ON THE S-PROCEDURE AND SOME VARIANTS

A THESIS

SUBMITTED TO THE DEPARTMENT OF INDUSTRIAL ENGINEERING

AND THE INSTITUTE OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Kürşad Derinkuyu

July, 2004

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Mustafa Çelebi Pınar (Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Mustafa Akgül

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Oya-Ekin Karaşan

Approved for the Institute of Engineering and Science:

Prof. Dr. Mehmet B. Baray
Director of the Institute

ABSTRACT

ON THE S-PROCEDURE AND SOME VARIANTS

Kürşad Derinkuyu
M.S. in Industrial Engineering
Supervisor: Prof. Dr. Mustafa Çelebi Pınar
July, 2004

In this thesis, we deal with the S-procedure that corresponds to verifying that the minimum of a quadratic function over constraints consisting of quadratic functions is positive. 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. We improve an error bound in the Approximate S-Lemma used in establishing levels of conservatism results for approximate robust counterparts. Moreover we extend the S-procedure and obtain some general results in this field. Finally, we get a bound similar to Nesterov's bound for trust region subproblem, which consists in minimizing an indefinite quadratic function subject to a norm-1 constraint by using the Approximate S-Lemma.

Keywords: S-procedure, Approximate S-Lemma, Extended S-procedure, robust optimization, (conic) quadratic programming.

ÖZET

S-PROSEDÜR VE BAZI ÇEŞİTLERİ HAKKINDA

Kürşad Derinkuyu
Endüstri Mühendisliği, Yüksek Lisans
Tez Yöneticisi: Prof. Dr. Mustafa Çelebi Pınar
Temmuz, 2004

Bu tezde ikinci dereceden fonksiyon kısıtları olan ikinci dereceden fonksiyonun pozitif olduğunu tetkik eden S-prosedür ile ilgilendik. S-prosedür kontrol teori ve sağlam optimizasyon analizinde etkili bir araçtır. Ayrıca doğrusal matris eşitsizliklerinin (ya da kısmi belirli [semi-definite] programlamalarının) yeniden formüle edilmesi ve ikinci dereceden programlama analizinde kullanılmaktadır. Yaklaşık sağlam tamamlayıcılar için tutuculuk sonuçları derecesinin tesis edilmesinde kullanılan Yaklaşık S-Önermesinde hata sınırını geliştirdik. Bundan başka S-prosedürü genişlettik ve bu alanda genel sonuçlar elde ettik. Son olarak, Yaklaşık S-Önermesi kullanarak norm-1 kısıtı olan ikinci dereceden fonksiyonun en aza indirgenmesine dayanan güvenilir bölge [trust region] alt problemleri için Nesterov'un sonucuna benzer sonuç elde ettik.

Anahtar sözcükler: S-prosedür, Yaklaşık S-Önerme, Genişletilmiş S-prosedür, sağlam optimizasyon, (konik) ikinci dereceden programlama .

To my family...

Acknowledgement

I would like to express my sincere gratitude to my supervisor Prof. Dr. Mustafa Çelebi Pınar for his invaluable guidance, instructive comments and everlasting trust during my graduate study. He has been supervising me with patience and his great helps bring this thesis to an end.

I am also indebted to Assoc. Prof. Dr. Mustafa Akgül and Assist. Prof. Dr. Oya-Ekin Karaşan for showing keen interest to the subject matter and accepting to read and review the thesis.

I am grateful to Assoc. Prof. Dr. Azer Kerimov and Prof. Dr. Gerhard Wilhelm Weber for their recommendations and guidance.

I would like to thank to Selçuk Gören, Hakan Gültekin and Emrah Zarifoğlu for their friendship, encouragement and academic support. I would like to extend my thanks to Ayşegül Altın, Oğuz Atan, Zümbül Bulut, Esra Büyüktahtakın and Mustafa Rasim Kılınç for their continuous morale support and friendship.

Finally, I would like to express my deepest gratitude to my family for their endless love and understanding.

Contents

Chapter 1

Introduction

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 [28] without theory. Theoretical foundation of S-procedure was made in 1971 by Yakubovich and his students [40].

S-procedure takes its importance from the quadratic and convexity world by linking them to one another. One can suppose that these two old and well known fields of mathematics are not related with each other, despite their surprising proximity. During many years, both convex sets and quadratic maps were in the interest of active research by being in the center of many problems of interest. Hence finding a relationship between these two fields is important for future research. At this point, S-procedure comes to help us to fulfill this need.

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 one quadratic function, we refer to it as S-Lemma and if there are at least two quadratic inequalities in the constraints, we refer to it as S-procedure. Yakubovich [40] proves the S-Lemma and gives a definition of S-procedure. Polyak [32] gives a result related to S-procedure for problems with two constraints.

Although the S-Lemma was proved in 1971, people have begun to find results about the convexity problems of quadratic functions since 1918. From Toeplitz-Hausdorff [37, 20] theorem to today's complicated theorems, many important results are available. In this period, not only the S-Lemma was improved, but also two new areas were introduced, called approximate S-Lemma and extended S-procedure.

Approximate S-Lemma developed by Ben-Tal *et.al.* [8] establishes a bound for problems with more than one constraints of quadratic type. Their result also implies the S-Lemma of Yakubovich. Extended S-procedure is a new term coined by this thesis and implies both the theorems of Yakubovich and Polyak. This procedure is a corollary of Au-Yeung and Poon [2], and Barvinok's [3] theorems.

In this thesis, firstly we give two results about bounds of approximate S-Lemma. Then the thesis is interested in the construction of extended S-procedure. The thesis also deals with an example of trust region subproblem, which consists in minimizing a quadratic function subject to an L_1 norm constraint as an example of application of approximate S-Lemma. In this thesis, Corollaries 22, 24 and 25 are new for the extended S-procedure. Moreover, Lemmas 26 and 28 are new for the approximate S-Lemma.

The remainder of this study is organized as follows: Chapter 2 is devoted to provide a background on the S-procedure with an extensive review of literature. In Chapter 3, our results for approximate S-Lemma and extended S-procedure are given. In Chapter 4, evaluation of the results and an instance of a problem as an example of approximate S-Lemma is considered, whereas the last chapter is devoted to concluding remarks and future research directions.

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, p)-matrices. If $A \in S_n^{\mathbf{R}}$ and $X \in M_{n,p}(\mathbf{R})$, then $\langle\langle AX, X\rangle\rangle = \langle\langle A, XX^T\rangle\rangle :=$ trace of $A^T(XX^T)$.

Chapter 2

Background

S-procedure is one of the fundamental tools of optimal control and robust optimization. It is related with 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. Because of that, we should talk about its history to realize its importance before discussing what the S-procedure is.

In 1918, O. Toeplitz [37] introduced the idea of the numerical range 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 C^n (It is also called the numerical range of A). He also conjectured that the numerical range itself is convex, and one year later, F. Hausdorff [20] proved it. The Toeplitz-Hausdorff theorem is a main theorem for its extensions in the numerical range and it is applied in many fields of mathematics. This theorem can be formulated as: let

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

Then, the set W is convex, which is the first assertion on convexity of quadratic maps.

For the real field, the first result was obtained by Dines [13] in 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

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

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

The next result was obtained by Brickman [11]. He proved that the image of the unit sphere for the $n \geq 3$ (for any real symmetric matrices A and B),

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

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

These three papers are the main papers of 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 which is bounded by quadratic inequalities. The first work on this area is Finsler's Theorem [18](known as Débreu's lemma). Calabi [12] also proved this result independently in studying differential geometry and matrix differential equations by giving new and short topological proof. (A unilateral version of this theorem is proved by Yuan [41], 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 [40] saw the relation between the convex world and quadratic maps, and proved the S-Lemma. Then, this lemma became popular in the control area. There exist several methods to prove it but we want to give here

a proof that uses Dines' theorem to understand the link between convexity and the S-Lemma. (One can consult Nemirovski's [30] book (pp. 132–135) or Sturm and Zhang's [35] paper to see a different proofs).

