Math. Meth. Oper. Res. (2006) 64: 55–77
DOI 10.1007/s00186-006-0070-8

ORIGINAL ARTICLE

Kürşad Derinkuyu · Mustafa Ç. Pınar

On the S-procedure and some variants

Received: 15 August 2004 / Accepted: 24 August 2005 / Published online: 9 June 2006

Abstract We give a concise review and extension of S-procedure that is an instrumental tool in control theory and robust optimization analysis. We also discuss the approximate S-Lemma as well as its applications in robust optimization.

Keywords S-Procedure · S-Lemma · Robust optimization · Control theory

1 Introduction

The purpose of this paper is to give a concise review of recent developments related to the S-procedure in a historical context as well as to offer a new extension. S-procedure is an instrumental tool in control theory and robust optimization analysis. It is also used in linear matrix inequality (or semi-definite programming) reformulations and analysis of quadratic programming. It was given in 1944 by Lure and Postnikov without any theoretical justification. Theoretical foundations of S-procedure were laid in 1971 by Yakubovich and his students (Yakubovich 1971).

S-procedure deals with the nonnegativity of a quadratic function on a set described by quadratic functions and provides a powerful tool for proving stability of nonlinear control systems. For simplicity, if the constraints consist of a single quadratic function, we refer to it as S-Lemma. If there are at least two quadratic inequalities in the constraint set, we use the term of S-procedure. Yakubovich

The many suggestions and detailed corrections of an anonymous referee are gratefully acknowledged.

K. Derinkuyu
Lehigh University, Bethlehem, USA

M. Ç. Pınar (✉)
Bilkent University, Ankara, Turkey
E-mail: mustafap@bilkent.edu.tr

(1971) was the first to prove the S-Lemma and to give a definition of S-procedure. Recently, Polyak (1998) gave a result related to S-procedure for problems involving two quadratic functions in the constraint set.

Although the S-Lemma was proved in 1971, results on the convexity problems of quadratic functions were already there since 1918. From Toeplitz (1918) and Hausdorff (1919) theorem to more recent results, many important contributions to the field are available. In this period, not only the S-Lemma was improved, but also a new result was introduced, called the approximate S-Lemma. The approximate S-Lemma developed by Ben-Tal et al. (2002) establishes a bound for problems with more than one constraints of quadratic type. Their result also implies the S-Lemma of Yakubovich.

In the present paper we offer yet another generalization of the S-procedure referred to as the extended S-procedure (a term coined in this paper), that implies both the theorems of Yakubovich and Polyak. This procedure is obtained as a corollary of Au-Yeung and Poon (1979), and Barvinok's (1995) theorems.

Although papers concerning the S-procedure abound (as well as many that make use of it), it appears that a summary review of the subject encompassing the latest developments still remains unavailable to the research community. The present paper should be considered an attempt to fulfill this need.

The remainder of this study is organized as follows: Sect. 2 provides a background on the S-procedure with an extensive (although not pretending to be exhaustive) review of literature. In Sect. 3, our exposition of approximate S-Lemma and extended S-procedure are given. Section 4 is devoted to a critical evaluation. Section 5 gives concluding remarks. Open problems are also pointed out in the last two sections.

Notation We work in a finite dimensional (Euclidian) setting $\mathbf{R}^n$, with the standard inner product denoted by $\langle .,. \rangle$ and associated norm denoted by $\|.\|$. We use $S_n^{\mathbf{R}}$ to denote $(n \times n)$ symmetric real matrices. For $A \in S_n^{\mathbf{R}}$, $A \succeq 0$ $(A \succ 0)$ means A is positive semi-definite (positive definite). Also we use $M_{n,p}(\mathbf{R})$ to denote the space of real $(n \times p)$-matrices. If $A \in S_n^{\mathbf{R}}$ and $X \in M_{n,p}(\mathbf{R})$, then $\langle\langle AX, X\rangle\rangle = \langle\langle A, XX^{\mathrm{T}}\rangle\rangle := TrAXX^{\mathrm{T}} =$ trace of $A^{\mathrm{T}}(XX^{\mathrm{T}})$.

2 Background

S-procedure is one of the fundamental tools of control theory and robust optimization. It is related to several mathematical fields such as numerical range, convex analysis and quadratic functions. Since it is at the crossroads of several fields, efforts were undertaken to improve it or to understand its structure. Therefore, it is only natural to begin with its history to appreciate its importance.

In 1918, Toeplitz introduced the idea of the numerical range ($W(A)$) of a complex $(n \times n)$ matrix A in the "Das algebraische Analogon zu einem Satze von Fejér". For a quadratic form z^*Az, he proved that it has a convex boundary for z belonging to the unit sphere in the space C^n of complex n-tuples (it is also called the numerical range of A). He also conjectured that the numerical range itself is convex. One year later, Hausdorff (1919) proved it. The Toeplitz–Hausdorff theorem is a very important result due to its extensions in the numerical range, and it is applied in many fields of mathematics. This theorem can be formulated as: let

$$W(A) = \{ z^* A z \mid \|z\| = 1 \}.$$

Then, the set W is convex in the set C of complex numbers. This result is the first assertion on convexity of quadratic maps.

For the real field, the first result was obtained by Dines (1941) for two real quadratic forms. Dines proved that for two dimensional image of $\mathbf{R}^n$ and for any real symmetric matrices A and B, the set

$$D = \{ (\langle Ax, x\rangle, \langle Bx, x\rangle) \mid x \in \mathbf{R}^n \}$$

is a convex cone where $\langle Ax, x\rangle = x^{\mathrm{T}} Ax$, and that under some additional assumption it is closed.

The next important result was obtained by Brickman (1961). He proved that the image of the unit sphere for the $n \geq 3$ (for any real symmetric matrices A and B),

$$B = \{ (\langle Ax, x\rangle, \langle Bx, x\rangle) \mid \|x\| = 1 \}$$

is a convex compact set in $\mathbf{R}^2$.

These three papers are the main contributions on the numerical range, and mathematicians tried in several ways to generalize them. Before explaining these developments, let us look at our main subject: S-procedure.

S-procedure deals with nonnegativity of a quadratic form under quadratic inequalities. The first result in this area is Finsler's Theorem (Finsler 1936/37) (also known as Débreu's lemma). Calabi (1964) also proved this result independently in studying differential geometry and matrix differential equations by giving a new and short topological proof (A unilateral version of this theorem was proved by Yuan (1990)).

Theorem 1 *The theorem of Finsler* (1936), *Calabi* (1964)
For $n \geq 3$, let $A, B \in S_n^{\mathbf{R}}$. Then the following are equivalent:

(*i*) $\langle Ax, x\rangle = 0$ *and* $\langle Bx, x\rangle = 0$ *implies* $x = 0$.
(*ii*) $\exists \mu_1, \mu_2 \in \mathbf{R}$ *such that* $\mu_1 A + \mu_2 B \succ 0$.

In 1971, Yakubovich proved the S-Lemma which became very popular in control theory. There exist several methods to prove it but we want to give here a proof that uses Dines' theorem to emphasize the link between convexity and the S-Lemma which is a separation theorem for convex sets. (One can consult Nemirovski's (2002) book (pp. 132–135) or Luo et al. (2003) paper for different proofs).

Theorem 2 (*S-Lemma*) *Let A,B be symmetric $n \times n$ matrices, and assume that the quadratic inequality*

$$x^{\mathrm{T}} Ax \geq 0$$

is strictly feasible(there exists $\overline{x}$ such that $\overline{x}^{\mathrm{T}} A\overline{x} > 0$). Then the quadratic inequality:

$$x^{\mathrm{T}} Bx \geq 0$$

is a consequence of it, i.e.,

$$x^{\mathrm{T}}Ax \geq 0 \Rightarrow x^{\mathrm{T}}Bx \geq 0$$

if and only if there exists a nonnegative λ *such that*

$$B \succeq \lambda A.$$

Proof The sufficiency part is immediately proved. Therefore, let us assume that $x^{\mathrm{T}}Bx \geq 0$ is a consequence of $x^{\mathrm{T}}Ax \geq 0$. Let

$$S = \{(x^{\mathrm{T}}Ax, x^{\mathrm{T}}Bx) : x \in \mathbf{R}^n\}$$

and

$$U = \{(u_1, u_2), u_1 \in \mathbf{R}_+, u_2 \in \mathbf{R}_{--}\}.$$

S is a convex set by Dines' theorem while U is a convex cone. Since their intersection is empty, a separating hyperplane exists. I.e., there exists nonzero $c = (c_1, c_2) \in \mathbf{R}^2$, such that $(c, s) \leq 0, \forall s \in S$ and $(c, u) \geq 0, \forall u \in U$. For $(0, -1) \in U$ we have $c_2 \leq 0$. For $(1, -\alpha) \in U$ where α is a small positive number arbitrarily chosen, we obtain $c_1 \geq \alpha c_2$. Letting α tend to zero, we get $c_1 \geq 0$. Since there exists $\overline{x}$ such that $\overline{x}^{\mathrm{T}}A\overline{x} > 0$, and by the separation argument we have $c_1 x^{\mathrm{T}}Ax + c_2 x^{\mathrm{T}}Bx \leq 0$ for all $x \in \mathbf{R}^n$, we can write

$$c_1\overline{x}^{\mathrm{T}}A\overline{x} + c_2\overline{x}^{\mathrm{T}}B\overline{x} \leq 0.$$

Since we have $c_1 \geq 0, \overline{x}^{\mathrm{T}}A\overline{x} > 0, \overline{x}^{\mathrm{T}}B\overline{x} \geq 0$ by hypothesis, and that c_1 and c_2 cannot both be zero, the last inequality implies that $c_2 < 0$. Therefore, we obtain: $x^{\mathrm{T}}Bx \geq -c_1/c_2 x^{\mathrm{T}}Ax$ for all $x \in \mathbf{R}^n$, which is equivalent to $B \succeq \lambda A$ after defining $\lambda = -c_1/c_2$. This completes the proof of the necessity part. Hence, the result is proved.

