Math. Program., Ser. A
DOI 10.1007/s10107-017-1145-4

FULL LENGTH PAPER

SOCP reformulation for the generalized trust region subproblem via a canonical form of two symmetric matrices

Rujun Jiang[1] · **Duan Li**[1] · **Baiyi Wu**[2]

Received: 11 January 2016 / Accepted: 1 April 2017

Abstract We investigate in this paper the generalized trust region subproblem (GTRS) of minimizing a general quadratic objective function subject to a general quadratic inequality constraint. By applying a simultaneous block diagonalization approach, we obtain a congruent canonical form for the symmetric matrices in both the objective and constraint functions. By exploiting the block separability of the canonical form, we show that all GTRSs with an optimal value bounded from below are second order cone programming (SOCP) representable. Our result generalizes the recent work of Ben-Tal and den Hertog (Math. Program. 143(1–2):1–29, 2014), which establishes the SOCP representability of the GTRS under the assumption of the simultaneous diagonalizability of the two matrices in the objective and constraint functions. We then derive a closed-form solution for the GTRS when the two matrices are not simultaneously diagonalizable. We further extend our method to two variants of the GTRS in which the inequality constraint is replaced by either an equality constraint or an interval constraint.

Keywords Trust region subproblem · Canonical form · Quadratically constrained quadratic programming

✉ Baiyi Wu
baiyiwu@outlook.com

Rujun Jiang
rjjiang@se.cuhk.edu.hk

Duan Li
dli@se.cuhk.edu.hk

[1] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

[2] School of Finance, Guangdong University of Foreign Studies, Guangzhou, Guangdong, China

Published online: 18 April 2017

Mathematics Subject Classification 90C20 · 90C26

1 Introduction

We consider in this paper the following generalized trust region subproblem (GTRS):

$$\begin{aligned}
(\mathrm{P})\ \ \min \quad & f(x)=\frac{1}{2}x^TDx+e^Tx\\
\text{s.t.} \quad & h(x)=\frac{1}{2}x^TAx+b^Tx+c\leq 0,
\end{aligned}$$

where A and D are $n\times n$ symmetric matrices but not necessarily positive semi-definite, $b, e\in\mathbb{R}^n$ and $c\in\mathbb{R}$.

The GTRS has been widely investigated in the optimization literature and includes the classical trust region subproblem (TRS) as its special case where the constraint reduces to a unit ball, i.e., $x^Tx\leq 1$. The classical TRS is fundamental in trust region methods for nonlinear optimization problems, see [5]. Other prominent applications of the classical TRS can be found in regularization and robust optimization, see [2]. Rendl and Wolkowicz [20] derive a fast algorithm to solve the classical TRS by exploiting the properties of a semi-definite programming (SDP) reformulation. The past two decades have witnessed theoretical results and numerous methods developed for solving the GTRS under various assumptions, see, for example, [4,7,15–17,19,20,23,28]. Sturm and Zhang [23] further reveal that problem (P) admits an SDP reformulation under the Slater condition. Still, the relatively large computational complexity of SDP algorithms prevents them from scaling to large-scale problems. Most fast algorithms [16,19,20] for the GTRS are developed under a regular condition that there exists a $\lambda\in\mathbb{R}$ such that $D+\lambda A\succ 0$, together with some other mild conditions. Recently, Ben-Tal and den Hertog [2] show that if the two matrices in both the objective and constraint functions are simultaneously diagonalizable (SD), the GTRS can be then transformed into an equivalent second order cone programming (SOCP) problem formulation. Simultaneous diagonalizability is actually a more general condition than the regular condition [7,11]. Conditions for simultaneous diagonalizability and corresponding algorithms are investigated recently in [12]. However, the literature lacks deep investigation about the general situations when the two matrices are not SD. To the best of our knowledge, the paper [11] represents an exception with some results for GTRS when the two matrices are not SD. This recognition of the state-of-the-art motivates our investigation in this paper.

Our contribution is twofold: we give a theoretical result for problem (P) when the two matrices are not SD and provide a stable algorithm for numerical implementation using properties revealed from the canonical form. More specifically, we prove that all GTRSs with an optimal value bounded from below are SOCP representable. To obtain the SOCP representation, we invoke and extend the congruent canonical form in Uhlig [24]. We first transform the two matrices into their canonical form of block diagonal matrices via congruence, and then make use of the block separability of the matrices in the canonical form to derive the SOCP reformulation. In particular, we derive necessary conditions from the canonical form for the GTRS to be bounded from below, and then show that the problem can further be transformed to an SOCP reformulation under such necessary conditions. Moreover, we can determine the attainableness of the optimal value by examining the associated coefficients in the linear terms in both the objective and constraint functions. In fact, our method using simultaneous block diagonalization is a generalization of the simultaneous diagonalizability in [2]. We simplify our result further to achieve a closed-form solution for problem (P) under the canonical form. Although this nice theoretical result cannot be directly applied to solve problem (P) as there is no stable algorithm in the current literature for computing the canonical form when the quadratic forms are not SD, we succeed to show in this paper that, by using the properties of the canonical form, we can derive a stable procedure to solve problem (P) without computing the canonical form.

We also extend our approach to two variants of problem (P) where the constraint is replaced by either an equality constraint,

$$\begin{aligned}
\text{(EP)}\quad \min \quad & f(x)=\frac{1}{2}x^TDx+e^Tx\\
\text{s.t.} \quad & h(x)=\frac{1}{2}x^TAx+b^Tx+c=0,
\end{aligned}$$

or an interval constraint,

$$\begin{aligned}
\text{(IP)}\quad \min \quad & f(x)=\frac{1}{2}x^TDx+e^Tx\\
\text{s.t.} \quad & c_1\le h(x)=\frac{1}{2}x^TAx+b^Tx\le c_2.
\end{aligned}$$

Moré [16] presents a method for problem (EP) by using the saddle point optimality condition under some mild assumptions. Xia et al. [26] transform the problem (EP) to an SDP reformulation by using the S-lemma with equality under the conditions that $A\neq 0$ and $h(x)$ can take both positive and negative values. One application of problem (EP) is time of arrival problem [10]. Stern and Wolkowicz [22] propose a method for problem (IP) under $b=0$ and the regular condition. By assuming $b=0$ and the simultaneous diagonalizability of A and D, Ben-Tal and Teboulle [4] derive the hidden convexity of problem (IP) and thus transform the problem to an SOCP reformulation. Ye and Zhang [27] further show that problem (IP) admits an SDP reformulation if both the primal and dual Slater conditions are satisfied. Recently, strong duality conditions of (IP) are studied in Pong and Wolkowicz [19] and a fast method is provided under the regular condition. Wang and Xia [25] further simplify the conditions in [19] and develop the S-lemma with interval bounds to solve (IP). Ben-Tal and den Hertog [2] further show that (IP) can be solved as an SOCP when A and D are SD without the assumption of $b=0$. Note that (IP) includes the equality constrained problem (EP) as a special case when setting $c_1=c_2$. Salahi and Taati [21] also derive an efficient algorithm for solving (IP) under the SD condition. On the other hand, we will discuss in latter sections that solution methods for (EP) can also be used to solve (IP). Essentially, we will show that some slightly modified versions of our previous results for problem (P) hold true for the equality constrained problem (EP) and the interval constrained problem (IP).

To summarize, we derive necessary conditions for problem (P) and its variants with equality constraint or interval constraint to be bounded from below and further transform the problems to their SOCP reformulations by exploiting the block separability of a canonical form of matrix pencils. Besides, we also derive the conditions for the attainableness of the optimal value. We finally provide a numerically stable method to solve problem (P) without computing the canonical form when the quadratic forms are not SD.

We organize our paper as follows. In Sect. 2, we introduce and extend a canonical form for the matrix pair in the quadratic forms and investigate the canonical form to understand the hidden structure for solving problem (P). Further more, we derive a stable method to solve problem (P) by using the structure of the canonical form, when the quadratic forms are not SD. In Sect. 3, we extend our methods to problems (EP) and (IP). In Sect. 4, we carry out numerical tests to demonstrate the effectiveness of our methods. Finally, we conclude our paper in Sect. 5.

Notations Throughout this paper, I_m represents the $m \times m$ identical matrix. The notation $\mathbb{R}^n$ represents the n dimensional vector space. For symmetric matrices A and B, $A \succeq B$ denotes that matrix $A - B$ is positive semi-definite. We denote the *Moore–Penrose pseudoinverse* of matrix A by A^+. We use $\operatorname{sign}(x)$ to denote the sign of a real number x, i.e., $\operatorname{sign}(x) = 1$, if $x \geq 0$, otherwise $\operatorname{sign}(x) = -1$. We use $\dim A$ to denote the dimension of a square matrix A, and use $A_{k:l,k:l}$ to denote the submatrix of matrix A by selecting the rows $k, k+1, \ldots, l$, and the columns $k, k+1, \ldots, l$. We also denote by $\operatorname{diag}(A_1, \ldots, A_k)$ the block diagonal matrix

$$\operatorname{diag}(A_1, \ldots, A_k) = \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_k \end{pmatrix}.$$

We denote by E the anti-diagonal matrix and by F the lower striped matrix,

$$E = \begin{pmatrix} 0 & & & 1 \\ & & \cdot & \\ & \cdot & & \\ \cdot & & & \\ 1 & & & 0 \end{pmatrix}, \quad F = \begin{pmatrix} & & & & 0 \\ & & & 0 & 1 \\ & & \cdots & & \\ & 0 & 1 & & \\ 0 & 1 & & & \end{pmatrix}. \tag{1}$$

We use $J(\lambda, m)$ to denote an $m \times m$ Jordan block

$$\begin{pmatrix} \lambda & \varkappa & & \\ & \cdot & \cdot & \\ & & \lambda & \varkappa \\ & & & \lambda \end{pmatrix}.$$

If the eigenvalue is a real number, i.e., $\lambda \in \mathbb{R}$, then $\varkappa = 1$ for $m \geq 2$, while $J = (\lambda)$ for $m = 1$. If the eigenvalues form a complex pair, i.e., $a \pm bi$, then $\lambda = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$, with a and $b \in \mathbb{R}$, and $b \neq 0$, and $\varkappa = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ for $m \geq 4$, while $J = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ for $m = 2$.

Let $v(\mathrm{P})$ denote the optimal value of problem (P). For an optimization problem $\min\{f(x) \mid x \in X\}$ with a nonempty feasible set X, we define its optimal value as $v = \inf\{f(x) \mid x \in X\}$. For any $\epsilon > 0$, we call a solution $\bar{x} \in X$ an ϵ optimal solution, if $f(\bar{x}) - v \leq \epsilon$.

2 SOCP reformulation for GTRS

In this section, we first transform any two symmetric matrices into a canonical form of two block diagonal matrices via congruence. Then we explore different cases of the canonical form with respect to the boundedness of the optimal value and its attainability and transform problem (P) to an equivalent SOCP problem if its optimal value is bounded from below. We further derive a closed-form solution when the quadratic forms are not SD. Finally, we show that although there is no stable algorithm in the current literature for computing the canonical form, we can still derive a stable algorithm to solve problem (P) by exploiting the properties of the canonical form.

2.1 Canonical form of two symmetric matrices

We invoke the following lemmas from Uhlig [24] to obtain a canonical form of any two real symmetric matrices.

Lemma 1 (Theorem 1 in [24]) *Let A and D be two $n \times n$ real symmetric matrices. Suppose A is nonsingular. Let $A^{-1}D$ have a Jordan normal form $diag(J_1, \ldots, J_k)$, where $J_1, \ldots, J_k$ are Jordan blocks either with real eigenvalues or with complex eigenvalues. Then there exists an $n \times n$ real congruent matrix S such that*

$$S^T AS = \mathrm{diag}(\varepsilon_1 E_1, \ldots, \varepsilon_k E_k)$$

and

$$S^T DS = \mathrm{diag}(\varepsilon_1 E_1 J_1, \ldots, \varepsilon_k E_k J_k),$$

where $\varepsilon_i = \pm 1$ and E_i is defined in (1). Furthermore, the signs of ε_i, $i = 1, \ldots, k$, are uniquely (up to permutations) associated with the Jordan blocks, J_i, $i = 1, \ldots, k$. In particular, $\varepsilon_i = 1$ if J_i has a pair of complex eigenvalues.

Lemma 2 (Theorem 2 in [24]) *Let A and D be two singular real symmetric matrices and assume that there exists a $\mu \in \Re$ such that $C = A + \mu D$ is nonsingular. Let*

$$J = \operatorname{diag}(J(\lambda_1, n_1), \ldots, J(\lambda_k, n_k), J(0, n_{k+1}), \ldots, J(0, n_p), J(1/\mu, n_{p+1}), \ldots, J(1/\mu, n_m))$$

be the Jordan normal form of $C^{-1}D$. Then there exists an $n \times n$ real congruent matrix S such that

$$S^T AS = \operatorname{diag}(\tau_1 E_1, \ldots, \tau_k E_k, \tau_{k+1} E_{k+1}, \ldots, \tau_p E_p, \tau_{p+1} F_{p+1}, \ldots, \tau_m F_m) \quad (2)$$

and

$$S^T DS = \operatorname{diag}(\tau_1 E_1 J(\kappa_1, n_1), \ldots, \tau_k E_k J(\kappa_k, n_k), \tau_{k+1} F_{k+1}, \ldots, \tau_p F_p, \tau_{p+1} E_{p+1}, \ldots, \tau_m E_m), \quad (3)$$

where E_i and F_i are defined in (1), respectively, $\tau_i = \pm 1$, $i = 1, \ldots, m$, $\dim E_i = \dim F_i = n_i$, $i = k+1, \ldots, m$, and $\kappa_i = \lambda_i/(1 - \mu\lambda_i)$, $i = 1, \ldots, k$. The signs of τ_i are uniquely (up to permutations) determined by the associated Jordan blocks $J(\kappa_i, n_i)$, E_i or F_i, $i = 1, \ldots, k$. In particular, $\tau_i = 1$ if $J(\lambda_i, n_i)$ has a pair of complex eigenvalues, $i = 1, \ldots, k$. Furthermore, $p - k \geq 1$ and $m - p \geq 1$.

