FIRST-ORDER ALGORITHMS FOR A CLASS OF FRACTIONAL OPTIMIZATION PROBLEMS *

NA ZHANG † AND QIA LI ‡

Abstract. We consider in this paper a class of single-ratio fractional minimization problems, in which the numerator part of the objective is the sum of a nonsmooth nonconvex function and a smooth nonconvex function while the denominator part is a nonsmooth convex function. Besides, the three functions involved in the objective are all nonnegative. To the best of our knowledge, this class of problems has seldom been carefully investigated in the literature and existing methods in general fractional optimization are not suitable for solving this problem. In this work, we first derive its first-order necessary optimality condition, by using the first-order operators of the three functions involved. Then we develop first-order algorithms, namely, the proximity-gradient-subgradient algorithm (PGSA), PGSA with monotone line search (PGSA_{ML}) and PGSA with nonmonotone line search (PGSA_{NL}). It is shown that any accumulation point of the sequence generated by them is a critical point of the problem under mild assumptions. Moreover, we establish global convergence of the sequence generated by PGSA or PGSA_{ML} and analyze its convergence rate, by further assuming the local Lipschitz continuity of the nonsmooth function in the numerator part, the smoothness of the denominator part and the Kurdyka-Lojasiewicz property of the objective. The proposed algorithms are applied to the sparse generalized eigenvalue problem associated with a pair of symmetric positive semidefinite matrices and the corresponding convergence results are obtained according to their general convergence theorems. We perform some preliminary numerical experiments to demonstrate the efficiency of the proposed algorithms.

Key words. fractional optimization, first-order algorithms, proximity algorithms

AMS subject classifications. 90C26, 90C30, 65K05

1. Introduction. A fractional optimization problem is the problem which minimizes or maximizes an objective involving one or several ratios of functions. Fractional optimization problems arise from various applications in many fields, such as economics [17, 29], wireless communication [33, 39, 40], artificial intelligence [4, 14] and so on. It has been extensively studied in the literature four categories of fractional optimization problems concerning minimizing a single ratio of two functions over a closed convex set. They are named according to the functions in the numerator and denominator: linear or quadratic fractional problems if both functions are linear or quadratic; convex-concave fractional problems if the numerator is convex and the denominator is concave; convex-convex fractional problems if both functions are convex. We refer the readers to [31, 32, 34], for an overview on the single-ratio fractional optimization.

In this paper, we consider a class of single-ratio fractional minimization problems in the form of

\[
\min \left\{ \frac{f(x) + h(x)}{g(x)} : x \in \Omega \right\},
\]

where \( f : \mathbb{R}^n \to \mathbb{R}_+ := [0, +\infty) \) is proper, lower semicontinuous on \( \mathbb{R}^n \) and continuous

---

*This research is supported in part by the Natural Science Foundation of China under grants 11971499 and 11701189, and by the Fundamental Research Funds for the Central Universities of China.

†Department of Applied Mathematics, College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, P. R. China.

‡Guangdong Province Key Laboratory of Computational Science, School of Data and Computer Sciences, Sun Yat-sen University, Guangzhou 510275, P. R. China (liqia@mail.sysu.edu.cn). Questions, comments, or corrections to this document may be directed to that email address.
on its domain, \( g : \mathbb{R}^n \to \mathbb{R}_+ := [0, +\infty) \) is continuous and convex on \( \mathbb{R}^n \), \( h : \mathbb{R}^n \to \mathbb{R}_+ \) is Lipschitz differentiable with a Lipschitz constant \( L > 0 \), and the set \( \Omega := \{ x \in \mathbb{R}^n : g(x) \neq 0 \} \) is nonempty. Moreover, problem (1.1) is assumed to have at least one optimal solution. It is obvious that both \( f \) and \( h \) are possibly nonconvex, while \( f \) and \( g \) can be nonsmooth. Problem (1.1) does not belong to any of the four categories of fractional minimization problems aforementioned. This class of optimization problems subsumes a wide range of application models, e.g., the sparse generalized eigenvalue problem (SGEP)\[6, 35\].

Now we turn to the algorithmic aspect of problem (1.1). To the best of our knowledge, this class of problems has seldom been studied in the literature and existing methods in general fractional optimization are not suitable for solving problem (1.1). Global optimization methods, e.g., branch and bound algorithms \[7, 18\], play an important role in directly solving fractional optimization problems. However, the variable \( x \) of problem (1.1) is usually high dimensional in modern machine learning models. Thus, it is not practical to apply global optimization methods due to their expensive computational cost. For single fractional optimization problems, the variable transformation and parametric approach have been proposed to overcome the algorithmic difficulties caused by the ratio involved. In \[9\], Charnes and Cooper first suggest a variable transformation by which a linear fractional problem is reduced to a linear program. In fact, with the help of that variable transformation, any convex-concave fractional minimization problem can be equivalently reduced to a convex minimization problem. Since problem (1.1) is not a convex-concave fractional minimization problem, through the variable transformation it remains nonconvex and in general difficult to solve. Hence, the variable transformation approach is not suitable for dealing with problem (1.1). Another widely used method for fractional optimization is the parametric approach, which takes good advantage of the relationship between a fractional problem and its associated parametric problem \[13, 16\]. Many efficient algorithms have been developed based on the parametric approach, see, for example, \[13, 15, 25, 27, 28\]. When they are applied to problem (1.1), most of these algorithms require to solve in each iteration a parametric subproblem in the form of

\[
\min \{ f(x) + h(x) - cg(x) : x \in \Omega \},
\]

where \( c \in \mathbb{R} \) is determined by the previous iteration. We notice that this parametric subproblem is nonconvex and nonsmooth, then in general it is hard to obtain a global optimal solution of problem (1.2). As a consequence, we can not directly utilize existing algorithms based on the parametric approach for problem (1.1).

In this work, we propose new iterative numerical algorithms for solving problem (1.1). In each iteration of the proposed algorithms, we mainly make use of the proximity operator of \( f \), the gradient of \( h \) and the subgradient of \( g \) at the current iterate. When the above first-order operations are easy to compute, our algorithms perform efficiently. Our contributions are summarized below.

- By Fréchet subdifferentials of \( f \), \( g \) and the gradient of \( h \), we derive a first-order necessary optimality condition for problem (1.1) and thus introduce the definition of its critical points.
- Based on the first-order optimality condition aforementioned, we develop for problem (1.1) three first-order numerical algorithms, namely, proximity-gradient-subgradient algorithm (PGSA), PGSA with monotone line search (PGSA-ML) and PGSA with nonmonotone line search (PGSA-NL). Under mild assumptions on problem (1.1), we prove that any accumulation point of
the sequence generated by any of the proposed algorithms is a critical point of problem (1.1). In addition, we show global convergence of the entire sequence generated by PGSA or PGSA_{ML}, by further assuming that $f$ is locally Lipschitz in its domain, $g$ is differentiable with a locally Lipschitz continuous gradient and the objective in problem (1.1) satisfies the Kurdyka-Lojasiewicz property. The convergence rate of PGSA and PGSA_{ML} are also estimated according to the Kurdyka-Lojasiewicz property.

- We identify SGEP associated with a pair of symmetric positive semidefinite matrices as a special case of problem (1.1) and apply the proposed algorithms to SGEP. We obtain the convergence results of the proposed algorithms for SGEP, by validating SGEP satisfies all the conditions needed in their general convergence theorems. In addition, we show that the sequence generated by PGSA or PGSA_{ML} converges R-linearly, if the corresponding critical point to which the sequence converges is a local minimizer of SGEP.

The remaining part of this paper is organized as follows. In Section 2, we introduce notation and some necessary preliminaries. Section 3 is devoted to a study of first-order necessary optimality conditions for problem (1.1). In Section 4, we propose the PGSA and give its convergence analysis. In Section 5, we develop PGSA with line search (PGSA_{L}), including PGSA_{ML} and PGSA_{NL}, and study their convergence property. We specify in Section 6 the proposed algorithms and convergence results obtained in Sections 4 and 5 to the sparse generalized eigenvalue problem. In Section 7, some numerical results are presented to demonstrate the efficiency of the proposed algorithms. Finally, we conclude this paper in the last section.

2. Notation and preliminaries. We start by our preferred notations. We denote by $\mathbb{N}$ the set of nonnegative integers. For a positive integer $n$, we let $\mathbb{N}_n := \{1, 2, \ldots, n\}$ and $\mathbf{0}_n$ be the $n$-dimensional vector. For $x \in \mathbb{R}$, let $[x]_+ := \max\{0, x\}$. By $\mathbb{S}^n$ (resp., $\mathbb{S}^n_+$) we denote the set of all $n \times n$ symmetric positive semidefinite (resp., definite) matrices. Given $H \in \mathbb{S}^n$, the weighted inner product of $x, y \in \mathbb{R}^n$ is defined by $\langle x, y \rangle_H := \langle x, Hy \rangle$ and the weighted $\ell_2$-norm of $x \in \mathbb{R}^n$ is defined by $\|x\|_H := \sqrt{\langle x, x \rangle_H}$. For an $n \times n$ matrix $A$, we denote by $\|A\|_2$ the matrix 2-norm of $A$. For $\Lambda \subseteq \mathbb{N}_n$, let $|\Lambda|$ be the number of elements in $\Lambda$. We denote by $x_{\Lambda} \in \mathbb{R}^{|\Lambda|}$ the sub-vector of $x$ whose indices are restricted to $\Lambda$. We also denote by $A_{\Lambda \Lambda}$ the $(|\Lambda| \times |\Lambda|)$ sub-matrix formed from picking the rows and columns of $A$ indexed by $\Lambda$. For a function $\varphi : \mathbb{R}^n \to \mathbb{R}$ and $t \in \mathbb{R}$, let $\text{lev}(\varphi, t) := \{x \in \mathbb{R}^n : \varphi(x) \leq t\}$.

For $x \in \mathbb{R}^n$, let $\text{supp}(x)$ be the support of $x$, that is, $\text{supp}(x) := \{i \in \mathbb{N}_n : x_i \neq 0\}$. Given $\delta > 0$, we let $B(x, \delta) := \{z \in \mathbb{R}^n : \|z - x\|_2 < \delta\}$ and $U(x, \delta) := \{z \in \mathbb{R}^n : |z_i - x_i| < \delta, \forall i \in \mathbb{N}_n\}$. In addition, $B(x, \delta)$ denotes the set $\{z \in \mathbb{R}^n : 0 < \|z - x\|_2 < \delta\}$. For any closed set $S \subseteq \mathbb{R}^n$, the distance from $x \in \mathbb{R}^n$ to $S$ is defined by $\text{dist}(x, S) := \inf\{\|x - z\|_2 : z \in S\}$. The indicator function on $S$ is defined by $\iota_S(x) := \begin{cases} 0, & \text{if } x \in S, \\ +\infty, & \text{otherwise}. \end{cases}$

In the remaining part of this section, we present some preliminaries on the Fréchet subdifferential $[30]$ and the Kurdyka-Lojasiewicz (KL) property $[2]$. These concepts play a central role in our theoretical and algorithmic developments.

2.1. Fréchet subdifferential and its calculus for the quotient of two functions. Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a proper function. The domain of $\varphi$ is defined by
\[ \text{dom}(\varphi) := \{ x \in \mathbb{R}^n : \varphi(x) < +\infty \}. \]

The Fréchet subdifferential of \( \varphi \) at \( x \in \text{dom}(\varphi) \), denoted by \( \partial \varphi(x) \), is defined by
\[
\partial \varphi(x) := \left\{ y \in \mathbb{R}^n : \liminf_{z \to x, z \neq x} \frac{\varphi(z) - \varphi(x) - \langle y, z - x \rangle}{\| z - x \|_2} \geq 0 \right\}.
\]

The set \( \partial \varphi(x) \) is convex and closed. If \( x \notin \text{dom}(\varphi) \), we let \( \partial \varphi(x) = \emptyset \). We say \( \varphi \) is Fréchet subdifferentiable at \( x \in \mathbb{R}^n \) when \( \partial \varphi(x) \neq \emptyset \). If \( \varphi \) is convex, then \( \partial \varphi(x) \) reduces to the classical subdifferential in convex analysis, i.e.,
\[
\partial \varphi(x) = \{ y \in \mathbb{R}^n : \varphi(z) - \varphi(x) - \langle y, z - x \rangle \geq 0, \forall z \in \mathbb{R}^n \}.
\]

Apart from the Fréchet subdifferential, we also need the notion of limiting (Fréchet) subdifferentials. The limiting subdifferential of \( \varphi \) at \( x \in \text{dom}(\varphi) \) is defined by
\[
\hat{\partial} \varphi(x) := \{ y \in \mathbb{R}^n : \exists x^k \to x, \varphi(x^k) \to \varphi(x), \ y^k \in \partial \varphi(x^k) \to y \}.
\]

It is straightforward that \( \partial \varphi(x) \subseteq \hat{\partial} \varphi(x) \) for all \( x \in \mathbb{R}^n \).

We next recall some simple and useful calculus results on \( \partial \) and \( \hat{\partial} \). For any \( \alpha > 0 \) and \( x \in \mathbb{R}^n \), \( \partial(\alpha \varphi)(x) = \alpha \partial \varphi(x) \) and \( \hat{\partial}(\alpha \varphi)(x) = \alpha \hat{\partial} \varphi(x) \). Let \( \varphi_1, \varphi_2 : \mathbb{R}^n \to \mathbb{R} \) be proper and lower semicontinuous and \( x \in \text{dom}(\varphi_1 + \varphi_2) \). Then, \( \partial \varphi_1(x) + \partial \varphi_2(x) \subseteq \partial(\varphi_1 + \varphi_2)(x) \). If \( \varphi_2 \) is differentiable at \( x \), then \( \partial \varphi_2(x) = \{ \nabla \varphi_2(x) \} \) and \( \partial(\varphi_1 + \varphi_2)(x) = \partial \varphi_1(x) + \nabla \varphi_2(x) \). Furthermore, if \( \varphi_2 \) is continuously differentiable at \( x \), then \( \hat{\partial} \varphi_2(x) = \{ \nabla \varphi_2(x) \} \) and \( \hat{\partial}(\varphi_1 + \varphi_2)(x) = \hat{\partial} \varphi_1(x) + \nabla \varphi_2(x) \).

We next present some results of the Fréchet subdifferential for the quotient of two functions. To this end, we first recall the calmness condition.

