KL property of exponent 1/2 of quadratic functions under nonnegative zero-norm constraints and applications

Shaohua Pan, Yuqia Wu and Shujun Bi

April 9, 2019

Abstract

This paper focuses on the quadratic optimization over two classes of nonnegative zero-norm constraints: nonnegative zero-norm sphere constraint and zero-norm simplex constraint, which have important applications in nonnegative sparse eigenvalue problems and sparse portfolio problems, respectively. We establish the KL property of exponent 1/2 for the extended-valued objective function of these nonconvex and nonsmooth optimization problems, and use this crucial property to develop a globally and linearly convergent projection gradient descent (PGD) method. Numerical results are included for nonnegative sparse principal component analysis and sparse portfolio problems with synthetic and real data to confirm the theoretical results.

Keywords: KL property of exponent 1/2; zero-norm constraint; sphere; simplex set

1 Introduction

For a given set $C \subset \mathbb{R}^p$, denote by $\delta_C$ the indicator function over the set $C$, i.e., $\delta_C(z) = 0$ if $z \in C$ and otherwise $\delta_C(z) = +\infty$. Let $\Omega := \{ x \in \mathbb{R}^p \mid \|x\|_0 \leq \kappa \}$ for an integer $\kappa \geq 1$ be the zero-norm constraint set, and $\mathbb{R}^p_+$ be the nonnegative orthant cone in $\mathbb{R}^p$. We are interested in the nonnegative zero-norm constrained quadratic optimization problems

$$\min_{x \in \mathbb{R}^p} \left\{ \Psi(x) := x^T A x + \delta_{\mathbb{R}^p_+ \cap \Omega \cap S}(x) \right\}$$

and

$$\min_{x \in \mathbb{R}^p} \left\{ \Theta(x) := x^T B x + \delta_{\Delta \cap \Omega}(x) \right\},$$

where $A$ is a $p \times p$ real symmetric matrix, $B$ is a $p \times p$ positive semidefinite matrix, and $S := \{ x \in \mathbb{R}^p \mid \|x\| = 1 \}$ and $\Delta := \{ x \in \mathbb{R}^p_+ \mid \langle e, x \rangle = 1 \}$ are the unit sphere and the

*(shhpan@scut.edu.cn) School of Mathematics, South China University of Technology, Guangzhou.
†Corresponding author. School of Mathematics, South China University of Technology, Guangzhou.
‡(bishj@scut.edu.cn) School of Mathematics, South China University of Technology, Guangzhou.
simplex set in $\mathbb{R}^p$. In the sequel, $\mathcal{S}$ and $\Delta$ also denote the unit sphere and the simplex set in $\mathbb{R}^l$ with $l \neq p$. When the nonnegativity in (1) is removed, it reduces to
\[
\min_{x \in \mathbb{R}^p} \left\{ \Phi(x) := x^T A x + \delta_{\mathbb{R}^n \cap \mathcal{S}}(x) \right\}.
\]

The model (2) mainly arises from the sparse portfolio problem. Since Brodie et al. [9] added the $\ell_1$-norm to the classical Markowitz model to obtain sparse portfolios, various types of sparse regularizer are incorporated into the Markowitz model. Here, we follow the same line as in [21, Section 4.1] to use directly the zero-norm for the sparsity. The typical application of models (1) and (3) comes from sparse principal component analysis (PCA). The model (3) is first proposed by Moghaddam et al. [26] and receives an active research in the past ten years. A variety of methods [15, 47, 1, 16, 35, 22] have been developed for solving it; for example, the SPCA algorithm developed by Zou et al. [47] with iterative elastic net regression, the semidefinite program relaxation method proposed in [1], the generalized power methods developed by Journée et al. [16], the augmented Lagrangian approach by Lu and Zhang [22], and the truncated power method proposed by Yuan and Zhang [41]. In some applications from economics, biology and computer vision, nonnegativity is also required to reduce the risk of a portfolio [14], increase the robustness of biological systems [6] or extract the relevant parts from images [18]. Thus, it is necessary to enforce nonnegativity in conjunction with sparsity on the computed components. The model (1) was proposed in [3] with the target at such applications. It should be pointed out that the model (1) often appears as a module in some matrix factorization algorithms for nonnegative low rank optimization problems (see [43]).

For the case that $A$ is negative semidefinite, Asteris et al. [3] developed an algorithm for the problem (1) by extending the spannogram framework to the nonnegative sparse PCA and provided provable approximation guarantees, Sigg and Buhmann [33] proposed an expectation maximization algorithm to solve (1) with an $\ell_1$-norm ball constraint instead of the zero-norm constraint, and Yang and Xu [39] later provided a unified framework for outlier-robust PCA-like algorithms, which involves solving a sequence of problems (1) with updated $A$. Recently, Liu et al. [21] developed a successive difference-of-convex approximation method for a class of nonconvex and nonsmooth optimization problems including the problem (1)-(3). For the above mentioned methods, to the best of our knowledge, no linear rate of convergence is established. The motivation of this work is to provide a globally and linearly convergent PGD method for (1)-(3).

For nonconvex and nonsmooth optimization problems, it is notoriously difficult to achieve the convergence of the whole sequence generated by an algorithm to a stationary point. In fact, for the problems with complicated nonconvex and nonsmooth structure, the characterization of stationary points is even not easy. In the past several years, it has witnessed the successful application of the KL property of the extended-valued objective function in analyzing the global convergence of algorithms (see, e.g., [4, 5, 7, 8]). In particular, the KL property of exponent $1/2$ plays a crucial role in achieving the linear rate of convergence. As recently discussed in [28], for the structured semiconvex function and the primal lower nice function, the KL property of exponent $1/2$ is usually weaker
than the metric subregularity of their subdifferential operators or the Luo-Tseng error bound [34], which are the common regularity for deriving the linear convergence of first-order algorithms (see, e.g., [24, 23, 10, 37, 46]). Thus, an interesting direction is to identify the class of functions with the KL property of exponent 1/2. We notice that some positive progress has been made in this direction; for example, Li and Pong [19, 40] developed some calculation rules for the exponent of KL property, Liu et al. [20] verified the KL property of exponent 1/2 for the quadratic function restricted to Stiefel manifold, and Zhang et al. [45] achieved the KL property of exponent 1/2 over the global optimal solution set for several regularized matrix factorization functions.

The main contribution of this paper is to establish the KL property of exponent 1/2 for the nonconvex and nonsmooth functions \( \Phi, \Psi \) and \( \Theta \). The function \( \Phi \) differs from the zero-norm regularized quadratic function of [38] in the replacement of the zero-norm function by the indicator function of \( \Omega \), but its KL property of exponent 1/2 requires a completely different analysis technique. As an application of this key property, we also develop a globally and linearly convergent PGD method for the problem \((1)-(3)\). Numerical comparisons with the approximate Spannogram [3] for nonnegative sparse PCA on synthetic and real data and numerical results for sparse portfolio problem are included to confirm the theoretical findings and validate the efficiency of the PGD method.

2 Notations and preliminaries

Throughout this paper, for an extended real-valued function \( f: \mathbb{R}^p \to (-\infty, +\infty) \), we say that \( f \) is proper if \( \text{dom} \, f := \{ x \in \mathbb{R}^p \mid f(x) < \infty \} \) is nonempty; use the notation \( x' \to x \)

to signify \( x' \to x \) and \( f(x') \to f(x) \), and write \( [\alpha \leq f \leq \beta] := \{ x \in \mathbb{R}^p \mid \alpha \leq f(x) \leq \beta \} \)

for \( \alpha \in \mathbb{R} \) and \( \beta \in \mathbb{R} \). For a given set \( C \subseteq \mathbb{R}^p \), the notation \( \left.x' \to x \right|_C \) to signify \( C \ni x' \to x \).

For a given \( \tau \in \mathbb{R}^p \) and \( r > 0 \), \( B(\tau, r) \) denotes the closed ball centered at \( \tau \) with radius \( r \), \( [\tau] \) means the subspace spanned by \( \tau \) and \([\tau]^\perp \) denotes its orthogonal complement. For a vector \( z \in \mathbb{R}^p \), \( z^\perp \) means the vector obtained by arranging the entries of \( z \) in a decreasing order, \( z^{\kappa, \perp} \in \mathbb{R}^\kappa \) is the vector composed of the first \( \kappa \) entries of \( z^\perp \) and \( \text{supp}(z) \) means the index set of nonzero entries of \( z \). For a matrix \( X \in \mathbb{S}^p \) and an index set \( J \subseteq \{1, \ldots, p\} \), \( X_{J} \in \mathbb{S}^{|J|} \) means the submatrix consisting of the entries \( X_{ij} \) for \((i, j) \in J \times J \). Likewise, for a vector \( z \in \mathbb{R}^p \), \( z_j \in \mathbb{R}^{|J|} \) is the vector consisting of the entries \( z_j \) for \( j \in J \). The notation \( e \) denotes a vector of all ones whose dimension is known from the context.

2.1 Generalized subdifferentials

**Definition 2.1** (see [31, Definition 8.3]) Consider a function \( f: \mathbb{R}^p \to (-\infty, +\infty) \) and a point \( x \in \text{dom} \, f \). The regular subdifferential of \( f \) at \( x \), denoted by \( \partial f(x) \), is defined as

\[
\partial f(x) := \left\{ v \in \mathbb{R}^p \mid \liminf_{x' \to x \atop x' \neq x} \frac{f(x') - f(x) - \langle v, x' - x \rangle}{\|x' - x\|} \geq 0 \right\};
\]
Let \( (ii) \) and the former is always convex but the latter is generally nonconvex. When invoking Definition 2.1 \( \hat{\partial}f \) and is precisely the subdifferential of \( f \) at \( x \) in the sense of convex analysis.

Remark 2.1 (i) At each \( x \in \text{dom} f \), \( \hat{\partial}f(x) \) and \( \partial f(x) \) are both closed with \( \hat{\partial}f(x) \subseteq \partial f(x) \), and the former is always convex but the latter is generally nonconvex. When \( f \) is convex, \( \hat{\partial}f(x) = \partial f(x) \) and is precisely the subdifferential of \( f \) at \( x \) in the sequel, we denote by \( \text{crit} f \) the set of critical points of \( f \). By [31, Theorem 10.1], we know that a local minimizer of \( f \) is necessarily a critical point of \( f \).

2.2 Normal cones to the set \( \mathbb{R}^p_+ \cap \Omega \cap S \)

First, we recall from [31] the concept of the (regular) normal cone to a set \( C \subseteq \mathbb{R}^p \).