The idea of this proof is used in many papers about the subject. It is also used in the first two results in the next section. At this point, we divide our review into two sub-areas. Firstly, we try to generalize this theorem to obtain more complicated cases. Then we look at a new area recently developed by Ben-Tal et al. (2002) to obtain approximate version of the general result.

2.1 Review of research on the S-procedure

The first attack to generalize the above theorems was made by Hestenes and McShane (1940). They generalized the theorem of Finsler (1936).

Theorem 3 *The theorem of Hestenes and McShane* (1940)

Assume that $x^{\mathrm{T}}Sx > 0$ *for all nonzero* x *such that* $\{x \in \mathbf{R}^n | \bigcap_{i=1}^r (\langle T_i x, x\rangle = 0)\}$. *Let* T_i *be such that* $\sum_i a_i T_i$ *is indefinite for any nontrivial choice of* $a_i \in \mathbf{R}$. *Moreover assume that for any subspace* $L \subseteq \mathbf{R}^n \setminus \bigcap_{i=1}^r (\langle T_i x, x\rangle = 0)$ *there are constants* $b_i \in \mathbf{R}$ *such that* $x^T \left(\sum_i b_i T_i\right) x > 0$ *for all nonzero* $x \in L$. *Then, there exists* $c \in \mathbf{R}^{r+1}$ *that;*

$$c_0 S + c_1 T_1 + \cdots + c_r T_r \succ 0.$$

For $r = 1$ *only the first assumption needs to be made.*

There are several papers in this area by Au-Yeung (1969), Dines (1942, 1943), John (1938), Kühne (1964), Taussky (1967) and others. One of the benefits of Finsler, and Hestenes and McShane's theorems is the appearence of positive definiteness of a linear combination of matrices a la S-procedure. One can find a review of related results covering the period until 1979 in a nice survey by Uhlig (1979).

To generalize the S-Lemma, researchers either replace vector variables with matrix variables, or make additional assumptions. First, we look into the first category and among these theorems, we deal with a most popular unpublished result: the theorem of Bohnenblust (unpublished Manuscript) on the joint positive definiteness of matrices. Although this theorem can be stated for the field of complex numbers and the skew field of real quaternions, we only deal with the field of real numbers.

Theorem 4 *The theorem of Bohnenblust*
Let $1 \leq p \leq n-1, m < \frac{(p+1)(p+2)}{2} - \delta_{n,p+1}$ *and* $A_1, \ldots, A_m \in S_n^{\mathbf{R}}$. *Suppose* $(0, \ldots, 0) \notin W_p(A_1, \ldots, A_m)$ *where*

$$W_p(A_1, \ldots, A_m) = \left\{ \left(\sum_{i=1}^{p} x_i^{\mathrm{T}} A_1 x_i, \ldots, \sum_{i=1}^{p} x_i^{\mathrm{T}} A_m x_i \right) : x_i \in \mathbf{R}^n, \sum_{i=1}^{p} x_i^{\mathrm{T}} x_i = 1 \right\}.$$

Then there exist $\alpha_1, \ldots, \alpha_m \in \mathbf{R}$ *such that the matrix* $\sum_1^m \alpha_i A_i$ *is positive definite. (*$\delta_{n,p+1}$ *is Kronecker delta).*

With the help of this theorem, Au-Yeung and Poon (1979) showed the extension of Brickman's and Toeplitz-Hausdorff theorem in 1979, and Poon (1997) gives the final version of this result in 1997. Here is the Au-Yeung and Poon theorem for real cases:

Theorem 5 *The theorem of Au-Yeung and Poon* (1979) [*Extension of Brickman* (1961) *using Bohnenblust*]
Let $1 \leq p \leq n-1, m < \frac{(p+1)(p+2)}{2} - \delta_{n,p+1}$ *and* $A_1, \ldots, A_m \in S_n^{\mathbf{R}}$. *Then,*

$$\{(\langle\langle A_1 X, X\rangle\rangle, \langle\langle A_2 X, X\rangle\rangle, \ldots, \langle\langle A_m X, X\rangle\rangle) | X \in M_{n,p}(\mathbf{R}), \|X\| = 1\}$$

is a convex compact subset of $\mathbf{R}^m$. (*$\delta_{i,j}$ is equal to one when $i = j$, otherwise zero*). ($\|.\|$ *denotes the Schur-Frobenius norm on* $M_{n,p}(\mathbf{R})$, *derived from* $\langle\langle ., .\rangle\rangle$).

Here $\langle\langle AX, X\rangle\rangle = Tr AXX^{\mathrm{T}} = \sum_{i=1}^{p} x_i^{\mathrm{T}} A x_i$ and x_i denotes the columns of X. A corollary of this theorem is given in the paper of Hiriart-Urruty and Torki (2002).

Theorem 6 *Corollary* (*Hiriart-Urruty and Torki* 2002) *of the theorem of Poon* (1997)
Let $A_1, A_2, \ldots, A_m \in S_n^{\mathbf{R}}$ *and let*

$$p := \left\{ \begin{array}{ll} \lfloor \frac{\sqrt{8m+1}-1}{2} \rfloor & \text{if } \frac{n(n+1)}{2} \neq m+1 \\ \lfloor \frac{\sqrt{8m+1}-1}{2} \rfloor + 1 & \text{if } \frac{n(n+1)}{2} = m+1 \end{array} \right\}$$

(*thus* $p = 1$ *when* $m = 2$ *and* $n \geq 3$, $p = 2$ *when* $m = 2$ *and* $n = 2$, *etc.*) *Then the following are equivalent:*

$$(i)\quad \left\{\begin{array}{l}\langle\langle A_1X, X\rangle\rangle = 0\\ \langle\langle A_2X, X\rangle\rangle = 0\\ \cdot\\ \cdot\\ \cdot\\ \langle\langle A_mX, X\rangle\rangle = 0\end{array}\right\} \Rightarrow (X = 0).$$

(*ii*) *There exists* $\mu_1, \ldots, \mu_m \in \mathbf{R}$ *such that*

$$\sum_{i=1}^{m} \mu_i A_i \succ 0.$$

We note that the paper by Hiriart-Urruty and Torki (2002) gives a good overview of the convexity of quadratic maps and poses several open problems.

Barvinok's (1995) gave another theorem extending the Dines's and Toeplitz–Hausdorff theorem while working on distance geometry.

Theorem 7 *The theorem of Barvinok*(1995)[*Extension of Dines*(1941)] *Let* $A_1, A_2, \ldots, A_m \in S_n^{\mathbf{R}}$, *and let* $p := \lfloor\frac{\sqrt{8m+1}-1}{2}\rfloor$. *Then*

$$\{(\langle\langle A_1X, X\rangle\rangle, \langle\langle A_2X, X\rangle\rangle, \ldots, \langle\langle A_mX, X\rangle\rangle)|X \in M_{n,p}(\mathbf{R})\}$$

is a convex cone of $\mathbf{R}^m$.

Papers of Poon and Barvinok are important for our extension results because we use them for the extended S-procedure in section 3. Now we give the definition of both S-procedure and extended S-procedure and turn our interest to results about S-procedure without extension but using additional assumptions. The definition of S-procedure is given by Yakubovich (1971) and his students in 1971. Before talking about related papers on S-procedure, let us define the S-procedure and extended S-procedure:

Definition 8 (*S-procedure and Extended S-procedure*)
Define

$$\begin{aligned} q_i(X) \;=\; &\sum_{j=1}^{p} x_j^{\mathrm{T}} Q_i x_j + 2b_i^{\mathrm{T}} \sum_{j=1}^{p} x_j + c_i,\\ &Q_i \in S_n^{\mathbf{R}},\; i = 0, \ldots, m,\; j = 1, \ldots, p,\; X = (x_1, \ldots, x_p)\\ F := \{&X \in M_{n,p}(\mathbf{R}) : q_i(X) \geq 0,\; i = 1, \ldots, m\}, \end{aligned}$$

$q_i(x_j)$ *is called quadratic function and if* b_i *and* c_i *are zero, then it is called quadratic form. Now consider the following conditions:*

(S_1) $q_0(X) \geq 0 \quad \forall X \in F$
(S_2) $\exists s \in \mathbf{R}_+^m : q_0(X) - \sum_{i=1}^{m} s_i q_i(X) \geq 0,\ \forall X \in M_{n,p}(\mathbf{R})$

Method of verifying (S_1) *using* (S_2) *is called S-procedure for* $p = 1$ *and called extended S-procedure for* $p > 1$.

