CARPATHIAN J. MATH. **34** (2018), No. 3, 449 - 457
Online version at http://carpathian.ubm.ro
Print Edition: ISSN 1584 - 2851 Online Edition: ISSN 1843 - 4401

*Dedicated to Professor Yeol Je Cho on the occasion of his retirement*

# A note on the accelerated proximal gradient method for nonconvex optimization

HUIJUAN WANG and HONG-KUN XU

ABSTRACT. We improve a recent accelerated proximal gradient (APG) method in [Li, Q., Zhou, Y., Liang, Y. and Varshney, P. K., *Convergence analysis of proximal gradient with momentum for nonconvex optimization*, in Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017] for nonconvex optimization by allowing variable stepsizes. We prove the convergence of the APG method for a composite nonconvex optimization problem under the assumption that the composite objective function satisfies the Kurdyka-Łojasiewicz property.

## 1. INTRODUCTION

In this paper we consider a composite optimization problem of the form

$$\min_{x\in H} F(x) := f(x) + g(x), \tag{1.1}$$

where $H = \mathbb{R}^d$ is a Euclidean $d$-space, and $f$ and $g$ are proper, lower-semicontinuous functions from $H$ to $(-\infty, \infty]$.

In the convex case (i.e., $f$ and $g$ are convex functions), the proximal gradient method [6] can well be used to solve (1.1); moreover, Nesterov's acceleration technique can be used to speed up the rate of convergence from $O(\frac{1}{k})$ to $O(\frac{1}{k^2})$ [3]. However, this remains open for nonconvex optimization.

Very recently in [10], Li, et al proposed a new algorithm which is known as an accelerated proximal gradient (APG) method which constructs three sequences $(x_k)$, $(y_k)$ and $(v_k)$ in such a way that $x_k$ is produced from $y_k$ through the composite of the proximal mapping of $g$ and the gradient $\nabla f$ of $f$ ($f$ is assumed to have Lipschitz continuous gradient), and $v_k$ is simply a linear combination of $x_k$ and $x_{k-1}$. Li, et al proved that their algorithm guarantees, under certain conditions, that the sequence $(x_k)$ is bounded, each cluster point of $(x_k)$ is a critical point of $F$, and $F$ is constant on the set of cluster points of $(x_k)$. They also obtained errors on the residual of $F(x_k) - \inf F$ under the uniformized Kurdyka-Łojasiewicz property with desingularizing function $\varphi(t) = ct^\theta$, where $c$ is a constant and $\theta \in (0, 1]$.

We continue working in this line by improving the algorithm and results of Li, et al [10] in twofold. First we allow the stepsizes to vary with the iteration steps and obtain the same convergence results of [10, Theorem 3]. Secondly, we prove convergence with finite length of our algorithm under the Kurdyka-Łojasiewicz property with a genaral desingularizing function, which is not discussed in Li, et al [10].

Received: 17.09.2017. In revised form: 13.06.2018. Accepted: 15.07.2018
2010 *Mathematics Subject Classification.* 90C26, 90C30, 90C46.
Key words and phrases. *accelerated proximal gradient method, composite nonconvex optimization, Kurdyka-Łojasiewicz property, subdifferential, critical point.*
Corresponding author: Hong-Kun Xu; xuhk@hdu.edu.cn

## 2. PRELIMINARIES

Let $d \geq 1$ be a given integer and consider the Euclidean $d$-space $\mathbb{R}^d$ with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$ (i.e., $\| \cdot \|_2$). By $\Gamma(\mathbb{R}^d)$ we denote the family of all functions $f : \mathbb{R}^d \to (-\infty, \infty] =: \overline{\mathbb{R}}$ which are proper and lower-semicontinuous (lsc).

