General inertial smoothing proximal gradient algorithm for the relaxation of matrix rank minimization problem

Jie Zhang\(^1\) and Xinmin Yang\(^2\)*

\(^1\)College of Mathematics, Sichuan University, Chengdu, 610065, China.
\(^2\)*School of Mathematical Sciences, Chongqing Normal University, Chongqing, 401331, China.

*Corresponding author. E-mail: xmyang@cqnu.edu.cn;
authors: zjieabc@163.com; xmyang@cqnu.edu.cn;

Abstract

We consider the exact continuous relaxation model of matrix rank minimization problem proposed by Yu and Zhang (Comput. Optim. Appl. 1-20, 2022). Motivated by the inertial technique, we propose a general inertial smoothing proximal gradient algorithm (GIMSPG) for this kind of problems. It is shown that the singular values of any accumulation point have a common support set and the nonzero singular values have a unified lower bound. Besides, the zero singular values of the accumulation point can be achieved within finite iterations. Moreover, we prove that any accumulation point of the sequence generated by the GIMSPG algorithm is a lifted stationary point of the continuous relaxation model under the flexible parameter constraint. Finally, we carry out numerical experiments on random data and image data respectively to illustrate the efficiency of the GIMSPG algorithm.

Keywords: Smoothing approximation, Proximal gradient method, Inertial, Rank minimization problem

1 Introduction

In the recent years, much work has been dedicated to the matrix rank minimization problem which emerge in many applications, especially in computer
vision [38] and matrix completion [27]. Lots of models, methods and its variants have been studied, one can refer to the literatures [7, 10–12, 14, 18–21, 37], in detail, the matrix rank minimization problem can be represented as

$$
\min_{X \in \mathbb{R}^{m \times n}} \mathcal{F}(X) := f(X) + \lambda \cdot \text{rank}(X),
$$

(1)

where \text{rank}(X) denotes the number of nonzero singular values.

In this paper, we concentrate on the exact continuous relaxation of matrix rank minimization problem proposed in [34], that is the following nonconvex relaxation problem,

$$
\min_{X \in \mathbb{R}^{m \times n}} \mathcal{F}(X) := f(X) + \lambda \Phi(X),
$$

(2)

where \(m \geq n\) and \(f : \mathbb{R}^{m \times n} \rightarrow [0, \infty)\) is convex and not necessarily smooth, \(\lambda\) is a positive parameter. \(\Phi(X) = \sum_{i=1}^{n} \phi(\sigma_i(X))\) is an exact continuous relaxation of the matrix rank minimization problem. The capped-\(\ell_1\) function \(\phi\) with given \(v > 0\) is

$$
\phi(t) = \min \{1, t/v\}, \quad t \geq 0.
$$

(3)

It is observed that \(\phi\) in (3) can be seen as a DC function, that is

$$
\phi(t) = \frac{t}{v} - \max \{\theta_1(t), \theta_2(t)\},
$$

(4)

with \(\theta_1(t) = 0\), \(\theta_2(t) = t/v - 1(t \geq 0)\) and \(\mathcal{D}(t) = \{i \in \{1, 2\} : \theta_i(t) = \max\{\theta_1(t), \theta_2(t)\}\}\). When the matrix \(X\) is to be diagonal, the problem (2) is reduced to the exact continuous relaxation model of \(\ell_0\) regularization problem which was given in [5], i.e.,

$$
\min_{x \in \mathbb{R}^n} \mathcal{F}(x) := f(x) + \lambda \Phi(x),
$$

(5)

where \(\Phi(x) = \sum_{i=1}^{n} \phi(x_i)\) and \(\phi\) is defined as \(\phi(t) = \min \{1, |t|/v\}\) for \(t \in \mathbb{R}\).

In [5], one proposed the efficient smoothing proximal gradient method to solve this kind of problem with box constraint.

In [36], the authors established the smoothing proximal gradient method with extrapolation (SPGE) algorithm for solving (5). The extrapolation coefficient can be obtained \(\sup_\beta \beta_k = 1\). Further, under more strict condition of the extrapolation coefficient which includes the extrapolation coefficient in sFISTA with the fixed restart [25], it is proved that any accumulation point of the sequence generated by SPGE is a lifted stationary point of (5). Besides, the convergence rate based on the proximal residual is developed. Since the penalty item of problem (2) is nonconvex for a fixed \(d\) in (12), the framework of SPGE algorithm in [36] can not be directly applied to the matrix case.
As we know, incorporating inertial item which is also called extrapolation item to the proximal gradient method is a popular technique to improve the efficiency of proximal gradient algorithm for solving the following composition optimization problem,

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

where \( f : \mathbb{R}^n \to \mathbb{R} \) is a smooth with Lipschitz continuous gradient and possibly nonconvex function, \( g : \mathbb{R}^n \to (-\infty, +\infty] \) is a proper closed, and convex function. The composition structure that one item is smooth convex and the other is convex makes the proximal gradient method [26] widely used. Specifically, the updating scheme can be read as:

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

where \( t_k \) is the stepsize. Throughout this paper, the proximal mapping [26] of \( \lambda g \) is defined as

\[
\text{prox}_{\lambda g}(u) := \arg \min_{x \in \mathbb{R}^n} \left\{ g(x) + \frac{1}{2\lambda} \|x - u\|^2 \right\},
\]

where \( \lambda > 0 \) and \( u \in \mathbb{R}^n \). The accelerated proximal gradient method for convex optimization problems have been well studied [23, 24]. In particular, Beck and Teboulle [2] established the remarkable Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) for the convex case which was based on the Nesterov’s method [22, 24]. Some variants of FISTA have been developed by choosing appropriate extrapolation parameters, we refer the readers to [8, 17] and their references therein. In [32], under the proper parameter constraints, Wu and Li established the proximal gradient algorithm with different extrapolation (PGe) coefficients on the proximal step and gradient step for solving the problem (6) when the smooth component is nonconvex and the nonsmooth component is convex. The general framework is given as follows:

\[
\begin{align*}
    y^k &= x^k + \alpha_k (x^k - x^{k-1}), \\
    z^k &= x^k + \beta_k (x^k - x^{k-1}), \\
    x^{k+1} &= \arg \min_{x \in \mathbb{R}^n} \left\{ g(x) + \langle \nabla f(z^k), x - y^k \rangle + \frac{1}{2\lambda_k} \|x - y_k\|^2 \right\},
\end{align*}
\]

where \( \alpha_k, \beta_k \) are the extrapolation coefficients, and \( \lambda_k \) is the stepsize. In [31], the authors further developed the inertial Bregman proximal gradient method for minimizing the sum of two possible nonconvex functions. In their method, the generalized Bregman distance replaces the Euclidean distance and two different inertial items on the proximal step and gradient step are taken respectively.
For matrix case, Toh and Yun [29] proposed the accelerated proximal gradient algorithm for the following convex problem:

$$\min_{X \in \mathbb{R}^{m \times n}} F(X) = f(X) + g(X),$$

where $f$ is a convex and smooth function with Lipschitz continuous gradient and $g$ is a proper closed convex function. The problem arises in applications such as in matrix completion problems [29], multi-task learning [1] and principal component analysis (PCA) [21, 33]. When the matrix is diagonal, the problem is reduced to the vector optimization problem (6).

Since optimization problem (2) is a matrix generalization of the vector optimization problem (5), it is natural for us to explore the possibility of extending some of the algorithms developed for problem (5) to solve problem (2). We observe that the loss function in above require a restrictive assumption that the function $f$ is smooth, it is difficult to apply the proximal gradient method directly for problem (2). Fortunately, the smoothing approximation methods proposed in [9] can overcome the dilemma. Smoothing approximation method is an efficient method which is widely used in other problems. One can refer to [30, 35] and their references therein. Especially, in [3], the author proposed the smoothing fast iterative shrinkage thresholding algorithm (sFISTA) for the convex problem with the following extrapolation coefficient under the vector case, i.e.,

$$\begin{align*}
\alpha_k &= \beta_k = \frac{t_k - 1}{t_k + 1}, \\
t_{k+1} &= \frac{1 + \sqrt{1 + 4t_k^2 \mu_k \mu_{k+1}}}{2},
\end{align*}$$

where $t_0 = 1$ and $\mu_0 \in (0, 1)$. They gave the global convergence rate $O(\ln k/k)$ on the objective function.

Motivated by the smoothing method and the general accelerated technique in (7), we propose a general inertial smoothing proximal gradient method (GIMSPG) for matrix case. Thanks to the special structure of penalty item of (2), though the subproblem is nonconvex in the GIMSPG algorithm, it has a closed form solution. Most recently, Li and Bian [16] studied a class of sparse group $\ell_0$ regularized optimization problem and gave an exact continuous relaxation model of it. They proposed the difference of convex (DC) algorithms for solving the relaxation model which has the DC structure. Moreover, they discussed the zero elements of the accumulation point has finite iterations. Inspired by this, under certain conditions of the parameters, we prove that the zero singular values of any accumulation point of the iterates generated by the GIMSPG algorithm have the same locations and the nonzero singular values are always not less than a fixed value. Besides, the GIMSPG algorithm has the ability identifying the zero singular values of any accumulation point within finite iterations. Furthermore, we show that any accumulation point of
the sequence generated by GIMSPG algorithm is a lifted stationary point of the continuous relaxation problem (2).

The outline of this paper is as follows. Some preliminaries are presented in Section 2. Then we establish the GIMSPG algorithm framework and discuss its convergence in Section 3. In Section 4, we conduct numerical experiments to illustrate the efficiency of the proposed GIMSPG algorithm. Finally, we draw some conclusions in Section 5.

