

S-Lemma with Equality and Its Applications

Yong Xia · Shu Wang · Ruey-Lin Sheu

Received: date / Accepted: date

Abstract Let $f(x) = x^T A x + 2a^T x + c$ and $h(x) = x^T B x + 2b^T x + d$ be two quadratic functions having symmetric matrices A and B . The S-lemma with equality asks when the unsolvability of the system $f(x) < 0, h(x) = 0$ implies the existence of a real number μ such that $f(x) + \mu h(x) \geq 0, \forall x \in \mathbb{R}^n$. The problem is much harder than the inequality version which asserts that, under Slater condition, $f(x) < 0, h(x) \leq 0$ is unsolvable if and only if $f(x) + \mu h(x) \geq 0, \forall x \in \mathbb{R}^n$ for some $\mu \geq 0$. In this paper, we show that the S-lemma with equality does not hold only when the matrix A has exactly one negative eigenvalue and $h(x)$ is a non-constant linear function ($B = 0, b \neq 0$). As an application, we can globally solve $\inf\{f(x) : h(x) = 0\}$ as well as the two-sided generalized trust region subproblem $\inf\{f(x) : l \leq h(x) \leq u\}$ without any condition. Moreover, the convexity of the joint numerical range $\{(f(x), h_1(x), \dots, h_p(x)) : x \in \mathbb{R}^n\}$ where f is a (possibly non-convex) quadratic function and $h_1(x), \dots, h_p(x)$ are affine functions can be characterized using the newly developed S-lemma with equality.

This research was supported by Taiwan National Science Council under grant 102-2115-M-006-010, by National Center for Theoretical Sciences (South), by National Natural Science Foundation of China under grant 11471325 and by Beijing Higher Education Young Elite Teacher Project 29201442.

Y. Xia

State Key Laboratory of Software Development Environment, LMIB of the Ministry of Education, School of Mathematics and System Sciences, Beihang University, Beijing, 100191, P. R. China E-mail: dearyxia@gmail.com

S. Wang

School of Mathematics and System Sciences, Beihang University, Beijing, 100191, P. R. China E-mail: wangshu.0130@163.com

R. L. Sheu

Department of Mathematics, National Cheng Kung University, Taiwan E-mail: rsheu@mail.ncku.edu.tw

Keywords S-lemma · Slater condition · Quadratically constrained quadratic program · Generalized trust region subproblem · Joint numerical range · Hidden convexity

Mathematics Subject Classification (2010) 90C20, 90C22, 90C26

1 Introduction

Let $f(x) = x^T A x + 2a^T x + c$ and $h(x) = x^T B x + 2b^T x + d$ be two quadratic functions having symmetric matrices A and B . In 1971, Yakubovich [36,37] proved a fundamental result, which we call the *classical S-lemma* in this paper. It asserts that, given any pair of quadratic functions (f, h) , if $h(x) \leq 0$ satisfies Slater's condition, namely, there is an $\bar{x} \in \mathbb{R}^n$ such that $h(\bar{x}) < 0$, the following two statements are always equivalent:

- (S₁) $(\forall x \in \mathbb{R}^n) \ h(x) \leq 0 \implies f(x) \geq 0$.
- (S₂) There exists a $\mu \geq 0$ such that $f(x) + \mu h(x) \geq 0, \forall x \in \mathbb{R}^n$.

In the following, let us denote the equivalence by (S₁)~(S₂) to mean that both statements are true or false synchronously.

Notice that (S₂) trivially implies (S₁), but the other way around from (S₁) to (S₂) is not so obvious. Yakubovich's proof (see also [27]) had to rely on (i) a homogenization scheme which transforms the nonhomogeneous system of (S₁) and (S₂) into homogeneous ones; (ii) if f and h are quadratic forms, the joint numerical range $\{(f(x), h(x)) : x \in \mathbb{R}^n\}$ is convex [9]; and (iii) Slater's condition for applying the separation theorem to provide the existence of $\mu \geq 0$.

The classical S-lemma is a powerful tool especially in control theory and robust optimization. See recent surveys in [8,27]. It is a form of the celebrated Farkas lemma [19]. It can be applied to solve and show the hidden convexity of the quadratic programming with a single quadratic inequality constraint (QP1QC) under Slater's condition. The connection is easily illustrated by

$$\begin{aligned}
 v(\text{QP1QC}) &= \inf_{x \in \mathbb{R}^n} \{f(x) : h(x) \leq 0\} \\
 &= \sup_{\lambda \in \mathbb{R}} \left\{ \lambda : \left\{ x \in \mathbb{R}^n \mid f(x) < \lambda, \ h(x) \leq 0 \right\} = \emptyset \right\} \\
 &= \sup_{\lambda \in \mathbb{R}} \left\{ \lambda : (\exists \mu \geq 0) \ f(x) - \lambda + \mu h(x) \geq 0, \forall x \in \mathbb{R}^n \right\} \quad (1) \\
 &= \sup_{\lambda \in \mathbb{R}, \mu \geq 0} \left\{ \lambda : \begin{bmatrix} A + \mu B & a + \mu b \\ a^T + \mu b^T & c + \mu d - \lambda \end{bmatrix} \succeq 0 \right\} \quad (2)
 \end{aligned}$$

where $v(\cdot)$ is the optimal value of problem (\cdot) and the notation $X \succeq 0$ implies that X is positive semidefinite. Notice that the key step (1) is due to (S₁)~(S₂), and the SDP result in (2) shows the hidden convexity of (QP1QC).

The classical S-lemma has several forms of generalization. Jeyakuma et al. in [21] extended \mathbb{R}^n in (S₁) and (S₂) to any linear manifold. The result was

applied to characterize the global optimality of (QP1QC) equipped with additional linear equality constraints. When three quadratic functions are considered, under the assumption that there exists a positive definite matrix pencil of the three quadratic functions, Jeyakuma et al. in [21] also gave a type of alternative theorem involving only strict inequalities. Generalizations of the classical S-lemma having two or more different h 's are referred to as *S-procedure*. See [8, 27] for a survey. Particularly, Fradkov and Yakubovich [13] proved that the strong duality holds for nonconvex quadratic optimization with two quadratic constraints in *complex* variables [3]. Jeyakumar et al. in [20] proved that, in the absence of Slater's condition but assuming that $\{x : h(x) \leq 0\} \neq \emptyset$, $(S_1) \sim (S_2)$ cannot be true for all f 's. However, (f, g) still satisfies a weaker equivalence $(S_1) \sim (S_2)$, the so-called *regularized S-lemma* below:

$$\begin{aligned} (S_1) \quad & (\forall x \in \mathbb{R}^n) \quad h(x) \leq 0 \implies f(x) \geq 0. \\ (S_2^r) \quad & (\forall \epsilon > 0) \quad (\exists \lambda_\epsilon \geq 0) \quad (\forall x \in \mathbb{R}^n) \quad f(x) + \lambda_\epsilon h(x) + \epsilon(x^T x + 1) \geq 0. \end{aligned}$$

It is weaker because (S_2) implies (S_2^r) by letting $\lambda_\epsilon = \mu \geq 0$, $\forall \epsilon > 0$.

This paper studies the *S-lemma with equality*. In the formality, it replaces $h \leq 0$ in (S_1) by $h = 0$ and $\mu \geq 0$ in (S_2) by $\mu \in \mathbb{R}$. It asks, for what pairs of (f, h) the following two statements are equivalent (i.e. $(E_1) \sim (E_2)$):

$$\begin{aligned} (E_1) \quad & (\forall x \in \mathbb{R}^n) \quad h(x) = 0 \implies f(x) \geq 0. \\ (E_2) \quad & \text{There exists a } \mu \in \mathbb{R} \text{ such that } f(x) + \mu h(x) \geq 0, \forall x \in \mathbb{R}^n. \end{aligned}$$

With the replacement, (S_1) is relaxed to (E_1) whereas (S_2) is relaxed to (E_2) . Though it is easy to see that $(E_1) \sim (E_2)$ does not always hold, yet it is by no means trivial to characterize conditions that (f, h) should satisfy to make $(E_1) \sim (E_2)$ correct. This is going to be the main theme of this paper.

To begin with the discussion, we always assume that $\{x : h(x) = 0\} \neq \emptyset$. In literature, the first variant of the S-lemma with equality was proposed by Finsler [12] in 1937 for the homogeneous system. Different proofs of Finsler's Theorem can be found in [14, 24]. It states that the statement

$$(\forall x \in \mathbb{R}^n, x \neq 0) \quad x^T B x = 0 \implies x^T A x > 0 \quad (3)$$

is true if and only if there exists a $\mu \in \mathbb{R}$ such that $A + \mu B \succ 0$. However, from the simple example that $f(x_1, x_2) = x_1 x_2$, $h(x_1, x_2) = -x_1^2$, we see that $h(x_1, x_2) = -x_1^2 = 0$ implies that $f(x_1, x_2) = 0$ so (E_1) is true but (3) is not valid. Finsler's Theorem then asserts that there is no $\mu \in \mathbb{R}$ such that $A + \mu B \succ 0$, but, indeed, there is no $\mu \in \mathbb{R}$ such that $f(x) + \mu h(x) = x_1 x_2 - \mu x_1^2 \geq 0$, $\forall x \in \mathbb{R}^n$. Thus (E_2) is false and $(E_1) \not\sim (E_2)$. The example shows that the equivalence of (E_1) and (E_2) is in general not true, even just for a simple homogeneous system by a slight generalization from Finsler's Theorem. We notice that in this example $h(x_1, x_2) = -x_1^2 \leq 0$ satisfies Slater's condition so that $(S_1) \sim (S_2)$. It does not satisfy the following "two-side" Slater's condition though.

Assumption 1 $h(x)$ takes both positive and negative values, i.e., there are $x', x'' \in \mathbb{R}^n$ such that $h(x') < 0 < h(x'')$.

Assumption 1 is obviously stricter than the usual Slater's condition. However, even when it is imposed, $(E_1) \sim (E_2)$ may still be invalid. For example, let $f(x_1, x_2) = x_1^2 - x_2^2$, $h(x_1, x_2) = x_2$ where $h(x_1, x_2) = x_2$ satisfies Assumption 1. We can check that $h(x_1, x_2) = x_2 = 0$ implies that $f(x_1, 0) = x_1^2 \geq 0$, but there is still no $\mu \in \mathbb{R}$ such that $f + \mu h = x_1^2 - x_2^2 + \mu x_2 \geq 0$, $\forall x_1, x_2 \in \mathbb{R}$. So $(E_1) \not\sim (E_2)$. However, since h satisfies Slater's condition, there must be $(S_1) \sim (S_2)$. A simple verification shows that " $h(x_1, x_2) = x_2 \leq 0 \not\Rightarrow f(x_1, x_2) = x_1^2 - x_2^2 \geq 0$ " so that both (S_1) and (S_2) are false in this case.

There were several attempts trying to establish some affirmative results for $(E_1) \sim (E_2)$, but they all came with an incomplete sufficient condition. Analogous to the role of Slater's condition in $(S_1) \sim (S_2)$, all sufficient conditions for $(E_1) \sim (E_2)$ are subject to Assumption 1. In literature, those sufficient conditions are

S-Condition 1 ([27]): $h(x)$ is strictly concave (or strictly convex).

S-Condition 2 ([3], Thm. A.2): There is an $\eta \in \mathbb{R}$ such that $A \succeq \eta B$.

S-Condition 3 ([33], Corollary 6): $h(x)$ is homogeneous.

S-Condition 4 ([25]): $h(0) = 0$ and there exists $\zeta \in X = \{x \in \mathbb{R}^n : h(x) = 0\}$ such that

$$(\forall x \in \mathbb{R}^n) \quad x^T B x = 0 \implies (B\zeta + b)^T x = 0. \quad (4)$$

We defer the discussion about the relations among S-Conditions 1 - 4 to Appendix for readers who are interested, but consider the following example, which shows that none of S-Conditions 1 - 4 can become necessary.

Let $f(x_1, x_2) = -x_1^2 - x_2^2$, $h(x_1, x_2) = x_2$. We check that $h(x_1, x_2) = x_2$ satisfies Assumption 1. For (E_1) , $h(x_1, x_2) = x_2 = 0$ does not imply that $f(x_1, 0) = -x_1^2 \geq 0$, $\forall x_1 \in \mathbb{R}$. For (E_2) , there is no $\mu \in \mathbb{R}$ such that $f + \mu h = -x_1^2 - x_2^2 + \mu x_2 \geq 0$, $\forall x_1, x_2 \in \mathbb{R}$. Since both (E_1) and (E_2) are false, $(E_1) \sim (E_2)$ in this example. However, this affirmative example cannot be characterized by either S-Condition above. We can verify that h is not strictly convex (concave); since $B = 0$, $A \prec 0$, there is no $\eta \in \mathbb{R}$ such that $A \succeq \eta B$; $h = 0$ is a hyperplane so it is not homogeneous; finally $h(0) = 0$, but due to $B = 0$, $b \neq 0$, S-Condition 4 does not hold.

From the above discussion, we learn that $(E_1) \sim (E_2)$ is a much harder problem than the classical S-lemma $(S_1) \sim (S_2)$. No attempt has been successfully made so far. Our approach is to break down the problem into several cases, each of which encompasses a unique algebraic and/or geometrical feature of (f, h) for easy analysis. We found that, when $h(x)$ takes both positive and negative values, the statement (E_1) (i.e. $\inf_{h(x)=0} f(x) \geq 0$ holds) *nearly implies* that there is a scalar μ to adjust the size and/or the direction of h such that the linear combination $f(x) + \mu h(x)$ becomes convex with an attainable minimum (on the entire \mathbb{R}^n , not just on $h(x) = 0$). The only exception is when $h(x) = 0$ is a "flat" $n - 1$ dimensional hyperplane on which f is convex while on the remaining dimension (complement to $h(x) = 0$) f becomes concave. If described in the algebraic way, under Assumption 1, if $B \neq 0$, there must be $(E_1) \sim (E_2)$. The only type of examples that can have $(E_1) \not\sim (E_2)$ is when,

and only when $B = 0$, $b \neq 0$, the matrix A has exactly one negative eigenvalue, and the matrix in (17) is positive semi-definite. The result explains why $f(x_1, x_2) = -x_1^2 - x_2^2$, $h(x_1, x_2) = x_2$ gets an affirmative result (the matrix A has two negative eigenvalues), while $f(x_1, x_2) = x_1^2 - x_2^2$, $h(x_1, x_2) = x_2$ gets a negative assertion ($B = 0$, $b^T = (0, 1/2)$, A has exactly one negative eigenvalue, and the matrix in (17) is $\text{Diag}(1, 0) \succeq 0$). Theorem 3 in Section 3 gives the complete statement as well as the proof for the S-lemma with equality to hold under Assumption 1, whereas the characterization of the theorem when Assumption 1 fails will be treated in Section 2.