**Definition 2.1 (Calmness condition [30]).** The function \( \varphi : \mathbb{R}^n \to \mathbb{R} \) is said to satisfy the calmness condition at \( x \in \text{dom}(\varphi) \), if there exists \( \kappa > 0 \) and a neighborhood \( O \) of \( x \), such that
\[
|\varphi(u) - \varphi(x)| \leq \kappa\| u - x \|_2
\]
for all \( u \in O \).

**Proposition 2.2 (Subdifferential calculus for quotient of two functions).** Let \( f_1 : \mathbb{R}^n \to \mathbb{R} \) be proper and \( f_2 : \mathbb{R}^n \to \mathbb{R} \). Define \( \rho : \mathbb{R}^n \to \mathbb{R} \) at \( x \in \mathbb{R}^n \) as
\[
\rho(x) := \begin{cases} 
  f_1(x) / f_2(x), & \text{if } x \in \text{dom}(f_1) \text{ and } f_2(x) \neq 0, \\
  +\infty, & \text{else}.
\end{cases}
\]

Let \( x \in \text{dom}(\rho) \) with \( a_1 := f_1(x) \geq 0 \) and \( a_2 := f_2(x) > 0 \). If \( f_1 \) is continuous at \( x \) relative to \( \text{dom}(\rho) \) and \( f_2 \) satisfies the calmness condition at \( x \), then
\[
\partial \rho(x) = \frac{\partial(a_2 f_1 - a_1 f_2)(x)}{a_2^2}.
\]

If \( f_2 \) is further differentiable at \( x \), then
\[
\partial \rho(x) = \frac{a_2 \partial f_1(x) - a_1 \nabla f_2(x)}{a_2^2}.
\]

The proof is given in the Appendix A.
2.2. Kurdyka-Łojasiewicz (KL) property.

Definition 2.3 (KL property [2]). A proper function \( \varphi : \mathbb{R}^n \to \bar{\mathbb{R}} \) is said to satisfy the KL property at \( \hat{x} \in \text{dom}(\hat{\partial} \varphi) \) if there exist \( \eta \in (0, +\infty) \), a neighborhood \( O \) of \( \hat{x} \) and a continuous concave function \( \phi : [0, \eta) \to \mathbb{R}^+ \), such that:

(i) \( \phi(0) = 0 \),

(ii) \( \phi \) is continuously differentiable on \( (0, \eta) \) with \( \phi' > 0 \),

(iii) For any \( x \in O \cap \{ x \in \mathbb{R}^n : \varphi(\hat{x}) < \varphi(x) < \varphi(\hat{x}) + \eta \} \), there holds \( \phi'(\varphi(x) - \varphi(\hat{x})) \text{dist}(0, \hat{\partial} \varphi(x)) \geq 1 \).

A proper lower semicontinuous function \( \varphi : \mathbb{R}^n \to \bar{\mathbb{R}} \) is called a KL function if \( \varphi \) satisfies the KL property at all points in \( \text{dom}(\hat{\partial} \varphi) \). For connections between the KL property and the well-known error bound theory [21, 22, 26], we refer the interested readers to [8]. A framework for proving global sequential convergence using the KL property is provided in [3]. We review this result in the next proposition.

Proposition 2.4. Let \( \varphi : \mathbb{R}^n \to \bar{\mathbb{R}} \) be a proper lower semicontinuous function. Consider a sequence satisfies the following three conditions:

(i) (Sufficient decrease condition.) There exists \( a > 0 \) such that

\[
\varphi(x^{k+1}) + a\|x^{k+1} - x^k\|^2_2 \leq \varphi(x^k)
\]

holds for any \( k \in \mathbb{N} \);

(ii) (Relative error condition.) There exists \( b > 0 \) and \( \omega^{k+1} \in \hat{\partial} \varphi(x^{k+1}) \) such that

\[
\|\omega^{k+1}\|_2 \leq b\|x^{k+1} - x^k\|_2
\]

holds for any \( k \in \mathbb{N} \);

(iii) (Continuity condition.) There exists a subsequence \( \{x^{k_j} : j \in \mathbb{N}\} \) and \( x^* \) such that

\[
x^{k_j} \to x^* \text{ and } \varphi(x^{k_j}) \to \varphi(x^*), \text{ as } j \to \infty.
\]

If \( \varphi \) satisfies the KL property at \( x^* \), then \( \sum_{k=1}^{\infty} \|x^k - x^{k-1}\|_2 < +\infty \), \( \lim_{k \to \infty} x^k = x^* \) and \( 0 \in \hat{\partial} \varphi(x^*) \).

3. First-order necessary optimality condition. In this section, we establish a first-order necessary optimality condition for local minimizers of problem (1.1). For convenience, we define \( F : \mathbb{R}^n \to \bar{\mathbb{R}} \) at \( x \in \mathbb{R}^n \) as

\[
F(x) := \begin{cases} 
\frac{f(x) + h(x)}{g(x)}, & \text{if } x \in \Omega \cap \text{dom}(f), \\
+\infty, & \text{else}.
\end{cases}
\]

Then, problem (1.1) can be written as

\[
\min \{ F(x) : x \in \mathbb{R}^n \}.
\]

From the generalized Fermat’s rule [30, Theorem 10.1], we know that if \( x^* \) is a local minimizer of problem (1.1) then \( 0 \in \partial F(x^*) \). Since \( g \) is not necessarily differentiable, by Proposition 2.2 for \( x \in \text{dom}(F) \), \( \partial F(x) \) can not be represented by subdifferential of \( f \) and \( g \) and the gradient of \( h \). Therefore, we have to derive the first-order optimality condition on a different manner.

Our idea is to take advantage of the parametric programing. With the help of the parametric problem, we obtain the first-order necessary optimality condition of
local minimizers of $F$. To this end, we first characterize local and global minimizers of problem (1.1) by those of its corresponding parametric problem. The result is presented in the next proposition and the proof is given in Appendix B.

**Proposition 3.1.** Let $x^* \in \text{dom}(F)$ and $c_* = F(x^*)$. Then, $x^*$ is a local (resp., global) minimizer of problem (1.1) if and only if $x^*$ is a local (resp., global) minimizer of the following problem:

$$
(3.1) \quad \min \{ f(x) + h(x) - c_* g(x) : x \in \Omega \}.
$$

We next present an important inequality, which plays a crucial role in deducing the first-order optimality condition.

**Lemma 3.2.** Let $x^* \in \text{dom}(F)$ be a local minimizer of problem (3.1) with $c_* = F(x^*)$. Then, there exists $\delta > 0$ such that for any $x \in B(x^*, \delta) \cap \text{dom}(F)$ and any $y^* \in \partial g(x^*)$, there holds

$$
(3.2) \quad f(x^*) - f(x) - \langle \nabla h(x^*) - c_* y^*, x - x^* \rangle + \frac{L}{2} \| x - x^* \|_2^2 
$$

**Proof.** Since $x^*$ is a local minimizer of problem (3.1), there exists $\delta > 0$ such that for any $x \in B(x^*, \delta) \cap \text{dom}(F)$, there holds

$$
(3.2) \quad f(x^*) + h(x^*) - c_* g(x^*) \leq f(x) + h(x) - c_* g(x).
$$

Due to the Lipschitz continuity of $\nabla h$, convexity of $g$ and $c_* \geq 0$, it follows that, for any $x \in \mathbb{R}^n$ and $y^* \in \partial g(x^*)$, $h(x) \leq h(x^*) + \langle \nabla h(x^*), x - x^* \rangle + \frac{L}{2} \| x - x^* \|_2^2$ and $c_* g(x^*) + \langle c_* y^*, x - x^* \rangle \leq c_* g(x)$. By summing (3.2) and those two inequalities, we get this lemma. \[ \square \]

Now, we are ready to present the first-order necessary optimality condition for problem (1.1).

**Theorem 3.3.** Let $x^* \in \text{dom}(F)$ be a local minimizer of problem (1.1) and $c_* = F(x^*)$, then $c_* \partial g(x^*) \subseteq \partial f(x^*) + \nabla h(x^*)$.

**Proof.** From Proposition 3.1, $x^*$ is a local minimizer of problem (3.1). By Lemma 3.2, we have that $x^*$ is a local minimizer of the following problem, for all $y^* \in \partial g(x^*)$,

$$
\min \left\{ f(x) + \langle \nabla h(x^*) - c_* y^*, x - x^* \rangle + \frac{L}{2} \| x - x^* \|_2^2 : x \in \Omega \right\}.
$$

Because $g$ is continuous on $\mathbb{R}^n$, $\Omega$ is an open subset of $\mathbb{R}^n$. Thus, $x^*$ is an interior point of $\Omega$. Therefore, $0 \in \partial f(x^*) + \nabla h(x^*) - c_* y^*$ for all $y^* \in \partial g(x^*)$. This implies that $c_* \partial g(x^*) \subseteq \partial f(x^*) + \nabla h(x^*)$. We complete the proof. \[ \square \]

Inspired by the above theorem, we define a critical point of $F$ as follows.

**Definition 3.4 (Critical point of $F$).** Let $x^* \in \text{dom}(F)$ and $c_* = F(x^*)$. We say that $x^*$ is a critical point of $F$ if

$$
0 \in \partial f(x^*) + \nabla h(x^*) - c_* \partial g(x^*).
$$

**Remark 3.5.** According to Proposition 2.2, it should be emphasized that we could neither say a critical point $x^*$ of $F$ defined in Definition 3.4 satisfies $0 \in \partial F(x^*)$, nor vice versa. However, by assuming that $g$ is differentiable, we have for $x \in \text{dom}(F)$

$$
\partial F(x) = \frac{g(x)(\partial f(x) + \nabla h(x)) - (f(x) + h(x))\nabla g(x)}{g^2(x)}.
$$
In this case, the statement that $x^*$ is a critical point of $F$ (Definition 3.4) coincides with that $0 \in \partial F(x^*)$.

By Theorem 3.3, if $x^*$ is a local minimizer of $F$, then $x^*$ is a critical point of $F$. In the remaining part of this paper, we dedicate to developing iterative numerical algorithms to find critical points of $F$.

4. The proximity-gradient-subgradient algorithm (PGSA) for solving problem (1.1). This section is devoted to designing numerical algorithms for solving problem (1.1). We first propose an iterative scheme for solving problem (1.1), according to the first-order optimality condition. Then, we establish the convergence of objective function values and the subsequential convergence under a mild assumption. Finally, by making additional assumptions on $f$, $g$ and assuming the level boundedness and KL property of the objective, we prove the convergence of the whole sequence generated by the proposed algorithm.

From Theorem 3.3, a local minimizer of problem (1.1) must be a critical point of $F$. Thus, our task becomes developing an algorithm with accumulation point being a critical point of $F$. To this end, we introduce the notion of proximity operators. For a proper and lower semicontinuous function $\varphi : \mathbb{R}^n \to \bar{\mathbb{R}}$, the proximity operator of $\varphi$ at $x \in \mathbb{R}^n$, denoted by $\text{prox}_\varphi(x)$, is defined by

$$\text{prox}_\varphi(x) := \arg \min \{ \varphi(y) + \frac{1}{2}\|y - x\|^2 : y \in \mathbb{R}^n \}.$$  

The operator $\text{prox}_\varphi$ is single-valued when $\varphi$ is convex and may be set-valued as $\varphi$ is nonconvex. With the help of the proximity of operator, we derive a sufficient condition for a critical point of $F$ in the following proposition.

**Proposition 4.1.** If $x^* \in \text{dom}(F)$ satisfies

$$x^* \in \text{prox}_{\alpha f}(x^* - \alpha \nabla h(x^*) + \alpha c^* y^*)$$

for some $\alpha > 0$, $y^* \in \partial g(x^*)$ and $c^* = F(x^*)$, then $x^*$ is a critical point of $F$.

**Proof.** By the proximity operator and the Fermat’s rule, (4.1) leads to

$$0 \in \alpha \partial f(x^*) - \alpha \nabla h(x^*) + \alpha c^* y^*,$$

which implies that $x^*$ is a critical point of $F$. \(\square\)

Inspired by Proposition 4.1, we propose the following first-order algorithm, which is stated in Algorithm 1. Since Algorithm 1 involves in the proximity operator of $f$, the gradient of $h$ and the subgradient of $g$, we refer to it as the proximity-gradient-subgradient algorithm (PGSA).

**Algorithm 1** proximity-gradient-subgradient algorithm (PGSA) for solving (1.1)

**Step 0.** Input $x^0 \in \text{dom}(F)$, $0 < \alpha \leq \alpha_k \leq \bar{\alpha} < 1/L$, for $k \in \mathbb{N}$. Set $k \leftarrow 0$.

**Step 1.** Compute

$$(k+1) \in \partial g(x^k),$$

$$c_k = \frac{f(x^k) + h(x^k)}{g(x^k)},$$

$$x^{k+1} \in \text{prox}_{\alpha_k f}(x^k - \alpha_k \nabla h(x^k) + \alpha_k c_k y^{k+1}).$$

**Step 2.** Set $k \leftarrow k + 1$ and go to Step 1.

We remark here that $\alpha_k$ is required to be in $(0, 1/L)$ to ensure $x^k \in \text{dom}(F)$ for all $k \in \mathbb{N}$. As a result the objective function value $c_k$ is well-defined. We will give the detailed proof in Lemma 4.2.
4.1. Convergence of objective function value. In this subsection, we prove that the objective function value \( \{ F(x^k) : k \in \mathbb{N} \} \) decreases and converges. We first establish below a lemma, which plays a crucial role in the convergence analysis.