**Definition 2.2** Consider a point \( \overline{x} \in C \). The regular normal cone to \( C \) at \( \overline{x} \) is given by

\[
\hat{N}_C(\overline{x}) := \left\{ v \in \mathbb{R}^p \mid \limsup_{x \to \overline{x}, x \neq \overline{x}} \langle v, x - \overline{x} \rangle \leq 0 \right\};
\]

while the normal cone to the set \( C \) at \( \overline{x} \), denoted by \( N_C(\overline{x}) \), is defined as

\[
N_C(\overline{x}) := \limsup_{x \to \overline{x}} \hat{N}_C(x) = \left\{ v \in \mathbb{R}^p \mid \exists x^k \to \overline{x}, v^k \to v \text{ with } v^k \in \hat{N}_C(x^k) \right\}.
\]

**Remark 2.2** (i) Notice that \( \hat{N}_C(\overline{x}) \) and \( N_C(\overline{x}) \) are closed and \( \hat{N}_C(\overline{x}) \subseteq N_C(\overline{x}) \). Also, \( \hat{N}_C(\overline{x}) \) is the negative polar of the tangent cone to \( C \) at \( \overline{x} \), i.e., \( \hat{N}_C(\overline{x}) = [T_C(\overline{x})]^\circ \), and hence is always convex. When \( C \) is convex, \( \hat{N}_C(\overline{x}) = N_C(\overline{x}) = \{ z \in \mathbb{R}^p \mid \langle z, x - \overline{x} \rangle \leq 0 \} \).

(ii) From [31, Exercise 8.14], it holds that \( \hat{\partial}C(\overline{x}) = \hat{N}_C(\overline{x}) \) and \( \partial C(\overline{x}) = N_C(\overline{x}) \). Together with [38, Lemma 2.2], it follows that \( \hat{N}_S(\overline{x}) = N_S(\overline{x}) = \{ \omega \overline{x} \mid \omega \in \mathbb{R} \} \).

The following lemma presents a characterization for the (regular) normal cone to \( \Omega \).

**Lemma 2.1** Consider an arbitrary \( x \in \Omega \). Write \( J = \text{supp}(x) \) and \( \overline{J} = \{ 1, \ldots, p \} \setminus J \).

(i) If \( \|x\|_0 = \kappa \), then \( \hat{N}_\Omega(x) = \{ v \in \mathbb{R}^p \mid v_J = 0 \} = N_\Omega(x) \).

(ii) If \( \|x\|_0 < \kappa \), then \( \{ 0 \} = \hat{N}_\Omega(x) \subseteq N_\Omega(x) = \Gamma \), where the set \( \Gamma \) is defined by

\[
\Gamma = \left\{ v \in \mathbb{R}^p \mid \exists \hat{J} \subseteq \overline{J} \text{ with } |\hat{J}| = \kappa - |J| \text{ such that } v_{J \cup \hat{J}} = 0 \right\}. \tag{4}
\]
Proof: (i) Take an arbitrary $v \in \hat{\mathcal{N}}_{\Omega}(x)$. By Definition 2.2, for any $z \rightarrow x$ and $z \neq x$, we have $\langle v, z - x \rangle \leq 0$. We next argue by contradiction that $v_J = 0$. If not, there exists some $i \in \text{supp}(x)$ such that $v_i \neq 0$. For each $k$, let $z^k = (z^k_1, \ldots, z^k_p)^T$ with $z^k_j$ defined by

$$z^k_j := \begin{cases} x_j & \text{if } \text{supp}(x) \ni j \neq i; \\ x_j + \text{sign}(v_i)\frac{1}{k} & \text{if } j = i; \\ 0 & \text{if } j \notin \text{supp}(x) \end{cases} \quad \text{for } j = 1, \ldots, p. \quad (5)$$

Clearly, $z^k \rightarrow x$ and $z^k \neq x$, but $\langle v, z^k - x \rangle = |v_i|/k > 0$, which gives a contradiction to $\langle v, z - x \rangle \leq 0$ for any $z \rightarrow x$ and $z \neq x$. Consequently, $\hat{\mathcal{N}}_{\Omega}(x) \subseteq \{ v \in \mathbb{R}^p \mid v_J = 0 \}$. Conversely, take an arbitrary $\xi \in \{ v \in \mathbb{R}^p \mid v_J = 0 \}$. Notice that there exists $\delta > 0$ such that for all $z \in B(x, \delta)$, $\text{supp}(z) \supseteq \text{supp}(x)$. Hence, for all $z \in B(x, \delta) \cap \Omega$, we have $\text{supp}(z) = \text{supp}(x)$, and then $\langle \xi, z - x \rangle = 0$. Thus, $\limsup_{z \rightarrow x, z \neq x} \frac{\langle v, z - x \rangle}{\|z - x\|} = 0$.

This, by Definition 2.2, shows that $\xi \in \hat{\mathcal{N}}_{\Omega}(x)$. By the arbitrariness of $\xi$, it follows that $\{ v \in \mathbb{R}^p \mid v_J = 0 \} \subseteq \hat{\mathcal{N}}_{\Omega}(x)$. Thus, we establish the first equality. Since $\hat{\mathcal{N}}_{\Omega}(x) \subseteq \hat{\mathcal{N}}_{\Omega}(x)$, to establish the second equality, it suffices to argue that $\mathcal{N}_{\Omega}(x) \subseteq \hat{\mathcal{N}}_{\Omega}(x)$. To this end, take an arbitrary $v \in \mathcal{N}_{\Omega}(x)$. Then, there exist sequences $z^k \rightarrow x$ and $v^k \rightarrow v$ such that $v^k \in \hat{\mathcal{N}}_{\Omega}(z^k)$ for each $k$. From the above arguments, $\text{supp}(z^k) = \text{supp}(x)$ for all sufficiently large $k$. Together with $v^k \in \hat{\mathcal{N}}_{\Omega}(z^k)$ and the first equality, we have $v^k_0 = 0$ for all sufficiently large $k$, and then $v_J = 0$. Thus, the inclusion $\mathcal{N}_{\Omega}(x) \subseteq \hat{\mathcal{N}}_{\Omega}(x)$ follows.

(ii) We first establish the first equality. Notice that $\{ 0 \} \subseteq \hat{\mathcal{N}}_{\Omega}(x)$. Suppose that there exists $0 \neq v \in \hat{\mathcal{N}}_{\Omega}(x)$. Without loss of generality, let $v_i \neq 0$ for some $i \in \{ 1, \ldots, p \}$. We proceed the arguments by the following two cases: $i \in \text{supp}(x)$ and $i \notin \text{supp}(x)$.

**Case 1:** $i \in \text{supp}(x)$. For each $k$, let $z^k = (z^k_1, \ldots, z^k_p)^T$ with $z^k_j$ ($j = 1, \ldots, p$) defined by (5). Clearly, for all sufficiently large $k$, $\|z^k\|_0 = \|x\|_0$, which implies $z^k \rightarrow x$. However,

$$\limsup_{z \rightarrow x} \frac{\langle v, z - x \rangle}{\|z - x\|} \geq \lim_{k \rightarrow \infty} \frac{\langle v, z^k - x \rangle}{\|z^k - x\|} = \lim_{k \rightarrow \infty} \frac{|v_i|}{k \|z^k - x\|} = \frac{|v_i|}{\|x\|_0} > 0.$$  

By Definition 2.2, this yields a contradiction to the fact that $v \in \hat{\mathcal{N}}_{\Omega}(x)$.

**Case 2:** $i \notin \text{supp}(x)$. For each $k$, let $z^k = (z^k_1, \ldots, z^k_p)^T$ with $z^k_j$ ($j = 1, \ldots, p$) defined by

$$z^k_j := \begin{cases} x_j & \text{if } \text{supp}(x) \ni j; \\ \text{sign}(v_i)\frac{1}{k} & \text{if } j = i; \\ 0 & \text{if } j \notin \text{supp}(x) \end{cases} \quad \text{for } j = 1, \ldots, p.$$  

Since $\|x\|_0 < \kappa$, we have $z^k \rightarrow x$, but $\limsup_{z \rightarrow x} \frac{\langle v, z - x \rangle}{\|z - x\|} \geq \lim_{k \rightarrow \infty} \frac{|v_i|}{k \|z^k - x\|} = \frac{|v_i|}{\|x\|_0} > 0$, a contradiction to $v \in \hat{\mathcal{N}}_{\Omega}(x)$. This shows that the equality $\{ 0 \} = \hat{\mathcal{N}}_{\Omega}(x)$ holds.
Since the inclusion is trivial, the rest only focuses on the second equality. For this purpose, let $\xi$ be an arbitrary point from $N_\Omega(x)$. By Definition 2.2, there exist sequences $z^k \to x$ and $\xi^k \to \xi$ with $\xi^k \in \bar{N}_\Omega(z^k)$ for each $k$. Since $z^k \to x$, there exists $\bar{k} \in \mathbb{N}$ such that for all $k \geq \bar{k}$, $\text{supp}(z^k)$ keep unchanged, say, $\text{supp}(z^k) = I$, and $I \supset \text{supp}(x) = J$. When $|I| = \kappa$, from part (i) and $\xi^k \in \bar{N}_\Omega(z^k)$, we have $\xi^k \perp 0$ for $k \geq \bar{k}$, and then $\xi_I = 0$; and moreover, since now $I$ includes $J$ strictly, $\xi_J = 0$ and there exists $\hat{J} \in \hat{J}$ with $|\hat{J}| = \kappa - |J|$ such that $\xi_{\hat{J}} = 0$. When $|I| < \kappa$, from $\bar{N}_\Omega(x) = \{0\}$, we have $\xi = 0$, which implies that $\xi_J = 0$ and there exists $\hat{J} \in \hat{J}$ with $|\hat{J}| = \kappa - |J|$ such that $\xi_{\hat{J}} = 0$. This show that $\xi \in \Gamma$, and consequently $N_\Omega(x) \subseteq \Gamma$. Next we argue that $\Gamma \subseteq N_\Omega(x)$. Take an arbitrary $\xi \in \Gamma$. Then, $\xi_J = 0$ and there is $\hat{J} \in \hat{J}$ with $|\hat{J}| = \kappa - |J|$ such that $\xi_{\hat{J}} = 0$. For each $k$, let $z^k = (z^k_1, \ldots, z^k_p)^T$ and $\xi^k = (\xi^k_1, \ldots, \xi^k_p)^T$ with $z^k$ and $v^k$ defined by

$$z^k := \begin{cases} x_i & \text{if } i \in J; \\ \frac{1}{k} & \text{if } i \in \hat{J}; \\ 0 & \text{if } i \in \mathbb{R} \setminus \hat{J} \setminus \hat{J}. \end{cases} \quad \text{and} \quad \xi^k := \begin{cases} 0 & \text{if } i \in J \cup \hat{J}; \\ \xi_i + \frac{k}{1} & \text{if } i \in \mathbb{R} \setminus \hat{J}. \end{cases}$$

Clearly, $\Omega \ni z^k \to x$ with $\|z^k\|_0 = \kappa$ and $\xi^k \to \xi$. Also, $\xi^k \in \bar{N}_\Omega(z^k)$ holds for each $k$ by part (i). Thus, we have $\xi \in \bar{N}_\Omega(x)$, and then $\Gamma \subseteq N_\Omega(x)$. The proof is completed. \(\square\)

**Remark 2.3 (a)** Since the zero-norm is regular by [38, Lemma 2.1], the result of Lemma 2.1(i) can also be obtained by invoking [31, Proposition 10.3] and [38, Lemma 2.1].