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

$$x^T Ax \geq 0$$

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

$$x^T Bx \geq 0$$

is a consequence of it,i.e.,

$$x^T Ax \geq 0 \Rightarrow x^T Bx \geq 0$$

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

$$B \succeq \lambda A.$$

Proof: From Dines' theorem:

$$S_1 = \{s_1 := (x^T Ax \geq 0, x^T Bx \geq 0) : x \in \mathbf{R}^n\}$$

and

$$S_2 = \{s_2 := (x^T Ax \geq 0, x^T Bx < 0) : x \in \mathbf{R}^n\}$$

are convex. Since their intersection is empty, a separating hyperplane exists. In other words, there exists nonzero $c = (c_1, c_2) \in \mathbf{R}^2$, such that $(c, s_1) \geq 0$, $\forall s_1 \in S_1$ and $(c, s_2) \leq 0$, $\forall s_2 \in S_2$. From second inequality, $c_1 \leq 0, c_2 \geq 0$. From first inequality, for $\forall x \in \mathbf{R}^n$,

$$c_1 x^T Ax + c_2 x^T Bx \geq 0.$$

We know that there exists $\overline{x}$ such that $\overline{x}^T A\overline{x} > 0$ and $c_1 \leq 0$, so c_2 cannot be equal to zero. Finally, dividing inequalities by c_2, and defining $\lambda = -\frac{c_1}{c_2}$, we obtain:

$$B \succeq \lambda A.$$

The converse is trivial:

$$x^T Bx \geq \lambda(x^T Ax) \geq 0.$$

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 the subject 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.*[8] to obtain approximate version of the general result.

2.1 Review of Research on the S-procedure

The first attack to generalize these theorems was made by Hestenes and McShane [21] in 1940. They generalized the theorem of Finsler (1936).

Theorem 3 *The theorem of Hestenes and McShane(1940) [Extension of Finsler(1936)]*

Assume that $x^T Sx > 0$ *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* $x^T(\sum_i b_i T_i)x > 0$ *for all nonzero* $x \in L$. *Then, there exists* $c \in \mathbf{R}^{r+1}$ *that;*

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

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

There are several papers in this area by Au-Yeung [1], Dines [14, 15], John [24], Kühne [26], Taussky [36] and others. One of the benefits of Finsler, and Hestenes and McShane's theorems is understanding how we obtain an assumption of positive definiteness of a linear combination of matrices if we take it in the S-procedure. One can find these theorems until 1979 in a good survey written by

Uhlig [38]. To generalize the S-Lemma, researchers either extend the set of vectors to the set of matrices or make additional assumptions. First, we search in the first category and among these theorems, we deal with one of the most popular unpublished papers: the theorem of Bohnenblust [9] on the joint positiveness of matrices. Although this theorem can be written 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* $A_1, ..., A_m \in S_n^{\mathbf{R}}$. *Suppose* $(0,,,0) \notin W_p(A_1, ..., A_m)$ *where*

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

Then there exist $\alpha_1, ..., \alpha_m \in \mathbf{R}$ *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 [2] showed the extension of Brickman's and Toeplitz-Hausdorff theorem in 1979, and Poon [33] gives the final version of this result in 1997. Here is the Au-Yeung and Poon's 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* $A_1, ..., A_m \in S_n^{\mathbf{R}}$. *Then,*

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

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 $\langle\langle AX, X\rangle\rangle = TrAXX^T = \sum_{i=1}^{p} x_i^T A x_i$ and x_i denotes the columns of X. A corollary of this theorem is given in the paper of Hiriart-Urruty and Torki [22] to show a different perspective of it in 2002:

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

$$\textit{(i)} \left\{ \begin{array}{l} \langle\langle A_1 X, X\rangle\rangle = 0 \\ \langle\langle A_2 X, X\rangle\rangle = 0 \\ . \\ . \\ . \\ \langle\langle A_m X, X\rangle\rangle = 0 \end{array} \right\} \Rightarrow (X = 0).$$

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

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

In 1995, Barvinok [3] gives another theorem that extends the Dines's and Toeplitz-Hausdorff theorem while working on distance geometry.

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

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

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 Chapter 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 [40] and his students in 1971. Before talking about related papers on S-procedure, let us define the S-procedure and extended S-procedure in our notation:

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

Define

$$q_i(x) = \sum_{j=1}^{p} x_j^T Q_i x_j + 2b_i^T \sum_{j=1}^{p} x_j + c_i,\ Q_i \in S_n,\ i = 0, ..., m,\ j = 1, ..., p,\ x = (x_1, ..., x_p)$$

$$F := \{x_j \in \mathbf{R}^n : q_i(x) \geq 0,\ i = 1, ..., m,\ j = 1, ..., p\},$$

$q_i(x_j)$ 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) $q_0(x) \geq 0\ \forall x \in F$

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

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 reverse is in generally false. If $S_1 \Leftrightarrow S_2$, the S-procedure is called lossless and this condition appears only in some special cases. If we define the S-procedure as a method of verifying S_1 using S_2, computation of S_2 is much easier than computation of S_1. For this reason the S-procedure is important and popular.

The first paper we review in this field is the paper of Megretsky and Treil [29] in 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, $\sigma_j : L \to \mathbf{R} (j = 0, 1, ..., m)$ be continuous time-invariant quadratic forms. Suppose that there exists $f_* \in L$ such that $\sigma_1(f_*) > 0, ..., \sigma_m(f_*) > 0$.
Then the following statements are equivalent:

(i) $\sigma_0(f) \leq 0$ for all $f \in L$ such that $\sigma_1(f) > 0, ..., \sigma_m(f) > 0$;

(ii) There exists $\tau_j \geq 0$ such that

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

for all $f \in L$.

Although this theorem gives us the S-procedure, time-invariant quadratic forms are very specific for this area. Moreover, one can find another convexity result for commutative matrices in the paper of Fradkov, 1973 [19] (Detailed information about commutative matrices can be obtained from Matrix Analysis book of Horn and Johnson [23]).

Theorem 10 *Theorem of Fradkov,1973*
If matrices $A_1, ..., A_m$ commute and m quadratic forms $f_i(x) = \langle A_i x, x\rangle, x \in \mathbf{R}^n, i = 1, ..., m$ are given. Then

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

is a closed convex cone for all m,n.

Despite of Megretsky and Treil, and Fradkov's papers, they are not the only angles to deal with the S-procedure. Extension for matrix variables is another viewpoint of this problem given by Luo *et.al.* [27] which uses quadratic matrix inequalities 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*

$$\begin{bmatrix} H & F+GX \\ (F+GX)^T & C+X^TB+B^TX+X^TAX \end{bmatrix} \succeq 0 \quad \textit{for all} \quad I - X^TDX \succeq 0$$

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

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

Neither Megretsky and Treil, and Fradkov's results nor extension of Luo *et.al.* satisfy the S-procedure in general. Therefore the S-procedure is still an open problem for us.

In 1998, Polyak [32] succeeded in proving the quadratic form case 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,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:*

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

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* $\alpha_i, i = 0, 1, 2$ *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 $\tau_1 \geq 0, \tau_2 \geq 0$*:*

$$A_0 \preceq \tau_1 A_1 + \tau_2 A_2$$

$$\alpha_0 \geq \tau_1 \alpha_1 + \tau_2 \alpha_2.$$

Polyak's theorem is good but not enough for our complicated world. Because of that, researchers begin to work on quadratic equations in a different way to get lower and upper bounds for optimal values of quadratic functions with quadratic constraints. Recently a new lemma was proved by Ben-Tal, Nemirovski and Roos [8] called Approximate S-Lemma.

2.2 Review of Research on the Approximate S-Lemma

In this section of this chapter, 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. With this method, the reader may appreciate the importance of approximate S-Lemma.