**Lemma 4.2.** The sequence \( \{ x^k : k \in \mathbb{N} \} \) generated by PGSA falls into \( \text{dom}(F) \) and satisfies

\[
(4.2) \quad f(x^{k+1}) + h(x^{k+1}) + \frac{1}{2\alpha_k} \| x^{k+1} - x^k \|_2^2 \leq c_k g(x^{k+1}).
\]

**Proof.** We prove inequality (4.2) and \( x^k \in \text{dom}(F) \) by induction. First, the initial point \( x^0 \) is in \( \text{dom}(F) \). Suppose \( x^k \in \text{dom}(F) \) for some \( k \in \mathbb{N} \). From PGSA and the definition of proximity operators, we get

\[
f(x^{k+1}) + \frac{1}{2\alpha_k} \| x^{k+1} - (x^k - \alpha_k \nabla h(x^k) + \alpha_k c_k y^{k+1}) \|_2^2 \\
\leq f(x^k) + \frac{1}{2\alpha_k} \| \nabla h(x^k) - \alpha_k c_k y^{k+1} \|_2^2,
\]

which implies that

\[
(4.3) \quad f(x^{k+1}) + \frac{1}{2\alpha_k} \| x^{k+1} - x^k \|_2^2 + \langle x^{k+1} - x^k, \nabla h(x^k) - c_k y^{k+1} \rangle \leq f(x^k).
\]

Since \( \nabla h \) is Lipschitz continuous with constant \( L \), there holds

\[
h(x^{k+1}) \leq h(x^k) + \langle \nabla h(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \| x^{k+1} - x^k \|_2^2.
\]

Due to the convexity of \( g \) and \( c_k \geq 0 \), it follows that

\[
(4.5) \quad c_k g(x^k) + \langle c_k y^{k+1}, x^{k+1} - x^k \rangle \leq c_k g(x^{k+1}).
\]

By summing (4.3), (4.4) and (4.5), we obtain (4.2) from \( c_k g(x^k) = f(x^k) + h(x^k) \).

Assume that \( x^{k+1} \notin \Omega \). We know \( x^{k+1} \notin \Omega \) and \( g(x^{k+1}) = 0 \) due to \( x^{k+1} \in \text{dom}(f) \) and \( \text{dom}(F) = \Omega \cap \text{dom}(f) \). By the fact \( f + h \geq 0 \) and \( 0 < \alpha_k < 1/L \), we deduce that \( x^{k+1} = x^k \) from (4.2). This contradicts to \( x^k \in \text{dom}(F) \) and thus implies \( x^{k+1} \notin \text{dom}(F) \). Therefore, we conclude \( x^k \in \text{dom}(F) \) for all \( k \in \mathbb{N} \).

With the help of Lemma 4.2, we get the main result of this subsection.

**Theorem 4.3.** Let \( \{ x^k : k \in \mathbb{N} \} \) be generated by PGSA. Then, the following statements hold:

(i) \( F(x^{k+1}) + \frac{1}{2\alpha_k} \| x^{k+1} - x^k \|_2^2 \leq F(x^k) \) for \( k \in \mathbb{N} \);

(ii) \( \lim_{k \to \infty} c_k = \lim_{k \to \infty} F(x^k) = c \) with \( c \geq 0 \);

(iii) \( \lim_{k \to \infty} \frac{1}{g(x^{k+1})} \| x^{k+1} - x^k \|_2^2 = 0 \).

**Proof.** From Lemma 4.2, \( g(x^k) \neq 0 \) for all \( k \in \mathbb{N} \). Thus, (4.2) in Lemma 4.2 implies Item (i) due to \( g(x^{k+1}) > 0 \). Item (ii) follows immediately by \( F \geq 0 \) and \( 0 < \alpha_k < 1/L \). Item (iii) is a direct consequence of Item (i) and Item (ii). We complete the proof. 

\[\square\]
4.2. Subsequential convergence. In this subsection, we consider the subsequential convergence of PGSA. We begin with a mild assumption.

Assumption 1. Functions $f + h$ and $g$ do not attain $0$ simultaneously.

Under Assumption 1, we will show that any accumulation point of $\{x^k : k \in \mathbb{N}\}$ is a critical point of $F$. To this end, we first prove that any accumulation point of $\{x^k : k \in \mathbb{N}\}$ belongs to $\text{dom}(F)$.

Lemma 4.4. Suppose Assumption 1 holds and let $\{x^k : k \in \mathbb{N}\}$ be generated by PGSA. Then any accumulation point of $\{x^k : k \in \mathbb{N}\}$ belongs to $\text{dom}(F)$.

Proof. Let $x^*$ be an accumulation point of $\{x^k : k \in \mathbb{N}\}$ and $\{x^{k_j} : j \in \mathbb{N}\}$ be a subsequence such that $\lim_{j \to \infty} x^{k_j} = x^*$. By the definition of $F$, it suffices to prove that $x^* \in \text{dom}(f)$ and $x^* \in \Omega$ respectively.

First, by Lemma 4.2 and $0 < \alpha_k < 1/L$, for all $j \in \mathbb{N}$ it holds that

\begin{equation}
(4.6) \quad f(x^{k_j}) + h(x^{k_j}) \leq c_{k_j - 1}g(x^{k_j}).
\end{equation}

Since $f$, $g$ and $h$ are continuous on $\text{dom}(F)$, from Item (ii) in Theorem 4.3 we obtain that

\begin{equation}
(4.7) \quad f(x^*) + h(x^*) = \lim_{j \to \infty} f(x^{k_j}) + h(x^{k_j}) \leq cg(x^*).
\end{equation}

We conclude that $x^* \in \text{dom}(f)$ by the fact that $c$, $g(x^*)$ and $h(x^*)$ are finite. Assume that $g(x^*) = 0$. Then (4.7) indicates $f(x^*) + h(x^*) = 0$, which contradicts to Assumption 1. Therefore, we know $g(x^*) \neq 0$, that is, $x^* \in \Omega$. This completes the proof.

We are now ready to present the main result of this subsection.

Theorem 4.5. Suppose Assumption 1 holds. Let $\{x^k : k \in \mathbb{N}\}$ be generated by PGSA. Then any accumulation point of $\{x^k : k \in \mathbb{N}\}$ is a critical point of $F$.

Proof. Let $\{x^{k_j} : j \in \mathbb{N}\}$ be a subsequence such that $\lim_{j \to \infty} x^{k_j} = x^*$. From Theorem 4.3 (i) and $\alpha_k \leq \bar{\alpha}$, we have

\begin{equation}
F(x^{k_j}) + 1/\bar{\alpha} - L/2g(x^{k_j})\|x^{k_j} - x^{k_j-1}\|^2_2 \leq F(x^{k_j-1}).
\end{equation}

Using Item (ii) of Theorem 4.3, Lemma 4.4, $\bar{\alpha} < 1/L$ and the continuity of $g$ at $x^*$, we conclude $\lim_{j \to \infty} \|x^{k_j} - x^{k_j-1}\| = 0$ and $\lim_{j \to \infty} x^{k_j-1} = x^*$. Since $g$ is a real-valued convex function and $\{x^{k_j-1} : j \in \mathbb{N}\}$ is bounded, we know that $\{y^{k_j} : j \in \mathbb{N}\}$ is also bounded. Without loss of generality we may assume $\lim_{j \to \infty} y^{k_j}$ and $\lim_{j \to \infty} \alpha_{k_j - 1}$ exist.

In addition, $\lim_{j \to \infty} y^{k_j} = y^*$ belongs to $\partial g(x^*)$ due to the closeness of $\partial g$. From the iteration of PGSA, we have

\begin{equation}
(4.8) \quad x^{k_j} \in \text{prox}_{\alpha_{k_j - 1}} f(x^{k_j-1} - \alpha_{k_j - 1} \nabla h(x^{k_j-1}) + \alpha_{k_j - 1} c_{k_j - 1} y^{k_j}).
\end{equation}

As $\nabla h$ and $f$ is continuous at $x^*$, we obtain (4.1) by passing to the limit in the above relation with $\alpha = \lim_{j \to \infty} \alpha_{k_j - 1}$. By Proposition 4.1, $x^*$ is a critical point of $F$.

4.3. Global sequential convergence. We investigate in this subsection the global convergence of the entire sequence $\{x^k : k \in \mathbb{N}\}$ generated by PGSA. We shall show $\{x^k : k \in \mathbb{N}\}$ converges to a critical point of $F$ under suitable assumptions. To this end, we need to introduce three assumptions as follows:
**Assumption 2.** Function $F$ is level bounded.

**Assumption 3.** Function $f$ is locally Lipschitz continuous on $\text{dom}(f)$.

**Assumption 4.** Function $g$ is continuously differentiable on $\Omega$ with a locally Lipschitz continuous gradient.

Our analysis in this subsection mainly makes use of Proposition 2.4 which is based on KL property. If $F$ is assumed to satisfy the KL property, from Proposition 2.4 we can establish the global convergence of PGSA by showing the lower semicontinuity of $F$, the boundedness of the sequence generated, and Items (i)-(ii) in Proposition 2.4. We shall prove these results in the following three lemmas.

**Lemma 4.6.** Suppose Assumption 1 holds. Then, $F$ is a lower semicontinuous function.

**Proof.** If $x \in \Omega$, it holds that $0 < g(x) = \lim_{y \rightarrow x} g(y)$. Since $f$ is lower semicontinuous and $h$ is continuous, we immediately have $F(x) \leq \liminf_{y \rightarrow x} f(y)$. If $x \notin \Omega$, we obtain that $F(x) = +\infty$ and $0 = g(x) = \lim_{y \rightarrow x} g(y)$. Due to Assumption 1, $0 < f(x) + h(x) \leq \liminf_{y \rightarrow x} f(y) + h(y)$. Thus, $\liminf_{y \rightarrow x} F(x) = +\infty$ from the fact that $g \geq 0$. Therefore, we have $F(x) = \liminf_{y \rightarrow x} F(y)$. This completes the proof.

**Lemma 4.7.** Suppose that Assumptions 1 and 2 hold. Let $\{x^k : k \in \mathbb{N}\}$ be generated by PGSA. Then the following statements hold:

(i) $\{x^k : k \in \mathbb{N}\}$ is bounded;

(ii) $F(x^{k+1}) + \frac{a}{2}\|x^{k+1} - x^k\|^2_2 \leq F(x^k)$ for $k \in \mathbb{N}$, where $a := (1/\alpha - L)/M > 0$ with $M := \sup\{g(x) : x \in \text{lev}(F,c_0)\}$.

**Proof.** By Theorem 4.3 (i), we have for all $k \in \mathbb{N}$, $x^k \in \text{lev}(F,c_0)$. Then the boundedness of $\{x^k : k \in \mathbb{N}\}$ follows immediately from Assumption 2. Assumption 1 ensures the lower semicontinuity of $F$. Hence, the set $\text{lev}(F,c_0)$ is closed and bounded. Since $g$ is continuous, we know $M$ is finite. This together with Theorem 4.3 (i) and $\alpha_k < \alpha$ yields Item (ii).

**Lemma 4.8.** Let $\{x^k : k \in \mathbb{N}\}$ be generated by PGSA. Suppose Assumptions 1-4 hold. Then there exist $b > 0$ and $\omega^{k+1} \in \partial F(x^{k+1})$ such that $\|\omega^{k+1}\|_2 \leq b\|x^{k+1} - x^k\|_2$ for all $k \in \mathbb{N}$.

**Proof.** By Lemmas 4.2 and 4.4, we know $x^k \in \Omega$ for any $k \in \mathbb{N}$ and any accumulation point $x^*$ of $\{x^k : k \in \mathbb{N}\}$ satisfies $g(x^*) > 0$. Thus, there exists $t > 0$ such that $g(x^k) \geq t$, since $\{x^k : k \in \mathbb{N}\}$ is bounded and $g$ is continuous on $\Omega$. Let $S \subseteq \mathbb{R}^n$ be a bounded closed subset satisfying $\{x^k : k \in \mathbb{N}\} \subseteq S \subseteq \text{dom}(F)$. Then it is easy to check that $\forall g$ and $F$ are globally Lipschitz continuous on $S$. We denote the Lipschitz constant of $\nabla g$ and $F$ by $L$ and $\tilde{L}$ respectively.

From the iteration of PGSA and the differentiability of $g$, we obtain that $x^k - x^{k+1} - \alpha_k \nabla h(x^k) + \alpha_k c_k \nabla g(x^k) \in \alpha_k \partial f(x^{k+1})$, which implies that

\[
\frac{1}{\alpha_k g(x^{k+1})} (x^k - x^{k+1}) - \nabla h(x^k) - \frac{c_k g(x^{k+1})}{g(x^k)} \nabla g(x^k) \in \partial f(x^{k+1}).
\]

From Assumptions 3-4 and Proposition 2.2, we have on $\text{dom}(\partial F)$

\[
\partial F = \frac{g(\partial f + \nabla h) - (f + h) \nabla g}{g^2}.
\]
The above relation and (4.9) suggest that \( \omega^{k+1} \in \partial F(x^{k+1}) \) with
\[
\omega^{k+1} := \frac{1}{\alpha_k g(x^{k+1})} (x^k - x^{k+1}) - \frac{\nabla h(x^k)}{g(x^{k+1})} + \frac{\nabla h(x^{k+1})}{g(x^{k+1})} + \frac{c_k}{g(x^{k+1})} \nabla g(x^k) - \frac{c_{k+1}}{g(x^{k+1})} \nabla g(x^{k+1}).
\]
By a direct computation, it follows that
\[
\|\omega^{k+1}\|_2 \leq \left( \frac{1}{\alpha_k t} + \frac{L + c_k \hat{L}}{t} + \frac{\|\nabla g(x^{k+1})\|_2 \hat{L}}{t} \right) \|x^{k+1} - x^k\|_2.
\]
(4.10)

From Theorem 4.3, we see that \( c_k \leq c_0 \) for \( k \in \mathbb{N} \). Since \( \{x^k : k \in \mathbb{N}\} \) is bounded and \( \nabla g \) is continuous on \( \Omega \), there exists \( \beta > 0 \) such that \( \|\nabla g(x^{k+1})\|_2 \leq \beta \) for all \( k \in \mathbb{N} \). Due to \( \alpha_k \geq \alpha > 0 \), we obtain finally from (4.10) that \( \|\omega^{k+1}\|_2 \leq b\|x^{k+1} - x^k\|_2 \) for all \( k \in \mathbb{N} \), where \( b := (1/\alpha + L + c_0 \hat{L} + \beta \hat{L})/a \). We complete the proof due to \( \partial F(x^{k+1}) \subseteq \partial F(x^{k+1}) \). \( \square \)

Now we are ready to present the main result of this subsection.

**Theorem 4.9.** Suppose that Assumptions 1-4 hold and \( F \) satisfies the KL property at any point in dom(\( F \)). Let \( \{x^k : k \in \mathbb{N}\} \) be generated by PGSA. Then \( \sum_{k=1}^{\infty} \|x^k - x^{k-1}\|_2 < +\infty \) and \( \{x^k : k \in \mathbb{N}\} \) converges to a critical point of \( F \).