**Remark 2.3 (b)** Lemma 2.1 implies that $N_\Omega(x) \subseteq [x]^\perp$, and $N_\Omega(x) = [x]^\perp$ if $\|x\|_0 = \kappa$. Moreover, it shows that $\Omega$ is Clarke regular only at those points with zero-norm equal to $\kappa$.

Next, by invoking Lemma 2.1, we present the characterization of the normal cones to the composite sets $\mathbb{R}^p_+ \cap \Omega$, $S \cap \Omega$ and $\mathbb{R}^p_+ \cap \Omega \cap S$ one by one.

**Proposition 2.2** Consider an arbitrary point $x \in \mathbb{R}^p_+ \cap \Omega$. Then, it holds that

$$N_{\mathbb{R}^p_+ \cap \Omega}(x) \subseteq N_{\mathbb{R}^p_+}(x) + N_\Omega(x) \subseteq [x]^\perp. \quad (6)$$

If, in addition, $\|x\|_0 = \kappa$, then $N_{\mathbb{R}^p_+ \cap \Omega}(x) = N_{\mathbb{R}^p_+}(x) + N_\Omega(x) = \bar{N}_{\mathbb{R}^p_+ \cap \Omega}(x) = [x]^\perp$.

**Proof:** Define the multifunction $F : \mathbb{R}^p \Rightarrow \mathbb{R}^p \times \mathbb{R}^p$ by $F(z) := (\Omega - z) \times (\mathbb{R}^p_+ - z)$ for $z \in \mathbb{R}^p$. Notice that the set $\Omega$ is a union of finitely many polyhedral convex sets. It is not hard to check that the multifunction $F$ is polyhedral, i.e., its graph is the union of finitely many polyhedral convex sets. From [29, Proposition 1], it follows that $F$ is metrically subregular at any point $(z, y) \in \text{gph}F$. Along with [13, Section 3.1], the first inclusion in (6) follows. Notice that $N_{\mathbb{R}^p_+}(x) = \mathbb{R}^p_+ \cap [x]^\perp$. Together with Lemma 2.1, the second inclusion in (6) holds. Now assume that $\|x\|_0 = \kappa$. Then, it holds that

$$N_{\mathbb{R}^p_+}(x) + N_\Omega(x) \supseteq N_{\mathbb{R}^p_+ \cap \Omega}(x) \supseteq \bar{N}_{\mathbb{R}^p_+ \cap \Omega}(x) \supseteq \bar{N}_{\mathbb{R}^p_+}(x) + N_\Omega(x) = N_{\mathbb{R}^p_+}(x) + N_\Omega(x)$$

where the first inclusion is due to (6), the third inclusion is due to [31, Corollary 10.9] and Remark 2.3, and the equality is using the regularity of $\Omega$ and the convexity of $\mathbb{R}^p_+$. By Lemma 2.1(i), we have $[x]^\perp \subseteq N_{\mathbb{R}^p_+}(x) + N_\Omega(x)$. The proof is completed. \(\square\)
Lemma 2.2 Consider an arbitrary point \( x \in \mathcal{S} \cap \Omega \). Then, it holds that
\[
\mathcal{N}_{\mathcal{S} \cap \Omega}(x) \subseteq \mathcal{N}_{\mathcal{S}}(x) + \mathcal{N}_{\Omega}(x). 
\]
If, in addition, \( \|x\| = \kappa \), then \( \mathcal{N}_{\mathcal{S} \cap \Omega}(x) = \mathcal{N}_{\mathcal{S}}(x) + \mathcal{N}_{\Omega}(x) = \hat{\mathcal{N}}_{\mathcal{S} \cap \Omega}(x) \).

Proof: Let \( u \in \mathcal{N}_{\mathcal{S}}(x) \) and \( v \in \mathcal{N}_{\Omega}(x) \) be such that \( u + v = 0 \). By Remark 2.3, there exists \( \omega \in \mathbb{R} \) such that \( u = \omega x \), and hence \( \omega x + v = 0 \). By Remark 2.3(b), we have \( \langle v, x \rangle = 0 \). Together with \( \omega x + v = 0 \), it is easy to obtain \( \omega = 0 \). This implies that \( u = v = 0 \). The first part of the inclusions then follows by [31, Theorem 6.42]. The second part holds by using the Clarke regularity of the sets \( \Omega \) and \( \mathcal{S} \) and following the same arguments as those for the second part of Proposition 2.1.

Remark 2.5 By the proof of Proposition 2.2, if the normal cone to a set \( C \) satisfies \( \mathcal{N}_{C}(z) \subseteq \{z\}^\perp \) for each \( z \in C \), then \( \mathcal{N}_{C \cap \mathcal{S}}(x) \subseteq \mathcal{N}_{C}(x) + \mathcal{N}_{\mathcal{S}}(x) \) for all \( x \in \mathbb{R}^p \).

For each \( x \in \mathbb{R}^p_+ \cap \Omega \cap \mathcal{S} \), by (6) \( \mathcal{N}_{\mathbb{R}^p_+ \cap \Omega}(x) \subseteq \{x\}^\perp \). Using Remark 2.5, Proposition 2.1 and the Clarke regularity of \( \mathbb{R}^p_+, \Omega \) and \( \mathcal{S} \) yields the following conclusion.

Proposition 2.3 Consider an arbitrary point \( x \in \mathbb{R}^p_+ \cap \Omega \cap \mathcal{S} \). Then, it holds that
\[
\mathcal{N}_{\mathbb{R}^p_+ \cap \Omega \cap \mathcal{S}}(x) \subseteq \mathcal{N}_{\mathbb{R}^p_+ \cap \Omega}(x) + \mathcal{N}_{\mathcal{S}}(x) \subseteq \mathcal{N}_{\mathbb{R}^p_+}(x) + \mathcal{N}_{\Omega}(x) + \mathcal{N}_{\mathcal{S}}(x). \quad (7)
\]
If, in addition, \( \|x\| = \kappa \), then \( \mathcal{N}_{\mathbb{R}^p_+ \cap \Omega \cap \mathcal{S}}(x) = \mathcal{N}_{\mathbb{R}^p_+}(x) + \mathcal{N}_{\Omega}(x) + \mathcal{N}_{\mathcal{S}}(x) \).

2.3 Normal cones to the set \( \Omega \cap \Delta \)

We first provide the characterization of the tangent cone and normal cone to the set \( \Delta \).

Lemma 2.2 Consider an arbitrary \( x \in \Delta \). Let \( J = \text{supp}(x) \) and \( \overline{J} = \{1, \ldots, p\} \setminus J \). Then
\[
\mathcal{T}_{\Delta}(x) = \left\{ h \in \mathbb{R}^p \mid e^\top h = 0, h_{\overline{J}} \in \mathbb{R}_{\overline{J}}^+ \right\},
\]
\[
\mathcal{N}_{\Delta}(x) = \left\{ \xi \in \mathbb{R}^p \mid \exists \omega \in \mathbb{R} \text{ s.t. } \xi_J = \omega e_J, \xi_{\overline{J}} \leq \omega e_{\overline{J}} \right\}. 
\]

Proof: By the definition of the tangent cone (see [31, Definition 6.1]), it is easy to check that \( \mathcal{T}_{\Delta}(x) \subseteq \{ h \in \mathbb{R}^p \mid e^\top h = 0, h_{\overline{J}} \in \mathbb{R}_{\overline{J}}^+ \} \). Now take an arbitrary \( h \) from the set on the right hand side. For each \( k \), define \( t_k := k^{-1} \) and \( h^k := h \). Clearly, for all sufficiently large \( k \), \( x + t_k h^k \in \mathbb{R}^p_+ \) and \( \langle e, x + t_k h^k \rangle = 1 \), which implies that \( h \in \mathcal{T}_{\Delta}(x) \). By the arbitrariness of \( h \), \( \mathcal{T}_{\Delta}(x) \supseteq \{ h \in \mathbb{R}^p \mid e^\top h = 0, h_{\overline{J}} \in \mathbb{R}_{\overline{J}}^+ \} \). The first equality holds.
Fix an arbitrary $\xi \in [T_\Delta(\overline{x})]^\circ$. Then, $h^* = 0$ is optimal to the following linear program
\[
\max_{h \in \mathbb{R}^p} \{ (\xi, h) \mid e^T h = 0, \ h \in \mathbb{R}_+^p \}
\]
which, by the duality theory of the linear program, is equivalent to saying that the dual
\[
\min_{\omega \in \mathbb{R}} \{ 0 \mid \xi J = \omega e_J, \xi J \leq \omega e_J \}
\]
has a nonempty feasible set. Since $[T_\Delta(\overline{x})]^\circ = N_\Delta(\overline{x})$, the second equality follows.

\textbf{Proposition 2.4} Consider an arbitrary point $x \in \Delta \cap \Omega$. Then, it holds that
\[
N_{\Omega \cap \Delta}(x) \subseteq N_\Omega(x) + N_\Delta(x),
\]
and furthermore, the inclusion becomes an equality if in addition $\|x\|_0 = \kappa$.

\textbf{Proof:} Notice that $\Omega$ is a union of finite many polyhedral convex set and $\Delta$ is a polyhedral convex set. The inclusion follows by Remark 2.4. Since $\Omega$ is Clarke regular at $x$ when $\|x\|_0 = \kappa$ and $\Delta$ is convex, the second part follows by [31, Corollary 10.9].

\section{Kurdyka-Lojasiewicz property}

\textbf{Definition 2.3} Let $f : \mathbb{R}^p \to (-\infty, +\infty]$ be a proper function. The function $f$ is said to have the Kurdyka-Lojasiewicz (KL) property at $x \in \text{dom} \partial f$ if there exist $\eta \in (0, +\infty]$, a continuous concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ satisfying the following two conditions
\begin{enumerate}
    \item [(i)] $\varphi(0) = 0$ and $\varphi$ is continuously differentiable on $(0, \eta)$;
    \item [(ii)] for all $s \in (0, \eta)$, $\varphi'(s) > 0$,
\end{enumerate}
and a neighborhood $U$ of $\overline{x}$ such that for all $x \in U \cap [f(\overline{x}) < f < f(\overline{x}) + \eta]$,
\[
\varphi'(f(x) - f(\overline{x}))\text{dist}(0, \partial f(x)) \geq 1.
\]
If the corresponding $\varphi$ can be chosen as $\varphi(s) = c s^{1/2}$ for some $c > 0$, then $f$ is said to have the KL property of exponent $1/2$ at $\overline{x}$. If $f$ has the KL property of exponent $1/2$ at each point of $\text{dom} \partial f$, then $f$ is called a KL function of exponent $1/2$.

\textbf{Remark 2.6} By [4, Lemma 2.1], a proper function has the KL property of exponent $1/2$ at any noncritical point. Hence, to show that it is a KL function of exponent $1/2$, it suffices to check whether it has the KL property of exponent $1/2$ at each critical point.

From [38, Proposition 3.1], we know that the quadratic function over the unit sphere has the KL property of exponent $1/2$, i.e., the following conclusion holds.

