The Smoothing Proximal Gradient Algorithm with Extrapolation for the Relaxation of \( L_0 \) Regularization Problem

Jie Zhang\textsuperscript{a} and Xinmin Yang\textsuperscript{b} and Gaoxi Li\textsuperscript{c} and Ke Zhang\textsuperscript{d}

\textsuperscript{a} College of Mathematics, Sichuan University, Chengdu 610065, China
\textsuperscript{b} School of Mathematics Science, Chongqing Normal University, Chongqing 401331, China
\textsuperscript{c} School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing, China
\textsuperscript{d} School of Mathematics and Computer Science, Chongqing College of International Business and Economics, Chongqing, China

\textbf{ARTICLE HISTORY}
Compiled December 3, 2021

\textbf{ABSTRACT}
In this paper, we consider the exact continuous relaxation model of \( l_0 \) regularization problem which was given by Bian and Chen (SIAM J. Numer. Anal 58(1): 858-883, 2020) and propose a smoothing proximal gradient algorithm with extrapolation (SPGE) for this kind of problem. We show that any accumulation point of the sequence generated by SPGE algorithm is a lifted stationary point of the relaxation model for a fairly general choice of extrapolation parameter. Moreover, it is shown that the whole sequence generated by SPGE algorithm converges to a lifted stationary point of the relaxation model. The convergence rate of \( o\left(\frac{1}{k^{1-\sigma}}\right)\) with respect to squared proximal residual is established. Finally, we conduct three important numerical experiments to illustrate the faster efficiency of the SPGE algorithm compared with the smoothing proximal gradient (SPG) algorithm proposed by Bian and Chen (SIAM J. Numer. Anal 58(1):858-883, 2020).

\textbf{KEYWORDS}
Smoothing approximation and Proximal gradient method and Extrapolation and \( l_0 \) regularization problem

\section{1. Introduction}

In this paper, we concentrate on the exact continuous relaxation model of \( l_0 \) regularization problem which was proposed in \cite{5}, i.e.,

\[
\min_{x \in \mathcal{X}} \mathcal{F}(x) := f(x) + \lambda \Phi(x),
\]

where \( \mathcal{X} = \{ x \in \mathbb{R}^n : l \leq x \leq u \} \), \( l, u \in \{ 0, \pm \infty \}^n \) with \( l \leq 0 \leq u \) and \( l < u \), \( \Phi(x) = \sum_{i=1}^{n} \phi(x_i) \), \( f : \mathbb{R}^n \rightarrow [0, \infty) \) is convex, not necessarily smooth and has Lipschitz continuous constant \( L_f \) on \( \mathcal{X} \). \( \lambda \) is a positive parameter keeping a trade-off between data fidelity and sparsity. The problem (1) is an exact continuous relaxation of the \( l_0 \) regularization problem and the \( l_0 \) regularization problem is specifically expressed as

\[
\min_{x \in \mathcal{X}} \mathcal{F}_{l_0}(x) := f(x) + \lambda \| x \|_0,
\]

CONTACT G.X.Li. Author. Email: ligaoxicn@126.com
where \( \| \cdot \|_0 \) denotes the so called \( l_0 \) norm. For a vector \( x \in \mathbb{R}^n \), \( \| x \|_0 = |\mathcal{A}(x)| \), where \( \mathcal{A}(x) = \{ i \in \{1, 2, 3, \ldots, n \} : x_i \neq 0 \} \) and \( \| x \|_0 \) is the cardinality of the vector \( x \). We say that \( x \in \mathbb{R}^n \) is sparse if \( |\mathcal{A}(x)| \ll n \). Given a parameter \( v > 0 \), the capped-\( l_1 \) function \( \phi \) is

\[
\phi(t) = \min \{1, |t|/v \}.
\]

Moreover, \( \phi \) in (3) is a DC function which can be represented in the following form,

\[
\phi(t) = \frac{|t|}{v} - \max \{ \theta_1(t), \theta_2(t), \theta_3(t) \},
\]

with \( \theta_1(t) = 0, \theta_2(t) = t/v - 1, \theta_3(t) = -t/v - 1 \) and \( \mathcal{D}(t) = \{ i = \{1, 2, 3 \} : \theta_i(t) = \max\{\theta_1(t), \theta_2(t), \theta_3(t)\} \} \).

In [5], Bian and Chen proposed SPG algorithm which combines the smoothing method and proximal gradient algorithm to solve the exact continuous relaxation model of the \( l_0 \) regularization problem. It is well known that the proximal gradient method is always an usual and efficient method to deal with the following problem

\[
\min_{x \in \mathbb{R}^n} \mathcal{F}(x) := f(x) + g(x),
\]

where \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is a smooth and possibly nonconvex function, \( g : \mathbb{R}^n \rightarrow (-\infty, +\infty] \) is a proper closed, nonsmooth and possibly nonconvex function. This kind of problem includes many optimization models deriving from various applications such as machine learning [12, 16], signal and image processing [14], compressed sensing [13] and so on. The proximal mapping [27] of \( g \) is often assumed to be easily computed for all \( t > 0, u \in \mathbb{R}^n \) and is defined as follows

\[
\text{prox}_{tg}(u) := \text{argmin} \left\{ g(x) + \frac{1}{2t} \| x - u \|^2 \right\}.
\]

Since the low cost of the proximal gradient method, a great many of researchers devote to it and aim to improve its efficiency. One common and efficient technique is to perform extrapolation, where momentum terms involving the previous iterations are added to the current iteration. One of the accelerated versions of the proximal gradient method was introduced by Ochs et al. [24]. They proposed to incorporate the inertial force which is also called extrapolation item into proximal step and studied the inertial proximal gradient algorithm for nonconvex optimization (iPiano). It can be seen as a nonsmooth split version of the Heavy-ball method from Polyak [28], the update system of the algorithm is

\[
y^k = x^k + \beta_k (x^k - x^{k-1}), \\
x^{k+1} = \text{prox}_{t_k g}(y_k - t_k \nabla f(x_k)),
\]

where \( \beta_k, t_k \) are the extrapolation coefficient and stepsize respectively satisfying some rules and \( \text{prox} \) is the proximal operator. The convergence rate \( O(1/k) \) with respect to squared proximal residual is obtained and global convergence is exhibited based on the Kurdyka-Lojasiewicz (K-L) property. Another acceleration method adds the inertial force not only to proximal step but also to gradient step which was raised by Nesterov [20, 22], the general framework is as follows:

\[
y^k = x^k + \beta_k (x^k - x^{k-1}), \\
x^{k+1} = \text{prox}_{t_k g}(y_k - t_k \nabla f(y_k)),
\]
where $\beta_k \in [0,1]$ is the extrapolation coefficient, $t_k$ is the stepsize depending on the Lipschitz constant of $\nabla f(x)$. Based on the Nesterov’s idea, Beck and Teboulle [1] developed a specific recurrence relation and exhibited the remarkable Fast Iterative Shrinkage Thresholding Algorithm (FISTA) for the convex case and the convergence rate $O(\frac{1}{k^2})$ of the objective function is obtained. Besides, some variants of FISTA are established by choosing appropriate extrapolation parameters, we refer the reader to [1, 7, 17] and their references therein. We also observe that Wen et al. [31] established the proximal gradient algorithm with extrapolation (PGe) for the case with gradient Lipschitz continuously differentiable but nonconvex loss function. The results about the iterative sequence and objective function sequence converging R-linearly to a stationary point of the problem are obtained under the error bound condition [18].

From the literatures mentioned above, all of the efficient accelerated algorithms require a restrictive assumption that $f$ is a smooth function. Actually, the function $f$ in problem (1) is always nonsmooth in many practical applications, such as the $l_1$ loss function

$$f(x) = \frac{1}{m}||Ax - b||_1,$$

and the censored regression problem

$$f(x) = \frac{1}{pm} \sum_{i=1}^{m} \{|\max\{A_i x - c_i, 0\} - b_i|^p\},$$

where $p \in [1, 2]$, $A_i^T \in R^m$ and $c_i, b_i \in R, i = 1, 2, \cdots, m$. Few researchers have focused on the accelerated algorithm when $f$ is nonsmooth, for this, the nature consideration is to smooth it by virtue of the smoothing method which is a well recognized technique for the nonsmooth optimization problem. One can solve it by an appropriate smooth approximation and solve a sequence of smooth problems gradually approaching to the original problem. Smooth approximation method for optimization problems involves many kinds of problems, one can refer to the literatures [6, 8–10, 15, 23, 29, 30, 33] and their references therein. It needs to mention that Nesterov [19] proposed an accelerated gradient algorithm based on a smoothing technique for problem (5) with $g(x) = 0$ and improved the efficiency estimate from $O(\frac{1}{\epsilon^2})$ to $O(\frac{1}{\epsilon})$. We also observe that [2] proposed a unifying framework that combines smoothing approximation based on the extension of the Moreau-infimal convolution technique with fast first order algorithm and proved its efficiency estimate.

Recently, we observe that Wu et al. [32] put forward a smoothing fast iterative hard thresholding (SFIHT) algorithm solving $l_0$ regularized nonsmooth convex problem (2) which together the smoothing approximations, extrapolation techniques with iterative hard thresholding methods. The extrapolation coefficient can be achieved $\sup_k \beta_k = 1$ in the SFIHT algorithm. They discussed the convergence of the SFIHT algorithm and obtained the convergence rate $O(\frac{1}{k^2})$ of the objective function.