Now we turn to the applications of the S-lemma with equality. The importance of the S-lemma with equality was not understood by us until we eventually realized from [25] that it is the key to solve the following long-standing interval bounded generalized trust region subproblem (studied in [29] by Pong and Wolkowicz and also by many other researchers)

$$\begin{aligned} \text{(GTRS)} \quad & \inf f(x) \\ \text{s.t.} \quad & l \leq h(x) \leq u. \end{aligned} \tag{5}$$

If $f(x)$ is convex and the optimal solution to the unconstrained problem $\min_{x \in \mathbb{R}^n} f(x)$ happens to be feasible to (5) (by checking whether $\nabla f(x) = 0$, $l \leq h(x) \leq u$ has a solution), then the optimal value $v(\text{GTRS}) = \min_{x \in \mathbb{R}^n} f(x)$. Otherwise, the optimal solution is located at one of the two boundaries:

$$v(\text{GTRS}) = \min \left\{ \inf_{h(x)=l} f(x), \inf_{h(x)=u} f(x) \right\}.$$

It reduces (GTRS) to

$$\begin{aligned} \text{(QP1EQC)} \quad & \inf f(x) \\ \text{s.t.} \quad & h(x) = 0. \end{aligned}$$

With the same idea for solving (QP1QC) by applying the classical S-lemma in (1)-(2), (QP1EQC) can be similarly proved to have the strong duality by the S-lemma with equality. We also discuss when (QP1EQC) becomes unbounded below, and give a necessary and sufficient condition to characterize when (QP1EQC) is attainable. As a consequence, both (QP1EQC) and (GTRS) are completely analyzed.

We remark that (QP1EQC) itself has many interesting applications, including the double well potential optimization problem [10,35], the time of arrival geolocation problem [15] and unbiased least squares optimization for system identification [26]. In particular, the double well potential model came from numerical approximations to the generalized Ginzburg-Landau functionals [18]. It minimizes the following special type of multi-variate polynomial of degree 4:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \left(\frac{1}{2} \|Qx - c\|^2 - d \right)^2 + \frac{1}{2} x^T A x - a^T x$$

where $Q \neq 0$ is an $m \times n$ matrix. We can reformulate it as an example of (QP1EQC):

$$\begin{aligned} \text{(DWP)} \quad & \min_{x \in \mathbb{R}^n, z \in \mathbb{R}} \frac{1}{2}z^2 + \frac{1}{2}x^T A x - a^T x \\ & \text{s.t.} \quad \frac{1}{2}\|Qx - c\|^2 - d - z = 0. \end{aligned}$$

In the constraint function, the variable z does not have a second order term, so it is not strictly concave (or strictly convex). When A has a negative eigenvalue whose eigenvector is not in the range of $Q^T Q$, there is no $\eta \in \mathbb{R}$ such that $A \succeq \eta Q^T Q$. Moreover, (DWP) is in general non-homogeneous. Finally, the model obviously fails an equivalent statement of S-Condition 4 in (64), which we derive in Appendix. In other words, none of the existing results leads a complete answer to (DWP).

Finally in Section 5, by applying the S-lemma with equality, we obtain a *necessary and sufficient* description for the convexity of the joint numerical range $S = \{(f(x), h_1(x), \dots, h_p(x)) : x \in \mathbb{R}^n\}$ where h_1, \dots, h_p are all affine functions. Dines in 1941 [9] showed that the joint numerical range $M = \{(f(x), h(x)) : x \in \mathbb{R}^n\}$ is a convex subset in \mathbb{R}^2 when f and h are quadratic forms. Dines' theorem later became a major machinery for Yakubovich to prove the famous classical S-lemma in 1971. However, it has been shown by Polyak in 1998 [28] that Dines' result is in general not true even when one of f and h is affine. Beck [2] has shown that, when f is strictly convex and $p \leq n-1$, S is closed and convex, but S is non-convex when $p = n$ and h_1, h_2, \dots, h_n have linearly independent normals. Our necessary and sufficient conditions for S being convex cover Polyak's counterexample and Beck's result as special cases.

Below we highlight a list of main contributions in the paper.

- When Assumption 1 fails, Theorem 1 characterizes the necessary and sufficient conditions under which $(E_1) \sim (E_2)$. (Sect. 2)
- When Assumption 1 fails, Theorem 2 gives a “regularized S-lemma with equality” $(E_1) \sim (E_2^r)$. The classical regularized S-lemma $(S_1) \sim (S_2^r)$ by Jeyakumar et al. in [20] is shown to be a direct consequence of $(E_1) \sim (E_2^r)$. (Sect. 2)
- When Assumption 1 holds, Theorem 3 characterizes the necessary and sufficient conditions under which $(E_1) \sim (E_2)$. (Sect. 3)
- As an application of Theorem 3, under Slater's condition, the classical S-lemma $(S_1) \sim (S_2)$ is shown to be a direct consequence of $(E_1) \sim (E_2)$. (Sect. 3)
- As an application of Theorem 3, problems (QP1EQC) as well as (GTRS) are completely solved without any condition. (Sect. 4)
- As an application of Theorem 3, Beck's result [2] about the convexity of $S = \{(f(x), h_1(x), \dots, h_p(x)) : x \in \mathbb{R}^n\}$ with f being strictly convex and h_i 's being affine can be now generalized to any quadratic function f in Theorem 9. (Sect. 5)

- The most general version of the “regularized S-lemma with equality” is given in Corollary 1 without any condition. (Sect. 3)

Throughout the paper, we always assume that $\{x \in \mathbb{R}^n : h(x) = 0\} \neq \emptyset$. Notation $A \succeq (\preceq) B$ denotes that the matrix $A - B$ is positive (negative) semidefinite. $A \succ (\prec) B$ means that the matrix $A - B$ is positive (negative) definite. S_+^n represents the space of all $n \times n$ positive semidefinite symmetric matrices. $A \bullet B = \text{Tr}(AB^T) = \sum_{i,j=1}^n a_{ij}b_{ij}$ stands for the standard inner product of two symmetric matrices A, B . The null and range space of B is denoted by $\mathcal{N}(B)$ and $\mathcal{R}(B)$, respectively; and B^+ is the Moore-Penrose generalized inverse of B . Denote by I_n the identity matrix of dimension n ; and by $\text{Diag}(a)$ the diagonal matrix with a being its diagonal vector. The notation $v(\cdot)$ denotes the optimal value of a particularly mentioned optimization problem (\cdot) .

2 The S-lemma with equality when Assumption 1 fails

We observe that Assumption 1 is violated and $\{x \in \mathbb{R}^n : h(x) = 0\} \neq \emptyset$ if and only if

$$\min_x h(x) = 0 \text{ or } \max_x h(x) = 0.$$

Namely, h is convex(concave) and the set $\{x : h(x) = 0\}$ consists of all the minimizers (maximizers) of $h(x)$ such that

$$\{x : h(x) = 0\} = \{x : \min_x (\max_x h(x)) = 0\} = \{-B^+b + Zy : y \in \mathbb{R}^m\} \quad (6)$$

where $Z \in \mathbb{R}^{n \times m}$ is a matrix basis of $\mathcal{N}(B)$, and

$$\begin{aligned} \min_x (\max_x h(x)) &= (-B^+b + Zy)^T B (-B^+b + Zy) + 2b^T (-B^+b + Zy) + d \\ &= -b^T B^+b + d = 0. \end{aligned}$$

We then investigate $(E_1) \sim (E_2)$ under the following property:

Proposition 1 *Assumption 1 is violated if and only if*

$$B \succeq (\preceq) 0, \quad b \in \mathcal{R}(B) \text{ and } -b^T B^+b + d = 0. \quad (7)$$

We first discuss the special case that both f and h are homogeneous (i.e., $a = c = b = d = 0$), while the non-homogeneous case will be proved using the homogeneous result.

Suppose h is homogeneous. Then, (7) is reduced to $B \succeq (\preceq) 0$ and (6) becomes $h(x) = 0 \iff x = Zy$. Therefore,

$$(E_1) \iff f(Zy) \geq 0, \quad \forall y.$$

Suppose f is also homogeneous. Then,

$$(\text{homogeneous}) \quad (E_1) \iff Z^T A Z \succeq 0. \quad (8)$$

On the other hand, when f and h are quadratic forms,

$$(\text{homogeneous}) \quad (E_2) \iff (\exists \mu \in \mathbb{R}) \quad A + \mu B \succeq 0. \quad (9)$$

Therefore, suppose Assumption 1 is violated and both f and h are quadratic forms, homogeneous version of $(E_1) \sim (E_2)$ is the same as $(8) \sim (9)$.

Since (9) trivially implies (8), it is sufficient to show that (8) implies (9). When $B \succ 0$, $A + \mu B \succ 0$ for any sufficiently large μ , so (9) is trivially true. We let $B \succeq 0$ but not definite, i.e., $Z \neq 0$. Then there are two possibilities:

- (a) Suppose $Z^T A Z \succ 0$. For $x^T B x = 0$, we have $x = Zy$ for some $y \neq 0$ and $x^T A x = y^T Z^T A Z y > 0$. In other words, the system

$$x^T A x \leq 0, \quad x^T B x = 0, \quad x \neq 0,$$

has no solution. By Finsler's Theorem, (9) must be true.

- (b) Suppose $Z^T A Z \succeq 0$ but not definite. Anstreicher and Wright [1] proved in 2000 that (8) is equivalent to (9) if and only if $\mathcal{N}(Z^T A Z) = \mathcal{N}(Z^T A^2 Z)$.

In summary, if f, h are homogeneous and either $h(x) \geq 0$ or $h(x) \leq 0$, the S-lemma with equality holds for one of the following three situations: (i) $B \succ (\prec) 0$; (ii) $B \succeq (\preceq) 0$ and $Z^T A Z \succ 0$; and (iii) $B \succeq (\preceq) 0$, $Z^T A Z \succeq 0$ and $\mathcal{N}(Z^T A Z) = \mathcal{N}(Z^T A^2 Z)$. Notice that, in Introduction, we have seen $(E_1) \not\sim (E_2)$ for the example $f(x_1, x_2) = x_1 x_2$, $h(x_1, x_2) = -x_1^2$. The reason now is clear that $B \preceq 0$, $Z^T A Z = 0$, $\mathcal{N}(Z^T A Z) = \mathbb{R}$, but $\mathcal{N}(Z^T A^2 Z) = \{0\}$.

For nonhomogeneous f and h , we first have the following proposition.

Proposition 2 *Suppose Assumption 1 is violated and f, h are nonhomogeneous. Then,*

$$(E_1) \iff W = \tilde{Z}^T \tilde{A} \tilde{Z} = \begin{bmatrix} Z^T A Z & Z^T a - Z^T A B^+ b \\ a^T Z - b^T B^+ A Z & b^T B^+ A B^+ b - 2a^T B^+ b + c \end{bmatrix} \succeq 0, \quad (10)$$

where Z is a matrix basis of $\mathcal{N}(B)$, and

$$\tilde{Z} = \begin{bmatrix} Z & 0 \\ 0 & 1 \end{bmatrix}, \quad \tilde{A} = \begin{bmatrix} A & a - A B^+ b \\ a^T - b^T B^+ A^T & b^T B^+ A B^+ b - 2a^T B^+ b + c \end{bmatrix}. \quad (11)$$

Proof. Since f is non-homogeneous and Assumption 1 fails, we have

$$\begin{aligned} \inf_{h(x)=0} f(x) &= f(-B^+ b + Zy) \\ &= (-B^+ b + Zy)^T A (-B^+ b + Zy) + 2a^T (-B^+ b + Zy) + c \\ &= \begin{bmatrix} y \\ 1 \end{bmatrix}^T W \begin{bmatrix} y \\ 1 \end{bmatrix} \geq 0, \quad \forall y \in \mathbb{R}^m \end{aligned}$$

where W is defined in (10). Then, the matrix W is positive semi-definite since, for any $\gamma \neq 0$,

$$\begin{bmatrix} y \\ \gamma \end{bmatrix}^T W \begin{bmatrix} y \\ \gamma \end{bmatrix} = \gamma^2 \begin{bmatrix} y/\gamma \\ 1 \end{bmatrix}^T W \begin{bmatrix} y/\gamma \\ 1 \end{bmatrix} \geq 0, \quad \forall y \in \mathbb{R}^m;$$

and also for $\gamma = 0$,

$$\begin{bmatrix} y \\ 0 \end{bmatrix}^T W \begin{bmatrix} y \\ 0 \end{bmatrix} = \lim_{\gamma \rightarrow 0} \begin{bmatrix} y \\ \gamma \end{bmatrix}^T W \begin{bmatrix} y \\ \gamma \end{bmatrix} \geq 0, \quad \forall y \in \mathbb{R}^m.$$

□

Notice that (10) is the homogeneous representation for the nonhomogeneous inequality $f(-B^+b + Zy) \geq 0$ by lifting one more dimension to \tilde{A} . The next result is the main theorem of this section.

Theorem 1 *Suppose Assumption 1 is violated and the same notations as in Proposition 2 are adopted. Then, $(E_1) \sim (E_2)$ if and only if one of the following conditions is satisfied:*

- (a) $\tilde{Z}^T \tilde{A} \tilde{Z} \succ 0$;
- (b) $\tilde{Z}^T \tilde{A} \tilde{Z} \succeq 0$ and $\mathcal{N}(\tilde{Z}^T \tilde{A} \tilde{Z}) = \mathcal{N}(\tilde{Z}^T \tilde{A}^2 \tilde{Z})$.