Note that always $S_2 \Rightarrow S_1$. Indeed,

$$q_0(x) \geq \sum_{i=1}^{m} s_i q_i(x) \geq 0.$$

Unfortunately, the converse is in general false. If $S_1 \Leftrightarrow S_2$, the S-procedure is called lossless. However, this condition is fulfilled only in some special cases.

The first paper reviewed on the S-Procedure with additional assumptions is the paper of Megretsky and Treil (1993). They prove the S-procedure for the continuous time-invariant quadratic forms.

Let $L^2 = L^2((0,\infty); \mathbf{R}^n)$ be the standard Hilbert space of real vector-valued square-summable functions defined on $(0,\infty)$. A subspace $L \in L^2$ is called time invariant if for any $f \in L$, and $\tau > 0$ the function f^τ,defined by $f^\tau(s) = 0$ for $s \leq \tau$, $f^\tau(s) = f(s-\tau)$ for $s > \tau$, belongs to L. Similarly, a functional $\sigma : L \to \mathbf{R}$ is called time invariant if $\sigma(f^\tau) = \sigma(f)\ \forall f \in L, \tau > 0$.

Theorem 9 *The S-procedure losslessness theorem of Megretsky and Treil*(1993)
Let $L \subset L^2$ be time invariant subspace, and $\sigma_j : L \to \mathbf{R}(j = 0, 1, \ldots, m)$ be continuous time-invariant quadratic forms. Suppose that there exists $f_* \in L$ such that $\sigma_1(f_*) > 0, \ldots, \sigma_m(f_*) > 0$.
Then the following statements are equivalent:

(*i*) $\sigma_0(f) \leq 0$ *for all $f \in L$ such that $\sigma_1(f) > 0, \ldots, \sigma_m(f) > 0$;*
(*ii*) *There exists $\tau_j \geq 0$ such that*

$$\sigma_0(f) + \tau_1\sigma_1(f) + \cdots + \tau_m\sigma_m(f) \leq 0$$

for all $f \in L$.

Although this theorem gives us the S-procedure, time-invariant quadratic forms are very domain specific. Moreover, one can find another convexity result for commutative matrices in the paper of Fradkov (1973) (Detailed information about commutative matrices can be obtained from the book Matrix Analysis by Horn and Johnson (1990)).

Theorem 10 *Theorem of Fradkov*
Let m quadratic forms $f_i(x) = \langle A_i x, x\rangle, x \in \mathbf{R}^n, i = 1, \ldots, m$ be given. If matrices $A_1, \ldots, A_m$ commute, then

$$F_m = \{(f_1(x), \ldots, f_m(x))^{\mathrm{T}} : x \in \mathbf{R}^n\} \subset \mathbf{R}^m$$

is a closed convex cone for all m,n.

In addition to Megretsky and Treil, and Fradkov's papers, yet further extensions of the S-procedure exist. A result in this direction was proved recently by Luo et al. (2003) where quadratic matrix inequalities were used instead of linear matrix inequalities.

Theorem 11 *Theorem of Luo et al.* (2003)
The data matrices (A, B, C, D, F, G, H) satisfy the robust fractional quadratic matrix inequality

$$\begin{bmatrix} H & F+GX \\ (F+GX)^{\mathrm{T}} & C+X^{\mathrm{T}}B + B^{\mathrm{T}}X+X^{\mathrm{T}}AX \end{bmatrix} \succeq 0 \quad \text{for all } X \text{ with } I - X^{\mathrm{T}}DX \succeq 0$$

if and only if there is $t \geq 0$ such that

$$\begin{bmatrix} H & F & G \\ F^{\mathrm{T}} & C & B^{\mathrm{T}} \\ G^{\mathrm{T}} & B & A \end{bmatrix} - t \begin{bmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & -D \end{bmatrix} \succeq 0.$$

Unfortunately, neither Megretsky and Treil, and Fradkov's results nor extension of Luo et al. imply the S-procedure in general.

Polyak (1998) succeeded in proving a version of S-procedure for $m = 2$ by making an additional assumption, and it is the most valuable result found recently in this field. He first proved the following theorem to obtain the S-procedure for $m = 2$:

Theorem 12 *Convexity result of Polyak in* 1998 (*relies on Brickman's theorem,* 1961)
For $n \geq 3$ the following assertions are equivalent:

(*i*) *There exists $\mu \in \mathbf{R}^3$ such that*

$$\mu_1 A_1 + \mu_2 A_2 + \mu_3 A_3 \succ 0.$$

(*ii*) *For $f_i(x) = \langle A_i x, x \rangle$, $x \in \mathbf{R}^n$, $i = 1, 2, 3$, the set:*

$$F = \{(f_1(x), f_2(x), f_3(x))^{\mathrm{T}} : x \in \mathbf{R}^n\} \subset \mathbf{R}^3$$

is an acute (contains no straight lines), closed convex cone.

This nice theorem and its beautiful proof bring us the following S-procedure for quadratic forms, $m = 2$.

Theorem 13 *Polyak's theorem,* 1998 (*uses separation lemma*)
Suppose $n \geq 3$, $f_i(x) = \langle A_i x, x \rangle$, $x \in \mathbf{R}^n$, $i = 0, 1, 2$, real numbers α_i, $i = 0, 1, 2$ and there exist $\mu \in \mathbf{R}^2$, $x^0 \in \mathbf{R}^n$ such that

$$\mu_1 A_1 + \mu_2 A_2 \succ 0$$

$$f_1(x^0) < \alpha_1, \; f_2(x^0) < \alpha_2.$$

Then

$$f_0(x) \leq \alpha_0 \;\; \forall x : f_1(x) \leq \alpha_1, \; f_2(x) \leq \alpha_2$$

holds if and only if there exist $\tau_1 \geq 0$, $\tau_2 \geq 0$:

$$\begin{aligned} A_0 &\preceq \tau_1 A_1 + \tau_2 A_2 \\ \alpha_0 &\geq \tau_1 \alpha_1 + \tau_2 \alpha_2. \end{aligned}$$

A related line of work on the optimality conditions for the minimization of quadratic functions subject to two quadratic inequalities can also be followed by starting to trace back the subject from the very nice, and relatively recent paper of Peng and Yuan (1997). To keep the paper at a reasonable length, we do not review these results here.

Unfortunately, Polyak's theorem, although very elegant, is not sufficient to deal with certain problems of robust optimization as we shall see in the next section. Recently a new result in this direction was proved by Ben-Tal et al. (2002) referred to as the approximate S-Lemma which we review next.

2.2 Review of research on the approximate S-Lemma

In this section, we not only deal with the approximate S-Lemma but also concentrate on its impact on robust systems of uncertain quadratic and conic quadratic problems whereby the reader may appreciate the importance of approximate S-Lemma.

S-Lemma has been widely used within the robust optimization paradigm of Ben-Tal et al. (2004, 1998, 2000), Boyd et al. (1994) and El-Ghaoui et al. (1998) to find robust counterparts for uncertain convex optimization problems under an ellipsoidal model of the uncertain parameters. Now we concentrate on approximate S-Lemma, so we use the same notation as the paper of Ben-Tal (2002). Before beginning to talk about the subject, we need additional notation and definitions about robust methodology and conic quadratic problems. For conic programming, Ben-Tal's lecture notes (2002) are an excellent reference.

Definition 14 *Let $K \subseteq \mathbf{R}^n$ be a closed pointed convex cone with nonempty interior. For given data $A \in M_{n,p}(\mathbf{R})$, $b \in \mathbf{R}^n$ and $c \in \mathbf{R}^p$, optimization problem of the form*

$$\min_{x \in \mathbf{R}^p} \{c^{\mathrm{T}} x : Ax - b \in K\} \tag{1}$$

is a conic problem (CP). When the data (A, b) belong to uncertain set U, the problem

$$\left\{ \min_{x \in \mathbf{R}^p} \{c^{\mathrm{T}} x : Ax - b \in K\} : (A, b) \in U \right\} \tag{2}$$

is called uncertain conic problem (UCP) and the problem

$$\min_{x \in \mathbf{R}^p} \{c^{\mathrm{T}} x : Ax - b \in K : \forall (A, b) \in U\} \tag{3}$$

is called robust counterpart (RC).