**Proof.** From Theorem 4.5, it suffices to prove that \( \sum_{k=1}^{\infty} \|x^k - x^{k-1}\|_2 < +\infty \) and \( \{x^k : k \in \mathbb{N}\} \) is convergent. According to Proposition 2.4, we obtain this theorem from Lemmas 4.6-4.8 and Lemma 4.4 immediately. \( \square \)

**4.4. Convergence rate.** Finally, we consider the convergence rate of PGSA. To this end, we further assume \( F \) is a KL function with the corresponding \( \phi \) (see Definition 2.3) taking the form \( \phi(s) = ds^{1 - \theta} \) for some \( d > 0 \) and \( \theta \in [0, 1) \). Then under the assumption of Theorem 4.9, we can estimate the convergence rate of PGSA, following a similar line of arguments to other convergence rate analysis based on the KL property; see, for example, [1, 36, 38].

**Theorem 4.10.** Suppose that Assumptions 1-4 hold. Let \( \{x^k : k \in \mathbb{N}\} \) be generated by PGSA and suppose that \( \{x^k : k \in \mathbb{N}\} \) converges to \( x^* \). Assume further that \( F \) satisfies the KL property at \( x^* \) with \( \phi(s) = ds^{1 - \theta} \) for some \( d > 0 \) and \( \theta \in [0, 1) \), then the following statements hold:
\begin{enumerate}
  \item [(i)] If \( \theta = 0 \), \( \{x^k : k \in \mathbb{N}\} \) converges to \( x^* \) finitely;
  \item [(ii)] If \( \theta \in (0, 1/2] \), \( \|x^k - x^*\|_2 \leq c_1 r^k \), \( \forall k \geq K_1 \) for some \( K_1 > 0 \), \( c_1 > 0 \), \( \tau \in (0, 1) \);
  \item [(iii)] If \( \theta \in (1/2, 1) \), \( \|x^k - x^*\|_2 \leq c_2 k^{-(1-\theta)/(2\theta - 1)} \), \( \forall k \geq K_2 \), for some \( K_2, c_2 > 0 \).
\end{enumerate}

Here we omit the proof for Theorem 4.10, since it can be performed very similarly to those for other optimization algorithms (see, for example, the proof of [1, Theorem 2]). We remark that as is pointed out in [2], all proper semialgebraic functions satisfy the KL property with \( \phi(s) = ds^{1 - \theta} \) for some \( d > 0 \) and \( \theta \in [0, 1) \). Consequently, both Theorems 4.9 and 4.10 are applicable when \( F \) is a semialgebraic function. Indeed, the objective functions are semialgebraic in a wide range of sparse optimization problems, including the sparse generalized eigenvalue problem (6.3) which will be studied in detail in Section 6.

**5. PGSA with line search.** In this section, we incorporate a line search scheme for adaptively choosing \( \alpha_k \) into PGSA. In PGSA, the step size \( \alpha_k \) should be less than \( 1/L \) for all \( k \in \mathbb{N} \) to ensure the convergence. However, this step size may be too small
in the case of large \( L \) and thus leads to slow convergence of PGSA. To speed up the convergence, we take advantage of the line search technique in [20, 37] to enlarge the step size and meanwhile guarantee the convergence of the algorithm. The PGSA with line search is summarized in Algorithm 2 (PGSA_L).

\begin{algorithm}[H]
\begin{algorithmic}
\caption{PGSA with line search (PGSA_L) for problem (1.1)}
\end{algorithmic}
\end{algorithm}

\begin{algorithmic}
\State \textbf{Step 0.} Input \( x^0 \in \text{dom}(F), a > 0, 0 < \underline{\alpha} < \bar{\alpha}, 0 < \eta < 1 \), and an integer \( N \geq 0 \).
\State Set \( k \leftarrow 0 \).
\State \textbf{Step 1.} \( y^{k+1} \in \partial g(x^k) \), \( c_k = \frac{f(x^k) + h(x^k)}{g(x^k)} \),
\State Choose \( \alpha_{k,0} \in [\underline{\alpha}, \bar{\alpha}] \).
\State \textbf{Step 2.} For \( m = 0, 1, \ldots \), do
\State \hspace{1em} \( \alpha_k = \alpha_{k-1} \eta^m \), \( \bar{x}^{k+1} \in \text{prox}_{\alpha_k f}(x^k - \alpha_k \nabla h(x^k) + \alpha_k c_k y^{k+1}) \),
\State \hspace{1em} If \( \bar{x}^{k+1} \) satisfies \( \bar{x}^{k+1} \in \text{dom}(F) \) and
\State \hspace{2em} \( F(\bar{x}^{k+1}) \leq \max_{|k-N|_+ \leq j \leq k} c_j - \frac{a}{2} \| \bar{x}^{k+1} - x^k \|_2^2 \),
\State \hspace{1em} \( \text{set } x^{k+1} = \bar{x}^{k+1} \) and go to \textbf{Step 3}.
\State \textbf{Step 3.} \( k \leftarrow k + 1 \) and go to \textbf{Step 1}.
\end{algorithmic}

From inequality (5.1), \( \{F(x^k) : k \in \mathbb{N}\} \) is monotone when \( N = 0 \), while it is generally nonmonotone when \( N > 0 \). For convenience of presentation, we call the algorithm PGSA with monotone line search (PGSA_ML) if \( N = 0 \) and PGSA with nonmonotone line search (PGSA_NL) if \( N > 0 \). Let \( \Delta x := x^k - x^{k-1} \), \( \Delta h := \nabla h(x^k) - \nabla h(x^{k-1}) \). Motivated from [5, 20, 37], we adopt a very popular choice of \( \alpha_{k,0} \) in the following formula

\begin{equation}
\alpha_{k,0} = \max \left\{ \alpha, \min \left( \bar{\alpha}, \frac{\| \Delta x \|_2}{\| \Delta h \|_2} \right) \right\}, \quad \text{if } \langle \Delta x, \Delta h \rangle \neq 0, \quad \text{else.}
\end{equation}

This initial step size can be viewed as an adaptive approximation of \( 1/L \) via some local curvature information of \( h \).

Next, we study the convergence property of PGSA_L. To this end, we define \( \tau : \mathbb{N} \to \mathbb{N} \) at \( k \in \mathbb{N} \) as \( \tau(k) := \max \{ i : i \in \arg \max \{ F(x^i) : |k-N|_+ \leq j \leq k \} \} \). The following lemma tells that PGSA_L is well defined and the sequence generated by PGSA_L is bounded under Assumption 2.

\textbf{Lemma 5.1.} \textit{Suppose that Assumption 2 holds and let } \( M := \sup \{ g(x) : x \in \text{lev}(F, c_0) \} \). \textit{Then, the following statements hold:}

(i) \textit{Step 2 of PGSA_L terminates at some } \( \alpha_k \geq \bar{\alpha} \text{ in most } T \text{ iterations, where } \bar{\alpha} := \eta/(aM + L), T := \left\lfloor \frac{-\log(\bar{\alpha} (aM + L))}{\log \eta} \right\rfloor + 1 \right\rfloor; \)

(ii) \textit{\( x^k \in \text{lev}(F, c_0) \) for all } \( k \in \mathbb{N} \);

(iii) \textit{\( \{F(x^{\tau(k)}) : k \in \mathbb{N}\} \) is nonincreasing.}

\textbf{Proof.} Assumption 2 ensures the boundedness of \( \text{lev}(F, c_0) \). Thus, we know \( M \) is finite thanks to the continuity of \( g \). In view of the updating rule for \( \alpha_k \) in Step 2 and \( \alpha_{k,0} \leq \bar{\alpha} \), after \( T \) iterations, we have \( \alpha_k \leq 1/(aM + L) = \bar{\alpha} / \eta \) for any \( k \in \mathbb{N} \).

We proceed by induction on \( k \). It is obvious that \( x^0 \in \text{lev}(F, c_0) \). Now, assume that for \( j = 0, 1, \ldots, k \), \( x^j \) has already been generated and \( x^j \in \text{lev}(F, c_0) \). In order
to prove Item (i), it suffices to show that if $\alpha_k \leq \tilde{a}/\eta$, then $\tilde{x}^{k+1} \in \text{dom}(F)$ and the following inequality holds
\begin{equation}
F(\tilde{x}^{k+1}) \leq c_k - \frac{a}{2} \|\tilde{x}^{k+1} - x^k \|^2.
\end{equation}
By Theorem 4.3 and $\alpha_k \leq 1/(aM + L) < 1/L$, we have $x^{k+1} \in \text{dom}(F)$ and
\begin{equation}
F(\tilde{x}^{k+1}) \leq c_k - \frac{1}{\alpha_k - L} \|x^{k+1} - x^k \|^2 \leq c_k - \frac{aM}{2g(\tilde{x}^{k+1})} \|\tilde{x}^{k+1} - x^k \|^2,
\end{equation}
which indicates that $F(\tilde{x}^{k+1}) \leq c_k \leq c_0$ and thus $\tilde{x}^{k+1} \in \text{lev}(F, c_0)$. Invoking $g(\tilde{x}^{k+1}) \leq M$, we obtain inequality (5.3) from (5.4).

We next prove $x^{k+1} \in \text{lev}(F, c_0)$ and $F(x^{\tau(j+1)}) \leq F(x^{\tau(j)})$ for $j \leq k$. By (5.1), we have $F(x^{\tau(j)}) \leq F(x^{\tau(j)})$ for $j \leq k$. Thus, for $j \leq k$,
\begin{align*}
F(x^{\tau(j+1)}) &= \max_{[j+1-N],+1 \leq i \leq j+1} F(x^i) \\
&= \max \left\{ F(x^{j+1}), \max_{[j+1-N],+1 \leq i \leq j} F(x^i) \right\} \\
&\leq \max \{ F(x^{j+1}), F(x^{\tau(j)}) \} \\
&= F(x^{\tau(j)}).
\end{align*}
This yields that $F(x^{k+1}) \leq F(x^{\tau(k)}) \leq F(x^{\tau(0)}) = c_0$. We complete the proof immediately. \(\Box\)

With the help of Lemma 5.1, we establish the subsequential convergence results of PGSA_{ML} in the next theorem.

**Theorem 5.2.** Suppose that Assumptions 1 and 2 hold. Let $\{x^k : k \in \mathbb{N}\}$ be generated by PGSA_{ML}. Then any accumulation point of $\{x^k : k \in \mathbb{N}\}$ is a critical point of $F$.

**Proof.** Let $x^*$ be an accumulation point of $\{x^k : k \in \mathbb{N}\}$. According to the proof of Theorem 4.5, it suffices to show $\{F(x^k) : k \in \mathbb{N}\}$ converges and $\lim_{k \to \infty} \|x^{k+1} - x^k\|_2 = 0$. By Lemma 5.1, $\{F(x^{\tau(k)}) : k \in \mathbb{N}\}$ is decreasing and $F \geq 0$. Hence, we have that $\lim_{k \to \infty} F(x^{\tau(k)}) = \xi$ for some $\xi \geq 0$. Since $f$ is continuous on $\text{dom}(f)$ and $\text{lev}(F, c_0)$ is closed and bounded, we deduce that $F$ is uniformly continuous on $\text{lev}(F, c_0)$. Noting that $\{x^k : k \in \mathbb{N}\} \subseteq \text{lev}(F, c_0)$ and proceeding as in the proof of [37, Lemma 4] starting from [37, Equation (34)], one can prove that $\lim_{k \to \infty} F(x^k) = \xi$ and $\lim_{k \to \infty} \|x^{k+1} - x^k\|_2 = 0$. We complete the proof. \(\Box\)

Under Assumptions 1-4 and assuming $F$ satisfies the KL property, we can prove the global convergence of the entire sequence generated by PGSA_{ML}.

**Theorem 5.3.** Suppose that Assumptions 1-4 hold and $F$ satisfies the KL property at any point in $\text{dom}(F)$. Let $\{x^k : k \in \mathbb{N}\}$ be generated by PGSA_{ML}. Then $\sum_{k=1}^{\infty} \|x^k - x^{k-1}\|_2 < +\infty$ and $\{x^k : k \in \mathbb{N}\}$ converges to a critical point of $F$.

**Proof.** From Theorem 5.2, it suffices to prove that $\sum_{k=1}^{\infty} \|x^k - x^{k-1}\|_2 < +\infty$ and $\{x^k : k \in \mathbb{N}\}$ is convergent. According to Proposition 2.4, we need to verify Items (i)-(iii) of the proposition, the boundedness of $\{x^k : k \in \mathbb{N}\}$ and that $F$ is lower semicontinuous.

First, the boundedness of $\{x^k : k \in \mathbb{N}\}$ and lower semicontinuity of $F$ follow from Lemma 5.1 and Lemma 4.6, respectively. Items (i) and (iii) of Proposition 2.4
are direct consequence of Lemma 5.1 and Theorem 5.2. Proposition 2.4 (ii) can be obtained by a proof similar to that of Lemma 4.8. Therefore, we complete the proof.

The convergence rate analysis of PGSA ML is almost the same as that of PGSA in Theorem 4.10. Here, we omit the details and present the corresponding results in the next theorem.

**Theorem 5.4.** Suppose that Assumptions 1-4 hold. Let \( \{x_k : k \in \mathbb{N}\} \) be generated by PGSA ML and suppose that \( \{x_k : k \in \mathbb{N}\} \) converges to \( x^* \). Assume further that \( F \) satisfies the KL property at \( x^* \) with \( \phi(s) = ds^{1-\theta} \) for some \( d > 0 \) and \( \theta \in [0,1) \), then Items (i)-(iii) of Theorem 4.10 hold.

### 6. Applications to sparse generalized eigenvalue problem

In this section, we identify SGEP associated with a pair of symmetric positive semidefinite matrices as a special case of problem (1.1) and apply our proposed algorithms. Then we establish the global sequential (resp. subsequential) convergence of the sequence generated by PGSA and PGSA ML (resp. PGSA NL) for SGEP. In addition, under suitable assumptions, we estimate the convergence rate of PGSA and PGSA ML.