Proof. Since Assumption 1 fails, by (7), we have $d = b^T B^+ b$. Then,

$$(\exists \mu \in \mathbb{R}) \quad f(x) + \mu h(x) \geq 0, \quad \forall x \in \mathbb{R}^n \iff \begin{bmatrix} A & a \\ a^T & c \end{bmatrix} + \mu \begin{bmatrix} B & b \\ b^T & b^T B^+ b \end{bmatrix} \succeq 0.$$

Using the invertible matrix $\begin{bmatrix} I & 0 \\ -b^T B^+ & 1 \end{bmatrix}$ to homogenize h , we obtain

$$\begin{aligned} (E_2) &\iff (\exists \mu \in \mathbb{R}) \quad \begin{bmatrix} A & a \\ a^T & c \end{bmatrix} + \mu \begin{bmatrix} B & b \\ b^T & b^T B^+ b \end{bmatrix} \succeq 0 \\ &\iff (\exists \mu \in \mathbb{R}) \quad \begin{bmatrix} I & 0 \\ -b^T B^+ & 1 \end{bmatrix} \begin{bmatrix} A & a \\ a^T & c \end{bmatrix} \begin{bmatrix} I & -B^+ b \\ 0 & 1 \end{bmatrix} + \mu \begin{bmatrix} B & 0 \\ 0 & 0 \end{bmatrix} \succeq 0 \\ &\iff (\exists \mu \in \mathbb{R}) \quad \tilde{A} + \mu \begin{bmatrix} B & 0 \\ 0 & 0 \end{bmatrix} \succeq 0 \end{aligned}$$

where \tilde{A} was defined in (11). Let $\tilde{B} = \begin{bmatrix} B & 0 \\ 0 & 0 \end{bmatrix} \succeq 0$. Notice that it is not definite and $\tilde{Z} = \begin{bmatrix} Z & 0 \\ 0 & 1 \end{bmatrix}$ the basis matrix for the null space of \tilde{B} . By applying the homogeneous version (8)~(9) to \tilde{A} and \tilde{B} , the conclusion of the theorem follows immediately. □

Recall that, the classical S-lemma $(S_1) \sim (S_2)$ relies on Slater's condition. When the condition is absent, the regularized S-lemma $(S_1) \sim (S_2^r)$ is a substitute for $(S_1) \sim (S_2)$ ([20]). Following the similar idea and as a direct consequence of Theorem 1, we can also formulate the following regularized version of S-lemma with equality in the absence of Assumption 1:

Theorem 2 *Suppose Assumption 1 is violated. Then, the following two statements are equivalent:*

$$(E_1) \quad (\forall x \in \mathbb{R}^n) \quad h(x) = 0 \implies f(x) \geq 0.$$

$$(E_2^r) \quad (\forall \epsilon > 0)(\exists \lambda_\epsilon)(\forall x \in \mathbb{R}^n) \quad f(x) + \lambda_\epsilon h(x) + \epsilon(x^T x + 1) \geq 0.$$

Proof. Since $(E_2^r) \implies (E_1)$ is trivial by letting $\epsilon \rightarrow 0$, it is sufficient to prove the converse. To this end, for all $\epsilon > 0$, consider $f_\epsilon(x) = f(x) + \epsilon(x^T x + 1) = x^T(A + \epsilon I)x + 2a^T x + (c + \epsilon)$ and $h(x)$. By (E_1) , we first have

$$(\forall \epsilon > 0)(\forall x \in \mathbb{R}^n) \quad h(x) = 0 \implies f(x) \geq 0 \implies f_\epsilon(x) \geq 0. \quad (12)$$

Secondly, from Proposition 2, (E_1) also implies that $W \succeq 0$ in (10). Then, due to $Z^T Z = I$ and $Z^T B^+ = 0$, there is

$$\begin{aligned} & \begin{bmatrix} Z^T(A + \epsilon I)Z & Z^T a - Z^T(A + \epsilon I)B^+ b \\ a^T Z - b^T B^+(A + \epsilon I)Z & b^T B^+(A + \epsilon I)B^+ b - 2a^T B^+ b + (c + \epsilon) \end{bmatrix} \\ &= \begin{bmatrix} Z^T A Z & Z^T a - Z^T A B^+ b \\ a^T Z - b^T B^+ A Z & b^T B^+ A B^+ b - 2a^T B^+ b + c \end{bmatrix} + \epsilon \begin{bmatrix} I & 0 \\ 0 & \|B^+ b\|^2 + 1 \end{bmatrix} \succ 0. \end{aligned}$$

In other words, Case (a) in Theorem 1 holds for $f_\epsilon(x)$ and $h(x)$. The S-lemma with equality for the pair $f_\epsilon(x)$ and $h(x)$ thus implies that, from (12), there must be an associated $\lambda_\epsilon \in \mathbb{R}$ for any $\epsilon > 0$ such that

$$f_\epsilon(x) + \lambda_\epsilon h(x) = f(x) + \lambda_\epsilon h(x) + \epsilon(x^T x + 1) \geq 0, \quad \forall x \in \mathbb{R}^n,$$

which proves the theorem. \square

Theorem 2 generalizes the regularized S-lemma. Consider $\hat{h}(x, z) = h(x) + z^2 = 0$. Assume that $h(x) \leq 0$ fails Slater's condition. Then, $\hat{h}(x, z) = h(x) + z^2 \geq 0, \forall x \in \mathbb{R}^n, \forall z \in \mathbb{R}$ fails Assumption 1. Since (S_1) can be equivalently rephrased as

$$(\forall x \in \mathbb{R}^n, \forall z \in \mathbb{R}) \quad \hat{h}(x, z) = h(x) + z^2 = 0 \implies \hat{f}(x, z) = f(x) \geq 0, \quad (13)$$

we can apply Theorem 2 to (\hat{f}, \hat{h}) and obtain

$$(\forall \epsilon > 0)(\exists \lambda_\epsilon)(\forall x \in \mathbb{R}^n, \forall z \in \mathbb{R}) \quad f(x) + \lambda_\epsilon(h(x) + z^2) + \epsilon(x^T x + 1) \geq 0. \quad (14)$$

Let $z \rightarrow \infty$ in (14), we have $\lambda_\epsilon \geq 0$ for all $\epsilon > 0$. The validity of (S_2^r) , and hence $(S_1) \sim (S_2^r)$, is concluded from setting $z = 0$ in (14).

The original proof of the regularized S-lemma in [20], based on Brickman's theorem [6] and Martínez-Legaz's result [23], is a bit tedious. Our argument by applying Theorem 2 can be viewed as a direct consequence of Finsler's Theorem [12], which is more direct and simple.

3 The S-lemma with equality when Assumption 1 holds

Throughout this section, we always assume that h takes both positive and negative values (Assumption 1). It is frequent to consider the homogenized version by introducing a new variable $t \in \mathbb{R}$ as follows:

$$\begin{aligned} \tilde{f}(x, t) &= x^T A x + 2ta^T x + ct^2, \\ \tilde{h}(x, t) &= x^T B x + 2tb^T x + dt^2. \end{aligned}$$

If $t \neq 0$, (E_2) implies that

$$(\exists \mu \in \mathbb{R})(\forall x \in \mathbb{R}^n) \tilde{f}(x, t) + \mu \tilde{h}(x, t) = t^2 \left(f\left(\frac{x}{t}\right) + \mu h\left(\frac{x}{t}\right) \right) \geq 0.$$

For $t = 0$, there is

$$(\exists \mu \in \mathbb{R})(\forall x \in \mathbb{R}^n) \tilde{f}(x, 0) + \mu \tilde{h}(x, 0) = \lim_{t \rightarrow 0} \tilde{f}(x, t) + \mu \tilde{h}(x, t) \geq 0.$$

Consequently, the validity of (E_2) implies that of its homogenized version

$$(\tilde{E}_2) \quad (\exists \mu \in \mathbb{R})(\forall x \in \mathbb{R}^n, \forall t \in \mathbb{R}) \tilde{f}(x, t) + \mu \tilde{h}(x, t) \geq 0,$$

and vice versa (by setting $t = 1$). So we have $(E_2) \sim (\tilde{E}_2)$. On the other hand, by the S-lemma with equality for a homogeneous quadratic system under Assumption 1 and S-Condition 3, $(\tilde{E}_2) \sim (\tilde{E}_1)$ below:

$$(\tilde{E}_1) \quad (\forall x \in \mathbb{R}^n, \forall t \in \mathbb{R}) \quad x^T Bx + 2tb^T x + dt^2 = 0 \implies x^T Ax + 2ta^T x + ct^2 \geq 0.$$

Comparing (\tilde{E}_1) with (E_1) , we only know they are equivalent when $t \neq 0$, by rewriting

$$(\forall x \in \mathbb{R}^n, t \neq 0) \quad x^T Bx + 2tb^T x + dt^2 = 0 \implies x^T Ax + 2ta^T x + ct^2 \geq 0 \quad (15)$$

as

$$(\forall x \in \mathbb{R}^n) \quad \left(\frac{x}{t}\right)^T B \left(\frac{x}{t}\right) + 2b^T \left(\frac{x}{t}\right) + d = 0 \implies \left(\frac{x}{t}\right)^T A \left(\frac{x}{t}\right) + 2a^T \left(\frac{x}{t}\right) + c \geq 0.$$

Therefore, if we are to argue that $(E_1) \sim (\tilde{E}_1)$, it amounts to finding conditions under which $(E_1) \implies (\tilde{E}_1)$ for $t = 0$. That is, we need conditions for the following compound statement to hold:

$$\left\{ (E_1) : \inf_{h(x)=0} f(x) \geq 0 \right\} \implies \left\{ (\forall x \in \mathbb{R}^n) \quad x^T Bx = 0 \implies x^T Ax \geq 0 \right\}. \quad (16)$$

We now present and prove the necessary and sufficient conditions for the S-lemma with equality to be valid.

Theorem 3 *Under Assumption 1 that $h(x)$ takes both positive and negative values, the S-lemma with equality holds except that A has exactly one negative eigenvalue, $B = 0$, $b \neq 0$ and*

$$\begin{bmatrix} V^T A V & V^T (Ax_0 + a) \\ (x_0^T A + a^T) V & f(x_0) \end{bmatrix} \succeq 0, \quad (17)$$

where $x_0 = -\frac{d}{2b^T b}b$, $V \in \mathbb{R}^{n \times (n-1)}$ is the matrix basis of $\mathcal{N}(b)$.

Proof. The statement of the theorem is the same as $(E_1) \not\sim (E_2)$ if, and only if $B = 0$, $b \neq 0$, A has exactly one negative eigenvalue and (17) holds. The proof is organized in the following sequence:

(Sufficiency) Suppose $B = 0$, $b \neq 0$, A has exactly one negative eigenvalue and (17) holds. We show that (16) is a false statement. That is, $(E_1) \not\sim (E_2)$.

(Necessity) Suppose (16) is a false statement such that (E_1) is true but

$$(\forall x \in \mathbb{R}^n) \quad x^T B x = 0 \implies x^T A x \geq 0 \quad (18)$$

fails. Through a case-by-case analysis (cases (a), (b), (c-1), (c-2) and (c-3) below), we show that there must be $B = 0$. Once this is obtained, it follows immediately that $b \neq 0$, (17) holds, and A has exactly one negative eigenvalue.

(Proof for sufficiency) From $B = 0$, $b \neq 0$, we observe that the set $\{x \in \mathbb{R}^n : h(x) = 0\}$ is a linear variety of $n - 1$ dimension

$$\{x \in \mathbb{R}^n : h(x) = 0\} = \{x_0 + Vy : y \in \mathbb{R}^{n-1}\} \quad (19)$$

with $x_0 = -\frac{d}{2b^T b}b$ being a particular solution of $h(x) = 0$ and V is a matrix basis of $\mathcal{N}(b)$. Then, (E_1) becomes an unconstrained minimization problem

$$\inf_{y \in \mathbb{R}^{n-1}} \{f(x_0 + Vy) = f(x_0) + 2(x_0^T A + a^T)Vy + y^T V^T A V y\} \geq 0 \quad (20)$$

where “ ≥ 0 ” comes from (17). That is, (E_1) is true. However, from $B = 0$ and $A \not\geq 0$, we know (18) is wrong so that $(E_1) \not\sim (E_2)$ and the sufficiency is proved.

(Proof for necessity) We defer the proof for $B = 0$ to later discussions but first assume that it is true. We show that, when $B = 0$, (E_1) is true and (18) fails, it follows immediately that $b \neq 0$, (17) holds, and A has exactly one negative eigenvalue.

Since $h(x)$ takes both positive and negative values, $B = 0$ implies that $b \neq 0$ and then the set $\{x \in \mathbb{R}^n : h(x) = 0\}$ is the $n - 1$ dimensional linear variety (19). Since (E_1) is true, we obtain (20), which gives (17) and $V^T A V \succeq 0$. As $B = 0$ and (18) fails, we have $A \not\geq 0$. Then, the matrix A must have exactly one negative eigenvalue.

In the remaining part of the necessity proof, we show $B = 0$ by contradiction. That is, we are going to argue that “if $B \neq 0$ and (18) fails, then $\inf_{h(x)=0} f(x) = -\infty$ (so (E_1) fails too).”

Since (18) fails, there is a $v \in \mathbb{R}^n$ such that

$$v^T B v = 0, \quad v^T A v < 0. \quad (21)$$

This v is to be utilized to construct a curve $\{y + \alpha(y)v : y \in \mathbb{R}^n, \alpha(y) \in \mathbb{R}\}$ on $\{x : h(x) = 0\}$ upon which f is unbounded below.

For this purpose, we derive the following formulae.

$$\begin{aligned} h(y + \alpha(y)v) &= (y + \alpha(y)v)^T B (y + \alpha(y)v) + 2b^T (y + \alpha(y)v) + d \\ &= h(y) + 2y^T (Bv)\alpha(y) + 2(b^T v)\alpha(y); \end{aligned} \quad (22)$$

and

$$\begin{aligned} f(y + \alpha(y)v) &= (y + \alpha(y)v)^T A(y + \alpha(y)v) + 2a^T(y + \alpha(y)v) + c \\ &= f(y) + 2y^T(Av)\alpha(y) + 2(a^T v)\alpha(y) + (v^T Av)(\alpha(y))^2. \end{aligned} \quad (23)$$

From (22), we can construct $\{y + \alpha(y)v : y \in \mathbb{R}^n, \alpha(y) \in \mathbb{R}\}$ on $\{x : h(x) = 0\}$ in different ways based on three cases: Case (a) for $b^T v = 0$; Case (b) for $b^T v \neq 0, Bv = 0$ and Case (c) for $b^T v \neq 0, Bv \neq 0$. For convenience, we may assume $d = 0$, i.e., $h(0) = 0$. If this is not the case, we choose a nonzero vector $x' \in \mathbb{R}^n$ such that $h(x') = 0$ and then translate the origin there.

- (a) Suppose $b^T v = 0$. In this case, we choose $y = 0$ and let $\alpha(y)$ be any real number α in (22). Due to $h(0) = 0$, it is easy to see that $h(\alpha v) = 0$ for all α . Therefore, the set $\{x : h(x) = 0\}$ contains a straight line $\{\alpha v : \alpha \in \mathbb{R}\}$ which goes through the origin in the direction of v . The values of f on this line can be easily read from (23) to have

$$\inf_{\alpha} \{f(\alpha v) = c + 2(a^T v)\alpha + (v^T Av)\alpha^2\} = -\infty$$

where $v^T Av < 0$ due to (21). Therefore, we have $\inf_{h(x)=0} f(x) = -\infty$.