To the best of our knowledge, there no scholars have studied the accelerated algorithm for the relaxation model (1) of the $l_0$ regularized problem when $f$ is nonsmooth. Benefitted from acceleration technique and smoothing methods mentioned above, we propose a SPGE algorithm for the relaxation model (1) of (2). Since the extrapolation item is added, the convergence can’t be easily got directly. But we can discuss its convergence by constructing the appropriate auxiliary sequence. We mainly analyze the convergence of the SPGE algorithm when the extrapolation term satisfying $\beta_k \in [0, \sqrt{(1-a\mu_k+\mu_k)}]$, with $0 < a < 1$ which is a fairly general choice of extrapolation parameters, in particular, the parameters can be chosen as in s-FISTA proposed by Bian [4] with fixed restart [25]. Though the range of the extrapolation term slightly smaller compare with [32], the better theoretical results can be obtained. The result that the sequence
\( \{ \| x^k - x^{k-1} \| \} \) is summable and the whole sequence \( \{ x^k \} \) is convergent. Further, the convergence rate with respect to proximal residual is obtained.

The paper is organised as follows. In Section 2, we present some brief and useful preliminaries derived from [5]. In Section 3, we propose the SPGE algorithm and analysis its the convergence behaviour. It is proved that any accumulation point of a subsequence generated by SPGE algorithm is a lifted stationary point of the continuous relaxation model. Further, the whole sequence \( \{ x^k \} \) generated by SPGE algorithm converges to a lifted stationary point of the relaxation model (1). Besides, the convergence rate as regards the squared proximal residual is obtained. In Section 4, we show that the efficiency of the proposed SPGE algorithm. Finally, we give some conclusions in Section 5.

Notations. Given a vector \( x \in R^n \), \( \| x \|_2 \) is replaced by \( \| x \| \). Denote \( N = \{0, 1, 2, \cdots \} \), \( D_n = \{ d \in R^n : d_i \in \{1, 2, 3 \} , i = 1, 2, \cdots , n \} \). For a nonempty, closed, convex set \( X \subseteq R^n \), \( N_X(x) \) represents the normal cone to \( X \) at \( x \in X \). \( 1_n \in R^n \) be the all ones vector and \( e_i \in R^n \) be the \( i \) column of the \( n \)-dimensional identity matrix. Denote \( \partial f(x) \) be a Clarke subgradient of \( f \) at \( x \in R^n \) where \( f(x) : R^n \rightarrow R \) is a locally Lipschitz continuous function.

2. Preliminaries and main results

In this section, we first briefly recall the definition which will be useful for the analysis in this paper.

Definition 2.1. (lifted stationary point [5, 26]) We say that \( x \in X \) is a lifted stationary point of (1) if there exist \( d_i \in D_i(x_i) \) for \( i = 1, 2, \cdots , n \) such that
\[
\lambda \sum_{i=1}^{n} \theta'_d(x_i) e_i \in \partial f(x) + \frac{\lambda}{v} \partial (\sum_{i=1}^{n} |x_i|) + N_X(x).
\] (6)

We make the following assumption in the convergence analysis.

Assumption 1. The parameter \( v \) in (3) satisfies \( 0 < v < \bar{v} := \frac{\lambda}{L_f} \).

Assumption 2. \( \mathcal{F} \) in (1)(or \( \mathcal{F}_{l_0} \) in (2)) is level bounded on \( X \).

Next we present the following lemmas which characterise the lifted stationary point of (1) under the Assumption 1,2 and show the relationship between the lifted stationary point, local minizers and global minizers of (1),(2).

Lemma 2.2. ([5]) If \( \mathcal{F} \) is a lifted stationary point of (1), then the vector \( d_\mathcal{F} = (d_1^\mathcal{F}, \cdots , d_n^\mathcal{F}) \in \prod_{i=1}^{n} D(x_i) \) satisfying (6) is unique. In particular, for \( i = 1, 2, \cdots , n \),
\[
d_i^\mathcal{F} = \begin{cases} 
1 & \text{if } |x_i| < v, \\
2 & \text{if } x_i \geq v, \\
3 & \text{if } x_i \leq -v.
\end{cases}
\] (7)

Lemma 2.3. ([5]) If \( \mathcal{F} \) is a lifted stationary point of (1), then it has the lower bound property, i.e., if \( x_i \in (v, v) \), then \( x_i = 0, \forall i = 1, \cdots , n \).

Lemma 2.4. ([5]) If \( \mathcal{F} \) is a lifted stationary point of (1), then it is a local minimizer of (2) and the objective functions have the same value at \( \mathcal{F} \).
Lemma 2.5. \((\ref{5})\) \(\overline{x}\) is a lifted stationary point of \((1)\) and satisfies \(|x_i| > v\) when \(x_i \neq 0, \forall i\) if and only if it is a local minimizer of \((1)\).

Lemma 2.6. \((\ref{5})\) \(\overline{x}\) is a global minimizer of \((1)\) if and only if it is a global minimizer of \((2)\) and the objective functions have the same value at \(\overline{x}\).

3. The smoothing proximal gradient algorithm with extrapolation and its convergence analysis

Combining the smoothing method and the proximal gradient method with extrapolation, we propose the SPGE algorithm and discuss its convergence in this section. For convenience, we first exhibit the definition of smoothing function deriving from \([5]\).

Definition 3.1. We call \(\tilde{f} : \mathbb{R}^n \times [0, \mu] \to \mathbb{R}\) with \(\mu > 0\) a smoothing function of the function \(f\) in \((1)\) if \(\tilde{f}(\cdot, \mu)\) is continuous differentiable in \(\mathbb{R}^n\) for any fixed \(\mu > 0\) and satisfies the following conditions:

\begin{enumerate}
  \item \(\lim_{z \to x, \mu \downarrow 0} \tilde{f}(z, \mu) = f(x), \quad \forall x \in \mathcal{X};\)
  \item \(\tilde{f}(x, \mu)\) is convex with respect to \(x\) in \(\mathcal{X}\) and for any fixed \(\mu > 0;\)
  \item \(\{\lim_{z \to x, \mu \downarrow 0} \nabla \tilde{f}(z, \mu)\} \subseteq \partial f(x), \quad \forall x \in \mathcal{X};\)
  \item there exists a positive constant \(\kappa\) such that
  \[|\tilde{f}(x, \mu_2) - \tilde{f}(x, \mu_1)| \leq \kappa|\mu_1 - \mu_2|, \quad \forall x \in \mathcal{X}, \mu_1, \mu_2 \in [0, \overline{\mu}];\]
  \item especially, \(|\tilde{f}(x, \mu) - f(x)| \leq \kappa \mu, \quad \forall x \in \mathcal{X}, \mu \in (0, \overline{\mu}];\)
\end{enumerate}

Remark 1. The definition in \((v)\) says that \(\nabla_x \tilde{f}(\cdot, \mu)\) is Lipschitz continuous on \(3\mathcal{X}\) since the extrapolation point \(y^k = x^k + \beta_k(x^k - x^{k-1})\) in the following algorithm may not belong to \(\mathcal{X}\) and we need to restrict it on \(3\mathcal{X}\).

For a given \(d = (d_1, d_2, \cdots, d_n)^T \in \mathbb{D}^n\), let
\[
\Phi^d(x) := \sum_{i=1}^{n} \left( \frac{|x_i|}{v} - \theta_{d_i}(x_i) \right),
\]
then
\[
\Phi(x) = \min_{d \in \mathbb{D}^n} \Phi^d(x), \quad \forall x \in \mathcal{X}.
\]

And especially for a fixed \(\overline{x} \in \mathcal{X}\), \(\Phi(\overline{x}) = \Phi^{d\overline{x}}(\overline{x})\), where \(d\overline{x}\) is defined in \((7)\).

\[
\overline{F}^d(x, \mu) := \tilde{f}(x, \mu) + \lambda \Phi^d(x), \quad \overline{F}(x, \mu) := \tilde{f}(x, \mu) + \lambda \Phi(x),
\]
where \(\tilde{f}\) is a smoothing function of \(f\). For any fixed \(\mu > 0\), and \(d \in \mathbb{D}^n\), \(\overline{F}^d(x, \mu)\) is convex function.
By (9), it holds that
\[
\bar{F}^d(x, \mu) \geq \bar{F}(x, \mu) \quad \forall d \in \mathbb{D}^n, x \in \mathcal{X}, \mu \in (0, \bar{\mu}].
\] (10)

In the following, we focus on the constrained convex optimization problem with a given smoothing parameter \(\mu > 0\), \(d \in \mathbb{D}^n\):
\[
\bar{F}^d(x, \mu) := \bar{f}(x, \mu) + \lambda \Phi^d(x).
\] (11)

Let \(d \in \mathbb{D}^n, w \in \mathbb{R}^n \) and \(\tau > 0\). Due to the fact that penalty item of (11) is piecewise linear, the proximal operator of \(\tau \Phi^d\) on \(\mathcal{X}\)
\[
\text{prox}_{\tau \Phi^d}(w) = \arg\min_{x} \left\{ \tau \Phi^d(x) + \frac{1}{2} \|x - w\|^2 \right\}
\]
has a closed form solution \(\hat{x} = (\hat{x}_1, \cdots, \hat{x}_n)\) with \(\hat{x}_i = \min\{\max\{l_i, h_i\}, u_i\}, \forall i = 1, 2, \cdots, n\), where
\[
h_i = \begin{cases} 
0 & \text{if } |\tilde{w}_i| \leq \tau/v, \\
\tilde{w}_i - \tau/v, & \text{if } \tilde{w}_i > \tau/v, \\
\tilde{w}_i + \tau/v, & \text{if } \tilde{w}_i < -\tau/v,
\end{cases}
\] (12)

where
\[
\tilde{w}_i = \begin{cases} 
w_i, & \text{if } d_i = 1, \\
w_i + \tau/v, & \text{if } d_i = 2, \\
w_i - \tau/v, & \text{if } d_i = 3.
\end{cases}
\]

Based on the proximal gradient method, we consider the approximation \(Q_{d^k}(x, y^k, \mu_k)\) of \(\bar{F}^d(\cdot, \mu_k)\) and the given point \(y^k\) is involved two previous iterations, specifically as follows:
\[
\begin{cases} 
y^k = x^k + \beta_{k-1}(x^k - x^{k-1}), \\
Q_{d^k}(x, y^k, \mu_k) = f(y^k, \mu_k) + (x - y^k, \nabla f(y^k, \mu_k)) + \frac{1}{2} L \mu_k^{-1} \|x - y^k\|^2 + \lambda \Phi^{d^k}(x),
\end{cases}
\]

where \(\beta_{k-1}\) is an extrapolation parameter, \(L\) is any constant satisfying \(L \geq \tilde{L}\). It is observed that the problem \(\min_x Q_{d^k}(x, y^k, \mu_k)\) has a unique minimizer \(\hat{x}\) since \(Q_{d^k}(x, y^k, \mu_k)\) is a strongly convex function with respect to \(x\) for any fixed \(d^k, \mu_k, y^k\), where \(d^k = dx^k, \mu_k\) is the smoothing parameter. The minimizer can be obtained by (12) with \(\tau = \lambda L^{-1} \mu_k\) and \(w^k = y^k - L^{-1} \mu_k \nabla f(y^k, \mu_k)\).