### 2.1. **Subdifferential of Nonconvex Functions.**

**Definition 2.1.** Let $f \in \Gamma(\mathbb{R}^d)$ and $x \in \mathrm{dom}(f)$ be given. We say that $x^* \in \mathbb{R}^d$ is a Fréchet derivative of $f$ at $x$ if

$$\liminf_{z \to x} \frac{f(z) - f(x) - \langle x^*, z - x \rangle}{\|z - x\|} \geq 0.$$

The set of Fréchet derivatives of $f$ at $x$, denoted $\hat{\partial} f(x)$, is said to be the Fréchet differential of $f$ at $x$.

The (Mordukhovich) limiting-subdifferential (or simply, subdifferential) of $f$ at $x$, denoted $\partial f(x)$, is defined as

$$\partial f(x) = \{x^* \in H : \exists\, x_k^* \to x^*,\ x_k^* \in \hat{\partial} f(x_k) \text{ with } x_k \xrightarrow{f} x\}.$$

Here "$\xrightarrow{f}$" means $f$-convergence, that is, $x_k \xrightarrow{f} x$ if and only if $x_k \to x$ and $f(x_k) \to f(x)$.

**Definition 2.2.** We say that a point $x$ is a critical point of $f$ if $0 \in \partial f(x)$. The lazy slope of $f$ at a point $x$ is defined as

$$|\partial f(x)| := \inf\{\|z\| : z \in \partial f(x)\} = \mathrm{dist}(0, \partial f(x)).$$

**Proposition 2.1.** [7] *Let $f, g \in \Gamma(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$ be given.*

(i) *We have $\widehat{\partial} f(x) \subset \partial f(x)$. Moreover, $\widehat{\partial} f(x)$ is convex and $\partial f(x)$ is closed (not necessarily convex). If $f$ is convex, then both sets are reduced to the subdifferential in the sense of convex analysis.*

(ii) *If the sequences $(x_k)$ an $(y_k)$ are such that $x_k \xrightarrow{f} x$, $y_k \to y$, and $y_k \in \partial f(x^k)$ for all $k$, then $y \in \partial f(x)$.*

(iii) *The Fermat's rule remains true: if $x$ is a local minimizer of $f$, then $x$ is a critical point (or stationary point) of $f$, that is, $0 \in \partial f(x)$.*

(iv) *If $g$ is continuously differentiable, then $\partial(f + g)(x) = \partial f(x) + \nabla g(x)$.*

(v) *We have that $\partial f$ is closed in the sense that if $\{(x_k, y_k)\}$ is a sequence in the graph of $\partial f$, $G(\partial f) := \{(z, w) : z \in \mathrm{dom}(\partial f),\ w \in \partial f(z)\}$, such that $x_k \xrightarrow{f} x$ and $y_k \to y$, it follows that $(x, y) \in G(\partial f)$.*

(vi) *If $x_k \xrightarrow{f} x$ and $\liminf_{k \to \infty} |\partial f(x_k)| = 0$, then $x$ is a critical point of $f$.*

### 2.2. **Kurdyka-Łojasiewicz Property.**

The Kurdyka-Łojasiewicz property [8, 9] plays a central part in the nonconvex optimization theory.

**Definition 2.3.** [2] We say that a function $f \in \Gamma(\mathbb{R}^d)$ satisfies the Kurdyka-Łojasiewicz property (KŁP) at $x^* \in \mathrm{dom}(\partial f)$ if there exist $\eta \in (0, \infty]$, a neighborhood $U$ of $x^*$, and a continuous concave function $\varphi : [0, \eta) \to \mathbb{R}^+$ such that

(i) $\varphi(0) = 0$,
(ii) $\varphi \in C^1(0, \eta)$,
(iii) $\varphi'(t) > 0$ for all $t \in (0, \eta)$,
(v) there holds the Kurdyka-Łojasiewicz inequality:

$$\varphi'(f(x) - f(x^*))|\partial f(x)| \geq 1 \tag{2.2}$$

for all $x \in U \cap \{x : f(x^*) < f(x) < f(x^*) + \eta\}$.

We say that $f \in \Gamma(\mathbb{R}^d)$ is a KŁ-function provided it satisfies KŁP at each point $x^* \in \mathrm{dom}(\partial f)$.