A feasible/optimal solution of (RC) is called a robust feasible/optimal solution of (UCP). Surely, the difficulty of problem is closely related to the uncertain set U which is

$$U = (A^0, b^0) + W$$

where (A^0, b^0) is a nominal data and W is a compact convex set, symmetric with respect to the origin.(W is interpreted as the perturbation set). If the uncertain set U is too complex, we need an approximation to bracket the optimal value of the problem in acceptable bounds. If the set $\mathcal{X}$ is the set of robust feasible solutions, then we can define it as

$$\mathcal{X} = \{x \in \mathbf{R}^p : Ax - b \in K \ \ \forall (A, b) \in (A^0, b^0) + W\}.$$

Also with an additional vector u, let the set $\mathcal{R}$ be

$$\mathcal{R} := \{(x, u) : Px + Qu + r \in \hat{K}\}$$

for a vector r, some matrices P and Q, and a pointed closed convex nonempty cone $\hat{K}$ with nonempty interior.

Definition 15 *$\mathcal{R}$ is an approximate robust counterpart of $\mathcal{X}$ if the projection of $\mathcal{R}$ onto the space of x-variables, i.e., the set $\hat{\mathcal{R}} \subseteq \mathbf{R}^p$ given by*

$$\hat{\mathcal{R}} := \{x : Px + Qu + r \in \hat{K} \textit{ for some } u\},$$

is contained in $\mathcal{X}$:

$$\hat{\mathcal{R}} \subseteq \mathcal{X}.$$

To measure the approximation error between $\hat{\mathcal{R}}$ and $\mathcal{X}$, one can shrink $\mathcal{X}$ until it fits into $\hat{\mathcal{R}}$. To do this, we should increase the size of uncertain set U as

$$U_\rho = \{(A^0, b^0) + \rho W\}, \ \ \rho \geq 1.$$

Then the new set of robust feasible solutions corresponding to U_ρ is:

$$\mathcal{X}_\rho = \{x \in \mathbf{R}^p : Ax - b \in K \ \ \forall (A, b) \in U_\rho\}.$$

If ρ is sufficiently large, the new robust feasible set becomes a subset of $\hat{\mathcal{R}}$. More precisely we have:

Definition 16 *The smallest ρ to obtain $\mathcal{X}_\rho \subseteq \hat{\mathcal{R}}$, i.e.*

$$\rho^* = inf_{\rho \geq 1}\{\rho : \mathcal{X}_\rho \subseteq \hat{\mathcal{R}}\},$$

is called the level of conservativeness of the approximate robust counterpart $\mathcal{R}$.

Finally we get

$$\mathcal{X}_\rho \subseteq \hat{\mathcal{R}} \subseteq \mathcal{X}.$$

After all of these definitions, now it is time to turn our interest to the uncertain quadratic constraint (it can also be written as a conic quadratic form):

$$x^{\mathrm{T}} A^{\mathrm{T}} A x \leq 2b^{\mathrm{T}} x + c \quad \forall (A, b, c) \in U_\rho,$$

where;

$$U_\rho = \left\{ (A, b, c) = (A^0, b^0, c^0) + \sum_{l=1}^{L} y_l (A^l, b^l, c^l) : y \in \rho V \right\},$$

and

$$V = \{y \in \mathbf{R}^L : y^{\mathrm{T}} Q_k y \leq 1, \ \ k = 1, \ldots, K\},$$

with $Q_k \succeq 0$ for each k and $\sum_{k=1}^{K} Q_k \succ 0$.

At this point, let us give an example to understand where the S-Lemma enters the system from the paper of Ben-Tal and Nemirovski (1998) (Theorem 3.2 in their paper).(It is also discussed in the paper of El-Ghaoui and Lebret (1997)). For the case $K = 1$, Q_1 is identity matrix:

Theorem 17 *For $A^l \in M_{n,p}(\mathbf{R})$, $b^l \in \mathbf{R}^p$, $c^l \in \mathbf{R}$, $l = 0, \ldots, L$ a vector $x \in \mathbf{R}^p$ is a solution of*

$$x^{\mathrm{T}} A^{\mathrm{T}} A x \le 2b^{\mathrm{T}} x + c \quad \forall (A, b, c) \in U_{simple}, \tag{4}$$

where

$$U_{simple} = \left\{ (A, b, c) = (A^0, b^0, c^0) + \sum_{l=1}^{L} y_l (A^l, b^l, c^l) : \|y\|^2 \le 1 \right\},$$

if and only if for some nonnegative λ, the pair (x, λ) is a solution of the following linear matrix inequality (LMI):

$$\left[\begin{array}{c|ccc|c} c^0 + 2x^{\mathrm{T}} b^0 - \lambda & \frac{1}{2}c^1 + x^{\mathrm{T}} b^1 & \cdots & \frac{1}{2}c^L + x^{\mathrm{T}} b^L & (A^0 x)^{\mathrm{T}} \\ \hline \frac{1}{2}c^1 + x^{\mathrm{T}} b^1 & \lambda & & & (A^1 x)^{\mathrm{T}} \\ \vdots & & \ddots & & \vdots \\ \frac{1}{2}c^L + x^{\mathrm{T}} b^L & & & \lambda & (A^L x)^{\mathrm{T}} \\ \hline (A^0 x) & (A^1 x) & \cdots & (A^L x) & I_n \end{array}\right] \succeq 0.$$

Proof Using uncertain set, (4) can be written as:

$$-x^{\mathrm{T}} \left[A^0 + \sum_{l=1}^{L} y_l A^l \right]^{\mathrm{T}} \left[A^0 + \sum_{l=1}^{L} y_l A^l \right] x + 2 \left[b^0 + \sum_{l=1}^{L} y_l b^l \right]^{\mathrm{T}} x + \left[c^0 + \sum_{l=1}^{L} y_l c^l \right] \ge 0 \quad \forall (y : \|y\|^2 \le 1).$$

Taking $\tau \le 1$,

$$-x^{\mathrm{T}} \left[A^0 \tau + \sum_{l=1}^{L} y_l A^l \right]^{\mathrm{T}} \left[A^0 \tau + \sum_{l=1}^{L} y_l A^l \right] x + 2\tau \left[b^0 \tau + \sum_{l=1}^{L} y_l b^l \right]^{\mathrm{T}} x + \tau \left[c^0 \tau + \sum_{l=1}^{L} y_l c^l \right] \ge 0 \quad \forall ((\tau, y) : \|y\|^2 \le \tau^2).$$

Clearly, If $\tau^2 - \|y\|^2 \ge 0$ then the first inequality holds. Now the S-Lemma enters the system and links these inequalities because both sides can be written as a single matrix. From S-Lemma, we can write

$$-x^{\mathrm{T}} \left[A^0 \tau + \sum_{l=1}^{L} y_l A^l \right]^{\mathrm{T}} \left[A^0 \tau + \sum_{l=1}^{L} y_l A^l \right] x + 2\tau \left[b^0 \tau + \sum_{l=1}^{L} y_l b^l \right]^{\mathrm{T}} x + \tau \left[c^0 \tau + \sum_{l=1}^{L} y_l c^l \right] - \lambda(\tau^2 - \|y\|^2) \ge 0$$

which is the same as

$$(\tau, y^{\mathrm{T}}) \left[\begin{pmatrix} c^0 + 2x^{\mathrm{T}}b^0 & \frac{1}{2}c^1 + x^{\mathrm{T}}b^1 & . & . & \frac{1}{2}c^L + x^{\mathrm{T}}b^L \\ \frac{1}{2}c^1 + x^{\mathrm{T}}b^1 & & & & \\ . & & & & \\ . & & & & \\ \frac{1}{2}c^L + x^{\mathrm{T}}b^L & & & & \end{pmatrix} \right.$$

$$\left. - \begin{pmatrix} (A^0x)^{\mathrm{T}} \\ (A^1x)^{\mathrm{T}} \\ . \\ . \\ (A^Lx)^{\mathrm{T}} \end{pmatrix} (A^0x, A^1x, .., A^Lx) \right] \begin{pmatrix} \tau \\ y \end{pmatrix}$$

$$+(\tau, y^{\mathrm{T}}) \begin{pmatrix} -\lambda & & & & \\ & \lambda & & & \\ & & . & & \\ & & & . & \\ & & & & \lambda \end{pmatrix} \begin{pmatrix} \tau \\ y \end{pmatrix} \geq 0.$$

From Schur lemma given below, we obtain the matrix in the theorem. This completes the proof.

Lemma 18 (*Schur complement lemma*) *Suppose A, B, C, D are respectively $n \times n$, $n \times p$, $p \times n$ and $p \times p$ matrices, and D is invertible. Let*

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

so that M is a $(n + p) \times (n + p)$ matrix. Then the Schur complement of the block D of the matrix M is the $n \times n$ matrix

$$A - BD^{-1}C.$$

Let D be positive definite. Then M is positive semi-definite if and only if the Schur complement of D in M is positive semi-definite.

Clearly, the proof completely depends on the S-Lemma. However the S-Lemma works only for a single quadratic form. Therefore we need a somehow different theorem that also works for the cases $K > 1$. Although it does not give an equivalence result as above, it gives reasonable bounds for us to work on more complicated problems. Now it is time to state this lemma and to see how it works.