Assume that \( A, B \) are both \( n \times n \) symmetric positive semidefinite matrices and any \( r \times r \) principal sub-matrix of \( B \) is positive definite for some integer \( r \in [1,n] \). If there exist \( \lambda^* \in \mathbb{R} \) and \( x^* \in \mathbb{R}^n \), such that \( Ax^* = \lambda^* B x^* \), then \( x^* \) is called the generalized eigenvector with respect to the generalized eigenvalue \( \lambda^* \) of the matrix pair \( (A,B) \). Obviously, the leading generalized eigenvector with respect to the largest generalized eigenvalue can be obtained by solving the following optimization problem

\[
\max \left\{ \frac{x^TAx}{x^TBx} : \|x\|_2 = 1, \ x^TBx \neq 0, \ x \in \mathbb{R}^n \right\}.
\]

In the context of sparse modeling, it is natural to incorporate the sparsity constraint into problem (6.1). This leads to the SGEP:

\[
\max \left\{ \frac{x^TAx}{x^TBx} : \|x\|_2 = 1, \ \|x\|_0 \leq r, \ x^TBx \neq 0, \ x \in \mathbb{R}^n \right\},
\]

where the \( \ell_0 \) function \( \|\cdot\|_0 \) counts the number of nonzero components in a vector. The SGEP covers several statistical learning models, such as the sparse principal component analysis [12, 41], sparse fisher discriminant analysis [11, 24], sparse sliced inverse regression [10, 19] and so on. One can easily check that the optimal solution set of SGEP is completely the same as that of the following minimization problem

\[
\min \left\{ \frac{x^TBx}{x^TAx} : \|x\|_2 = 1, \ \|x\|_0 \leq r, \ x^TAx \neq 0, \ x \in \mathbb{R}^n \right\}.
\]

Thus, problem (6.3) is another formulation of SGEP. We also notice that problem (6.3) is not a classical quadratic fractional problem due to its nonconvex constraints. In fact, problem (6.3) is a special case of the general optimization problem (1.1) with \( f \) being the indicator function on the set \( \{ x \in \mathbb{R}^n : \|x\|_0 \leq r, \ \|x\|_2 = 1 \} \). 

\[
g(x) = x^TAx, \quad h(x) = x^TBx \quad \text{for} \ x \in \mathbb{R}^n.
\]

Therefore, the proposed PGSA and PGSA NL can be directly applied to problem (6.3). For convenience of presentation, we denote the constraint set \( \{ x \in \mathbb{R}^n : \|x\|_0 \leq r, \ \|x\|_2 = 1 \} \) in problem (6.3) by \( C \) and define \( G : \mathbb{R}^n \to \mathbb{R} \) at \( x \in \mathbb{R}^n \) as

\[
G(x) := \begin{cases} \frac{x^TBx}{x^TAx}, & \text{if } x \in C \text{ and } x^TAx \neq 0, \\ +\infty, & \text{else}. \end{cases}
\]
6.1. Critical points of problem (6.3). In this subsection, we have a closer look at the critical points of problem (6.3). We begin with the following lemma concerning the Fréchet subdifferential of the indicator function \( \iota_C \).

**Lemma 6.1.** Let \( x \in C \) and \( \Lambda \) be the support of \( x \), then the following statements hold:

(i) \( \partial \iota_C(x) = \begin{cases} \{ v \in \mathbb{R}^n : v = tx, t \in \mathbb{R} \} & \text{if } \|x\| < r, \\
\{ v \in \mathbb{R}^n : v_\Lambda = tx_\Lambda, t \in \mathbb{R} \} & \text{else.} \end{cases} \)

(ii) For any \( v \in \partial \iota_C(x) \), there exists \( t \in \mathbb{R} \), such that \( v_\Lambda = tx_\Lambda \). In particular, if \( r = n \), i.e., \( C = \{ x \in \mathbb{R}^n : \|x\|_2 = 1 \} \), then \( \partial \iota_C(x) = \{ v \in \mathbb{R}^n : v = tx, t \in \mathbb{R} \} \).

The proof of Lemma 6.1 is given in Appendix C. With the help of Lemma 6.1, we characterize the relationship between the critical points of \( G \) and the generalized eigenvectors of matrix pair \( (A, B) \) or the related sub-matrix pair of \( (A, B) \).

**Proposition 6.2.** Let \( x^* \in \text{dom}(G) \) and \( \Lambda \) be the support of \( x^* \). Then \( x^* \) is a critical point of \( G \) if and only if one of the following statements hold:

(i) \( |\Lambda| < r \) and \( x^* \) is an unit generalized eigenvector with respect to the generalized eigenvalue \( 1/G(x^*) \) of the matrix pair \( (A, B) \), i.e., \( Bx^* = G(x^*)Ax^* \);

(ii) \( |\Lambda| = r \) and \( x^*_\Lambda \) is an unit generalized eigenvector with respect to the generalized eigenvalue \( 1/G(x^*) \) of the matrix pair \( (A_\Lambda, B_\Lambda) \), i.e., \( B_\Lambda x^*_\Lambda = G(x^*)A_\Lambda x^*_\Lambda \).

**Proof.** According to Definition 3.4, \( x^* \) is a critical point of \( G \) if and only if

\[
0 \in \partial \iota_C(x^*) + 2Bx^* - 2G(x^*)Ax^*.
\]

We first prove Item (i). Assume that \( |\Lambda| < r \). By Lemma 6.1, the inclusion (6.4) is equivalent to the following relation

\[
d_1x^* + 2Bx^* - 2G(x^*)Ax^* = 0
\]

for some \( d_1 \in \mathbb{R} \). Multiplying \( (x^*)^T \) on both sides of the above equality, we get that \( d_1 = 0 \). This proves Item (i).

Next, we prove Item (ii). Suppose that \( |\Lambda| = r \). Invoking Lemma 6.1 in this case, inclusion (6.4) implies that there exist \( d_2 \in \mathbb{R} \) and \( v \in \mathbb{R}^n \) such that \( v_\Lambda = d_2x^*_\Lambda \) and

\[
v + 2Bx^* - 2G(x^*)Ax^* = 0.
\]

This yields that

\[
d_2x^*_\Lambda + 2B_\Lambda x^*_\Lambda - 2G(x^*)A_\Lambda x^*_\Lambda = 0.
\]

Multiplying \( (x^*_\Lambda)^T \) on both sides of the above equality, we immediately obtain \( d_2 = 0 \) and

\[
B_\Lambda x^*_\Lambda = G(x^*)A_\Lambda x^*_\Lambda.
\]

Conversely, if \( x^* \) satisfies (6.7), set \( v \in \mathbb{R}^n \) to be the vector that \( v_\Lambda = 0 \) and \( v_{A^C} = 2(G(x^*)Ax^* - Bx^*)_{A^C} \). Then, \( v \in \partial \iota_C(x^*) \) and (6.6) holds, that imply inclusion (6.4). We then complete the proof.

6.2. Implementation and convergence of PGSA and PGSA Cũng for problem (6.3). In this subsection, we discuss the implementation of PGSA and PGSAAlso for problem (6.3) and then establish their convergence results.

We note that the proposed algorithms for problem (6.3) mainly involve the computation of proximity operator associated with \( \iota_C \) and the gradients of \( x^TAx \) and
Thus, the computational cost in these algorithms relies heavily on \( \text{prox}_{\phi} \), which is exactly the projection operator onto \( C \), denoted here by \( \text{proj}_C \). We next show that \( \text{proj}_C \) has a closed form and thus can be efficiently computed. To this end, we first recall the projection operator onto the set \( \{ x \in \mathbb{R}^n : \|x\|_0 \leq r \} \), denoted by \( T_r(x) \). It is well-known that for \( x \in \mathbb{R}^n \), \( (T_r(x))_i = x_i \) for the \( r \) largest components in absolute value of \( x \) and \( (T_r(x))_i = 0 \) else. Since the \( r \) largest components may not be uniquely defined, \( T_r \) is a set-valued operator. With the help of \( T_r \) and Proposition 4.3 in [23], we can immediately obtain the closed form of \( \text{proj}_C \) in the following proposition.

**Proposition 6.3.** Given \( x \in \mathbb{R}^n \), then

\[
\text{proj}_C(x) = \begin{cases} 
\{ \frac{y}{\|y\|_2} : y \in T_r(x) \}, & \text{if } x \neq 0, \\
C, & \text{else}.
\end{cases}
\]

Next, we investigate the convergence property of PGSA and PGSA\(_L\) for problem (6.3) based on the general convergence results presented in Section 4.3 and Section 5. To this end, we shall verify Assumptions 1-4 hold for problem (6.3) and \( G \) is a KL function. First, since \( B_\Lambda \) is symmetric positive definite for any subset \( \Lambda \subseteq \mathbb{N}_n \) with \( |\Lambda| \leq r \), then \( \iota_C(x) + x^T Bx \) does not attain 0 for all \( x \in \mathbb{R}^n \). Second, the level boundedness of \( G \) follows from the boundedness of \( C \). In addition, it is obvious that \( \iota_C \) is locally Lipschitz continuous on \( C \) and \( x^T Ax \) is continuously differentiable with a Lipschitz continuous gradient. Finally, according to [2, section 4.3], one can easily check that the function \( G \) is a semialgebraic function and thus satisfies the KL property. Therefore, in view of Theorems 4.9, 5.2 and 5.3, we immediately obtain the following two theorems regarding the convergence of PGSA and PGSA\(_L\) for problem (6.3).

**Theorem 6.4.** Let \( \{x_k : k \in \mathbb{N}\} \) be generated by PGSA and PGSA\(_L\) (PGSA\(_L\) with \( N = 0 \)) for problem (6.3). Then \( \{x_k : k \in \mathbb{N}\} \) converges globally to a critical point of \( G \).

**Theorem 6.5.** Let \( \{x_k : k \in \mathbb{N}\} \) be generated by PGSA\(_NL\) (PGSA\(_L\) with \( N > 0 \)) for problem (6.3). Then \( \{x_k : k \in \mathbb{N}\} \) is bounded and any of its accumulation points is a critical point of \( G \).

### 6.3. Convergence rate of PGSA and PGSA\(_L\) for problem (6.3)

In this subsection, we consider the convergence rate of \( \{x_k : k \in \mathbb{N}\} \) generated by PGSA and PGSA\(_L\) for problem (6.3). By Theorem 6.4, the sequence \( \{x_k : k \in \mathbb{N}\} \) converges to \( x^* \), which is a critical point of \( G \). According to Theorems 4.10 and 5.4, we can further estimate the convergence rate of \( \{x_k : k \in \mathbb{N}\} \) by showing that \( G \) satisfies the KL property at \( x^* \) with \( \phi(s) = ds^{1-\theta} \) for some \( d > 0 \) and \( \theta \in (0,1) \).

To this end, we first prove that the objective function of the generalized eigenvalue problem (without sparsity constraint) satisfies the KL property at its global minimizers with the corresponding \( \phi(s) = ds^{d} \) for some \( d > 0 \) in the following proposition.

**Proposition 6.6.** Given \( D \in \mathbb{S}^n \) and \( E \in \mathbb{S}^m_+ \), let \( \varphi : \mathbb{R}^m \to \mathbb{R} \) be defined at \( x \in \mathbb{R}^m \) as

\[
\varphi(x) := \begin{cases}
\frac{x^TDx}{2}, & \text{if } \|x\|_2 = 1 \text{ and } x^TDx \neq 0, \\
+\infty, & \text{else}.
\end{cases}
\]

Let \( \hat{x} \in \arg\min\{ \varphi(x) : x \in \mathbb{R}^m \} \). Then \( \varphi \) satisfies the KL property at \( \hat{x} \) with the corresponding \( \phi(s) = ds^\frac{d}{2} \) for some \( d > 0 \), i.e., there exist \( d > 0, \eta \in (0, +\infty) \) and a
neighborhood \( U \) of \( \hat{x} \), such that for any \( x \in U \cap \{z \in \mathbb{R}^m : \varphi(x) < \varphi(z) < \varphi(\hat{x}) + \eta\} \),

\[
\text{dist}(0, \hat{\varphi}(x)) \geq \frac{2}{d} \sqrt{\varphi(x) - \varphi(\hat{x})}.
\]

**Proof.** By Lemma 6.1 (ii), we have

\[
\hat{\varphi}(x) = \left\{ tx + \frac{2Ex - 2\varphi(x)Dx}{x^T Dx} : t \in \mathbb{R} \right\}.
\]

Using the fact that \( \langle x, \frac{2Ex - 2\varphi(x)Dx}{x^T Dx} \rangle = 0 \), we deduce that

\[
\text{dist}(0, \hat{\varphi}(x)) = \left\| \frac{2Ex - 2\varphi(x)Dx}{x^T Dx} \right\|_2.
\]

Let \( U \) be a neighborhood of \( \hat{x} \) such that for all \( x \in U \), there hold \( \frac{1}{2} \hat{x}^T \hat{x} \leq x^T Dx \leq 2\hat{x}^T \hat{x} \). Without loss of generality, we may assume \( \omega = \frac{1}{2} \hat{x}^T \hat{x} \leq x^T Dx \leq 2\hat{x}^T \hat{x} \) and \( x^T \hat{x} \neq 0 \). Then, for any \( x \in U \cap \text{dom}(\varphi) \), it holds that

\[
\text{dist}(0, \hat{\varphi}(x)) \geq \sqrt{\frac{\mu}{\lambda_1^2 \hat{x}^T \hat{x}}} \|Ex - \varphi(x)Dx\|_{E^{-1}},
\]

where \( \mu > 0 \) is the smallest eigenvalue of \( E \). Denote by \( \lambda_i \) the \( i \)-th largest eigenvalue of \( E^{-1}D \) for \( i \in \mathbb{N}_m \). The hypothesis \( \hat{x} \in \arg\min\{\varphi(x) : x \in \mathbb{R}^m\} \) yields that \( E^{-1}D\hat{x} = \lambda_1 \hat{x} \) and \( \lambda_1 = 1/\varphi(\hat{x}) \). Hence, for all \( x \in U \cap \text{dom}(\varphi) \), we obtain that

\[
\varphi(x) - \varphi(\hat{x}) = \varphi(x) - \frac{1}{\lambda_1} = \frac{\lambda_1 x^T Ex - x^T Dx}{\lambda_1 x^T Dx} \leq \frac{2}{\lambda_1^2 \hat{x}^T \hat{x}}(\lambda_1 x^T Ex - x^T Dx).
\]

Based on (6.9) and (6.10), it suffices to show that for all \( x \in U \cap \text{dom}(\varphi) \),

\[
\|Ex - \varphi(x)Dx\|^2_{E^{-1}} \geq d_1(\lambda_1 x^T Ex - x^T Dx)
\]

for some \( d_1 > 0 \).

