conclusion is not true, there exists a $\{k_j\}$ and $k_j\to\infty$ such that

$$2\left(f\left(x_{k_j}\right)-f\left(y_{k_j}\right)-\left\langle\nabla f\left(x_{k_j}\right),x_{k_j}-y_{k_j}\right\rangle\right)>\frac{\mu_0}{\lambda_{k_j}}\left\|x_{k_j}-y_{k_j}\right\|^2 \tag{70}$$

holds. Then, based on the scheme of adaptive nonmonotone stepsize, we have

$$\lambda_{k_j+1}=\frac{\mu_1\cdot\left\|x_{k_j}-y_{k_j}\right\|^2}{2\left|f\left(x_{k_j}\right)-f\left(y_{k_j}\right)-\left\langle\nabla f\left(x_{k_j}\right),x_{k_j}-y_{k_j}\right\rangle\right|}. \tag{71}$$

From the above two formulas, easy to obtain

$$\left\|x_{k_j}-y_{k_j}\right\|^2<\frac{2\lambda_{k_j}}{\mu_0}\left|f\left(x_{k_j}\right)-f\left(y_{k_j}\right)-\left\langle\nabla f\left(x_{k_j}\right),x_{k_j}-y_{k_j}\right\rangle\right|=\frac{\mu_1\lambda_{k_j}}{\mu_0\lambda_{k_j+1}}\left\|x_{k_j}-y_{k_j}\right\|^2 \tag{72}$$

There have a contradiction because of

$$\frac{\mu_1\lambda_{k_j}}{\mu_0\lambda_{k_j+1}}\to\frac{\mu_1}{\mu_0}<1. \tag{73}$$

Therefore, (11) will holds constantly after a finite step $\hat{k}$.

References

1. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. Ser. B 116(1-2), 5-16 (2009)
2. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projectionmethods for non-convex problems: an approach based on the Kurdyka??ojasiewicz inequality. Math. Oper. Res. 35(2), 438-457 (2010)
3. Attouch, H., Bolte, J., Svaiter, 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(1-2), 91-129 (2013)
4. 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)
5. 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(2), 471-507 (2017)
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. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. 17(4), 1205-1223 (2006)
12. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. 2, 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. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. 2(1), 183-202 (2009)
15. Bolte, J., Daniilidis, A., Ley, O., Mazet, L.: Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. Trans. Am. Math. Soc. 362(6), 3319-3363 (2010)
16. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, Berlin (2011)
17. Bolte, J., Sabach, S., Teboulle. M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. 146(1-2), 459-494 (2014)
18. 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)
19. Bot, R.I., Csetnek, E.R., Lszl, S.C.: An inertial forward-backward algorithm for the minimization of the sum of two nonconvex functions. EURO J. Comput. Optim. 4, 3-25 (2016)
20. Bot, R.I., Csetnek, E.R., Lszl, S.C.: Approaching nonsmooth non-convex minimization through second-order proximal-gradient dynamical systems. J. Evol. Equ. 18(3), 1291-1318 (2018)
21. 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)
22. 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)
23. Chen, G.H.G., Rockafellar, R.T.: Convergence rates in forward-backward splitting. SIAM J. Optim. 7(2), 421-444 (1997)
24. Combettes, P.L., Pesquet, J.C.: Proximal splitting methods in signal processing. In: Bauschke, H.H., Burachik, R.S., Combettes, P.L., Elser, V., Luke, D.R., Wolkowicz, H. (eds.) Fixed-point Algorithms for Inverse Problems in Science and Engineering, pp.185-212. Springer (2011)
25. 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. App. (2013)
26. 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)
27. 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)
28. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory 52(44), 1289-1306 (2006)

29. Donghwan, K., Jeffrey, A.F.: Another look at the fast iterative shrinkage/thresholding algorithm (FISTA). SIAM J. Optim. 28, 223-250 (2018)
30. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96, 1348-1360 (2001)
31. Frankel, P., Garrigos, G., Eypouquet. J.P.: Splitting methods with variable metric for Kurdyka-Łojasiewicz functions and general convergence rates. J. Optim. Theory Appl. 165, 874-900 (2014)
32. Hien, L.T.K., Gillis, N., Patrinos, P.: Inertial block mirror descent method for non-convex non-smooth optimization (2019). arXiv:1903.01818
33. 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)
34. Kurdyka, K.: On gradients of functions definable in o-minimal structures. Ann. I. Fourier. 48(3), 769-783 (1998)
35. 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
36. 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
37. Li, H., Lin, Z.: Accelerated proximal gradient methods for nonconvex programming. In: Proceedings of NeurIPS, 379-387 (2015)
38. Lu, H., Freund, R.M., Nesterov, Y.: Relatively smooth convex optimization by first-order methods, and applications. SIAM J. Optim. 28, 333-354 (2018)
39. Maingé, P.E., Gobinddass, M.: Convergence of one-step projected gradient methods for variational inequalities. J. Optim. Theory Appl. 171(1), 146-168 (2016)
40. 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
41. Nesterov, Y.: A method for solving the convex programming problem with convergence rate $O\left(\frac{1}{k^2}\right)$. Dokl. Akad. Nauk SSSR. 269, 543-547 (1983)
42. Ochs, P., Chen, Y., Brox, T., Pock, T.: Inertial proximal algorithm for nonconvex optimization. SIAM J. Imaging Sci. 7(2), 1388-1419 (2014)
43. Osher, S., Burger, M., Goldfarb, D., Xu, J., Yin, W.: An iterative regularization method for total variationbased image restoration. Multiscale Model. Simul. 4, 460-489 (2005)
44. Ochs, P., Fadili, J., Brox, T.: Non-smooth non-convex Bregman minimization: unification and new algorithms. J. Optim. Theory Appl. 181, 244-278 (2019)
45. Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Optim. 1(3), 127-239 (2014)
46. Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Optim. 1(3), 127-239 (2014)
47. Palomar, D.P., Eldar, Y.C.: Convex Optimization in Signal Processing and Communications. Cambridge University Press, Cambridge (2010)
48. Sra, S., Nowozin, S., Wright, S.J.: Optimization for Machine Learning. MIT Press, Cambridge (2012)
49. 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)
50. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Math. Program. 117, 387-423 (2009)
51. Teboulle, M.: A simplified view of first order methods for optimization. Math. Program. 170, 67-96 (2018)
52. Themelis, A., Stella, L., Patrinos, P.: Forward-backward envelope for the sum of two nonconvex functions: further properties and nonmonotone linesearch algorithms. SIAM J. Optim. 28, 2274-2303 (2018)
53. 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)

54. 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
55. 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)
56. Xu, F.M., Lu, Z.S., Xu, Z.B.: An efficient optimization approach for a cardinality-constrained index tracking problem. Optim. Method. Softw. 31(2), 258-271 (2016)
57. Yin, W., Osher, S., Goldfarb, D., Darbon, J.: Bregman iterative algorithms for l-minimization with applications to compressed sensing. SIAM J. Imaging Sci. 1, 143-168 (2008)
58. Zeng L.M., Xie. J.: Group variable selection via SCAD-l2. Statistics. 48, 49-66 (2014)