- (b) Suppose $Bv = 0$ and $b^T v \neq 0$. For any $y \in \mathbb{R}^n$, let

$$\alpha(y) = -\frac{h(y)}{2b^T v}.$$

Then, $y + \alpha(y)v$ satisfies $h(y + \alpha(y)v) = h(y) + 2(b^T v)\alpha(y) = 0$. Since $B \neq 0$, there is an eigenvector $u \neq 0$ corresponding to a nonzero eigenvalue σ of B , i.e., $Bu = \sigma u$. Consider y to be the line $\{\gamma u : \gamma \in \mathbb{R}\}$ spanned by u . Then, $\Gamma = \{\gamma u + \alpha(\gamma u)v : \gamma \in \mathbb{R}\}$ is a curve on $\{x : h(x) = 0\}$. The values of f on Γ can be verified to be unbounded below by

$$\begin{aligned} &\inf_{\gamma \in \mathbb{R}} \{f(\gamma u + \alpha(\gamma u)v)\} \\ &= \inf_{\gamma \in \mathbb{R}} \left\{ f(\gamma u) - 2(\gamma Au + a)^T v \frac{h(\gamma u)}{2b^T v} + v^T Av \frac{h^2(\gamma u)}{(2b^T v)^2} \right\} \\ &= -\infty \end{aligned}$$

since the coefficient $\frac{v^T Av}{(2b^T v)^2} < 0$ and $h^2(\gamma u) = (\sigma\gamma^2\|u\|^2 + 2\gamma b^T u)^2$ is a polynomial of degree 4 in γ whereas $f(\gamma u)$ is only of degree 2 and $\gamma h(\gamma u)$ is of degree 3.

- (c) Suppose $Bv \neq 0$ and $b^T v \neq 0$. In this case, B is indefinite; otherwise $B \succeq (\preceq) 0$, $v^T Bv = 0$ would imply that $Bv = 0$. Without loss of generality, assume that B is diagonal after performing the eigenvalue decomposition on B . We also assume that $B = \text{Diag}(B_{ii})$ where

$$B_{ii} = \begin{cases} 1, & i \in I, \\ -1, & i \in J, \\ 0, & i \in \{1, 2, \dots, n\} \setminus (I \cup J). \end{cases}$$

It will soon be clear that only the signs of the entries matter. Since B is indefinite, both I and J are non-empty, i.e., $\#I \geq 1$ and $\#J \geq 1$. It follows that the rank of B is at least 2. When $\text{Rank}(B) = 2$, the homogeneous quadratic surface $x^T Bx = 0$ is the union of two vertical-like hyperplanes (a type of cylindroid in geometry), which will be dealt with separately. When $\text{Rank}(B) = 3$, $x^T Bx = 0$ is a second order cone (a double circular cone). When $\text{Rank}(B) > 3$, it is sure that $x^T Bx = 0$ is not the union of hyperplanes since there is at least one three-dimensional second order cone embedded as a cross section.

- (c-1) Suppose $\text{Rank}(B) \geq 3$. We assume that $I = \{i_1, i_2, \dots, i_m\}$ and $J = \{j_1, j_2, \dots, j_k\}$ with $m \geq 2$, $k \geq 1$. We first claim that it is always possible to choose one v such that $v^T Bv = 0$, $v^T Av < 0$ and $v_i^2 \neq v_j^2$ for some $i \in I$ and $j \in J$. If the vector v satisfying (21) does not meet the requirement, namely $v_i^2 = v_j^2$ for all $i \in I$ and $j \in J$, we can perturb v by ϵw where $\epsilon > 0$ is a sufficiently small constant; and

$$w_i = \begin{cases} 1, & i \in \{i_2, j_1\} \\ 0, & \text{o.w.} \end{cases}$$

Then, $Bw = 0$; $(v + \epsilon w)^T B(v + \epsilon w) = 0$; $(v + \epsilon w)^T A(v + \epsilon w) < 0$; and $(v_{i_1} + \epsilon w_{i_1})^2 = v_{i_1}^2 \neq (v_{j_1} + \epsilon)^2 = (v_{j_1} + \epsilon w_{j_1})^2$. So we can assume that $v^T Bv = 0$, $v^T Av < 0$, $Bv \neq 0$, $b^T v \neq 0$ and $v_{i_1}^2 \neq v_{j_1}^2$.

Define a straight line

$$x_\beta = \beta(v_{j_1})e_1 + \beta(v_{i_1})e_2, \quad \forall \beta \in \mathbb{R}. \quad (24)$$

Since $x_\beta^T Bv = 0$, $\forall \beta \in \mathbb{R}$, the line x_β and the vector v are conjugate with respect to B . By defining

$$\alpha(x_\beta) = -\frac{h(x_\beta)}{2b^T v},$$

and by the conjugacy, we see from (22) that

$$h(x_\beta + \alpha(x_\beta)v) = h(x_\beta) + 2(b^T v)\alpha(x_\beta) = 0$$

and $x_\beta + \alpha v$ is a curve on $h(x) = 0$ upon which $f(x)$ is unbounded below as

$$\begin{aligned} & \inf_{h(x_\beta + \alpha v) = 0} \{f(x_\beta + \alpha v) = f(x_\beta) + 2(Ax_\beta + a)^T v\alpha + v^T A v \alpha^2\} \\ &= \inf_{\beta \in \mathbb{R}} \left\{ f(x_\beta) - 2(Ax_\beta + a)^T v \frac{h(x_\beta)}{2b^T v} + v^T A v \frac{h^2(x_\beta)}{(2b^T v)^2} \right\} \\ &= -\infty, \end{aligned}$$

where the coefficient $\frac{v^T A v}{(2b^T v)^2} < 0$ and due to $(v_{j_1})^2 - (v_{i_1})^2 \neq 0$,

$$h^2(x_\beta) = (\beta^2(v_{j_1})^2 - \beta^2(v_{i_1})^2 + 2\beta b_{i_1} v_{j_1} + 2\beta b_{j_1} v_{i_1})^2$$

is a polynomial of degree 4 in β whereas $f(x_\beta)$ is only of degree 2 and $\beta h(x_\beta)$ is of degree 3. Finally, we remark that when the diagonal elements of B are not just 0, 1, -1, the line x_β defined in (24) can be adjusted through the linear combination of e_1 and e_2 to maintain the conjugacy to v and the rest of the proof follows immediately.

- (c-2) Suppose $\text{Rank}(B) = 2$ with $B_{11} = 1$, $B_{22} = -1$ and $B_{ii} = 0$, for $i \geq 3$. In this subcase (c-2), we handle $\tilde{b} = (b_3, \dots, b_n)^T \neq 0$ whereas leaving $\tilde{b} = 0$ to (c-3) next. Since $x^T B x = 0$ is the union of two cylindroid hyperplanes and $\tilde{b} \neq 0$, we show that there is an oblique cross section of $h(x) = 0$, which contains a straight line in the direction of v for any v satisfying (21) and $Bv \neq 0$. Since it must be $v_1^2 = v_2^2 \neq 0$, the line has a formula

$$l(t) = \begin{pmatrix} 0 \\ y \\ \tilde{z} \end{pmatrix} + tv, \quad t \in \mathbb{R} \quad \text{and} \quad h(l(t)) = 0$$

where

$$y = \frac{b^T v}{v_2}; \quad \tilde{z} = \frac{(\frac{b^T v}{v_2})^2 - \frac{2b_2(b^T v)}{v_2}}{2\tilde{b}^T \tilde{b}} \tilde{b}.$$

Denoting $x_0^T = (0, y, \tilde{z}^T)$ and substituting $l(t)$ in $f(x)$ yields

$$f(x_0 + tv) = (v^T A v)t^2 + 2(x_0^T A v + a^T v)t + f(x_0),$$

which tends to $-\infty$ as $|t| \rightarrow \infty$ due to $v^T A v < 0$.

- (c-3) Suppose that $\text{Rank}(B) = 2$ with $B_{11} = 1$, $B_{22} = -1$, $B_{ii} = 0$, for $i \geq 3$ and $\tilde{b} = 0$. Then, the optimization problem $\inf_{h(x)=0} f(x)$ is unconstrained in the last $n - 2$ variables. Denote

$$A = \begin{bmatrix} A_1 & A_2 \\ A_2^T & A_3 \end{bmatrix}$$

where $A_1 \in \mathbb{R}^{2 \times 2}$. Let $A_3 = U^T \Sigma U$ be the eigenvalue decomposition with U orthogonal and $\Sigma = \text{Diag}(\sigma_i)$. The Moore-Penrose generalized inverse of A_3 is $A_3^+ = U^T \Sigma^+ U$, where $\Sigma^+ = \text{Diag}(\sigma_i^{-1})$ when $\sigma_i \neq 0$ and $0^{-1} = 0$. Define

$$W = \begin{bmatrix} I_2 & -A_2 A_3^+ \\ 0 & U \end{bmatrix}.$$

Then we have

$$W \begin{bmatrix} A_1 & A_2 \\ A_2^T & A_3 \end{bmatrix} W^T = \begin{bmatrix} \hat{A}_1 & 0 \\ 0 & \Sigma \end{bmatrix}, \quad W B W^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0_{n-2} \end{bmatrix} = B$$

where $\hat{A}_1 = A_1 - A_2 A_3^+ A_2^T$. Introducing the coordinate change

$$\begin{pmatrix} z \\ y \end{pmatrix} = W^{-T} x; \quad \begin{pmatrix} \hat{a}_z \\ \hat{a}_y \end{pmatrix} = W a; \quad \hat{b} = \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \end{pmatrix} = W b$$

where $z \in \mathbb{R}^2$, $y \in \mathbb{R}^{n-2}$ yields

$$\begin{aligned} \inf_{h(x)=0} f(x) &= \inf z^T \hat{A}_1 z + 2\hat{a}_z^T z + y^T \Sigma y + 2\hat{a}_y^T y + c \\ \text{s.t. } z^T \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} z + 2\hat{b}^T z &= 0, \quad z \in \mathbb{R}^2, \quad y \in \mathbb{R}^{n-2}. \end{aligned} \quad (25)$$

Obviously, due to $B_{ii} = 0$, $i \geq 3$ and $\tilde{b} = (b_3, b_4, \dots, b_n) = 0$, the variable y is unconstrained. If there is some $\sigma_i < 0$ (in which case $A_3 \not\subseteq 0$); or some $\sigma_i = 0$ but $(\hat{a}_y)_i \neq 0$ (in which case at least one of the two columns of A_2^T is not in the range of A_3), the problem $\inf_{h(x)=0} f(x)$ is surely $-\infty$. Therefore, we only have to concentrate on the case that, for all $i \in \{3, 4, \dots, n\}$, either $\sigma_i > 0$ or $\sigma_i = (\hat{a}_y)_i = 0$. That is, the function

$$y^T \Sigma y + 2\hat{a}_y^T y + c$$

is a convex sum-of-squares quadratic with no pure linear terms, which has a global optimal solution $y^* = -\Sigma^+ \hat{a}_y$. Substituting y^* into (25), it reduces to the following quadratic optimization problem with two variables:

$$\inf_{h(x)=0} f(x) = \inf_{z \in \mathbb{R}^2} z^T \hat{A}_1 z + 2\hat{a}_z^T z - \hat{a}_y^T \Sigma^+ \hat{a}_y + c \quad (26)$$

$$\text{s.t. } z^T \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} z + 2\hat{b}^T z = 0. \quad (27)$$

Notice that the equation (27) can be used to solve either z_1 or z_2 . To see this, we first suppose that $|\hat{b}_1| \geq |\hat{b}_2|$. By introducing

$$\delta = \sqrt{\hat{b}_1^2 - \hat{b}_2^2} - \hat{b}_1, \quad \hat{z}_1 = z_1 - \delta, \quad \hat{z}_2 = z_2 - \hat{b}_2,$$

we obtain a new expression for (27):

$$z_1^2 - z_2^2 + 2\hat{b}_1 z_1 + 2\hat{b}_2 z_2 = \hat{z}_1^2 - \hat{z}_2^2 + 2(\hat{b}_1 + \delta)\hat{z}_1 \quad (28)$$

in which there is no linear term of \hat{z}_2 . Similarly for $|\hat{b}_1| < |\hat{b}_2|$, we can eliminate the linear term of \hat{z}_1 by letting

$$\delta = \sqrt{\hat{b}_2^2 - \hat{b}_1^2} + \hat{b}_2, \quad \hat{z}_1 = z_1 + \hat{b}_1, \quad \hat{z}_2 = z_2 - \delta. \quad (29)$$

We also notice that the transformation (28) or (29) only affect the linear term \hat{a}_z of (26) but not the second order term \hat{A}_1 . Consequently, we may assume $|\hat{b}_1| \geq |\hat{b}_2| = 0$ so that (27) can be solved as

$$z_2 = \pm \sqrt{z_1^2 + 2\hat{b}_1 z_1}. \quad (30)$$

Denote $\hat{A}_1 = \begin{bmatrix} \hat{a}_{11} & \hat{a}_{12} \\ \hat{a}_{12} & \hat{a}_{22} \end{bmatrix}$ and substitute (30) into (26). It becomes

$$\begin{aligned} \inf_{h(x)=0} f(x) &\leq \inf_{|z_1| \gg 1} \left\{ \hat{a}_{11} z_1^2 \pm 2\hat{a}_{12} z_1 \sqrt{z_1^2 + 2\hat{b}_1 z_1 + \hat{a}_{22}(z_1^2 + 2\hat{b}_1 z_1)} \right. \\ &\quad \left. + 2(\hat{a}_z)_1 z_1 \pm 2(\hat{a}_z)_2 \sqrt{z_1^2 + 2\hat{b}_1 z_1} - \hat{a}_y^T \Sigma^+ \hat{a}_y + c \right\} \\ &= \inf_{|z_1| \gg 1} \left\{ (\hat{a}_{11} - 2|\hat{a}_{12}| + \hat{a}_{22}) z_1^2 + O(z_1) \right\} \end{aligned}$$