Ben-Tal et al. (2002) proved the following result; see Lemma A.6, pp. 554–559. (Ben-Tal et al. also showed that the approximate S-Lemma implies the usual S-Lemma).

Lemma 19 (*Approximate S-Lemma*). *Let R, R_0, $R_1, \ldots, R_k$ be symmetric $n \times n$ matrices such that*

$$R_1, \ldots, R_k \succeq 0, \tag{5}$$

and assume that

$$\exists \lambda_0, \lambda_1, \ldots, \lambda_k \geq 0 \; s.t. \sum_{k=0}^{K} \lambda_k R_k \succ 0. \tag{6}$$

Consider the following quadratically constrained quadratic program,

$$\mathrm{QCQ} = \max_{y \varepsilon R^n} \{ y^{\mathrm{T}} R y : y^{\mathrm{T}} R_0 y \le r_0, y^{\mathrm{T}} R_k y \le 1, k = 1, \ldots, K \} \tag{7}$$

and the semidefinite optimization problem

$$\mathrm{SDP} = \min_{\mu_0, \mu_1, \ldots, \mu_K} \{ r_0 \mu_0 + \textstyle\sum_{k=1}^{K} \mu_k : \sum_{k=0}^{K} \mu_k R_k \succeq R, \mu \ge 0 \}. \tag{8}$$

Then

(*i*) *If problem* (7) *is feasible, then problem* (8) *is bounded below and*

$$\mathrm{SDP} \ge \mathrm{QCQ}. \tag{9}$$

Moreover, there exists $y_* \in \mathbf{R}^n$ *such that*

$$y_*^{\mathrm{T}} R y_* = \mathrm{SDP}, \tag{10}$$

$$y_*^{\mathrm{T}} R_0 y_* \le r_0, \tag{11}$$

$$y_*^{\mathrm{T}} R_k y_* \le \tilde{\rho}^2, \ k = 1, \ldots, K, \tag{12}$$

where

$$\tilde{\rho} := (2 \log(6 \sum_{k=1}^{K} rank\ R_k))^{\frac{1}{2}}, \tag{13}$$

if R_0 *is a dyadic matrix (that can be written on the form* xx^{T}, $x \in \mathbf{R}^n$*) and*

$$\tilde{\rho} := (2 \log(16 n^2 \sum_{k=1}^{K} rank\ R_k))^{\frac{1}{2}} \tag{14}$$

otherwise.

(*ii*) *If*

$$r_0 > 0, \tag{15}$$

then (7) *is feasible, problem* (8) *is solvable, and*

$$0 \le \mathrm{QCQ} \le \mathrm{SDP} \le \tilde{\rho}^2 \mathrm{QCQ}. \tag{16}$$

After giving the approximate S-Lemma, now we are ready to work on more complicated uncertainty sets which are e.g., the cases $K > 1$, from the paper of Ben-Tal et al. (2002). Let us begin by defining the corresponding robust feasible set:

$$\mathcal{X}_\rho = \{ x : x^{\mathrm{T}} A^{\mathrm{T}} A x \le 2 b^{\mathrm{T}} x + c \quad \forall (A, b, c) \in U_\rho \},$$

where

$$U_\rho = \left\{ (A, b, c) = (A^0, b^0, c^0) + \rho \sum_{l=1}^{L} y_l (A^l, b^l, c^l) : y^{\mathrm{T}} Q_k y \le 1, \ k = 1, \ldots, K \right\}.$$

Note that the robust counterpart of uncertain quadratic constraint with the intersection-of-ellipsoids ($\cap$-ellipsoid) uncertainty U_ρ is, in general NP-hard to form. In fact, not only this, but also the problem of robust feasibility check is NP-hard. (Ben-Tal et al. 2002, pp. 539).

To combine the sets of $\mathcal{X}_\rho$ and U_ρ, we need additional notation:

$$a[x] = A^0 x,\ c[x] = 2x^{\mathrm{T}} b^0 + c^0,\ A_\rho[x] = \rho(A^1 x, \ldots, A^L x),$$

and

$$b_\rho[x] = \rho \begin{bmatrix} x^{\mathrm{T}} b^1 \\ \cdot \\ \cdot \\ \cdot \\ x^{\mathrm{T}} b^L \end{bmatrix},\ d_\rho = \frac{1}{2}\rho \begin{bmatrix} c^1 \\ \cdot \\ \cdot \\ \cdot \\ c^L \end{bmatrix}.$$

Then one may easily verify that $x \in \mathcal{X}^\rho$ holds if and only if

$$y^{\mathrm{T}} Q_k y \le 1, k = 1, \ldots, K$$
$$\Rightarrow (a[x] + A_\rho[x]y)^{\mathrm{T}}(a[x] + A_\rho[x]y) \le 2(b_\rho[x] + d_\rho)^{\mathrm{T}} y + c[x].$$

If y satisfies the above, $-y$ also does. Therefore we can write:

$$y^{\mathrm{T}} Q_k y \le 1, k = 1, \ldots, K$$
$$\Rightarrow y^{\mathrm{T}} A_\rho[x]^{\mathrm{T}} A_\rho[x] y \pm 2y^{\mathrm{T}}(A_\rho[x]^{\mathrm{T}} a[x] - b_\rho[x] - d_\rho) \le c[x] - a[x]^{\mathrm{T}} a[x].$$

If we take the $t^2 \le 1$, the inequality can be rewritten as

$$t^2 \le 1,\ y^{\mathrm{T}} Q_k y \le 1, k = 1, \ldots, K$$
$$\Rightarrow y^{\mathrm{T}} A_\rho[x]^{\mathrm{T}} A_\rho[x] y + 2t y^{\mathrm{T}}(A_\rho[x]^{\mathrm{T}} a[x] - b_\rho[x] - d_\rho) \le c[x] - a[x]^{\mathrm{T}} a[x].$$

If there exists $\lambda_k \ge 0, k = 1, \ldots, K$, we can join these inequalities such that for all t and for all y:

$$\sum_{k=1}^{K} \lambda_k y^{\mathrm{T}} Q_k y + \left(c[x] - a[x]^{\mathrm{T}} a[x] - \sum_{k=1}^{K} \lambda_k \right) t^2$$
$$\ge y^{\mathrm{T}} A_\rho[x]^{\mathrm{T}} A_\rho[x] y + 2t y^{\mathrm{T}}(A_\rho[x]^{\mathrm{T}} a[x] - b_\rho[x] - d_\rho).$$

Surely, our new inequality needs more conditions than the first one. Therefore if the last inequality holds, then the previous one also holds. If we write our inequality in matrix form, we obtain

$$\exists \lambda \ge 0 \text{ s.t. } \begin{bmatrix} c[x] - a[x]^{\mathrm{T}} a[x] - \sum_{k=1}^{K} \lambda_k & (A_\rho[x]^{\mathrm{T}} a[x] - b_\rho[x] - d_\rho)^{\mathrm{T}} \\ (A_\rho[x]^{\mathrm{T}} a[x] - b_\rho[x] - d_\rho) & \sum_{k=1}^{K} \lambda_k Q_k - A_\rho[x]^{\mathrm{T}} A_\rho[x] \end{bmatrix} \succeq 0.$$

From the Schur complement lemma, we obtain the following theorem:

Theorem 20 *The set $\mathcal{R}_\rho$ of (x, λ) satisfying $\lambda \geq 0$ and*

$$\begin{bmatrix} c[x] - \sum_{k=1}^{K} \lambda_k & (-b_\rho[x] - d_\rho)^{\mathrm{T}} & a[x]^{\mathrm{T}} \\ (-b_\rho[x] - d_\rho) & \sum_{k=1}^{K} \lambda_k Q_k & -A_\rho[x]^{\mathrm{T}} \\ a[x] & -A_\rho[x] & I_M \end{bmatrix} \succeq 0 \tag{17}$$

is an approximate robust counterpart of the set $\mathcal{X}_\rho$ of robust feasible solutions of uncertain quadratic constraint.

Although we obtained an approximate robust counterpart we still do not know the level of conservativeness of this set. Now, we will see the relationship between level of conservativeness and approximate S-Lemma.