**Remark 2.** It is need to be emphasized that the following convergence analysis is discussed under the extrapolation coefficient \(\beta_k\) condition which is
\[
0 \leq \beta_k^2 \leq (1 - a \frac{\mu_{k+1}}{\mu_k}) \frac{\mu_{k+1}}{\mu_k}, \quad \text{with } 0 < a \ll 1.
\] (16)

**Remark 3.** The sequence \(\{H(x^{k+1}, x^k, \mu_k)\}\) in the SPGE algorithm framework is specifically defined as in (18).
Algorithm 1 SPGE algorithm

1: **Data:** $\sigma \in (0,1)$, $\alpha > 0$,
   **Initialization:** $x^{-1} = x^0 \in \mathcal{X}$, $\mu_{-1} = \mu_0 \in (0,\bar{\mu})$, $L > \bar{L}$. Set $k = 0$.
2: **while** a termination criterion is not met **do**
3: **step 1:** Choose $\beta_{k-1} \in [0, \sqrt{\frac{\mu_k}{\mu_{k-1}}})$, and let $d^k = d^{x^k}$, where $d^{x^k}$ is defined in (7).
4: **step 2:** Compute
   \[
   \begin{aligned}
   y^k &= x^k + \beta_{k-1}(x^k - x^{k-1}), \\
   x^{k+1} &= \text{argmin}_{x \in \mathcal{X}} Q_{d^k}(x, y^k, \mu_k);
   \end{aligned}
   \] (13)
5: **step 3:** If
   \[
   H(x^{k+1}, x^k, \mu_k) + \kappa \mu_k - (H(x^k, x^{k-1}, \mu_{k-1}) + \kappa \mu_{k-1}) \leq -\alpha \mu_k^2,
   \] (14)
   set $\mu_{k+1} = \mu_k$, otherwise, set
   \[
   \mu_{k+1} = \frac{\mu_0}{(k+1)^\sigma},
   \] (15)
6: set $k := k + 1$, and return to Step 1.
7: **end while

Remark 4. From the above algorithm framework, it can be found that the subproblem (13) boils down to the SPG algorithm when $\beta_k = 0$, therefore, the SPG algorithm is a special case of the SPGE algorithm. Besides, when $\beta_k$ is chosen as (16), that is $\sup \beta_k \leq \sqrt{1 - a} < 1$ with $0 < a \ll 1$, it is more general and contains one special extrapolation coefficient deriving from the smoothing fast iterative shrinkage thresholding algorithm (s-FISTA) which was proposed in [4] with fixed restart or adaptive restart system [25] for solving subproblem (13). That is, one reset the initial point $t_{-1} = 1$, $\mu_{-1} = \mu_0$ after fixed iterations or restart the algorithm when $\langle y^{k-1} - x^k, x^k - x^{k-1} \rangle > 0$.

The extrapolation parameters in experiments of Section 4 is chosen as in the way of the fixed restart. The extrapolation coefficient in s-FISTA is defined as follows, for any $k \geq 0$

\[
t_k = \frac{1 + \sqrt{1 + 4 \frac{\mu_{k-1}^2}{\mu_k^2} + 1}}{2}, \quad \beta_{k-1} = \frac{t_{k-1} - 1}{t_k},
\] (17)

where $t_{-1} = 1$.

In what follows we analyze the basic convergence behaviors of the iterative sequence generated by the SPGE algorithm. Though it gets much more difficult to prove directly, added the extrapolation, we can develop the convergence analysis by constructing the following auxiliary sequence. Suppose $\{x^k\}, \{y^k\}, \{\mu_k\}$ are the sequences generated by SPGE algorithm, define

\[
H(x^{k+1}, x^k, \mu_k) := \bar{F}(x^{k+1}, \mu_k) + \tau_k \|x^{k+1} - x^k\|^2,
\] (18)

where $\tau_k = \frac{L \mu_{k-1}}{4} + \frac{L \beta_k^2 \mu_{k-1}^2}{4} + \frac{1}{4}$. It needs to be emphasised that the convergence analysis of the SPGE algorithm is heavily based on the auxiliary sequence.
Lemma 3.2. (well defined of the SPGE algorithm) The proposed SPGE algorithm is well defined, and the sequences \( \{ x^k \}, \{ \mu_k \} \) generated by the SPGE have the following properties.

(i) \( \{ x^k \} \subseteq X \).

(ii) there are infinite elements in \( \mathcal{N}^* \) and \( \lim_{k \to +\infty} \mu_k = 0 \), where \( \mathcal{N}^* = \{ k \in \mathbb{N} : \mu_{k+1} \neq \mu_k \} \).

**Proof.** The proof is similar to Lemma 3.2 in [5], the results can be easily obtained. \( \square \)

Now we show that the auxiliary sequence \( \{ H(x^{k+1}, x^k, \mu_k) + \kappa \mu_k \} \) is nonincreasing when \( \beta_k \in [0, \sqrt{\frac{\mu_{k+1}}{\mu_k}}) \).

**Lemma 3.3.** For any \( k \in \mathbb{N} \), if \( \beta_k \in [0, \sqrt{\frac{\mu_{k+1}}{\mu_k}}) \), then

\[
H(x^{k+1}, x^k, \mu_k) + \kappa \mu_k - (H(x^k, x^{k-1}, \mu_{k-1}) + \kappa \mu_{k-1}) \\
\leq (\tau_k - \frac{1}{2} L \mu_k^{-1}) \| x^{k+1} - x^k \|^2 - (\tau_{k-1} - \frac{1}{2} L \mu_{k-1}^{-1} \beta_{k-1}^2) \| x^k - x^{k-1} \|^2, \tag{19}
\]

and the sequence \( \{ H(x^{k+1}, x^k, \mu_k) + \kappa \mu_k \} \) is a nonincreasing sequence.

**Proof.** From the strong convexity of \( Q_{d^k}(x, y^k, \mu_k) \) with the strong convex coefficient \( L \mu_k^{-1} \), we have

\[
\lambda \Phi^{d^k}(x^{k+1}) \leq \lambda \Phi^{d^k}(x) + \langle x - x^{k+1}, \nabla \tilde{f}(y^k, \mu_k) \rangle + \frac{1}{2} L \mu_k^{-1} \| x - y^k \|^2 \\
- \frac{1}{2} L \mu_k^{-1} \| x^{k+1} - y^k \|^2 - \frac{1}{2} L \mu_k^{-1} \| x - x^{k+1} \|^2, \tag{20}
\]

Since \( \nabla_x \tilde{f}(\cdot, \mu) \) has Lipschitz constant \( \tilde{L} \mu_k^{-1} \) from the Definition 3.1(v), it follows that

\[
\tilde{f}(x^{k+1}, \mu_k) \leq \tilde{f}(y^k, \mu_k) + \langle x^{k+1} - y^k, \nabla \tilde{f}(y^k, \mu_k) \rangle + \frac{1}{2} L \mu_k^{-1} \| x^{k+1} - y^k \|^2. \tag{21}
\]

Combining (20) with (21), we have

\[
\tilde{F}^{d^k}(x^{k+1}, \mu_k) \leq \lambda \Phi^{d^k}(x) + \tilde{f}(y^k, \mu_k) + \langle x - y^k, \nabla \tilde{f}(y^k, \mu_k) \rangle + \frac{1}{2} L \mu_k^{-1} \| x - y^k \|^2 \\
- \frac{1}{2} L \mu_k^{-1} \| x - x^{k+1} \|^2.
\]

Using the convexity of \( \tilde{f}(x, \mu_k) \) with respect to \( x \), the above inequality can be rewritten as

\[
\tilde{F}^{d^k}(x^{k+1}, \mu_k) \leq \tilde{F}^{d^k}(x, \mu_k) + \frac{1}{2} L \mu_k^{-1} \| x - y^k \|^2 - \frac{1}{2} L \mu_k^{-1} \| x - x^{k+1} \|^2, \tag{22}
\]

Setting \( x = x^k, d^k = d^k \), according to \( \tilde{F}^{d^k}(x^{k+1}, \mu_k) \geq \tilde{F}(x^{k+1}, \mu_k) \), we have

\[
\tilde{F}(x^{k+1}, \mu_k) \leq \tilde{F}^{d^k}(x^k, \mu_k) + \frac{1}{2} L \mu_k^{-1} \| x^k - y^k \|^2 - \frac{1}{2} L \mu_k^{-1} \| x^k - x^{k+1} \|^2. \tag{23}
\]
Together the Definition 3.1(iv) with the above inequality, we deduce

\[
\tilde{F}(x^{k+1}, \mu_k) + \kappa \mu_k \leq \tilde{F}(x^k, \mu_{k-1}) + \kappa \mu_{k-1} + \frac{1}{2} L \mu_k^{-1} \|x^k - y^k\|^2 - \frac{1}{2} L \mu_k^{-1} \|x^k - x^{k+1}\|^2,
\]