Next we generalize the results in the previous two lemmas to general situations where we do not assume the existence of the nonsingular matrix pencil for two symmetric matrices.

Theorem 1 *For any two $n \times n$ real symmetric matrices A and D, there exists an $n \times n$ real invertible matrix S such that*

$$S^T AS = \operatorname{diag}(\tau_1 E_1, \ldots, \tau_k E_k, \tau_{k+1} E_{k+1}, \ldots, \tau_p E_p, \tau_{p+1} F_{p+1}, \ldots, \tau_m F_m, 0, \ldots, 0) \quad (4)$$

and

$$S^T DS = \operatorname{diag}(\tau_1 E_1 J(\kappa_1, n_1), \ldots, \tau_k E_k J(\kappa_k, n_k), \tau_{k+1} F_{k+1}, \ldots, \tau_p F_p, \tau_{p+1} E_{p+1}, \ldots, \tau_m E_m, 0 \ldots, 0), \quad (5)$$

where $\dim E_i = \dim F_i = n_i$, $i = k+1, \ldots, m$, and $\tau_i = \pm 1$, $i = 1, \ldots, m$. The signs of τ_i are uniquely (up to permutations) determined by the associated Jordan blocks $J(\kappa_i, n_i)$, E_i or F_i. The values of κ_i are uniquely (up to permutations) determined by the associated Jordan blocks $J(\kappa_i, n_i)$, $i = 1, \ldots, k$. In particular, $\tau_i = 1$ if $J(\lambda_i, n_i)$ has a pair of complex eigenvalues, $i = 1, \ldots, k$.

Proof Based on the results in Lemmas 1 and 2, we only need to consider in the proof the case where A and D are both singular and there does not exist a $\mu \in \mathbb{R}$ such that $A + \mu D$ is nonsingular.

We can always find a congruent matrix Q_1 such that $\bar{A} \triangleq Q_1^T A Q_1 = \operatorname{diag}(A_1, 0, \ldots, 0)$, where A_1 is a $q \times q$ diagonal matrix and $q = \operatorname{rank}(A)$. Denote

$$\bar{D} \triangleq Q_1^T D Q_1 = \begin{pmatrix} D_1 & D_2 \\ D_2^T & D_3 \end{pmatrix},$$

where D_1 is a $q \times q$ matrix. We can always find a congruent matrix S such that $S^T D_3 S = \operatorname{diag}(D_6, 0, \ldots, 0)$, where D_6 is a nonsingular $s \times s$ diagonal matrix. Let $Q_2 \triangleq \operatorname{diag}(I_q, S)$. Then $\hat{A} \triangleq Q_2^T \bar{A} Q_2 = \bar{A}$, and

$$\hat{D} \triangleq Q_2^T \bar{D} Q_2 = \begin{pmatrix} D_1 & D_4 & D_5 \\ D_4^T & D_6 & 0 \\ D_5^T & 0 & 0 \end{pmatrix}.$$

Let

$$Q_3 \triangleq \begin{pmatrix} I_q & 0 & 0 \\ -D_6^{-1} D_4^T & I_s & 0 \\ 0 & 0 & I_{n-q-s} \end{pmatrix}.$$

Then,

$$\tilde{D} \triangleq Q_3^T \hat{D} Q_3 = \begin{pmatrix} D_1 - D_4 D_6^{-1} D_4^T & 0 & D_5 \\ 0 & D_6 & 0 \\ D_5^T & 0 & 0 \end{pmatrix},$$

and $\tilde{A} \triangleq Q_3^T \hat{A} Q_3 = \hat{A} = \bar{A}$. We can always choose a $\mu \in \mathbb{R}$ such that the first q columns of $\tilde{D} + \mu \tilde{A}$ are linearly independent. For example, we can choose $\mu = \max_{i=1,\ldots,q} \sum_{j=1}^{q} |b_{ij}|/|a_{ii}| + 1$, where b_{ij} is the element in the ith row and the jth column of $D_1 - D_4 D_6^{-1} D_4^T$ and a_{ii} is the ith diagonal element of A_1. Then $\mu A_1 + D_1 - D_4 D_6^{-1} D_4^T$ is nonsingular.

If the columns in D_5 are linearly independent, then $\tilde{D} + \mu \tilde{A}$ is nonsingular and thus $D + \mu A$ is nonsingular, which contradicts our assumption of no nonsingular matrix pencil. Thus the columns in D_5 are linearly dependent. We can always find a congruent matrix Q_4 such that $\check{A} = Q_4^T \tilde{A} Q_4 = \hat{A} = \bar{A}$, and

$$\check{D} \triangleq Q_4^T \tilde{D} Q_4 = \begin{pmatrix} D_1 - D_4 D_6^{-1} D_4 & 0 & D_5' & 0 \\ 0 & D_6 & 0 & 0 \\ D_5'^T & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

where D_5' is of full column rank. Let

$$A' \triangleq \begin{pmatrix} A_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad D' \triangleq \begin{pmatrix} D_1 - D_4 D_6^{-1} D_4^T & 0 & D_5' \\ 0 & D_6 & 0 \\ D_5'^T & 0 & 0 \end{pmatrix}.$$

Then there exists a μ such that $D' + \mu A'$ is nonsingular. From Lemma 2 we know that A' and D' can be congruent to the canonical form in (2) and (3). So A and D can be congruent to the canonical form in (4) and (5). □

Remark 1 If A and D are both singular and there does not exist a $\mu \in \mathbb{R}$ such that $A + \mu D$ is nonsingular, then the number of the common 0 terms in the lower right part of (4) and (5) is equal to $n - \operatorname{rank}(A' + \mu D')$. The common 0 terms in fact correspond to the common null space of D and A.

From Lemmas 1, 2 and Theorem 1, we know that (4) and (5) represent a canonical form for any two real symmetric matrices A and D via congruence. Without loss of generality, we assume before Sect. 2.3 that matrices A and D in problem (P) satisfy:

$$
\begin{aligned}
A &= \operatorname{diag}(A_1, A_2, \ldots, A_s)\\
&= \operatorname{diag}\left(\tau_1 E_1, \ldots, \tau_k E_k, \tau_{k+1} E_{k+1}, \ldots, \tau_p E_p, \tau_{p+1} F_{p+1}, \ldots, \tau_m F_m, 0, \ldots, 0\right),\\
D &= \operatorname{diag}(D_1, D_2, \ldots, D_s)\\
&= \operatorname{diag}(\tau_1 E_1 J(\kappa_1, n_1), \ldots, \tau_k E_k J(\kappa_k, n_k), \tau_{k+1} F_{k+1}, \ldots, \tau_p F_p, \tau_{p+1} E_{p+1},\\
&\qquad \ldots, \tau_m E_m, 0 \ldots, 0).
\end{aligned}
$$

Note that we have four kinds of block pairs (A_i, D_i): $(\tau_i E_i, \tau_i E_i J(\kappa_i, n_i))$, $(\tau_i E_i, \tau_i F_i)$, $(\tau_i F_i, \tau_i E_i)$ and $(0, 0)$. In fact, the second kind of block pairs is a special case of the first kind with $\kappa_i = 0$ due to $E_i J(0, n_i) = F_i$. We call the first two kinds of block pairs type A block pairs, the third kind of block pairs type B block pairs and the last one type C block pairs.

Remark 2 Using another canonical form in [14], a similar idea appeared in [1] in March 2016 for solving the GTRS. The canonical form in [14] is essentially the same as that in Theorem 1 generalized from [24] except that the canonical form in [14] expresses the blocks associated with complex eigenvalues separately.

2.2 Assumptions

Without loss of generality, we make the following assumptions.

Assumption 1 (i) There is at least one feasible solution in problem (P); (ii) The following three conditions do not hold true at the same time: $A \succeq 0$, $b \in \operatorname{Range}(A)$ and $c = \frac{1}{2} b^T A^+ b$.

Note that problem (P) is infeasible if and only if $A \succeq 0$, $b \in \operatorname{Range}(A)$ and $c = \frac{1}{2} b^T A^+ b + k$ for some $k > 0$, which leads to $h(x) = \frac{1}{2}(x + A^+ b)^T A (x + A^+ b) + k > 0$. If $A \succeq 0$, $b \in \operatorname{Range}(A)$ and $c = \frac{1}{2} b^T A^+ b$, then $h(x) = \frac{1}{2} x^T A x + b^T x + c = \frac{1}{2}(x + A^+ b)^T A (x + A^+ b) \geq 0$. Thus, the inequality constraint in problem (P) becomes an equality constraint which means all the feasible solutions are in the boundary. Actually, Assumption 1 is equivalent to the Slater condition, i.e., there exists an $\bar{x}$ such that $h(\bar{x}) < 0$. Moreover, when the three conditions in (ii) hold together, problem (P) reduces to an unconstrained quadratic problem: decompose A as $A = L^T L$, where $L \in \mathbb{R}^{r \times n}$ with r being the rank of A. Then, the constraint becomes $(x + A^+ b)^T L^T L (x + A^+ b) = 0$ and thus $L(x + A^+ b) = 0$. Rewrite $x = -A^+ b + V y$, where $V \in \mathbb{R}^{n \times (n-r)}$ is a matrix basis of the null space of L

and $y \in \mathbb{R}^{n-r}$. The problem then reduces to an unconstrained quadratic optimization problem.

We then, w.l.o.g., make similar assumptions for (EP) and (IP).

Assumption 2 (i) There is at least one feasible solution in (EP); (ii) The following conditions do not hold true at the same time: $A \succeq 0$ or $A \preceq 0$, $b \in \text{Range}(A)$ and $c = \frac{1}{2}b^T A^+ b$; and (iii) $A \neq 0$.

Assumption 2 is equivalent to a "two-side" Slater condition (Assumption 1 in [26]) plus the condition $A \neq 0$. In fact, if $A = 0$, we can transform the constraint $b^T x + c = 0$ to another quadratic equality constraint which satisfies all the three conditions in (i), i.e., $(b^T x + c)^2 = 0$. So problem (EP) can be transformed to an unconstrained quadratic optimization problem using the null space representation of $L(x + A^+ b) = 0$, by decomposing A as $A = L^T L$, when (ii) is violated, or using the null space representation of $b^T x + c = 0$ when (iii) is violated.

Similarly, we have the following Assumption 3.

Assumption 3 (i) There is at least one feasible solution in (IP); (ii) The following conditions do not hold true at the same time: $A \succeq 0$ or $A \preceq 0$, $b \in \text{Range}(A)$ and $c_i = \frac{1}{2}b^T A^+ b$ for $i = 1$ or 2; and (iii) $A \neq 0$.

It is also interesting to compare Assumption 3 to Assumption 2.1 in [19].

Assumption 4 (Assumption 2.1 in [19]) (i) $A \neq 0$; (ii) (IP) is feasible; (iii) The following relative interior constraint qualification holds

$$c_1 < tr(A\hat{X}) - 2b^T\hat{x} < c_2 \text{ for some } \hat{X} \succ \hat{x}\hat{x}^T;$$

(iv) (IP) is bounded from below; (v) dual problem of (IP) is feasible.

It is easy to see that (i) and (iii) in Assumption 3 are just (ii) and (i), respectively, in Assumption 4. It is shown in [25] that (iii) in Assumption 4 is equivalent to the existence of an $\bar{x} \in \mathbb{R}^n$ such that $c_1 < \frac{1}{2}\bar{x}^T A\bar{x} - b^T\bar{x} < c_2$, which is further equivalent to (ii) in Assumption 3. So our Assumption 3 is just Assumption 2.1 in [19] without the boundedness assumption.

2.3 SOCP reformulation from the canonical form

Let us recall the S-lemma [18], which states the equivalence of the following two statements under Slater condition:

(S_1) $(\forall x \in \mathbb{R}^n)\ h(x) \leq 0 \Rightarrow f(x) \geq 0$.
(S_2) $\exists\ \mu \geq 0$ such that $f(x) + \mu h(x) \geq 0,\ \forall x \in \mathbb{R}^n$.

In general, $f(x)$ can be represented as $f(x) = \frac{1}{2}x^T Dx + e^T x + v$ with an additional constant v in (S_1) and (S_2), and we use this representation of $f(x)$ in the following of this section when discussing the S-lemma. The connection between problem (P) and the S-lemma is illustrated in [26] as follows,

$$
\begin{aligned}
v(P) &= \inf_{x\in\mathbb{R}^n}\{\, f(x) \mid h(x) \le 0\}\\
&= \sup_{\eta\in\mathbb{R}}\{\, \eta : \{x \in \mathbb{R}^n \mid f(x) < \eta, h(x) \le 0\} = \emptyset\}\\
&= \sup_{\eta\in\mathbb{R}}\{\, \eta \mid \exists\mu \ge 0 \text{ such that } f(x) - \eta + \mu h(x) \ge 0,\ \forall x \in \mathbb{R}^n\}\\
&= \sup_{\eta\in\mathbb{R},\mu\ge 0}\{\, \eta \mid \begin{pmatrix} D+\mu A & e+\mu b \\ e^T + \mu b^T & 2(v+\mu c-\eta)\end{pmatrix} \succeq 0\}.
\end{aligned} \tag{6}
$$

In fact, (6) leads to a necessary and sufficient condition under which problem (P) is bounded from below in the following lemma.

Lemma 3 ([11]) *Under Assumption 1, problem (P) is bounded from below if and only if the following system has a solution for* λ*:*

$$
D + \lambda A \succeq 0,\ \lambda \ge 0,\ e + \lambda b \in \text{Range}(D + \lambda A). \tag{7}
$$

With Lemma 3, we have the following theorems.