Theorem 21 *The level of conservativeness of the approximate robust counterpart $\mathcal{R}$ (as given by 17) of the set $\mathcal{X}$ is at most*

$$\tilde{\rho} := (\, 2 \log(6 \sum_{k=1}^{K} \text{rank } R_k))^{\frac{1}{2}}, \tag{18}$$

Proof We have to show that when x cannot be extended to a solution (x, λ), then there exists $\zeta_* \in \mathbf{R}^n$ such that

$$\zeta_*^{\mathrm{T}} Q_k \zeta_* \leq 1, \quad k = 1, \ldots, K \tag{19}$$

and

$$\tilde{\rho}^2 \zeta_*^{\mathrm{T}} A_\rho[x]^{\mathrm{T}} A_\rho[x] \zeta_* + 2\tilde{\rho} \zeta_*^{\mathrm{T}} (A_\rho[x]^{\mathrm{T}} a[x] - b_\rho[x] - d_\rho) > c[x] - a[x]^{\mathrm{T}} a[x]. \tag{20}$$

The proof is based on approximate S-Lemma, so we need to work with the following notation. Let

$$R = \left[\begin{array}{c|c} 0 & (A_\rho[x]^{\mathrm{T}} a[x] - b_\rho[x] - d_\rho)^{\mathrm{T}} \\ \hline A_\rho[x]^{\mathrm{T}} a[x] - b_\rho[x] - d_\rho & A_\rho[x]^{\mathrm{T}} A_\rho[x] \end{array} \right],$$

$$R_0 = \left[\begin{array}{c|c} 1 & 0^{\mathrm{T}} \\ \hline 0 & 0 \end{array} \right], \; R_k = \left[\begin{array}{c|c} 0 & 0^{\mathrm{T}} \\ \hline 0 & Q_k \end{array} \right],$$

and $r_0 = 1$. Note that $R_1, \ldots, R_K$ are positive semidefinite, and

$$R_0 + \sum_{k=1}^{K} R_k = \left[\begin{array}{c|c} 1 & 0^{\mathrm{T}} \\ \hline 0 & \sum_{k=1}^{K} Q_k \end{array} \right] \succ 0.$$

Therefore conditions of approximate S-Lemma are satisfied, the estimate is valid.

Case I In the first case we will prove that the following two conditions cannot appear at the same time for our case written at the beginning of the proof. Inequalities are:

$$R \preceq \sum_{k=0}^{K} \lambda_k R_k, \tag{21}$$

$$\sum_{k=0}^{K} \lambda_k \leq c[x] - a[x]^{\mathrm{T}} a[x]. \tag{22}$$

Note Ben-Tal et al. try to prove this case by claiming: assumption that x cannot be extended to a solution of (17) implies that x cannot be extended to a solution of uncertain quadratic constraint. However this claim is erroneous because the uncertain quadratic constraint set is larger than the set (17). Therefore, x cannot be extended to a solution of (17), but may be extended to a solution of uncertain quadratic constraint. Hence we change this part of the proof and we claim that these two inequalities cause x to be a solution of (17), which contradicts our assumption.

Let us turn to the proof with the new claim. Assume that there exist $\lambda_0, \ldots, \lambda_k \geq 0$ such that

$$R \prec \sum_{k=0}^{K} \lambda_k R_k,$$

$$\sum_{k=0}^{K} \lambda_k \leq c[x] - a[x]^{\mathrm{T}} a[x].$$

From assumption x cannot be extended to a solution of (17). On the other hand, we have

$$(t, y^{\mathrm{T}}) R \begin{pmatrix} t \\ y \end{pmatrix} \leq \sum_{k=0}^{K} \lambda_k (t, y^{\mathrm{T}}) R_k \begin{pmatrix} t \\ y \end{pmatrix} \quad \forall t, y$$

or

$$(t, y^{\mathrm{T}}) \begin{pmatrix} 0 & (A_p[x]^{\mathrm{T}} a[x] - b_p[x] - d_p)^{\mathrm{T}} \\ (A_p[x]^{\mathrm{T}} a[x] - b_p[x] - d_p) & A_p[x]^{\mathrm{T}} A_p[x] \end{pmatrix} \begin{pmatrix} t \\ y \end{pmatrix}$$
$$\leq \lambda_0 t^2 + \sum_{k=1}^{K} \lambda_k y^{\mathrm{T}} Q_k y$$

or, equivalently

$$\lambda_0 t^2 + \sum_{k=1}^{K} \lambda_k y^{\mathrm{T}} Q_k y - 2t y^{\mathrm{T}} (A_p[x]^{\mathrm{T}} a[x] - b_p[x] - d_p) - y^{\mathrm{T}} A_p[x]^{\mathrm{T}} A_p[x] y \geq 0. \tag{23}$$

We know that

$$\sum_{k=0}^{K} \lambda_k \leq c[x] - a[x]^{\mathrm{T}} a[x],$$

$$\lambda_0 + \sum_{k=1}^{K} \lambda_k \leq c[x] - a[x]^{\mathrm{T}} a[x],$$

$$\lambda_0 \leq c[x] - a[x]^{\mathrm{T}} a[x] - \sum_{k=1}^{K} \lambda_k.$$

From (23) and taking $-t$ instead of t, we obtain

$$\left(c[x] - a[x]^{\mathrm{T}}a[x] - \sum_{k=1}^{K}\lambda_k\right)t^2$$
$$+ \sum_{k=1}^{K}\lambda_k y^{\mathrm{T}} Q_k y + 2ty^{\mathrm{T}}\left(A_p[x]^{\mathrm{T}}a[x] - b_p[x] - d_p\right) - y^{\mathrm{T}}A_p[x]^{\mathrm{T}}A_p[x]y \geq 0,$$

or,

$$\exists\lambda \geq 0, \text{ s.t. } (t, y^{\mathrm{T}})\begin{pmatrix} c[x]-a[x]^{\mathrm{T}}a[x]-\sum_{k=1}^{K}\lambda_k & (A_p[x]^{\mathrm{T}}a[x]-b_p[x]-d_p)^{\mathrm{T}} \\ (A_p[x]^{\mathrm{T}}a[x]-b_p[x]-d_p) & \sum_{k=1}^{K}\lambda_k Q_k - A_p[x]^{\mathrm{T}}A_p[x] \end{pmatrix}\begin{pmatrix} t \\ y \end{pmatrix} \geq 0, \forall t, y.$$

However x is extended to a solution of (17), so it contradicts with our assumption. Case I cannot occur.

Case II There is no $\lambda_0, \ldots, \lambda_K \geq 0$ that satisfies both (21) and (22). Hence from approximate S-Lemma:

$$\text{SDP} > c[x] - a[x]^{\mathrm{T}}a[x]. \tag{24}$$

There exists $y_* = (t_*, \eta_*)$ such that

$$\begin{aligned} y_*^{\mathrm{T}} R_0 y_* &= t_*^2 \leq r_0 = 1, \\ y_*^{\mathrm{T}} R_k y_* &= \eta_*^{\mathrm{T}} Q_k \eta_* \leq \tilde{\rho}^2, \quad k = 1, \ldots, K, \\ y_*^{\mathrm{T}} R y_* &= \eta_*^{\mathrm{T}} A_\rho[x]^{\mathrm{T}} A_\rho[x]\eta_* + 2t_*\eta_*^{\mathrm{T}}(A_\rho[x]^{\mathrm{T}}a[x] - b_\rho[x] - d_\rho) = \text{SDP} \\ &> c[x] - a[x]^{\mathrm{T}}a[x], \end{aligned}$$

from (24). Setting $\overline{\eta} = \tilde{\rho}^{-1}\eta_*$, these inequalities turn into

$$|t_*| \leq 1,$$
$$\overline{\eta}^{\mathrm{T}} Q_k \overline{\eta} \leq 1, \quad k = 1, \ldots, K,$$
$$\tilde{\rho}^2\overline{\eta}^{\mathrm{T}} A_\rho[x]^{\mathrm{T}} A_\rho[x]\overline{\eta} + 2\tilde{\rho}\overline{\eta}^{\mathrm{T}} t_*(A_\rho[x]^T a[x] - b_\rho[x] - d_\rho)$$
$$> c[x] - a[x]^{\mathrm{T}}a[x].$$

If $(t_*, \overline{\eta})$ is a solution of this system, then $\zeta_* = \overline{\eta}$ or $\zeta_* = -\overline{\eta}$ is a solution of (19), (20). This completes the proof.

Although the background on S-Lemma, S-procedure and approximate S-Lemma is vast, we tried to give the main theorems we deemed important here and explain them by giving some examples. In the next section, we give some results that strongly rely on these theorems.

3 The extended S-procedure

We defined the Extended S-procedure (8) in the previous section. Now we prove some related results by using the Barvinok, and Au-Yeung and Poon theorems.

3.1 Corollary for Barvinok's Theorem (1995)

In this subsection, we deal with changing Barvinok's result into the form of an extended S-procedure. If we define the function $f(X)$ whose ith component is $f_i(X) = (\langle\langle A_i X, X\rangle\rangle$, with $i = 0, 1, \ldots, m-1$ and $X \in M_{n,p}(\mathbf{R})$, then the theorem of Barvinok can be written as:

Theorem 22 *Let $A_0, A_1, \ldots, A_{m-1} \in S_n^{\mathbf{R}}$, and let $p := \lfloor \frac{\sqrt{8m+1}-1}{2} \rfloor$. Then*

$$\{(f_0(X), f_1(X), \ldots, f_{m-1}(X)) | X \in M_{n,p}(\mathbf{R})\}$$

is a convex cone of $\mathbf{R}^m$.