Then it has that

\[
\begin{align*}
\tilde{F}(x^{k+1}, \mu_k) + \kappa \mu_k + \tau_k \|x^k - x^{k+1}\|^2 & \leq \tilde{F}(x^k, \mu_{k-1}) + \kappa \mu_{k-1} + \tau_{k-1} \|x^{k-1} - x^{k}\|^2 \\
& \quad + (\tau_k - \frac{1}{2} L \mu_k^{-1}) \|x^k - x^{k+1}\|^2 - (\tau_{k-1} - \frac{1}{2} \beta_{k-1} L \mu_k^{-1}) \|x^k - x^{k-1}\|^2.
\end{align*}
\]

By the definition of \(\tau_k\) from (18) and \(\beta_k \in [0, \sqrt{\frac{\mu_{k+1}}{\mu_k}}]\), we deduce that

\[
\tau_k - \frac{1}{2} L \mu_k^{-1} < 0, \quad \tau_{k-1} - \frac{1}{2} L \mu_k^{-1} \beta_{k-1}^2 > 0,
\]

and then the sequence \(\{H(x^{k+1}, x^k, \mu_k) + \kappa \mu_k\}\) is a nonincreasing sequence. This completes the proof.

\[
\square
\]

**Corollary 3.4.** For any \(k \in \mathbb{N}\), \(\beta_k \in [0, \sqrt{\frac{\mu_{k+1}}{\mu_k}}]\), \(\beta_k\) is chosen as (16), then there exists \(\zeta \in \mathbb{R}\) satisfying

\[
\zeta := \lim_{k \to +\infty} H(x^{k+1}, x^k, \mu_k) + \kappa \mu_k,
\]

and

\[
\lim_{k \to +\infty} H(x^{k+1}, x^k, \mu_k) + \kappa \mu_k = \lim_{k \to +\infty} H(x^{k+1}, x^k, \mu_k) = \lim_{k \to +\infty} \tilde{F}(x^k, \mu_{k-1}) = \lim_{k \to +\infty} F(x^k) = \zeta. \tag{24}
\]

Moreover, for any \(k \in \mathbb{N}\), the sequence \(\{x^k\}\) is bounded.

**Proof.** From (18), Definition 3.1(iv) and Lemma 2.6, we have

\[
H(x^{k+1}, x^k, \mu_k) + \kappa \mu_k \geq \tilde{F}(x^{k+1}, \mu_k) + \kappa \mu_k \geq F(x^{k+1}) \geq \min_{x \in X} F(x) = \min_{x \in X} F_{l_0}(x).
\]

Together the lower bounded of the sequence \(\{H(x^{k+1}, x^k, \mu_k) + \kappa \mu_k\}\) with its nonincreasing, then it is convergent and there exists \(\zeta \in \mathbb{R}\) satisfying

\[
\zeta := \lim_{k \to +\infty} H(x^{k+1}, x^k, \mu_k) + \kappa \mu_k.
\]

By virtue of \(\lim_{k \to +\infty} \mu_k = 0\) and Definition 3.1(i), we have

\[
\lim_{k \to +\infty} H(x^{k+1}, x^k, \mu_k) + \kappa \mu_k = \lim_{k \to +\infty} H(x^{k+1}, x^k, \mu_k) = \zeta. \tag{25}
\]

9
Next, we are ready to prove \( \lim_{k \to +\infty} \tilde{F}(x^k, \mu_{k-1}) = \zeta \). From (19), we deduce

\[
|\tilde{F}(x^k, \mu_{k-1}) - \zeta| \\
= |H(x^k, x^{k-1}, \mu_{k-1}) - \zeta - \tau_{k-1} \|x^k - x^{k-1}\|^2| \\
\leq |H(x^k, x^{k-1}, \mu_{k-1}) - \zeta| + \frac{\tau_{k-1}}{\tau_{k-1} - \frac{1}{2}L \mu_k \beta_k^{2 - 1}} (H(x^k, x^{k-1}, \mu_{k-1}) \\
+ \kappa \mu_{k-1} - (H(x^{k+1}, x^k, \mu_k) + \kappa \mu_k)) \\
\leq |H(x^k, x^{k-1}, \mu_{k-1}) - \zeta| + \frac{1 + \beta_k^{2 - 1} \mu_k}{1 - \beta_k^{2 - 1} \mu_k} \mu_{k-1} \|x^k - x^{k-1}\|^2 (H(x^k, x^{k-1}, \mu_{k-1}) - \zeta) \\
\leq |H(x^k, x^{k-1}, \mu_{k-1}) - \zeta| + \frac{2 - a}{a} (H(x^k, x^{k-1}, \mu_{k-1}) - \zeta),
\]

where the first inequality uses the triangle inequality and lemma 3.3, the second one derives from the nonincreasing of \( \{H(x^k, x^{k+1}, \mu_k) + \kappa \mu_k\} \) and \( H(x^k, x^{k-1}, \mu_{k-1}) + \kappa \mu_{k-1} \geq \zeta \), the last one is due to \( \beta_k^{2 - 1} \mu_k \leq 1 - a \mu_k \leq 1 - a \) and \( \frac{1}{1 - \beta_k^{2 - 1} \mu_k} \leq \frac{1}{a} \). When \( k \) is sufficiently large, the right of the inequality (26) tends to 0, thus

\[
\lim_{k \to +\infty} \tilde{F}(x^k, \mu_{k-1}) = \zeta.
\]

and \( \lim_{k \to +\infty} \tilde{F}(x^k) = \lim_{k \to +\infty} \tilde{F}(x^k, \mu_{k-1}) = \zeta \) from Definition 3.1(i). Using the nonincreasing property of the sequence \( \{H(x^k, x^{k+1}, \mu_k) + \kappa \mu_k\} \), we have

\[
\mathcal{F}(x^{k+1}) \leq \tilde{F}(x^{k+1}, \mu_k) + \kappa \mu_k \leq H(x^{k+1}, x^k, \mu_k) + \kappa \mu_k \leq H(x^1, x^0, \mu_0) + \kappa \mu_0 < +\infty.
\]

Since \( \mathcal{F} \) is level bounded on \( \mathcal{X} \) according to the Assumption 2, we easily have that the sequence \( \{x^k\} \) is bounded for any \( k \in \mathbb{N} \). \hfill \Box

**Corollary 3.5.** When \( \beta_k \) is chosen as (16), it holds that

\[
\sum_{k=0}^{+\infty} \|x^{k+1} - x^k\|^2 < +\infty,
\]

and

\[
\sum_{k=0}^{+\infty} \mu_k^{-1} \|x^{k+1} - x^k\|^2 < +\infty, \quad \sum_{k=0}^{+\infty} \mu_{k+1}^{-1} \|x^{k+1} - x^k\|^2 < +\infty.
\]

**Proof.** Summing the inequality (19) from \( k = 0 \) to \( k = K \), we get

\[
\sum_{k=0}^{K} (-\tau_k + \frac{1}{2}L \mu_k^{-1}) \|x^{k+1} - x^k\|^2 \\
\leq H(x^0, x^{-1}, \mu_1) - H(x^{K+1}, x^K, \mu_K) + \kappa (\mu_1 - \mu_K),
\]

10
that is
\[
\sum_{k=0}^{K} \frac{L}{4} (\mu_k^{-1} - \beta_k^2 \mu_{k+1}^{-1}) \|x^{k+1} - x^k\|^2
\]
\[
= \sum_{k=0}^{K} \frac{L}{4} (\mu_{k+1}^{-1} - \beta_k^2) \mu_{k+1}^{-1} \|x^{k+1} - x^k\|^2
\]
\[
\leq H(x^0, x^{-1}, \mu_{-1}) - H(x^{K+1}, x^K, \mu_K) + \kappa (\mu_{-1} - \mu_K).
\]
Since \(\beta_k^2 \mu_k \mu_{k+1} \leq 1 - a \mu_k \mu_{k+1}\), \(\mu_k\) is nonincreasing and \(\mu_k \in (0, 1)\) for any \(k \in \mathbb{N}\), it holds that
\[
\beta_k^2 \frac{\mu_k}{\mu_{k+1}} \leq 1 - a,
\]
thus,
\[
\mu_k^{-1} - \beta_k^2 \mu_k^{-1} \geq a, \quad \mu_{k+1}^{-1} - \beta_k \geq a.
\]
When \(K\) is large enough, since \(\mu_k\) is nonincreasing, it holds that
\[
\sum_{k=0}^{+\infty} \|x^{k+1} - x^k\|^2 \leq H(x^0, x^{-1}, \mu_{-1}) + \kappa \mu_{-1} - \zeta < +\infty,
\]
\[
\sum_{k=0}^{+\infty} \mu_k^{-1} \|x^{k+1} - x^k\|^2 \leq \sum_{k=0}^{+\infty} \mu_k^{-1} \|x^{k+1} - x^k\|^2
\]
\[
\leq H(x^0, x^{-1}, \mu_{-1}) + \kappa \mu_{-1} - \zeta < +\infty.
\]
The proof is completed.

Next, the following theorem is given to illustrate the convergence of the subsequence.

**Theorem 3.6.** Suppose \(\beta_k\) is chosen as (16), any accumulation point of \(\{x^k : k \in \mathbb{N}^s\}\) is a lifted stationary point of (1).