The KŁ inequality (2.2) (at a single point) can, in some circumstances, be extended to a compact set, as shown below.

**Lemma 2.1.** [5, Lemma 6] (Uniformized KŁ property) *Let $f \in \Gamma(\mathbb{R}^d)$ and let $\Omega \subset \mathbb{R}^d$ be a nonempty compact set. Assume that $f$ is constant on $\Omega$ and satisfies the KŁ property at each point of $\Omega$. Then there exist $\varepsilon > 0$ and $\eta > 0$, and $\varphi$ satisfying properties (i)-(iii) of Definition 2.3 such that for all $\bar{u} \in \Omega$ and all $u \in \mathbb{R}^d$ with the property:*

$$u \in \{u \in \mathbb{R}^d : \mathrm{dist}(u, \Omega) < \varepsilon\} \cap \{f(\bar{u}) < f(u) < f(\bar{u}) + \eta\}, \tag{2.3}$$

*the following uniformized KŁ inequality holds:*

$$\varphi'(f(u) - f(\bar{u}))|\partial f(u)| \geq 1. \tag{2.4}$$

More discussions on can be found in [1, 2, 7, 4]

### 2.3. **Proximal Mappings.**

**Definition 2.4.** Let $f \in \Gamma(\mathbb{R}^d)$ and let $\lambda > 0$. The proximal mapping of $f$ (of index $\lambda$) is defined as

$$\mathrm{prox}_{\lambda f}(x) := \arg\min \left\{ f(y) + \frac{1}{2\lambda}\|y - x\|^2 : y \in \mathbb{R}^d \right\}, \quad x \in \mathbb{R}^d. \tag{2.5}$$

Note that if $f$ is, in addition, convex, then $\mathrm{prox}_{\lambda f}$ is single-valued and well defined over the entire space $\mathbb{R}^d$. However, in the general nonconvex case, $\mathrm{prox}_{\lambda f}$ is set-valued and may be defined on a subset of $\mathbb{R}^d$ (more details can be found [12]).

The following inequality (2.7) is widely used in optimization theory (see [11]). However, for the sake of completeness, we include a proof.

**Lemma 2.2.** *Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and its gradient $\nabla f$ is $L$-Lipschitz continuous for some constant $L \geq 0$:*

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\| \quad \textit{for all } x, y \in \mathbb{R}. \tag{2.6}$$

*Then we have*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2 \quad \textit{for all } x, y \in \mathbb{R}. \tag{2.7}$$

*Proof.* Let $x, y \in \mathbb{R}$ and define a function $\varphi$ by

$$\varphi(t) := f(x + t(y - x)), \quad t \in \mathbb{R}.$$

Then $\varphi'(t) = \langle \nabla f(x + t(y - x)), y - x\rangle$. It turns out that

$$\begin{aligned} f(y) - f(x) &= \int_0^1 \varphi'(t)dt = \int_0^1 \langle \nabla f(x + t(y - x)), y - x\rangle dt \\ &= \langle \nabla f(x), y - x\rangle + \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x\rangle dt. \end{aligned}$$

Using the Lipschitz condition (2.6), we get

$$f(y) - f(x) \leq \langle \nabla f(x), y - x\rangle + L\|y - x\|^2 \int_0^1 t dt$$

and the desired inequality (2.7) follows immediately. □

**Lemma 2.3.** *Let $f, g \in \Gamma(\mathbb{R}^d)$ and set $F = f + g$. Let $\lambda > 0$ be given. Assume that the gradient $\nabla f$ of $f$ is $L$-Lipschitz continuous. Then, for any $u \in \mathbb{R}^d$ and setting*

$$\hat{u} := \mathrm{prox}_{\lambda g}(u - \lambda \nabla f(u)),$$

*we have*

$$F(\hat{u}) \leq F(u) - \frac{1}{2}\left(\frac{1}{\lambda} - L\right)\|\hat{u} - u\|^2. \tag{2.8}$$