\textbf{Lemma 2.3} Fix an arbitrary positive integer $m$. For any given $H \in \mathbb{S}^m$, define
\[
g(z) := z^T H z + \delta_\mathbb{S}(z) \quad \text{for} \ z \in \mathbb{R}^m.
\]
Then $g$ is a KL function of exponent $1/2$, and at each $\overline{z} \in \mathbb{R}^m$, $\partial g(\overline{z}) = 2 H \overline{z} + [\overline{z}]$. 

8
3 KL property of exponent 1/2 of Φ and Ψ

In this section, we shall show that both Φ and Ψ are the KL functions of exponent 1/2. Firstly, from [31, Exercise 10.10] and Proposition 2.2, for any \( x \in S \cap \Omega \),
\[
\partial \Phi(x) = 2Ax + N_{S \cap \Omega}(x) \subseteq 2Ax + N_S(x) + N_\Omega(x),
\]
and if in addition \( \|x\|_0 = \kappa \), then the inclusion becomes an equality.

**Proposition 3.1** The \( \Phi \) defined in equation (3) is a KL function of exponent 1/2.

**Proof:** By Remark 2.6, it suffices to argue that \( \Phi \) has the KL property of exponent 1/2 at each \( x \in \text{crit} \Phi \). Fix an arbitrary \( \varpi \in \text{crit} \Phi \). We proceed the arguments by two cases.

**Case 1:** \( \|\varpi\|_0 = \kappa \). Write \( J = \text{supp}(\varpi) \). Let \( g \) be the function defined as in Lemma 2.3 with \( H = A_I \). By Lemma 2.3, \( g \) is a KL function of exponent 1/2. So, there exist \( \delta > 0, \eta > 0 \) and \( c > 0 \) such that for all \( z \in B(\varpi, \delta) \cap [g(\varpi) < \Phi(\varpi) + \eta] \)
\[
\text{dist}(0, \partial g(z)) \geq c\sqrt{g(z) - g(\varpi)}.
\]
Take an arbitrary \( x \) from the set \( B(\varpi, \delta) \cap [\Phi(\varpi) < \Phi(\varpi) + \eta] \). Clearly, \( x \in \Omega \cap S \). In addition, by reducing \( \delta \) if necessary, we have \( \text{supp}(x) \supseteq \text{supp}(\varpi) \). Along with \( \|x\|_0 \leq \kappa \), \( \text{supp}(x) = \text{supp}(\varpi) = J \) and \( \|x\|_0 = \kappa \). The inclusion in (8) becomes an equality. Then
\[
\text{dist}(0, \partial \Phi(x)) = \min_{\zeta \in N_S(x), \xi \in N_\Omega(x)} 2Ax + \zeta + \xi
\]
\[
= \min_{\omega \in \mathbb{R}, \xi \in N_\Omega(x)} 2Ax + \omega x + \xi
\]
\[
\geq \min_{\omega \in \mathbb{R}} \|2A_Ix + \omega x\| = \text{dist}(0, \partial g(x))
\]
where the second equality is by Remark 2.3(ii), the inequality is due to Lemma 2.1(i) and \( \text{supp}(x) = J \), and the last equality is by the definition of \( g \). On the other hand, from \( \text{supp}(x) = \text{supp}(\varpi) = J \) and the expressions of \( \Phi \) and \( g \), it follows that
\[
\Phi(x) - \Phi(\varpi) = x^TAx - \varpi^TA\varpi = x^TA_Ix - \varpi^TA_I\varpi = g(x) - g(\varpi)
\]
which, along with \( x \in [\Phi(\varpi) < \Phi(\varpi) + \eta] \), implies that \( x \in [g(\varpi) < \Phi(\varpi) + \eta] \).
Notice that \( \|x - x\| = \|x - \varpi\| \leq \delta \). From inequalities (9) and (10), we have
\[
\text{dist}(0, \partial \Phi(x)) \geq \text{dist}(0, \partial g(x)) \geq c\sqrt{g(x) - g(\varpi)} = c\sqrt{\Phi(x) - \Phi(\varpi)}
\]
By the arbitrariness of \( x \), the function \( \Phi \) has the KL property of exponent 1/2 at \( \varpi \).

**Case 2:** \( \|\varpi\|_0 < \kappa \). By the continuity, there exists \( \delta_1 > 0 \) such that for all \( x' \in B(\varpi, \delta_1) \), \( \text{supp}(x') \supseteq \text{supp}(\varpi) \). Let \( I := \{I \mid \{1, \ldots, p\} \supseteq I \supseteq \text{supp}(\varpi)\} \). For each \( I \in I \), define
\[
g_I(z) := z^TA_{II}z + \delta_S(z) \quad \forall z \in \mathbb{R}^{|I|}.
\]
By Lemma 2.3, \( g_I \) is a KL function of exponent 1/2. Therefore, there exist \( \delta_I > 0, \eta_I > 0 \) and \( c_I > 0 \) such that for all \( z \in \mathbb{B}(\mathbf{x}_I, \delta_I) \cap [g_I(\mathbf{x}_I) < g_I(\mathbf{x}_I) + \eta_I] \),
\[
\text{dist}(0, \partial g_I(z)) \geq c_I \sqrt{g_I(z) - g_I(\mathbf{x}_I)}. \tag{12}
\]
Notice that \( \mathcal{I} \) includes finite index sets. We set \( \delta = \min(\delta_1, \min_{I \in \mathcal{I}} \delta_I), \eta := \min_{I \in \mathcal{I}} \eta_I \) and \( c := \min_{I \in \mathcal{I}} c_I \). Take an arbitrary \( x \in \mathbb{B}(\mathbf{x}, \delta) \cap [\Phi(\mathbf{x}) < \Phi < \Phi(\mathbf{x}) + \eta] \). Clearly, \( x \in \mathcal{S} \cap \Omega \) and \( J := \text{supp}(x) \supseteq \text{supp}(\mathbf{x}) \). From the inclusion (8), it follows that
\[
\text{dist}(0, \partial \Phi(x)) \geq \min_{\zeta \in \mathcal{N}_{\mathbf{x}}(x), \xi \in \mathcal{N}_{\Omega}(x)} \|2A x + \zeta + \xi\|
= \min_{\omega \in \mathbb{R}, \zeta \in \mathcal{N}_{\Omega}(x)} \|2A x + \omega x + \xi\|
\geq \min_{\omega \in \mathbb{R}} \|2A_J x_J + \omega x_J\| = \text{dist}(0, \partial g_J(x_J)) \tag{13}
\]
where the first equality is due to Remark 2.3(ii), the second inequality is due to Lemma 2.1 and \( \text{supp}(x) = J \), and the last equality is by the definition of \( g_J \). In addition, from \( J = \text{supp}(x) \supseteq \text{supp}(\mathbf{x}) \) and the expressions of \( \Phi \) and \( g_J \), it follows that
\[
\Phi(x) - \Phi(\mathbf{x}) = x^T Ax - \mathbf{x}^T A \mathbf{x} = x^T A_J x_J - \mathbf{x}^T A_{\mathbf{x}} = g_J(x_J) - g_J(\mathbf{x}_J)
\]
which, by \( x \in [\Phi(\mathbf{x}) < \Phi < \Phi(\mathbf{x}) + \eta] \), implies that \( x_J \in [g(\mathbf{x}_J) < g_J < g_J(\mathbf{x}_J) + \eta] \). Notice that \( \|x_J - \mathbf{x}_J\| = \|x - \mathbf{x}\| \leq \delta \). Thus, from inequalities (13) and (12),
\[
\text{dist}(0, \partial \Phi(x)) \geq \text{dist}(0, \partial g_J(x_J)) \geq c \sqrt{g_J(x_J) - g_J(\mathbf{x}_J)} = c \sqrt{\Phi(x) - \Phi(\mathbf{x})}.
\]
By the arbitrariness of \( x \), the function \( \Phi \) has the KL property of exponent 1/2 at \( \mathbf{x} \).

The above arguments, together with the arbitrariness of \( \mathbf{x} \) in \( \text{crit} \Phi \), show that the function \( \Phi \) has the KL property of exponent 1/2. The proof is then completed.

Notice that the KL property of exponent 1/2 is lack of stability in the following sense: if \( f : \mathbb{R}^n \to (-\infty, +\infty] \) is a KL function of exponent 1/2, then its linear perturbation \( \tilde{f}(\cdot) = f(\cdot) + \langle u, \cdot \rangle \) for some \( u \in \mathbb{R}^p \) may not be a KL function of exponent 1/2. Hence, even armed with Proposition 3.1, one generally can not expect the KL property of exponent 1/2 of the function \( \Psi(\cdot) = \Phi(\cdot) + \delta \_p^\mathbb{R} \) over the set \( \text{crit} \Psi \). However, since \( \mathcal{N}_{\mathbb{R}_+ \cap \Omega}(x) \) shares with a key property of \( \mathcal{N}_{\Omega}(x) \), i.e., \( \mathcal{N}_{\mathbb{R}_+ \cap \Omega}(x) \subseteq [x]^\perp \), we can achieve this goal.

**Proposition 3.2** The \( \Psi \) defined in equation (1) is a KL function of exponent 1/2.

The proof of Proposition 3.2 is included in Appendix. Although the arguments are similar to those for Proposition 3.1, we include it for completeness.

## 4 KL property of exponent 1/2 of \( \Theta \)

From Proposition 2.4 and [31, Exercise 10.10], for any \( x \in \Delta \cap \Omega \), it holds that
\[
\partial \Theta(x) = 2Bx + \mathcal{N}_{\Delta \cap \Omega}(x) \subset 2Bx + \mathcal{N}_{\Delta}(x) + \mathcal{N}_\Omega(x), \tag{14}
\]
and the inclusion will becomes an equality if in addition \( \|x\|_0 = \kappa \). In order to establish the main result of this section, we also need the following crucial lemma.
Lemma 4.1 For a given positive semidefinite matrix $H \in \mathbb{S}_+^m$, let $h(z) := z^T H z + \delta_\Delta(z)$ for $z \in \mathbb{R}^m$. Then, $h$ is a KL function of exponent $1/2$.

Proof: Since $h$ is a convex function and its subdifferential operator $\partial h$ is a polyhedral multifunction from $\mathbb{R}^m$ to $\mathbb{R}^m$, from [29, Proposition 1] we know that $\partial f$ is metrically subregular at each point $(x, y) \in \text{gph}\partial f$. Together with [28, Theorem 3.1], it follows that $h$ is a KL function of exponent $1/2$. The proof is completed. \hfill \qed

Proposition 4.1 The $\Theta$ defined in equation (2) is a KL function of exponent $1/2$.

Proof: Fix an arbitrary $\pi \in \text{crit}\Theta$. We proceed the arguments by two cases as below.