**Proof.** According to Corollary 3.4, it follows that \(\{x^k : k \in \mathbb{N}^s\}\) is bounded, so it has an accumulation point which might as well denote as \(\bar{x}\), i.e., there exists a subsequence \(\{x^{k_i}\}_{k_i \in \mathbb{N}^s}\) of \(\{x^k\}_{k \in \mathbb{N}^s}\) such that \(x^{k_i} \to \bar{x}\) as \(i \to +\infty\). For any \(k \in \mathbb{N}^s\), the inequality (14) does not hold. Together this with (19), we get
\[
(-\tau_{k_i} + \frac{1}{2} L \mu^{-1}_{k_i}) \|x^{k_{i+1}} - x^{k_i}\|^2 \leq \alpha \mu^{-2}_{k_i},
\]
that is,
\[
\frac{L}{4} (1 - \beta_{k_i}^2 \frac{\mu_{k_i}}{\mu_{k_i+1}}) \mu_{k_i}^{-1} \|x^{k_{i+1}} - x^{k_i}\|^2 \leq \alpha \mu^{-2}_{k_i},
\]
\[
\frac{L}{4} (\mu_{k_i}^{-1} \mu_{k_i+1} - \beta_{k_i}^2) \mu_{k_i+1}^{-1} \|x^{k_{i+1}} - x^{k_i}\|^2 \leq \alpha \mu^{-2}_{k_i}.
\]
Multiplying $\mu_k^{-1}$ and $\mu_k^{-1}$ on both sides of the above inequalities respectively, it holds that

$$
\begin{align*}
\frac{L}{4} (1 - \beta_k^2 \frac{\mu_k}{\mu_{k+1}}) \mu_k^{-2} \|x_k^{i+1} - x_k^i\|^2 \leq \alpha \mu_k, \quad (27) \\
\frac{L}{4} (\mu_k^{-1} \mu_{k+1} - \beta_k^2 \mu_k^{-2}) \|x_k^{i+1} - x_k^i\|^2 \leq \alpha \mu_k \mu_{k+1}^{-1}. \quad (28)
\end{align*}
$$

Since $\beta_k^2 \frac{\mu_k}{\mu_{k+1}} \leq 1 - a \frac{\mu_k}{\mu_{k+1}}$, $\mu_k$ is nonincreasing and $\mu_k \in (0, 1)$, it holds that $\beta_k^2 \frac{\mu_k}{\mu_{k+1}} \leq 1 - a \leq 1 - a \mu_k$. Then

$$
1 - \beta_k^2 \frac{\mu_k}{\mu_{k+1}} \geq a, \quad \mu_k^{-1} \mu_{k+1} - \beta_k^2 \geq a.
$$

Moreover, $\lim_{i \to +\infty} \mu_k = 0$. Taking limits on (27), (28), the right hand of them are tends to 0, then we get

$$
\lim_{i \to +\infty} \mu_k^{-1} \|x_k^{i+1} - x_k^i\| = 0, \quad \lim_{i \to +\infty} \mu_k^{-1} \|x_k^{i+1} - x_k^i\| = 0,
$$

and then

$$
\lim_{i \to +\infty} \mu_k^{-1} (x_k^{i+1} - y_k^i) = \lim_{i \to +\infty} \mu_k^{-1} ((x_k^{i+1} - x_k^i) - \beta_k (x_k^i - x_k^{i-1})) = 0. \quad (29)
$$

By the first order necessary optimality condition of (13), we have

$$
\langle \nabla \tilde{f}(x_k^i, \mu_k) + L \mu_k^{-1} (x_k^{i+1} - y_k^i) + \lambda \xi_k^i, x - x_k^{i+1} \rangle \geq 0, \quad (30)
$$

where $\xi_k^i \in \partial \Phi d^k_i (x_k^{i+1}), x \in \mathcal{X}$. Due to $\lim_{i \to +\infty} x_k^i = \bar{x}, \lim_{i \to +\infty} \|x_k^{i+1} - x_k^i\| = 0$, then $\lim_{i \to +\infty} x_k^{i+1} = \bar{x}$. Since the finiteness of the set $\{d_k^i : i \in \mathbb{N}\}$, there is a subsequence $\{k_{i_j}\}$ of $\{k_i\}$ and $\bar{d} \in \mathcal{D}(\bar{x})$ such that $d^{k_j} = \bar{d}, \forall j \in \mathbb{N}$. According to the upper semicontinuity of $\partial \Phi \bar{d}$ and $\lim_{j \to +\infty} x^{k_j} = \lim_{j \to +\infty} y^{k_j} = \bar{x}$, we obtain

$$
\left\{ \lim_{j \to +\infty} \xi^{k_j} : \xi^{k_j} \in \partial \Phi d^{k_j} (x^{k_j+1}) \right\} \subseteq \partial \Phi \bar{d}(\bar{x}). \quad (31)
$$

Combining (29), (30), (31) with the Definition 3.1(iii), letting $k_i = k_{i_j}$ in (30) and $j \to +\infty$, it deduces that there exist $\bar{\psi} \in \partial f(\bar{x})$ and $\zeta^{\bar{d}} \in \partial \Phi \bar{d}(\bar{x})$ satisfying the following

$$
\langle \bar{\psi} + \lambda \zeta^{\bar{d}}, x - \bar{x} \rangle \geq 0, \forall x \in \mathcal{X}. \quad (32)
$$

Thus, we have that $\bar{x}$ is a lifted stationary point of (1) by $\bar{d} \in \mathcal{D}(\bar{x})$, the definition of $\Phi \bar{d}$ and the convexity of $\mathcal{X}$. Then the proof is completed. \qed

**Theorem 3.7.** Suppose $\beta_k$ is chosen as in (16), it has that:

(i) for any $k \in \mathbb{N}$, $\mathcal{A}(x^{k+1}) \subseteq \mathcal{A}(x^k) \cup \mathcal{A}(x^{k-1})$,

where $\mathcal{A}(x) = \{i = \{1, 2, \cdots, n\} : x_i \neq 0\}$;

(ii) $\|\nabla \tilde{f}(y^k, \mu_{k-1})\| < \frac{1}{2} (\lambda/v + Lf)$ when $k$ is sufficiently large;

(iii) for any $k \in \mathbb{N}$, there exists $K \in \mathbb{N}$ such that for all $k \geq K$
Therefore, from (33) and (34), we have that
\[ \| x^{k+1} - x^k \| + \| x^k - x^{k-1} \| \geq C \mu_k, \]
where \( C = \min \left\{ \frac{\lambda}{v}, \frac{3}{4} L^{-1}(\lambda/v - L_f) \right\} \).

(b) \( \sum_k \| x^{k+1} - x^k \| < +\infty \), and we have \( x^k \to x^* \) as \( k \to +\infty \), where \( x^* \) is a lifted stationary point of (1).

\[ \| \nabla f(x^k) \| \leq \frac{1}{\mu} \| x^k - x^{k-1} \| \]

Proof. (i) We just only need to prove that if \( x_i^{k-1} = x_i^k = 0 \), we have \( x_i^{k+1} = 0 \). Suppose \( x_i^{k-1} = x_i^k = 0 \), then \( d_i^k = 1 \) and \( y_i^k = x_i^k + \beta_{k-1}(x_i^k - x_i^{k-1}) = 0 \). According to (12) and \( v < \frac{\lambda}{L_f} \), we have
\[ |y_i^k - L^{-1}\mu_k \nabla_i f(y_i^k, \mu_k)| \leq L^{-1}\mu_k \| \nabla f(y_i^k, \mu_k) \| < (\lambda L^{-1}\mu_k)/v. \]
Thus, \( x_i^{k+1} = 0 \) from (12), so the result is proved.
(ii) The verification is similar to the Lemma 3.6 in [5], it is easy to get the result.
(iii) (a) Letting \( u^k = y^k - L^{-1}\mu_k \nabla f(y^k, \mu_k) \), for any \( k \geq K, j \in \{1, 2, \cdots, n\} \), either \( |x_j^k| < v \) or \( |x_j^k| \geq v \).

First we discuss the index \( j \) with respect to \( 0 < |x_j^k| < v \) on different three cases.

Case 1. If \( |w_j^k| \leq \frac{\lambda}{v} L^{-1}\mu_k \), then we have \( x_j^{k+1} = 0 \), which means that \( d_j^{k+1} = 1 \) from (7). We also discuss it on three different cases.

Case (1): If \( |w_j^k| \leq \frac{\lambda}{v} L^{-1}\mu_{k+1} \), then \( x_j^{k+2} = 0 \). From \( A(x^{k+3}) \subseteq A(x^{k+2}) \cup A(x^{k+1}) \), we get that \( x_j^{k+3} \equiv 0 \) for sufficiently large \( k \).

Case (2): If \( w_j^{k+1} > \frac{\lambda}{v} L^{-1}\mu_{k+1} \), we have
\[
|x_j^{k+2}| = |x_j^{k+2} - x_j^{k+1}| \geq |x_j^{k+2} - y_j^{k+1}| - |y_j^{k+1} - x_j^{k+1}|
= |x_j^{k+2} - y_j^{k+1}| - \beta_k |x_j^k|
\geq \frac{\lambda}{v} L^{-1}\mu_{k+1} - |L^{-1}\mu_{k+1} \nabla_j f(y^{k+1}, \mu_{k+1})| - |x_j^k|
\geq \frac{1}{2} \left( \frac{\lambda}{v} - L_f \right) L^{-1}\mu_{k+1} - |x_j^k|,
\]
where the second inequality derives from (12), \( 0 \leq \beta_k < 1 \) and the third inequality derives from (ii) of this Lemma.