*Proof.* We have

$$\begin{aligned}\hat{u} &= \arg\min_z g(z) + \frac{1}{2\lambda}\|z - u + \lambda \nabla f(u)\|^2 \\ &= \arg\min_z g(z) + \frac{1}{2\lambda}\|z - u\|^2 + \langle z - u, \nabla f(u)\rangle.\end{aligned}$$

Since $\hat{u}$ is a minimizer of the function

$$z \mapsto \psi(z) := g(z) + \frac{1}{2\lambda}\|z - u\|^2 + \langle z - u, \nabla f(u)\rangle \tag{2.9}$$

it turns out that

$$g(\hat{u}) + \frac{1}{2\lambda}\|\hat{u} - u\|^2 + \langle \hat{u} - u, \nabla f(u)\rangle \leq g(u). \tag{2.10}$$

On the other hand, since $\nabla f$ is $L$-Lipschitz, we can use Lemma 2.2 to get the inequality:

$$f(\hat{u}) \leq f(u) + \langle \nabla f(u), \hat{u} - u\rangle + \frac{L}{2}\|\hat{u} - u\|^2. \tag{2.11}$$

Adding up (2.10) and (2.11) immediately yields (2.8). □

2.4. **Convergence Lemma.** We also need the following basic result regarding convergence of nonnegative series.

**Lemma 2.4.** *Let $\{a_n\}$ be a sequence of real nonnegative numbers such that*

$$a_{k+1} \leq \gamma a_k + b_k, \quad k \geq 0, \tag{2.12}$$

*where $\gamma \in [0, 1)$ and $b_k \geq 0$ such that $\sum_{k=0}^{\infty} b_k < \infty$. Then $\sum_{k=0}^{\infty} a_k < \infty$.*

## 3. Main Results

Consider the following composite optimization problem

$$\min_{x \in \mathbb{R}^d} F(x) := f(x) + g(x), \tag{3.13}$$

where $f, g \in \Gamma(\mathbb{R}^d)$.

The following accelerated proximal gradient (APG) algorithm is introduced by Li, et al [10, Algorithm 3].

**Algorithm 1** (Algorithm APG nonconvex problem)

Input: $y_1 = x_0$, $\beta_k = \frac{k}{k+3}$, $\lambda < \frac{1}{L}$
**for** $k = 1, 2, \cdots$ **do**
  $x_k = \mathrm{prox}_{\lambda g}(y_k - \lambda \nabla f(y_k))$.
  $v_k = x_k + \beta_k(x_k - x_{k-1})$.
  if $F(x_k) \leq F(v_k)$, then $y_{k+1} = x_k$,
  else if $F(x_k) \geq F(v_k)$, then $y_{k+1} = v_k$.
  end if
**end for**

The following result regarding Algorithm 1 is proved in [10].

**Theorem 3.1.** [10, Theorem 1] *Let the following assumptions be satisfied:*

(A1) $f, g \in \Gamma_0(\mathbb{R}^d)$, $\inf_{x\in\mathbb{R}^d} F(x) > -\infty$, *and for each* $\alpha \in \mathbb{R}$, *the sublevel set* $\{x \in \mathbb{R}^d : F(x) \leq \alpha\}$ *is bounded;*

(A2) $f$ *has a continuous gradient* $\nabla f$ *that is* $L$*-Lipschitz continuous, i.e.,*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad x, y \in \mathbb{R}^d.$$

*Let* $\{x_k\}$ *be a generated by Algorithm 1 with stepsize* $\lambda < \frac{1}{L}$. *Then*

(i) $\{x_k\}$ *is a bounded sequence;*

(ii) *The set* $\Omega$ *of limit points of* $\{x_k\}$ *forms a compact set, on which the objective function* $F$ *is constant;*

(iii) *All elements of* $\Omega$ *are critical points of* $F$.

Below we slightly improve the above algorithm by allowing the stepsizes to depend on the steps, that is, we take $\lambda := \lambda_k$, where $k$ is the number of iterations. We also readjust the parameter $\beta_k$ (in our convergence proof we actually only require $\beta_k \leq 1 - \beta$ for some $\beta \in (0, 1)$).