Theorem 2 *Consider the case where a type A block pair* $(\tau_i E_i, \tau_i E_i J(\kappa_i, n_i))$ *exists in problem* (P)*. If the size of the associated Jordan block* $J(\kappa_i, n_i)$ *is greater than 2 and the associated eigenvalue of the Jordan block is real, then the objective value of* (P) *is unbounded from below, i.e.,* $v(\text{P}) = -\infty$*.*

Proof If the size of the associated Jordan block $J(\kappa_i, n_i)$ is greater than 2, then $\tau_i(E_i J(\kappa_i, n_i) + \mu E_i)$ takes the following form

$$
\tau_i(E_i J(\kappa_i, n_i) + \mu E_i) = \tau_i \begin{pmatrix} & & & \kappa_i+\mu \\ & & \kappa_i+\mu & 1 \\ & \cdots & \cdots & \\ \kappa_i+\mu & 1 & & \end{pmatrix}.
$$

Since the $(n_i - 1) \times (n_i - 1)$ principal minor

$$
\tau_i \begin{pmatrix} & & \kappa_i+\mu & 1 \\ & \cdots & \cdots & \\ \kappa_i+\mu & 1 & & \\ 1 & & & \end{pmatrix}
$$

is non-positive semi-definite when its size $n_i - 1$ is greater than or equal to 2, $D_i + \mu A_i = \tau_i(E_i J(\kappa_i, n_i) + \mu E_i)$ cannot be positive semi-definite. Thus, there is no μ such that $D + \mu A = \text{diag}(D_1 + \mu A_1, \ldots, D_m + \mu A_m, 0, \ldots, 0) \succeq 0$. So (7) is infeasible and by Lemma 3 we have $v(\text{P}) = -\infty$. □

Using similar proofs, we have the following theorems.

Theorem 3 *Consider the case where a type A block pair* $(\tau_i E_i, \tau_i E_i J(\kappa_i, n_i))$ *exists in problem* (P)*. If the eigenvalues of the associated Jordan block* $J(\kappa_i, n_i)$ *form a complex pair, then the objective value of problem* (P) *is unbounded from below, i.e.,* $v(\text{P}) = -\infty$*.*

Proof If the eigenvalues of the associated Jordan block $J(\kappa_i, n_i)$ form a complex pair, then there does not exist a $\mu \in \mathbb{R}$ such that

$$\tau_i(E_i J(\kappa_i, n_i) + \mu E_i) = \tau_i \begin{pmatrix} & & & b_i & a_i + \mu \\ & & & a_i + \mu & -b_i \\ & & \cdots & & \\ b_i & a_i + \mu & & & \\ a_i + \mu & -b_i & & & \end{pmatrix} \succeq 0,$$

because the 4×4 principal minor (if $n_i = 4k$ for some positive integer k),

$$\tau_i \begin{pmatrix} & & b_i & a_i + \mu \\ & & a_i + \mu & -b_i \\ b_i & a_i + \mu & & \\ a_i + \mu & -b_i & & \end{pmatrix}$$

or the 2×2 principal minor (if $n_i = 4k + 2$ for some positive integer k),

$$\tau_i \begin{pmatrix} b_i & a_i + \mu \\ a_i + \mu & -b_i \end{pmatrix}$$

is non-positive semi-definite. So (7) is infeasible and by Lemma 3 we get $v(\mathrm{P}) = -\infty$. □

Theorem 4 *Consider the case where a type B block pair* $(\tau_i F_i, \tau_i E_i)$ *exists in problem* (P). *If* $\dim F_i \geq 2$*, then problem* (P) *is unbounded from below, i.e.,* $v(\mathrm{P}) = -\infty$.

Proof If the size of the associated Jordan block $J(\kappa_i, n_i)$ is larger than or equal to 2, then there does not exist a $\mu \in \mathbb{R}$ such that

$$\tau_i(E_i + \mu F_i) = \tau_i \begin{pmatrix} & & & & 1 \\ & & & 1 & \mu \\ & & \cdots & & \\ & 1 & \mu & & \\ 1 & \mu & & & \end{pmatrix} \succeq 0.$$

So (7) is infeasible and by Lemma 3 we have $v(\mathrm{P}) = -\infty$. □

Remark 3 Assumption 1 is necessary in Theorem 4. Otherwise the S-lemma does not hold and we have the following counter example: $\min f(x) = x_1 x_2$ subject to $h(x) = x_2^2 \leq 0$. The problem has a size 2 block pair $(F_{2\times 2}, E_{2\times 2})$ but a finite optimal value of $\min_{h(x) \leq 0} f(x) = 0$.

So if problem (P) has a finite optimal solution, then any type B block pairs are of size 1 and any type A block pairs are of a size less than or equal to 2 and the eigenvalues in the associated Jordan blocks are real. Now let us consider a type A block pair with size 2, and, without loss of generality, let it be the first block $(A_1, D_1) = (\tau_1 E_1, \tau_1 E_1 J_1(\lambda, 2))$ with

$$E_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \ E_1 J_1(\lambda, 2) = \begin{pmatrix} 0 & \lambda \\ \lambda & 1 \end{pmatrix}.$$

Define $D = \mathrm{diag}(D_1, D_J)$, $A = \mathrm{diag}(A_1, A_J)$, $I = \{1, 2\}$, $J = \{3, \ldots, n\}$, $e_I = (e_1, e_2)^T$, $e_J = (e_3, \ldots, e_n)^T$, $b_I = (b_1, b_2)^T$, $b_J = (b_3, \ldots, b_n)^T$, $z = (x_1, x_2)^T$, and $y = (x_3, \ldots, x_n)^T$. We can then represent problem (P) as follows,

$$\begin{aligned} \text{(RP)} \min \quad & \frac{1}{2} y^T D_J y + e_J^T y + \frac{1}{2} z^T D_1 z + e_I^T z \\ \text{s.t.} \quad & \frac{1}{2} y^T A_J y + b_J^T y + \frac{1}{2} z^T A_1 z + b_I^T z + c \le 0. \end{aligned}$$

The term in the constraint associated with (A_1, D_1) is

$$\frac{1}{2} z^T A_1 z + b_I^T z = \frac{1}{2} \tau_1 z^T E_1 z + b_I^T z = \tau_1 z_1 z_2 + b_1 z_1 + b_2 z_2, \tag{8}$$

and the term in the objective function associated with (A_1, D_1) is

$$\frac{1}{2} z^T D_1 z + e_I^T z = \frac{1}{2} \tau_1 z^T E_1 J_1(\lambda, 2) z + e_I^T z = \tau_1 \lambda z_1 z_2 + \frac{1}{2} \tau_1 z_2^2 + e_1 z_1 + e_2 z_2. \tag{9}$$

Without loss of generality, we further assume $b_1 = b_2 = 0$. Since otherwise when letting $z_1' = z_1 + \tau_1 b_2$ and $z_2' = z_2 + \tau_1 b_1$, the constraint function will become $\frac{1}{2} y^T A_J y + b_J^T y + \frac{1}{2} \tau_1 z'^T A_1 z' + c'$, where $c' = c - \tau_1 b_1 b_2$, and the objective function will become $\frac{1}{2} y^T D_J y + e_J^T y + \frac{1}{2} \tau_1 z'^T D_1 z' + e'^T_I z' + p$, where $e'_1 = e_1 - \lambda b_1$, $e'_2 = e_2 - b_1 - \lambda b_2$, and $p = -e_1 \tau_1 b_2 - e_2 \tau_1 b_1 + \tau_1 \lambda b_1 b_2 + \frac{1}{2} \tau_1 b_1^2$. Note that $\tau_1 = \pm 1$ according to Theorem 1.

From now on, we assume that the coefficients in b corresponding to any 2×2 type A Jordan block pair are 0.

Theorem 5 *Consider the case where there exists a type A block pair* $(\tau_1 E_1, \tau_1 E_1 J_1(\lambda, 2))$ *in problem* (P) *and the eigenvalue of the associated Jordan block* $J_1(\lambda, 2)$ *is real. Assume there is a feasible solution* $\bar{x} = (\bar{z}^T, \bar{y}^T)^T$ *and let* $\pi = \tau_1 \bar{z}_1 \bar{z}_2$. *Let* $\rho = \inf\{(9) \mid (8) = \tau_1 z_1 z_2 \le \pi\}$. *We have the following three cases:*

1. *When* $\tau_1 = 1$. *If* $(\lambda \le 0,\ e_1 = 0,\ e_2 \ne 0)$ *or* $(\lambda = 0,\ e_1 = 0,\ e_2 = 0,\ \pi \ge 0)$ *or* $(\lambda < 0,\ e_1 = 0,\ e_2 = 0,\ \pi = 0)$, *then* $\rho = \lambda\pi - \frac{1}{2} e_2^2$ *and the infimum is attainable;*
2. *When* $\tau_1 = 1$. *If* $(\lambda = 0,\ e_1 = 0,\ e_2 = 0,\ \pi < 0)$ *or* $(\lambda < 0,\ e_1 = 0,\ e_2 = 0,\ \pi \ne 0)$, *then* $\rho = \lambda\pi - \frac{1}{2} e_2^2$ *and the infimum is unattainable;*
3. *Otherwise,* $\rho = -\infty$ *and thus problem* (P) *is unbounded from below.*

Proof We consider the problem in the following cases:

- When $\tau_1 = 1$, (8) becomes $z_1 z_2 \le \pi$ and (9) becomes $\lambda z_1 z_2 + \frac{1}{2} z_2^2 + e_1 z_1 + e_2 z_2$. We then have the following cases corresponding to the values of λ, e_1, e_2 and π.
  - When $\lambda > 0$, set $z_1 = -\frac{M}{\lambda} - M$ and $z_2 = M$, where $M \in \mathbb{R}$ is chosen such that $-(1+\frac{1}{\lambda})M^2 \le \pi$. Then $z_1 z_2 \le \pi$ and (9) $= -(\lambda+\frac{1}{2})M^2 - (\frac{e_1}{\lambda}+e_1-e_2)M \to -\infty$ when $M \to \infty$ (case 3)
  - When $\lambda = 0$, we have the following subcases:
    - When $e_1 \neq 0$, set $z_1 = -\text{sign}(e_1)M$ and $z_2 = -\text{sign}(e_1)\frac{\pi}{M}$, where $M \in \mathbb{R}$ and $z_1 z_2 = \pi$. Then $\rho = -|e_1|M + C_1 \cdot \frac{1}{M} + C_2 \cdot \frac{1}{M^2} \to -\infty$ when $M \to +\infty$, where $C_i, i = 1, 2$, are the reduced constants (case 3)
    - When $e_1 = 0$, (9) becomes $\frac{1}{2} z_2^2 + e_2 z_2 = \frac{1}{2}(z_2 + e_2)^2 - \frac{1}{2} e_2^2 \ge -\frac{1}{2} e_2^2 \Rightarrow \rho \ge -\frac{1}{2} e_2^2$.
      - If $e_2 \neq 0$, set $z_1 = -\frac{\pi}{e_2}$ and $z_2 = -e_2$, then $z_1 z_2 = \pi$ and (9) $= -\frac{1}{2} e_2^2$. Thus $\rho = -\frac{1}{2} e_2^2$ and the infimum is attainable (case 1)
      - If $e_2 = 0$ and $\pi \ge 0$, we can set z_1 at any real value and $z_2 = 0$ such that $z_1 z_2 \le \pi$ and thus $\rho = -\frac{1}{2} e_2^2$ and the infimum is attainable (case 1)
      - If otherwise $e_2 = 0$ and $\pi < 0$, we cannot set $z_2 = -e_2 = 0$, which contradicts the constraint $z_1 z_2 \le \pi < 0$. So the infimum is unattainable. But we can set $z_1 = M\pi$ and $z_2 = \frac{1}{M}$ $(M \in \mathbb{R})$ such that $z_1 z_2 = \pi$ and (9) $= \frac{1}{2M^2} \to 0$ when $M \to \infty$. Thus $\rho = \lim\limits_{M \to \infty} \frac{1}{2M^2} - \frac{1}{2} e_2^2 = 0$ but the infimum is unattainable (case 2)
  - When $\lambda < 0$, we have the following subcases:
    - When $e_1 \neq 0$, set $z_1 = -\text{sign}(e_1)M$ and $z_2 = -\frac{\text{sign}(e_1)\pi}{M}$ $(M \in \mathbb{R})$ such that $z_1 z_2 = \pi$ and then (9) $= -|e_1|M + C_1 + C_2 \cdot \frac{1}{M} + C_3 \cdot \frac{1}{M^2} \to -\infty$ when $M \to \infty$, where $C_i,\ i = 1, 2, 3$, are the reduced constants. Thus $\rho = -\infty$ (case 3)
    - When $e_1 = 0$, (9) $= \lambda z_1 z_2 + \frac{1}{2} z_2^2 + e_2 z_2 \ge \lambda\pi + \frac{1}{2}(z_2 + e_2)^2 - \frac{1}{2} e_2^2 \ge \lambda\pi - \frac{1}{2} e_2^2$. Next we show that $\rho = \lambda\pi - \frac{1}{2} e_2^2$. We first note that, to achieve $\lambda z_1 z_2 = \lambda\pi$ in the above inequality, we need to set $z_1 z_2 = \pi$.
      - If $e_2 \neq 0$, set $z_1 = -\frac{\pi}{e_2}$ and $z_2 = -e_2$, such that $z_1 z_2 = \pi$ and then (9) $= \lambda\pi - \frac{1}{2} e_2^2$ (case 1)
      - If $e_2 = 0$ and $\pi \neq 0$, we cannot set $z_2 = -e_2 = 0$, which contradicts the constraint $z_1 z_2 = \pi \neq 0$. So the infimum is unattainable. But we can set $z_1 = M\pi$ and $z_2 = \frac{1}{M}$ $(M \in \mathbb{R})$ such that $z_1 z_2 = \pi$ and (9) $= \lambda\pi + \frac{1}{2M^2} \to \lambda\pi$ when $M \to \infty$. So $\rho = \lambda\pi - \frac{1}{2} e_2^2 = \lambda\pi$ and the infimum is unattainable (case 2)
      - If $e_2 = 0$ and $\pi = 0$, we can set z_1 at any real value and $z_2 = 0$ and thus attain the infimum $\rho = \lambda\pi - \frac{1}{2} e_2^2 = \lambda\pi$ (case 1)
- When $\tau_1 = -1$, set $z_1 = -\frac{\pi}{M}$ and $z_2 = M$ $(M \in \mathbb{R})$ such that $\tau_1 z_1 z_2 = -z_1 z_2 = \pi$ and (9) $= -\frac{1}{2} M^2 + C_1 \cdot M + C_2 \cdot \frac{1}{M} + C_3 \to -\infty$ when $M \to \infty$, where $C_i,\ i = 1, 2, 3$, are the reduced constants. Thus $\rho = -\infty$ (case 3)

Since $\inf\{(9) \mid (8) = \tau_1 z_1 z_2 \leq \pi\}$ is a subproblem of (RP), if there is a feasible solution $(\bar{z}^T, \bar{y}^T)^T$ for (RP) with $\bar{z}_1\bar{z}_2 = \pi$, and $\rho = -\infty$, then $v(\mathrm{P}) = v(\mathrm{RP}) = -\infty$. □

Remark 4 1. Case 2 in the above theorem is the only case where problem (P) is bounded from below but its infimum is unattainable.