Notations. Let $\mathbb{R}^{m \times n}$ be the matrix space of $m \times n$ with the standard inner product, i.e., for any $X, Y \in \mathbb{R}^{m \times n}(m \geq n)$, $\langle X, Y \rangle = \text{tr}(X^T Y)$, where $\text{tr}(X^T Y)$ is the trace of the matrix $X^T Y$. For any matrix $X \in \mathbb{R}^{m \times n}$, the Frobenius norm and nuclear norm are respectively denoted by $\|X\|_F = \sqrt{\text{tr}(X^T X)}$, $\|X\|_* = \sum_{i=1}^{n} \sigma_i(X)$ where $\sigma(X) := (\sigma_1(X), \ldots, \sigma_n(X))^T$ is the singular values of $X$ with $\sigma_1(X) \geq \sigma_2(X) \geq \ldots \geq \sigma_n(X) \geq 0$. Denote $\mathcal{A}(\sigma(X)) = \{ i \in \{1, 2, \ldots, n\} : \sigma_i(X) \neq 0 \}$, $\|X\|_1 = \|\text{vec}(X)\|_1$, where $\text{vec}(X)$ is the vectorization operation of a matrix $X$. Denote $\partial f(X)$ be a Clarke subgradient of $f$ at $X \in \mathbb{R}^{m \times n}$ where $f(X) : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ is a locally Lipschitz continuous function. $\mathcal{D}(x)$ is a diagonal matrix generated by vector $x$. $E_i = \mathcal{D}(e_i)$ where $e_i$ is the unit vector and the $i$-th element is 1. Denote $\mathbb{Q}^n$ be the set of $n \times n$ dimension unitary orthogonal matrix and $\mathbb{S}^n$ be the real and symmetric $n \times n$ matrices. For vector $x \in \mathbb{R}^n$, $\|x\|_1 = \sum_{i=1}^{n} |x_i|$. Denote $\mathbb{N} = \{0, 1, 2, \ldots\}$, $\mathbb{D}^n = \{ d \in \mathbb{R}^n : d_i \in \{1, 2\}, i = 1, 2, \ldots, n \}$.

2 Preliminaries

In this section, we first recall some definitions and primary results which are the basis for the rest of the discussion. First, we give the form of the Clarke subdifferential of $\Phi$ at $X \in \mathbb{R}^{m \times n}$ followed by [15], that is

$$\partial \Phi(X) = \{ U \mathcal{D}(x)V^T : x \in \partial \sum_{i=1}^{n} \phi(\sigma_i(X)), U, V \in \mathcal{M}(X) \},$$

where $\mathcal{M}(X) = \{ (U, V) \in \mathbb{Q}^{m \times m} \times \mathbb{Q}^{n \times n} : U^T U = V^T V = I, X = U \mathcal{D}(\sigma(X))V^T \}.$

Definition 1 (Lifted stationary point [34]) We say that $X \in \mathbb{R}^{m \times n}$ is a lifted stationary point of (2) if there exist $d_i \in \mathcal{D}(\sigma_i(X))$ for $i = 1, 2, \ldots, n$ such that

$$\lambda \sum_{i=1}^{n} \theta_{d_i}^L(\sigma_i(X))E_i \in \{ U \partial f(X)V^T + \frac{\lambda}{\nu} \mathcal{D}(\partial \|x\|_1 |_{x=\sigma_i(X)}) : (U, V) \in \mathcal{M}(X) \},$$

(10)

where $\sigma_i(X)$ is the $i$th largest singular value of $X$.
Assumption 1 \( f \) is Lipschitz continuous on \( \mathbb{R}^{m \times n} \) with the Lipschitz continuous constant \( L_f \).

Assumption 2 The parameter \( v \) in (3) satisfies \( 0 < v < \bar{v} := \frac{\lambda}{L_f} \).

Assumption 3 \( \mathcal{F} \) in (2)(or \( \mathcal{F}_{\ell_0} \) in (1) ) is level bounded on \( \mathbb{R}^{m \times n} \).

Lemma 1 ([34]) If \( \bar{X} \) is a lifted stationary point of (2), then the vector \( d_{\bar{X}} = (d_{\bar{X}}^1, \ldots, d_{\bar{X}}^n) \in \prod_{i=1}^n \mathcal{D}(\sigma_i(\bar{X})) \) satisfying (10) is unique. In particular, for \( i = 1, 2, \ldots, n \), it has that
\[
\begin{cases}
1, & \text{if } \sigma_i(\bar{X}) < v, \\
2, & \text{if } \sigma_i(\bar{X}) \geq v.
\end{cases}
\]

Lemma 2 ([34]) \( \bar{X} \) is a global minimizer of (2) if and only if it is a global minimizer of (1) and the objective functions have the same value at \( \bar{X} \).

Lemma 3 ([34]) For a given \( W \in \mathbb{R}^{m \times n} \) and \( \tau > 0 \), suppose \( U \mathcal{D}(w)V^T \) be the singular value decomposition of \( W \) and \( \hat{x} = \text{prox}_{\tau \Phi_d}(w) \), then \( \hat{X} = U \mathcal{D}(\hat{x})V^T \) is an optimal solution of the problem
\[
\min_X \{ \tau \Phi_d(X) + \frac{1}{2} \| X - W \|^2 \}.
\]

3 The general inertial smoothing proximal gradient algorithm and its convergence analysis

In order to overcome the nondifferentiability of the convex loss function \( f \) in (2), we use a smoothing function defined as in [34] to approximate convex function \( f \).

Definition 2 We call \( \tilde{f} : \mathbb{R}^{m \times n} \times [0, \bar{\mu}] \to \mathbb{R} \) with \( \bar{\mu} > 0 \) a smoothing function of the function \( f \) in (2) if \( \tilde{f}(\cdot, \mu) \) is continuous differentiable in \( \mathbb{R}^{m \times n} \) for any fixed \( \mu > 0 \) and satisfies the following conditions:
(i) \( \lim_{Z \to X, \mu \downarrow 0} \tilde{f}(Z, \mu) = f(X) \quad \forall X \in \mathbb{R}^{m \times n} \);
(ii) \( \tilde{f}(X, \mu) \) is convex with respect to \( X \) for any fixed \( \mu > 0 \);
(iii) \( \{ \lim_{Z \to X, \mu \downarrow 0} \nabla Z \tilde{f}(Z, \mu) \} \subseteq \partial f(X) \quad \forall X \in \mathbb{R}^{m \times n} \);
(iv) there exists a positive constant \( \kappa \) such that
\[
|\tilde{f}(X, \mu_2) - \tilde{f}(X, \mu_1)| \leq \kappa |\mu_1 - \mu_2|, \quad \forall X \in \mathbb{R}^{m \times n}, \mu_1, \mu_2 \in [0, \bar{\mu}],
\]
especially,
\[
|\tilde{f}(X, \mu) - f(X)| \leq \kappa \mu, \quad \forall X \in \mathbb{R}^{m \times n}, 0 < \mu \leq \bar{\mu}.
\]
(v) there exists a constant $\tilde{L} > 0$ such that for any $\mu \in (0, \bar{\mu}]$, $\nabla_X \tilde{f}(\cdot, \mu)$ is Lipschitz continuous with Lipschitz constant $\tilde{L}\mu^{-1}$.

Denote

$$\Phi^d(X) := \sum_{i=1}^{n} \left( \sigma_i(X) \frac{\sigma_i(X)}{v} - \theta_{d_i}(\sigma_i(X)) \right) = (d_1, d_2, \ldots, d_n)^T \in \mathbb{D}^n. \quad (12)$$

It can be easily got that

$$\Phi(X) = \min_{d \in \mathbb{D}^n} \Phi^d(X) \quad \forall X \in \mathbb{R}^{m \times n}, \quad (13)$$

and $\Phi(\bar{X}) = \Phi^{d^X}(\bar{X})$ with fixed $\bar{X}$ where $d^X$ is defined in (11).

For the discussion convenience, we give the notations in the following way

$$\tilde{F}^d(X, \mu) := \tilde{f}(X, \mu) + \lambda \Phi^d(X) \quad \text{and} \quad \tilde{F}(X, \mu) := \tilde{f}(X, \mu) + \lambda \Phi(X), \quad (14)$$

where $\tilde{f}$ is a smoothing function of $f$, $\mu > 0$, and $d \in \mathbb{D}^n$. By the formulation of (13), we have

$$\tilde{F}^d(X, \mu) \geq \tilde{F}(X, \mu), \quad \forall d \in \mathbb{D}^n, X \in \mathbb{R}^{m \times n}, \mu \in (0, \bar{\mu}]. \quad (14)$$

In the following, we focus on the following nonconvex optimization problem with the given smoothing parameter $\mu > 0$, and $d \in \mathbb{D}^n$:

$$\tilde{F}^d(X, \mu) = \tilde{f}(X, \mu) + \lambda \Phi^d(X). \quad (15)$$

When the matrix $X$ is the diagonal matrix, since the penalty item of (14) is piecewise linearity function, the proximal operator of $\tau \Phi^d$ on $\mathbb{R}^n$ has a closed form solution with the given vectors $d \in \mathbb{D}^n$, $w \in \mathbb{R}^n$, and a positive number $\tau > 0$, in detail, the following proximal operator

$$\hat{x} = \arg \min_{x \in \mathbb{R}^n} \left\{ \tau \Phi^d(x) + \frac{1}{2} \|x - w\|^2 \right\} \quad (16)$$

can be calculated by $\hat{x}_i = \max \left\{ \hat{w}_i - \frac{\tau}{v}, 0 \right\}$, $\forall i = 1, 2, \ldots, n$, where

$$\hat{w}_i = \begin{cases} w_i, & \text{if } d_i = 1, \\
 w_i + \tau/v, & \text{if } d_i = 2. \end{cases} \quad (17)$$

Motivated by the smoothing approximation technique, the proximal gradient algorithm and the efficiency of inertial method, we consider the following
general intrial smoothing proxiaml gradient method of $\tilde{F}^d_k(\cdot, \mu_k)$, i.e.,

$$
\begin{cases}
Y^k = X^k + \alpha_k (X^k - X^{k-1}), \\
Z^k = X^k + \beta_k (X^k - X^{k-1}), \\
Q^d_k (X, Y^k, Z^k, \mu_k) = \langle X - Y^k, \nabla \tilde{f}(Z^k, \mu_k) \rangle + \frac{1}{2} h_k \mu_k^{-1} \| X - Y^k \|^2 + \lambda \Phi^d_k (X),
\end{cases}
$$

where $\alpha_k, \beta_k$ are different extrapolation parameters and certain conditions are required for them, $h_k$ is the parameter depending on $\tilde{L}$. Especially, when $\alpha_k = \beta_k = 0$, the GIMSPG algorithm is reduced to the MSPG algorithm.