**Algorithm 2** (Algorithm APG nonconvex problem with variable stepsizes)

Input: $y_1 = x_0$, $\beta_k = \frac{1}{k+1}$, $\lambda_k < \frac{1}{L}$
**for** $k = 1, 2, \cdots$ **do**
  $x_k = \mathrm{prox}_{\lambda_k g}(y_k - \lambda_k \nabla f(y_k))$.
  $v_k = x_k + \beta_k(x_k - x_{k-1})$.
  if $F(x_k) \leq F(v_k)$, then $y_{k+1} = x_k$,
  else if $F(x_k) > F(v_k)$, then $y_{k+1} = v_k$.
  end if
**end for**

The main results in this paper show that the conclusions of Theorem 3.1 remain true for variable stepsizes, and moreover, convergence of the trajectories is guaranteed if, in addition, the composite function $F$ satisfies the Kurdyka-Łojasiewicz property.

**Theorem 3.2.** *Consider a sequence* $\{x_k\}$ *generated by Algorithm 2. Assume the conditions (A1) and (A2) of Theorem 3.1 hold, and in addition, the stepsize sequence* $\{\lambda_k\}$ *satisfies the property:* $0 < a \leq \lambda_k \leq b < \dfrac{1}{L}$ *for all* $k$. *Then the following conclusions hold.*

(i) $\{x_k\}$ *is a bounded sequence;*

(ii) *The set* $\Omega$ *of limit points of* $\{x_k\}$ *forms a compact set, on which the objective function* $F$ *is constant;*

(iii) *All elements of* $\Omega$ *are critical points of* $F$.

*Moreover, if, in addition,* $F = f + g$ *is a KL function, then* $\{x_k\}$ *converges to a critical point of* $F$ *with finite length, that is,*

$$\sum_{k=0}^{\infty} \|x_{k+1} - x_k\| < \infty. \tag{3.14}$$

*Proof.* Apply (2.8) to the case where $\lambda := \lambda_k$ and $u := y_k$ to find that

$$F(x_k) \leq F(y_k) - \frac{1}{2}\left(\frac{1}{\lambda_k} - L\right)\|x_k - y_k\|^2. \tag{3.15}$$

In particular, $F(x_k) \leq F(y_k)$ for $\lambda_k < \frac{1}{L}$.

It is a straightforward observation from the definition of Algorithm 2 that $F(y_{k+1}) \leq F(x_k)$ which together with (3.15) immediately results in that

$$F(y_{k+1}) \leq F(x_k) \leq F(y_k) \leq F(x_{k-1}) \leq \cdots \leq F(y_1) \leq F(x_0). \tag{3.16}$$

Consequently, $\lim_{k\to\infty} F(x_k) = \lim_{k\to\infty} F(y_k)$ exists. Moreover, by (A1), we know that $\{x_k\}$ and $\{y_k\}$ are bounded. Rewrite (3.15) as

$$\frac{1}{2}\left(\frac{1}{\lambda_k} - L\right)\|x_k - y_k\|^2 \leq F(y_k) - F(x_k) \leq F(x_{k-1}) - F(x_k).$$

Since $\lambda_k \leq b < \frac{1}{L}$ for all $k$, this implies that

$$\frac{1}{2}\left(\frac{1}{b} - L\right)\sum_{k=1}^{\infty}\|x_k - y_k\|^2 \leq F(x_0) - \lim_{k\to\infty} F(x_k) < \infty.$$

In particular,

$$\lim_{k\to\infty}\|x_k - y_k\| = 0. \tag{3.17}$$

Now let $\Omega$ be the set of cluster points of $\{x_k\}$, that is,

$$\Omega \equiv \omega(\{x_k\}) := \{x \in \mathbb{R}^d : \exists\, x_{k_i} \to x\}.$$

The boundedness of $\{x_k\}$ ensures that $\Omega \neq \emptyset$, and due to (3.17), we also have $\Omega = \omega(\{y_k\})$.