2. If two matrices A and D are SD via congruence, which covers most conditions discussed in the existing literature [2,7,16], then the optimal value is attainable when problem (P) is bounded from below.
3. All the bounded cases require $e_1 = 0$ in the linear terms in the objective function associated with 2 × 2 Jordan blocks.

The following theorem is a direct result of Theorems 2–5.

Theorem 6 *If problem* (P) *has an optimal value bounded from below, then:*

1. $\dim E_i \leq 2,\ i = 1, \ldots, p$, $\dim E_i = 1,\ i = p+1, \ldots, m$, *and there is no complex eigenvalue pair in* $J(\kappa_i, n_i)$*;*
2. *If for some index* i, $\dim E_i = 2$, *then the* i*th block satisfies the conditions in either case 1 or case 2 in Theorem 5.*

Note that the conditions in items 1 and 2 of Theorem 6 are necessary for problem (P) to be bounded from below and we assume that these conditions hold in the following discussion of this section. Rearrange the block pairs with single elements to the upper left part of the diagonal in the canonical form and express A and D in the following forms,

$$A = \mathrm{diag}\left(\alpha_1, \ldots, \alpha_l, E_1, \ldots, E_{\frac{n-l}{2}}\right), \tag{10}$$

$$D = \mathrm{diag}(\delta_1, \ldots, \delta_l, E_1 J(\zeta_1, 2), \ldots, E_{\frac{n-l}{2}} J(\zeta_{\frac{n-l}{2}}, 2)), \tag{11}$$

where l is the number of block pairs with size 1 in the canonical form of (4) and (5). In the following, we assume that A and D are in the form of (10) and (11) and further assume $b_i = 0,\ i = l+1, \ldots, n$, as we discussed after (9).

Moreover, from item 3 in Remark 4, we have $e_{l+2j-1} = 0,\ j = 1, \ldots, \frac{n-l}{2}$. Then problem (P) can be reduced to the following form:

$$\begin{aligned}
(\mathrm{P}_1) \quad \min \quad f(x) = & \sum_{i=1}^{l}\left(\frac{1}{2}\delta_i x_i^2 + e_i x_i\right) \\
& + \sum_{j=1,\ldots,\frac{n-l}{2}} \left(\zeta_j x_{l+2j-1} x_{l+2j} + \frac{1}{2}x_{l+2j}^2 + e_{l+2j} x_{l+2j}\right) \\
\text{s.t.} \quad h(x) = & \sum_{i=1}^{l}\left(\frac{1}{2}\alpha_i x_i^2 + b_i x_i\right) + \sum_{j=1,\ldots,\frac{n-l}{2}} (x_{l+2j-1} x_{l+2j}) + c \leq 0.
\end{aligned}$$

Theorem 7 *Assume that items 1 and 2 in Theorem 6 are satisfied, then* $v(\mathrm{P}) = v(\mathrm{P}_1) = v(\mathrm{P}_2)$*, where* (P_2) *is the following SOCP problem:*

$$
\begin{aligned}
(\mathrm{P}_2)\quad \min\quad & \sum_{i=1}^{l}(\delta_i y_i + e_i x_i) + \sum_{j=1}^{\frac{n-l}{2}} \zeta_j z_j + c_0 \\
\text{s.t.}\quad & \sum_{i=1}^{l}(\alpha_i y_i + b_i x_i) + \sum_{j=1}^{\frac{n-l}{2}} z_j + c \le 0, \\
& \frac{1}{2}x_i^2 - y_i \le 0,\ \forall i = 1, 2, \ldots, l, \\
& x, y \in \mathbb{R}^l,\ z \in \mathbb{R}^{\frac{n-l}{2}},
\end{aligned}
$$

where $c_0 = -\sum_{j=1,\ldots,\frac{n-l}{2}} \frac{1}{2}e_{l+2j}^2$.

More specifically, if (P_2) *admits an optimal solution, then there exists an optimal solution* $(\bar{x}, \bar{y}, \bar{z})$ *to* (P_2) *with* $\frac{1}{2}\bar{x}_i^2 = \bar{y}_i,\ i = 1, 2, \ldots, l$. *Moreover, we can find an optimal solution (or an* ϵ *optimal solution)* $\tilde{x}$ *to* (P_1) *with*

$$
\begin{aligned}
&\tilde{x}_i = \bar{x}_i, i = 1, \ldots, l, \\
&\tilde{x}_{l+2j} = \begin{cases} 1/M & \textit{if} \begin{cases} \zeta_j = 0, e_{l+2j} = 0, \bar{z}_j < 0, \\ \textit{or}\ \zeta_j < 0, e_{l+2j} = 0, \bar{z}_j \neq 0, \end{cases} \\ -e_{l+2j} & \textit{otherwise}, \end{cases} \qquad j = 1, \ldots, \tfrac{n-l}{2}, \\
&\tilde{x}_{l+2j-1} = \tfrac{\bar{z}_j}{\tilde{x}_{l+2j}},\ j = 1, \ldots, \tfrac{n-l}{2}.
\end{aligned} \tag{12}
$$

In addition, if (P_1) *is bounded from below, then the optimal value of* (P_1) *is unattainable if and only if* $\zeta_j = 0,\ e_{l+2j} = 0,\ \bar{z}_j < 0$ *or* $\zeta_j < 0,\ e_{l+2j} = 0,\ \bar{z}_j \neq 0$. *In this case, for any* $\epsilon > 0$, *there exists an* ϵ *optimal solution* $\tilde{x}$ *such that* $f(\tilde{x}) - v(\mathrm{P}_1) < \epsilon$ *with a sufficiently large* $M > 0$.

Proof Because of Theorem 6, (P_1) is equivalent to (P). And the main differences between (P_1) and (P_2) are the terms associated to the 2×2 Jordan blocks. Let us consider how to simplify the terms associated with the 2×2 Jordan blocks. According to Assumption 1, (P_1) is feasible. For any feasible solution $\hat{x}$ of (P_1), we let $\pi_j = \hat{x}_{l+2j-1}\hat{x}_{l+2j}$. Now let us concentrate on problem $\inf\{\zeta_j x_{l+2j-1}x_{l+2j} + \frac{1}{2}x_{l+2j}^2 + e_{l+2j}x_{l+2j} \mid x_{l+2j-1}x_{l+2j} = \pi_j\} = \inf\{\zeta_j\pi_j + \frac{1}{2}x_{l+2j}^2 + e_{l+2j}x_{l+2j} \mid x_{l+2j-1}x_{l+2j} = \pi_j\} = \inf\{\zeta_j\pi_j + \frac{1}{2}(x_{l+2j}^2 + e_{l+2j})^2 - \frac{1}{2}e_{l+2j}^2 \mid x_{l+2j-1}x_{l+2j} = \pi_j\}$. Thus setting $x_{l+2j} = -e_{l+2j}$ (if $e_{l+2j} = 0$, set $x_{l+2j} = \frac{1}{M}$ as in the proof of Theorem 5) and $x_{l+2j-1} = \frac{\pi_j}{x_{l+2j}}$, the objective function $(\zeta_j x_{l+2j-1}x_{l+2j} + \frac{1}{2}x_{l+2j}^2 + e_{l+2j}x_{l+2j})$ has an infimum $\zeta_j\pi_j - \frac{1}{2}e_{l+2j}^2$ under the constraint $x_{l+2j-1}x_{l+2j} = \pi_j$, which is linear with the cross term $x_{l+2j-1}x_{l+2j} = \pi_j$.

Using such a separability, we denote

$$
z_j = x_{l+2j-1}x_{l+2j} \text{ and } c_0 = -\sum_{j=1,\ldots,\frac{n-l}{2}} \frac{1}{2}e_{l+2j}^2,
$$

and have the following problem which has the same objective value with (P$_1$):

$$\begin{aligned}(\mathrm{P}_3)\ \min\quad & \sum_{i=1}^{l}\left(\frac{1}{2}\delta_i x_i^2 + e_i x_i\right) + \sum_{j=1}^{\frac{n-l}{2}} \zeta_j z_j + c_0\\ \text{s.t.}\quad & \sum_{i=1}^{l}\left(\frac{1}{2}\alpha_i x_i^2 + b_i^T x_i\right) + \sum_{j=1}^{\frac{n-l}{2}} z_j + c \le 0.\end{aligned}$$

Moreover, if there is an optimal solution $(\bar{x}, \bar{z})$ of (P$_3$), we can also find an optimal solution (or an ϵ optimal solution) $\tilde{x}$ of (P$_1$) in the form of (12). In this case, the optimal value of (P$_1$) is unattainable if and only if $\zeta_j = 0,\ e_{l+2j} = 0, \bar{z}_j < 0$ or $\zeta_j < 0,\ e_{l+2j} = 0,\ \bar{z}_j \neq 0$ from Theorem 5. Furthermore, for any $\epsilon > 0$, if we set $M \ge \sqrt{\frac{1}{2\epsilon}}$, then $f(\tilde{x}) - v(\mathrm{P}_1) = \frac{1}{2M^2} \le \epsilon$.

Introducing $y_i = \frac{1}{2}x_i^2,\ i = 1, 2, \ldots, l$, (P$_3$) is then equivalent to the following (P$_4$):

$$\begin{aligned}(\mathrm{P}_4)\ \ \min\quad & \sum_{i=1}^{l}(\delta_i y_i + e_i x_i) + \sum_{j=1}^{\frac{n-l}{2}} \zeta_j z_j + c_0\\ \text{s.t.}\quad & \sum_{i=1}^{l}(\alpha_i y_i + b_i x_i) + \sum_{j=1}^{\frac{n-l}{2}} z_j + c \le 0,\\ & \frac{1}{2}x_i^2 - y_i = 0,\ \forall i = 1, 2, \ldots, l,\\ & x, y \in \mathbb{R}^l,\ z \in \mathbb{R}^{\frac{n-l}{2}}.\end{aligned}$$

We next prove the equivalence of (P$_4$) and (P$_2$) in two parts:

1. If (P$_2$) is unbounded from below, then (P$_4$) is unbounded from below.
2. If (P$_2$) has an optimal solution (x^*, y^*, z^*), then we can always find a solution $(\bar{x}, \bar{y}, \bar{z})$ with $\bar{y}_i = \frac{1}{2}\bar{x}_i^2,\ i = 1, \ldots, l$ and $\bar{z} = z^*$, which is optimal not only to (P$_2$) but also to (P$_4$).

The first part is proved in the following Lemma 4. Now let us prove part 2. Note that if (P$_2$) is bounded from below, then there must exist an optimal solution (x^*, y^*, z^*) since Slater condition is satisfied. Denote

$$J := \left\{ i : \frac{1}{2}\left(x_i^*\right)^2 < y_i^*,\ i = 1, \ldots, l \right\}.$$

If $J = \emptyset$, then (x^*, y^*, z^*) is also an optimal solution of (P$_4$). If $J \neq \emptyset$, by Theorem 3 in [2], we can transform the optimal solution (x^*, y^*, z^*) of (P$_2$) to an optimal solution $(\bar{x}, \bar{y}, z^*)$ of (P$_2$) with $\bar{y}_i = \frac{1}{2}\bar{x}_i^2, i = 1, \ldots, l$, and $(\bar{x}, \bar{y}, z^*)$ is also a feasible solution of (P$_4$), since $\bar{y}_i = \frac{1}{2}\bar{x}_i^2, i = 1, \ldots, l$. So $v(\text{P}_2) \geq v(\text{P}_4)$. But (P$_2$) is a relaxation of (P$_4$), so $v(\text{P}_2) \leq v(\text{P}_4)$. Thus $v(\text{P}_2) = v(\text{P}_4)$ and $(\bar{x}, \bar{y}, z^*)$ is optimal to (P$_4$). □

Lemma 4 *If* (P$_2$) *is unbounded from below, then* (P$_4$) *is unbounded from below.*

Proof We only need to prove that (P$_3$) (since (P$_4$) is equivalent to (P$_3$)) is bounded from below implies that (P$_2$) is bounded from below.

In this proof, we only consider the cases with no z term in (P$_2$) and (P$_3$), since z only appears in the linear terms in both the objective and constraint functions, which can be regarded as a special case of the x variable (i.e., the coefficients before z_j^2 are $0, j = 1, \ldots, \frac{n-l}{2}$).

Denote the Lagrangian function of (P$_3$) as $L(x, \nu) = f(x) + \nu h(x)$, and the dual function as $\theta(\nu) = \min\limits_x L(x, \nu)$, where $\nu \geq 0$. If (P$_3$) is bounded from below, from the S-lemma (as Slater condition holds here), we know there is no duality gap between the primal problem (P$_3$) and its Lagrangian dual problem of $\max\limits_{\nu \geq 0} \theta(\nu)$, i.e., there exists $(\bar{x}, \bar{\nu})$ such that $\min\limits_{h(x) \leq 0} f(x) = f(\bar{x}) = \theta(\bar{\nu}) = \max\limits_{\nu \geq 0} \theta(\nu)$. So $(\bar{x}, \bar{\nu})$ is a saddle point of the Lagrangian function $L(x, \nu)$. Then $f(\bar{x}) = \min\limits_{x \in \mathbb{R}^n} L(x, \bar{\nu})$, $h(\bar{x}) \leq 0$, $\bar{\nu} \geq 0$, $\bar{\nu} h(\bar{x}) = 0$. From $\min\limits_{x \in \mathbb{R}^n} L(x, \bar{\nu}) = \min\limits_{x \in \mathbb{R}^n} \sum_{i=1}^{l} (\frac{1}{2}(\delta_i + \bar{\nu}\alpha_i)x_i^2 + (e_i + \bar{\nu} b_i)x_i) + c_0 + \bar{\nu} c = f(\bar{x})$, we get $\delta_i + \bar{\nu}\alpha_i \geq 0$ and $(\delta_i + \bar{\nu}\alpha_i)\bar{x}_i + (e_i + \bar{\nu} b_i) = 0$ and if, in addition, $\delta_i + \bar{\nu}\alpha_i = 0$, we have $e_i + \bar{\nu} b_i = 0$, $i = 1, \ldots, l$. So $(\bar{x}, \bar{\nu})$ satisfies the KKT conditions of (P$_3$), i.e., $(\delta_i + \bar{\nu}\alpha_i)\bar{x}_i + e_i + \bar{\nu} b_i = 0, i = 1, \ldots, l, \bar{\nu} \geq 0, \bar{\nu} h(\bar{x}) = 0$.