Case(3): If \( w_j^{k+1} < -\frac{\lambda}{v} L^{-1}\mu_{k+1} \), similar to the process of Case 1(2), it holds that
\[
|x_j^{k+2}| \geq \frac{1}{2} \left( \frac{\lambda}{v} - L_f \right) L^{-1}\mu_{k+1} - |x_j^k|,
\]
(33)

Therefore, from (33) and (34), we have that
\[
\| x^{k+2} - x^{k+1} \|^2 + \| x^{k+1} - x^k \|^2
\geq |x_j^{k+2} - 0|^2 + |x_j^k - 0|^2
\geq \frac{1}{8} \left( \frac{\lambda}{v} - L_f \right)^2 (L^{-1}\mu_{k+1})^2,
\]
where the second inequality derives from \( 2(a^2 + b^2) \geq (a + b)^2 \) for any \( a, b \geq 0 \). According to
2(a + b)^2 \geq a^2 + b^2 \text{ for any } a, b \geq 0 \text{ again, we have}

\|x^{k+2} - x^{k+1}\| + \|x^{k+1} - x^k\| \geq \frac{1}{4} (\frac{\lambda}{\nu} - L_f) L^{-1} \mu_{k+1}.

(35)

Case 2. If \( w_j^k > \frac{\lambda}{\nu} L^{-1} \mu_k \), similar to the process of Case 1(2), it holds that

\[ |x_j^{k+1} - x_j^k| \geq \frac{1}{2} (\frac{\lambda}{\nu} - L_f) L^{-1} \mu_k - |x_j^k - x_j^{k-1}|, \]

(36)

Case 3. If \( w_j^k < -\frac{\lambda}{\nu} L^{-1} \mu_k \), similar to the process of Case 1(3), we have

\[ |x_j^{k+1} - x_j^k| \geq \frac{1}{2} (\frac{\lambda}{\nu} - L_f) L^{-1} \mu_k - |x_j^k - x_j^{k-1}|. \]

(37)

Thus, by using \( 2(a + b)^2 \geq a^2 + b^2 \) for any \( a, b \geq 0 \), from (36) and (37), we get

\[ \|x^{k+1} - x^k\| + \|x^k - x^{k-1}\| \geq \frac{1}{4} (\frac{\lambda}{\nu} - L_f) L^{-1} \mu_k. \]

(38)

Next, we focus on the index \( j \) satisfying \( x_j^k = 0 \), which means that \( d_j^k = 1 \). Analogously, from the analysis of Case 1(2) and Case 1(3) of \( 0 < |x_j^k| < v \), we easily have \( x_j^{k+1} = 0 \) or \( \|x^{k+1} - x^k\| + \|x^k - x^{k-1}\| \geq \frac{1}{4} (\frac{\lambda}{\nu} - L_f) L^{-1} \mu_k. \)

At last, we discuss the index \( j \) with respect to \( |x_j^k| \geq v \), which means that \( d_j^k = 2 \) or 3. Due to the analysis about \( d_j^k = 2 \) or 3 is similar, we just discuss it when \( d_j^k = 2 \).

Case 1. If \( |w_j^k + \frac{\lambda}{\nu} L^{-1} \mu_k| \leq \frac{\lambda}{\nu} L^{-1} \mu_k \), then we have \( x_j^{k+1} = 0 \), which means that \( d_j^{k+1} = 1 \). We also discuss it on three different cases.

Case (1): If \( |w_j^{k+1}| \leq \frac{\lambda}{\nu} L^{-1} \mu_{k+1} \), then \( x_j^{k+2} = 0 \). From \( \mathcal{A}(x^{k+3}) \subseteq \mathcal{A}(x^{k+2}) \cup \mathcal{A}(x^{k+1}) \), we get that \( x_j^{k+3} = 0 \) for sufficiently large \( k \).

Case(2): If \( w_j^{k+1} > \frac{\lambda}{\nu} L^{-1} \mu_{k+1} \), similar to the analysis of Case 1(2) of \( 0 < |x_j^k| < v \), we have the result (33).

Case(3): If \( w_j^{k+1} < -\frac{\lambda}{\nu} L^{-1} \mu_{k+1} \), similar to the analysis of Case 1(3) of \( 0 < |x_j^k| < v \), we have the result (34).

Therefore, for case 1(2) and case 1(3) with respect to \( |x_j^k| \geq v \), we have (35).

Case 2. If \( w_j^k + \frac{\lambda}{\nu} L^{-1} \mu_k > \frac{\lambda}{\nu} L^{-1} \mu_k \),

\[ |x_j^{k+1} - x_j^k| \geq |x_j^{k+1} - y_j^k| - |y_j^k - x_j^k| \\
= |x_j^{k+1} - y_j^k| - \beta_{k-1} |x_j^k - x_j^{k-1}| \\
\geq |L^{-1} \mu_k \nabla \tilde{f}(y^k, \mu_k)| - |x_j^k - x_j^{k-1}| \]

it holds that

\[ |x_j^{k+1} - x_j^k| \geq L^{-1} \mu_k |\nabla_j \tilde{f}(y^k, \mu_k)| - |x_j^k - x_j^{k-1}| \]

Since \( \tilde{f}(y^k, \mu_k) \) is gradient lipschitz continuous on \( 3\mathcal{X} \) with respect to \( y^k \), there exist a constant
where the first inequality comes from the result (iii)(a) of this Lemma, the second one derives from $(a + b)^2 \leq 2(a^2 + b^2)$ for any $a, b \geq 0$, the third one due to the nonincreasing of $\mu_k$ and the last one comes from Corollary 3.5. From the above analysis, then we get

$$C \sum_{k \in M_2} \|x^{k+1} - x^k\| \leq C \sum_{k \in M_2} (\|x^{k+1} - x^k\| + \|x^k - x^{k-1}\|)$$

$$< +\infty.$$ (42)
that is,
\[ \sum_{k \in M_2} \| x^{k+1} - x^k \| < +\infty. \]
Similarly,
\[ \sum_{k \in M_3} \| x^{k+1} - x^k \| \geq +\infty. \]
Thus,
\[ \sum_k \| x^{k+1} - x^k \| \leq \sum_{k \in M_1} \| x^{k+1} - x^k \| + \sum_{k \in M_2} \| x^{k+1} - x^k \| + \sum_{k \in M_3} \| x^{k+1} - x^k \| < +\infty. \]
Moreover, since \( \{ x^k \} \) is bounded, then we have \( x^k \to x^* \) as \( k \to +\infty \), where \( x^* \) is a lifted stationary point of (1).

\[ \text{Remark 5.} \] In fact, the global convergence of the SPG algorithm in [5] can also be obtained by the idea of discussing each case according to the index of sequence value.

At the last part of the section, we aim at proving the local convergence rate of SPGE with respect to the proximal residual, where the proximal residual is defined as
\[ r(x) := x - \text{prox}_g(x - \nabla f(x)), \] (43)
where \( g \) is convex (possibly nonsmooth) function and \( f \) is a smooth (possibly nonconvex) function. Obviously, \( r(x) = 0 \) is equivalent to \( x = \text{prox}_g(x - \nabla f(x)) \) from [27], which characterize the first order optimality condition of \( \min_{x \in \mathbb{R}^n} f(x) + g(x) \). Thus, it can also be seen as a quality to measure the convergence of the sequence \( \{ x^k \} \) to a critical point. First, we present some preparatory work.

\[ \text{Lemma 3.8.} \] ([21, 24]) Suppose \( y, z \in \mathbb{R}^n, \alpha > 0 \). Define the functions
\[ p_g(\alpha) := \frac{1}{\alpha} \| \text{prox}_{\alpha g}(y - \alpha z) - y \| \]
and
\[ q_g(\alpha) := \| \text{prox}_{\alpha g}(y - \alpha z) - y \|. \]
Then \( p_g(\alpha) \) is a decreasing function of \( \alpha \), and \( q_g(\alpha) \) is increasing in \( \alpha \).

\[ \text{Theorem 3.9.} \] Suppose that \( \{ x^k \} \) is a sequence generated by SPGE algorithm and \( r(x^k) \) is a proximal residual defined by (43). Then for any \( k \in \mathbb{N} \) and \( \sigma \in (0,1) \), when \( \bar{\mu} < L \), it holds that
\[ \min_{0 \leq k \leq K} \| r(x^k) \|^2 \leq \frac{C_1}{(K + 1)^{1-\sigma}}, \]
where $C_1 = L^2C_0$ and $C_0 = 8(H(x^0, x^{-1}, \mu_{-1}) - \zeta)$. 

**Proof.** First, we reformulate $Q_{\phi^e}(x, y^k, \mu_k)$ in the form of proximal type, that is

$$\arg\min_{x \in \mathcal{X}} Q_{\phi^e}(x, y^k, \mu_k) = \prox_{L^{-1}\mu_k \lambda \Phi^e}(y^k - L^{-1}\mu_k \nabla \tilde{f}(y^k, \mu_k)).$$

From Lemma 3.8, we have that

$$r(x^k) = \|\prox_{\lambda \Phi^e}(x^k - \nabla \tilde{f}(x^k, \mu_k)) - x^k\|
\leq \frac{1}{\min\{1, L^{-1}\mu_k\}} \|\prox_{L^{-1}\mu_k \lambda \Phi^e}(x^k - L^{-1}\mu_k \nabla \tilde{f}(x^k, \mu_k)) - x^k\|,$$

Since $\mu_k \leq \bar{\mu} < L$, it has that

$$L^{-1}\mu_k r(x^k) \leq \|\prox_{\lambda \Phi^e L^{-1}\mu_k}(x^k - L^{-1}\mu_k \nabla \tilde{f}(x^k, \mu_k)) - x^k\|
\leq \|\prox_{\lambda \Phi^e L^{-1}\mu_k}(x^k - L^{-1}\mu_k \nabla \tilde{f}(x^k, \mu_k)) - \prox_{\lambda \Phi^e L^{-1}\mu_k}(y^k - L^{-1}\mu_k \nabla \tilde{f}(y^k, \mu_k))\|
+ \|\prox_{\lambda \Phi^e L^{-1}\mu_k}(y^k - L^{-1}\mu_k \nabla \tilde{f}(y^k, \mu_k)) - x^k\|
\leq \|x^k - y^k\| + L^{-1}\bar{L}\|x^k - y^k\| + \|x^{k+1} - x^k\|
\leq 2\|x^k - y^k\| + \|x^{k+1} - x^k\|,$$

where the third inequality derives from the nonexpansiveness of the proximal operator and the lipschitz continuous of the gradient $\nabla \tilde{f}(\cdot, \mu_k)$.