Case 1: $\|\pi\|_0 = \kappa$. Write $J = \text{supp}(\pi)$. Let $h$ be defined as in Lemma 4.1 with $H = BJJ$. By Lemma 4.1, $h$ is a KL function of exponent $1/2$. Hence, there exist $\delta > 0$, $\eta > 0$ and $c > 0$ such that for all $z \in \mathbb{B}(\pi, \delta) \cap [h(\pi_j) < h(\pi_j) + \eta]$,

$$\text{dist}(0, \partial h(z)) \geq c\sqrt{h(z) - h(\pi_j)}. \quad (15)$$

Take an arbitrary $x \in \mathbb{B}(\pi, \delta) \cap [\Theta(\pi) < \Theta < \Theta(\pi) + \eta]$. Clearly, $x \in \Omega \cap \Delta$. In addition, by reducing $\delta$ if necessary, we have $\text{supp}(x) \supseteq \text{supp}(\pi)$. Together with $\|x\|_0 \leq \kappa$, $\text{supp}(x) = \text{supp}(\pi) = J$ and $\|x\|_0 = \kappa$. The inclusion in (14) becomes an equality. Then

$$\text{dist}(0, \partial \Theta(x)) = \min_{\zeta \in \mathbb{N}_\Delta(x), \xi \in \mathbb{N}_0(x)} \|2Bx + \zeta + \xi\|
\geq \min_{\zeta \in \mathbb{N}_\Delta(x), \xi \in \mathbb{N}_0(x)} \|(2Bx + \zeta + \xi)_J\|
\geq \min_{\omega \in \mathbb{R}} \|2BJJx_J + \omega e_J\|$$

where the last inequality is due to $\xi_J = 0$ and $\zeta_J = \omega' e_J$ for some $\omega' \in \mathbb{R}$, implied by $\text{supp}(x) = J$. From Lemma 2.2 and $\text{supp}(x) = J$, it follows that $\mathbb{N}_\Delta(x_J) = \{|\tau e_J| : \tau \in \mathbb{R}\}$, which along with $\partial h(x_J) = 2BJJx_J + \mathbb{N}_\Delta(x_J)$ implies that

$$\text{dist}(0, \partial h(x_J)) = \min_{\tau \in \mathbb{R}} \|2BJJx_J + \tau e_J\|.$$ 

In addition, from $\text{supp}(x) = \text{supp}(\pi) = J$ and the expressions of $\Theta$ and $h$, it follows that

$$\Theta(x) - \Theta(\pi) = x^T Bx - \pi^T B\pi = x_J^T BJJx_J - \pi_J^T BJJ\pi_J = h(x_J) - h(\pi_J)$$

which, along with $x \in [\Theta(\pi) < \Theta < \Theta(\pi) + \eta]$, implies that $x_J \in [h(\pi_J) < h(\pi_J) + \eta]$. Notice that $\|x_J - \pi_J\| = \|x - \pi\| \leq \delta$. Combining the last three equations with (15) yields

$$\text{dist}(0, \partial \Theta(x)) \geq \text{dist}(0, \partial h(x_J)) \geq c\sqrt{h(x_J) - h(\pi_J)} = c\sqrt{\Theta(x) - \Theta(\pi)}.$$ 

By the arbitrariness of $x$, the function $\Theta$ has the KL property of exponent $1/2$ at $\pi$.

Case 2: $\|\pi\|_0 < \kappa$. By the continuity, there exists $\delta_1 > 0$ such that for all $x' \in \mathbb{B}(\pi, \delta_1)$, $\text{supp}(x') \supseteq \text{supp}(\pi)$. Let $I := \{I \mid \{1, \ldots, p\} \supseteq I \supseteq \text{supp}(\pi)\}$. For each $I \in I$, define

$$h_I(z) := z^T B_{II} z + \delta_\Delta(z) \quad \forall z \in \mathbb{R}^{|I|}.$$
By Lemma 4.1, \( h_I \) is a KL function of exponent 1/2. So, there exist \( \delta_I > 0, \eta_I > 0 \) and \( c_I > 0 \) such that for all \( z \in B(\overline{\tau}, \delta_I) \cap [h_I(\overline{\tau}) < h_I(z) < h_I(\overline{\tau}) + \eta_I] \),

\[
\text{dist}(0, \partial h_I(z)) \geq c_I \sqrt{h_I(z) - h_I(\overline{\tau})}.
\] (17)

Notice that \( I \) includes finite index sets. We set \( \delta = \min(\delta_I, \min_{I \in \mathcal{I}} \delta_I), \eta := \min_{I \in \mathcal{I}} \eta_I \) and \( c := \min_{I \in \mathcal{I}} c_I \). Take an arbitrary \( x \in B(\overline{\tau}, \delta) \cap [\Theta(\overline{\tau}) < \Theta < \Theta(\overline{\tau}) + \eta]. \) Clearly, \( x \in \Delta \cap \Omega \) and \( J := \text{supp}(x) \supseteq \text{supp}(\overline{\tau}). \) From the inclusion (14), it follows that

\[
\text{dist}(0, \partial \Theta(x)) \geq \min_{\zeta \in \mathcal{N}_\Delta(x), \xi \in \mathcal{N}_\Omega(x)} \|2Bx + \zeta + \xi\| \\
\geq \min_{\zeta \in \mathcal{N}_\Delta(x), \xi \in \mathcal{N}_\Omega(x)} \|(2Bx + \zeta + \xi)_J\| \\
\geq \min_{\omega \in \mathbb{R}} \|2B_J x_J + \omega e_J\| \\
\geq \min_{\tau \in \mathbb{R}} \|2B_J x_J + \tau e_J\|.
\] (18)

where the last inequality is due to \( \xi_j = 0 \) and \( \zeta_j = \omega' e_J \) for some \( \omega' \in \mathbb{R} \) implied by \( \text{supp}(x) = J \). From Lemma 2.2 and \( J = \text{supp}(x) \), it follows that \( \mathcal{N}_\Delta(x_J) = \{\tau e_J \mid \tau \in \mathbb{R}\} \), which together with \( \partial h_J(x_J) = 2B_J x_J + \mathcal{N}_\Delta(x_J) \) implies that

\[
\text{dist}(0, \partial h_J(x_J)) = \min_{\tau \in \mathbb{R}} \|2B_J x_J + \tau e_J\|.
\]

In addition, from \( J \supseteq \text{supp}(\overline{\tau}) \) and the expressions of \( \Theta \) and \( h_J \), it follows that

\[
\Theta(x) - \Theta(\overline{\tau}) = x^T B x - \overline{\tau}^T B \overline{\tau} = x^T_J B_J x_J - \overline{\tau}^T_J B_J \overline{\tau}_J = h_J(x_J) - h_J(\overline{\tau}_J)
\]

which, by \( x \in [\Theta(\overline{\tau}) < \Theta < \Theta(\overline{\tau}) + \eta] \), implies that \( x_J \in [h_J(\overline{\tau}_J) < h_J(x_J) < h_J(\overline{\tau}_J) + \eta] \). Notice that \( \|x_J - \overline{\tau}_J\| = \|x - \overline{\tau}\| \leq \delta \). Thus, from inequalities (18) and (17),

\[
\text{dist}(0, \partial \Theta(x)) \geq \text{dist}(0, \partial h_J(x_J)) \geq c \sqrt{h_J(x_J) - h_J(\overline{\tau}_J)} = c \sqrt{\Theta(x) - \Theta(\overline{\tau})}.
\]

By the arbitrariness of \( x \), the function \( \Theta \) has the KL property of exponent 1/2 at \( \overline{\tau} \).

The above arguments, together with the arbitrariness of \( \overline{\tau} \) in \( \text{crit} \Theta \), show that the function \( \Theta \) has the KL property of exponent 1/2. The proof is then completed.  \( \square \)

5 **Globally and linearly convergent PGD**

The projection operator onto the set \( \mathbb{R}^p_+ \cap \Omega \cap S \) is a set-valued mapping defined by

\[
P(u) := \arg \min_{x \in \mathbb{R}^p} \left\{ \frac{1}{2} \|x - u\|^2 \right\} \text{ s.t. } x \geq 0, \|x\|_0 \leq \kappa, \|x\| = 1.
\] (19)

The following lemma gives a characterization for \( P \). Since the proof is easy, we omit it.
Lemma 5.1 Let $\mathcal{P}$ be the multifunction from $\mathbb{R}^p$ to $\mathbb{R}^p$ defined by (19). Define

$$Q(u) := \arg \min_{x \in \mathbb{R}^p} \left\{ \frac{1}{2} \|x - u\|^2 \mid x \geq 0, \|x\|_0 \leq \kappa, \|x\| = 1 \right\} \text{ for } u \in \mathbb{R}^p. \quad (20)$$

Fix an arbitrary $z \in \mathbb{R}^p$. Let $P$ be a $p \times p$ permutation matrix with $z = Pz^1$. Then, it holds that $\mathcal{P}(z) = PQ(z)$. Write $I := \text{supp}(z^1)$ and $\overline{I} := \{1, 2, \ldots, p\} \setminus I$. If $z^1_i < 0$, then $Q(z) = \{e_i \mid i \in J(z)\}$ with $J(z) = \{i \mid z^1_i = z^1_i\}$; if $z^1_i = 0$, then it holds that $Q(z) = \{u \in \mathbb{R}^p \mid \|u\|_I = 1, u_I \geq 0, \|u\|_0 \leq \kappa, u_I = 0\}$; if $z^1_i > 0$, then $Q(z) = \{(z^1_i, 0) \mid \|z^1_i\| \in \mathbb{R}^p\}$.

The following lemma shows that the projection onto the set $\Delta \cap \Omega$ is also available.

Lemma 5.2 Fix a vector $z \in \mathbb{R}^p$. Let $P$ be a permutation matrix such that $z^1 = Pz$. If

$$y^* = \arg \min_{y \in \mathbb{R}^p} \left\{ \frac{1}{2} \|y - z^{k-1}\|^2 \mid \langle e, y \rangle = 1, y \in \mathbb{R}_+^p \right\}, \quad (21)$$

then $P^{-1}(y^*; 0)$ is a global optimal solution to the following problem

$$\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|\theta - z\|^2 \mid \|\theta\|_0 \leq \kappa, \langle e, \theta \rangle = 1, \theta \in \mathbb{R}_+^p \right\}. \quad (22)$$

Proof: Notice that the minimization problem in (22) is equivalent to the problem

$$\min_{u \in \mathbb{R}^p} \left\{ \frac{1}{2} \|u - z^k\|^2 \mid \|u\|_0 \leq \kappa, \langle e, u \rangle = 1, u \in \mathbb{R}_+^p \right\}, \quad (23)$$

in the sense that if $u^*$ is a global optimal solution of (23), then $P^{-1}u^*$ is globally optimal to (22), and conversely, if $\theta^*$ is a global optimal solution of (22), then $P\theta^*$ is globally optimal to (23). The desired result then follows from Lemma 1 in Appendix.

Motivated by Lemma 5.1-5.2, we apply the following PGD method for solving the problem (1)-(3), where $\Xi$ is one of the sets $\mathbb{R}_+^p \cap \Omega \cap S$, $\Omega \cap S$ and $\Delta \cap \Omega$.