Based on (16) and Lemma 3, we get that the problem $\min Q^d_k (X, Y^k, Z^k, \mu_k)$ has a minimizer $\hat{X} = UD(\hat{x})V^T$ with $\hat{x} = \text{prox}_{\tau \Phi^d_k (w)}$, $\tau = \lambda \mu_k h_k^{-1}$ and $W^k = UD(w^k)V^T$, $W^k = Y^k - h_k^{-1} \mu_k \nabla \tilde{f}(Z^k, \mu_k)$.

**Assumption 4** (Parameters constraints) The parameters $\alpha_k, \beta_k, h_k$ in the GIMSPG algorithm satisfy the following conditions: for any $0 < \varepsilon \ll 1$, $\alpha_k$ is nonincreasing and $\alpha_k \in \left[0, \frac{1 - 2^2 \varepsilon}{1 + 2 \sigma} \right)$, $\beta_k \in [0, 1]$, $\{h_k\}$ is nonincreasing and satisfies $h_0^{-1} \leq h_k^{-1} \leq \min \{ \frac{1 - \alpha_k - (\alpha_k + \varepsilon) \mu_k}{(1 - \beta_k) L}, \frac{\alpha_k}{\beta_k L} \}$.

**Algorithm 1** GIMSPG algorithm

1. **Input:** $X^{-1} = X^0 \in \mathbb{R}^{m \times n}$, $\mu_{-1} = \mu_0 \in (0, \bar{\mu}]$, $\sigma \in (0, 1)$, choosing the parameters $\alpha_k, \beta_k, h_k$ satisfying the Assumption 4. Set $k = 0$.
2. **while** a termination criterion is not met **do**
3. **step 1:** Let $d^k = d^X_k$, where $d^X_k$ is defined as in (11).
4. **step 2:** Compute

$$
\begin{cases}
Y^k = X^k + \alpha_k (X^k - X^{k-1}), \\
Z^k = X^k + \beta_k (X^k - X^{k-1}), \\
X^{k+1} \in \arg \min_{X \in \mathbb{R}^{m \times n}} Q^d_k (X, Y^k, Z^k, \mu_k).
\end{cases}
$$

5. **step 3:** If

$$H_{\delta_k+1} (X^{k+1}, X^k, \mu_{k+1}, \mu_k) - H_{\delta_k} (X^k, X^{k-1}, \mu_k, \mu_{k-1}) \leq - \alpha \mu_k^2,
$$

set $\mu_{k+1} = \mu_k$, otherwise, set

$$
\mu_{k+1} = \frac{\mu_0}{(k + 1) \sigma},
$$

set $k := k + 1$, and return to Step 1.
6. **end while**
Remark 1 The parameter $h_k$ is bounded. Indeed, if $\frac{1-\alpha_k - (\alpha_k + \varepsilon) \frac{\mu_k}{\mu_{k+1}}}{(1-\beta_k)L} > \frac{1}{L}$ and $\frac{\alpha_k}{\beta_k} L > \frac{1}{L}$, then it has that $\alpha_k + (\alpha_k + \varepsilon) \frac{\mu_k}{\mu_{k+1}} < \beta_k < \alpha_k$ which is impossible, then $\frac{1}{L}$ can be seen as a upper bound of $h_k^{-1}$.

Remark 2 From Assumption 4, when $\frac{1-\alpha_k - (\alpha_k + \varepsilon) \frac{\mu_k}{\mu_{k+1}}}{(1-\beta_k)L} = \frac{\alpha_k}{\beta_k} L$, we have the largest range of the stepsize. In this case, $\alpha_k = \beta_k \frac{(1-\varepsilon) \frac{\mu_k}{\mu_{k+1}}}{1+ \beta_k \frac{\mu_k}{\mu_{k+1}}}$. Moreover, from $\inf \frac{\mu_k}{\mu_{k+1}} = 1$ and $\sup \frac{\mu_k}{\mu_{k+1}} = 2^\sigma$, we can take $\alpha_k = \frac{0.98 \beta_k}{1+2^\sigma \beta_k}$ and $h_k = \frac{(1+2^\sigma \beta_k)L}{0.98}$. It needs to mention that the parameters in the GIMSPG algorithm can be selected adaptively. However, how to select appropriate parameters is beyond the scope of this paper.

Next, we are ready to discuss the convergence by constructing the auxiliary sequence. For any $k \in \mathbb{N}$, define

$$H_{\delta_{k+1}}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) := \tilde{F}(X^{k+1}, \mu_k) + \kappa \mu_k + \delta_{k+1} \mu_{k+1}^{-1} \|X^{k+1} - X^k\|^2,$$

where $\{X^k\}$, $\{\mu_k\}$ are the sequences generated by GIMSPG algorithm and $\delta_{k+1} = \frac{h_{k+1} \alpha_{k+1}}{2}$. It needs to mention that the sequence is essential to the convergence analysis of the GIMSPG algorithm. Now, we start by showing that GIMSPG algorithm is well defined.

**Lemma 4** The proposed GIMSPG algorithm is well defined, and the sequence $\{\mu_k\}$ generated by the GIMSPG has the property that there are infinite elements in $\mathcal{N}^s$ and $\lim_{k \to +\infty} \mu_k = 0$, where $\mathcal{N}^s := \{k \in \mathbb{N} : \mu_{k+1} \neq \mu_k\}$.

**Proof** We can easily get the result from [36] in Lemma 4. $\square$

Then we are intend to prove that the auxiliary sequence $H_{\delta_{k+1}}(X^{k+1}, X^k, \mu_{k+1}, \mu_k)$ is monotonically decreasing.

**Lemma 5** For any $k \in \mathbb{N}$, suppose $\{X^k\}$ be the sequence generated by GIMSPG algorithm, it holds that

$$H_{\delta_{k+1}}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) - H_{\delta_k}(X^k, X^{k-1}, \mu_k, \mu_{k-1}) \leq \left(\frac{\tilde{L} - h_k}{2} + \frac{h_k \alpha_k - \beta_k \tilde{L}}{2}\right) \mu_k^{-1} \|X^{k+1} - X^k\|^2 + \frac{\tilde{L} \beta_k (\beta_k - 1)}{2} \mu_k^{-1} \|X^k - X^{k-1}\|^2.$$

(22)
Moreover, when the parameters $\alpha_k, \beta_k, h_k$ satisfy Assumption 4, it holds that

\[
H_{\delta_{k+1}}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) - H_{\delta_k}(X^k, X^{k-1}, \mu_k, \mu_{k-1}) \leq \frac{-\varepsilon h_k}{2} \mu_{k+1}^{-1} \|X^{k+1} - X^k\|^2, \tag{23}
\]

and the sequence $H_{\delta_{k+1}}(X^{k+1}, X^k, \mu_{k+1}, \mu_k)$ is nonincreasing.

**Proof** From the optimality of subproblem (18), we have

\[
\langle X^{k+1} - Y^k, \nabla \tilde{f}(Z^k, \mu_k) \rangle + \frac{1}{2} h_k \mu_k^{-1} \|X^{k+1} - Y^k\|^2 + \lambda \Phi^{dk}(X^{k+1}) \\
\leq \langle X^k - Y^k, \nabla \tilde{f}(Z^k, \mu_k) \rangle + \frac{1}{2} h_k \mu_k^{-1} \|X^k - Y^k\|^2 + \lambda \Phi^{dk}(X^k). \tag{24}
\]

Since $\nabla \tilde{f}(\cdot, \mu_k)$ is Lipschitz continuous with modulus $\bar{L} \mu_k^{-1}$, it follows from the Definition 2(v) that

\[
\tilde{f}(X^{k+1}, \mu_k) \leq \tilde{f}(Z^k, \mu_k) + \langle X^{k+1} - Z^k, \nabla \tilde{f}(Z^k, \mu_k) \rangle + \frac{1}{2} \bar{L} \mu_k^{-1} \|X^{k+1} - Z^k\|^2. \tag{25}
\]

Moreover, by the convexity of $\tilde{f}(\cdot, \mu_k)$, it holds that

\[
\tilde{f}(Z^k, \mu_k) + \langle X^k - Z^k, \nabla \tilde{f}(Z^k, \mu_k) \rangle \leq \tilde{f}(X^k, \mu_k). \tag{26}
\]

Combining (24), (25) and (26), we have

\[
\tilde{f}^{dk}(X^{k+1}, \mu_k) - \tilde{f}^{dk}(X^k, \mu_k) \leq h_k \mu_k^{-1} \langle X^k - Y^k, X^{k+1} - X^k \rangle - \frac{1}{2} h_k \mu_k^{-1} \|X^{k+1} - X^k\|^2 + \tilde{L} \mu_k^{-1} \|X^{k+1} - Z^k\|^2. \tag{27}
\]

Denote $\Delta_k := X^k - X^{k-1}$, we have $\alpha_k \Delta_k = Y^k - X^k$, $\beta_k \Delta_k = Z^k - X^k$, $\beta_k \Delta_k - \Delta_{k+1} = Z^k - X^{k+1}$. Then it follows from (27) that,

\[
\tilde{f}^{dk}(X^{k+1}, \mu_k) - \tilde{f}^{dk}(X^k, \mu_k) \leq h_k \mu_k^{-1} \langle \alpha_k \Delta_k, \Delta_{k+1} \rangle + \frac{\bar{L} \mu_k^{-1}}{2} \|\beta_k \Delta_k - \Delta_{k+1}\|^2 - \frac{1}{2} h_k \mu_k^{-1} \|\Delta_{k+1}\|^2 \\
= \frac{\bar{L} - h_k \mu_k^{-1}}{2} \|\Delta_{k+1}\|^2 + \frac{\bar{L} \beta_k^2 \mu_k^{-1}}{2} \|\Delta_k\|^2 + (h_k \alpha_k - \beta_k \bar{L}) \mu_k^{-1} \langle \Delta_k, \Delta_{k+1} \rangle \\
\leq \frac{\bar{L} - h_k}{2} \|\Delta_{k+1}\|^2 + \frac{h_k \alpha_k - \beta_k \bar{L}}{2} \mu_k^{-1} \|\Delta_{k+1}\|^2 + \left(\frac{\bar{L} \beta_k^2}{2} + \frac{h_k \alpha_k - \beta_k \bar{L}}{2}\right) \mu_k^{-1} \|\Delta_k\|^2, \tag{28}
\]
where the second inequality comes from the Assumption 4 and Cauchy-Schwartz inequality.