Now let $\bar{x} \in \Omega$ and let $\{x_{k_i}\}$ be a subsequence of $\{x_k\}$ such that $x_{k_i} \to \bar{x}$. By definition of the algorithm, we have

$$x_k = \arg\min_{z\in\mathbb{R}^d} g(z) + \frac{1}{2\lambda}\|z - (y_k - \lambda\nabla f(y_k))\|^2. \tag{3.18}$$

By the optimality condition, we obtain

$$0 \in \partial g(x_k) + \frac{1}{\lambda_k}(x_k - y_k) + \nabla f(y_k).$$

Equivalently,

$$\frac{1}{\lambda_k}(y_k - x_k) - \nabla f(y_k) \in \partial g(x_k). \tag{3.19}$$

Applying (3.19) to the subsequence $\{k_i\}$ we get

$$\frac{1}{\lambda_{k_i}}(y_{k_i} - x_{k_i}) - \nabla f(y_{k_i}) \in \partial g(x_{k_i}). \tag{3.20}$$

With no loss of generality (up to a further convergent subsequence of $\{\lambda_{k_i}\}$ if necessary), we may assume $\lambda_{k_i} \to \bar{\lambda} \in [a, b]$.

Now since $x_{k_i} \to \bar{x}$, $y_{k_i} \to \bar{x}$, and $\frac{1}{\lambda_{k_i}}(y_{k_i} - x_{k_i}) \to 0$, we may take the limit in (3.20) as $i \to \infty$ and by the closedness of the subdifferential $\partial g$ of $g$ to obtain

$$-\nabla f(\bar{x}) \in \partial g(\bar{x}).$$

This is rewritten as $0 \in \partial F(\bar{x}) = \nabla f(\bar{x}) + \partial g(\bar{x})$. Hence, $\bar{x}$ is a critical point of $F$.

We finally verify that $F$ is constant on $\Omega$; it suffices to show that

$$F(\bar{x}) = \lim_{k\to\infty} F(x_k). \tag{3.21}$$

Here $\bar{x} \in \Omega$ and $x_{k_i} \to \bar{x}$ as above. Since $F(\bar{x}) = f(\bar{x}) + g(\bar{x})$ and since $f$ is continuous, all we need to prove is that

$$\lim_{i\to\infty} g(x_{k_i}) = g(\bar{x}).$$

On the one hand, from (3.18) we immediately deduce that

$$g(x_{k_i}) \le g(\bar{x}) + \frac{1}{2\lambda_{k_i}}(\|\bar{x} - y_{k_i} + \lambda_{k_i}\nabla f(y_{k_i})\|^2 - \|x_{k_i} - y_{k_i} + \lambda_{k_i}\nabla f(y_{k_i})\|^2)$$

$$= g(\bar{x}) + \frac{1}{2\lambda_{k_i}}(\|\bar{x} - y_{k_i}\|^2 - \|x_{k_i} - y_{k_i}\|^2) + \langle \bar{x} - x_{k_i}, \nabla f(y_{k_i})\rangle. \tag{3.22}$$

Since $x_{k_i} \to \bar{x}$, $\|x_{k_i} - y_{k_i}\| \to 0$, and $\{\lambda_{k_i}\}$ is bounded away from 0 from below, it turns out from (3.22) that $\limsup_{i\to\infty} g(x_{k_i}) \le g(\bar{x})$.

On the other hand, the lower semicontinuity of $g$ implies that $g(\bar{x}) \le \liminf_{i\to\infty} g(x_{k_i})$. Consequently, we have verified that $\lim_{i\to\infty} g(x_{k_i}) = g(\bar{x})$ exists. Furthermore, since $f$ is continuous, we have $\lim_{k\to\infty} F(x_k) = \lim_{i\to\infty} F(x_{k_i}) = \lim_{i\to\infty}(f(x_{k_i}) + g(x_{k_i})) = f(\bar{x}) + g(\bar{x}) = F(\bar{x})$. This proves (3.21).