It remains to show, due to $v^T B v = 0$, $v^T A v < 0$, $B v \neq 0$ and that we are in the case $\Sigma \succeq 0$, the leading term $\hat{a}_{11} - 2|\hat{a}_{12}| + \hat{a}_{22} < 0$. We first perform the coordinate change $\hat{v} = W^{-T} v$ to give

$$v^T B v = 0 \implies \hat{v}^T \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0_{n-2} \end{bmatrix} \hat{v} = 0 \implies \hat{v}_1^2 = \hat{v}_2^2; \quad (31)$$

$$v^T A v < 0 \implies \hat{v}^T \begin{bmatrix} \hat{A}_1 & 0 \\ 0 & \Sigma \end{bmatrix} \hat{v} < 0 \implies \begin{bmatrix} \hat{v}_1 \\ \hat{v}_2 \end{bmatrix}^T \hat{A}_1 \begin{bmatrix} \hat{v}_1 \\ \hat{v}_2 \end{bmatrix} < 0. \quad (32)$$

Moreover, $B v \neq 0$ implies that $B W^T \hat{v} = B \hat{v} \neq 0$, i.e., either $\hat{v}_1 \neq 0$ or $\hat{v}_2 \neq 0$. From (31), it must be $\hat{v}_1 = \hat{v}_2 \neq 0$ or $\hat{v}_1 = -\hat{v}_2 \neq 0$. It follows from (32) that one of the following equations holds:

$$\begin{aligned} \hat{a}_{11} + \hat{a}_{22} + 2\hat{a}_{12} &< 0, \quad \text{for } \hat{v}_1 = \hat{v}_2; \\ \hat{a}_{11} + \hat{a}_{22} - 2\hat{a}_{12} &< 0, \quad \text{for } \hat{v}_1 = -\hat{v}_2. \end{aligned}$$

Combining together, we have $\hat{a}_{11} + \hat{a}_{22} < 2|\hat{a}_{12}|$. The final subcase and the entire proof of the theorem is thus complete. \square

Theorem 3 generalizes the classical S-lemma $(S_1) \sim (S_2)$ under Slater's condition. First, like (13), we rewrite (S_1) in terms of $\hat{h}(x, z) = h(x) + z^2$ and $\hat{f}(x, z) = f(x)$. Let \bar{x} be a Slater point of $h(x) \leq 0$. Then, Assumption 1 can be easily checked to hold by

$$\hat{h}(\bar{x}, 0) = h(\bar{x}) < 0, \quad \hat{h}(\bar{x}, 1 - h(\bar{x})) = h(\bar{x}) + (1 - h(\bar{x}))^2 > 0.$$

Moreover, $\hat{h}(x, z)$ has a non-zero Hessian that $\nabla^2 \hat{h} \neq 0$. By Theorem 3,

$$(\exists \mu \in \mathbb{R})(\forall x \in \mathbb{R}^n, \forall z \in \mathbb{R}) \quad f(x) + \mu \hat{h}(x, z) = f(x) + \mu h(x) + \mu z^2 \geq 0. \quad (33)$$

Then, $\mu \geq 0$ follows from $z \rightarrow \infty$ and setting $z = 0$ in (33) yields (S_2) . Therefore, under Slater's condition, (S_1) implies (S_2) by Theorem 3.

Finally, when Assumption 1 holds but $B = 0$, the S-lemma with equality could possibly fail. In that case, we can formulate a modified version of regularized S-lemma.

Theorem 4 *Suppose that Assumption 1 holds but $B = 0$. The statement*

$$h(x) = 0 \implies f(x) \geq 0, \forall x \in \mathbb{R}^n \quad (34)$$

is true if and only if

$$(\forall \epsilon > 0)(\exists \lambda_\epsilon)(\forall x \in \mathbb{R}^n) f(x) + \lambda_\epsilon(h(x))^2 + \epsilon(x^T x + 1) \geq 0, \quad (35)$$

Proof. It is sufficient to prove (34) implies (35). First, (34) implies that

$$(h(x))^2 = 0 \implies f(x) \geq 0, \forall x \in \mathbb{R}^n.$$

Since Assumption 1 cannot hold for h^2 , we apply Theorem 2 to (f, h^2) and complete the proof. \square

Theorems 2, 3 and 4 can be packed together to yield the following regularized S-lemma with equality.

Corollary 1 *The statement*

$$h(x) = 0 \implies f(x) \geq 0, \forall x \in \mathbb{R}^n$$

is true if and only if

$$(\forall \epsilon > 0)(\exists \lambda_\epsilon)(\forall x \in \mathbb{R}^n) f(x) + \lambda_\epsilon(h(x))^{\phi(B)} + \epsilon(x^T x + 1) \geq 0,$$

where

$$\phi(B) = \begin{cases} 1, & \text{if } B \neq 0, \\ 2, & \text{if } B = 0. \end{cases}$$

4 Application to solve (QP1EQC)

Moré in 1993 published an early result on the saddle point optimality condition for (QP1EQC) under mild conditions [24]. It states:

Theorem 5 ([24], Thm 3.2) *Under Assumption 1 and $B \neq 0$, a vector x^* is a global minimizer of (QP1EQC) if and only if $h(x^*) = 0$ and there is a multiplier $\mu^* \in \mathbb{R}$ such that the Kuhn-Tucker condition*

$$Ax^* + a + \mu^*(Bx^* + b) = 0 \quad (36)$$

*is satisfied with the second order condition $A + \mu^*B \succeq 0$.*

However, (36) may not always have a pair of solution (x^*, λ^*) since (QP1EQC) could be unbounded below or have an unattainable optimal value. Even when the problem has an attainable minimum, algorithmic approach for computing (x^*, λ^*) from (36) often required strong conditions such as the existence of a positive definite matrix pencil $A + \mu B \succ 0$ (also known as the dual Slater condition), e.g. [24, 29, 31, 38]. The dual Slater condition is not practical since it is stricter than the two matrices A and B being simultaneously diagonalizable via congruence (SDC) [16]. By the S-lemma with equality, we can now solve

(QP1EQC) directly by the standard SDP relaxation (because it is tight) and a rank-one decomposition procedure (if necessary) without resorting to the Kuhn-Tucker condition (36) and without any assumption. We also analyze (QP1EQC) when it is unbounded below; or is unattainable.

In fact, when $B = 0$, the constraint is just $2b^T x + d = 0$. When Assumption 1 fails, there must be $B \succeq (\preceq) 0$ so that the constraint is reduced to the first order condition $Bx + b = 0$. By the null space representation of $2b^T x + d = 0$ or $Bx + b = 0$, (QP1EQC) becomes an unconstrained problem. It would then be either unbounded below or a convex unconstrained problem with an attainable optimal solution.

Now it remains to consider (QP1EQC) for $B \neq 0$ and under Assumption 1. Applying the S-lemma with equality (Theorem 3), we can recast (QP1EQC) as the following semidefinite programming problems (SDP):

$$v(\text{QP1EQC}) = \sup_{s \in \mathbb{R}} \{s : \{x \in \mathbb{R}^n | f(x) - s < 0, h(x) = 0\} = \emptyset\} \quad (37)$$

$$= \sup_{s \in \mathbb{R}} \{s : (\exists \mu \in \mathbb{R}) f(x) - s + \mu h(x) \geq 0, \forall x \in \mathbb{R}^n\} \quad (38)$$

$$= \sup_{s, \mu \in \mathbb{R}} \left\{ s : \begin{bmatrix} A + \mu B & a + \mu b \\ a^T + \mu b^T & c + \mu d - s \end{bmatrix} \succeq 0 \right\} \quad (39)$$

$$\leq \inf_{X \in S_+^n} \left\{ \begin{bmatrix} A & a \\ a^T & c \end{bmatrix} \bullet X : \begin{bmatrix} B & b \\ b^T & d \end{bmatrix} \bullet X = 0, X_{n+1, n+1} = 1 \right\} \quad (40)$$

$$\leq \inf_{x \in \mathbb{R}^n} \left\{ \begin{bmatrix} A & a \\ a^T & c \end{bmatrix} \bullet \begin{pmatrix} x \\ 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}^T : \begin{bmatrix} B & b \\ b^T & d \end{bmatrix} \bullet \begin{pmatrix} x \\ 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}^T = 0 \right\} \quad (41)$$

$$= v(\text{QP1EQC}).$$

Note that (38) is equivalent to the Lagrangian dual of (QP1EQC)

$$(\text{LD}) \quad \sup_{\mu \in \mathbb{R}} \left\{ \inf_{x \in \mathbb{R}^n} L(x, \mu) := f(x) + \mu h(x) \right\}. \quad (42)$$

The equation (39) is the SDP reformulation of (42) which is known as Shor relaxation scheme [30]. The inequality (40) follows from the conic weak duality. Eventually, all inequalities above become equalities and they prove the strong duality (no duality gap between (QP1EQC) and its Lagrange dual), as well as the tight SDP relaxation.

The strong duality result (37)-(41) does not rely on the existence of an optimal solution x^* to (QP1EQC) and thus it is more general than Theorem 5. It is possible that the strong duality holds like $v(\text{QP1EQC}) = v(\text{LD}) = -\infty$ for an unbounded (QP1EQC). When $v(\text{QP1EQC}) = -\infty$, the SDP reformulation (39) of the Lagrange dual (LD) is surely infeasible. The converse is also true. When $v(\text{QP1EQC}) > -\infty$, by the S-lemma with equality, there is some $\mu \in \mathbb{R}$ such that $f(x) - v(\text{QP1EQC}) + \mu h(x) \geq 0, \forall x \in \mathbb{R}^n$. The dual must be feasible.

Moreover, when $v(\text{QP1EQC}) > -\infty$, due to the tight SDP relaxation (40), an optimal solution x^* of (QP1EQC) can be found if, and only if the primal SDP relaxation (40) attains the optimal solution at some $X^* \in S_+^n$. Then we

can employ the standard rank-one decomposition procedure [32] to generate a rank-one solution out of X^* for (QP1EQC). Notice that the strong duality (37)-(41) does not warrant (QP1EQC) and its primal SDP relaxation (40) an attainable optimal solution though.

We will show in Theorem 6 that, when $v(\text{QP1EQC}) = v(\text{LD}) > -\infty$, the dual SDP (39) is not only feasible but the value $v(\text{LD})$ is always attainable. Moreover, the primal problem (QP1EQC) is unattainable if and only if its dual feasible set is a single point set $\{(v(\text{QP1EQC}), \mu^*)\}$ at which either (43) or (44) happens. We first prove a lemma for the dual attainment property.

Lemma 1 *Under Assumption 1 and $B \neq 0$, if (QP1EQC) has an optimal solution x^* , then the dual SDP (39) also has an optimal solution (s^*, μ^*) such that the primal-dual pair (x^*, μ^*) satisfies the Kuhn-Tucker condition (36).*

Proof. Let x^* be an optimal solution of (QP1EQC). According to Theorem 5, there is a μ^* such that $A + \mu^*B \succeq 0$, $a + \mu^*b \in \mathcal{R}(A + \mu^*B)$ and thus

$$x^* = \arg \min f(x) + \mu^*h(x).$$

Let $s^* = f(x^*) + \mu^*h(x^*) = f(x^*)$. Since $s^* \leq f(x) + \mu^*h(x)$, $\forall x \in \mathbb{R}^n$, the pair (s^*, μ^*) is dual feasible to (39) (or to (38)). Suppose (s, μ) is any dual feasible pair such that $s \leq f(x) + \mu h(x)$, $\forall x \in \mathbb{R}^n$. There must also be $A + \mu B \succeq 0$, $a + \mu b \in \mathcal{R}(A + \mu B)$ and $x(\mu) = -(A + \mu B)^+(a + \mu b)$ such that

$$\begin{aligned} s &\leq \inf\{f(x) + \mu h(x) : x \in \mathbb{R}^n\} \\ &= f(x(\mu)) + \mu h(x(\mu)) \\ &\leq f(x^*) + \mu h(x^*) \\ &= f(x^*) \\ &= s^*. \end{aligned}$$

Therefore, the dual SDP (39) has an optimal solution (s^*, μ^*) and the primal-dual pair (x^*, μ^*) satisfies the Kuhn-Tucker condition (36). \square

Theorem 6 *Under Assumption 1, $B \neq 0$ and $v(\text{QP1EQC}) > -\infty$, the dual SDP (39) always has an optimal solution (s^*, μ^*) . Moreover, the infimum of (QP1EQC) is unattainable when, and only when the dual SDP (39) possesses a unique feasible μ^* ; and at μ^* the two functions $f(x)$, $h(x)$ together satisfy*

$$\text{either } V^T B V \succeq 0, h(y_0) - (B y_0 + b)^T V (V^T B V)^+ V^T (B y_0 + b) > 0, \quad (43)$$

$$\text{or } V^T B V \preceq 0, h(y_0) - (B y_0 + b)^T V (V^T B V)^+ V^T (B y_0 + b) < 0, \quad (44)$$

where $y_0 = -(A + \mu^*B)^+(a + \mu^*b)$, V is the matrix basis of $\mathcal{N}(A + \mu^*B)$.

Proof. Since $v(\text{QP1EQC}) > -\infty$, there is some $\mu \in \mathbb{R}$ such that $f(x) - v(\text{QP1EQC}) + \mu h(x) \geq 0$, $\forall x \in \mathbb{R}^n$. It is clear that such an μ satisfies

$$\begin{cases} A + \mu B \succeq 0, \\ a + \mu b \in \mathcal{R}(A + \mu B). \end{cases}$$

We first consider that the matrix pencil $I_{\succeq}(A, B) = \{\mu : A + \mu B \succeq 0\}$ is a single-point set $\{\mu^*\}$. Then, any dual feasible (s, μ^*) must satisfy

$$s \leq \inf\{f(x) + \mu^* h(x) : x \in \mathbb{R}^n\} = f(x(\mu^*)) + \mu^* h(x(\mu^*))$$

where $x(\mu^*) = -(A + \mu^* B)^+(a + \mu^* b)$. Then, (s^*, μ^*) with $s^* = f(x(\mu^*)) + \mu^* h(x(\mu^*))$ is the dual optimal solution. Obviously, $s^* = v(\text{QP1EQC})$. Moreover, when $I_{\succeq}(A, B) = \{\mu^*\}$, all the Kuhn-Tucker points of (36) can be completely specified by

$$x^*(y) = -(A + \mu^* B)^+(a + \mu^* b) + Vy = y_0 + Vy, \quad \forall y.$$

Observe that

$$h(x^*(y)) = h(y_0 + Vy) = y^T(V^T BV)y + 2(y_0^T BV + b^T V)y + h(y_0).$$

In case of (43), $h(x)$ restricted on the set of Kuhn-Tucker points is convex and

$$\min_y h(x^*(y)) = h(y_0) - (By_0 + b)^T V(V^T BV)^+ V^T (By_0 + b) > 0.$$

It indicates that the quadratic equation $h(x^*(y)) = 0$ has no solution. By Theorem 5, (QP1EQC) cannot have an optimal solution. Since $v(\text{QP1EQC}) > -\infty$, it is unattainable. The other case (44) can be analogously argued.