By using separation lemma of convex analysis, we obtain the following corollary:

Corollary 23 *Let $A_0, A_1, \ldots, A_{m-1} \in S_n^{\mathbf{R}}$, and let $p := \lfloor \frac{\sqrt{8m+1}-1}{2} \rfloor$. Assume there exists $X^0 \in M_{n,p}(\mathbf{R})$, such that*

$$f_i(X^0) = (\langle\langle A_i X^0, X^0\rangle\rangle > 0, \quad i = 1, \ldots, m-1. \tag{25}$$

Then

$$f_0(X) \geq 0 \; \forall X : f_i(X) \geq 0, \quad i = 1, \ldots, m-1. \tag{26}$$

holds if and only if there exists $\tau_i \geq 0$ for $i = 1, \ldots, m-1$:

$$f_0(X) \geq \sum_{i=1}^{m-1} \tau_i f_i(X). \tag{27}$$

Proof We proceed exactly as in the proof of the S-Lemma (Theorem 2). Since the sufficiency part is again easy, we concentrate on the necessity. Let

$$S = \{(\langle\langle\langle A_0 X, X\rangle\rangle, \langle\langle A_1 X, X\rangle\rangle, \ldots, \langle\langle A_{m-1} X, X\rangle\rangle) : X \in M_{n,p}(\mathbf{R})\}$$

and

$$U = \mathbf{R}_{--} \times \mathbf{R}_+^{m-1}.$$

S is a convex set by Barvinok's theorem (Theorem 7). Since the intersection of S and U is empty, a separating hyperplane exists. I.e., there exists nonzero $c = (c_0, c_1, \ldots, c_{m-1}) \in \mathbf{R}^m$, such that $(c, s) \leq 0, \forall s \in S$ and $(c, u) \geq 0$, $\forall u \in U$. Using similar arguments to those in the proof of Theorem 2 we obtain $c_0 \leq 0$, and $c_1 \geq 0, \ldots, c_{m-1} \geq 0$. From first inequality, for $\forall X \in M_{n,p}(\mathbf{R})$,

$$c_0\langle\langle A_0 X, X\rangle\rangle + c_1\langle\langle A_1 X, X\rangle\rangle + \cdots + c_{m-1}\langle\langle A_{m-1} X, X\rangle\rangle \leq 0.$$

We know that there exists X^0 such that $f_i(X^0) = (\langle\langle A_i X^0, X^0\rangle\rangle > 0$ and $c_i \geq 0$ for $i = 1, \ldots, m-1$, so c_0 cannot be zero by arguments similar to those used in the proof of Theorem 2. Therefore, defining $\tau_i = -\frac{c_i}{c_0}$, we obtain:

$$f_0(X) \geq \sum_{i=1}^{m-1} \tau_i f_i(X).$$

The proof is complete.

Clearly, there exists a link between the S-procedure and convexity provided by the separation lemma.

3.2 Corollary for Au-Yeung and Poon (1979) and Poon's Theorem (1997)

The next theorem we deal with is the theorem of Au-Yeung and Poon(1979) that strongly relies on Bohnenblust's unpublished paper. With same definition of $f(X)$ as in the first corollary, we can rewrite this theorem as follows.

Theorem 24 *Let the integer p be defined as*

$$p := \left\{ \begin{array}{ll} \lfloor \frac{\sqrt{8(m-1)+1}-1}{2} \rfloor & \text{if } \frac{n(n+1)}{2} \neq m \\ \lfloor \frac{\sqrt{8(m-1)+1}-1}{2} \rfloor + 1 & \text{if } \frac{n(n+1)}{2} = m \end{array} \right\},$$

and $A_0, \ldots, A_{m-1} \in S_n^{\mathbf{R}}$. *Then,*

$$\{(f_0(X), f_1(X), \ldots, f_{m-1}(X)) | X \in M_{n,p}(\mathbf{R}), \|X\| = 1\}$$

is a convex compact subset of $\mathbf{R}^m$.

First, we establish the following corollary by using the procedure of Polyak's proof in the paper Polyak (1998):

Corollary 25 *Let* $A_0, A_1, \ldots, A_m \in S_n^{\mathbf{R}}$, *and let p be defined as in theorem of Au-Yeung and Poon. Also* $f_i(X) = (\langle\langle A_i X, X \rangle\rangle$, *with* $i = 0, 1, \ldots, m$. *If there exists* $\mu \in \mathbf{R}^{m+1}$ *such that;*

$$\sum_{i=0}^{m} \mu_i f_i(X) > 0, \quad i = 0, \ldots, m, \tag{28}$$

then the set

$$F = \{(f_0(X), f_1(X), \ldots, f_m(X)) | X \in M_{n,p}(\mathbf{R})\}$$

is convex.

Proof We proceed as in Polyak (1998). If $f \in F$, $f = f(X) = (f_0(X), f_1(X), \ldots, f_m(X))$, for $\lambda > 0$, then $\lambda f = f(\sqrt{\lambda} X) \in F$, thus F is a cone.

With respect to linear transformations of a space, the convexity property is invariant. Also, a linear combination of quadratic forms is a quadratic form. Therefore there exists a linear map $g = Tf$ such that $g_m = \sum_{i=0}^{m} \mu_i f_i(X) > 0$ and $G = \{g(X) : X \in M_{n,p}(\mathbf{R})\}$ is convex if and only if F is convex.

Also by making a nonsingular linear transformation (it does not change G), we can assume that $g_m = \|X\|^2$ where $\|X\|^2 = \sum_{i=1}^{p} \|x_i\|^2$ with $n \times 1$ vectors x_i. We know that from Polyak's paper it is nonsingular linear transformation when X is a one column matrix. Therefore we have nothing but summation of nonsingular linear transformations which is also in this case a nonsingular linear transformation. (It has still the characteristic of being injective, $\|X\|^2 = 0 \Leftrightarrow X = 0$, and of being surjective as its range equals $\mathbf{R}_+ \cup \{\mathbf{0}\}$). From the Theorem of Au-Yeung and Poon we have

$$H = \{((g_0(X), g_1(X), \ldots, g_{m-1}(X)))^{\mathrm{T}} | X \in M_{n,p}(\mathbf{R}), \|X\| = 1\}$$

is convex, but also $G = \{\lambda Q, \lambda \geq 0\}$ where

$$Q = \{(h_0, h_1, \ldots, h_{m-1}, 1)^{\mathrm{T}} : h \in H\}.$$

Hence, G is convex. Therefore F is convex.

The previous result leads to the following corollary.

Corollary 26 *Let* $A_0, A_1, \ldots, A_m \in S_n^{\mathbf{R}}$*, and let*

$$p := \left\{ \begin{array}{ll} \lfloor \frac{\sqrt{8m+1}-1}{2} \rfloor & \text{if } \frac{n(n+1)}{2} \neq m+1 \\ \lfloor \frac{\sqrt{8m+1}-1}{2} \rfloor + 1 & \text{if } \frac{n(n+1)}{2} = m+1 \end{array} \right\}.$$

Assume there exists $X^0 \in M_{n,p}(\mathbf{R})$*, such that*

$$f_i(X^0) = (\langle\langle A_i X^0, X^0 \rangle\rangle > 0, \quad i = 1, \ldots, m. \tag{29}$$

and that

$$\sum_{i=1}^{m} \mu_i f_i(X) > 0. \tag{30}$$

Then

$$f_0(X) \geq 0 \quad \forall X : \quad f_i(X) \geq 0, \quad i = 1, \ldots, m. \tag{31}$$

holds if and only if there exists $\tau_i \geq 0$ *for* $i = 1, \ldots, m$*:*

$$f_0(X) \geq \sum_{i=1}^{m} \tau_i f_i(X). \tag{32}$$

Proof The proof is identical to that in the corollary of Barvinok's theorem given above.

These corollaries are extended versions of Yakubovich and Polyak's S-procedures. However none of them gives a better solution for the case $p = 1$. In other words, we still fall back to the classical results when X is a one column matrix.

4 Further research on approximate S-Lemma

In this section, we summarize briefly our efforts to improve bounds of the approximate S-Lemma. For the dyadic case which is of interest for robust optimization, we obtained only a partial result.

Ben-Tal et al. (2002) give the following conjecture to improve the dyadic case which is the main ingredient for proving approximation results in robust quadratically constrained programs and conic quadratic programs.