Letting $d^k = dX^k$ and according to $\tilde{F}(X^{k+1}, \mu_k) \geq \tilde{F}(X^{k+1}, \mu_k)$, we have

$$\tilde{F}(X^{k+1}, \mu_k) + \frac{\alpha_{k+1} h_{k+1} \mu_k^{-1}}{2} \| \Delta_{k+1} \|^2 - (\tilde{F}(X^k, \mu_k) \geq \frac{\alpha_k h_k \mu_k^{-1}}{2} \| \Delta_k \|^2)$$

$$\leq (\frac{\tilde{L} - h_k}{2} + \frac{h_k \alpha_k - \beta_k \tilde{L}}{2}) \mu_k^{-1} + \frac{\alpha_{k+1} h_{k+1} \mu_k^{-1}}{2} \| \Delta_{k+1} \|^2$$

$$+ (\frac{\tilde{L}^2 + h_k \alpha_k - \beta_k \tilde{L}}{2}) \mu_k^{-1} - \frac{\alpha_k h_k \mu_k^{-1}}{2} \| \Delta_k \|^2.$$

Combining (29), (30) with the nonincreasing of $h_k, \alpha_k$, we get

$$H_{\delta_{k+1}}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) - H_{\delta_k}(X^k, X^{k-1}, \mu_k, \mu_{k-1})$$

$$\leq (\frac{\tilde{L} - h_k}{2} + \frac{h_k \alpha_k - \beta_k \tilde{L}}{2}) \mu_k^{-1} + \frac{\alpha_k h_k \mu_k^{-1}}{2} \| \Delta_{k+1} \|^2 + \frac{\tilde{L}(\beta_k^2 - \beta_k)}{2} \mu_k^{-1} \| \Delta_k \|^2.$$

According to the parameters constraint in Assumption 4, we easily have that $(\tilde{L} - h_k + h_k \alpha_k - \beta_k \tilde{L}) \mu_k^{-1} + \alpha_k h_k \mu_k^{-1} \leq h_k \mu_k^{-1} \epsilon, \beta_k^2 - \beta_k \leq 0$. Hence, the inequality (23) holds. So the desired results are obtained. \(\square\)

**Corollary 1** For any $k \in \mathbb{N}$, under the Assumption 4, there exists $\zeta \in \mathbb{R}$ satisfying

$$\zeta := \lim_{k \to +\infty} H_{\delta_{k+1}}(X^{k+1}, X^k, \mu_{k+1}, \mu_k).$$

Further, it holds that

$$\lim_{k \to +\infty} H_{\delta_{k+1}}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) = \lim_{k \to +\infty} \tilde{F}(X^k, \mu_{k-1}) = \lim_{k \to +\infty} \mathcal{F}(X^k) = \zeta.$$

(31)

and the sequence $\{X^k\}$ is bounded.

**Proof** By the definition of $H_{\delta_{k+1}}$ from (21), we obtain that

$$H_{\delta_{k+1}}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) \geq \tilde{F}(X^{k+1}, \mu_k) + \kappa \mu_k \geq \mathcal{F}(X^{k+1}) \geq \min_{X \in \mathbb{R}^{m \times n}} \mathcal{F}(X) = \min_{X \in \mathbb{R}^{m \times n}} \mathcal{F}_{\ell_0}(X),$$

where the last equality holds due to the result that the global minimizers of problem (2) and problem (1) is equivalent from Lemma 2, further, according to
the fact that \( \{H_{\delta+1}(X^{k+1}, X^k, \mu_{k+1}, \mu_k)\} \) is nonincreasing, then there exists \( \zeta \in \mathbb{R} \) such that

\[
\lim_{k \to +\infty} H_{\delta+1}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) = \zeta, \quad H_{\delta+1}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) \geq \zeta, \forall k \in \mathbb{N}.
\]

(32)

Next, note that

\[
|\tilde{F}(X^{k+1}, \mu_k) + \kappa \mu_k - \zeta| = |H_{\delta+1}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) - \zeta| - \frac{\alpha \mu_{k+1} h_{k+1}^2 \|\Delta_{k+1}\|^2}{2} \\
\leq |H_{\delta+1}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) - \zeta| + \frac{\alpha \mu_{k+1} h_{k+1}}{2 \varepsilon h_k}, \varepsilon h_k \mu_{k+1}^2 \|\Delta_{k+1}\|^2 \\
\leq |H_{\delta+1}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) - \zeta| + \frac{\alpha \mu_{k+1} h_{k+1}}{2 \varepsilon h_k} (H_{\delta_0}(X^k, X^{k-1}, \mu_k, \mu_{k-1}) - \zeta)
\]

(33)

where the first inequality holds due to the triangle inequality and the second one derives from (23) and (32). When \( k \) is sufficiently large, the right of the above inequality tends to 0 by means of (32) and the boundedness of \( \frac{\alpha \mu_{k+1} h_{k+1}}{2 \varepsilon h_k} \). Therefore, it holds that

\[
\lim_{k \to +\infty} H_{\delta+1}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) = \lim_{k \to +\infty} \tilde{F}(X^{k+1}, \mu_k) = \lim_{k \to +\infty} F(X^k) = \zeta.
\]

From the nonincreasing property of the sequence \( \{H_{\delta+1}(X^{k+1}, X^k, \mu_{k+1}, \mu_k)\} \) again, it holds that

\[
F(X^{k+1}) \leq \tilde{F}(X^{k+1}, \mu_k) + \kappa \mu_k \leq H_{\delta+1}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) \\
\leq H_{\delta_0}(X^0, X^{-1}, \mu_0, \mu_{-1}) < +\infty.
\]

Then we get that \( \{X^k\} \) is bounded for any \( k \in \mathbb{N} \) from the level bounded of \( F \) by the Assumption 3.

\[\square\]

**Corollary 2** Under the parameters constraint in Assumption 4, it holds that

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

Moreover, \( \lim_{k \to +\infty} \|X^k - X^{k-1}\| = 0 \).
Proof Summing up the inequality (23) from \( k = 0 \) to \( k = K \) for any \( K \in \mathbb{N} \), it yields that
\[
\sum_{k=0}^{K} \frac{\varepsilon h_k}{2} \mu_{k+1}^{-1} \|X^{k+1} - X^k\|^2 \leq H_{\delta_0}(X^0, X^{-1}, \mu_0, \mu_{-1}) - H_{\delta_{K+1}}(X^K, X_K, \mu_{K+1}, \mu_K),
\]
when \( K \) is sufficiently large, from (32) and the boundedness of \( h_k \), we have
\[
\sum_{k=0}^{+\infty} \mu_{k+1}^{-1} \|X^{k+1} - X^k\|^2 < +\infty.
\]
Since \( \{\mu_k\} \) is a nonincreasing sequence, it can be easily got that
\[
\sum_{k=0}^{+\infty} \mu_{k+1}^{-1} \|X^{k+1} - X^k\|^2 \leq \sum_{k=0}^{+\infty} \mu_{k+1}^{-1} \|X^{k+1} - X^k\|^2 < +\infty.
\]
Further, from \( \mu_k \leq \bar{\mu} < 1 \) for any \( k \in \mathbb{N} \), we have
\[
\lim_{k \to +\infty} \|X^{k+1} - X^k\| = 0.
\]
So the proof is completed. \( \square \)

**Theorem 1** Suppose \( \{X^k\} \) be the sequence generated by the GIMSPG algorithm. Then, we have that

(i) for any \( k \in \mathbb{N}^* \), \( d^k \) in Algorithm 1 only changes finite number of times;

(ii) for any accumulation point \( \bar{X} \) and \( X^* \) of \( \{X^k, k \in \mathbb{N}^*\} \), it follows that \( \mathcal{A}^c(\sigma(\bar{X})) = \mathcal{A}^c(\sigma(X^*)) \), moreover, there exists \( J \in \mathbb{N} \), for any \( i \geq J \), it holds that
\[
\|\sigma(X^k_i)_{\mathcal{A}^c(\sigma(\bar{X}))} - \sigma(X_{\mathcal{A}^c(\sigma(\bar{X}))})\| = 0,
\]
where \( \{X^k_i\} \) is a subsequence of \( \{X^k : k \in \mathbb{N}^*\} \) such that \( \lim_{i \to +\infty} X^k_i = \bar{X} \).

**Proof** (i) From the first order optimality condition of (18), there exist \( U, V \in \mathcal{M}(X^{k+1}) \) such that
\[
U(\nabla \tilde{f}(Z^k, \mu_k) + h_k \mu_k^{-1}(X^{k+1} - Y^k))V^T + \frac{\lambda}{v} \partial\|\sigma(X^{k+1})\|_1 - \lambda \sum_{j=1}^{n} \nabla \theta_{d_j}^\ast(\sigma_j(X^{k+1}))E_j = 0.
\]
(34)
From (23) and for any \( k \in \mathcal{N}^s \), the inequality (19) does not hold, then we get

\[
\frac{\varepsilon h_k}{2} \mu_{k+1}^{-1} \| X^{k+1} - X^k \|^2 \leq \alpha \mu_k^2.
\]

That is,

\[
\sqrt{\frac{\varepsilon h_k}{2} \mu_{k+1}^{-1} \| X^{k+1} - X^k \|} \leq \sqrt{\alpha \mu_k^2 \mu_{k+1}^{-1}}.
\]

According to \( \lim_{k \to +\infty} \mu_k = 0 \), the nonincreasing of \( \mu_k \) and the boundedness of \( h_k \), we have that

\[
\lim_{k \to +\infty} \mu_{k+1}^{-1} \| X^{k+1} - X^k \| = 0, \quad \lim_{k \to +\infty} \mu_k^{-1} \| X^{k+1} - X^k \| = 0.
\]

(35)

Further, since

\[
\mu_k^{-1} \| X^{k+1} - Y^k \| \leq \mu_k^{-1} \| X^{k+1} - X^k \| + \mu_k^{-1} \| X^k - X^{k-1} \|,
\]

we have \( \lim_{k \to +\infty} \mu_k^{-1} \| X^{k+1} - Y^k \| = 0 \). From the boundedness of \( h_k \), we get that

\[
\lim_{k \to +\infty} h_k \mu_k^{-1} \| X^{k+1} - Y^k \| = 0,
\]

then there exists \( K \in \mathbb{N} \) such that for any \( k \geq K \), it has that

\[
h_k \mu_k^{-1} \| X^{k+1} - Y^k \| < \varepsilon,
\]

(37)

where \( \varepsilon = \frac{1}{2} (\frac{\lambda}{v} - L_f) \).