Algorithm 1 PGD method for the problem (1)-(3)

Initialization: Select $\tau = \frac{1}{\|z\|^{\gamma}}$ for $\gamma > 0$ and an initial point $x^0 \in \Xi$. Set $k = 0$.

while the stopping conditions are not satisfied do

$$x^{k+1} \in \arg \min_{x \in \Xi} \left\{ \frac{1}{2} \|x - (x^k - \tau Mx^k)\|^2 \right\} \text{ with } M = A \text{ or } B. \quad (24)$$

end while

Due to the compactness of the set $\Xi$, clearly, Algorithm 1 is well defined. Next we establish the properties of the sequence $\{x^k\}$ generated by Algorithm 1. Unless otherwise stated, the function $F$ appearing in the sequel means one of $\Psi$, $\Phi$ and $\Theta$. 

13
Lemma 5.3 Let \( \{x^k\}_{k \in \mathbb{N}} \) be the sequence generated by Algorithm 1. Then, it holds that
\[
F(x^{k+1}) \leq F(x^k) - \gamma \|x^{k+1} - x^k\|^2,
\]
and hence \( \sum_{k=1}^{\infty} \|x^{k+1} - x^k\|^2 < \infty \), which implies that \( \lim_{k \to \infty} \|x^{k+1} - x^k\| = 0 \).

**Proof:** By the optimality of \( x^{k+1} \) and the feasibility of \( x^k \) to the problem (24), we have
\[
2 \tau (M x^k, x^{k+1} - x^k) + \|x^{k+1} - x^k\|^2 \leq 0,
\]
or equivalently,
\[
\frac{2}{\|M\|} (M x^k, x^{k+1} - x^k) + \|x^{k+1} - x^k\|^2 \leq -\frac{\gamma}{\|M\|} \|x^{k+1} - x^k\|^2.
\]
Notice that \( \{x^k\}_{k \in \mathbb{N}} \subset \Xi \). Together with the definition of \( F \), it follows that
\[
F(x^{k+1}) - F(x^k) = \langle x^{k+1}, M x^{k+1} \rangle - \langle x^k, M x^k \rangle
\]
\[
\leq 2 \langle M x^k, x^{k+1} - x^k \rangle + \|M\| \|x^{k+1} - x^k\|^2
\]
\[
\leq \|M\| \|M\|^{-1} \langle M x^k, x^{k+1} - x^k \rangle + \|x^{k+1} - x^k\|^2
\]
\[
\leq -\gamma \|x^{k+1} - x^k\|^2.
\]
This yields the desired result. The proof is completed. \( \square \)

For each \( k \in \mathbb{N} \), by the optimality of \( x^k \) to the problem (24) and [31, Exercise 8.8(c)],
\[
0 \in x^k - x^{k-1} + \tau M x^{k-1} + N_{\Xi}(x^k).
\]
Along with the definition of \( F \), one may upper bound the elements of \( \partial F(x^k) \) as below.

**Lemma 5.4** Let \( \{x^k\} \) be the sequence generated by Algorithm 1. For each \( k \in \mathbb{N} \), define
\[
u^k := 2 M (x^k - x^{k-1}) - 2 \tau^{-1} (x^k - x^{k-1}).
\]
Then, \( \nu^k \in \partial F(x^k) \) for each \( k \) and
\[
\|\nu^k\| \leq 2 \|M\| + \gamma \|x^k - x^{k-1}\|.
\]

**Proposition 5.1** Let \( \{x^k\}_{k \in \mathbb{N}} \) be the sequence generated by Algorithm 1, and denote by \( \varpi(x^0) \) the set of limit points of \( \{x^k\}_{k \in \mathbb{N}} \). Then, the following assertions hold:

(i) \( \emptyset \neq \varpi(x^0) \subseteq \text{crit} F \) and \( \lim_{k \to \infty} \text{dist}(x^k, \varpi(x^0)) = 0 \);

(ii) \( \varpi(x^0) \) is a nonempty, compact and connected set;

(iii) The function \( F \) is finite and constant on \( \varpi(x^0) \).

**Proof:** (i) Since \( \{x^k\}_{k \in \mathbb{N}} \subset \Xi, \varpi(x^0) \neq \emptyset \). Take an arbitrary \( x^* \) from \( \varpi(x^0) \). Then, there exists a subsequence \( \{x^{k_q}\}_{q \in \mathbb{N}} \) such that \( x^{k_q} \to x^* \) as \( q \to \infty \). From Lemma 5.3, it follows that \( x^{k_q-1} \to x^* \) as \( q \to \infty \). Since \( \{x^k\} \subset \Xi \), we have \( F(x^{k_q}) = \langle x^{k_q}, M x^{k_q} \rangle \), and hence \( \lim_{q \to \infty} F(x^{k_q}) = \Psi(x^*) \). By Lemma 5.4, for each \( q \in \mathbb{N} \), it holds that
\[
2 M (x^{k_q} - x^{k_q-1}) - 2 \tau^{-1} (x^{k_q} - x^{k_q-1}) \in \partial F(x^{k_q}).
\]
Taking the limit $q \to \infty$ and using Remark 2.1(ii) yields $0 \in \partial F(x^*)$. Thus, we obtain $\varpi(x^0) \subseteq \text{crit } F$, and part (i) follows. Using the same arguments as those for [7, Lemma 5(iii)] yields part (ii). We next prove part (iii). Take an arbitrary $x^* \in \varpi(x^0)$. So, there exists a subsequence $\{x^{kq}\}_{q \in \mathbb{N}}$ such that $x^{kq} \to x^*$ as $q \to \infty$. By Lemma 5.3, the sequence $\{F(x^{kq})\}_{k \in \mathbb{N}}$ is convergent, and denote its limit by $\omega^*$. Since $\lim_{q \to \infty} F(x^{kq}) = F(x^*)$, we have $F(x^*) = \omega^*$. By the arbitrariness of $x^*$ in $\varpi(x^0)$, part (iii) then follows. □

Recall that $F$ is a KL function of exponent $1/2$ by Proposition 3.1, 3.2 and 4.1. By invoking Lemma 5.3-5.4 and Proposition 5.1 and following the similar arguments as those for [4, Theorem 3.2 & 3.4] or [7, Theorem 1], we have the following convergence result.

**Theorem 5.1** Let $\{x^k\}_{k \in \mathbb{N}}$ be the sequence generated by Algorithm 1. Then,

(i) the sequence $\{x^k\}_{k \in \mathbb{N}}$ has a finite length, i.e., $\sum_{k=1}^{\infty} \|x^{k+1} - x^k\| < \infty$.

(ii) the sequence $\{x^k\}_{k \in \mathbb{N}}$ converges R-linearly to a critical point of $F$;

(iii) the sequence $\{F(x^k)\}_{k \in \mathbb{N}}$ is R-linearly convergent.

6 Numerical experiments

In this section, we empirically evaluate the performance of Algorithm 1 on synthetic and real datasets by making numerical comparisons with the approximate Spannogram [3] proposed for nonnegative sparse PCA. It is known that Spannogram outperforms previous algorithms, for example, the NSPCA [44] and the EM algorithm [33]. All numerical results are computed by a laptop running on 64-bit Windows Operating System with an Intel(R) Core(TM) i7-7700 CPU 2.8GHz and 16.00 GB RAM.

In the rest of this section, for a given covariance matrix $\Sigma$, we compute the first $k \geq 1$ nonnegative sparse PCs of $\Sigma$ with Algorithm 1 in the following way:

(S1) Take $\Sigma^0 = \Sigma$. Solve the problem (1) for $A = -\Sigma^0$ with Algorithm 1 to yield the first nonnegative sparse PC $v^1$ of $\Sigma$. Set $k = 1$.

(S2) Generate $\Sigma^k$ by removing the contribution of $v^k$ from $\Sigma^{k-1}$ with the projection deflation method in [25], and then solve the problem (1) for $A = -\Sigma^k$ with Algorithm 1 to yield the $(k+1)$th nonnegative sparse PC $v^{k+1}$ of $\Sigma$.

(S3) Set $k \leftarrow k + 1$, and go to Step (S2).

Throughout the experiment, for Algorithm 1 we choose $\gamma = 10^{-5}$ and use the absolute value of the first largest eigenvector of $-A$ (and $B$ for the problem (2)) as the starting point; while for Spannogram we adopt the default parameter and starting point as in the code, which can be downloaded from http://megasthenis.github.io/.
6.1 Synthetic examples

We shall evaluate the performance of Algorithm 1 by applying it to data whose covariance matrix $\Sigma$ has sparse eigenvectors. Suppose that the data from $\mathbb{R}^p$ is such that the $q$ ($q < p$) leading eigenvectors of the covariance matrix $\Sigma$ are sparse. The data matrix $X \in \mathbb{R}^{n \times p}$ is generated in the same way as in [32]. Let $v^1, \ldots, v^q$ be the first $q$ eigenvectors which are specified to be sparse and orthonormal. To generate the remaining $p-q$ eigenvectors which are not specified to be sparse, we form $\tilde{V} = [v^1 \cdots v^q \hat{V}]$ where $\hat{V} \in \mathbb{R}^{p \times (q-p)}$ is drawn from the standard normal distribution and apply the Gram-Schmidt orthogonalization method to $\tilde{V}$ to obtain an orthogonal matrix $V = [v^1 \cdots v^q v^{q+1} \cdots v^p]$; generate the positive real numbers $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ to form the positive definite diagonal matrix $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$; let $Z$ be $n$ samples which are random draw from the standard normal distribution $N(0, I_p)$; and set $X = V\Lambda^{1/2}Z$. Then, $\text{cov}(X) = \Sigma = V\Lambda V^T$.

Example 6.1 Take $q = 2$ and let $v^1 \in \mathbb{R}^p$ and $v^2 \in \mathbb{R}^p$ with nonzero entries specified as

$v^1 = \cdots = v^8 = \frac{1}{\sqrt{10}}, v^9 = v^{10} = -\frac{1}{\sqrt{10}}, v^{11} = \cdots = v^{18} = \frac{1}{\sqrt{10}}, v^{19} = v^{20} = -\frac{1}{\sqrt{10}}.$

We take $\lambda_1 = 20, \lambda_2 = 10$ and $\lambda_i = 1$ for $i = 3, 4, \ldots, 2000$. Table 1 reports the simulation results of the sample size $n = 200$ and $n = 500$ yielded by Algorithm 1 for solving (1) with $A = -\Sigma$ and Spannogram, respectively, which are the average of the results obtained by running two methods for 200 random problems.

Table 1: Numerical results of Algorithm 1 and Spannogram for Example 6.1

| $n$ | Methods   | $v^1$ Angle | Correct(%) | $v^2$ Angle | Correct(%) | Time(s) |
|-----|-----------|-------------|------------|-------------|------------|---------|
| 200 | Algorithm 1 | 26.707 | 1 | 27.004 | 0.995 | 0.46 |
|     | Spannogram | 26.709 | 1 | 27.026 | 0.995 | 0.77 |
| 500 | Algorithm 1 | 26.621 | 1 | 26.623 | 1 | 0.39 |
|     | Spannogram | 26.693 | 1 | 26.702 | 1 | 1.84 |

In Table 1, Angle column reports the angle between the nonnegative sparse PCs and the real one (since the PCs yielded by the methods may be different from the true one), and Correct column denotes the percentage of correctly identified zero loadings. From Table 1, we see that Algorithm 1 yields smaller angles than Spannogram does, although the percentage of correctly identified zero loadings are the same, and as the sample size increases, the percentage of correct identification of two methods increases.