Next we can construct a KKT point of (P$_2$) from the saddle point $(\bar{x}, \bar{\nu})$. Define $\bar{y}_i = \frac{1}{2}\bar{x}_i^2$ and $\bar{\mu}_i = \delta_i + \bar{\nu}\alpha_i \geq 0, i = 1, \ldots, l$. Then $(\bar{x}, \bar{y}, \bar{\mu}, \bar{\nu})$ satisfies the KKT condition of (P$_2$): $\delta_i + \bar{\nu}\alpha_i - \bar{\mu}_i = 0, e_i + \bar{\nu} b_i + \bar{\mu}_i \bar{x}_i = 0, \bar{\mu}_i(\frac{1}{2}\bar{x}_i^2 - \bar{y}_i) = 0$, $i = 1, \ldots, l, \bar{\nu}(\sum_{i=1}^{l}(\alpha_i \bar{y}_i + b_i \bar{x}_i) + c) = 0$. Thus $(\bar{x}, \bar{y})$ is a global optimal solution of (P$_2$) because of the convexity of (P$_2$). So we conclude that (P$_2$) is bounded from below. □

Theorem 7 shows us that problem (P) is equivalent to a convex quadratic problem. It is well known that convex quadratic form is in fact SOCP representable (see page 96, [3]), and thus problem (P) is further equivalent to an SOCP problem. More specifically, when we decompose Q as $Q = P^T P$, $x^T Q x + p^T x + q \leq 0$ is equivalent to $\left\| \begin{pmatrix} Px \\ \frac{p^T x + q + 1}{2} \end{pmatrix} \right\| \leq \frac{1 - (p^T x + q)}{2}$.

Further investigation suggests that we can obtain a closed-form solution for Problem (P$_2$) when there exists a 2×2 block pair in the canonical form.

Proposition 1 *Problem* (P$_2$) *has a closed-form solution when $l < n$, i.e., there exists at least one 2×2 block in the canonical form of* (4) *and* (5)*:*

1. *If $\zeta_i s$ are not equal, $v(\text{P}_2) = -\infty$;*
2. *If $\zeta_1 = \zeta_2 = \cdots = \zeta_p > 0$, $v(\text{P}_2) = -\infty$;*
3. *If $\zeta_1 = \zeta_2 = \cdots = \zeta_p = 0$, problem* (P$_2$) *admits a closed-form solution. If $\delta_i > 0$ or $\delta_i = e_i = 0$, $i = 1, \ldots, \frac{n-l}{2}$, by defining $\frac{0}{0} = 0$, we have $x_i = -\frac{e_i}{\delta_i}$, $y_i =$*

$\frac{1}{2}x_i^2$, $i = 1, \ldots, \frac{n-l}{2}$. With properly choosing z satisfying the constraint, we have $v(\mathrm{P}_2) = -\sum_i^l \frac{e_i^2}{2\delta_i} + c_0$. Otherwise $v(\mathrm{P}_2) = -\infty$.

4. *If $\zeta_1 = \zeta_2 = \cdots = \zeta_p < 0$, problem* ($\mathrm{P}_2$) *admits a closed-form solution. If $\delta_i - \zeta_1\alpha_i > 0$ or $\delta_i - \zeta_1\alpha_i = e_i - \zeta_1 b_i = 0$, $i = 1, \ldots, \frac{n-l}{2}$, we have $x_i = -\frac{e_i - \zeta_1 b}{\delta_i - \zeta_1\alpha_i}$, $y_i = \frac{1}{2}x_i^2$, $i = 1, \ldots, \frac{n-l}{2}$. With properly choosing z satisfying the constraint, we have $v(\mathrm{P}_2) = -\sum_i^l \frac{(e_i - \zeta_1 b_i)^2}{2(\delta_i - \zeta_1\alpha_i)} + c_0 - \zeta_1 c$. Otherwise $v(\mathrm{P}_2) = -\infty$.*

Moreover, we can get a closed-form solution to problem (P) *from a solution to problem* (P_2).

Proof We prove this proposition for the following four cases.

1. If ζ_is are not equal, problem (P_2) is a linear problem in terms of z with one linear constraint, where the parameters before z in the objective are different but the parameters before z in the constraint are the same. This yields $v(\mathrm{P}_2) = -\infty$.
2. If $\zeta_1 = \zeta_2 = \cdots = \zeta_p > 0$, for any feasible solution (x, y, z), letting $z_j \to -\infty$, $j = 1, \ldots, \frac{n-l}{2}$, still keeps the feasibility, thus yielding $v(\mathrm{P}_2) = -\infty$.
3. If $\zeta_1 = \zeta_2 = \cdots = \zeta_p = 0$, the linear constraint is redundant. If $\delta_i > 0$ or $\delta_i = e_i = 0$, $i = 1, \ldots, \frac{n-l}{2}$, by defining $\frac{0}{0} = 0$, problem (P_2) then admits a closed-form solution $x_i = -\frac{e_i}{\delta_i}$, $y_i = \frac{1}{2}x_i^2$, $i = 1, \ldots, \frac{n-l}{2}$. By properly choosing z such that the constraint is satisfied, we have $f(x, z) = -\sum_i^l \frac{e_i^2}{2\delta_i} + c_0$. Otherwise $v(\mathrm{P}_2) = -\infty$.
4. If $\zeta_1 = \zeta_2 = \cdots = \zeta_p < 0$, for any optimal solution (x, y, z), the linear constraint must take equality, i.e., $\sum_{i=1}^l(\alpha_i y_i + b_i x_i) + \sum_{j=1}^{\frac{n-l}{2}} z_j + c = 0$. Otherwise, enlarging z such that the linear constraint takes equality will lead to a smaller objective value, which contradicts the optimality. So we can add $-\zeta_1[\sum_{i=1}^l(\alpha_i y_i + b_i x_i) + \sum_{j=1}^{\frac{n-l}{2}} z_j + c]$ to the objective, which yields $\sum_{i=1}^l[(\delta_i - \alpha_i\zeta_1)y_i + (e_i - \zeta_i b_i)x_i] + c_0 - \zeta_1 c$, and the linear constraint is then redundant since we can set z at any value that makes the constraint satisfied. Now the case reduces to case 3 and we have a closed-form solution. □

Example 1 Consider the following problem:

$$\begin{aligned} \min \quad & -x_1x_2 + 0.5x_2^2 - x_3^2 + x_4^2 + 2x_2 - x_4 \\ \text{s.t.} \quad & x_1x_2 + x_3^2 + 0.75x_4^2 \le 1.25, \end{aligned}$$

where the related matrices can be expressed as

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1.5 \end{pmatrix}, \quad D = \begin{pmatrix} 0 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix},$$

$e = (0, 2, 0, -1)^T$, $b = 0$ and $c = -1.25$. Note that A and D are not SD, but in the canonical form in (4) and (5). According to Theorem 7, we get the following equivalent convex quadratic problem reformulation,

$$\begin{aligned} \min \quad & -z_1 - 2y_3 + 2y_4 - x_4 - 2 \\ \text{s.t.} \quad & z_1 + 2y_3 + 1.5y_4 \le 1.25 \\ & \frac{1}{2}x_3^2 - y_3 \le 0 \\ & \frac{1}{2}x_4^2 - y_4 \le 0. \end{aligned}$$

1. Solving the above convex quadratic problem via any standard solver (we here use Sedumi in CVX), yields the optimal solution $z_1^* = -12.9773$, $x_3^* = 0$, $x_4^* = 0.2857$, $y_3^* = 7.0830$ and $y_4^* = 0.0408$. Note $\frac{1}{2}(x_3^*)^2 - y_3^* = -7.0830 < 0$. Using the transformation method in Theorem 3 in [2], we obtain a new solution $(\bar{x}, \bar{y}, \bar{z})$, with $\bar{x}_3 = \pm\sqrt{2y_3^*} = \pm 3.7638$, $\bar{z}_1 = z_1^*$, $\bar{x}_3 = x_3^*$, $\bar{y}_3 = y_3^*$ and $\bar{y}_4 = y_4^*$. By applying Theorem 7, we get $\tilde{x}_2 = -2$, $\tilde{x}_1 = \frac{\bar{z}_1}{\tilde{x}_2} = 6.4886$, $\tilde{x}_3 = \bar{x}_3$ and $\tilde{x}_4 = \bar{x}_4$. So we obtain an optimal solution $\tilde{x} = (6.4886, -2, \pm 3.7638, 0.2857)^T$ to the origin problem, with an optimal value -3.3929.
2. We can also apply case 4 in Proposition 1 to achieve a closed-form solution for Example 1. Since $\zeta_1 = -1$ determines the Lagrangian multiplier associated with the first constraint as 1, the Lagrangian associated with the first constraint becomes

$$\begin{aligned} &(-z_1 - 2y_3 + 2y_4 - x_4 - 2) + (z_1 + 2y_3 + 1.5y_4 - 1.25) \\ &\quad = 3.5y_4 - x_4 - 3.25. \end{aligned}$$

   Then we obtain a closed-form solution $x_3 = 0$, $x_4 = \frac{2}{7}$, $y_4 = \frac{2}{49}$ and the objective value is $-\frac{95}{28} = -3.3929$. Set z_1 such that $z_1 + 2y_3 + 1.5y_4 - 1.25 = 0$, i.e., $z_1 = \frac{233}{196}$. Setting $x_2 = -2$ gives rise to $x_1 = \frac{z_1}{x_2} = -\frac{233}{392}$. It is easy to check that $x = \left(-\frac{233}{392}, -2, 0, \frac{2}{7}\right)$ is another optimal solution.

2.4 Solving the GTRS without computing the canonical form

Although the canonical form provides rich structural information for solving problem (P), unfortunately, there is no numerically stable method in the current literature to compute the canonical form. In fact, there is also no stable algorithm to compute the canonical form in [14]. This fact is consistent with the nonexistence of stable Jordan decomposition methods, see Chapter 7 in [8,13]. Large Jordan blocks are difficult to handle due to an unstableness of computing Jordan blocks. On the other hand, problem (P) itself also has some problematic cases, i.e., a small perturbation of the matrices may yield a significant change of the optimal value, see [19], which, we believe, is closely related to the unstableness of the Jordan decomposition methods. Our analysis above sheds some light on the relationship between the unstable cases of the canonical form with $k \times k$ blocks with $k \ge 2$ (or the Jordan normal form with $k \times k$ blocks with $k \ge 2$) and the problematic cases of problem (P).

Although we do not have a numerically stable algorithm to compute the canonical form, we are able to find an alternative way to utilize the structural information provided by the canonical form and solve the GTRS in a stable way. Define $I_{PSD} = \{\lambda : \ D + \lambda A \succeq 0, \ \lambda \geq 0\}$. It is shown in [16] that I_{PSD} is either an interval or a singleton. Although the SD case is more general than the case where I_{PSD} is an interval, it is hard to simultaneously diagonalize the two matrices when I_{PSD} is a singleton. Due to Lemma 3, GTRS is bounded from below only if I_{PSD} is nonempty. Hence, we solve the GTRS in the following two cases: (i) I_{PSD} is an interval and (ii) I_{PSD} is a singleton.

Let us first consider the case where I_{PSD} is an interval. First let us assume that (A, D) is a nonsingular matrix pair. In this case, simultaneous diagonalization of A and D can be done in the following way: (i) First find a λ such that $A + \lambda D \succ 0$; Then apply Cholesky decomposition to obtain L such that $L^T L = A + \lambda D$ (So $(L^{-1})^T(A + \lambda D)L^{-1} = I$); (ii) Apply spectral decomposition to $(L^{-1})^T DL$ with some orthogonal matrix P, which makes both $P^T(L^{-1})^T DLP$ and $P^T(L^{-1})^T ALP = I - \lambda P^T(L^{-1})^T DLP$ diagonal. A special case of the above algorithm for the case where either A or D is positive definite can be found in Chapter 11.6.5 in [6] and [8,21]. To this end, we only need to consider the case that I_{PSD} is an interval and (A, D) is a singular matrix pair. Let $U \in \mathbb{R}^{n\times m}$ be the common null space of A and D, and $U_\perp$ be the orthogonal complement of U in $\mathbb{R}^n$. Then there exists a λ with $\lambda \geq 0$ such that $U_\perp^T(D + \lambda A)U_\perp \succ 0$. Similar to the nonsingular matrix pencil case, we can compute the congruent matrix $R \in \mathbb{R}^{(n-m)\times(n-m)}$ such that $R^T U_\perp^T DU_\perp R$ and $R^T U_\perp^T AU_\perp R$ are both diagonal. Moreover, $Q = [U \ U_\perp]\mathrm{diag}(I_m, R)$ makes both $Q^T AQ$ and $Q^T DQ$ diagonal and thus problem (P) can be solved by the SOCP reformulation in [2]. Our numerical tests in Sect. 4 show that the SOCP reformulation (with the cost of the SD approach taken into consideration) is much more efficient than the SDP reformulation.