If there exist \( k_0 \geq k \) and \( j \in \{1, 2, \ldots, n\} \) such that \( \sigma_j(X^{k_0}) < v \), then \( d_j^{k_0} = 1 \) from (11) which yields that \( \nabla \theta_{d_j^{k_0}} (\sigma_j(X^{k_0+1})) = 0 \).

Next, we will prove \( \sigma_j(X^{k_0+1}) = 0 \) by contradiction. If \( \sigma_j(X^{k_0+1}) \neq 0 \), from (37), Assumption 2 and Definition 2(iii), then it holds that

\[
(U \nabla f(Z^k, \mu_k)V^T)_{jj} + h_k \mu_k^{-1} (U(X^{k+1} - Y^k)V^T)_{jj} + \frac{\lambda}{v} \neq 0,
\]

which contradicts (34). In conclusion, if there exist \( k_0 \geq k \) and \( j \in \{1, 2, \ldots, n\} \) satisfying \( \sigma_j(X^{k_0}) < v \), then \( \sigma_j(X^{k_0+1}) = 0 \). Then we get that \( \sigma_j(X^k) \equiv 0 \) for any \( \tilde{k} > k_0 \). Therefore, for sufficiently large \( \tilde{k} \), it holds that \( \sigma_j(X^k) \equiv 0 \) or \( \sigma_j(X^k) \geq v \) for \( j = 1, \ldots, n \). So \( d^k \) only changes finite number of times.

(ii) Suppose \( \bar{X} \) and \( X^* \) be any two acummalation points of \( \{X^k\} \), then we get \( \mathcal{A}(\sigma(\bar{X})) = \mathcal{A}(\sigma(X^*)) \) by (i) of this Theorem. From the boundedness of \( \{X^k\} \), then there exists a subsequence \( \{X^{k_i}, k_i \in \mathcal{N}^s\} \) of \( \{X^k, k \in \mathcal{N}^s\} \) such that \( \lim_{i \to +\infty} X^{k_i} = \bar{X} \). Then there exists \( J \in \mathbb{N} \), for any \( i \geq J \), it has that

\[
\| \sigma(X^{k_i})_{\mathcal{A}}(\sigma(\bar{X})) - \sigma(\bar{X})_{\mathcal{A}}(\sigma(\bar{X})) \| = 0,
\]

where
so the proof is completed. □

In the following, we establish the subsequence convergence result.

**Theorem 2** Under the Assumption 4, any accumulation point of \( \{X^k : k \in \mathbb{N}^s\} \) is a lifted stationary point of problem (2).

**Proof** By the boundedness of \( \{X^k : k \in \mathbb{N}^s\} \) from Corollary 1, letting \( \bar{X} \) be an accumulation point of \( \{X^k : k \in \mathbb{N}^s\} \), then there exists a subsequence \( \{X^{k_i}\}_{k_i \in \mathbb{N}^s} \) of \( \{X^k\}_{k \in \mathbb{N}^s} \) satisfying \( \lim_{i \to +\infty} X^{k_i} \to \bar{X} \).

Using the first order necessary optimality condition of (18), we get that

\[
0 \in \nabla \tilde{f}(Z^{k_i}, \mu^{k_i}) + h_{k_i} \mu^{-1}_{k_i} (X^{k_i} + 1 - Y^{k_i}) + \lambda \xi^{k_i} \quad \forall \xi^{k_i} \in \partial \Phi_{d^{k_i}}(X^{k_i} + 1),
\]

(38)

By virtue of \( \lim_{i \to +\infty} X^{k_i} = \bar{X} \) and the fact that elements in \( \{d^{k_i} : i \in \mathbb{N}\} \) are finite, there is a subsequence \( \{k_{i_j}\} \) of \( \{k_i\} \) and \( \bar{d} \in \mathcal{D}(\sigma(\bar{X})) \) such that \( d^{k_{i_j}} = \bar{d}, \forall j \in \mathbb{N} \). Further, from \( \lim_{j \to +\infty} \|X^{k_{i_j}+1} - X^{k_{i_j}}\| = 0, \) (35) and the upper semicontinuity of \( \partial \Phi_{d^{k_{i_j}}} \), we get that

\[
\lim_{j \to +\infty} X^{k_{i_j}+1} = \bar{X}, \quad \lim_{j \to +\infty} Z^{k_{i_j}} = \bar{X}, \quad \lim_{j \to +\infty} \mu^{-1}_{k_{i_j}} \|X^{k_{i_j}+1} - Y^{k_{i_j}}\| = 0,
\]

and

\[
\left\{ \lim_{j \to +\infty} \xi^{k_{i_j}} : \xi^{k_{i_j}} \in \partial \Phi_{d^{k_{i_j}}} (X^{k_{i_j}+1}) \right\} \subseteq \partial \Phi_{\bar{d}}(\bar{X}),
\]

(39)

taking \( k_i = k_{i_j} \) in (38) and letting \( j \to +\infty \), together (35), (38), (39) with the Definition (2)(iii), it deduces that there exist \( \bar{\psi} \in \partial f(\bar{X}) \) and \( \bar{\xi} \in \partial \Phi_{\bar{d}}(\bar{X}) \) satisfying the following

\[
0 = \bar{\psi} + \lambda \bar{\xi}.
\]

This completes the proof. □

**Remark 3** It’s well known that, in the vector case, the bregman distance is a generalized form of the Euclidean distance. One can refer to the literatures [4, 28] which have studied the theoretical analysis and methods based on Bregman distance. Similarly, in the matrix case, for any \( X, Y \in \mathbb{S}^n \), the Euclidean distance \( \frac{1}{2} \|X - Y\|^2 \) used in solving the subproblem (18) can be replaced by general Bregman regularization distance \( D_\phi(X, Y) \) where \( \phi : \mathbb{S}^n \to \mathbb{R} \) is a convex and continuously differentiable function which has been studied in [13, 19].
For any $X, Y \in \mathbb{S}^n$, the Bregman distance on matrix space is defined as

$$D_{\phi}(X, Y) := \phi(X) - \phi(Y) - tr(\nabla^T \phi(X)(X - Y)).$$

Specially, when $\phi(X) = \frac{1}{2}\|X\|^2$, we have $D_{\phi}(X, Y) = \frac{1}{2}\|X - Y\|^2$. If the function $\phi$ is strong convex with the strong convex modulus $\varrho$, then $D_{\phi}(X, Y) \geq \frac{\varrho}{2}\|X - Y\|^2$. In this case, the subproblem in (18) is generalized as follows:

$$\begin{align*}
Y^k &= X^k + \alpha_k(X^k - X^{k-1}), \\
Z^k &= X^k + \beta_k(X^k - X^{k-1}), \\
X^{k+1} &= \arg\min_{X \in \mathbb{S}^n} Q'_{d_k}(X, Y^k, Z^k, \mu_k)
\end{align*}$$

where $Q'_{d_k}(X, Y^k, Z^k, \mu_k) = \langle X, \nabla \tilde{f}(Z^k, \mu_k) \rangle - h_k \mu_k^{-1}(Y_k - X_k) + h_k \mu_k^{-1}D_{\phi}(X, X^k) + \lambda \Phi_{d_k}(X)$. In this case, the auxiliary sequence defined as (21) is descending under the Assumption 5 and the proof is presented in Appendix 5. The parameter constraints in Assumption 4 is generalized as follows:

**Assumption 5** for any $0 < \varepsilon \ll 1$, $\alpha_k \in [0, \frac{\varrho - 2^2 \varepsilon}{1 + 2^2 \varepsilon}), \beta_k \in [0, 1], \{h_k\}$ is nonincreasing and satisfies $h_0^{-1} \leq h_k^{-1} \leq \min\{\frac{\varrho - \alpha_k - (\alpha_k + \varepsilon) h_k}{(1 - \beta_k) L}, \frac{\alpha_k}{\beta_k L}\}$.

Particularly, if the convex function $\phi$ is a strong convex function with the modulus $\varrho$ satisfying $\varrho \geq (1 + 2^2)$, then $\alpha_k \in [0, 1), \beta_k \in [0, 1]$. This means that the parameter $\alpha_k$ and $\beta_k$ can be chosen as the sFISTA system (9) with fixted restart.

### 4 Numerical Experiments

In this section, we aim to verify the efficiency of the GIMSPG algorithm compared with the MSPG algorithm [34], FPCA [19], SVT [7] and VBMFL1 [37] on matrix completion problem, i.e.,

$$\min_{X \in \mathbb{R}^{m \times n}} \mathcal{F}(X) := \|P_\Omega(X - M)\|_1 + \lambda \cdot \text{rank}(X),$$

where $M \in \mathbb{R}^{m \times n}$ and $\Omega$ is the subset of the index set of the matrix $M$, $P_\Omega(\cdot)$ is the projection operator which projects onto the subspace of sparse matrices with nonzero entries confined to the index subset $\Omega$. The goal of this problem is to find the missing entries of the partially observed low-rank matrix $M$ based on the known elements $\{M_{ij} : i, j \in \Omega\}$. All the numerical experiments are carried out on 1.80GHz Core i5 PC with 12GB of RAM.