S-Lemma has been widely used within the robust optimization paradigm of Ben-Tal and Nemirovski and co-authors [6, 5, 7] and El-Ghaoui and co-authors [17, 10] 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 *et.al.* [8]. Before beginning to talk about the subject, we need additional notations and definitions about robust methodology and conic quadratic problems.(For conic programming, Ben-Tal's [4] book is a good 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*

$$\min_{x \in \mathbf{R}^p} \{c^T x : Ax - b \in K\} \tag{2.1}$$

is a conic problem (CP). When the data of the constraint (A, b) is coming from uncertain set U, the problem

$$\{\min_{x \in \mathbf{R}^p} \{c^T x : Ax - b \in K\} : (A, b) \in U\} \tag{2.2}$$

is called uncertain conic problem (UCP) and the problem

$$\min_{x \in \mathbf{R}^p} \{c^T x : Ax - b \in K : \forall (A, b) \in U\} \tag{2.3}$$

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 with 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 put 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 plane 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 find a subset for $\hat{\mathcal{R}}$, the set $\mathcal{X}$ should shrink. 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 $\hat{\mathcal{R}}$. Let us give the formal definition of these words:

Definition 16 *The smallest ρ to obtain:*

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

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 into the uncertain quadratic constraint (It can also be written as a conic quadratic form):

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

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

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

for each $Q_k \succeq 0$ 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 [6] in 1998.(Theorem 3.2 in that paper).(It is also discussed in the paper of El Ghaoui and Lebret [16]). For the case $K = 1$, Q_1 is identity matrix:

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

$$x^T A^T A x \leq 2b^T x + c \quad \forall (A, b, c) \in U_{simple}, \tag{2.4}$$

where

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

if and only if for some $\lambda \in \mathbf{R}$ the pair (x, λ) is a solution of the following linear matrix inequality (LMI):

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

Its proof can be seen in the appendix part of this thesis. Clearly, the proof completely depends on the S-Lemma. However the S-Lemma works only for single quadratic form. Therefore we need a somehow different theorem that also works for the cases $K > 1$. Of course, it does not give exact results as above, but it gives reasonable bounds for us to work on more complicated problems. Now it is time to obtain this lemma and see how it works.

Ben-Tal *et al.* proved the following result; see [8] Lemma A.6, pp.554–559. (Ben-Tal *et al.* also showed that the Approximate S-Lemma implies the usual S-Lemma).

Lemma 18 *(Approximate S-Lemma). Let $R, R_0, R_1, ..., R_k$ be symmetric $n \times n$ matrices such that*

$$R_1, ..., R_k \succeq 0, \tag{2.5}$$

and assume that

$$\exists \lambda_0, \lambda_1, ..., \lambda_k \geq 0 \; s.t. \; \sum_{k=0}^{K} \lambda_k R_k \succ 0. \tag{2.6}$$

Consider the following quadratically constrained quadratic program,

$$QCQ = \max_{y \varepsilon R^n} \{ \; y^T R y \;\; : \;\; y^T R_0 y \leq r_0, y^T R_k y \leq 1, k = 1, ..., K \; \} \tag{2.7}$$

and the semidefinite optimization problem

$$SDP = \min_{\mu_0, \mu_1, ..., \mu_K} \{ \; r_0 \mu_0 + \textstyle\sum_{k=1}^{K} \mu_k \;\; : \;\; \sum_{k=0}^{K} \mu_k R_k \succeq R, \mu \geq 0 \; \}. \tag{2.8}$$

Then

(i) If problem (2.7) is feasible, then problem (2.8) is bounded below and

$$SDP \geq QCQ. \tag{2.9}$$

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

$$y_*^T R y_* = SDP, \tag{2.10}$$

$$y_*^T R_0 y_* \leq r_0, \tag{2.11}$$

$$y_*^T R_k y_* \leq \tilde{\rho}^2,\ k = 1, ..., K, \tag{2.12}$$

where

$$\tilde{\rho} := (\ 2log(\ 6\sum_{k=1}^{K} rank\, R_k\ )\ )^{\frac{1}{2}}, \tag{2.13}$$

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

$$\tilde{\rho} := (\ 2log(\ 16n^2\sum_{k=1}^{K} rank\, R_k\ )\ )^{\frac{1}{2}} \tag{2.14}$$

otherwise.

(ii) If

$$r_0 > 0, \tag{2.15}$$

then (2.7) is feasible, problem (2.8) is solvable, and

$$0 \leq QCQ \leq SDP \leq \tilde{\rho}^2 QCQ. \tag{2.16}$$

The proof of this lemma (due to Ben-Tal, Nemirovski and Roos) is given in the appendix. After giving the theorem, now we are ready to work on more complicated uncertainty sets which are the cases $K > 1$, from the paper of Ben-Tal *et al.* [8]. Let us begin by defining the robust feasible set of them:

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

where

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

Note that the robust counterpart of uncertain quadratic constraint with the $\cap$-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.*, pp. 539 [8]).

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

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

and

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

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

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

If y satisfies the above, $-y$ also does. Therefore it is the same as:

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

If we take the $t^2 \leq 1$, inequality can be written as;

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

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

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

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

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

From the Schur complement (see appendix (30)), we will obtain the following theorem:

Theorem 19 *The set $\mathcal{R}_\rho$ of (x, λ) satisfying $\lambda \geq 0$ and*

$$\begin{bmatrix} c[x]\text{-}\sum_{k=1}^{K} \lambda_k & (\text{-}b_\rho[x] - d_\rho)^T & a[x]^T \\ (\text{-}b_\rho[x] - d_\rho) & \sum_{k=1}^{K} \lambda_k Q_k & \text{-}A_\rho[x]^T \\ a[x] & \text{-}A_\rho[x] & I_M \end{bmatrix} \succeq 0 \tag{2.17}$$

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

Now we get the approximate robust counterpart but 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 20 *The level of conservativeness of the approximate robust counterpart $\mathcal{R}$ (as given by 2.17) of the set $\mathcal{X}$ is at most*

$$\tilde{\rho} := (\ 2log(\ 6 \sum_{k=1}^{K} rank\ R_k\ )\ )^{\frac{1}{2}}, \tag{2.18}$$

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

$$\zeta_*^T Q_k \zeta_* \leq 1, \quad k = 1, ..., K \tag{2.19}$$

and

$$\tilde{\rho}^2 \zeta_*^T A_\rho[x]^T A_\rho[x] \zeta_* + 2\tilde{\rho} \zeta_*^T (A_\rho[x]^T a[x] - b_\rho[x] - d_\rho) > c[x] - a[x]^T a[x]. \tag{2.20}$$

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

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

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

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

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

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:

$$R \preceq \sum_{k=0}^{K} \lambda_k R_k, \tag{2.21}$$

$$\sum_{k=0}^{K} \lambda_k \leq c[x] - a[x]^T a[x]. \tag{2.22}$$

Note: Ben-Tal *et.al.* try to prove this case by claiming: assumption that x cannot be extended to a solution of (2.17) implies that x cannot be extended to a solution of uncertain quadratic constraint. However this claim is wrong because the uncertain quadratic constraint set is larger than the set (2.17). Therefore, x cannot be extended to a solution of (2.17), but may be extended to a solution of uncertain quadratic constraint. Hence we change this part of the proof and instead of it, we claim that these two inequalities cause x to be a solution of (2.17), which contradicts our assumption.

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

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

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

From assumption x cannot be extended to a solution of (2.17). On the other hand, we have

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

or

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

or, equivalently

$$\lambda_0 t^2 + \sum_{k=1}^{K} \lambda_k y^T Q_k y - 2ty^T(A_p[x]^T a[x] - b_p[x] - d_p) - y^T A_p[x]^T A_p[x] y \geq 0 \quad (2.23)$$

We know that

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

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

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

From (2.23) and taking $-t$ instead of t, we obtain

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

or,

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

However x is extended to a solution of (2.17), so it contradicts with our assumption. Case I cannot occur.