6.2 Feature extraction of faces

In this part, we use the well-known Yale face database [17] to evaluate the performance of Algorithm 1. The database contains 165 gray-scale images of 15 individuals. Among others, there are 11 images for per subject, and each image corresponds
to a different facial expression or configuration. The data can be downloaded from http://cvc.yale.edu/projects/yalefaces/yale-faces.html. The original pixel size of the face image is $320 \times 243$, but we here crop the image into $100 \times 100$ pixel size for numerical testing. The first row of Figure 1 consists of five samples of the Yale face dataset, and the second row lists the corresponding recovery images yielded by the standard PCA, i.e., the five eigenvectors of the covariance matrix of the dataset. We reconstruct the images in the first row by taking $A$ as their negative covariance matrices and using Algorithm 1 to solve (1) under different sparsity, and make numerical comparisons with Spannogram.

Figure 1: Five samples and the image reconstructed by the standard PCA

![Figure 1](image1.png)

Figure 2 lists the images reconstructed by the two methods, where the images in group A, C and E are reconstructed in terms of the first five nonnegative sparse PCs yielded by Algorithm 1 under sparsity $\kappa = 1000, 2000$ and $3000$, respectively, while the images in group B, D and F are reconstructed in terms of the first five nonnegative sparse PCs yielded by Spannogram under the corresponding sparsity. We see that the two methods have comparable performance in the reconstruction of these images.

6.3 Unsupervised gene selection

Gene expression data from DNA microarrays provides the expression level of thousand of genes across several hundreds of experiments. Since the number of genes is larger than the sample size, it is necessary to reduce the dimension of the data. It is well known that the PCA is usually a linear combination of all genes, and has a difficulty in reducing dimensionality and hence in explaining. In this part, we apply Algorithm 1 for computing the nonnegative sparse PCs, and evaluate its performance with the Leukemia dataset [2] (see http://portals.broadinstitute.org/cgi-bin/cancer/publications).
Figure 2: Five face images reconstructed by Algorithm 1 and Spannogram
We use Algorithm 1 and Spannogram to compute the first nonnegative sparse PC of the sample covariance matrix of the Leukemia dataset, containing 72 samples and 12582 genes. Figure 3 depicts the explained variance of the first nonnegative sparse PC yielded by the two methods under different sparsity. One may see that for the sparsity \( \kappa < 500 \), the proportion of explained variance yielded by Algorithm 1 is a little lower than that of Spannogram, but for the sparsity \( \kappa \geq 500 \), the proportions of explained variance given by Algorithm 1 is a little higher than that of Spannogram. Figure 4 plots the cumulative variance of the first \( k \) nonnegative sparse PCs yielded by Algorithm 1 and Spannogram under a fixed sparsity. We see that for \( \kappa = 1000 \), the cumulative variance for the first 10 nonnegative sparse PCs are almost the same for the two methods, but for \( \kappa = 2000 \), the cumulative variance of Algorithm 1 is a little higher than that of Spannogram.

From the previous numerical comparisons, we conclude that Algorithm 1 has a little better performance than Spannogram does for computing nonnegative sparse eigenvectors in terms of the correct identification ratio or the proportion of the explained variance. For those problems with a small or medium scale, Algorithm 1 requires less computing time than Spannogram does (see Table 1), but for those problems with a large scale, say, this gene selection problem, Algorithm 1 requires more ten times computing time than Spannogram does since the former needs to estimate the largest eigenvector.
6.4 Sparse portfolio selection

In this part, we apply Algorithm 1 for computing the sparse portfolio with the real market data same as in [11], which is available from http://host.uniroma3.it/docenti/cesarone/DataSets.htm. The data consists of five real-world datasets: DJIA, HSI, STOXX50, NDX and FTSE, and each dataset consists of daily prices from Thomson Results Datastream. We evaluate the performance of Algorithm 1 in terms of out-of-sample portfolio variance (Var), out of-sample portfolio Sharpe ratio (SR), and portfolio turnover. During the testing, we convert the daily prices into weekly ones, and adopt the rolling-horizon procedure similar to [12] but with weekly rebalancing to compute the sparse portfolio. Table 2 summarizes the results of Var, SR and Turnover when applying Algorithm 1 for solving the problem (2) with $\kappa = 10$ and estimate window 52.

Table 2: Performance of Algorithm 1 for the sparse portfolio

|        | DJIA | FTSE100 | HSI | STOXX50 | NAS |
|--------|------|---------|-----|---------|-----|
| Var($\times 10^{-4}$) | 3.83 | 3.22    | 3.66| 4.50    | 4.44|
| SR     | 0.093| 0.100   | 0.110| 0.064   | 0.116|
| Turnover | 0.189| 0.393   | 0.158| 0.200   | 0.289|
7 Conclusions

We have established the KL property of exponent $1/2$ for the extended-valued objective functions $\Phi$, $\Psi$ and $\Theta$ from nonnegative sparse eigenvalue problems and sparse portfolio problems, and as a byproduct of this result, provided the global and linear convergence for the PGD method. Since the KL property of exponent $1/2$ is lack of stability, its research is generally tough for those functions with complicated structure and is required case by case. In our future research work, we shall explore the KL property of exponent $1/2$ of other extended-valued objective functions associated to sparse and low-rank optimization.

References

[1] A. d’Aspremont, L. El Ghaoui, M. I. Jordan and G. R. Lanckriet, A direct formulation for sparse PCA using semidefinite programming, SIAM Review, 49(2007): 434-448.

[2] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub and S. J. Korsmeyer Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature Genetics, 30(2001): 41-47.

[3] M. Asteris, D. Papailiopoulos and A. Dimakis, Nonnegative sparse PCA with provable guarantees, International Conference on Machine Learning, 2014.

[4] H. Attouch, J. Bolte, P. Redont and A. Soubeyran, Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality, Mathematics of Operations Research, 35(2010): 438-457.

[5] H. Attouch, J. Bolte and B. F. 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): 91-129.

[6] L. Badea and D. Tiliwea, Sparse factorizations of gene expression data guided by binding data, Pacific Symposium on Biocomputing, 2005: 447.

[7] J. Bolte, S. Sabach and M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Mathematical Programming, 146(2014): 459-494.

[8] J. Bolte, T. P. Nguyen, J. Peypouquet and B. W. Suter, From error bounds to the complexity of first-order descent methods for convex functions, Mathematical Programming, 165(2017): 471-507.
[9] J. Brodie, I. Daubechies, C. De Mol, D. Giannone and I. Loris, Sparse and stable Markowitz portfolios, Proceedings of the National Academy of Sciences, 106(2009): 12267-12272.

[10] Y. Cui, D. F. Sun and K. C. Toh, On the asymptotic superlinear convergence of the augmented Lagrangian method for semidefinite programming with multiple solutions, arXiv:1610.00875, 2016.

[11] A. Carleo, F. Cesaroni, A. Gheno and J. M. Ricci, Approximating exact expected utility via portfolio efficient frontiers, Decisions in Economics and Finance, 40(2017): 115-143.

[12] V. DeMiguel, L. Garlappi, F. J. Nogales and R. Uppal, A generalized approach to portfolio optimization: improving performance by constraining portfolio norms, Management Science, 55(2009): 798-812.

[13] A. D. Ioffe and J. V. Outrata, On metric and calmness qualification conditions in subdifferential calculus, Set-Valued Analysis, 16(2008): 199-227.

[14] R. Jagannathan and T. S. Ma, Risk reduction in large portfolios: Why imposing the wrong constraints helps, Journal of Finance, 58(2003): 1651-1684.

[15] I. T. Jolliffe, N. T. Trendafilov and M. Uddin, A modified principal component technique based on the lasso, Journal of Computational and Graphical Statistics, 12(2003): 531-547.

[16] M. Journée, Y. Nesterov, P. Richtárik and R. Sepulchre, Generalized power method for sparse principal component analysis, Journal of Machine Learning Research, 11(2010): 517-553.

[17] N. Kwak, Principal component analysis based on L1-norm maximization, IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(2008): 1672-1680.

[18] D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature, 401(1999): 788-791.

[19] G. Y. Li and T. K. Pong, Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods, Foundations of Computational Mathematics, 18(2018): 1199-1232.

[20] H. K. Liu, W. J. Wu and A.M.-C. So, Quadratic optimization with orthogonality constraints: explicit Łojasiewicz exponent and linear convergence of line-search methods, Proceedings of Machine Learning Research, 46(2016): 1158-1167.

[21] T. X. Liu, T. K. Pong and A. Takeda, A successive difference-of-convex approximation method for a class of nonconvex nonsmooth optimization problems, Mathematical Programming, 2017, 10.1007/s10107-018-1327-8.
[22] Z. S. Lu and Y. Zhang, *An augmented Lagrangian approach for sparse principal component analysis*, Mathematical Programming, 135(2012): 149-193.

[23] Z. Q. Luo and P. Tseng, *On the linear convergence of descent methods for convex essentially smooth minimization*, SIAM Journal on Control and Optimization, 30(1992): 408-425.

[24] F. J. Luque, *Asymptotic convergence analysis of the proximal point algorithm*, SIAM Journal on Control and Optimization, 22(1984): 277-293.

[25] L. W. Mackey, *Deflation methods for sparse PCA*, Advances in Neural Information Processing Systems 21 (NIPS 2008).

[26] B. Moghaddam, Y. Weiss and S. Avidan, *Spectral bounds for sparse PCA: exact and greedy algorithms*, International Conference on Neural Information Processing Systems, 2005.

[27] B. Moghaddam, Y. Weiss and S. Avidan, *Generalized spectral bounds for sparse lda*, In Proceedings of the 23rd International Conference on Machine Learning (ICML’06), pages 641-648, 2006.

[28] S. H. Pan, D. D. Zhang and Y. L. Liu, *Metric subregularity of subdifferential and KL property of exponent 1/2*, submitted to Journal of Optimization Theory and Applications, 2018.

[29] S. M. Robinson, *Some continuity properties of polyhedral multifunctions*, Mathematical Programming Study, 14(1981): 206–214.

[30] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

[31] R. T. Rockafellar and R. J-B. Wets, *Variational Analysis*, Springer, 1998.

[32] H. Shen and J. Z. Huang, *Sparse principal component analysis via regularized low rank matrix approximation*, Journal of Multivariate Analysis, 99(2008): 1015-1034.

[33] C. D. Sigg and J. M. Buhmann, *Expectation-maximization for sparse and non-negative PCA*, Proceedings of the 25th international conference on Machine learning (ICML), 2008: 960-967.

[34] P. Tseng and S. Yun, *A coordinate gradient descent method for nonsmooth separable minimization*, Mathematical Programming, 117(2009): 387-423.