In the following, we denote $\bar{X}$ the output of the GIMSPG algorithm and use “Iter” “time” to present the number of iterations and the CPU time in
seconds respectively. We pick $X^0 = P_{\Omega}(M)$ as the initial point. The stopping standard is set as
\[
\frac{\|X^{k+1} - X^k\|}{\|X^k\|} \leq 10^{-4}.
\]
For the following $\ell_1$ loss function,
\[
f(x) = \|x - b\|_1, \quad \text{with } x, b \in \mathbb{R}^n,
\]
the smoothing function is defined as
\[
\tilde{f}(x, \mu) = \sum_{i=1}^{n} \tilde{\theta}(x_i - b_i, \mu) \quad \text{with} \quad \tilde{\theta}(s, \mu) = \begin{cases} |s| & \text{if } |s| > \mu, \\ \frac{s^2}{2 \mu} + \frac{\mu}{2} & \text{if } |s| \leq \mu. \end{cases}
\]
(42)

In the numerical simulation, the non-Gaussian noise model is set as the typical two-component Gaussian mixture model (GMM), the specific form of the probability density function is
\[
p_v(i) = (1 - c)N(0, \sigma_A^2) + cN(0, \sigma_B^2),
\]
(43)
where $N(0, \sigma_A^2)$ indicates as the general noise disturbance with variance $\sigma_A^2$ and $N(0, \sigma_B^2)$ represents the outliers with a large variance $\sigma_B^2$. The parameter $c$ trade off between them.

The parameters in the GIMSPG algorithm are defined as follows: $\tilde{L} = 1$, $\sigma = 0.9$, $\mu_0 = 1$, $\alpha = 1$, $\lambda = 20$. From Remark 2, we can take $\alpha_k = \frac{0.98 \beta_k}{1 + 2^\gamma \beta_k}$ and $h_k = \frac{(1 + 2^\gamma \beta_k) \tilde{L}}{0.98}$.

Moreover, in order to chose an appropriate value of $\beta_k$, we perform the experiments to test the influence of different values of $\beta_k$ on GIMSPG algorithm from the RMSE, the last value $\mu$ of smoothing factor and time. The details are presented in Figure 1. From it, we observe that the higher value of $\beta_k$, the used time is less, but RMSE and the last value of the smoothing factor is larger. For the sake of balance, in the following experiments, we chose $\beta_k = 0.4$ in Section 4.1 and 4.2. In Section 4.3, we chose $\beta_k = 0.3$.

### 4.1 Matrix completion on the random data

In this subsection, we perform the numerical experiments on random generated data. In order to make a fair comparison, our results are averaged over 20 independent tests. Besides, the way of the generated data is same as [34], in detail, we first randomly generate matrices $M_L \in \mathbb{R}^{m \times r}$ and $M_R \in \mathbb{R}^{n \times r}$ with i.i.d. standard Gaussian entries and let $M = M_L M_R^T$. We then sample a subset $\Omega$ with sample ration $sr$ uniformly at random, where $sr = \frac{|\Omega|}{mn}$. In the following experiments, the parameters in the GMM noise are set as $\sigma_A^2 = 0.0001$, $\sigma_B^2 = 0.1$, $c = 0.1$. The rank of the matrix $M$ is set as $r = 30$ and the sampling ration $sr$ are given three different cases $sr = 0.2$, $sr = 0.6$, $sr = 0.8$. The size of the square matrix $m$ increases from 100 to 200 with increment 10.
The root-mean-square error (RMSE) is defined as an error estimation criterion,

$$\text{RMSE} := \sqrt{\frac{\|X - M\|^2}{mn}}. \quad (44)$$

The Figure 2 presents the last value of the objective function and the last value $\mu$ of smoothing factor for GIMSPG algorithm and MSPG algorithm under different matrix size and $sr = 0.2$, $sr = 0.6$, $sr = 0.8$. It can be easily seen that the last value of the objective function value of GIMSPG algorithm is lower than the MSPG algorithm except $sr = 0.2$. Besides, the last value $\mu$ of smoothing factor of GIMSPG algorithm are far lower than the MSPG algorithm, that is, the model in GIMSPG algorithm can be better approximation to the original problem. Moreover, we illustrate the efficiency of GIMSPG algorithm compared with MSPG, FPCA, SVT and VBMFL1 from RMSE and the running time $T$ two aspects under different matrix size for $sr = 0.2$, $sr = 0.6$, $sr = 0.8$ in Tables 1,2,3. From them, we see that the running time of GIMSPG algorithm is the least one in these algorithms. We also observe that as the dimension increases, so does the time. Compared with MSPG algorithm, when the sample ratio is lower, the RMSE of GIMSPG algorithm is better. And compared with the other algorithms, the GIMSPG algorithm always has the least value of RMSE.

Fig. 1: The time, the last value of $\mu$ and RMSE for $sr = 0.6$ and $m, n = 150$

4.2 Image inpainting

In this subsection, we aim to illustrate the efficiency of GIMSPG algorithm by solving a grayscale inpainting problem with non-Gaussian noise. The image inpainting problem is to fill the missing pixel values of the image at given pixel locations. In fact, the grayscale image can be seen as the matrix, and the image inpainting problem can be represented as the matrix completion problem if the matrix satisfies the property of the low rank. In our experiments, two
Fig. 2: The last value of the objective function (2) and smoothing factor \( \mu \) for 
\( sr = 0.2, 0.6, 0.8 \)

Table 1: Numerical results of the random matrix problem for \( sr = 0.8 \)

| \( m \) | RMSE | T | GIMSPG | MSPG | VBMFL1 | FPCA | SVT | GIMSPG | MSPG | VBMFL1 | FPCA | SVT |
|-------|------|---|--------|------|--------|------|-----|--------|------|--------|------|-----|
| 100   | 0.030| 0.031 | 0.051 | 0.551 | 0.242 | 0.32 | 1.01 | 3.33 | 5.13 | 3.70   |
| 110   | 0.029| 0.030 | 0.045 | 0.472 | 0.076 | 0.33 | 1.21 | 3.53 | 4.97 | 4.03   |
| 120   | 0.029| 0.029 | 0.038 | 0.352 | 0.049 | 0.33 | 1.26 | 3.56 | 5.55 | 4.31   |
| 130   | 0.028| 0.029 | 0.036 | 0.289 | 0.036 | 0.45 | 1.46 | 3.84 | 6.44 | 4.32   |
| 140   | 0.028| 0.028 | 0.035 | 0.284 | 0.028 | 0.51 | 1.78 | 3.56 | 6.72 | 4.51   |
| 150   | 0.028| 0.028 | 0.034 | 0.271 | 0.026 | 0.55 | 1.85 | 3.58 | 6.72 | 4.75   |
| 160   | 0.028| 0.028 | 0.037 | 0.271 | 0.023 | 0.62 | 2.21 | 4.06 | 7.74 | 4.52   |
| 170   | 0.028| 0.027 | 0.034 | 0.230 | 0.022 | 0.78 | 2.66 | 4.30 | 8.28 | 4.70   |
| 180   | 0.027| 0.026 | 0.032 | 0.209 | 0.022 | 0.84 | 2.89 | 4.16 | 8.86 | 4.41   |
| 190   | 0.025| 0.026 | 0.031 | 0.201 | 0.021 | 0.90 | 3.19 | 4.59 | 9.72 | 4.31   |
| 200   | 0.025| 0.025 | 0.030 | 0.196 | 0.020 | 1.04 | 3.47 | 4.94 | 10.03| 4.84   |

different grayscale images are considered which are from the USC-SIPI image database \(^1\). In detail, they are “Chart” and “Ruler”. “Chart” is an image with 256 \( \times \) 256 pixels and rank \( r = 185 \). “Ruler” is an image with 512 \( \times \) 512 with rank \( r = 67 \). To evaluate the performance of the GIMSPG algorithm under the non-Gaussian noise, the peak signal-to-noise ratio (PSNR) is considered as an evaluation metric which is defined as follows:

\[
\text{PSNR} := 10 \log_{10} \left( \frac{mn}{\| X - M \|_F^2} \right).
\]

We solve the image inpainting problem under three different sample ratios \( sr = 0.2, sr = 0.6, sr = 0.8 \). From the results in Table 4, we observe that the last value \( \mu \) of smoothing factor of GIMSPG algorithm is less than MSPG algorithm in each case. The used time \( T \) of GIMSPG is the least in these

\(^1\)http://sipi.usc.edu/database/.
Table 2: Numerical results of the random matrix problem for $sr = 0.6$

| m | RMSE | $T$ |
|---|---|---|
|   | GIMSPG | MSPG | VBMFL1 | FPCA | SVT | GIMSPG | MSPG | VBMFL1 | FPCA | SVT |
| 100 | 0.036 | 0.036 | 0.124 | 1.640 | 1.860 | 0.42 | 1.40 | 5.11 | 4.71 | 5.23 |
| 110 | 0.033 | 0.033 | 0.080 | 1.360 | 1.380 | 0.57 | 1.50 | 5.19 | 5.09 | 4.84 |
| 120 | 0.033 | 0.033 | 0.062 | 0.870 | 1.324 | 0.85 | 1.70 | 5.38 | 5.94 | 5.70 |
| 130 | 0.031 | 0.031 | 0.054 | 0.610 | 0.972 | 0.91 | 1.88 | 5.44 | 6.05 | 8.41 |
| 140 | 0.030 | 0.051 | 0.500 | 0.707 | 1.09 | 2.10 | 5.52 | 6.67 | 7.27 |
| 150 | 0.029 | 0.047 | 0.437 | 0.525 | 1.87 | 2.32 | 6.21 | 7.17 | 7.47 |
| 160 | 0.028 | 0.042 | 0.396 | 0.368 | 1.90 | 2.41 | 6.36 | 8.09 | 8.55 |
| 170 | 0.027 | 0.041 | 0.357 | 0.219 | 1.99 | 2.99 | 6.59 | 8.64 | 9.01 |
| 180 | 0.027 | 0.038 | 0.326 | 0.163 | 0.84 | 3.10 | 7.16 | 10.06 | 9.03 |
| 190 | 0.026 | 0.035 | 0.309 | 0.127 | 2.14 | 4.11 | 7.51 | 10.23 | 10.06 |
| 200 | 0.026 | 0.034 | 0.289 | 0.068 | 2.38 | 4.37 | 8.14 | 10.55 | 10.84 |