Case II. There is no $\lambda_0, ..., \lambda_K \geq 0$ that satisfies both (2.21) and (2.22). Hence from approximate S-Lemma:

$$SDP > c[x] - a[x]^T a[x]. \quad (2.24)$$

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

$$y_*^T R_0 y_* = t_*^2 \leq r_0 = 1,$$

$$y_*^T R_k y_* = \eta_*^T Q_k \eta_* \leq \tilde{\rho}^2, \quad k = 1, ..., K,$$

$$y_*^T R y_* = \eta_*^T A_\rho[x]^T A_\rho[x] \eta_* + 2t_* \eta_*^T (A_\rho[x]^T a[x] - b_\rho[x] - d_\rho) = SDP$$

$$> c[x] - a[x]^T a[x],$$

from (2.24). Setting $\overline{\eta} = \tilde{\rho}^{-1} \eta_*$, these inequalities turn into

$$|t_*| \leq 1,$$

$$\overline{\eta}^T Q_k \overline{\eta} \leq 1, \quad k = 1, ..., K,$$

$$\tilde{\rho}^2 \overline{\eta}^T A_\rho[x]^T A_\rho[x] \overline{\eta} + 2\tilde{\rho} \overline{\eta}^T t_* (A_\rho[x]^T a[x] - b_\rho[x] - d_\rho) > c[x] - a[x]^T a[x].$$

If $(t_*, \overline{\eta})$ is a solution of this system, then $\zeta_* = \overline{\eta}$ or $\zeta_* = -\overline{\eta}$ is a solution of (2.19)-(2.20). This completes the proof.

This concludes this chapter. Although the background on S-Lemma, S-procedure and approximate S-Lemma is vast, we tried to give the main theorems in our view here and explain them by giving some examples. In the next chapter, we give some results that strongly rely on these theorems.

Chapter 3

Results

In this chapter, we give some results about the extended S-procedure and approximate S-Lemma. Although we find two results by giving additional assumptions and extending S-procedure in the first section, we improve or attempt to improve the lemmas that are necessary for approximate S-Lemma in the other section.

3.1 Some Results on Extended S-procedure

We defined extended S-procedure (8) in the previous chapter. Now we show some results by using Barvinok, and Au-Yeung and Poon's theorems.

3.1.1 Corollary for Barvinok's Theorem(1995)

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

Theorem 21 *Let $A_0, A_2, ..., A_{m-1} \in S_n^{\mathbf{R}}$, and let $p := \lfloor \frac{\sqrt{8m+1}-1}{2} \rfloor$. Then*

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

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

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

Corollary 22 *Let $A_0, A_2, ..., A_{m-1} \in S_n^{\mathbf{R}}$, and let $p := \lfloor \frac{\sqrt{8m+1}-1}{2} \rfloor$. Assume there exists $X^0 \in M_{n,p}(\mathbf{R})$, such that*

$$f_i(X^0) = (\langle\langle A_i X^0, X^0 \rangle\rangle > 0, \quad i = 1, ..., m-1. \tag{3.1}$$

Then

$$f_0(X) \geq 0 \quad \forall X : \quad f_i(X) \geq 0, \quad i = 1, ..., m-1. \tag{3.2}$$

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

$$f_0(X) \geq \sum_{i=1}^{m-1} \tau_i f_i(X). \tag{3.3}$$

Proof: From Barvinok's theorem (7),

$$S_1 = \{s_1 := (\langle\langle A_0 X, X \rangle\rangle \geq 0, \langle\langle A_1 X, X \rangle\rangle \geq 0, ..., \langle\langle A_{m-1} X, X \rangle\rangle \geq 0) : X \in M_{n,p}(\mathbf{R})\}$$

and

$$S_2 = \{s_2 := (\langle\langle A_0 X, X \rangle\rangle < 0, \langle\langle A_1 X, X \rangle\rangle \geq 0, ..., \langle\langle A_{m-1} X, X \rangle\rangle \geq 0) : X \in M_{n,p}(\mathbf{R})\}$$

are convex. Since their intersection is empty, a separating hyperplane exists. In other words, there exists nonzero $c = (c_0, c_1, ..., c_{m-1}) \in \mathbf{R}^m$, such that $(c, s_1) \geq 0$, $\forall s_1 \in S_1$ and $(c, s_2) \leq 0$, $\forall s_2 \in S_2$. From second inequality, $c_0 \geq 0, c_1 \leq 0, ..., c_{m-1} \leq 0$. (If S_2 is empty, choose any point $d = (d_0, d_1, ..., d_{m-1}) \in \mathbf{R}^m$, where $d_0 < 0$, $d_i \geq 0$ for $i = 1, ..., m-1$ and obtain the same result for c). From first inequality, for $\forall X \in M_{n,p}(\mathbf{R})$,

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

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 $c_i \leq 0$ for $i = 1, ..., m-1$, so c_0 cannot be equal zero. Finally, dividing inequalities by c_0, and 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 converse is trivial:

$$f_0(X) \geq \sum_{i=1}^{m-1} \tau_i f_i(X) \geq 0.$$

The proof is complete. Clearly, there exists a strong relationship between the S-procedure and convexity. The link between these two fields is provided by the separation lemma.

3.1.2 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. Turning it to an extended S-procedure is not very hard. With same definition of $f(X)$ as in the first corollary, we can write this theorem as follows.

Theorem 23 *Let the integer p be defined as*

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

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

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

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

Firstly, we will show the following corollary by using the procedure of Polyak's proof in the paper [32]:

Corollary 24 *Let $A_0, A_1, ..., A_m \in S_n^{\mathbf{R}}$, 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 $i = 0, 1, ..., m$. If there exists $\mu \in \mathbf{R}^{m+1}$ such that;*

$$\sum_{i=0}^{m} \mu_i f_i(X) > 0, \quad i = 0, ..., m, \tag{3.4}$$

then the set

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

is convex.

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

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. Because when you fix $g_0 = f_0, g_1 = f_1, ...g_{m-1} = f_{m-1}$, the variable of g_m only depends on f_m.

Also we can make a nonsingular linear transformation and assume $g_m = \|X\|^2$. Because we can write $\|X\|^2 = \sum_{i=1}^{p} \|x_i\|^2$ 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 nonsingular linear transformation. (It has still the characteristic of being injective, $\|X\|^2 = 0 \Leftrightarrow X = 0$ and surjective, its range equals $\mathbf{R}_+ \cup \{\mathbf{0}\}$). From the Theorem of Au-Yeung and Poon;

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

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

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

Hence G is convex therefore F is convex. Therefore we can write following corollary:

Corollary 25 *Let $A_0, A_1, ..., A_m \in S_n^{\mathbf{R}}$, and let*

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

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

$$f_i(X^0) = (\langle\langle A_i X^0, X^0 \rangle\rangle > 0, \quad i = 1, ..., m. \tag{3.5}$$

and that

$$\sum_{i=1}^{m} \mu_i f_i(X) > 0. \tag{3.6}$$

Then

$$f_0(X) \geq 0 \quad \forall X : \quad f_i(X) \geq 0, \quad i = 1, ..., m. \tag{3.7}$$

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

$$f_0(X) \geq \sum_{i=1}^{m} \tau_i f_i(X). \tag{3.8}$$

Proof: The proof is the same as in corollary of Barvinok's theorem given above.

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

In the next section, we deal with some results about the approximate S-Lemma.

3.2 Some Results on Approximate S-Lemma

In this section, we attempt to improve bounds of the Approximate S-Lemma given by Ben-Tal *et.al.* both for dyadic case and general case. For dyadic case, our idea failed but we get some related result for relaxed case. Despite of it, we were able to improve the bound for general case. Firstly, we talk about the dyadic case.