For matrices \( D \) and \( E \), there exists an invertible \( m \times m \) matrix \( P = [p_1, p_2, \ldots, p_m] \) such that \( P^{-1}E^{-1}DP = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_m\} \) and \( P^TEP = I \) with \( p_1 = \frac{1}{\sqrt{\hat{x}^T \hat{x}}} \).

Without loss of generality, we may assume \( \lambda_1 > \lambda_2 \). Then for any \( x \in U \cap \text{dom}(\varphi) \), there exist \( a_i \in \mathbb{R} \), \( i \in \mathbb{N}_m \), such that \( x = \sum_{i=1}^m a_i p_i \). One can easily check that \( x^T Ex = \sum_{i=1}^m a_i^2 \), \( x^T Dx = x^T E^{-1}Dx = \sum_{i=1}^m a_i^2 \lambda_i \), and \( x^T DE^{-1}Dx = x^T DE^{-1}EE^{-1}Dx = \sum_{i=1}^m a_i^2 \lambda_i \). Therefore, by a direct computation we have that

\[
\|Ex - \varphi(x)Dx\|^2_{E^{-1}} = (Ex - \varphi(x)Dx)^T E^{-1}(Ex - \varphi(x)Dx)
\]

\[
= \frac{(x^T Ex)^3}{(x^T Dx)^2} \left( \frac{x^T DE^{-1}Dx}{x^T Ex} - \left( \frac{x^T Dx}{x^T Ex} \right)^2 \right)
\]

\[
\geq \frac{(\hat{x}^T \hat{x})^3}{32(\hat{x}^T \hat{x})^2} \left( \sum_{i=1}^m a_i^2 \lambda_i^2 \sum_{i=1}^m a_i^2 - \left( \sum_{i=1}^m a_i^2 \lambda_i \sum_{i=1}^m a_i^2 \right)^2 \right).
\]

Let \( \omega_i = \frac{a_i^2}{a_i^2} \) for \( i \in \mathbb{N}_m \). It is clear that \( \sum_{i=1}^m \omega_i = 1 \) and \( \omega_i \geq 0 \) for \( i \in \mathbb{N}_m \). Now, substituting \( \omega_i \) into inequality (6.12) and then using the strong convexity of \( \lambda^2 \),
we further obtain that
\[
\|Ex - \varphi(x)Dx\|_{E^{-1}}^2 \geq \frac{(\hat{x}^TE\hat{x})^3}{32(\hat{x}^TD\hat{x})^2} \left( \sum_{i=1}^{m} \omega_i \lambda_i^2 - \left( \sum_{i=1}^{m} \omega_i \lambda_i \right)^2 \right)
\]
\[
\geq \frac{(\hat{x}^TE\hat{x})^3}{32(\hat{x}^TD\hat{x})^2} \omega_1 (1 - \omega_1) \left( \lambda_1 - \frac{\sum_{i=2}^{m} \omega_i \lambda_i}{1 - \omega_1} \right)^2
\]
\[
\geq \frac{(\hat{x}^TE\hat{x})^3}{32(\hat{x}^TD\hat{x})^2} \left( \sum_{i=1}^{m} a_i^2 \right) \left( \sum_{i=1}^{m} a_i^2 - \sum_{i=1}^{m} \lambda_i \right)
\]
\[
= \frac{(\hat{x}^TE\hat{x})^3}{32(\hat{x}^TD\hat{x})^2} \left( \lambda_1 - \frac{\omega_1}{2} \sum_{i=1}^{m} a_i^2 \right)
\]
(6.13)

In addition, from \(p_1 = \frac{\hat{x}}{\hat{x}^TE\hat{x}}\), we get that
\[
x^TE\hat{x} = \left( \sum_{i=1}^{m} a_i p_i \right)^T (Ep_1 \hat{x}^TE\hat{x}) = a_1 \hat{x}^T E\hat{x}.
\]

This relation together with \(x^TE\hat{x} > \frac{1}{2} \hat{x}^TE\hat{x}\) implies that \(a_1 > 1/2\). Using this and noting that \(x^TEx < 2\hat{x}^TE\hat{x}\), we have from (6.13) that
\[
\|Ex - \varphi(x)Dx\|_{E^{-1}}^2 \geq \frac{\hat{x}^T E\hat{x}}{512(\hat{x}^TD\hat{x})^2} (\lambda_1 - \lambda_2)(\lambda_1 x^TEx - x^TDx).
\]
(6.14)

Combing inequalities (6.9), (6.10) and (6.14), we finally obtain that for all \(x \in U \cap \text{dom}(\varphi),\)
\[
\text{dist}(0, \hat{\partial} \varphi(x)) \geq \frac{\sqrt{\mu(\lambda_1 - \lambda_2)}}{32 \hat{x}^TD\hat{x}} \sqrt{\varphi(x) - \varphi(\hat{x})} = \frac{2}{\sqrt{\mu(\lambda_1 - \lambda_2)}} \frac{1}{\sqrt{\mu(\lambda_1 - \lambda_2)}} \sqrt{\varphi(x) - \varphi(\hat{x})}.
\]

Let \(d = \frac{6 \hat{x}^TD\hat{x}}{\sqrt{\mu(\lambda_1 - \lambda_2)}}\) and \(\eta = +\infty\), then we prove the desired result. \(\square\)

We next establish that \(G\) satisfies the KL property at some special points with the corresponding \(\phi(s) = ds^*\) for some \(d > 0\). Before that, we first present a lemma regarding local minimizers of \(G\). We recall \(z \in \mathbb{R}^n\) is a support optimal solution of problem (6.3) on \(\Lambda \subseteq \mathbb{N}_n\), if
\[
z \in \arg\min \{ G(x) : \text{supp}(x) \subseteq \Lambda \}.
\]

**Lemma 6.7.** Let \(\hat{x} \in \mathbb{R}^n\) and \(\hat{\Lambda} := \text{supp}(\hat{x})\). Then \(\hat{x}\) is a support optimal solution of problem (6.3) on any \(\Lambda \subseteq \mathbb{N}_n\) with \(|\Lambda| \leq r\) and \(\hat{\Lambda} \subseteq \Lambda\) if and only if \(\hat{x}\) is a local minimizer of \(G\).

**Proof.** One can easily check that if \(\hat{x}\) is a support optimal solution of problem (6.3) on any \(\Lambda \subseteq \mathbb{N}_n\) with \(|\Lambda| \leq r\) and \(\hat{\Lambda} \subseteq \Lambda\), then \(\hat{x}\) is a local minimizer of \(G\).

Next, we assume that \(\hat{x}\) is a local minimizer of \(G\). By Theorem 3.3 and Proposition 6.2, \(\hat{x}\) satisfies \(B_{\Lambda} \hat{x}_\Lambda = G(\hat{x})A_{\Lambda} \hat{x}_\Lambda\) for any \(\Lambda \subseteq \mathbb{N}_n\) with \(|\Lambda| \leq r\) and \(\hat{\Lambda} \subseteq \Lambda\).

We next prove this lemma by contradiction. Suppose there exists \(\Lambda \subseteq \Lambda \subseteq \mathbb{N}_n\) with \(|\Lambda| \leq r\) such that \(\hat{x} \notin \arg\min \{ G(x) : \text{supp}(x) \subseteq \Lambda \}\). This yields that \(\hat{x}_\Lambda \notin \text{supp}(\hat{x})\) for any \(\Lambda \subseteq \Lambda \subseteq \mathbb{N}_n\) with \(|\Lambda| \leq r\).
arg min \{ \frac{x^T B A x}{x^T A x} : \| x \|_2 = 1 \}, implying that 1/G(\tilde{x}) is not the largest eigenvalue of 
B^\Lambda_1 A_\Lambda. Then, there exists x^* \in C with \text{supp}(x^*) \subseteq \Lambda and \tilde{x}_\Lambda^T B_\Lambda x^*_\Lambda = 0 such that

x^*_\Lambda \in \arg min \{ \frac{x^T B_\Lambda x}{x^T A_\Lambda x} : \| x \|_2 = 1 \},

that is, B_\Lambda x^*_\Lambda = G(x^*) A_\Lambda x^*_\Lambda and G(x^*) is the largest eigenvalue of B^\Lambda_1 A_\Lambda.

Let x_\epsilon = \frac{\tilde{x} + \epsilon x^*}{\| \tilde{x} + \epsilon x^* \|_2} for \epsilon > 0. It is clear that supp(x_\epsilon) \subseteq \Lambda and x_\epsilon \in C. It follows that

G(x_\epsilon) = \frac{(\tilde{x}_\Lambda + \epsilon x^*_\Lambda)^T B_\Lambda (\tilde{x}_\Lambda + \epsilon x^*_\Lambda)}{(\tilde{x}_\Lambda + \epsilon x^*_\Lambda)^T A_\Lambda (\tilde{x}_\Lambda + \epsilon x^*_\Lambda)} = \frac{G(\tilde{x}) \tilde{x}_\Lambda^T A_\Lambda \tilde{x}_\Lambda + \epsilon^2 G(x^*) (x^*_\Lambda)^T A_\Lambda x^*_\Lambda}{\tilde{x}_\Lambda^T A_\Lambda \tilde{x}_\Lambda + \epsilon^2 (x^*_\Lambda)^T A_\Lambda x^*_\Lambda},

where the last equality holds since \tilde{x}_\Lambda^T B_\Lambda x^*_\Lambda = 0 implies \tilde{x}_\Lambda^T A_\Lambda x^*_\Lambda = 0. One can check that G(x_\epsilon) < G(\tilde{x}) for any \epsilon > 0 due to the relations that \tilde{G}(\tilde{x}) > G(x^*), \tilde{x}_\Lambda^T A_\Lambda \tilde{x}_\Lambda > 0 and (x^*_\Lambda)^T A_\Lambda x^*_\Lambda > 0. This contradicts to the fact that \tilde{x} is a local minimizer of G.

We obtain this lemma immediately. \( \square \)

Now, we are ready to prove that G satisfies the KL property at any local minimizer of G with the corresponding \phi(s) = ds^\perp for some d > 0.

\textbf{Proposition 6.8.} If \tilde{x} \in \mathbb{R}^n is a local minimizer of G, then G satisfies the KL property at \tilde{x} with the corresponding \phi(s) = ds^\perp for some d > 0.

\textbf{Proof.} Let \Lambda := \text{supp}(\tilde{x}) and obviously it holds that \| \Lambda \| \leq r. Given \Lambda \subseteq \mathbb{N}_n, let \varphi_\Lambda be the function \varphi which is defined in (6.8) with respect to \Lambda = D_\Lambda, E = B_\Lambda.

First, we consider the case of \| \Lambda \| = r. Since \tilde{x} is a local minimizer of G, by Lemma 6.7, \tilde{x}_\Lambda is a global minimizer of \varphi_\Lambda. By Proposition (6.6), there exists \delta_1 > 0 and d > 0, such that for all \epsilon \in U(\tilde{x}_\Lambda, \delta_1) \cap \{ \epsilon \in \mathbb{R}^n : \varphi_\Lambda(\tilde{x}_\Lambda) < \varphi_\Lambda(\epsilon) < +\infty \},

(6.16) \quad \text{dist}(0, \partial \varphi_\Lambda(\epsilon)) \geq \frac{2}{d} \sqrt{\varphi_\Lambda(\epsilon) - \varphi_\Lambda(\tilde{x}_\Lambda)}.

Let \delta := \min\{\delta_1, \delta_2\} with \delta_2 := \frac{1}{2} \min\{\| \tilde{x}_j \| : j \in \Lambda \}. The facts that \| \Lambda \| = r and G(\tilde{x}) = \varphi_\Lambda(\tilde{x}_\Lambda) imply that for all \epsilon \in U(\tilde{x}, \delta) \cap \{ \epsilon \in \mathbb{R}^n : G(\tilde{x}) < \inf G(\epsilon) < +\infty \},

(6.17) \quad \text{supp}(x) = \Lambda and x_\Lambda \in U(\tilde{x}_\Lambda, \delta_1) with \varphi_\Lambda(\tilde{x}_\Lambda) < G(x) = \varphi_\Lambda(x_\Lambda) < +\infty.

In addition, by Lemma 6.1, we have for any \epsilon \in \text{dom}(G) satisfying supp(x) = \Lambda,

(6.18) \quad \text{dist}(0, \partial G(x)) \geq \text{dist}(0, \{ t x_\Lambda + \frac{2B_\Lambda x_\Lambda - 2G(x) A_\Lambda x_\Lambda}{x_\Lambda^T A_\Lambda x_\Lambda} : t \in \mathbb{R} \}).

Therefore, using this inequality we further obtain from (6.16) and (6.17) that for all \epsilon \in U(\tilde{x}, \delta) \cap \{ \epsilon \in \mathbb{R}^n : G(\tilde{x}) < \inf G(\epsilon) < +\infty \},

(6.19) \quad \text{dist}(0, \partial G(x)) \geq \text{dist}(0, \partial \varphi_\Lambda(x_\Lambda)) \geq \frac{2}{d} \sqrt{\varphi_\Lambda(x_\Lambda) - \varphi_\Lambda(\tilde{x}_\Lambda)} = \frac{2}{d} \sqrt{G(x) - G(\tilde{x})},

which indicates that G satisfies the KL property at \tilde{x} with the corresponding \phi(s) = ds^\perp in the case of \| \Lambda \| = r.

Next, we consider the case of \| \Lambda \| < r. Because \tilde{x} is a local minimizer of G, from Lemma 6.7, we have \tilde{x}_\Lambda \in \arg min \{ \varphi_\Lambda(z) : z \in \mathbb{R}^{|\Lambda|} \} for any \Lambda \subseteq \Lambda \subseteq \mathbb{N}_n with
problem (6.3). All algorithms are conducted in Matlab.