[35] V. Q. Vu, J. Cho, J. Lei and K. Robe, *Fantope projection and selection: A near-optimal convex relaxation of sparse PCA*, Advances in Neural Information Processing Systems, 17(2013): 367-369.

[36] R. Varshavsky, A. Gottlieb, M. Linial and D. Horn, *Novel unsupervised feature filtering of biological data*, Bioinformatics, 22(2006): 507-513.
From [31, Exercise 10.10] and Proposition 2.3, for any $x \in \mathbb{R}_+^p \cap S \cap \Omega$, we have

$$\partial \Psi(x) = 2Ax + N_{\mathbb{R}_+^p \cap S \cap \Omega}(x) \subseteq 2Ax + N_{\mathbb{R}_+^p \cap \Omega}(x) + N_S(x), \quad (25)$$

and if in addition $\|x\|_0 = \kappa$, then the inclusion will become an equality.

**The proof of Proposition 3.2:** Fix an arbitrary point $\bar{x} \in \text{crit} \Psi$. We proceed the arguments by two cases $\|\bar{x}\|_0 = \kappa$ and $\|\bar{x}\|_0 < \kappa$ as below.
Case 1: \( \| \mathbf{x} \|_0 = \kappa \). Write \( J = \text{supp}(\mathbf{x}) \). Let \( g \) be the function defined as in Lemma 2.3 with \( H = A_J \). By Lemma 2.3, \( g \) is a KL function of exponent \( 1/2 \). So, there exist \( \delta > 0, \eta > 0 \) and \( c > 0 \) such that for all \( z \in B(\mathbf{x}_J, \delta) \cap [g(\mathbf{x}_J) < g < g(\mathbf{x}) + \eta] \),

\[
\text{dist}(0, \partial g(z)) \geq c_1 \sqrt{g(z) - g(\mathbf{x}_J)}. \tag{26}
\]

Let \( x \) be an arbitrary point from the set \( B(\mathbf{x}, \delta) \cap [\Psi(\mathbf{x}) < \Psi < \Psi(\mathbf{x}) + \eta] \). Clearly, \( x \in S \cap \Omega \cap \mathbb{R}^+_p \). In addition, by reducing \( \delta \) if necessary, we have \( \text{supp}(x) \supseteq \text{supp}(\mathbf{x}) \). Together with \( \|x\|_0 \leq \kappa \), it follows that \( \text{supp}(x) = \text{supp}(\mathbf{x}) = J \). By equation (25),

\[
\text{dist}(0, \partial \Psi(x)) \geq \min_{\zeta \in N_S(x), \xi \in N_{x^+}(x)} \|2Ax + \zeta + \xi\|
\]

\[
= \min_{\omega \in \mathbb{R}, \xi \in N_{x^+}(x)} \|2Ax + \omega x + \xi\|
\]

\[
\geq \min_{\omega \in \mathbb{R}} \|2A_J x_J + \omega x_J\| = \text{dist}(0, \partial g(x_J)) \tag{27}
\]

where the second inequality is by \( \text{supp}(x) = J \) and the inclusion (6), and the last equality is by the definition of \( g \). In addition, from \( \text{supp}(x) = \text{supp}(\mathbf{x}) = J \), it follows that

\[
\Psi(x) - \Psi(\mathbf{x}) = x^T A x - x^T A \mathbf{x} = x^T A_J x_J - A \mathbf{x}_J = g(x_J) - g(\mathbf{x}_J)
\]

which, along with \( x \in [\Psi(\mathbf{x}) < \Psi < \Psi(\mathbf{x}) + \eta] \), implies that \( x_J \in [g(\mathbf{x}_J) < g < g(\mathbf{x}_J) + \eta] \). Notice that \( \|x_J - \mathbf{x}_J\| = \|x - \mathbf{x}\| \leq \delta \). Thus, from inequalities (26) and (27), we have

\[
\text{dist}(0, \partial \Psi(x)) = \text{dist}(0, \partial g(x_J)) \geq c \sqrt{g(x_J) - g(\mathbf{x}_J)} = c \sqrt{\Psi(x) - \Psi(\mathbf{x})}.
\]

By the arbitrariness of \( x \), the function \( \Psi \) has the KL property of exponent \( 1/2 \) at \( \mathbf{x} \).

Case 2: \( \| \mathbf{x} \|_0 < \kappa \). By the continuity, there exists \( \delta_1 > 0 \) such that for all \( x' \in B(\mathbf{x}, \delta_1) \), \( \text{supp}(x') \supseteq \text{supp}(\mathbf{x}) \). Let \( \mathcal{I} \) be the set defined in the proof of Proposition 3.1, and for each \( I \in \mathcal{I} \), let \( g_I \) be the function defined by (11). By Lemma 2.3, for each \( I \in \mathcal{I} \), there exist \( \delta_I > 0, \eta_I > 0 \) and \( c_I > 0 \) such that for all \( z \in B(\mathbf{x}_I, \delta_I) \cap [g_I(\mathbf{x}) < g_I < g_I(\mathbf{x}) + \eta_I] \),

\[
\text{dist}(0, \partial g_I(z)) \geq c_I \sqrt{g_I(z) - g_I(\mathbf{x}_I)}. \tag{28}
\]

Set \( \delta = \min(\delta_1, \min_{I \in \mathcal{I}} \delta_I), \eta = \min_{I \in \mathcal{I}} \eta_I \) and \( c := \min_{I \in \mathcal{I}} c_I \). Let \( x \) be an arbitrary point from the set \( B(\mathbf{x}, \delta) \cap [\Psi(\mathbf{x}) < \Psi < \Psi(\mathbf{x}) + \eta] \). Clearly, \( x \in S \cap \Omega \cap \mathbb{R}^+_p \). Hence, \( \|x\|_0 \leq \kappa \). Write \( J = \text{supp}(x) \). Clearly, \( J \supseteq \text{supp}(\mathbf{x}) \). From equation (25), we have

\[
\text{dist}(0, \partial \Psi(x)) \geq \min_{\zeta \in N_S(x), \xi \in N_{x^+}(x)} \|2Ax + \zeta + \xi\|
\]

\[
= \min_{\omega \in \mathbb{R}, \xi \in N_{x^+}(x)} \|2Ax + \omega x + \xi\|
\]

\[
\geq \min_{\omega \in \mathbb{R}} \|2A_J x_J + \omega x_J\| = \text{dist}(0, \partial g_J(x_J)) \tag{29}
\]

25
where the second inequality is due to \( \text{supp}(x) = J \) and \( \xi_J = 0 \) implied by the inclusion (6), and the last one is by the definition of \( g_J \). In addition, from \( J \supseteq \sigma(\mathbf{T}) \),

\[
\Psi(x) - \Psi(\mathbf{T}) = x^T A x - \mathbf{T}^T A \mathbf{T} = x^T A_J x_J - \mathbf{T}^T A_J \mathbf{T}_J = g_J(x_J) - g_J(\mathbf{T}_J)
\]

which by \( x \in [\Psi(\mathbf{T}) < \Psi < \Psi(\mathbf{T}) + \eta] \) implies that \( x_J \in [g(\mathbf{T}_J) < g_J < g_J(\mathbf{T}_J) + \eta] \). Notice that \( \|x_J - \mathbf{T}_J\| = \|x - \mathbf{T}\| \leq \delta \). Thus, from inequalities (28) and (29),

\[
\text{dist}(0, \partial \Psi(x)) \geq \text{dist}(0, \partial g_J(x_J)) \geq c \sqrt{g_J(x_J) - g_J(\mathbf{T}_J)} = c \sqrt{\Psi(x) - \Psi(\mathbf{T})}.
\]

By the arbitrariness of \( x \), the function \( \Psi \) has the KL property of exponent \( 1/2 \) at \( \mathbf{T} \).

The last two cases, along with the arbitrariness of \( \mathbf{T} \) in \( \text{crit}\Psi \), show that the function \( \Psi \) is a KL function of exponent \( 1/2 \). The proof is then completed. \( \square \)

**Lemma 1** For any given integer \( s \geq 1 \) and \( z \in \mathbb{R}^n \), we consider the following problems

\[
\min_{u \in \mathbb{R}^n} \left\{ \frac{1}{2} \|u - z^{\perp}\|^2 \text{ s.t. } \|u\|_0 \leq s, \langle e, u \rangle = 1, u_i \geq 0, i = 1, 2, \ldots, n \right\} \tag{30}
\]

and

\[
\min_{y \in \mathbb{R}^s} \left\{ \frac{1}{2} \|y - z^{\perp}\|^2 \text{ s.t. } \langle e, y \rangle = 1, y_i \geq 0, i = 1, 2, \ldots, s \right\}. \tag{31}
\]

If \( y^* \) is the unique optimal solution to the problem (31), then \( (y^*; 0) \in \mathbb{R}^n \) is a global optimal solution of the problem (30); conversely, if \( u^* \) is a global optimal solution of (30), then \( (u^*)^{s \perp} \) is the unique optimal solution to (31).

**Proof:** Let \( y^* = (y_1^*, \ldots, y_s^*)^T \) be the unique optimal solution to (31). Take an arbitrary feasible point \( u \in \mathbb{R}^n \) of (30). Then, \( (u_1^*, \ldots, u_s^*)^T \) must be the feasible to (31), which implies that \( \sum_{i=1}^s (u_i^* - z_i^*)^2 \geq \sum_{i=1}^s (y_i^* - z_i^*)^2 \). Consequently, we have that

\[
\|u - z\|^2 \geq \|u^* - z\|^2 = \sum_{i=1}^s (u_i^* - z_i^*)^2 + \sum_{i=s+1}^n (z_i^*)^2 \geq \sum_{i=1}^s (y_i^* - z_i^*)^2 + \sum_{i=s+1}^n (z_i^*)^2 = \|(y^*; 0) - z\|^2.
\]

Together with the feasibility of \( (y^*; 0) \) to (30), this shows that \( (y^*; 0) \) is optimal to (30).

Conversely, let \( y \in \mathbb{R}^s \) be an arbitrary feasible point of (31). Clearly, \( (y; 0) \) is feasible to (30). This means that \( \|(y; 0) - z\|^2 \geq \|u^* - z\|^2 \), which is equivalent to saying that

\[
\|y - z^{\perp}\|^2 + \sum_{i=s+1}^n (z_i^*)^2 \geq \|u^* - z^{\perp}\|^2 \geq \|u^{s \perp} - z^{\perp}\|^2
\]

\[
= \sum_{i=1}^s (u_i^{s \perp} - z_i^{s \perp})^2 + \sum_{i=s+1}^n (z_i^{s \perp})^2.
\]

The last inequality is equivalent to saying that \( \|y - z^{s \perp}\|^2 \geq \sum_{i=1}^s (u_i^{s \perp} - z_i^{s \perp})^2 \). By the arbitrariness of \( y \) in the feasible set of (31) and the feasibility of \( (u_1^{s \perp}, \ldots, u_s^{s \perp})^T \) to (31), we immediately obtain the desired result. The proof is then completed. \( \square \)