Next, we show that, if $v(\text{QP1EQC}) > -\infty$ and $I_{\succeq}(A, B)$ is not a single-point set, an optimal solution to (QP1EQC) can be constructed. Then, by Lemma 1, we can conclude that the dual SDP (39) is always attained when $v(\text{QP1EQC}) > -\infty$. First, by an argument in the proof of Theorem 5.1 in [24], $I_{\succeq}(A, B)$ is an interval with an interior point. Denote by

$$I_{\succeq}(A, B) = [\mu_{\min}, \mu_{\max}], \quad \mu_{\min} < \mu_{\max}$$

whereas it is possible that $\mu_{\min} = -\infty$ and $\mu_{\max} = +\infty$. Since $I_{\succeq}(A, B)$ is an interval with a non-empty interior, by Theorem 3 (b) in [17], we have

$$\mathcal{N}(A + \mu B) = \mathcal{N}(A) \cap \mathcal{N}(B), \quad \forall \mu \in (\mu_{\min}, \mu_{\max}).$$

Let $V \in \mathbb{R}^{n \times r}$ be the basis matrix of $\mathcal{N}(A) \cap \mathcal{N}(B)$ and $U \in \mathbb{R}^{n \times (n-r)}$ be the orthonormal complementary subspace of V . Then we have

$$\begin{bmatrix} U^T \\ V^T \end{bmatrix} (A + \mu B) \begin{bmatrix} U & V \end{bmatrix} = \begin{bmatrix} U^T A U + \mu U^T B U & 0 \\ 0 & 0 \end{bmatrix}, \quad \forall \mu \in (\mu_{\min}, \mu_{\max}).$$

Let $u \in \mathbb{R}^{n-r}$. By $A + \mu B \succeq 0$, $(u^T U^T)(A + \mu B)(Uu) = 0$ if and only if $(A + \mu B)(Uu) = 0$. Since U is the orthogonal complement of V , it must be $u = 0$. In other words.

$$U^T A U + \mu U^T B U \succ 0, \quad \forall \mu \in (\mu_{\min}, \mu_{\max}). \quad (45)$$

With the $[U \ V]_{n \times n}$ coordinate change and the notation $0_{m \times r}$ for the $m \times r$ zero matrix; 0_n for the n -dimensional zero vector, we can recast the dual SDP (39) as

$$\begin{aligned} & \sup \left\{ s \in \mathbb{R} : \begin{bmatrix} U^T & 0_{n-r} \\ V^T & 0_r \\ 0_n^T & 1 \end{bmatrix} \begin{bmatrix} A + \mu B & a + \mu b \\ a^T + \mu b^T & c + \mu d - s \end{bmatrix} \begin{bmatrix} U & V & 0_n \\ 0_{n-r}^T & 0_r^T & 1 \end{bmatrix} \succeq 0 \right\} \\ &= \sup \left\{ s \in \mathbb{R} : \begin{bmatrix} U^T A U + \mu U^T B U & 0_{(n-r) \times r} & U^T(a + \mu b) \\ 0_{r \times (n-r)} & 0_{r \times r} & V^T(a + \mu b) \\ (a + \mu b)^T U & (a + \mu b)^T V & c + \mu d - s \end{bmatrix} \succeq 0 \right\} \\ &= \sup \left\{ s \in \mathbb{R} : V^T(a + \mu b) = 0, \begin{bmatrix} U^T A U + \mu U^T B U & U^T(a + \mu b) \\ (a + \mu b)^T U & c + \mu d - s \end{bmatrix} \succeq 0 \right\} \quad (46) \end{aligned}$$

Since $v(\text{QP1EQC}) > -\infty$, the dual SDP (39) is feasible. By (45), it implies that (46) admits a strict feasible point (μ, s) that satisfies the positive semi-definite constraint. By writing down the conic dual of (46):

$$\begin{aligned} (\text{UD}) \quad & \inf_{Y, z} \begin{bmatrix} U^T A U & U^T a \\ a^T U & c \end{bmatrix} \bullet Y + a^T V z \\ & \text{s.t.} \begin{bmatrix} U^T B U & U^T b \\ b^T U & d \end{bmatrix} \bullet Y + b^T V z = 0, \\ & Y_{n-r+1, n-r+1} = 1, \ Y \in S_+^{n-r+1}, \end{aligned}$$

and by the strong duality theorem, there must be

$$v(\text{QP1EQC}) = v(\text{LD}) = v(\text{Prob}(46)) = v(\text{UD}) > -\infty.$$

In particular, (UD) can be attained at an optimal solution, say (Y^*, z^*) . Let

$$Y^* = \begin{bmatrix} Y_{(n-r) \times (n-r)}^* & Y_{\{1:n-r\}, n-r+1}^* \\ Y_{\{1:n-r\}, n-r+1}^{*T} & Y_{n-r+1, n-r+1}^* \end{bmatrix}$$

where $Y_{\{1:n-r\}, n-r+1}^* = (Y_{1, n-r+1}^*, Y_{2, n-r+1}^*, \dots, Y_{n-r, n-r+1}^*)^T$. Then, define

$$X^* = \begin{bmatrix} U & V & 0_n \\ 0_{n-r}^T & 0_r^T & 1 \end{bmatrix} \begin{bmatrix} Y_{(n-r) \times (n-r)}^* & 0_{(n-r) \times r} & Y_{\{1:n-r\}, n-r+1}^* \\ 0_{r \times (n-r)} & \frac{1}{4} z^* z^{*T} & \frac{1}{2} z^* \\ Y_{\{1:n-r\}, n-r+1}^{*T} & \frac{1}{2} z^{*T} & 1 \end{bmatrix} \begin{bmatrix} U^T & 0_{n-r} \\ V^T & 0_r \\ 0_n^T & 1 \end{bmatrix} \succeq 0.$$

We can verify that

$$\begin{bmatrix} B & b \\ b^T & d \end{bmatrix} \bullet X^* = 0, \quad X_{n+1, n+1}^* = 1,$$

and

$$\begin{bmatrix} A & a \\ a^T & c \end{bmatrix} \bullet X^* = \begin{bmatrix} U^T A U & U^T a \\ a^T U & c \end{bmatrix} \bullet Y^* + a^T V z^* = v(\text{UD}) = v(\text{QP1EQC}).$$

In other words, X^* is an optimal solution of the primal SDP (40). Employing the standard rank-one decomposition procedure [32], we have shown that (QP1EQC) is attained. The proof is complete. \square

The following example is provided to illustrate the idea of Theorem 6.

Example 1 Consider

$$\begin{aligned} & \inf x_1^2 \\ & \text{s.t. } x_1 x_2 - 1 = 0, \end{aligned}$$

where $a = b = (0, 0)^T$, $c = 0$, $d = -1$ and

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix}.$$

The graph of the objective function is a paraboloid by extending the parabola $z = x_1^2$ parallel along the x_2 direction. The constraint is a hyperbola upon which there are a pair of traces on $z = x_1^2$, both asymptotically approaching 0 from the positive territory. Then, $v(\text{QP1EQC}) = 0$ but is unattainable.

The primal SDP (40) is

$$\begin{aligned} & \inf \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \bullet X \\ & \text{s.t. } \begin{bmatrix} 0 & 0.5 & 0 \\ 0.5 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} \bullet X = 0, \\ & X_{3,3} = 1, \quad X \succeq 0. \end{aligned}$$

Define

$$X(\epsilon) = \begin{bmatrix} \epsilon & 1 & 0 \\ 1 & \frac{1}{\epsilon} & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Then $X(\epsilon)$ is a feasible solution of the primal SDP (40) for any $\epsilon > 0$. The corresponding objective value is ϵ . Since $\lim_{\epsilon \rightarrow 0} X(\epsilon)$ does not have a limit, the optimal value of the primal SDP is 0, but unattainable.

The dual SDP (39) is

$$\begin{aligned} & \sup s \\ & \text{s.t. } \begin{bmatrix} 1 & 0.5\mu & 0 \\ 0.5\mu & 0 & 0 \\ 0 & 0 & -\mu - s \end{bmatrix} \succeq 0. \end{aligned}$$

Obviously, $\mu^* = 0$ is the only feasible dual solution, which forces $s \leq 0$. Then, the dual optimal value is 0, confirming the strong duality; and the dual optimal solution is $(s^*, \mu^*) = (0, 0)$, confirming that the dual must be attainable if the problem is bounded from below.

Moreover, we can verify that

$$\mathcal{N}(A + \mu^* B) = \mathcal{N}(A) = \text{span}\{(0, 1)^T\}$$

such that $V = (0, 1)^T$. A direct computation shows

$$y_0 = -(A + \mu^* B)^+(a + \mu^* b) = (0, 0)^T; \quad V^T B V = 0,$$

and hence

$$h(y_0 + V y) = h(y_0) - (B y_0 + b)^T V (V^T B V)^+ V^T (B y_0 + b) = -1,$$

which confirms (44) that none of the Kuhn-Tucker points are feasible.

Finally, we consider the interval bounded generalized trust region subproblem (GTRS), which is to deal with $\inf\{f(x) : l \leq h(x) \leq u, x \in \mathbb{R}^n\}$. Jeyakumar et al. [21] studied (GTRS) using Polyak's result [28] (see also Theorem 7 in Section 5). They assumed that h is homogeneous and strictly convex. Pong and Wolkowicz [29] proved the strong duality based on the the following assumptions:

1. $B \neq 0$.
2. (GTRS) is feasible.
3. The following relative interior constraint qualification holds

$$(\text{RICQ}) \quad l < B \bullet \hat{X} + 2b^T \hat{x} + d < u, \text{ for some } \hat{X} \succ \hat{x}\hat{x}^T.$$

4. (GTRS) is bounded below.
5. The dual problem of (GTRS) is feasible.

Very recently, Wang and Xia [34] showed that Assumption 3 above in [29] can be equivalently rephrased as a simple “strict feasibility” assumption that there exists an $\bar{x} \in \mathbb{R}^n$ such that $l < h(\bar{x}) < u$. Hence Assumption 2 in [29] is redundant. In addition, Assumption 4 naturally implies Assumption 5 by a similar result from Theorem 6 in this section. This has also been done in [34] by Wang and Xia. Indeed, based on the S-lemma with equality from an earlier arXiv version of this paper, they established the S-lemma with interval bounds. Under the strict feasibility assumption, the following two statements are equivalent $((I_1) \sim (I_2))$:

- (I₁) $l \leq h(x) \leq u \implies f(x) \geq 0, \forall x \in \mathbb{R}^n$.
- (I₂) $(\exists \mu \in \mathbb{R}) \quad f(x) + \mu_-(h(x) - u) + \mu_+(l - h(x)) \geq 0, \forall x \in \mathbb{R}^n$ where $\mu_+ = \max\{\mu, 0\}, \mu_- = -\min\{\mu, 0\}$.

except for the special case where A has exactly one negative eigenvalue, $B = 0$, $b \neq 0$ and there exists a $\nu \geq 0$ such that

$$\begin{bmatrix} V^T A V & \frac{1}{2b^T b} V^T A b & V^T a \\ \frac{1}{2b^T b} b^T A V & \frac{b^T A b}{(2b^T b)^2} + \nu & \frac{a^T b}{2b^T b} - \frac{\nu}{2}(l + u - 2d) \\ a^T V & \frac{a^T b}{2b^T b} - \frac{\nu}{2}(l + u - 2d) & c + \nu(l - d)(u - d) \end{bmatrix} \succeq 0, \quad (47)$$

with $V \in \mathbb{R}^{n \times (n-1)}$ being the matrix basis of $\mathcal{N}(b)$.

In summary, (GTRS) is now completely answered by the S-lemma with equality. When the strict feasibility holds and $B \neq 0$, (GTRS) has a tight SDP relaxation (the SDP formulation can be read in [29, 34]). In that case, if $v(\text{GTRS}) > -\infty$ and attainable, an optimal solution can be found by solving the SDP relaxation followed by a typical rank-one decomposition procedure. Otherwise, when (GTRS) fails to satisfy the strict feasibility or has $B = 0$, $h(x)$ reduces to be linear. Then, (GTRS) would be either a convex unconstrained optimization problem with its optimal solution residing in the interval $l \leq h(x) \leq u$; or can be determined by $v(\text{GTRS}) = \min \{ \inf_{h(x)=l} f(x), \inf_{h(x)=u} f(x) \}$. Since $h(x) = l$ (or $h(x) = u$) can be replaced by a system of linear equations, each of both becomes an unconstrained problem and could either be unbounded below or has an attainable optimal solution.

5 Application to Convexity of Joint Numerical Range

Dines in 1941 proved a fundamental but somehow surprising result in classical mathematical analysis that the joint numerical range $M = \{(f(x), h(x)) : x \in \mathbb{R}^n\} \subseteq \mathbb{R}^2$ is convex when f and h are quadratic forms [9]. He explained in the same paper that the observation on the convexity of M was motivated by Finsler's theorem (S-lemma) [12] in 1937 because “it asserts the existence of a supporting line of the map M which has contact with M only at $(0, 0)$. This suggests that the theorem is related to the theory of convex sets.” Dines also described the shape of M to be either the entire x - y plane or an angular sector of angle less than π , provided f, h have no common zero except $x = 0$.

Since then, the progress has been slow. Extension of Dines' result to the 2D image of nonhomogeneous functions f and h occurred much later in 1998 due to Polyak [28]. It may be stated as

Theorem 7 ([28]) *Let $f, h \in \mathbb{R}^n$ be nonhomogeneous quadratic functions. Suppose that $n \geq 2$ and there exists $\mu \in \mathbb{R}^2$ such that*

$$\mu_1 A + \mu_2 B \succ 0.$$

Then the set $M = \{(f(x), h(x)) : x \in \mathbb{R}^n\} \subseteq \mathbb{R}^2$ is closed and convex.

A counterexample was provided in the same paper [28] to show that the joint numerical range M is in general nonconvex for nonhomogeneous quadratic functions. Dines' theorem even fails when one of f and h is an affine function.

Example 2 ([28]) *Consider the following two quadratic functions in \mathbb{R}^2*

$$f(x) = 2x_1^2 - x_2^2, \quad h(x) = x_1 + x_2.$$

Then we can verify that

$$M = \{(f(x), h(x)) : x \in \mathbb{R}^n\} = \{(y_1, y_2) \in \mathbb{R}^2 : y_1 \geq -2y_2^2\}$$

which is nonconvex.