According to $\beta_{k-1} \in [0, 1)$, the above inequality holds as follows

$$L^{-1}\mu_k r(x^k) \leq 2\|x^k - x^{k-1}\| + \|x^{k+1} - x^k\|. \quad \text{(44)}$$

Squaring above inequality on both sides, using $2(a^2+b^2) \geq (a+b)^2$ for any $a, b \geq 0$ and multiplying $\mu_{k-1}^{-1}$ on both sides, it can be reformulated as the following

$$L^{-2}\mu_k r^2(x^k) \leq 8\mu_{k-1}^{-1}\|x^k - x^{k-1}\|^2 + 2\mu_{k-1}^{-1}\|x^{k+1} - x^k\|^2, \quad \text{(45)}$$

Summing up (45) from $k = 0$ to $K$, it yields

$$\sum_{k=0}^{K} L^{-2}\mu_k r^2(x^k) \leq \sum_{k=0}^{K} 8\mu_{k-1}^{-1}\|x^k - x^{k-1}\|^2 + \sum_{k=0}^{K} 2\mu_{k-1}^{-1}\|x^{k+1} - x^k\|^2,$$
Further, from the nonincreasing of $\mu_k$ and from Corollary 3.5, we have

\[
(L^{-2} \mu_K (K + 1)) \min_{0 \leq k \leq K} r^2(x^k) \\
\leq (L^{-2} (K + 1)) \min_{0 \leq k \leq K} \mu_k r^2(x^k) \\
\leq L^{-2} \sum_{k=0}^{K} \mu_k r^2(x^k) \\
\leq \sum_{k=0}^{K} 8 \mu_k^{-1} \|x^k - x^{k-1}\|^2 + \sum_{k=0}^{K} 2 \mu_k^{-1} \|x^{k+1} - x^k\|^2 \\
\leq C_0,
\]

where $C_0 = 8(H(x^0, x^{-1}, \mu_{-1}) - \zeta)$.

According to $\mu_k \geq \frac{\mu_0}{k^a}$, it has that

\[
\min_{0 \leq k \leq K} r^2(x^k) \leq \frac{C_1}{(K + 1)\mu_{K+1}} \leq \frac{C_1}{(K + 1)^{1-\sigma}},
\]

where $C_1 = L^2 C_0$, $0 < \sigma < 1$. This completes the proof.

**Remark 6.** From Theorem 3.6, it is observed that the convergence of a subsequence of $\mu_k^{-1} \|x^{k+1} - x^k\|$ is a sufficient optimality condition of the SPGE algorithm. Then from the inequality (44), that is

\[
L^{-1} r(x^k) < 2 \mu_k^{-1} \|x^k - x^{k-1}\| + \mu_k^{-1} \|x^{k+1} - x^k\|,
\]

it is concluded that the proximal residual $r(x^k)$ can be taken as a quality to measure the convergence of the SPGE algorithm.

### 4. Numerical Experiments

In this section, we perform numerical experiments on three different problems which are deriving from [5] to illustrate the efficiency of the SPGE algorithm and all the experiments are carried out on 1.80GHz Core i5 PC with 12GB of RAM. The stopping standard is set as

\[
\text{number of iterations} \leq \text{Maxiter} \quad \text{or} \quad \mu_k \leq \epsilon.
\]

In the following, we use "Iter" to present the number of iterations, "time" to present the CPU time in seconds, $\bar{x}$ and $\bar{r}$ to present the output of the SPG algorithm and SPGE algorithm respectively. The extrapolation item is $\beta_k \in [0, \sqrt{(1 - a \frac{\mu_{k+1}}{\mu_k}) \frac{\mu_{k+1}}{\mu_k}}]$ with $0 < a \ll 1$. The value of $a$ is chosen as $a = 0.0001$ in the following numerical experiment.

Before we show the experiments, we present the smoothing function for $l_1$ loss function and censored regression loss function in the following.

For the $l_1$ loss function,

\[
f(x) = \frac{1}{m} \|Ax - b\|_1,
\]

\[

\]

\[

\]
where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$. The smoothing function is defined as

$$\tilde{f}(x, \mu) = \frac{1}{m} \sum_{i=1}^{m} \tilde{\theta}(A_i x - b_i, \mu) \quad \text{with} \quad \tilde{\theta}(s, \mu) = \begin{cases} |s| & \text{if } |s| > \mu, \\ s^2 + \frac{\mu^2}{2} & \text{if } |s| \leq \mu. \end{cases}$$

(48)

For the nonsmooth convex censored regression loss function

$$f(x) = \frac{1}{pm} \sum_{i=1}^{m} \max \{A_i x - c_i, 0\} - b_i^p,$$

where $p \in [1, 2], A_i^T \in \mathbb{R}^n$ and $c_i, b_i \in \mathbb{R}$, $i = 1, \ldots, m$. When $p = 1$, the smoothing function is defined as

$$\tilde{f}(x, \mu) = \frac{1}{m} \sum_{i=1}^{m} \tilde{\phi}(A_i x, \mu) - b_i^\mu$$

with

$$\tilde{\phi}(s, \mu) = \begin{cases} \max \{s, 0\} & \text{if } |s| > \mu, \\ \frac{(s+\mu)^2}{4\mu} & \text{if } |s| \leq \mu. \end{cases}$$

(49)

Example 4.1. We first consider the following problem

$$\min_{0 \leq x_1, x_2 \leq 1} F_{t_0}(x_1, x_2) := |x_1 + x_2 - 1| + \lambda \|x\|_0.$$

(50)

Let $\mathcal{GM}$ denote the global minimizers of (50).

| $\lambda$ | $\mathcal{GM}$ | $v$ | $\bar{x}$ | $\tau$ | Iter | time |
|-----------|----------------|-----|----------|--------|------|------|
|          |                |     |          |        | SPG  | SPGE | SPG  | SPGE |
| 0.7/0.8/0.9 | (1, 0.1)     | 0.4/0.5/0.6 | (1, 0.1)/(1, 0.1) | (1, 0.1)/(1, 0.1) | 15/14/13 | 10/7/7 | 0.0336/0.0369/0.0341 | 0.0302/0.0319/0.0302 |
| 1/1/1     | (1, 0.1)/(0, 0) | 0.7/0.5/0.3 | (1, 0.1)/(1, 0.1) | (1, 0.1)/(1, 0.1) | 15/13/14 | 7/10/10 | 0.0314/0.0328/0.0353 | 0.0311/0.0316/0.0349 |
| 1.2/1.3/1.4 | (0, 0)       | 0.8/0.9/1 | (0, 0)/(0, 0) | (0, 0)/(0, 0) | 16/25/23 | 7/8/5 | 0.0299/0.0296/0.0296 | 0.0288/0.0283/0.0283 |

The parameters in SPG algorithm are set as default values and the parameters of SPGE algorithm are set as in the following:

$$L = \sqrt{2}, \sigma = 1, \alpha = 0.9, \text{Maxiter} = 10^4, \kappa = \frac{1}{2}, \epsilon = 10^{-3}, L_t = \sqrt{2}.$$
Example 4.2. Consider the linear regression problem, i.e.,

$$\min_{0 \leq x \leq 10} \mathcal{F}_b(x) := \frac{1}{m} \|Ax - b\|_1 + \lambda \|x\|_0,$$

(51)

where $b \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times n} (m \ll n)$.

The sensing matrix $A$, observation $b$ and the parameters for SPG algorithm are set as in same way in [5], one can refer to [5] in details. The parameters for SPGE algorithm are set as in the following:

$L = 2$, $\alpha = 1$, $\sigma = 0.9$, $\text{Maxiter} = 10^4$, $\kappa = \frac{1}{2}$, $\epsilon = 10^{-3}$.

Setting $\lambda = 18.8$, $L_f = \|A\|_\infty$ and $v = \min \left\{ \frac{\lambda}{L_f}, 10 \right\}$, $x_0 = 1.97 \ast \text{ones}(n, 1)$. The extrapolation term $\beta_k$ in the SPGE algorithm is chosen as in the way of (17) with the fixed restart systems and we perform a fixed restart every 500 iterations. We denote the SPGE algorithm as s-FISTA. At the initialize point, we set $\mu_{-1} = \mu_0 = \mu_1 = 50$ and $t_{-1} = 1$.

Figure 1. $m = 60$, $n = 120$ for problem (41)

Figure 2. $m = 80$, $n = 160$ for problem (41)
First, we present the effectiveness of s-FISTA with fixed restart schemes in Figure 1, 2, 3 with respect to $\|x^k - x^*\|$ and $|F(x^k) - F(x^*)|$ against the number of iterations, where $x^*$ is the true solution of the problem (51). From the part of (a) (b), it is observed that s-FISTA with fixed restart use fewer iterations as a faster rate than SPG algorithm. The Figure 4 exhibit the convergence of $\mu_k$ and present that the output of SPGE algorithm has the lower bound property and is very close to original signal $x^*$ with higher probability compared with SPG algorithm.