3.2.1 Partial Result for Dyadic Case

In the paper of Ben-Tal *et.al.*, they give a conjecture to improve dyadic case which is

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

$$Pr(\xi_i = 1) = Pr(\xi_i = -1) = 1/2 \tag{3.9}$$

then one has,

$$Pr(|\xi^T x| \leq 1) \geq 1/2 \tag{3.10}$$

This conjecture improve the bound to $\frac{1}{2}$ from $\frac{1}{3}$. We work on this conjecture by using n-dimensional geometry. However we only proved the following relaxed version of it:

Lemma 26 *Let* $x = \{x_1, , , , x_n\}$ *and* $\xi = \{\xi_1, , , , \xi_n\} \in \mathbf{R}^n$. *If* $\|x\|_2 = 1$ *and* $\|\xi\|_2^2 = n$ *then one has*

$$Pr(|\xi^T x| \leq 1) \geq 1/2. \tag{3.11}$$

Proof: This lemma is a relaxed version of conjecture, because the vectors ξ are equally distributed on the surface of hyper-sphere of $\|\xi\|_2^2 = n$. The conjecture wants that for any x, at least half of the vectors satisfies the inequality. However we prove that for any x, half of the surface of the hyper-sphere satisfies the inequality. We also prove 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. Now let us begin the proof.

The condition of $\|x\|_2 = 1$ and $\|\xi\|_2^2 = n$ are surfaces of hyper-sphere with radius 1 and $\sqrt{n}$, respectively, in multidimensional geometry. Firstly, we prove that for any ξ, more than half of the x vectors satisfy (3.11). In fact for any ξ, we get two such parallel hyperplanes passing through hyper-sphere $\|x\|_2 = 1$, so taking $\xi = (1, 1, ..., 1)$ does not hinder our general situation.

By taking $\xi = (1, 1, ..., 1)$, we get $(|\sum x_i| \leq 1)$ the volume between two hyperplanes. These two conditions define the surface of hyper-sphere between $\sum x_i = 1$ and $\sum x_i = -1$ hyperplanes. Therefore we should search the ratio on the surface of hyper-sphere which corresponds to our conditions. From symmetry, we can take the upper hyper-hemisphere which satisfies $\sum_{i=1}^{n} x_i \geq 0$ and get the ratio of surface under hyperplane $\sum_{i=1}^{n} x_i = 1$ to the surface of hyper-hemisphere.

The proof shows it directly for dimension one, two and three, then calculates it for the dimensions larger than 3.

N=1

$$\|x\|_2 = 1 \Rightarrow x = \mp 1$$
$$Pr(|\xi^T x| \leq 1) = 1 \geq 1/2.$$

Note: If x is bigger than 1, corresponding probability is zero. Hence, we cannot take such x for any dimension n bigger than 1. (Because the same result occurs for any dimension n when you give $x_i = 1$ and any $j \neq i$, $x_j = 0$ in other dimensions)

N=2 For $x = (x_1, x_2)$,

$$x_1^2 + x_2^2 = 1$$

.

It is a circle with radius 1. For semi-circle, the curve between

$$0 \leq x_1 + x_2 \leq 1$$

has 90 degrees which is half of semi-circle. (The other side of circle also gives same result).

N=3 For $x = (x_1, x_2, x_3)$,

$$x_1^2 + x_2^2 + x_3^2 = 1 \text{ and } x_1 + x_2 + x_3 \geq 0$$

is a hemisphere. We know that the distance between the origin and the plane that cuts the surface of hemisphere to two equal parts is $r/2$, which is $1/2$ for

our case. Also, the distance between origin and $x_1 + x_2 + x_3 = 1$ is $1/\sqrt{3}$ which is bigger than $1/2$, so the surface between $x_1 + x_2 + x_3 = 1$ and $x_1 + x_2 + x_3 = 0$ is more than half of total surface of hemisphere.

N $\geq$ 4

For dimensions larger than 3, we should introduce the surface contend geometric center of hyper-hemisphere (SCGCoH) which is, (pp. 137, Sommerville [34]) (distance from origin, $r = 1$)

$$\frac{\Gamma(n/2)}{\sqrt{\pi}\Gamma(n+1/2)}r. \tag{3.12}$$

Geometric center depends on both area and distance. One may imagine it as the center of gravity of surface of hyper-hemisphere. Now, we should prove following two statements which are enough to complete our proof for the fixed ξ.

- For the hyperplane which is parallel to ground of hyper-hemisphere, $\sum_{i=1}^{n} x_i \geq 0$, and passes through $SCGCoH$, the surface of sphere under this hyperplane is larger than or equal to the other part.
- The distance between origin and the SCGCoH is smaller than the distance between origin and the hyperplane $\sum_{i=1}^{n} x_i = 1$.

These two statements are enough to complete proof because first of them indicates the relationship between $SCGCoH$ and half of the surface of hyper-hemisphere, and second one shows the relationship between $SCGCoH$ and our cutting hyperplane.

Firstly, let us look at the first statement. For more than three dimensions, we know that (3.12) is less than $1/2$. In other words, the lower part is more concentrated than the upper part. Also we know that if we divide hyper-hemisphere into small equal length parts along the direction passing from the origin and perpendicular to the hyperplane, the parts nearer to the origin has larger area than the parts far away from it.

If we take a line passing from origin and the SCGCoH, the hyper-hemisphere encircles this line. Therefore we can suppose that the points of surface of hyper-hemisphere are on this line. By projecting these points to the line, we can calculate the SCGCOH.

Because the distance between origin and the SCGCoH is less than 1/2, let us subtract the points on the surface in the below (closer to origin) and above of hyperplane which has equal distance to hyperplane. At the end of this procedure, we already have points in the below part but there is no point between the SCGCoH and 2SCGCoH. Therefore, from definition of geometric center;

(Area Remaining Below)(Dist. Bet. SCGCoH and SCGCoH of Remaining Below Part)

=

(Area Remaining Above)(Dist. Bet. SCGCoH and SCGCoH of Remaining Above Part).

Because there is no point between the SCGCOH and 2SCGCoH on the above part after the procedure, the distance between the SCGCoH and SCGCoH of remaining above part is bigger than the distance between origin and the SCGCoH. Therefore, distance between the SCGCoH and SCGCoH of remaining below part is smaller than the other. (Observe that SCGCoH of remaining below part is between origin and the SCGCoH). Hence

$$\text{(Area Remaining Below)} > \text{(Area Remaining Above)},$$

$$\text{(Area Below)} > \text{(Area Above)}.$$

The proof of the first statement is complete. Secondly, let us show that the distance between origin and the hyperplane $x_1 + ... + x_n = 1$ which is $1/\sqrt{n}$, is bigger than the distance between origin and the SCGCoH. Before explaining the proof, let us introduce the notation double prime:

Definition 27 *Double Prime*

$$n!! = \begin{cases} n(n\text{-}2)...5.3.1 & if \;\; n > 0, odd \\ n(n\text{-}2)...6.4.2 & if \;\; n > 0, even \\ 1 & -1, 0 \end{cases}$$

The proof will be done by using induction. For n even, SCGCoH is $(r = 1)$

$$= \frac{\Gamma(n/2)}{\sqrt{\pi}\Gamma(\frac{n+1}{2})} = \frac{\frac{n-2}{2}!}{\sqrt{\pi}\frac{(n-1)!!\sqrt{\pi}}{2^{n/2}}} = \frac{(\frac{n}{2}-1)!2^{n/2}}{(n-1)!!\sqrt{\pi}}.$$

For $n = 4$,

$$\frac{(\frac{n}{2}-1)!2^{n/2}}{(n-1)!!\pi} = \frac{4}{3\pi} \leq \frac{1}{\sqrt{4}}.$$

Assume, it is true for $n = k$;

$$\frac{(\frac{k}{2}-1)!2^{k/2}}{(k-1)!!\pi} = \frac{(k-2)!!2}{(k-1)!!\pi} \leq \frac{1}{\sqrt{k}}.$$

Let us look at the case for $n = k + 2$;

$$\frac{(\frac{k}{2})!2^{\frac{k+2}{2}}}{(k+1)!!\pi} = \frac{(k)!!2}{(k+1)!!\pi} \overset{?}{\leq} \frac{1}{\sqrt{k+2}}$$

$$\frac{(k)!!2}{(k+1)!!\pi} = B\frac{k}{k+1} \overset{?}{\leq} \frac{1}{\sqrt{k+2}}$$

$$B\frac{k}{k+1} \leq \frac{k}{\sqrt{k}(k+1)} \overset{?}{\leq} \frac{1}{\sqrt{k+2}}$$

$$\Rightarrow \quad \sqrt{k(k+2)} \overset{?}{\leq} k+1$$

$$\Rightarrow \quad k^2 + 2k \overset{?}{\leq} k^2 + 2k + 1$$

$$\Rightarrow \quad 0 \leq 1$$

so it is true for even case.