Let \( \hat{\delta} := \min\{\delta_1 : \hat{A} \subseteq A \subseteq N_n, |A| \leq r\} \) and \( \hat{d} := \max\{|\Lambda| : \Lambda \subseteq N_n, |\Lambda| \leq r\} \). Set \( \Lambda := \text{supp}(x) \). Take any \( x \in U(\hat{x}, \hat{\delta}) \cap \{x \in \mathbb{R}^n : G(x) < G(x) < +\infty \} \) and set \( \Lambda := \text{supp}(x) \). Then we immediately see that \( \hat{A} \subseteq \Lambda \) with \( |\Lambda| \leq r \), \( G(x) = \varphi_L(x) \) and \( G(\hat{x}) = \varphi_L(\hat{x}) \). In addition, one can check that \( x_0 \in U(\hat{x}_L, \delta) \cap \{z \in \mathbb{R}^{|\Lambda|} : \varphi_L(\hat{x}_L) < \varphi_L(z) < +\infty\} \). Therefore, we have that

\[
\text{dist}(0, \hat{d}G(x)) = \text{dist}(0, \hat{d}\varphi_L(x_{\Lambda})) \geq \frac{2}{d_L} \sqrt{\varphi_L(x_{\Lambda}) - \varphi_L(\hat{x}_L)} = \frac{2}{d} \sqrt{G(x) - G(\hat{x})}.
\]

We then complete the proof.

With the help of Theorems 4.10, 5.4, 6.4 and Proposition 6.8, we can prove that if the limit point of \( \{x^k : k \in \mathbb{N}\} \) generated by PGSA or PGSA_{ML} for problem (6.3) is a local minimizer of \( G \), then \( \{x^k : k \in \mathbb{N}\} \) converges to it R-linearly. We present this result in the following theorem.

**Theorem 6.9.** Let \( \{x^k : k \in \mathbb{N}\} \) be generated by PGSA or PGSA_{ML} for problem (6.3) and \( x^* \) be the limit point of \( \{x^k : k \in \mathbb{N}\} \). If \( x^* \) is a local minimizer of \( G \), then \( \{x^k : k \in \mathbb{N}\} \) converges to \( x^* \) R-linearly.

If the initial point is close enough to a global minimizer of \( G \), we further have the following convergence result, concerning PGSA and PGSA_{ML} for problem (6.3).

**Corollary 6.10.** Let \( \hat{x} \in \mathbb{R}^n \) be a global minimizer of \( G \). Then there exists \( \delta > 0 \), such that the sequence \( \{x^k : k \in \mathbb{N}\} \) generated by PGSA or PGSA_{ML} for problem (6.3) with \( \|x^0 - \hat{x}\|_2 < \delta \) converges R-linearly to a global minimizer of \( G \).

**Proof.** By Theorem 6.4 and Theorem 2.12 in [3], there exists \( \delta > 0 \), such that \( \{x^k : k \in \mathbb{N}\} \), which starts from \( x^0 \) satisfying \( \|x^0 - \hat{x}\|_2 < \delta \), converges to a global minimizer \( \hat{x} \) of \( G \). We then obtain the desired result immediately from Theorem 6.9.

To close this section, we point out the relation between PGSA for problem (6.3) and an existing algorithm for SGEP. Very recently, in [35] the authors propose a truncated Rayleigh flow method (TRFM) for solving SGEP and show that TRFM converges R-linearly to a global minimizer of \( G \) when the initial point \( x^0 \) is close enough to that global minimizer. By appropriate reformulations, we observe that TRFM essentially coincides with PGSA for problem (6.3) with a constant step size in \( (0, 1/L) \). Although the convergence result in Corollary 6.10 seems very similar to that of TRFM, there are great differences between their convergence proof. The convergence of TRFM is established mainly from the viewpoint of statistics, while our convergence analysis for PGSA is primarily based on the KL property of the objective in problem (6.3). In addition, we show that for arbitrary starting points, PGSA converges to a critical point of \( G \), and the convergence rate is also R-linear when the critical point is a local minimizer of \( G \). However, in [35] there is no convergence guarantee for TRFM starting from an arbitrary initial point.

**7. Numerical experiments.** In this section, we conduct some numerical experiments to test the efficiency of our proposed algorithms, namely, PGSA, PGSA_{ML} and PGSA_{NL} for solving problem (6.3). All algorithms are conducted in Matlab R2015b on a Dell desktop with an Intel(R) Core(TM) i5-8500 CPU (3.00GHz) and 8GB of RAM.
In our experiments, we set the parameters of the aforementioned algorithms as follows. For PGSA, we set \( \alpha_k = 0.99/\|B\|_2 \) for \( k \in \mathbb{N} \). We notice that using this setting of parameters, PGSA for problem (6.3) is essentially the same as TRFM in [35] as mentioned at the end of Section 6. For PGSA\_ML and PGSA\_NL, we set \( a = 10^{-3}, \bar{a} = 10^{-8}, \tilde{a} = 10^8, \) and \( \eta = 0.5 \). Also, \( N \) is set to be 4 in PGSA\_NL. In addition, we choose \( \alpha_{0,r,0} = 0.99/\|B\|_2 \) and \( \alpha_{k,0} \) via formula (5.2) for \( k \in \mathbb{N} \).

In our experiments, we focus on binary classification problem using sparse Fisher’s discriminant analysis (SFDA) which is a special instance of SGEP. Given \( p \) data samples \( \{z^1, z^2, \ldots, z^p\} \) consisting of two distinct classes with \( n \) features, let \( \mathcal{I}_k \subseteq \mathbb{N} \) be the index set of samples in the \( k \)-th class and denote \( |\mathcal{I}_k| \) by \( p_k \) (\( k = 1 \) or 2). The within-class and between-class covariance matrices are defined as:

\[
\hat{\Sigma}_w := \frac{1}{p} \sum_{k=1}^{2} \sum_{i \in \mathcal{I}_k} (z^i - \hat{u}^k)(z^i - \hat{u}^k)^T \quad \text{and} \quad \hat{\Sigma}_b := \frac{1}{p} \sum_{k=1}^{2} p_k \hat{u}^k(\hat{u}^k)^T,
\]

where \( \hat{u}^k := \sum_{i \in \mathcal{I}_k} z^i/p_k \) for \( k = 1, 2 \). For an integer \( r \in [1, n] \), the SFDA seeks a sparse projection vector by solving the following problem:

\[
\max \left\{ \frac{x^T \hat{\Sigma}_b x}{x^T \hat{\Sigma}_w x} : \|x\|_0 \leq r, \|x\|_2 = 1, \ x \in \mathbb{R}^n \right\}.
\]

As mentioned in Section 6, the above problem can be reformulated as

\[
\min \left\{ \frac{x^T \hat{\Sigma}_w x}{x^T \hat{\Sigma}_b x} : \|x\|_0 \leq r, \|x\|_2 = 1, \ x^T \hat{\Sigma}_b x \neq 0, \ x \in \mathbb{R}^n \right\}.
\]

Clearly, problem (7.1) is a special case of problem (6.3) with \( A = \hat{\Sigma}_b \) and \( B = \hat{\Sigma}_w \). Therefore, the proposed algorithms are all applicable. Below we perform numerical tests of them for problem (6.3).

We consider a simulation setting similar to that of [35]. The samples of the \( k \)-th class are randomly generated following a Gaussian distribution with mean \( u^k \) and covariance \( \Sigma \) for \( k = 1 \) and 2. We set \( u^1 = 0, u^2_j = 0.5 \) for \( j \in \{2, 4, \ldots, 40\} \) and \( u^2_j = 0 \) otherwise. Meanwhile, we let \( \Sigma \) be a block diagonal matrix with five blocks, each of which is in the dimension \( \sqrt{n}/5 \times \sqrt{n}/5 \). The \( (j, j') \)-th entry of each block takes value \( 0.8^{|j-j'|} \). We fix \( p = 1000, p_1 = p_2 = 500 \) and use different values for \( n \in \{500, 1000, 2000\} \), while the sparsity rate \( r/n \) is varied from \( \{0.05, 0.1, 0.2\} \) for a fixed \( n \). For each \( (n, r) \), we generate 100 instances of two-class dataset randomly as described above. Then we perform all the algorithms for the corresponding problem (7.1). The initial point \( x^0 \) is chosen as \( x^0_i = 1/\sqrt{r} \) for \( i \in \mathbb{N}_r \) and \( x^0_i = 0 \) otherwise. All the algorithms are terminated when the number of iterations hits 6000 or the successive changes of the iterates are small enough, i.e., \( \|x^k - x^{k-1}\|_2 \leq 10^{-6} \).

Table 1 reports the computational results averaged over 100 random instances, when the iterates of the algorithms first satisfy \( |G(x^k) - G(x^{k-1})| < \epsilon \) for \( \epsilon \in \{10^{-6}, 10^{-7}, 10^{-8}\} \). In each row for a specific number, the three columns in a bracket give the averaged iteration number, CPU time and objective value from left to right respectively. We observe that PGSA\_ML and PGSA\_NL substantially outperform PGSA in terms of CPU time, while the objective values achieved by them are slightly lower than that by PGSA. This demonstrates the effectiveness of the line search schemes involved in PGSA\_ML and PGSA\_NL. Besides, the performance of PGSA\_ML is comparable to that of PGSA\_NL.
Table 1

Performance comparison among PGSA, PGSA\_ML and PGSA\_NL for problem (7.1). For a given $\epsilon$, the first column in the bracket gives the first iteration number $k$ such that $|G(x^k) - G(x^{k-1})| < \epsilon$, while the second and third columns give the corresponding CPU time (in second) and objective values ($\times 10^{-2}$) respectively.

| $\epsilon$ | $p$ | $r/p$ | PGSA (TRFM) | PGSA\_ML | PGSA\_NL |
|------------|-----|------|-------------|-----------|-----------|
| 10^{-6}    | 500 | 0.05 | (322, 0.05, 18.33) | (63, 0.03, 15.44) | (54, 0.03, 14.99) |
|            |     | 0.1  | (202, 0.04, 8.11)  | (50, 0.03, 7.87)  | (47, 0.02, 7.86)  |
|            |     | 0.2  | (207, 0.04, 6.94)  | (69, 0.03, 6.65)  | (65, 0.03, 6.61)  |
|            | 1000| 0.05 | (263, 0.32, 8.22)  | (51, 0.16, 7.78)  | (47, 0.16, 7.76)  |
|            |     | 0.1  | (222, 0.28, 6.70)  | (73, 0.18, 6.22)  | (65, 0.16, 6.17)  |
|            |     | 0.2  | (366, 0.39, 5.00)  | (109, 0.22, 4.49) | (94, 0.19, 4.48)  |
|            | 2000| 0.05 | (257, 3.60, 6.34)  | (72, 2.97, 5.69)  | (69, 2.90, 5.62)  |
|            |     | 0.1  | (356, 4.24, 4.43)  | (107, 3.35, 3.73) | (93, 3.23, 3.62)  |
|            |     | 0.2  | (794, 5.90, 1.93)  | (167, 3.61, 1.60) | (137, 3.34, 1.53) |
| 10^{-7}    | 500 | 0.05 | (422, 0.05, 18.32) | (79, 0.03, 15.44) | (66, 0.03, 14.98) |
|            |     | 0.1  | (314, 0.05, 8.10)  | (77, 0.03, 7.86)  | (69, 0.03, 7.85)  |
|            |     | 0.2  | (324, 0.05, 6.93)  | (103, 0.03, 6.62) | (88, 0.03, 6.58)  |
|            | 1000| 0.05 | (394, 0.42, 8.21)  | (80, 0.21, 7.76)  | (69, 0.20, 7.75)  |
|            |     | 0.1  | (358, 0.38, 6.67)  | (116, 0.23, 6.18) | (94, 0.21, 6.14)  |
|            |     | 0.2  | (581, 0.55, 4.95)  | (161, 0.28, 4.44) | (134, 0.25, 4.38) |
|            | 2000| 0.05 | (424, 4.28, 6.31)  | (119, 3.31, 5.63) | (98, 3.16, 5.59)  |
|            |     | 0.1  | (569, 5.19, 4.38)  | (165, 3.79, 3.64) | (134, 3.57, 3.56) |
|            |     | 0.2  | (1390, 8.37, 1.84) | (276, 4.43, 1.48) | (221, 4.00, 1.42) |
| 10^{-8}    | 500 | 0.05 | (526, 0.06, 18.32) | (96, 0.03, 15.44) | (79, 0.03, 14.98) |
|            |     | 0.1  | (434, 0.05, 8.10)  | (108, 0.03, 7.86) | (90, 0.03, 7.85)  |
|            |     | 0.2  | (450, 0.06, 6.92)  | (136, 0.04, 6.61) | (110, 0.04, 6.57) |
|            | 1000| 0.05 | (532, 0.53, 8.21)  | (107, 0.25, 7.76) | (92, 0.24, 7.75)  |
|            |     | 0.1  | (508, 0.49, 6.67)  | (155, 0.29, 6.16) | (122, 0.26, 6.13) |
|            |     | 0.2  | (790, 0.71, 4.94)  | (210, 0.35, 4.42) | (166, 0.30, 4.37) |
|            | 2000| 0.05 | (603, 5.02, 6.30)  | (165, 3.67, 5.61) | (130, 3.46, 5.57) |
|            |     | 0.1  | (820, 6.28, 4.36)  | (232, 4.34, 3.60) | (167, 3.89, 3.55) |
|            |     | 0.2  | (1946, 10.70, 1.81)| (384, 5.28, 1.45) | (300, 4.74, 1.38) |

Next, we study the convergence rate of the proposed algorithms. We verify that in all the tests the approximate solution obtained by each of the algorithms is a support optimal solution of problem (7.1) with $r$ nonzero elements as defined in (6.15) and thus is a local minimizer from Lemma 6.7. In view of Theorem 6.9, one can expect to see R-linear convergence of the sequence generated by PGSA and PGSA\_ML. We plot $\|x^k - x^*\|_2$ (in logarithmic scale) against the number of iterations in Figure 1, where $x^*$ is the approximated solution produced by the corresponding algorithm. It is obvious that the sequence generated by PGSA\_ML or PGSA\_NL converges much faster than that by PGSA. As can be seen from Figure 1, the sequence generated by PGSA or PGSA\_ML appears to converge R-linearly, which confirms with Theorem 6.9. Finally, we remark that although we have no theoretical results concerning the convergence rate or even convergence of the whole sequence generated by PGSA\_NL, that sequence also seems to converge R-linearly and its convergence rate is slightly faster than that of PGSA\_ML.
8. Conclusion. In this paper, we study a class of single-ratio fractional optimization problems that appears frequently in applications. The numerator part of the objective is the sum of a nonsmooth nonconvex function and a nonconvex smooth function, while the denominator part is a nonsmooth convex function. We derive a first-order necessary optimality condition for this problem and develop for it first-order algorithms, namely, PGSA, PGSA\_ML and PGSA\_NL. We show the sub-sequential convergence of the sequence generated by the proposed algorithms under mild assumptions. Moreover, we establish global convergence of the whole sequence generated by PGSA or PGSA\_ML and estimate the convergence rate by additional assumptions on the objective. The proposed algorithms are further applied to solve the sparse generalized eigenvalue problems and their convergence results for the problem are gained according to the general convergence theorems for them. Finally, we conduct some preliminary numerical experiments to illustrate the efficiency of the proposed algorithms.