Now we consider the case where I_{PSD} is a singleton. Note that the singleton case includes the non-SD case as a special case. Our goal for the singleton case is then to find a stable method to solve problem (P) to overcome the unstableness of computing the canonical form. Without loss of generality, we assume that $D+\lambda A$ forms a nonsingular matrix pencil for some $\lambda \in \mathbb{R}$. Otherwise, if (7) is violated, problem (P) is unbounded from below; if (7) is satisfied, similar to the interval case, problem (P) can be reduced to the the following reformulation,

$$\begin{aligned} \min \quad & \frac{1}{2}y^T U_\perp^T DU_\perp y + e^T U_\perp y + e^T Uz \\ \text{s.t.} \quad & \frac{1}{2}y^T U_\perp^T AU_\perp y + b^T U_\perp y + b^T Uz + c \leq 0, \end{aligned}$$

where $U \in \mathbb{R}^{n\times m}$ is the common null space of A and D, and $U_\perp$ is the orthogonal complement of U in $\mathbb{R}^n$. A closed form solution for this reformulation is given as $y = (U_\perp^T(D + \lambda A)U_\perp)^{-1}(e^T U_\perp + \lambda b^T U_\perp)$ and any $z \in \mathbb{R}^m$ such that the constraint being active.

In the following theorem, provided that $I_{PSD} = \{\lambda\}$ is a singleton, we derive a closed-form solution if the optimal objective value is attainable or an asymptotic solution if the optimal value is unattainable.

Theorem 8 *Assume that $D + \lambda A$ forms a nonsingular matrix pencil for some $\lambda \in \mathbb{R}$, $I_{PSD} = \{\lambda\}$ is a singleton, λ is the solution to system* (7) *and v^* is the optimal value of problem* (P). *Define a candidate solution $x^* = (D + \lambda A)^+(e + \lambda b)$. Define a matrix $V \in \mathbb{R}^{n\times m}$ $(n > m)$ such that $V = \text{null}(D + \lambda A)$ and $V_\perp$ is the orthogonal complement of V in $\mathbb{R}^n$. We can derive an optimal solution of problem (P) in the following two cases:*

1. *If $h(x^*) \le 0$, x^* is already an optimal solution to* (P).
2. *When $h(x^*) > 0$, we can solve problem (P) in the following cases:*
   (a) *$V^T AV$ has a negative eigenvalue. Let p be an eigenvector corresponding to an arbitrary negative eigenvalue. Then the quadratic equation*

$$h(x^* + \theta Vp) = \frac{1}{2}p^T V^T AVp\theta^2 + (Ax^* + b)^T Vp\theta + h(x^*) = 0$$

   has a solution $\tilde{\theta}$, and $\bar{x} = x^* + \tilde{\theta}Vp$ is an optimal solution to (P)*;*
   (b) *$V^T AV \succeq 0$ and there exists a nonzero vector $p \in$ null$(V^T AV)$ satisfying $(Ax^* + b)^T Vp \ne 0$. Then $\bar{x} = x^* + \tilde{\theta}Vp$ is an optimal solution to* (P), *where $\tilde{\theta}$ is a solution of the linear equation*

$$h\left(x^* + \theta Vp\right) = h\left(x^*\right) + \theta\left(Ax^* + b\right)^T Vp = 0.$$

   (c) *Otherwise $V^T AV \succeq 0$ and $(Ax^* + b)^T VK = 0$, where $K =$ null$(V^T AV)$. Then we can always identify a vector $p \in$ null$(V^T AV)$ such that $AVp \ne 0$. Then for any $\epsilon > 0$, there exists a $\varpi > 0$ and $q_1 \in \mathbb{R}^m$ such that there exists $\bar{x} = x^* + \frac{1}{\varpi}Vp + \varpi V_\perp q_1$ satisfying $h(\bar{x}) = 0$ and $f(\bar{x}) \le v^* + \epsilon$.*

Proof We only need to prove the three subcases in case 2. Let us define $g(x) = \frac{1}{2}x^T(D + \lambda A)x + (e + \lambda b)^T x + \lambda c$ and

$$\text{(G)} \quad \min g(x) \quad \text{s.t. } \frac{1}{2}x^T Ax + b^T x + c \le 0.$$

Note that $v(\text{G}) \le v(\text{P})$ since $\frac{1}{2}x^T Ax + b^T x + c \le 0$ and $\lambda \ge 0$. We also have $(e + \lambda b)^T V = 0$, otherwise (P) is unbounded from below due to Lemma 3. Now we are ready to prove all the three subcases.

(a) From the assumption, we know that $(D + \lambda A)Vp = 0$ and $p^T V^T AVp < 0$. The equation

$$h(x^* + \theta Vp) = \frac{1}{2}p^T V^T AVp\theta^2 + (Ax^* + b)^T Vp\theta + h(x^*) = 0$$

has two solutions for θ since $p^T V^T AVp < 0$ and $h(x^*) > 0$. With the definition of $\tilde{\theta}$, we know $x^* + \tilde{\theta}Vp$ is feasible to problem (P). Then due to $f(x^* + \tilde{\theta}Vp) =$

$g(x^*+\tilde{\theta}Vp)-\lambda h(x^*+\tilde{\theta}Vp)=g(x^*)$, we conclude that $x^*+\tilde{\theta}Vp$ is an optimal solution to both problems (G) and (P) since $h(x^*+\tilde{\theta}Vp)=0$.

(b) The equation $h(x^*+\theta Vp)=h(x^*)+(Ax^*+b)^TVp\theta=0$ is a linear equation of θ because $V^TAVp=0$. The parameter before the linear term, $(Ax^*+b)^TVp$, is not equal to zero and thus the linear equation admits a unique solution. Then due to $g(x^*+\tilde{\theta}p)=g(x^*)$, $x^*+\tilde{\theta}Vp$ is still optimal to (G). So by the definition of $\tilde{\theta}$, we conclude $x^*+\tilde{\theta}Vp$ is feasible and thus optimal to both problems (G) and (P) since $h(x^*+\tilde{\theta}Vp)=0$.

(c) Let R_1 and R_2 be the canonical forms of D and A in (4) and (5), respectively. Then $S^T(D+\lambda A)S=R_1+\lambda R_2$. Because the matrix pencil $D+\lambda A$ is nonsingular for some $\lambda\in\mathbb{R}$, there are no type C block pairs (common 0s) in the canonical forms. Since I_{PSD} is a singleton, $V^TAV\succeq 0$ implies that V^TAV has at least one zero eigenvalue. Otherwise $V^TAV\succ 0$, and from the canonical form, we know $R_1+(\lambda+\epsilon)R_2\succeq 0$ for some sufficiently small $\epsilon>0$, which contradicts $I_{PSD}=\{\lambda\}$. From Theorem 1, we know that there can only exist 1 × 1 or 2 × 2 block pairs. Together with the fact that V^TAV has at least one zero eigenvalue and nonexistence of type C block pair, we know there must exist at least one 2 × 2 block pair $(R_1^i,R_2^i)=\left(\begin{bmatrix}0&-\lambda\\-\lambda&1\end{bmatrix},\begin{bmatrix}0&1\\1&0\end{bmatrix}\right)$ and $R_1^i+\lambda R_2^i=\begin{bmatrix}0&0\\0&1\end{bmatrix}\succeq 0$. Let $z_0=\begin{pmatrix}1\\0\end{pmatrix}$. Then we have $z_0\in\text{null}(R_1^i+\lambda R_2^i)$, $z_0^TR_2^iz_0=0$ and $R_2^iz_0=\begin{pmatrix}0\\1\end{pmatrix}$. This implies that there exists a vector $p\in\text{null}(V^TAV)$ such that $AVp\neq 0$. Due to $p\in\text{null}(V^TAV)$, for any $q_1\in\mathbb{R}^m$, we have

$$p^TV^TAV_\perp q_1=p^TV^TAV_\perp q_1+p^TV^TAVq_2=p^TV^TA[V_\perp\ V]\begin{pmatrix}q_1\\q_2\end{pmatrix}$$

for any $q_2\in\mathbb{R}^{n-m}$. Define α with $\alpha_j=-\frac{h(x^*)+1}{(p^TVA)_j}$, and $\alpha_i=0$ for all $i\neq j$, where j is an arbitrary index such that $(p^TV^TA)_j\neq 0$. Letting $q=\begin{pmatrix}q_1\\q_2\end{pmatrix}=[V_\perp\ V]^{-1}\alpha$, we have

$$p^TV^TAV_\perp q_1=p^TV^TA\alpha=(p^TV^TA)_j\alpha_j=-(h(x^*)+1). \qquad (13)$$

Let $\bar{x}=x^*+\frac{1}{\varpi}Vp+\varpi V_\perp q_1$. Then for any $\epsilon>0$, due to $V=\text{null}(D+\lambda A)$, there exists a $\varpi_1>0$ such that the following inequality holds,

$$\begin{aligned}g(\bar{x})&=g(x^*)+\varpi^2(V_\perp q_1)^T(D+\lambda A)V_\perp q_1\\&\quad+\frac{1}{\varpi}[(D+\lambda A)x^*+(e+\lambda b)]^TVp+\varpi[(D+\lambda A)x^*+(e+\lambda b)]^TV_\perp q_1\\&\leq g(x^*)+\epsilon,\end{aligned}$$

for all $0<\varpi<\varpi_1$ (note that $(e+\lambda b)^TVp=0$ since $e+\lambda b\in\text{Range}(D+\lambda A)$). Meanwhile there exists a $\varpi_2>0$ such that the following inequality holds for all $0<\varpi<\varpi_2$,

$$\begin{aligned}h(\bar{x}) &= h\left(x^*\right) + \frac{1}{2\varpi^2}p^T V^T A V p + p^T V^T A V_\perp q_1 + \frac{1}{2}\varpi^2 q_1^T V_\perp^T A V_\perp q_1\\&\quad + \frac{1}{\varpi}\left(Ax^* + b\right)^T Vp + \varpi\left(Ax^* + b\right)^T V_\perp q_1\\&= h\left(x^*\right) + \left(-h\left(x^*\right) - 1\right) + \frac{1}{2}\varpi^2 q_1^T V_\perp^T A V_\perp q_1 + \varpi\left(Ax^* + b\right)^T V_\perp q_1\\&\le 0,\end{aligned}$$

which implies that $\bar{x}$ is feasible. In the above inequality, the term $p^T V^T A V p$ vanishes because $p \in \text{null}(V^T A V)$, $(Ax^* + b)^T Vp = 0$ as assumed, and $p^T V^T A V_\perp q_1 = -h(x^*) - 1$ due to (13). We complete the proof. □

Remark 5 When the dimension of $\text{null}(V^T A V)$ is k, we can conclude that there are k 2×2 blocks in the canonical form associated with the same type of block pairs. Moreover, if the k 2×2 blocks associate with type A block pairs, the associated $\kappa_i = \lambda$ for all i in (5).

It is interesting to notice that a sufficient condition, which is similar to case 2 in Theorem 8, for A and D to be SD is shown in Corollary 3.8 in [11], i.e., "$Q(A) \cap Q(D) = N(A) \cap N(D)$, with $\dim\{N(A) \cap N(D)\} \ne n - 2$", where $Q(A) = \{x : x^T Ax = 0\}$ and $N(A) = \{x : Ax = 0\}$. In fact, removing $\dim\{N(A) \cap N(D)\} \ne n - 2$, the above condition is still sufficient.

Proposition 2 *If $Q(A) \cap Q(D) = N(A) \cap N(D)$, then A and D are SD.*

Proof Let us first consider a case where A and D are not SD, i.e., there exist $E_k \in \mathbb{R}^{k \times k}$ and $E_k J_k \in \mathbb{R}^{k \times k}$ with $k \ge 2$ in the corresponding canonical form. Then define a k dimensional vector $v = (1\ 0\ \dots\ 0)^T$. It is easy to verify that $v^T E_i v = 0$ and $v^T E_i J_i v = 0$. This contradicts $Q(A) \cap Q(D) = N(A) \cap N(D)$ since $N(E_i) = \{0\}$. □

Remark 6 In fact, A and D may still be SD if $Q(A) \cap Q(D) \ne N(A) \cap N(D)$. A counter example is as follows: a nonsingular matrix pencil $A = \text{diag}(\alpha)$ and $D = \text{diag}(\delta)$ with $\alpha, \delta \in \mathbb{R}^k$, and the following linear system has a nonzero solution x^*,

$$\alpha^T x = 0, \quad \delta^T x = 0, \quad x \ge 0.$$

Then we have a nonzero $y^* \in Q(A) \cap Q(D)$ with $y_i = \sqrt{x_i}$, $i = 1, \dots, k$. On the other hand, we have $N(A) \cap N(D) = \{0\}$ since the matrix pencil is nonsingular. So we have $N(A) \cap N(D) \ne Q(A) \cap Q(D)$. A specifical example is: $A = \text{diag}(1, -1)$, $D = \text{diag}(-1, 1)$. And $Q(A) \cap Q(D) = \{t(1, 1), t \in \mathbb{R}\} \ne N(A) \cap N(D) = \{0\}$.

3 Extension to equality constrained and interval bounded variants of GTRS

This section extends the usage of the canonical form and the SOCP reformulation to the equality constrained problem (EP) and the interval bounded problem (IP).

3.1 GTRS with equality constraint

With the same notations as in Sect. 2, Theorems 2, 3 and 4 still hold here, which can be proved in a similar way by the S-lemma with equality [26]. However, Theorem 5 needs some modifications. In the following of this section, we still use (8) and (9) to denote the associated terms in both the constraint and objective functions, i.e., (8) $= \frac{1}{2}\tau_1 z^T E_I z = \tau_1 z_1 z_2 = \pi$ and (9) $= \tau_1 \lambda z_1 z_2 + \frac{1}{2}\tau_1 z_2^2 + e_1 z_1 + e_2 z_2$.