For n odd, SCGCoH is $(r = 1)$

$$= \frac{\Gamma(n/2)}{\sqrt{\pi}\Gamma(\frac{n+1}{2})} = \frac{\frac{(n-2)!!\sqrt{\pi}}{2^{\frac{n-1}{2}}}}{\sqrt{\pi}\frac{n-1}{2}!} = \frac{(n-2)!!}{\frac{n-1}{2}!2^{\frac{n-1}{2}}}.$$

For $n = 5$ we have

$$\frac{(n-2)!!}{\frac{n-1}{2}!2^{\frac{n-1}{2}}} = \frac{3}{8} \leq \frac{1}{\sqrt{5}}.$$

Assume, it is true for $n = k$;

$$\frac{(k-2)!!}{\frac{k-1}{2}!2^{\frac{k-1}{2}}} = \frac{(k-2)!!}{(k-1)!!} \leq \frac{1}{\sqrt{k}}.$$

Let look at the case for $n = k + 2$;

$$\begin{aligned}
\frac{(k)!!}{\frac{k+1}{2}!2^{\frac{k+1}{2}}} = \frac{(k)!!}{(k+1)!!} &\overset{?}{\leq} \frac{1}{\sqrt{k+2}} \\
= \frac{(k)!!}{(k+1)!!} = C\frac{k}{k+1} &\overset{?}{\leq} \frac{1}{\sqrt{k+2}} \\
= C\frac{k}{k+1} \leq \frac{k}{\sqrt{k}(k+1)} &\overset{?}{\leq} \frac{1}{\sqrt{k+2}} \\
\Rightarrow \quad \sqrt{k(k+2)} &\overset{?}{\leq} k+1 \\
\Rightarrow \quad k^2 + 2k &\overset{?}{\leq} k^2 + 2k + 1 \\
\Rightarrow \quad 0 &\leq 1.
\end{aligned}$$

It is true for odd case. The proof of the second statement is also complete.

Let us look at the other side which is for fixed x, whether more than half of the ξ vectors is in the system or not, which is $(|\xi^T x| \leq 1)$.

By taking $x = (1, 0, ..., 0)$, we get $(|\xi_1| \leq 1)$ the volume between two hyperplanes with distance 1 to the origin. In fact for any ξ, we get two such parallel hyperplanes. Because $\|\xi\|_2 = \sqrt{n}$ is a hyper-sphere, taking $x = (1, 0, ..., 0)$ does not create loss of generality.

For the previous case, the distance between origin and the hyperplane $x_1 + ... + x_n = 1$ is $1/\sqrt{n}$ and radius is 1. Now the distance between origin and the hyperplane $(|\xi_1| = 1)$ is 1 and radius is $\sqrt{n}$. We know that SCGCoH depends on radius linearly, so the previous proof is sufficient for these cases.

We proved that both fixed x, more than half of the ξ vectors satisfies the system and fixed ξ, more than half of the x vectors satisfies the system which is $(|\xi^T x| \leq 1)$.

We proved all the cases of our claim. This concludes the proof of the lemma.

3.2.2 Improvement Lemma for General Case

The following result is crucial in establishing the bound

$$\rho := (\ 2log(\ 4n \sum_{k=1}^{K} \text{rank } R_k\ )\ )^{\frac{1}{2}} \tag{3.13}$$

in an improved version of Lemma 34. The original result of Ben-Tal *et al.* in [8] has the right-hand side of Lemma 34 equal to $\frac{1}{8n^2}$. They conjectured that its right-hand side can be $\frac{1}{4}$. Unlike their proof which uses moments, our proof relies on recursive subdivision of a matrix into 4 submatrices.

Lemma 28 *Let B denote a symmetric $n \times n$ matrix and $\xi = \{\xi_1, , , , \xi_n\} \in \mathbf{R}^n$. The coordinates ξ_i of ξ are independently identically distributed random variables with*

$$Pr(\xi_i = 1) = Pr(\xi_i = -1) = 1/2 \tag{3.14}$$

then one has

$$Pr(\xi^T B \xi \leq TrB) \geq \frac{1}{2^{\lceil \log_2(n) \rceil}} > \frac{1}{2n}. \tag{3.15}$$

Proof: Consider the random variable

$$\gamma := \sum_{i<j} \xi_i \xi_j B_{ij} = \frac{1}{2}(\xi^T B \xi - TrB).$$

Then (3.15) is equivalent to

$$\omega := Pr(\gamma \leq 0) > \frac{1}{2n}.$$

Before beginning to construct the proof for general n, let us look at the cases $n = 1, 2, 3, 4$.

For $n = 1$, we have

$$\omega := Pr(\gamma \leq 0) = 1.$$

For $n = 2$, we have

$$\omega := Pr(\frac{\xi^T B \xi - TrB}{2} \leq 0) = Pr(\xi_1 \xi_2 B_{12} \leq 0) = Pr(-\xi_1 \xi_2 B_{12} \leq 0) = \frac{1}{2}.$$

For $n = 3$ assume there exists a 3×3 symmetric matrix C satisfying

$$Pr(\xi^T C\xi \leq TrC) < \frac{1}{4},$$
$$Pr(\xi^T C\xi > TrC) > \frac{3}{4}.$$

The latter inequality says that

$$Pr(\xi_1\xi_2 C_{12} + \xi_1\xi_3 C_{13} + \xi_2\xi_3 C_{23} > 0) > \frac{3}{4}.$$

Now, let C^1, C^2, C^3 be (3×3) symmetric matrices such that

$$C^1 = \begin{Bmatrix} C_{11} & -C_{12} & -C_{13} \\ -C_{21} & C_{22} & C_{23} \\ -C_{31} & C_{32} & C_{33} \end{Bmatrix}, \quad C^2 = \begin{Bmatrix} C_{11} & -C_{12} & C_{13} \\ -C_{21} & C_{22} & -C_{23} \\ C_{31} & -C_{32} & C_{33} \end{Bmatrix},$$
$$C^3 = \begin{Bmatrix} C_{11} & C_{12} & -C_{13} \\ C_{21} & C_{22} & -C_{23} \\ -C_{31} & -C_{32} & C_{33} \end{Bmatrix}$$

then we have

$$Pr(\xi^T C\xi > TrC) = Pr(\xi^T C^1\xi > TrC^1) = Pr(\xi^T C^2\xi > TrC^2) = Pr(\xi^T C^3\xi > TrC^3).$$

For multiplication of any vector ξ with matrix C^i for $i = 1, 2, 3$, there exists multiplication of another vector ξ with matrix C which gives same result. To see why this is true, simply change the ξ_1, ξ_2, ξ_3 elements of vector ξ with negative ones for C^1, C^2, C^3 respectively, and obtain same result as with matrix C .

If

$$Pr(\xi^T C\xi > TrC) = Pr(\xi_1\xi_2 C_{12} + \xi_1\xi_3 C_{13} + \xi_2\xi_3 C_{23} > 0) > \frac{3}{4},$$

and

$$Pr(\xi_1\xi_2 C^1_{12} + \xi_1\xi_3 C^1_{13} + \xi_2\xi_3 C^1_{23} > 0) > \frac{3}{4}.$$

At least half of the ξ vectors satisfy both of the inequalities above so it holds that

$$Pr(\xi_2\xi_3 C_{23} > 0) > \frac{1}{2} \quad (\frac{3}{4} + \frac{3}{4} - 1). \tag{3.16}$$

If we have

$$Pr(\xi_1\xi_2C_{12}^2+\xi_1\xi_3C_{13}^2+\xi_2\xi_3C_{23}^2>0)>\frac{3}{4}$$
$$Pr(\xi_1\xi_2C_{12}^3+\xi_1\xi_3C_{13}^3+\xi_2\xi_3C_{23}^3>0)>\frac{3}{4}$$

then

$$Pr(-\xi_2\xi_3C_{23}>0)>\frac{1}{2}\quad(\frac{3}{4}+\frac{3}{4}-1). \tag{3.17}$$

Both (3.16) and (3.17) cannot be true simultaneously. Therefore, we obtained a contradiction and finished the proof for $n=3$.