Finally we prove (3.14) under the additional condition that $F$ satisfies the KL property. Observe that the conclusions in part (ii) guarantee that

$$\lim_{k\to\infty} \operatorname{dist}(x_k, \Omega) = 0. \tag{3.23}$$

As previously, assume $x_{k_i} \to \bar{x}$; then we have proved that $\bar{x}$ is a critical point of $F$. We may assume $x_k \ne y_k$ (since, if $x_k = y_k$ for some $k$, $x_k$ is a critical point of $F$ and the iteration process is terminated); hence $F(x_k) < F(y_k)$, and furthermore, $F(x_{k+1}) < F(x_k)$ by (3.16). Recall that we have $F(\bar{x}) = \lim_{k\to\infty} F(x_k)$.

By (3.23) we can apply Lemma 2.1 to get

$$\varphi'(F(x_k) - F(\bar{x}))|\partial F(x_k)| \ge 1 \tag{3.24}$$

for all $k \ge k_0$. Here $k_0$ is big enough so that $\operatorname{dist}(x_k, \Omega) < \varepsilon$ for all $k \ge k_0$. Before further proceeding, we notice the following two facts:

- Fact 1: $F(x_k) \le F(y_k) - c_1\|x_k - y_k\|^2 \le F(x_{k-1}) - c_1\|x_k - y_k\|^2$, where $c_1 = \frac{1}{2}\left(\frac{1}{b} - L\right) > 0$. This follows from (3.15) and the fact that $\lambda_k \le b$.
- Fact 2: $\|w_k\| \le c_2\|x_k - y_k\|$, where $w_k = \nabla f(x_k) - \nabla f(y_k) + \frac{1}{\lambda_k}(y_k - x_k) \in \partial F(x_k)$, and $c_2 = L + \frac{1}{a}$. This is due to (3.19) and the facts that $\|\nabla f(x_k) - \nabla f(y_k)\| \le L\|x_k - y_k\|$ and $\lambda_k \ge a$.

Applying (3.24) and Fact 1, we derive that, for $k \ge k_0$,

$$\varphi'(F(x_k) - F(\bar{x})) \ge \frac{1}{\|w_k\|} \ge \frac{1}{c_2}\frac{1}{\|x_k - y_k\|}. \tag{3.25}$$

Since $\varphi$ is concave, we have the inequality:

$$\varphi(x) - \varphi(y) \ge \varphi'(x)(x - y), \quad x, y \in \mathbb{R}.$$

It follows that

$$\varphi(F(x_k) - F(\bar{x})) - \varphi(F(x_{k+1}) - F(\bar{x})) \ge \varphi'(F(x_k) - F(\bar{x}))(F(x_k) - F(x_{k+1}))$$
$$\ge \varphi'(F(x_k) - F(\bar{x}))c_1\|x_{k+1} - y_{k+1}\|^2.$$

This, combining with (3.25), yields

$$\varphi(F(x_k) - F(\bar{x})) - \varphi(F(x_{k+1}) - F(\bar{x})) \ge \frac{c_1}{c_2}\frac{\|x_{k+1} - y_{k+1}\|^2}{\|x_k - y_k\|}.$$

In other words,

$$\frac{\|x_{k+1} - y_{k+1}\|^2}{\|x_k - y_k\|} \le \frac{c_2}{c_1}[\varphi(F(x_k) - F(\bar{x})) - \varphi(F(x_{k+1}) - F(\bar{x}))]. \tag{3.26}$$