Theorem 9 *Consider the case where there exists a type A block pair* $(\tau_1 e_1, \tau_1 e_1 J_1(\lambda, 2))$ *in problem* (P) *and the eigenvalue of the associated Jordan block* $J_1(\lambda, 2)$ *is real. Assume there is a feasible solution* $\bar{x} = (\bar{z}^T, \bar{y}^T)^T$ *and let* $\pi = \tau_1 \bar{z}_1 \bar{z}_2$. *Let* $\rho = \inf\{$ (9) $\mid$ (8) $= \tau_1 z_1 z_2 = \pi\}$. *We have the following three cases:*

1. *When* $\tau_1 = 1$. *If* $(e_1 = 0, e_2 \neq 0)$ *or* $(e_1 = 0, e_2 = 0, \pi = 0)$, *then* $\rho = \lambda\pi - \frac{1}{2}e_2^2$ *and the infimum is attainable;*
2. *When* $\tau_1 = 1$. *If* $e_1 = 0, e_2 = 0, \pi \neq 0$, *then* $\rho = \lambda\pi - \frac{1}{2}e_2^2$ *and the infimum is unattainable;*
3. *Otherwise,* $\rho = -\infty$ *and thus* (EP) *is unbounded from below.*

Proof The proof is similar to that of Theorem 5. □

Theorem 10 *If the optimal value of problem* (EP) *is bounded from below, then:*

1. $\dim E_i \leq 2$, $i = 1, \ldots, p$, $\dim E_i = 1$, $i = p+1, \ldots, m$, *and there is no complex eigenvalue pair in* $J(\kappa_i, n_i)$;
2. *If for some index* i, $\dim E_i = 2$, *then the* i*th block satisfies case 1 or case 2 in Theorem* 9.

Note that the conditions in items 1 and 2 of Theorem 10 are necessary for problem (EP) to be bounded from below and we assume that the conditions hold in the following of this section. In the same way as the method in solving problem (P), we can then assume that A and D have the form in (10) and (11), $b_j = 0$, for $j = l+1, \ldots, n$, and $e_{l+2j-1} = 0,\ j = 1, \ldots, \frac{n-l}{2}$.

Similarly to Theorem 7, using the S-lemma with equality [26] under Assumption 2, we have the following theorem.

Theorem 11 *Assume that items 1 and 2 in Theorem* 10 *are satisfied, then problem* (EP) *has the same optimal value with the following SOCP reformulation:*

$$
\begin{aligned}
(\mathrm{EP}_1)\quad \min\quad & \sum_{i=1}^{l}(\delta_i y_i + e_i x_i) + \sum_{j=1}^{\frac{n-l}{2}} \zeta_j z_j + c_0 \\
\text{s.t.}\quad & \sum_{i=1}^{l}(\alpha_i y_i + b_i x_i) + \sum_{j=1}^{\frac{n-l}{2}} z_j + c = 0 \\
& \frac{1}{2}x_i^2 - y_i \leq 0,\ \forall i = 1, 2, \ldots, l, \\
& x, y \in \mathbb{R}^l,\ z \in \mathbb{R}^{\frac{n-l}{2}},
\end{aligned}
$$

where $c_0 = -\sum_{j=1,\ldots,\frac{n-l}{2}} \frac{1}{2} e_{l+2j}^2$.

More specifically, if (EP$_1$) *admits an optimal solution, then there exists an optimal solution* $(\bar{x}, \bar{y}, \bar{z})$ *to* (EP$_1$) *with* $\frac{1}{2}\bar{x}_i^2 = \bar{y}_i,\ i = 1, 2, \ldots, l$. *Moreover, we can find an optimal solution (or an* ϵ *optimal solution)* $\tilde{x}$ *to* (EP) *with*

$$\begin{aligned}
&\tilde{x}_i = \bar{x}_i,\ i = 1, \ldots, l,\\
&\tilde{x}_{l+2j} = \begin{cases} 1/M & \text{if } e_{l+2j} = 0,\ \bar{z}_j \neq 0, \\ -e_{l+2j} & \text{otherwise}, \end{cases} \qquad j = 1, \ldots, \tfrac{n-l}{2},\\
&\tilde{x}_{l+2j-1} = \frac{\bar{z}_j}{\tilde{x}_{l+2j}},\ j = 1, \ldots, \tfrac{n-l}{2}.
\end{aligned}$$

In addition, if (EP) *is bounded from below, the optimal value of* (EP) *is unattainable if and only if* $e_{l+2j} = 0$ *and* $\bar{z}_j \neq 0$. *In this case, for any* $\epsilon > 0$, *there exists an* ϵ *optimal solution* $\tilde{x}$ *such that* $f(\tilde{x}) - v(\text{EP}) < \epsilon$ *with a sufficient large* $M > 0$.

3.2 GTRS with interval constraint

Theorem 10 holds in this case and thus we can still assume, without loss of generality, A and D have the form in (10) and (11), $b_j = 0$, for $j = l+1, \ldots, n$, and $e_{l+2j-1} = 0,\ j = 1,\ \ldots,\ \frac{n-l}{2}$.

Theorem 12 *Assume that the conditions in items 1 and 2 in Theorem* 10 *are satisfied, problem* (IP) *has the same optimal value with the following SOCP problem:*

$$\begin{aligned}
(\text{IP}_1)\quad \min\quad & \sum_{i=1}^{l} (\delta_i y_i + e_i x_i) + \sum_{j=1}^{\frac{n-l}{2}} \zeta_j z_j + c_0\\
\text{s.t.}\quad & \bar{h}(x, y, z) = \sum_{i=1}^{l} (\alpha_i y_i + b_i x_i) + \sum_{j=1}^{\frac{n-l}{2}} z_j \geq c_1\\
& \bar{h}(x, y, z) = \sum_{i=1}^{l} (\alpha_i y_i + b_i x_i) + \sum_{j=1}^{\frac{n-l}{2}} z_j \leq c_2\\
& \frac{1}{2} x_i^2 - y_i \leq 0,\ \forall i = 1, 2, \ldots, l,\\
& x, y \in \mathbb{R}^l,\ z \in \mathbb{R}^{\frac{n-l}{2}},
\end{aligned}$$

$c_0 = -\sum_{j=1,\ldots,\frac{n-l}{2}} \frac{1}{2} e_{l+2j}^2$.

More specifically, if (IP$_1$) *admits an optimal solution, then there exists an optimal solution* $(\bar{x}, \bar{y}, \bar{z})$ *to* (IP$_1$) *with* $\frac{1}{2}\bar{x}_i^2 = \bar{y}_i,\ i = 1, 2, \ldots, l$. *Moreover, we can find an optimal solution (or an* ϵ *optimal solution) to (IP) with*

$$\begin{aligned}
&\tilde{x}_i = \bar{x}_i,\ i = 1, \ldots, l,\\
&\tilde{x}_{l+2j} = \begin{cases} 1/M & \textit{if}\, e_{l+2j} = 0,\ \bar{z}_j \neq 0,\\ -e_{l+2j} & \textit{otherwise}, \end{cases} \qquad j = 1, \ldots, \tfrac{n-l}{2},\\
&\tilde{x}_{l+2j-1} = \tfrac{\bar{z}_j}{\tilde{x}_{l+2j}},\ j = 1, \ldots, \tfrac{n-l}{2}.
\end{aligned}$$

In addition, if (IP) *is bounded from below, the optimal value of* (IP) *is unattainable if and only if* $e_{l+2j} = 0$ *and* $\bar{z}_j \neq 0$*. In this case, for any* $\epsilon > 0$*, there exists an* ϵ *optimal solution* $\tilde{x}$ *such that* $f(\tilde{x}) - v(\text{IP}) < \epsilon$ *with a sufficient large* $M > 0$.

Proof (IP) is equivalent to the following (IP$_2$):

$$\begin{aligned}
(\text{IP}_2)\quad \min\quad & \sum_{i=1}^{l}(\delta_i y_i + e_i x_i) + \sum_{j=1}^{\frac{n-l}{2}} \zeta_j z_j + c_0\\
\text{s.t.}\quad & -\sum_{i=1}^{l}(\alpha_i y_i + b_i x_i) - \sum_{j=1}^{\frac{n-l}{2}} z_j + c_1 \leq 0\\
& \sum_{i=1}^{l}(\alpha_i y_i + b_i x_i) + \sum_{j=1}^{\frac{n-l}{2}} z_j - c_2 \leq 0\\
& \frac{1}{2}x_i^2 - y_i = 0,\ \forall i = 1, 2, \ldots, l,\\
& x, y \in \mathbb{R}^l,\ z \in \mathbb{R}^{\frac{n-l}{2}}.
\end{aligned}$$

So we only need to prove the equivalence between (IP$_1$) and (IP$_2$).

By the S-lemma with interval bounds [25], similarly to Lemma 4, we know that if $v(\text{IP}_1)$ is unbounded from below, then $v(\text{IP}_2)$ is unbounded from below.

Now we consider the case where $v(\text{IP}_1)$ is bounded from below. Then there exists a global minimum (x^*, y^*, z^*) for (IP$_1$). The Fritz–John conditions of (IP$_1$) are stated as following: there exist $\nu_0 \geq 0,\ \nu_1 \geq 0,\ \nu_2 \geq 0,\ \mu_i \geq 0,\ i = 1, \ldots, l$, not all of which are zero, such that

$$\begin{aligned}
&\nu_0\delta_i - (\nu_1 - \nu_2)\alpha_i - \mu_i = 0,\ \forall i = 1, \ldots, l,\\
&\nu_0 e_i - (\nu_1 - \nu_2)b_i + \mu_i x_i^* = 0,\ \forall i = 1, \ldots, l,\\
&\nu_0\zeta_j - (\nu_1 - \nu_2) = 0,\ \forall j = 1, \ldots, \frac{n-l}{2}.
\end{aligned}$$

We assume that α_i and b_i are not both zero for $i = 1, \ldots, l$, otherwise x_i is a free variable only appearing in the objective function and then (IP$_1$) is either unbounded from below or can be reduced to a new problem without variable x_i. Moreover, we cannot take equality in both sides of the quadratic constraint, so there must exist at least one strict inequality. Then, from the last equation in Fritz–John conditions and the complementary slack conditions, we conclude $\nu_1(\bar{h}(x^*, y^*, z^*) - c_1) = 0,\ \nu_2(\bar{h}(x^*, y^*, z^*) -$

$c_2) = 0$, and one of ν_1 and ν_2 must be 0. Then from the first equation in Fritz–John conditions we know that if there exists some index i such that $\frac{1}{2}\left(x_i^*\right)^2 - y_i^* < 0$, together with the complementary slack conditions $\mu_i \left(\frac{1}{2}\left(x_i^*\right)^2 - y_i^*\right) = 0$, we conclude $\mu_i = 0$ and $\nu_0 > 0$ (otherwise $(\nu_1 - \nu_2)\alpha_i = (\nu_1 - \nu_2)b_i = 0 \Rightarrow \nu_1 - \nu_2 = 0 \Rightarrow \nu_1 = \nu_2 = 0$ and thus $\mu_i = 0$ for all i, which contradicts the fact that $\nu_0, \nu_1, \nu_2, \mu_i, i = 1, \ldots, l$, are not all zero). So the Fritz–John conditions is reduced to the KKT conditions. Because one of ν_1 and ν_2 is 0, Assumption 6 in [2] holds. Then, by applying Theorem 7 in [2], we can get another optimal solution $(\bar{x}, \bar{y}, \bar{z})$ to (IP$_1$) with $\frac{1}{2}\bar{x}_i^2 = \bar{y}_i,\ i = 1, 2, \ldots, l$, and $\bar{z} = z^*$, which is also optimal to (IP$_2$).

The remaining of the proof is similar to that of Theorem 7. □

Remark 7 Actually, problem (IP) must have an optimal solution on the boundary, except for the case where $D \succeq 0$, $e \in Range(D)$ and $x = -D^+e$ is in the interior of the interval constraint. This is because, if the optimal solution x^* is not on the boundary, then x^* must be a local minimum and (IP) has only one local minimum under the conditions that D is semi-definite positive, $e \in Range(D)$ and the local minimum is $x = -D^+e$. So we can first verify whether the conditions $D \succeq 0$ and $e \in Range(D)$ are satisfied and then check whether $x = -D^+e$ is in the interior of the constraint. Otherwise, the optimal solution must be on the boundary. Then we can separate the problem into two equality constrained problems with an equality constraint $h(x) = c_1$ or $h(x) = c_2$, and solve them with the methods for the equality constrained case. And the solution with the smaller optimal value of the above two equality constrained problems is the optimal solution to problem (IP). This fact is also investigated in [19].

Similarly to Proposition 1, we have the following proposition for problem (EP) ((IP), respectively).

Proposition 3 *Problem* (EP$_2$) *((*IP$_2$*), respectively) has a closed-form solution when $l < n$, i.e., there exists at least one 2×2 block in the canonical form of* (4) *and* (5)*:*

1. *If ζ_is are not equal, v(*EP$_2$*) $= -\infty$ (v(*IP$_2$*) $= -\infty$, respectively);*
2. *If $\zeta_1 = \zeta_2 = \cdots = \zeta_p$, problem* (EP$_2$) *((*IP$_2$*), respectively) admits a closed-form solution. If $\delta_i - \zeta_1\alpha_i > 0$ or $\delta_i - \zeta_1\alpha_i = e_i - \zeta_1 b_i = 0$, $i = 1, \ldots, \frac{n-l}{2}$, by defining $\frac{0}{0} = 0$, we have $x_i = -\frac{e_i - \zeta_1 b}{\delta_i - \zeta_1\alpha_i}$, $y_i = \frac{1}{2}x_i^2$, $i = 1, \ldots, \frac{n-l}{2}$. With properly choosing z satisfying the constraint, we have $f(x, z) = -\sum_i^l \frac{(e_i - \zeta_1 b_i)^2}{2(\delta_i - \zeta_1\alpha_i)} + c_0 - \zeta_1 c$. Otherwise v(*EP$_2$*) $= -\infty$ (v(*IP$_2$*) $= -\infty$, respectively).*

Moreover, we can get a closed-form solution to problem (EP) *((*IP*), respectively) from a solution to problem* (EP$_2$) *((*IP$_2$*), respectively).*

We also point out that the approach in Theorem 8 to solve the singleton case of I_{PSD} in Sect. 2.4 is still applicable to the two variants of the GTRS in this section. Since the main technique is almost the same, we omit the details here to save space.

4 Numerical results

In this section, we present computational results for solving problem (P) when I_{PSD} is either an interval or a singleton with a purpose to compare the efficiency of different methods. All the numerical tests were implemented in Matlab 2013a (8.1.0.604), 64 bit and was run on a Linux machine with 48 GB RAM, 2600 MHz cpu and 64-bit CentOS Release 5.5. The SDP problems in our numerical examples are modeled by CVX 2.1, and solved by Sedumi within CVX, as Sedumi is usually faster than the other solver SDPT3 in CVX. The SOCP problems are solved by CPLEX 12.6.1.