In 2007, Beck [2] studied the convexity of the image of mappings comprised of a strictly convex quadratic function and a set of affine functions.

Theorem 8 ([2]) *Let $h_i(x) = 2b_i^T x + d_i$ for $i = 1, \dots, p$, where $b_i \in \mathbb{R}^n, d_i \in \mathbb{R}$. Suppose $A \succ 0$. When $p \leq n - 1$, the set*

$$S = \{(f(x), h_1(x), \dots, h_p(x)) : x \in \mathbb{R}^n\} \subseteq \mathbb{R}^{p+1} \quad (48)$$

is closed and convex. Besides, when $p = n$ and b_1, \dots, b_n are linearly independent, S is nonconvex.

It happens that the newly developed S-lemma with equality can be used to give a theorem on the convexity of jointly numerical range S , sufficiently strong to cover both Example 2 and Theorem 8. Namely, we are able to say something similar to what Dines had done in [9], from the S-lemma with equality (cf. Finsler's theorem) to the convexity of S (cf. M).

Theorem 9 *Suppose $h_i(x) = 2b_i^T x + d_i$ for $i = 1, \dots, p$, where $b_i \in \mathbb{R}^n, d_i \in \mathbb{R}$. Let $P = [b_1, b_2, \dots, b_p]^T$ and $r = \text{rank}(P)$. Then, the set S defined in (48) is convex if and only if neither of the following two cases occurs:*

- (a) $V^T A V \succeq 0$, $V^T a \in \mathcal{R}(V^T A)$ and $W^T A W$ has a negative eigenvalue;
- (b) $V^T A V \preceq 0$, $V^T a \in \mathcal{R}(V^T A)$ and $W^T A W$ has a positive eigenvalue,

where $V \in \mathbb{R}^{n \times (n-r)}$ is the matrix basis of $\mathcal{N}(P)$ and $W \in \mathbb{R}^{n \times (n-\text{rank}(V^T A))}$ is the matrix basis of $\mathcal{N}(V^T A)$.

Proof. Let $u = (u_f, u_{h_1}, \dots, u_{h_p})$ and $v = (v_f, v_{h_1}, \dots, v_{h_p})$ be any two distinct points in S . That is, there exists $x_u, x_v \in \mathbb{R}^n$ and $x_u \neq x_v$ such that

$$u_f = f(x_u), \quad u_{h_i} = h_i(x_u); \quad v_f = f(x_v), \quad v_{h_i} = h_i(x_v), \quad i = 1, \dots, p.$$

Then, S is nonconvex if and only if we cannot find $x_\lambda \in \mathbb{R}^n$ such that

$$(f(x_\lambda), h_1(x_\lambda), \dots, h_p(x_\lambda)) = (1 - \lambda)u + \lambda v, \quad \lambda \in (0, 1).$$

Equivalently, the following system in terms of (x_λ, λ) has no solution

$$f(x_\lambda) = (1 - \lambda)u_f + \lambda v_f, \quad (49)$$

$$h_i(x_\lambda) = (1 - \lambda)u_{h_i} + \lambda v_{h_i}, \quad i = 1, \dots, p, \quad (50)$$

$$\lambda \in (0, 1). \quad (51)$$

Since h_1, \dots, h_p are affine, the equations (50) imply that

$$\begin{aligned} h_i(x_\lambda) &= 2b_i^T x_\lambda + d_i \\ &= (1 - \lambda)(2b_i^T x_u + d_i) + \lambda(2b_i^T x_v + d_i) \\ &= 2b_i^T ((1 - \lambda)x_u + \lambda x_v) + d_i, \quad i = 1, 2, \dots, p \end{aligned}$$

and thus x_λ has to lie on the affine space:

$$x_\lambda = (1 - \lambda)x_u + \lambda x_v + Vy, \quad \text{for some } y \in \mathbb{R}^{n-r},$$

where $V \in \mathbb{R}^{n \times (n-r)}$ is the matrix basis of $\mathcal{N}([b_1, b_2, \dots, b_p]^T)$. Substituting x_λ into (49), we obtain

$$\begin{aligned} f(x_\lambda) &= f((1-\lambda)x_u + \lambda x_v + Vy) \\ &= (1-\lambda)^2 x_u^T A x_u + \lambda^2 x_v^T A x_v + 2\lambda(1-\lambda)x_u^T A x_v + 2a^T((1-\lambda)x_u + \lambda x_v) \\ &\quad + y^T V^T A V y + 2((1-\lambda)x_u + \lambda x_v)^T A V y + 2a^T V y + c. \end{aligned} \quad (52)$$

Equate (52) with

$$(1-\lambda)f(x_u) + \lambda f(x_v) = (1-\lambda)(x_u^T A x_u) + \lambda x_v^T A x_v + 2a^T((1-\lambda)x_u + \lambda x_v) + c$$

to yield

$$\begin{aligned} &y^T V^T A V y + 2((1-\lambda)x_u + \lambda x_v)^T A V y + 2a^T V y \\ &= \lambda(1-\lambda)(x_u^T A x_u + x_v^T A x_v - 2x_u^T A x_v). \end{aligned}$$

This is a quadratic equation in variables $(\lambda, y) \in \mathbb{R}^{1+n-r}$. To simplify, let $\delta = x_u^T A x_u + x_v^T A x_v - 2x_u^T A x_v = (x_u - x_v)^T A (x_u - x_v)$ and

$$G = \begin{bmatrix} \delta & (x_v - x_u)^T A V \\ V^T A (x_v - x_u) & V^T A V \end{bmatrix}, \quad q = \begin{bmatrix} -\delta \\ 2V^T(Ax_u + a) \end{bmatrix} \quad (53)$$

to obtain

$$\begin{aligned} g(\lambda, y) &= \delta \lambda^2 - (2(x_u - x_v)^T A V y + \delta) \lambda + y^T V^T A V y + 2x_u^T A V y + 2a^T V y \\ &:= \begin{bmatrix} \lambda \\ y \end{bmatrix}^T G \begin{bmatrix} \lambda \\ y \end{bmatrix} + q^T \begin{bmatrix} \lambda \\ y \end{bmatrix}. \end{aligned}$$

With all the settings, the joint numerical range S is not convex if and only if the system (49)-(51) is unsolvable, which is equivalent to the system

$$\lambda^2 - \lambda < 0, \quad g(\lambda, y) = 0 \quad (54)$$

having no solution.

We can show that, when (54) is unsolvable, $g(\lambda, y)$ must take both positive and negative values. Should such a claim fail, we would have

$$G \succeq (\preceq) 0, \quad q \in \mathcal{R}(G), \quad q^T G^+ q = 0. \quad (55)$$

Since $G \succeq (\preceq) 0$ implies that $G^+ \succeq (\preceq) 0$, $q^T G^+ q = 0$ implies that $G^+ q = 0$. Therefore,

$$q \in \mathcal{N}(G^+) = \mathcal{N}(G).$$

However, from (55), we also know $q \in \mathcal{R}(G)$ and thus $q = 0$. By (53),

$$\delta = 0, \quad V^T(Ax_u + a) = 0.$$

Since $G \succeq (\preceq) 0$, all 2×2 principal minors consisting of $\delta = 0$ must have non-negative determinants. Then, the first row (column) of G should be identically

0. Namely, $V^T A(x_v - x_u) = 0$. The function g is thus reduced to $g(\lambda, y) = y^T V^T A V y$ so that (54) would be trivially solvable. It is a contradiction.

Since Assumption 1 holds and since the quadratic term of the function $\varphi(\lambda, y) = \lambda^2 - \lambda$ does not have exactly one negative eigenvalue, by the S-lemma with equality, the statement (54) has no solution if and only if there is a μ such that

$$\lambda^2 - \lambda + \mu g(\lambda, y) \geq 0, \quad \forall \lambda \in \mathbb{R}, y \in \mathbb{R}^{n-r},$$

which has the following positive semi-definite formulation:

$$\begin{bmatrix} 1 + \mu\delta & -\mu(x_u - x_v)^T A V & -\frac{1}{2} - \frac{1}{2}\mu\delta \\ -\mu V^T A(x_u - x_v) & \mu V^T A V & \mu V^T (A x_u + a) \\ -\frac{1}{2} - \frac{1}{2}\mu\delta & \mu(x_u^T A + a^T) V & 0 \end{bmatrix} \succeq 0. \quad (56)$$

By considering all 2×2 principal minors including the 0 element in (56), we obtain

$$-\frac{1}{2} - \frac{1}{2}\mu\delta = 0, \quad \mu V^T (A x_u + a) = 0, \quad \mu V^T A(x_u - x_v) = 0, \quad \mu V^T A V \succeq 0.$$

Therefore, $\mu \neq 0$, $\delta \neq 0$ and

$$V^T a \in \mathcal{R}(V^T A), \quad (57)$$

$$x_u - x_v \in \mathcal{N}(V^T A), \quad (58)$$

$$-\frac{1}{\delta} V^T A V \succeq 0. \quad (59)$$

Let $W \in \mathbb{R}^{n \times (n - \text{rank}(V^T A))}$ be the matrix basis of $\mathcal{N}(V^T A)$ and express (58) as

$$x_u - x_v = Wz, \quad \text{for some } z \in \mathbb{R}^{n - \text{rank}(V^T A)}.$$

Then, (59) implies that

$$\delta = (x_u - x_v)^T A(x_u - x_v) = z^T W^T A W z \begin{cases} < 0, & \text{if } V^T A V \succeq 0, \\ > 0, & \text{if } V^T A V \preceq 0, \end{cases}$$

which, together with (57), completes the proof. \square

By Theorem 9, we can now confirm that the joint numerical range in Example 2 is non-convex.

Example 3 Consider the nonconvex Example 2 where $n = 2$, $p = 1$:

$$A = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}, \quad a = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad b = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}.$$

Then we can compute

$$V = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix}, \quad V^T A = \begin{bmatrix} \sqrt{2} & \frac{\sqrt{2}}{2} \end{bmatrix}, \quad W = \begin{bmatrix} \frac{\sqrt{5}}{5} \\ -\frac{2\sqrt{5}}{5} \end{bmatrix}$$

and, by Case (a) in Theorem 9, $M = \{(2x_1^2 - x_2^2, x_1 + x_2) : (x_1, x_2) \in \mathbb{R}^2\}$ is nonconvex since

$$V^T AV = \frac{1}{2} \succeq 0, \quad V^T a = 0 \in \mathcal{R}(V^T A), \quad W^T AW = -\frac{2}{5}.$$

Next, we show that Theorem 9 implies Theorem 8 under $A \succ 0$.

– Suppose $p \leq n - 1$. We have

$$r = \text{rank}([b_1, b_2, \dots, b_p]^T) \leq n - 1. \quad (60)$$

Then, $\text{rank}(V) = n - r \geq 1$. Since V is a basis matrix consisting of at least one column and $A \succ 0$, we have $V^T AV \succ 0$ and $W^T AW \succeq 0$. It follows that neither Case (a) nor Case (b) in Theorem 9 holds. Therefore, the set S defined in (48) is convex. We remark that the condition $p \leq n - 1$ in Theorem 8 can be improved to (60).

– Suppose $p = n$ and b_1, \dots, b_n are linearly independent. We have $r = \text{rank}([b_1, b_2, \dots, b_p]^T) = n$, and $\text{rank}(V) = 0$, $\text{rank}(W) = n$. Then,

$$V^T AV = 0, \quad V^T a \in \mathcal{R}(V^T A), \quad W^T AW \succ 0.$$

By Case (b) in Theorem 9, the set S defined in (48) is nonconvex.

Finally in this section, we mention a weaker result than the convexity of $\{(f(x), h_1(x), \dots, h_p(x)) : x \in \mathbb{R}^n\}$, which asks for the convexity of

$$\{(f(x), h_1(x), \dots, h_p(x)) : x \in \mathbb{R}^n\} + \mathbb{R}_+^{p+1}, \quad (61)$$

where f, h_1, \dots, h_p are nonhomogeneous quadratic functions and \mathbb{R}_+^{p+1} is the nonnegative orthant of \mathbb{R}^{p+1} . For the special case $p = 1$, Tuy and Tuan showed that $\{(f(x), h_1(x)) : x \in \mathbb{R}^n\} + \mathbb{R}_+^2$ is always convex, see Corollary 10 of [33]. For a general p , suppose $f(x), h_1(x), \dots, h_p(x)$ are homogeneous quadratic functions and $\nabla^2 f(x), \nabla^2 h_1(x), \dots, \nabla^2 h_p(x)$ are all Z-matrices (A real symmetric matrix A is called a Z-matrix if $A_{ij} \leq 0$ for all $i \neq j$), Jeyakumar et al. [21] showed that

$$\{(f(x), h_1(x), \dots, h_p(x)) : x \in \mathbb{R}^n\} + \text{int}\mathbb{R}_+^{p+1}$$

is convex, where $\text{int}\mathbb{R}_+^{p+1}$ is the interior of \mathbb{R}_+^{p+1} . More recently, suppose $h_1(x)$ is a strictly convex quadratic function, $h_2(x), \dots, h_p(x)$ are affine linear functions, for any nonhomogeneous quadratic function $f(x)$, Jeyakumar and Li [22] showed that the set (61) is convex under the dimension assumption:

$$\dim \mathcal{N}(\nabla^2 f(x) - \lambda_{\min}(\nabla^2 f(x))I_n) \geq \dim \text{span}\{\nabla h_2(x), \dots, \nabla h_p(x)\} + 1,$$

where $\dim L$ denotes the dimension of the subspace L and $\lambda_{\min}(A)$ is the minimal eigenvalue of the matrix A .

Since the convexity of S guarantees the convexity of $S + \mathbb{R}_+^{p+1}$, Theorem 9 trivially implies a new sufficient condition for the set (61).

Corollary 2 Let $f(x) = x^T A x + 2a^T x + c$ and $h_i(x) = 2b_i^T x + d_i$ for $i = 1, \dots, p$. Suppose neither Case (a) nor Case (b) in Theorem 9 occurs, the set $\{(f(x), h_1(x), \dots, h_p(x)) : x \in \mathbb{R}^n\} + \mathbb{R}_+^{p+1}$ is convex.

Moreover, when $p = 1$ and $h_1(x)$ is an affine function, with the help of Theorem 9, we can see how $\{(f(x), h_1(x)) : x \in \mathbb{R}^n\} + \mathbb{R}_+^2$ becomes convex in the case that $\{(f(x), h_1(x)) : x \in \mathbb{R}^n\}$ is nonconvex.