Table 3: Numerical results of the random matrix problem for $sr = 0.2$

| m | RMSE | $T$ |
|---|---|---|
|   | GIMSPG | MSPG | VBMFL1 | FPCA | SVT | GIMSPG | MSPG | VBMFL1 | FPCA | SVT |
| 100 | 0.070 | 0.075 | 5.649 | 5.770 | 6.536 | 0.69 | 1.40 | 13.42 | 2.38 | 8.17 |
| 110 | 0.071 | 0.074 | 5.496 | 5.660 | 6.497 | 0.80 | 1.50 | 16.59 | 2.64 | 9.95 |
| 120 | 0.069 | 0.073 | 5.486 | 5.540 | 6.563 | 0.94 | 1.60 | 17.38 | 3.27 | 10.41 |
| 130 | 0.066 | 0.073 | 5.456 | 5.300 | 6.618 | 1.08 | 1.70 | 18.33 | 3.27 | 13.51 |
| 140 | 0.066 | 0.075 | 5.446 | 5.209 | 6.255 | 1.50 | 1.98 | 20.84 | 3.73 | 15.41 |
| 150 | 0.066 | 0.072 | 5.442 | 5.270 | 6.230 | 1.66 | 2.19 | 21.84 | 4.06 | 15.84 |
| 160 | 0.065 | 0.071 | 5.508 | 5.230 | 6.210 | 1.80 | 2.57 | 25.67 | 3.99 | 27.09 |
| 170 | 0.065 | 0.068 | 5.309 | 5.050 | 6.194 | 2.30 | 2.80 | 26.77 | 4.31 | 23.89 |
| 180 | 0.064 | 0.069 | 5.212 | 5.010 | 5.984 | 2.60 | 3.39 | 29.16 | 5.25 | 29.01 |
| 190 | 0.062 | 0.068 | 5.091 | 4.970 | 5.020 | 2.63 | 3.58 | 29.98 | 5.98 | 29.21 |
| 200 | 0.062 | 0.067 | 4.239 | 4.680 | 5.064 | 3.20 | 4.82 | 32.83 | 6.66 | 29.77 |

algorithms. The PSNR of GIMSPG algorithm is higher than MSPG for $sr = 0.2$ and obviously higher than algorithm in all the cases. From Figures 3, 4, one can see that the recoverability of GIMSPG algorithm is slightly better than that MSPG algorithm especially in the lower sample rate and apparently better than others.

4.3 MRI Volume Dataset

In this subsection, we perform the experiments on the MRI image which is of size $217 \times 181$ with 181 slices and we select the 38th slice for the experiments. The two sampling ratios $sr = 0.6$, $sr = 0.8$ are considered. The parameters in GMM noise are set as $\sigma_A^2 = 0.0001$, $\sigma_B^2 = 0.1$ and $c = 0.1$.

From Figure 5 and Table 5, we see that the recoverability of VBMFL1 is relatively better but it spends more time. Particularly, we notice that the PSNR of GIMSPG algorithm is higher than other algorithms except the VBMFL1 in
Table 4: Numerical results on the image data for $sr = 0.2, sr = 0.6, sr = 0.8$

| Image  | sr  | PSNR | T       | $\mu$   |
|--------|-----|------|---------|---------|
|        |     | GIMSPG | MSPG | VBMFL1 | FPCA | SVT | GIMSPG | MSPG | VBMFL1 | FPCA | SVT | MSPGE | MSPG |
| chart  | 0.2 | 18.04 | 17.16 | 15.96  | 15.60 | 9.16 | 9.78  | 18.10 | 18.67  | 0.53  | 2.72 | 0.022 | 0.062 |
|        | 0.6 | 29.13 | 29.01 | 19.80  | 18.80 | 10.49 | 2.19  | 12.59 | 33.10  | 14.16 | 2.78 | 0.035 | 0.131 |
|        | 0.8 | 30.84 | 30.18 | 20.63  | 19.50 | 10.65 | 1.55  | 3.16  | 47.94  | 13.64 | 3.58 | 0.107 | 0.234 |
| ruler  | 0.2 | 21.17 | 20.74 | 20.10  | 19.90 | 5.21  | 25.45 | 55.76 | 5.23   | 26.25 | 5.42 | 0.020 | 0.105 |
|        | 0.6 | 27.02 | 27.04 | 19.80  | 18.80 | 5.22  | 11.50 | 12.86 | 19.67  | 54.91 | 6.31 | 0.044 | 0.238 |
|        | 0.8 | 30.01 | 30.39 | 20.50  | 20.20 | 5.24  | 7.11  | 14.83 | 49.91  | 49.98 | 7.09 | 0.093 | 0.273 |

Table 5: Numerical results of the MRI volume dataset for $sr = 0.6, sr = 0.8$

| Image  | sr  | PSNR | T       | $\mu$   |
|--------|-----|------|---------|---------|
|        |     | GIMSPG | MSPG | VBMFL1 | FPCA | SVT | GIMSPG | MSPG | VBMFL1 | FPCA | SVT | MSPGE | MSPG |
| MRI    | 0.6 | 29.13 | 29.01 | 29.58  | 26.70 | 13.56 | 2.19  | 12.59 | 32.91  | 10.67 | 2.22 | 0.035 | 0.131 |
|        | 0.8 | 33.06 | 32.99 | 30.33  | 27.20 | 13.59 | 1.79  | 10.57 | 26.61  | 11.58 | 2.44 | 0.029 | 0.137 |
Fig. 3: Results of “Chart” recovery. The first column is the sample images with \(sr = 0.2, 0.6, 0.8\). The rest columns are the recovery images by the corresponding algorithms.

Fig. 4: Results of “Ruler” recovery. The first column is the sample images with \(sr = 0.2, 0.6, 0.8\). The rest columns are the recovery images by the corresponding algorithms.

Table 5. Besides, the used time \(T\) of GIMSPG is the least one. Compared with MSPG, the last value \(\mu\) of smoothing factor of GIMSPG algorithm is lower.

5 Conclusions

This paper mainly proposes the GIMSPG algorithm for the exact continuous model [34] of matrix rank minimization problem and discuss its convergence.
under the flexible parameters condition. It is shown that the singular values of the accumulation point have a common support set and the nonzero elements have a unified lower bound. Further, the zero singular values of the accumulation point can be achieved within finite iterations. Moreover, we prove that any accumulation point of the sequence generated by GIMSPG algorithm is a lifted stationary point of the continuous relaxation model under the flexible parameter constraint. Further, we generalize Euclidean distance to the Bregman distance in the GIMSPG algorithm and . Under the appropriate parameters, the range of the extrapolation parameters is large enough which also includes the sFISTA system (9) with fixed restart. At last, the superiority of the GIMSPG algorithm is verified on random data and real image data respectively.

### Appendix A

**Lemma 6** For any $k \in \mathbb{N}$, suppose $\{X^k\}$ be the sequence generated by GIMSPG, we have

$$
H_{\delta_{k+1}}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) - H_{\delta_k}(X^k, X^{k-1}, \mu_k, \mu_{k-1})
\leq ((\tilde{L} - \varrho h_k + \frac{h_k \alpha_k - \beta_k \tilde{L}}{2})\mu_k^{-1} + \frac{\alpha_k h_k \mu_{k+1}^{-1}}{2})\|X^{k+1} - X^k\|^2
+ \frac{\tilde{L} \beta_k (\beta_k - 1)}{2} \mu_k^{-1} \|X^k - X^{k-1}\|^2.
$$

When the parameters $\alpha_k, \beta_k, h_k$ satisfying Assumption 5, it holds that

$$
H_{\delta_{k+1}}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) - H_{\delta_k}(X^k, X^{k-1}, \mu_k, \mu_{k-1}) \leq -\frac{\varepsilon h_k}{2} \mu_{k+1}^{-1} \|X^{k+1} - X^k\|^2
$$

(A1)
Moreover, the sequence \( H_{\delta_{k+1}}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) \) is a nonincreasing sequence.

**Proof** By (40), we have

\[
\langle X^{k+1}, \nabla \tilde{f}(Z^k, \mu_k) - h_k \mu_k^{-1}(Y_k - X_k) \rangle + h_k \mu_k^{-1} D_\phi(X^{k+1}, X^k) + \lambda \Phi_d^k(X^{k+1}) \\
\leq \langle X^k, \nabla \tilde{f}(Z^k, \mu_k) - h_k \mu_k^{-1}(Y_k - X_k) \rangle + \lambda \Phi_d^k(X^k). \tag{A2}
\]

Since \( \nabla \tilde{f}(\cdot, \mu_k) \) has Lipschitz constant \( \tilde{L}\mu_k^{-1} \), it follows from the Definition 2(v) that

\[
\tilde{f}(X^{k+1}, \mu_k) \leq \tilde{f}(Z^k, \mu_k) + \langle X^{k+1} - Z^k, \nabla \tilde{f}(Z^k, \mu_k) \rangle + \frac{1}{2} \tilde{L}\mu_k^{-1} \|X^{k+1} - Z^k\|^2. \tag{A3}
\]

Moreover, \( \tilde{f}(X^k, \mu_k) \) is convex with respect to \( X^k \), we have

\[
\tilde{f}(Z^k, \mu_k) + \langle X^k - Z^k, \nabla \tilde{f}(Z^k, \mu_k) \rangle \leq \tilde{f}(X^k, \mu_k). \tag{A4}
\]

Combining (A2), (A3) and (A4), we have

\[
\tilde{F}^d_k(X^{k+1}, \mu_k) - \tilde{F}^d_k(X^k, \mu_k) \leq -h_k \mu_k^{-1} \langle Y_k - X^k, X^k - X^{k+1} \rangle - h_k \mu_k^{-1} D_\phi(X^{k+1}, X^k) + \frac{\tilde{L}\mu_k^{-1}}{2} \|X^{k+1} - Z^k\|^2.
\]