4.1 The case where I_{PSD} is an interval

For the case where I_{PSD} is an interval, we generate two sets of instances randomly.

In the first set, the matrices A and D are both diagonal and thus we already have the SOCP reformulation. The test instances of the first set were generated as follows: $A = \mathrm{diag}(a)$ with a random vector $a \in \mathbb{R}^n$ uniformly generated from $[0, 1]^n$, $D = I - A$. The parameters e, b were independent random vectors generated uniformly from $[0, 1]^n$ and c was a random scalar generated uniformly from $[0, 1]$.

In the second set, the parameters e, b and c were generated in the same way. But we set $A = P^T \mathrm{diag}(a) P$ and $D = P^T \mathrm{diag}(d) P$, where $P = E^T E + I$ and E was an $n \times n$ random matrix with uniformly distributed numbers.

Table 1 compares the computational times of SDP and SOCP reformulation for problem (P) for the first set. We run 10 instances for each n and report the average time as “CPU time”. The computational times of SDP and SOCP are represented as

Table 1 Computational results for SDP and SOCP with diagonal matrix inputs

n	CPU time (s)	
	SDP	SOCP
200	2.89	0.0372
400	10.7	0.0506
600	27.2	0.0444
800	56.7	0.0552
1000	782	0.0802
1200	1385	0.0899
1400	2081	0.102
1600	2930	0.154
1800	4216	0.162
2000	–	0.561
4000	–	0.829
8000	–	2.84
14,000		7.81
20,000	–	7.19
40,000	–	196

Table 2 Computational results for SDP and SOCP with general matrix inputs

n	CPU time (s)				
	SDP	SOCP (total)	SOCP	Computing λ	Diagtime
100	2.79	0.0855	0.0161	0.0209	0.0485
600	163	2.1559	0.0379	1.13	0.988
1000	808	6.9833	0.0733	4.25	2.66
1500	2403	20.414	0.124	14.5	5.79
2000	–	57.799	0.199	45.0	12.6
3000	–	235.53	0.430	188	47.1
4000	–	529.354	0.654	431	97.7
5000	–	1056.02	1.02	1036	19.0
6000	–	1767.21	1.51	1735	30.7
7000	–	2726.01	2.01	2677	47.0
8000	–	3945.64	2.64	3874	69.0

"SDP" and "SOCP", respectively. Table 1 shows that SOCP reformulation is much faster than the SDP reformulation as expected. In fact, the SDP problems would be solved in more than five hours and out of memory when $n > 2000$. On the other hand, the SOCP problems can be solved very fast even when $n = 40{,}000$.

Table 2 compares the computational times of SDP and SOCP reformulations for problem (P) for the second set. We use "SDP" and "SOCP" to denote the CPU time for solving the SDP reformulation in CVX and the SOCP reformulation in CPLEX, respectively. We simultaneously diagonalize the matrix pair (A, D) by combining the algorithm in [9] of finding λ and the SD algorithm in [6] for a matrix pair with a positive definite matrix pencil. We use "computing λ" to denote the time for computing λ via the algorithm in [9] and "diagtime" to denote the time for simultaneously diagonalizing a definite matrix and an arbitrary symmetric matrix, respectively. So the runtime of SD approach equals to the time of "computing λ" plus "diagtime". We also use "SOCP (total)" to denote the total CPU time for solving the GTRS via our SOCP reformulation, i.e., the summation of the time for the SD approach and the time for solving the pure SOCP problem. To obtain the SOCP reformulation, we need first to invoke the SD approach. Table 2 shows that the runtime of the SD approach is much larger than that of solving an SOCP problem for all n. The SOCP reformulation is still much faster than the SDP reformulation even after we include the time for the SD approach. It is clear that the computation of λ becomes the major part of runtime for the SOCP reformulation. Using the state-of-the-art method in [9], finding λ may cost larger than 1 h when $n \geq 8000$. This performance also indicates that the SOCP reformulation is efficient for moderate size problems.

We also run some numerical tests for Hard case 1 in [19] for (IP). The data was generated in the same way as in [21]. Since for Hard case 1, the method in [19] often fails in initialization and is generally slower than the DB algorithm in [21], we only compare the SOCP reformulation with the DB algorithm. In Table 3, we use SOCP

Table 3 Computational results for SOCP reformulation and DB algorithm

n	CPU time (s)		
	SOCP	DB	Diagtime
1000	2.67	2.63	2.61
2000	17.48	17.32	17.32
3000	54.20	54.00	53.79
4000	112.73	112.63	112.22
5000	199.59	199.52	198.87
6000	308.68	308.68	307.66
7000	407.73	407.26	406.33
8000	606.33	605.67	604.47
9000	852.76	852.02	850.57
10,000	1116.68	1115.63	1114.11
11,000	1485.69	1484.93	1482.58
12,000	1925.76	1924.91	1922.08

and DB to denote the CPU time for solving the SOCP reformulation in CPLEX and the DB algorithm, respectively. And "diagtime" denotes the time for simultaneously diagonalizing the two matrices A and D. Note that in this data set, one of A and D is positive definite and thus we do not need to compute λ. From Table 3, we conclude that the SOCP reformulation has a similar performance with the DB algorithm. Also note that the SD approach takes a major part of the runtime for both methods, larger than a percentage of 99%. The numerical results show that the pure SOCP problem can be solved in less than four seconds by CPLEX when $n \leq 12{,}000$.

4.2 The case where I_{PSD} is a singleton

We generate two sets of test problems for the case where I_{PSD} is a singleton. For the first set: we set $A_0 = \text{diag}(a, 1, -1)$ where a was uniformly generated from $[0, 1]^{n-2}$. We set $A = P^T A_0 P$ and $D = P^T P - A$, where P is a random nonsingular matrix generated by Matlab commands P = sprandsym(n, 0.5, 0.1, 2). The parameters c and b are generated in a similar way to Sect. 4.1, and $e = (D + A)e_0 - b$, where e_0 is the all one vector. We set similar parameters for the second set except that we set $A_0 = \text{diag}\left(a, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\right)$.

We still apply the algorithm in [9] to compute λ. But since the purpose of this algorithm is to find a definite matrix pencil and thus cannot be applied to a singular matrix pencil directly, we need some perturbation to the two matrices A and D. More specifically, to avoid numerical precision error, we can find a $\lambda \in I_{PSD}$ in the following way: first add ϵI with some sufficiently small constant $\epsilon > 0$ (in our numerical algorithm, we set $\epsilon = 10^{-8}$ and $\epsilon = 10^{-12}$ for sets 1 and 2, respectively) to D and A and then detect w_1 and w_2 with $w_1^2 + w_2^2 = 1$ such that $w_1(D + \epsilon I) + w_2(A + \epsilon I)$ is positive definite. Then in the numerical sense, we can conclude $w_1 D + w_2 A \succeq 0$.

Table 4 Computational results for solving the singleton case

Set-n	CPU time (s)		
	SDP	Theorem 8 (total)	Computing λ
1–500	117.1	5.8	5.5
1–1000	8689	37.8	35
1–2000	–	268.9	265.3
1–3000	–	1170.3	1162.4
1–4000	–	3390.0	3372.0
2–500	141.89	8.9	8.5
2–1000	8076	51.8	48.9
2–2000	–	314.1	312.9
2–3000	–	1497.5	1494.5
2–4000	–	3467.2	3461.5

Numerical results further show that our algorithm is quite stable and always outputs the same optimal value with the SDP reformulation.

Our numerical algorithm based on Theorem 8 outputs an optimal solution or an asymptotic solution for problem (P). We use "SDP" and 'Theorem 8 (total)" to denote the CPU time for solving the SDP reformulation in CVX and the CPU time of applying Theorem 8, including the computational time for finding λ, respectively. We still use "computing λ" to denote the time for computing λ via the algorithm in [9]. Table 4 shows that our algorithm performs much better than the SDP reformulation. It is also remarkable that the main runtime is in finding λ and the runtime of obtaining a closed-form solution is neglected when compared with the time of finding λ. This is mainly due to the latter takes only several steps of matrix multiplications and computing the null space of a matrix. It is also interesting to note that the SDP takes two to three hours to solve the test problem with $n = 1000$, which indicates that the singleton case is usually hard to solve. On the other hand, the method based on Theorem 8 can solve problem with size up to 4000 in less than one hour.

5 Conclusions

In this paper, we have successfully conducted a theoretical analysis for the GTRS. Particularly, we have derived the SOCP reformulation under the condition that the GTRS is bounded from below, using a canonical form via congruence of the two matrices in both the objective and constraint functions. While Ben-Tal and den Hertog investigate in [2] the simultaneous diagonalizability of the two matrices, we explore the simultaneous block diagonalizability, which applies to two arbitrary matrices. More specifically, we introduce and extend the canonical form for two real matrices in [24] to find a block diagonal form for two matrices in both the objective and constraint functions. Exploiting the separability of the block diagonal form of the two matrices, we show that problem (P) is SOCP representable if it is bounded from below. We also establish the attainableness of the problem from the canonical form without any additional calculation. Using the SOCP reformulation, we further show a closed-form solution for the GTRS when the quadratic forms are not SD. Although there is no stable algorithm in the current literature to compute the canonical form, we derive a stable algorithm in this paper to solve the GTRS only using the block structure of the canonical form. Our numerical results demonstrate the efficiency of our methods. We further extend our methods to solve two variants of problem (P), where the inequality constraint is replaced by either an equality constraint or an interval constraint.

One of our future research is to consider variants of the GTRS with additional linear inequality constraints or two general quadratic constraints.

Acknowledgements The authors gratefully acknowledge the support of Hong Kong Research Grants Council under Grant 14213716. The second author is also grateful to the support from Patrick Huen Wing Ming Chair Professorship of Systems Engineering and Engineering Management. The authors would like to thank the two anonymous reviewers for their insightful and constructive comments which help improve the paper significantly. The authors also appreciate Prof. N. Higham and Prof. M. Salahi for sharing with them the codes for detecting definite matrix pencils and the codes for DB algorithm, respectively.

References

1. Adachi, S., Nakatsukasa, Y.: Eigenvalue-based algorithm and analysis for nonconvex QCQP with one constraint (2016). http://www.keisu.t.u-tokyo.ac.jp/research/techrep/data/2016/METR16-07.pdf
2. Ben-Tal, A., den Hertog, D.: Hidden conic quadratic representation of some nonconvex quadratic optimization problems. Math. Program **143**(1–2), 1–29 (2014)
3. Ben-Tal, A., Nemirovski, A.: Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications, vol. 2. SIAM, Philadelphia (2001)
4. Ben-Tal, A., Teboulle, M.: Hidden convexity in some nonconvex quadratically constrained quadratic programming. Math. Program **72**(1), 51–63 (1996)
5. Conn, A.R., Gould, N.I., Toint, P.L.: Trust Region Methods, vol. 1. SIAM, Philadelphia (2000)
6. Datta, B.N.: Numerical Linear Algebra and Applications. SIAM, Philadelphia (2010)
7. 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)
8. Golub, G.H., Van Loan, C.F.: Matrix Computations, vol. 3. JHU Press, Baltimore (2012)
9. Guo, C.H., Higham, N.J., Tisseur, F.: An improved arc algorithm for detecting definite hermitian pairs. SIAM J. Matrix Anal. Appl. **31**(3), 1131–1151 (2009)
10. Hmam, H.: Quadratic optimization with one quadratic equality constraint. Tech. Rep., Warfare and Radar Division DSTO Defence Science and Technology Organisation, Report DSTO-TR-2416 (2010)
11. 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)
12. Jiang, R., Li, D.: Simultaneous diagonalization of matrices and its applications in quadratically constrained quadratic programming. SIAM J. Optim. **26**(3), 1649–1668 (2016)
13. Kågström, B., Ruhe, A.: An algorithm for numerical computation of the Jordan normal form of a complex matrix. ACM Trans. Math. Softw. **6**(3), 398–419 (1980)
14. Lancaster, P., Rodman, L.: Canonical forms for hermitian matrix pairs under strict equivalence and congruence. SIAM Rev. **47**(3), 407–443 (2005)
15. Martínez, J.M.: Local minimizers of quadratic functions on Euclidean balls and spheres. SIAM J. Optim. **4**(1), 159–176 (1994)
16. Moré, J.J.: Generalizations of the trust region problem. Optim. Methods Softw. **2**(3–4), 189–209 (1993)
17. Moré, J.J., Sorensen, D.C.: Computing a trust region step. SIAM J. Sci. Stat. Comput. **4**(3), 553–572 (1983)
18. Pólik, I., Terlaky, T.: A survey of the S-lemma. SIAM Rev. **49**(3), 371–418 (2007)
19. Pong, T.K., Wolkowicz, H.: The generalized trust region subproblem. Comput. Optim. Appl. **58**(2), 273–322 (2014)
20. Rendl, F., Wolkowicz, H.: A semidefinite framework for trust region subproblems with applications to large scale minimization. Math. Program **77**(1), 273–299 (1997)

21. Salahi, M., Taati, A.: An efficient algorithm for solving the generalized trust region subproblem. Comp. Appl. Math. (2016). doi:10.1007/s40314-016-0349-1
22. Stern, R.J., Wolkowicz, H.: Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations. SIAM J. Optim. **5**(2), 286–313 (1995)
23. Sturm, J.F., Zhang, S.: On cones of nonnegative quadratic functions. Math. Oper. Res. **28**(2), 246–267 (2003)
24. Uhlig, F.: A canonical form for a pair of real symmetric matrices that generate a nonsingular pencil. Linear Algebra Appl. **14**(3), 189–209 (1976)
25. Wang, S., Xia, Y.: Strong duality for generalized trust region subproblem: S-lemma with interval bounds. Optim. Lett. **9**(6), 1063–1073 (2015)
26. Xia, Y., Wang, S., Sheu, R.L.: S-lemma with equality and its applications. Math. Program **156**(1–2), 513–547 (2016)
27. Ye, Y., Zhang, S.: New results on quadratic minimization. SIAM J. Optim. **14**(1), 245–267 (2003)
28. Yuan, Y.: On a subproblem of trust region algorithms for constrained optimization. Math. Program **47**(1–3), 53–63 (1990)