Conjecture *Let* $x = \{x_1, \ldots, x_n\}$ *and* $\xi = \{\xi_1, \ldots, \xi_n\} \in \mathbf{R}^n$*. If* $\|x\|_2 = 1$ *and the coordinates* ξ_i *of* ξ *are independently identically distributed random variables with*

$$Pr(\xi_i = 1) = Pr(\xi_i = -1) = \frac{1}{2} \tag{33}$$

then one has

$$Pr(|\xi^{\mathrm{T}} x| \leq 1) \geq \frac{1}{2}. \tag{34}$$

This conjecture improves the bound to 1/2 from 1/3. We worked on this conjecture by using n-dimensional geometry. However, we only proved the following relaxed version of it Derinkuyu (2004):

Lemma 27 *Let* $x = \{x_1, \ldots, x_n\}$ *and* $\xi = \{\xi_1, \ldots, \xi_n\} \in \mathbf{R}^n$. *If* $\|x\|_2 = 1$ *and* $\|\xi\|_2^2 = n$ *then one has*

$$Pr(|\xi^{\mathrm{T}} x| \le 1) \ge \frac{1}{2}. \tag{35}$$

This lemma is a relaxed version of the above conjecture, because the vectors ξ are equally distributed on the surface of hyper-sphere of $\|\xi\|_2^2 = n$. The conjecture states that for any x, at least half of the vectors satisfies the inequality. However, we proved that for any x, half of the surface of the hyper-sphere satisfies the inequality. We also proved the opposite side of it. In other words, for any ξ, half of the surface of the hyper-sphere of x, which is $\|x\|_2 = 1$, satisfies the inequality. Since the proof is long and quite involved we omit it here.

5 Discussion

In this section, we give a critical evaluation of our results on extended S-Lemma and approximate S-Lemma.

For extended S-Lemma, we developed two corollaries from theorems of Barvinok and Poon. Although they resemble each other, we can get a better result from corollary of Poon if we have positive linear combination of given matrices.

For the corollary of Barvinok's theorem, the relationship between p and m is $p := \lfloor \frac{\sqrt{8m+1}-1}{2} \rfloor$. On the other hand, in the corollary of Poon's result, we have:

$$p := \left\{ \begin{array}{ll} \lfloor \frac{\sqrt{8(m-1)+1}-1}{2} \rfloor & \text{if } \frac{n(n+1)}{2} \neq m \\ \lfloor \frac{\sqrt{8(m-1)+1}-1}{2} \rfloor + 1 & \text{if } \frac{n(n+1)}{2} = m \end{array} \right\},$$

However, one needs additional assumption in the second case. In fact this assumption is equivalent to assuming positive definiteness of a linear combination of matrices. One can reach this result by observing $\langle\langle AX, X \rangle\rangle = \sum_{i=1}^{p} x^{\mathrm{T}} A x$ for $X \in M_{n,p}, x \in \mathbf{R}^n$. To obtain this positive definiteness, the corollary of Poon's theorem is given by Hiriart-Urruty and Torki that we explained in the background section. Also Polyak gives an analysis for $m = 2$ case. For generalization of this result, Uhlig's survey is a useful paper.

Although we extended the S-Lemma, it does not improve the S-Lemma of Yakubovich or Polyak for the cases $X \in M_{n,1}$. (Note that the corollary of Poon's result gives $m = 3$ for $p = 1$. It corresponds to quadratic function over two quadratic constraints in the S-procedure). Therefore, we have still problems for $m > 2$.

The improvement in the approximate S-Lemma defied our efforts and remains a difficult open problem.

6 Concluding remarks

In this study, we dealt with S-procedure and some of its variants that remain fundamental tools of different fields such as control theory and robust optimization. In general, S-procedure corresponds to verifying that the minimum of a non-convex function over a non-convex set is positive. This problem belongs in general to the class of NP-complete problems. Hence, to prove new theorems either in S-procedure by extending or giving extra assumptions or in approximate S-Lemma by narrowing the bounds will be valuable assets for the optimization and control communities.

For general case, we dealt with corollaries of the theorems of Barvinok and Poon to understand their meaning for S-procedure. This also highlighted the relationship between convex and quadratic worlds. In the corollary of Barvinok, we obtain the extended version of Yakubovich's theorem. However it does not give any improvement for classical vector case. On the other hand, we obtain a better result in the corollary of Poon's theorem, if we make an assumption of positive definiteness of a linear combination of matrices. This corollary also gives the same result as Polyak's theorem for classical vector case.

In the case of S-procedure, the best result due to Polyak is about $m = 2$ case. Polyak shows counterexamples in his paper that the assumptions he gives are not enough for the $m > 2$ case. Therefore we need additional assumptions to prove new results on $m > 2$ case. The problem in this area is to obtain the minimal assumptions satisfying the case $m > 2$. This problem is still open.

Then, we turned our interest into the approximate S-Lemma where our efforts failed to improve the result in the dyadic case, which is the case of interest for robust optimization. This also remains a major open problem.

References

Au-Yeung YH (1969) A theorem on a mapping from a sphere to the circle and the simultaneous diagonalisation of two hermitian matrices. Proc Am Math Soc 20:545–548

Au-Yeung YH, Poon YT (1979) A remark on the convexity and positive definiteness concerning hermitian matrices. Southeast Asian Bull Math 3:85–92

Barvinok AI (1995) Problems of distance geometry and convex properties of quadratic maps. Discrete Comput Geom 13:189–202

Ben-Tal A (2002) Conic and robust optimization. Technical report, Israel Institute of Technology, Technion

Ben-Tal A, Goryashko A, Guslitzer E, Nemirovski A (2004) Adjustable robust solutions to uncertain linear programs. Math Prog 99:351–376

Ben-Tal A, Nemirovski A (1998) Robust convex optimization. Math Oper Res 23:769–805

Ben-Tal A, Nemirovski A (2000) Lectures on modern convex optimization: analysis, algorithms and engineering applications. SIAM-MPS, Philadelphia

Ben-Tal A, Nemirovski A, Roos C (2002) Robust solutions of uncertain quadratic and conic-quadratic problems. SIAM J Optim 13:535–560

Boyd S, El-Ghaoui L, Feron E, Balakhrishnan V (1994) Linear matrix inequalities in systems and control theory. SIAM, Philadelphia

Brickman L (1961) On the fields of values of a matrix. Proc Am Math Soc 12:61–66

Calabi E (1964) Linear systems of real quadratic forms. Proc Am Math 15:844–846

Derinkuyu K (2004) On the s-procedure and some variants. Unpublished M.Sc. Thesis, Bilkent University

Dines LL (1941) On the mapping of quadratic forms. Bull Am Math Soc 47:494–498

Dines LL (1942) On the mapping of n quadratic forms. Bull Am Math Soc 48:467–471

Dines LL (1943) On linear combinations of quadratic forms. Bull Am Math Soc 49:388–393

El-Ghaoui L, Lebret H (1997) Robust solutions to least-squares problems with uncertain data. SIAM J Matrix Anal Appl 18(4):1035–1064

El-Ghaoui L, Oustry F, Lebret H (1998) Robust solutions to uncertain semidefinite programs. SIAM J Optim 9:33–52

Finsler P (1936/37) Über das vorkommen definiter und semidefiniter formen in scharen quadratischer formen. Comment Math Helv 9:188–192
Fradkov AL (1973) Duality theorems for certain nonconvex extremum problems. Siberian Math J 14:247–264
Hausdorff F (1919) Der wertvorrat einer bilinearform. Math Z 3:314–316
Hestenes MR, McShane EJ (1940) A theorem on quadratic forms and its application in the calculus of variations. Trans Am Math Soc 40:501–512
Hiriart-Urruty JB, Torki M (2002) Permanently going back and forth between the quadratic world and the convexity world in optimization. Appl Math Optim 45:169–184
Horn RA, Johnson CR (1990) Matrix analysis. Cambridge University Press, New York
John F (1938) A note on the maximum principle for elliptic differential equations. Bull Am Math Soc 44:268–271
Kühne R (1964) Über eine klasse j-selbstadjungierter operatoren. Math Ann 154:56–69
Luo Z-Q, Sturm J, Zhang S-Z (2003) Multivariate nonnegative quadratic mappings. Technical report, Chinese University of Hong-Kong
Lur'e AI, Postnikov VN (1944) On the theory of stability of control systems (in Russian). Appl Math Mech 8(3)
Megretsky A, Treil S (1993) Power distribution inequalities in optimization and robustness of uncertain systems. Math Syst Estimation Control 3:301–319
Nemirovski A (2002) Five lectures on modern convex optimization. Technical report, Israel Institute of Technology, Technion
Peng JM, Yuan Y-X (1997) Optimality conditions for the minimization of a quadratic with two quadratic constraints. SIAM J Optim 7:579–594
Polyak BT (1998) Convexity of quadratic transformations and its use in control and optimization. J Optim Theory Appl 99:553–583
Poon YT (1997) Generalized numerical ranges, joint positive definiteness and multiple eigenvalues. Proc Am Math Soc 125:1625–1634
Taussky O (1967) Positive-definite matrices. In: Shisha O (ed) Inequalities. Academic, New York, pp 309–319
Toeplitz O (1918) Das algebraische analogen zu einem satze von fejér. Math Z 2:187–197
Uhlig F (1979) A recurring theorem about pairs of quadratic forms and extension: a survey. Linear Algebra Appl 25:219–237
Yakubovich VA (1971) The s-procedure in nonlinear control theory. Vestnik Leningr Univ 4: 73–93 in Russian, 1971, No.1, 62–77
Yuan Y (1990) On a subproblem of trust region algorithms for constrained optimization. Math Program 47:53–63