Denote \( \Delta_k := X^k - X^{k-1} \), then it has \( \alpha_k \Delta_k := Y^k - X^k \), \( \beta_k \Delta_k := Z^k - X^k \), \( \beta_k \Delta_k - \Delta_{k+1} := Z^k - X^{k+1} \). That is,

\[
\tilde{F}^d_k(X^{k+1}, \mu_k) - \tilde{F}^d_k(X^k, \mu_k) \leq h_k \mu_k^{-1} \langle \alpha_k \Delta_k, \Delta_{k+1} \rangle + \frac{\tilde{L}\mu_k^{-1}}{2} \|\beta_k \Delta_k - \Delta_{k+1}\|^2 - \frac{1}{2} gh_k \mu_k^{-1} \|\Delta_{k+1}\|^2 \\
= \frac{(\tilde{L} - gh_k)\mu_k^{-1}}{2} \|\Delta_{k+1}\|^2 + \frac{\tilde{L}}{2} \beta_k^2 \mu_k^{-1} \|\Delta_k\|^2 + (h_k \alpha_k - \beta_k \tilde{L})\mu_k^{-1} \langle \Delta_k, \Delta_{k+1} \rangle \\
\leq \frac{(\tilde{L} - gh_k)}{2} + \frac{h_k \alpha_k - \beta_k \tilde{L}}{2} \mu_k^{-1} \|\Delta_{k+1}\|^2 + \frac{(\tilde{L} \beta_k^2 + h_k \alpha_k - \beta_k \tilde{L})}{2} \mu_k^{-1} \|\Delta_k\|^2 \tag{A6}
\]

where the second inequality comes from the Cauchy-Schwartz inequality.

Letting \( d^k = dX^k \) and according to \( \tilde{F}^d_k(X^{k+1}, \mu_k) \geq \tilde{F}(X^{k+1}, \mu_k) \), we have

\[
\tilde{F}(X^{k+1}, \mu_k) + \frac{\alpha_k + h_k \mu_k^{-1}}{2} \|\Delta_{k+1}\|^2 - (\tilde{F}(X^k, \mu_k) + \frac{\alpha_k h_k \mu_k^{-1}}{2} \|\Delta_k\|^2)
\]
\[
\leq \left( \left( \frac{\bar{L} - \varrho h_k}{2} + \frac{h_k \alpha_k - \beta_k \bar{L}}{2} \right) \mu_k^{-1} + \frac{\alpha_{k+1} h_{k+1} \mu_{k+1}^{-1}}{2} \right) \| \Delta_{k+1} \|^2 \\
+ \left( \left( \frac{\bar{L}^2 + h_k \alpha_k - \beta_k \bar{L}}{2} \right) \mu_k^{-1} - \frac{\alpha_k h_k \mu_k^{-1}}{2} \right) \| \Delta_k \|^2
\]
(A7)

By the Definition 2(iv) and the nonincreasing of \( \mu_k \), we easily have that
\[
\bar{F}(X^k, \mu_k) \leq \bar{F}(X^k, \mu_{k-1}) + \kappa(\mu_{k-1} - \mu_k).
\]
(A8)

Together (A7) with (A8), we get
\[
H_{\delta_{k+1}}(X^{k+1}, X^k, \mu_{k+1}, \mu_k) - H_{\delta_k}(X^k, X^{k-1}, \mu_k, \mu_{k-1}) \\
\leq \left( \left( \frac{\bar{L} - \varrho h_k}{2} + \frac{h_k \alpha_k - \beta_k \bar{L}}{2} \right) \mu_k^{-1} + \frac{\alpha_{k+1} h_{k+1} \mu_{k+1}^{-1}}{2} \right) \| \Delta_{k+1} \|^2 + \frac{\bar{L}(\beta_k^2 - \beta_k)}{2} \mu_k^{-1} \| \Delta_k \|^2
\]

According to the parameters constraint in Assumption 5, we easily have that
\[
(\bar{L} - \varrho h_k + h_k \alpha_k - \beta_k \bar{L}) \mu_k^{-1} + \alpha_k h_k \mu_k^{-1} < -\epsilon h_k \mu_k^{-1}, \beta_k^2 - \beta_k \leq 0.
\]
Hence, the inequality (A1) holds. So the desired result is obtained. \(\square\)

References

[1] Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Mach. Learn. 73, 243-272 (2008)

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

[3] Bian, W.: Smoothing accelerated algorithm for constrained nonsmooth convex optimization problems (in Chinese). Sci. Sin. Math. 50, 1651–1666 (2020)

[4] 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)

[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] Bouwmans,T., Zahzah, E., H.: Robust PCA via Principal Component Pursuit: A Review for a Comparative Evaluation in Video Surveillance. Comput Vis. Image Underst. 122, 22-34 (2014)

[7] Cai, J., Candes, E., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. Optim. 20, 1956-1982 (2008)
[8] 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)

[9] Chen, X.: Smoothing methods for nonsmooth, nonconvex minimization. Math. Program. 134, 71-99 (2012)

[10] Fornasier, M., Rauhut, H., Ward, R.: Low-rank matrix recovery via iteratively reweighted least squares minimization. SIAM J. Optim. 21, 1614-1640 (2011)

[11] He, Y., Wang, F., Li, Y., Qin, J., Chen, B.: Robust matrix completion via maximum correntropy criterion and half-quadratic optimization. IEEE T. Signal Proces. 68, 181-195 (2020)

[12] Ji, S., Ye, J.: An accelerated gradient method for trace norm minimization. Proceedings of the 26th annual international conference on machine learning. 457-464 (2009)

[13] Kulis, B., Sustik, M., A., Dhillon, I., S.: Low-Rank Kernel Learning with Bregman Matrix Divergences. J. Mach. Learn. Res. 10, 2009

[14] Lai, M., Xu, Y., Yin, W.: Improved iteratively reweighted least squares for unconstrained smoothed q minimization. SIAM J. Numer. Anal. 51, 927-957 (2013)

[15] Lewis, A., S., Sendov, H., S.: Nonsmooth analysis of singular values. Part II: Applications. Set-Valued Anal. 13, 243-264 (2005)

[16] Li, W., Bian, W., Toh, K., C.: DC algorithms for a class of sparse group ℓ_0 regularized optimization problems (2021). arXiv:2109.05251

[17] Liang, J., Schonlieb, C., B.: “Faster FISTA.” European Signal Processing Conference. (2018)

[18] Lu, Z., Zhang, Y., Lu, J.: ℓ_p Regularized low-rank approximation via iterative reweighted singular value minimization. Comput. Optim. Appl. 68, 619-642 (2017)

[19] Ma, S., Goldfarb, D., Chen, L.: Fixed point and Bregman iterative methods for matrix rank minimization. Math. Program. 128, 321-353 (2011)

[20] Ma, T., H., Lou, Y., Huang, T., Z.: Truncated ℓ_{1-2} models for sparse recovery and rank minimization. SIAM J. Imaging Sci. 10, 1346-1380 (2017)
[21] Mesbahi, M., Papavassilopoulos, G., P.: On the rank minimization problem over a positive semidefinite linear matrix inequality. IEEE T. Automat. Contr. 42, 239-243(1997)

[22] Nesterov, Y.: Introductory Lectures on Convex Programming. Kluwer Academic Publisher, Dordrecht (2004)

[23] Nesterov Y. Gradient methods for minimizing composite functions[J]. Mathematical programming, 2013, 140(1): 125-161.

[24] Nesterov, Y.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Dokl. Akad. Nauk. SSSR 269, 543-547 (1983)

[25] O’Donoghue, B., Candes, E.J.: Adaptive restart for accelerated gradient schemes. Found. Comput. Math. 15, 715-732 (2015)

[26] Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Mach. Learn. 1, 127-239 (2014)

[27] Recht, B., Fazel, M., Parrilo, P.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. Siam Rev. 52, 471-501 (2010)

[28] Teboulle, M.: A simplified view of first order methods for optimization. Math. Program. 170, 67-96 (2018)

[29] Toh, K., C., Yun, S.: An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. Pac. J. Optim. 6, 615-640 (2010)

[30] Wu, F., Bian, W., Xue, X.: Smoothing fast iterative hard thresholding algorithm for $\ell_0$ regularized nonsmooth convex regression problem (2021). arXiv:2104.13107

[31] Wu, Z., Li, C., Li, M., Lim, A.: Inertial proximal gradient methods with Bregman regularization for a class of nonconvex optimization problems. J. Global Optim. 79, 617-644 (2021)

[32] Wu, Z., Li, M.: General inertial proximal gradient method for a class of nonconvex nonsmooth optimization problems. Comput. Optim. Appl. 73, 129-158(2019)

[33] Xu, H., Caramanis, C., Sanghavi, S.: Robust PCA via outlier pursuit. IEEE Trans. Inf. Theory. 58, 3047-3064 (2012)

[34] Yu, Q., Zhang, X.: A smoothing proximal gradient algorithm for matrix rank minimization problem. Comput. Optim. Appl. 1-20 (2022)
[35] Zhang, C., Chen, X.: Smoothing projected gradient method and its applica-
tion to stochastic linear complementarity problems. SIAM J. Optim. 
20, 627-649(2009)

[36] Zhang, J., Yang, X., Li, G., Zhang, K. The smoothing proximal gradi-
ent algorithm with extrapolation for the relaxation of $l_0$ regularization 
problem (2021). arXiv:2112.01114.

[37] Zhao, Q., Meng, D., Xu, Z., Yan, Y.: $L_1$-norm low-rank matrix factor-
ization by variational Bayesian method. IEEE T. Neur. Net. Lear. 26, 
825-839(2015)

[38] Zheng, Y., Liu, G., Sugimoto, S., Yan, S., Okutomi, M.: Practical low-rank 
matrix approximation under robust $L_1$-norm. In 2012 IEEE Conference 
on Computer Vision and Pattern Recognition (pp. 1410-1417). IEEE