Fix $\gamma \in (0,1)$. It is then not hard to get from (3.26)

$$\|x_{k+1}-y_{k+1}\| \leq \gamma\|x_k-y_k\| + \frac{1}{\gamma}\frac{c_2}{c_1}[\varphi(F(x_k)-F(\bar{x})) - \varphi(F(x_{k+1})-F(\bar{x})]. \tag{3.27}$$

for all $k \geq 1$. By Lemma 2.4, (3.27) guarantees that

$$\sum_{k=1}^{\infty}\|x_k-y_k\| < \infty.$$

Note that $y_{k+1}$ is either $x_k$ if $F(x_k) \leq F(v_k)$ or $v_k = x_k+\beta_k(x_k-x_{k-1})$ if $F(x_k) > F(v_k)$. In the latter case, we have $\|x_k - x_{k+1}\| \leq \|y_{k+1}-x_{k+1}\| + \beta_k\|x_k-x_{k-1}\|$ and we have estimates on the partial sums:

$$\sum_{i=1}^{k}\|x_i-x_{i+1}\| \leq \sum_{i=1}^{k}\|y_{i+1}-x_{i+1}\| + \sum_{i=1}^{k}\beta_i\|x_i-x_{i-1}\|.$$

It turns out that

$$\sum_{i=1}^{k-1}(1-\beta_{i+1})\|x_i-x_{i+1}\| \leq \sum_{i=1}^{k}\|y_{i+1}-x_{i+1}\| + \beta_1\|x_1-x_0\|.$$

Since $\beta_{i+1} = \dfrac{1}{i+2}$, $1-\beta_{i+1} = \dfrac{i+1}{i+2} \geq \dfrac{2}{3}$ for $i \geq 1$. Consequently, we derive from the last inequality that

$$\sum_{i=1}^{\infty}\|x_i-x_{i+1}\| \leq \frac{3}{2}\left(\sum_{i=1}^{\infty}\|y_{i+1}-x_{i+1}\| + \|x_1-x_0\|\right) < \infty$$

and (3.14) is proved. □

**Acknowledgement**. We were grateful to the reviewers for their helpful suggestions and comments which improved the presentation of this manuscript.

## REFERENCES

[1] Attouch, H., Bolte, J., Redont, P. and Soubeyran, A., *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality*, Math. Operations Research, **35** (2010), No. 2, 438–457

[2] Attouch, H., Bolte, J. and 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** (2013), 91–129

[3] Beck A. and Teboulle, M., *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., **2** (2009), No. 1, 183–202

[4] Bolte, J., Daniilidis, A. and Lewis, A., *The Łojasiewicz inequality for nonsmooth sub-analytic functions with applications to subgradient dynamical systems*, SIAM J. Optim., **17** (2007), 1205–1223

[5] Bolte, J., Sabach, S. and Teboulle, M., *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Math. Program., Ser. A , **146** (2014), 459–494

[6] Combettes, P. L. and Wajs, R., *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., **4** (2005), No. 4, 1168–1200

[7] Frankel, P., Garrigos, G. and Peypouquet, J., *Splitting methods with variable Metric for Kurdyka-Łojasiewicz functions and general convergence rates*, J. Optim. Theory Appl., **65** (2015), No. 3, 874–900

[8] Kurdyka, K., *On gradients of functions definable in o-minimal structures*, Ann. Inst. Fourier (Grenoble), **48** (1998), No. 3, 769–783

[9] Łojasiewicz, S., Une propriete topologique des sous-ensembles analytiques reels, in: *Les Euations aux Derivees Partielles*, Editions du centre National de la Recherche Scientifique, Paris, 1963, pp. 87–89

[10] Li, Q., Zhou, Y., Liang, Y. and Varshney, P. K., *Convergence analysis of proximal gradient with momentum for nonconvex optimization*, in Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017

[11] Nesterov, Y. E., *Introductory Lectures on Convex Optimization: a basic course* Kluwer Academic Publishers, Massachusetts, 2004
[12] Rockafellar, R. T. and Wets, R. J.-B., *Variational Analysis*, Grundlehren der Mathematischen Wissenschaften, vol. **317**, Springer, Berlin, 1998

DEPARTMENT OF MATHEMATICS
SCHOOL OF SCIENCE
HANGZHOU DIANZI UNIVERSITY
HANGZHOU 310018 CHINA
*E-mail address*: 2805065050@qq.com
*E-mail address*: xuhk@hdu.edu.cn