Finally, we illustrate the efficiency of the SPGE algorithm from the time, relative error and successful rate aspects. The relative error is defined as in [5] and the successful rate is defined as the proportion of the number of elements $\{i : x^* - \bar{x} = \pm 0.01\}$ in the variable. The average results of 20 experiments when $A(\bar{x}) = A(x^*)$ are presented in Table 2. From Table 2, we view that SPGE algorithm works better than SPG algorithm as a whole and especially we observe that the used time of the SPGE algorithm is far less than the SPG algorithm.

Example 4.3. Consider censored regression problem, that is,

\[
\min_{0 \leq x \leq 1} \mathcal{F}_{b_0}(x) := \frac{1}{pm} \sum_{i=1}^{m} \left\{ \max \{A_i x - c_i, 0\} - b_i \right\} + \lambda \|x\|_0. \tag{52}
\]
Table 2. Numerical results of the SPGE and SPG algorithm for problem (51) ($m \ll n$)

| m   | n   | s   | SPG time | SPGE time | rel-err SPG | rel-err SPGE | suc-rat SPG | suc-rat SPGE |
|-----|-----|-----|----------|-----------|-------------|-------------|------------|-------------|
| 60  | 120 | 12  | 7.68     | 0.96      | 0.008       | 0.004       | 96%        | 95%         |
| 80  | 160 | 16  | 8.59     | 1.21      | 0.006       | 0.002       | 96%        | 98%         |
| 100 | 200 | 20  | 12.35    | 2.19      | 0.011       | 0.005       | 94.6%      | 95.5%       |

The goal is to find the true sparse signal $x^*$ about the system $\max \{Ax^* - c_i, 0\} \approx b$ with $A \in \mathbb{R}^{m \times n}$ ($m \gg n$). The parameters about SPG algorithm are set as default values, one can refer to [5] in details and the parameters about SPGE algorithm are set as in the following $L = 1.5$, $\alpha = 1$, $\sigma = 0.9$, Maxiter = $10^4$, $\epsilon = 0.01$ $\kappa = \frac{1}{2}$, $L_f = \|A\|_{\infty}$, $\lambda := \lambda_0 L_f$ where $\lambda_0 \in [0.001: 0.001: 0.1]$. The extrapolation term $\beta_k$ in the SPGE algorithm is chosen as in the way of (17) with the fixed restart systems and we perform a fixed restart every 500 iterations. At the initialize point, we set $\mu_{-1} = \mu_0 = \mu_1 = 1$ and $t_{-1} = 1$. Let $x^0 = 0.1 * \text{ones}(n, 1)$, $v = \min \left\{ \frac{\lambda}{L_f}, 1 \right\}$. We illustrate the efficiency of the SPGE algorithm from the time, relative error, sparsity regression rate and successful rate aspects. The relative error, sparsity regression rate is defined as in [5] and the successful rate is defined as in example 4.2. The average numerical results of 20 tests are performed to illustrate the efficiency of SPGE algorithm and the details are presented in Table 3. From Table 3, we see that SPGE algorithm takes much less time than SPG algorithm and other indexes are better than the SPG algorithm.

Table 3. Numerical results of the SPGE and SPG algorithm for problem (52) ($m \gg n$)

| m   | n   | s   | SPG time | SPGE time | rel-err SPG | rel-err SPGE | suc-rat SPG | suc-rat SPGE | $A(\pi)$ |
|-----|-----|-----|----------|-----------|-------------|-------------|------------|-------------|----------|
| 500 | 100 | 20  | 5.42     | 1.16      | 6.48e-03    | 2.43e-03    | 98%        | 99%        | 20.7     | 19.7     |
| 1000| 200 | 40  | 30.35    | 5.89      | 5.48e-03    | 2.59e-03    | 98.5%      | 99%        | 39.97    | 39.99    |
| 2000| 400 | 80  | 236.24   | 40.32     | 8.00e-03    | 5.86e-03    | 98%        | 99%        | 79.65    | 79.80    |

5. Conclusion

This paper mainly propose the SPGE algorithm for the exact continuous model of $l_0$ regularization problem [5] and discuss its convergence for a fairly general choice of extrapolation parameters, in particular, the parameters can be chosen as in s-FISTA proposed by Bian [4] with fixed restart when $\beta_k \in [0, \sqrt{(1 - a \frac{\mu_{k+1}}{\mu_k} \frac{\mu_{k+1}}{\mu_k})}]$ with $a \in (0, 1)$ sufficiently small but not equal 0. Especially, it holds that $\sum_{k=0}^{+\infty} \|x^{k+1} - x^k\| < +\infty$ and the whole sequence $\{x_k\}$ convergences to a lifted stationary point of (1). Further, the rough convergence rate of $o\left(\frac{1}{k^{1/\sigma}}\right)$ ($\sigma \in (0, 1)$) with respect to squared proximal residual is established. From the numerical experiments, we see that SPGE algorithm is more efficient than SPG algorithm. In particular, in first example, when $v = 1$, the SPGE algorithm can find a global minimizer on different cases. But for SPG algorithm, only $v \geq 0.5$, it can find the global minimizer. Besides, the iterations and the used time of the SPGE algorithm are less than the SPG algorithm in these examples.
References

[1] Beck, A., Teboulle, M.: A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. Siam J. Imaging Sci. 2, 183-202 (2009)
[2] Beck, A., Teboulle, M.: Smoothing and First Order Methods: A Unified Framework. SIAM J. Optim. 22, 557-580 (2012)
[3] Bertsekas, D. P., Nonlinear Programming, 2nd ed.(1999)
[4] Bian, W.: Smoothing accelerated algorithm for constrained nonsmooth convex optimization problems (in Chinese). Sci. Sin. Math. 50, 1651-1666 (2020)
[5] Bian, W., Chen, X.: A Smoothing Proximal Gradient Algorithm for Nonsmooth Convex Regression with Cardinality Penalty. SIAM J. Numer. Anal. 58,858-883 (2020)
[6] Bian, W., Chen, X.: Smoothing neural network for constrained non-Lipschitz optimization with applications. IEEE Trans. Neural Netw. Learn. Syst. 23, 399-411(2012)
[7] 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)
[8] Chen, X.: Smoothing methods for nonsmooth, nonconvex minimization. Math. Program. 134, 71-99(2012)
[9] Chen, X.: Fukushima M, A Smoothing Method for a Mathematical Program with P-Matrix Linear Complementarity Constraints. Comput. Optim. Appl. 27, 223-246(2004)
[10] Chen, X., Nashed, Z., Qi, L.: Smoothing methods and semismooth methods for nondifferentiable operator equations. SIAM Journal on Numerical Analysis,38,1200-1216(2000)
[11] Chen, X., Peng, J., Zhang, S.: Sparse solutions to random standard quadratic optimization problems. Math. Program. 141, 273-293(2013)
[12] Clarkson, K, L., Hazan, E., Woodruff, D, P.: Sublinear optimization for machine learning. J. ACM. 59, 1-49(2012)
[13] Donoho, D., L.: Compressed sensing. IEEE Trans. Inf. Theory. 52, 1289-1306(2006)
[14] Ghadimi, S., Lan, G.: Accelerated gradient methods for nonconvex nonlinear and stochastic programming. Math. Program. 156, 59-99 (2016)
[15] Hu, M., Fukushima, M.: Smoothing approach to Nash equilibrium formulations for a class of equilibrium problems with shared complementarity constraints. Comput. Optim. Appl. 52, 415-437(2012)
[16] Jain, P., Kar, P.: Non-convex Optimization for Machine Learning. Foundations and Trendsin Machine Learning. 10, 142-336(2017)
[17] Liang, J., Schonlieb, C, B.: Faster FISTA. In 2018 26th European Signal Processing Conference (EUSIPCO),IEEE. 1-9, Rome, (2018)
[18] Luo, Z., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Ann. Oper. Res. 46, 157-178 (1993)
[19] Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. 103, 127-152(2005)
[20] Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers, Boston (2004)
[21] Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. 140, 125-161(2013)
[22] Beck, A.:First-order methods in optimization. Society for Industrial and Applied Mathematics(2017)
[23] Nesterov, Y.: Smoothing technique and its applications in semidefinite optimization. Math. Program. 110, 245-259(2007)
[24] Ochs, P., Chen, Y., Brox, T., Pock, T.:iPiano: Inertial proximal algorithm for nonconvex optimization. SIAM J. Imaging Sci. 7,1388-1419(2014)
[25] O’Donoghue, B., Candes, E.J.: Adaptive restart for accelerated gradient schemes. Found. Comput. Math. 15, 715–732 (2015)
[26] Pang, J. S., Razaviyayn, M., Alvarado, A.: Computing B-stationary points of nonsmooth DC programs. Math. Oper. Res. 42, 95-118(2017)
[27] Parikh, N., Boyd, S.: Proximal algorithms. FNT in Optimization. 1, 127-239(2014)
[28] Polyak, B. T.: Some methods of speeding up the convergence of iteration methods. Usr Computational Mathematics and Mathematical Physics, 4, 1-17(1964)
[29] Polyak, R, A.: Nonlinear rescaling vs. smoothing technique in convex optimization. Math. Program.
92,197-235(2002)

[30] Qi, L., Sun, D., Zhou, G.: A new look at smoothing Newton methods for nonlinear complementarity problems and box constrained variational inequalities. Math. Program. 87, 1-35(2000)

[31] Wen, B., Chen, X., 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)

[32] Wu, F., Bian, W., Xue, X.: Smoothing fast iterative hard thresholding algorithm for $\ell_0$ regularized nonsmooth convex regression problem. arXiv preprint arXiv:2104.13107

[33] Zhang, C., Chen, X.: Smoothing projected gradient method and its application to stochastic linear complementarity problems. SIAM J. Optim. 20, 627-649(2009)