For $n=4$ assume there exists a 4×4 symmetric matrix D matrix violating (B.33), i.e., one has

$$Pr(\xi^TD\xi\leq TrD)<\frac{1}{4}$$
$$Pr(\xi^TD\xi>TrD)>\frac{3}{4}.$$

Let D^1,D^2,D^3,D^4 be 4×4 symmetric matrices defined as follows:

$$D^1=\begin{Bmatrix}D_{11}&-D_{12}&-D_{13}&-D_{14}\\-D_{21}&D_{22}&D_{23}&D_{24}\\-D_{31}&D_{32}&D_{33}&D_{34}\\-D_{41}&D_{42}&D_{43}&D_{44}\end{Bmatrix},\quad D^2=\begin{Bmatrix}D_{11}&-D_{12}&D_{13}&D_{14}\\-D_{21}&D_{22}&-D_{23}&-D_{24}\\D_{31}&-D_{32}&D_{33}&D_{34}\\D_{41}&-D_{42}&D_{43}&D_{44}\end{Bmatrix},$$

$$D^3=\begin{Bmatrix}D_{11}&D_{12}&-D_{13}&D_{14}\\D_{21}&D_{22}&-D_{23}&D_{24}\\-D_{31}&-D_{32}&D_{33}&-D_{34}\\D_{41}&D_{42}&-D_{43}&D_{44}\end{Bmatrix},\quad D^4=\begin{Bmatrix}D_{11}&D_{12}&D_{13}&-D_{14}\\D_{21}&D_{22}&D_{23}&-D_{24}\\D_{31}&D_{32}&D_{33}&-D_{34}\\-D_{41}&-D_{42}&-D_{43}&D_{44}\end{Bmatrix}$$

then it is immediate to observe that

$$Pr(\xi^TD\xi>TrD)=Pr(\xi^TD^k\xi>TrD^k),\quad k=1,2,3,4$$

since changing the ξ_1,ξ_2,ξ_3,ξ_4 elements of vector ξ with negative ones for D^1,D^2,D^3,D^4, respectively, we obtain the same result as with matrix D.

Now, if it is true that

$$Pr(\xi^TD^1\xi>TrD^1)=Pr(-\xi_1\xi_2D_{12}-\xi_1\xi_3D_{13}-\xi_1\xi_4D_{14}+\xi_2\xi_3D_{23}+\xi_2\xi_4D_{24}+\xi_3\xi_4D_{34}>0)>\frac{3}{4}$$

$$Pr(\xi^T D^2 \xi > TrD^2) = Pr(-\xi_1\xi_2 D_{12} + \xi_1\xi_3 D_{13} + \xi_1\xi_4 D_{14} - \xi_2\xi_3 D_{23} - \xi_2\xi_4 D_{24} + \xi_3\xi_4 D_{34} > 0) > \frac{3}{4}$$

then we have

$$Pr(-\xi_1\xi_2 D_{12} + \xi_3\xi_4 D_{34} > 0) > \frac{1}{2} \quad (\frac{3}{4} + \frac{3}{4} - 1). \tag{3.18}$$

On the other hand, if

$$Pr(\xi^T D^3 \xi > TrD^3) = Pr(+\xi_1\xi_2 D_{12} - \xi_1\xi_3 D_{13} + \xi_1\xi_4 D_{14} - \xi_2\xi_3 D_{23} + \xi_2\xi_4 D_{24} - \xi_3\xi_4 D_{34} > 0) > \frac{3}{4}$$

$$Pr(\xi^T D^4 \xi > TrD^4) = Pr(+\xi_1\xi_2 D_{12} + \xi_1\xi_3 D_{13} - \xi_1\xi_4 D_{14} + \xi_2\xi_3 D_{23} - \xi_2\xi_4 D_{24} - \xi_3\xi_4 D_{34} > 0) > \frac{3}{4}$$

then

$$Pr(+\xi_1\xi_2 D_{12} - \xi_3\xi_4 D_{34} > 0) > \frac{1}{2} \quad (\frac{3}{4} + \frac{3}{4} - 1). \tag{3.19}$$

Again, both (3.18) and (3.19) cannot be true at the same time. Therefore, we obtained a contradiction.

Up to this point, we proved the base cases. Because there is no off-diagonal element for $n = 1$, we can focus on the other base cases which are 2,3,4. If a larger matrix can be decomposed (for nonzero elements) into these smaller matrices from their diagonals, we can easily say that the resulting matrices also obey the 1/4 criteria that led to the contradictions above. If this decomposition is not possible, then there exists a matrix B which has nonzero elements:$\{B_{12}, B_{34}, B_{35}, B_{45}, -B_{67}, -B_{68}, -B_{69}, B_{78}, B_{79}, B_{89}, \}$. Moreover there exist B^1, B^2, B^3 matrices which give the same result as with matrix B and their nonzero elements are

$$\{-B_{12}, -B_{34}, -B_{35}, B_{45}, -B_{67}, B_{68}, B_{69}, -B_{78}, -B_{79}, B_{89}, \} \text{ for } B^1$$

$$\{B_{12}, -B_{34}, B_{35}, -B_{45}, B_{67}, -B_{68}, B_{69}, -B_{78}, B_{79}, -B_{89}, \} \text{ for } B^2$$

$$\{-B_{12}, B_{34}, -B_{35}, -B_{45}, B_{67}, B_{68}, -B_{69}, B_{78}, -B_{79}, -B_{89}, \} \text{ for } B^3.$$

By using the same steps above, we can get a contradiction. Therefore, our aim is to get these bases matrices from a larger matrix. To do this, we developed

the meiosis (a term borrowed from biology) procedure below which divides the matrix into four smaller submatrices at each step.

Let us now concentrate on $n > 4$. Assume that the result holds true for $n < k$, $k > 4$, and let us look at $n = k$ by using the contradiction method. Assume that the conclusion is false for $n = k$, i.e., there exists $k \times k$ symmetric matrix B such that

$$Pr(\xi^T B\xi \leq TrB) < \frac{1}{2^{\lceil \log_2(k) \rceil}}.$$

Now, we define the symmetric matrices B^1, B^2, B^3, B^4 and begin the meiosis procedure:

$$B^1_{ij} = \left\{ \begin{array}{ll} -B_{ij} & i = 1, ...\lceil \frac{k}{4} \rceil, j \neq 1, ...\lceil \frac{k}{4} \rceil, i < j \\ B^1_{ji} & i \neq j \\ B_{ij} & i = j, o.w. \end{array} \right\},$$

$$B^2_{ij} = \left\{ \begin{array}{ll} -B_{ij} & i = \lceil \frac{k}{4} \rceil + 1, ..., 2\lceil \frac{k}{4} \rceil, j \neq \lceil \frac{k}{4} \rceil + 1, ..., 2\lceil \frac{k}{4} \rceil, i < j \\ B^2_{ji} & i \neq j \\ B_{ij} & i = j, o.w. \end{array} \right\},$$

$$B^3_{ij} = \left\{ \begin{array}{ll} -B_{ij} & i = 2\lceil \frac{k}{4} \rceil + 1, ..., \min(3\lceil \frac{k}{4} \rceil, k), j \neq 2\lceil \frac{k}{4} \rceil + 1, ..., \min(3\lceil \frac{k}{4} \rceil, k), i < j \\ B^3_{ji} & i \neq j \\ B_{ij} & i = j, o.w. \end{array} \right\}.$$

If $\min(3\lceil \frac{k}{4} \rceil, k) = k$ then we let $B^4 = B$, otherwise we take

$$B^4_{ij} = \left\{ \begin{array}{ll} -B_{ij} & i = 3\lceil \frac{k}{4} \rceil + 1, ..., k, j \neq 3\lceil \frac{k}{4} \rceil + 1, ..., k, i < j \\ B^4_{ji} & i \neq j \\ B_{ij} & i = j, o.w. \end{array} \right\}.$$

For example if $n = 5$, we subdivide into $\{1, 2\}, \{3, 4\}, \{5\}, \{\}$ $(B^4 = B)$. For $n = 9$, we subdivide into $\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9\}, \{\}$ $(B^4 = B)$. For $n = 11$, we subdivide into $\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9\}, \{10, 11\}$.