Appendix A. Proof of Proposition 2.2.

Proof. When \( x \) is an isolated point of \( \text{dom}(\rho) \), it is trivial that \( \partial \rho(x) = \partial(a_2 f_1 - a_1 f_2) = \mathbb{R}^n \). We next consider the case that \( x \) is not an isolated point of \( \text{dom}(\rho) \). For any \( u \in \mathbb{R}^n \) and \( v \in \mathbb{R}^n \), a direct computation yields

\[
\frac{f_1(u) - a_1}{a_2} - \langle v, u - x \rangle \|u - x\|_2 = a_2 f_1(u) - a_1 f_2(u) - \langle a_2 v, u - x \rangle \frac{a_2}{a_2^2} \|u - x\|_2 + R(x, u),
\]

where \( R(x, u) = (a_2 - f_2(u))(a_2 f_1(u) - a_1 f_2(u))/(a_2^2 f_2(u)\|u - x\|_2) \). Since \( f_2 \) satisfies the calmness condition and \( f_1 \) is continuous at \( x \), we get that \( \liminf_{u \to x} R(x, u) = 0 \). Noting this fact and by the definition of Fréchet subdifferential, we have

\[
\partial \rho(x) = \left\{ v \in \mathbb{R}^n : \liminf_{u \to x} \frac{f_1(u) - a_1}{a_2} \frac{a_2}{a_2^2} \|u - x\|_2 \geq 0 \right\} = \left\{ v \in \mathbb{R}^n : \liminf_{u \to x} \frac{a_2 f_1(u) - a_1 f_2(u) - \langle a_2 v, u - x \rangle}{a_2^2 \|u - x\|_2} \geq 0 \right\} = \frac{\partial(a_2 f_1 - a_1 f_2)(x)}{a_2^2}.
\]

We complete the proof. □

Appendix B. Proof of Proposition 3.1.
Proof. We only need to prove the proposition holds for local minimizers, since the conclusion for global minimizers can be proven similarly.

Suppose \( x^* \) is a local minimizer of problem (1.1). Then, there exists \( \delta > 0 \) such that for any \( x \in B(x^*, \delta) \cap \text{dom}(F) \), there holds

\[
0 \leq \frac{f(x) + h(x)}{g(x)} - \frac{f(x^*) + h(x^*)}{g(x^*)}.
\]

This indicates that

\[
0 \leq f(x) + h(x) - \frac{f(x^*) + h(x^*)}{g(x^*)} g(x) = f(x) + h(x) - c(x) g(x)
\]

for all \( x \in B(x^*, \delta) \cap \text{dom}(F) \), since \( g(x) > 0 \). Due to the facts that the objective function value of problem (3.1) at \( x^* \) is 0, we have that \( x^* \) is a local minimizer of problem (3.1).

Conversely, if \( x^* \) is a local minimizer of problem (3.1), then (B.2) holds for \( x \in B(x^*, \delta) \cap \text{dom}(F) \) with some \( \delta > 0 \). By simple calculation, we obtain that (B.1) holds for \( x \in B(x^*, \delta) \cap \text{dom}(F) \). This implies that \( x^* \) is a local minimizer of problem (1.1). We then complete the proof. \( \square \)

Appendix C. Proof of Lemma 6.1.

Proof. By the definition of Fréchet subdifferential, we have that

\[
\partial_{tC}(x) = \left\{ v \in \mathbb{R}^n : \liminf_{y \to \mathcal{C}} \frac{(v^t x - y^t)}{2} \geq 0 \right\}.
\]

Let \( \Lambda := \text{supp}(x) \). We first prove Item (i). In the case that \( |\Lambda| = r \), there exists a neighborhood \( U \) of \( x \), such that \( \text{supp}(y) = \Lambda \) for all \( y \in U \cap \mathcal{C} \). Thus, we obtain that

\[
\partial_{tC}(x) = \left\{ v \in \mathbb{R}^n : \liminf_{y \to \mathcal{C}} \frac{(v^t x - y^t)}{2} \geq 0 \right\} = \left\{ v \in \mathbb{R}^n : v = tx, t \in \mathbb{R} \right\}.
\]

Next we consider the case when \( |\Lambda| < r \). For any \( t \in \mathbb{R} \), we have

\[
\lim_{y \to \mathcal{C}} \frac{(tx, x - y)}{2} = \lim_{y \to \mathcal{C}} \frac{t}{2} \left( \frac{x^2}{2} + \frac{y^2}{2} - 2x^Ty \right) = \lim_{y \to \mathcal{C}} t \frac{1 - x^Ty}{2(1 - x^Ty)}.\]

Hence, we see that \( \{ v \in \mathbb{R}^n : v = tx, t \in \mathbb{R} \} \subseteq \partial_{tC}(x) \). We further note that for any \( v \in \partial_{tC}(x) \),

\[
0 \leq \liminf_{y \to \mathcal{C}} \frac{(v^t x - y^t)}{2} \leq \liminf_{y \to \mathcal{C}} \frac{(v^t x - y^t)}{2} = \frac{|v^t x - y^t|}{2}.
\]

which indicates that \( v = tx \) for some \( t \in \mathbb{R} \). Finally, we show that for all \( v \in \partial_{tC}(x) \), \( v_j = 0 \) if \( j \notin \Lambda \). Otherwise, there exists \( \tilde{v} \in \partial_{tC}(x) \) and \( j_0 \notin \Lambda \) such that \( \tilde{v}_{j_0} \neq 0 \). Choose \( \{ y_k : k \in \mathbb{N} \} \) such that \( y^k_{j_0} = \sqrt{1 - 1/\|x\|^2} x_{j_0} \), \( y^k_{j_0} = v_{j_0} / (\|v_{j_0}\|) \), and \( y^k_{j_0} = 0 \) for \( j \in \Lambda \cup \{ j_0 \} \). Then we have that \( \{ y_k : k \in \mathbb{N} \} \subseteq \mathcal{C} \) and \( \lim_{k \to \infty} y^k = x \). One can verify that

\[
\lim_{k \to \infty} \frac{(\tilde{v}, x - y_k)}{2} = \frac{|\tilde{v}_{j_0}|}{2} < 0,
\]

which contradicts \( \tilde{v} \in \partial_{tC}(x) \). This proves Item (i).
We turn to Item (ii). Let \( v \in \partial I_C(x) \). Then there exist \( x^k \in C \) and \( v^k \in \partial I_C(x^k) \) such that \( \lim_{k \to \infty} v^k = v \). Thus, we have \( \Lambda \subseteq \text{supp}(x^k) \) when \( k \geq K \) for some \( K \in \mathbb{N} \). By item (i), there exists \( \{t_k \in \mathbb{R} : k \geq K \} \) such that \( v^k_\Lambda = t_k x^k_\Lambda \) for \( k \geq K \). Therefore, we deduce that \( v_\Lambda = t x_\Lambda \) for some \( t \in \mathbb{R} \). \( \square \)

REFERENCES

[1] **Hedy Attouch and Jerome Bolte**, On the convergence of the proximal algorithm for nondifferentiable functions involving analytic features, Mathematical Programming, 116 (2009), pp. 5–16.

[2] **Hedy Attouch, Jerome Bolte, Patrick Redont, and Antoine Soueyrev**, Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-lojasiewicz inequality, Mathematics of Operations Research, 35 (2010), pp. 438–457.

[3] **Hedy Attouch, Jerome Bolte, and Benar Fux Svaiter**, Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gauss-seidel methods, Mathematical Programming, 137 (2013), pp. 91–129.

[4] **Roberto Baldacci, Andrew Lim, Emiliano Traversi, and Roberto Wöfler Calvo**, Optimal solution of vehicle routing problems with fractional objective function, arXiv preprint arXiv:1804.03316, (2018).

[5] **Jonathan Barzilai and Jonathan M Borwein**, Two-point step size gradient methods, IMA Journal of Numerical Analysis, 8 (1988), pp. 141–148.

[6] **Amir Beck and Yonina C Eldar**, Sparsity constrained nonlinear optimization: Optimality conditions and algorithms, SIAM Journal on Optimization, 23 (2013), pp. 1480–1509.

[7] **Harold P Benson**, Fractional programming with convez quadratic forms and functions, European Journal of Operational Research, 173 (2006), pp. 351–369.

[8] **Jerome Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter**, From error bounds to the complexity of first-order descent methods for convex functions, Mathematical Programming, 165 (2017), pp. 471–507.

[9] **Abraham Charnes and William W Cooper**, Programming with linear fractional functionals, Naval Research logistics quarterly, 9 (1962), pp. 181–186.

[10] **Xin Chen, Changliang Zou, and R Dennis Cook**, Coordinate-independent sparse sufficient dimension reduction and variable selection, The Annals of Statistics, 38 (2010), pp. 3696–3723.

[11] **Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll**, Sparse discriminant analysis, Technometrics, 53 (2011), pp. 406–413.

[12] **Alexandre d'Aspremont, Francis Bach, and Laurent El Ghaoui**, Optimal solutions for sparse principal component analysis, Journal of Machine Learning Research, 9 (2008), pp. 1269–1294.

[13] **Werner Dinkelbach**, On nonlinear fractional programming, Management Science, 13 (1967), pp. 492–498.

[14] **Patrik O Hoyer**, Non-negative matrix factorization with sparseness constraints, Journal of Machine Learning Research, 5 (2004), pp. 1457–1469.

[15] **Tosihhide Ibaraki**, Parametric approaches to fractional programs, Mathematical Programming, 26 (1983), pp. 345–362.

[16] **Raj Jagannathan**, On some properties of programming problems in parametric form pertaining to fractional programming, Management Science, 12 (1966), pp. 609–615.

[17] **Hiroshi Konno and Michimori Inori**, Bond portfolio optimization by bilinear fractional programming, Journal of the Operations Research Society of Japan, 32 (1989), pp. 143–158.

[18] **Hiroshi Konno, Katsuhiro Tsuchiya, and Rei Yamamoto**, Minimization of the ratio of functions defined as sums of the absolute values, Journal of Optimization Theory and Applications, 135 (2007), pp. 399–410.

[19] **Lexin Li and Christopher J Nachtsheim**, Sparse sliced inverse regression, Technometrics, 48 (2006), pp. 502–510.

[20] **Zhaosong Lu and Zirui Zhou**, Nonmonotone enhanced proximal dc algorithms for a class of structured nonsmooth dc programming, SIAM Journal on Optimization, 29 (2019), pp. 2725–2752.

[21] **Zhi-Quan Luo and Jong-Shi Pang**, Error bounds for analytic systems and their applications, Mathematical Programming, 67 (1994), pp. 1–28.

[22] **Zhi-Quan Luo and Paul Tseng**, Error bounds and convergence analysis of feasible descent
methods: a general approach, Annals of Operations Research, 46 (1993), pp. 157–178.

[23] Ronny Luss and Marc Teboulle, Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint, SIAM Review, 55 (2013), pp. 65–98.

[24] Qing Mai, Hui Zou, and Ming Yuan, A direct approach to sparse discriminant analysis in ultra-high dimensions, Biometrika, 99 (2012), pp. 29–42.

[25] Jong-Shi Pang, A parametric linear complementarity technique for optimal portfolio selection with a risk-free asset, Operations Research, 28 (1980), pp. 927–941.

[26] Error bounds in mathematical programming, Mathematical Programming, 79 (1997), pp. 299–332.

[27] Jong-Shi Pang and Patrick Lee, A parametric linear complementarity technique for the computation of equilibrium prices in single commodity spatial model, Mathematical Programming, 20 (1981), pp. 81–102.

[28] Panos M Pardalos and Andrew T Phillips, Global optimization of fractional programs, Journal of Global Optimization, 1 (1991), pp. 173–182.

[29] Panos M Pardalos, Mattias Sandström, and Costas Zopounidis, On the use of optimization models for portfolio selection: A review and some computational results, Computational Economics, 7 (1994), pp. 227–244.

[30] R. Tyrrell Rockafellar and Roger J. B. Wets, Variational analysis, Springer, 2004.

[31] Siegfried Schaible, Fractional programming, Handbook of Global Optimization. Horst, R., Pardalos, PM (eds.), (1995), pp. 495–608.

[32] Siegfried Schaible and Toshidide Ibaraki, Fractional programming, European Journal of Operational Research, 12 (1983), pp. 325–338.

[33] Kaiming Shen and Wei Yu, Fractional programming for communication systems part i: Power control and beamforming, IEEE Transactions on Signal Processing, 66 (2018), pp. 2616–2630.

[34] Ioan M Stanču-Minasian, Fractional programming: theory, methods and applications, vol. 409, Springer Science & Business Media, 2012.

[35] Kean Ming Tan, Zhaoran Wang, Han Liu, and Tong Zhang, Sparse generalized eigenvalue problem: Optimal statistical rates via truncated Rayleigh flow, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80 (2018), pp. 1057–1086.

[36] Bo Wen, Xiaojun Chen, and Ting Kei Pong, A proximal difference-of-convex algorithm with extrapolation, Computational Optimization and Applications, 69 (2018), pp. 297–324.

[37] Stephen J Wright, Robert D Nowak, and Mario A Figueiredo, Sparse reconstruction by separable approximation, IEEE Transactions on Signal Processing, 57 (2009), pp. 2479–2493.

[38] Yangyang Xu and Wotao Yin, A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion, SIAM Journal on Imaging Sciences, 6 (2013), pp. 1758–1789.

[39] Alessio Zappone, Emil Björnson, Luca Sanguinetti, and Eduard Jorswieck, Globally optimal energy-efficient power control and receiver design in wireless networks, IEEE Transactions on Signal Processing, 65 (2017), pp. 2844–2859.

[40] Alessio Zappone, Luca Sanguinetti, and Mérouane Debbah, Energy-delay efficient power control in wireless networks, IEEE Transactions on Communications, 66 (2017), pp. 418–431.

[41] Hui Zou, Trevor Hastie, and Robert Tibshirani, Sparse principal component analysis, Journal of Computational and Graphical Statistics, 15 (2006), pp. 265–286.