Corollary 3 Let $f(x) = x^T A x + 2a^T x + c$ and $h_1(x) = 2b_1^T x + d_1$. Suppose $S := \{(f(x), h_1(x)) : x \in \mathbb{R}^n\}$ is nonconvex. Then, exactly one of the following cases happens:

- (i) The set S contains a parametric curve $C(t) = \{(u(t), v(t)) : t \in \mathbb{R}\}$ satisfying $\lim_{t \rightarrow \infty} u(t) = \lim_{t \rightarrow \infty} v(t) = -\infty$. It follows immediately that $S + \mathbb{R}_+^2 = \mathbb{R}^2$.
- (ii) $A = \alpha b_1 b_1^T$ for some $\alpha > 0$. Consequently, $S + \mathbb{R}_+^2$ is convex.

Proof. Since S is nonconvex, by Theorem 9, either Case (a) or Case (b) occurs.

Suppose Case (a) in Theorem 9 happens. Then A has a negative eigenvalue. Let $z \neq 0$ be the corresponding eigenvector, i.e., $z^T A z < 0$. Then, $b_1^T z \neq 0$. Otherwise, $b_1^T z = 0$ would imply that $z \in \mathcal{R}(V)$. It would follow from $V^T A V \succeq 0$ that $z^T A z \geq 0$, which is a contradiction. Without loss of generality, we assume that $b_1^T z < 0$ and define the parametric curve

$$C(t) = \{(u(t), v(t)) : u(t) = f(tz), v(t) = h_1(tz), t \in \mathbb{R}\}.$$

Then, $C(t) \subset S$ and

$$\lim_{t \rightarrow \infty} u(t) = \lim_{t \rightarrow \infty} f(tz) = \lim_{t \rightarrow \infty} t^2 z^T A z + 2ta^T z + c = -\infty; \quad (62)$$

$$\lim_{t \rightarrow \infty} v(t) = \lim_{t \rightarrow \infty} h_1(tz) = \lim_{t \rightarrow \infty} 2t(b_1^T z) + d_1 = -\infty, \quad (63)$$

which shows that (i) holds.

Suppose Case (b) in Theorem 9 happens. We know A must have a positive eigenvalue with $z \neq 0$ being the corresponding eigenvector. Since $V^T A V \preceq 0$ and $z^T A z > 0$, there must be $b_1^T z \neq 0$ and we again assume that $b_1^T z < 0$. In addition, if A also has a negative eigenvalue except for the positive one(s), by the same argument as in (62)-(63), we conclude immediately that (i) holds.

Otherwise, when $A \succeq 0$ but $V^T A V \not\preceq 0$, there exists a vector u such that $u^T V^T A V u < 0$. By continuity, we can find a sufficiently small $\beta > 0$ such that $(Vu + \beta z)^T A (Vu + \beta z) < 0$. By defining the curve

$$C(t) = \{(u(t), v(t)) : u(t) = f(t(Vu + \beta z)), v(t) = h_1(t(Vu + \beta z)), t \in \mathbb{R}\},$$

we also see that (i) holds since

$$\begin{aligned} & \lim_{t \rightarrow \infty} u(t) \\ &= \lim_{t \rightarrow \infty} t^2 (Vu + \beta z)^T A (Vu + \beta z) + 2ta^T (Vu + \beta z) + c \\ &= -\infty, \end{aligned}$$

and

$$\lim_{t \rightarrow \infty} v(t) = \lim_{t \rightarrow \infty} h_1(t(Vu + \beta z)) = \lim_{t \rightarrow \infty} 2t\beta(b_1^T z) + d_1 = -\infty.$$

Finally, it remains to show that $A \succeq 0$ and $V^T AV = 0$ lead to (ii). When this happens, we have $AV = 0$. Since V is the matrix basis for the null space of the $1 \times n$ matrix $[b_1^T]$, A must be of rank one with the form $A = \alpha b_1 b_1^T$ for some $\alpha > 0$. That is, (ii) holds. It follows from the convexity of $f(x)$ and $h_1(x)$ that $S + \mathbb{R}_+^2$ is convex. \square

6 Concluding remarks

This paper is devoted to a completely new understanding toward the S-lemma with equality. While the inequality version $(S_1) \sim (S_2)$ has been established under Slater's condition, it was not immediately clear whether the equality version $(E_1) \sim (E_2)$ could be a real obstacle. While the inequality version $(S_1) \sim (S_2)$ has been widely used in many applications, the important consequence of $(E_1) \sim (E_2)$, perhaps due to lack of an affirmative result, was not yet visualized before. As $(E_1) \sim (E_2)$ is not true in general, we really need to get down to the most subtle detail looking for a successful argument, many from geometrical observations on quadratic manifolds. This complicated essence as well as the poor accessibility to intuition make it hard to come out with an easy-to-grasp intuitive proof. The long-standing interval bounded generalized trust region subproblem (GTRS) has now been resolved thoroughly by the full characterization of the quadratic programming with a single quadratic equality constraint (QP1EQC). The relation between the S-lemma and the convexity of joint numerical ranges is now further strengthened, indicating a step forward to the duality theory for nonconvex optimization. We wish that the study can sparkle new idea for solving the nonconvex quadratic optimization problem with multiple constraints and polynomial optimization problems.

Appendix

We discuss the relations among S-Conditions 1,2,3 and 4. First, it is easy to see that the definiteness of B implies $A \succeq \eta B$ for some η . Therefore,

$$\text{S-Condition 1} \implies \text{S-Condition 2.}$$

Moreover, when B is definite, $x^T Bx = 0$ if and only if $x = 0$ and thus

$$\text{S-Condition 1} \implies \text{S-Condition 4.}$$

When $h(x)$ is homogeneous, there is $b = d = 0$ so that $h(0) = 0$. By choosing $\zeta = 0$,

$$\text{S-Condition 3} \implies \text{S-Condition 4}$$

It is not difficult to verify that neither S-Condition 2 nor S-Condition 4 can imply each other [25]. Consequently, neither S-Condition 2 nor S-Condition 4 is necessary for the S-lemma with equality.

We now show that the statement (4) in S-Condition 4 can be equivalently simplified as

$$\begin{cases} b \in \mathcal{R}(B), & \text{if } B \succeq 0 \text{ or } B \preceq 0, \\ B\zeta + b = 0, & \text{otherwise.} \end{cases} \quad (64)$$

Notice that (64) trivially implies (4), so it is sufficient to prove the converse.

Suppose $B \succeq 0$ or $B \preceq 0$. Then, $x^T Bx = 0 \iff Bx = 0$ and (4) can be recast as

$$Bx = 0 \implies b^T x = 0, \forall x \in \mathbb{R}^n,$$

which shows that $b \in \mathcal{R}(B)$.

Now assume that B is indefinite. We first rewrite (4) by

$$x^T Bx = 0 \implies x^T (B\zeta + b)(B\zeta + b)^T x = 0, \forall x \in \mathbb{R}^n.$$

Since $(B\zeta + b)(B\zeta + b)^T \succeq 0$, it is further equivalent to

$$x^T Bx = 0 \implies -x^T (B\zeta + b)(B\zeta + b)^T x \geq 0, \forall x \in \mathbb{R}^n.$$

Since B is indefinite, $h(x) = x^T Bx$ takes both positive and negative values. By the S-lemma with equality for homogeneous quadratic forms,

$$(\exists \mu \in \mathbb{R}) \quad -(B\zeta + b)(B\zeta + b)^T + \mu B \succeq 0. \quad (65)$$

Since $\mu B \succeq (B\zeta + b)(B\zeta + b)^T \succeq 0$ and B is indefinite, it must be $\mu = 0$ and thus $(B\zeta + b)(B\zeta + b)^T = 0$. It implies that $B\zeta + b = 0$.

Acknowledgments

The authors are grateful to the two anonymous referees for their valuable comments.

References

1. Anstreicher, K.M., Wright, M.H.: A note on the augmented Hessian when the reduced Hessian is semidefinite. *SIAM J. Optim.* **11**(1), 243–253 (2000)
2. Beck, A.: On the Convexity of a Class of Quadratic Mappings and its Application to the Problem of Finding the Smallest Ball Enclosing a Given Intersection of Ball, *J. Global Optim.* **39**, 113–126 (2007)
3. Beck, A., Eldar, Y.C.: Strong duality in nonconvex quadratic optimization with two quadratic constraint. *SIAM J. Optim.* **17**(3), 844–860 (2006)
4. Ben-Tal, A., Hertog, D.: Hidden conic quadratic representation of some nonconvex quadratic optimization problems. *Math. Program. Ser. A* **143**, 1–9 (2014)
5. Ben-Tal, A., Teboulle, M.: Hidden convexity in some nonconvex quadratically constrained quadratic programming. *Math. Program.* **72**, 51–63 (1996)
6. Brickman, L.: On the field of values of a matrix. *Proc. Am. Math. Soc.* **12**, 61–66 (1961)

7. Burer, S., Anstreicher, K.M.: Second-Order-Cone Constraints for Extended Trust-Region Subproblems. *SIAM J. Optim.* **23**(1), 432–451 (2013)
8. Derinkuyu, K., Pinar, M.Ç.: On the S-procedure and some variants. *Math. Meth. Oper. Res.* **64**, 55–77 (2006)
9. Dines, L.L.: On the mapping of quadratic forms. *Bull. Amer. Math. Soc.* **47**, 494–498 (1941)
10. Fang, S.C., Gao, D.Y., Lin, G.X., Sheu, R.L., Xing, W.: Double well potential function and its optimization in the n-dimensional real space – Part I, *Math. Mech. Solids* (2015) DOI:10.1177/1081286514566704
11. Feng, J.M., Lin, G.X., Sheu, R.L., Xia, Y.: Duality and solutions for quadratic programming over single non-homogeneous quadratic constraint. *J. Global Optim.* **54**(2), 275–293 (2012)
12. Finsler, P.: Über das vorkommen definiter und semidefiniter Formen in scharen quadratischer Formen. *Comment. Math. Helv.* **9**, 188–192 (1937)
13. Fradkov, A.L., Yakubovich, V.A.: The S-procedure and the duality relation in convex quadratic programming problems, *Vestnik Leningrad. Univ.* **1**, 81–87 (1973)
14. Hestenes, M.R.: *Optimization Theory*, John Wiley & Sons (1975)
15. Hmam, H.: Quadratic optimisation with one quadratic equality constraint. *Electronic Warfare and Radar Division DSTO Defence Science and Technology Organisation, Australia*, Report DSTO-TR-2416 (2010)
16. Horn, R., Johnson, C.R.: *Matrix Analysis*. Cambridge University Press, Cambridge, UK (1985)
17. Hsia, Y., Lin, G.X., Sheu, R.L.: A revisit to quadratic programming with one inequality quadratic constraint via matrix pencil. *Pac. J. Optim.* **10**(3), 461–481 (2014)
18. Jerrard R.L.: Lower bounds for generalized Ginzburg-Landau functionals, *SIAM J. Math. Anal.* **30**(4), 721–746 (1999)
19. Jeyakumar, V.: Farkas lemma: generalizations. *Encyclopedia of optimization.* **2**, 87–91, Kluwer Boston, USA (2000)
20. Jeyakumar, V., Huy, N.Q., Li, G.: Necessary and sufficient conditions for S-lemma and nonconvex quadratic optimization, *Optim. Eng.* **10**, 491–503 (2009)
21. Jeyakumar, V., Lee, G.M., Li, G.Y.: Alternative theorems for quadratic inequality systems and global quadratic optimization, *SIAM J. Optim.* **20**(2), 983–1001 (2009)
22. Jeyakumar, V., Li, G.Y.: Trust-Region Problems with Linear Inequality Constraints: Exact SDP Relaxation, Global Optimality and Robust Optimization, *Math. Program.* **147**(1-2), 171–206 (2014)
23. Martínez-Legaz, J.E. On Brickman’s theorem. *J. Convex Anal.* **12**, 139–143 (2005)
24. Moré, J.J.: Generalizations of the trust region problem. *Optim. Methods Softw.* **2**, 189–209 (1993)
25. Nguyen, V.B., Sheu, R.L., Xia, Y.: An SDP approach for quadratic fractional problems with a two-sided quadratic constraint. *Optim. Methods Softw.* (2015) DOI:10.1080/10556788.2015.1029575
26. Palanthandalam-Madapusi, H.J., Pelt, T.H.V., Bernstein, D.S.: Matrix pencils and existence conditions for quadratic programming with a sign-indefinite quadratic equality constraint. *J. Global Optim.* **45**(4), 533–549 (2009)
27. Pólik, I., Terlaky, T.: A Survey of the S-lemma. *SIAM Rev.* **49**(3), 371–418 (2007)
28. Polyak, B.T.: Convexity of quadratic transformations and its use in control and optimization. *J. Optimiz. Theory App.* **99**(3), 553–583 (1998)
29. Pong, T.K., Wolkowicz, H.: The generalized trust region subproblem. *Comput. Optim. Appl.* **58**(2), 273–322 (2014)
30. Shor, N.Z.: Quadratic optimization problems. *Sov. J. Comput. Syst. Sci.* **25**, 1–11 (1987)
31. Sturm, J.F., Wolkowicz, H.: Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations. *SIAM J. Optim.* **5**, 286–313 (1995)
32. Sturm, J.F., Zhang, S.: On cones of nonnegative quadratic functions. *Math. Oper. Res.* **28**(2), 246–267 (2003)
33. Tuy, H., Tuan, H.D.: Generalized S-lemma and strong duality in nonconvex quadratic programming. *J. Global Optim.* **56**(3): 1045–1072 (2013)
34. Wang, S., Xia, Y.: Strong Duality for Generalized Trust Region Subproblem: S-Lemma with Interval Bounds. *Optim. Lett.* (2014) DOI:10.1007/s11590-014-0812-0

-
35. Xia, Y., Sheu, R.L., Fang, S.C., Xing, W.: Double well potential function and its optimization in the n-dimensional real space – Part II, *Math. Mech. Solids* (2015) DOI:10.1177/1081286514566723
 36. Yakubovich, V.A.: S-procedure in nonlinear control theory. *Vestnik Leningrad. Univ.* **1**, 62–77 (1971) (in Russian)
 37. Yakubovich, V.A.: S-procedure in nonlinear control theory. *Vestnik Leningrad. Univ.* **4**, 73–93 (1977) (English translation)
 38. Ye, Y., Zhang, S.: New results on quadratic minimization. *SIAM J. Optim.* **14**, 245–267 (2003)