Now, we have

$$Pr(\xi^T B\xi > TrB) = Pr(\xi^T B^l \xi > TrB^l), \quad l = 1, 2, 3, 4$$

by negating the elements of vector ξ corresponding to the changed element of B^l. For example, for $n = 5$ and $B^l = B^1$, make (ξ_1, ξ_2) take negative of their values in multiplication with B. We obtain the same result as with B.

If

$$Pr(\xi^T B^1 \xi - TrB^1 > 0) > 1 - \frac{1}{2^{\lceil \log_2(k) \rceil}}$$

and

$$Pr(\xi^T B^2 \xi - TrB^2 > 0) > 1 - \frac{1}{2^{\lceil \log_2(k) \rceil}}$$

then we have

$$Pr(\xi^T (B^1 + B^2) \xi - Tr(B^1 + B^2) > 0) > 1 - \frac{2}{2^{\lceil \log_2(k) \rceil}} \tag{3.20}$$

On the other hand, if

$$Pr(\xi^T B^3 \xi - TrB^3 > 0) > 1 - \frac{1}{2^{\lceil \log_2(k) \rceil}}$$

and

$$Pr(\xi^T B^4 \xi - TrB^4 > 0) > 1 - \frac{1}{2^{\lceil \log_2(k) \rceil}}$$

then we have

$$Pr(\xi^T (B^3 + B^4) \xi - Tr(B^3 + B^4) > 0) > 1 - \frac{2}{2^{\lceil \log_2(k) \rceil}}. \tag{3.21}$$

From (3.20) and (3.21) we obtain

$$Pr(\xi^T (B^1 + B^2 + B^3 + B^4) \xi - Tr(B^1 + B^2 + B^3 + B^4) > 0) > 1 - \frac{4}{2^{\lceil \log_2(k) \rceil}} \tag{3.22}$$

which is the same as

$$Pr(\sum_{\substack{i<j \\ i<\lceil \frac{k}{4} \rceil \\ j \leq \lceil \frac{k}{4} \rceil}} \xi_i \xi_j B_{ij} + \sum_{\substack{i<j \\ \lceil \frac{k}{4} \rceil + 1 \leq i < 2\lceil \frac{k}{4} \rceil \\ \lceil \frac{k}{4} \rceil + 1 < j \leq 2\lceil \frac{k}{4} \rceil}} \xi_i \xi_j B_{ij} + \sum_{\substack{i<j \\ 2\lceil \frac{k}{4} \rceil + 1 \leq i < \min(3\lceil \frac{k}{4} \rceil, k) \\ 2\lceil \frac{k}{4} \rceil + 1 < j \leq \min(3\lceil \frac{k}{4} \rceil, k)}} \xi_i \xi_j B_{ij} +$$

$$\sum_{\substack{if\ \min(3\lceil \frac{k}{4} \rceil, k) = k,\ empty \\ otherwise, \quad i<j \\ 3\lceil \frac{k}{4} \rceil + 1 \leq i < k \\ 3\lceil \frac{k}{4} \rceil + 1 < j \leq k}} \xi_i \xi_j B_{ij} > 0) > 1 - \frac{4}{2^{\lceil \log_2(k) \rceil}}$$

$$= Pr(\sum_{\substack{i<j \\ i<\lceil \frac{k}{4} \rceil \\ j \leq \lceil \frac{k}{4} \rceil}} \xi_i \xi_j B_{ij} > 0) > 1 - \frac{1}{2^{\lceil \log_2(k) \rceil - 2}} = 1 - \frac{1}{2^{\lceil \log_2(\lceil \frac{k}{4} \rceil) \rceil}}$$

This gives a contradiction using the induction argument since the matrix in the last summation is a $\lceil \frac{k}{4} \rceil \times \lceil \frac{k}{4} \rceil$ matrix. More precisely, the last step consists in applying the meiosis procedure for each of the four summations until we get to the point where none of them contains more than a 4-dimensional matrix. Observe that all the summations are independent from one another. Therefore we can invoke the base cases $n = 2, 3, 4$ given at the beginning. The proof is complete.

This concludes the chapter. In Chapter 4, we talk about the impact of the previous results and present an example about using approximate S-Lemma.

Chapter 4

Evaluation

In this chapter, 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 better result from corollary of Poon if we have positive linear combination of given matrices.

For 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, it is:

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

However, you need additional assumption in the second case. In fact this assumption is same as positive definiteness of a linear combination of matrices. One can reach this result by observing $\langle\langle AX, X\rangle\rangle = \sum_{i=1}^{p} x^T Ax$ 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 chapter. Also Polyak gives an analysis for $m = 2$ case. For generalization of this result, Uhlig's survey is a useful paper.

Although we extend 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 $m > 2$.

In the second part of results, we dealt with approximate S-Lemma. We tried to improve the bounds for both dyadic and general case. Although we only obtain the relaxed version of conjecture of dyadic case, we improved the bound of general case.

In the dyadic case, our result does not carry any meaning to improve the bound. Because we just proved for the continuous case, but the lemma requires the discrete case. In fact, the conjecture is very strong. We know that there is no better bound for the inequality. From this conjecture, we also obtain the following sub-conjecture which is also an open problem.

Sub-Conjecture: *For* $x = (\frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}})$, $|\xi^T x| \leq 1$ *becomes* $|\sum_{i=1}^{k} x_i - \sum_{i=k+1}^{n} x_i| \leq 1$. *Therefore the probability of inequality is same as:*

$$\frac{\begin{pmatrix} n \\ \lceil \frac{n-\sqrt{n}}{2} \rceil \end{pmatrix} + ... + \begin{pmatrix} n \\ n - \lceil \frac{n-\sqrt{n}}{2} \rceil \end{pmatrix}}{2^n} \geq \frac{1}{2}.$$

The final result of previous chapter is about general case of approximate S-Lemma. In this case, we improve the bound from $\frac{1}{8n^2}$ to $\frac{1}{2n}$. Final version of approximation of general case is:

$$\rho := (\ 2log(\ 4n \textstyle\sum_{k=1}^{K} \text{rank } R_k\ )\ )^{\frac{1}{2}}, \tag{4.1}$$

which was previously

$$\rho := (\ 2log(\ 16n^2 \textstyle\sum_{k=1}^{K} \text{rank } R_k\ )\ )^{\frac{1}{2}}. \tag{4.2}$$

We offered an improvement from n^2 to n under the logarithm. Moreover the improvement is better for small $\sum_{k=1}^{K} \text{rank } R_k$, because it is fixed term in both

(4.1) and (4.2). In addition, if the matrix

$$M = \left[\begin{array}{c|c|c|c} A & & & \\ \hline & B & & \\ \hline & & C & \\ \hline & & & D \end{array}\right]$$

has square sub-matrices A, B, C, D on the diagonal which are zero matrices, then the procedure of the proof solves it in the first step and give the $\frac{1}{4}$ result what the conjecture says.

After explaining effects of results, let us continue with an example of application for approximate S-Lemma. From Chapter 2, we see that the approximate S-Lemma can be used to find relationship between uncertain quadratic constraints and their robust counterparts. Remember that we gave an example of a general case of uncertain quadratic constraints in the previous chapter. Now we deal with an another example that includes a specific uncertain set called norm-1 uncertainty. We define

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

where

$$U_\rho = \left\{\ (A,b,c) = (A^0,b^0,c^0) + \rho \textstyle\sum_{l=1}^{L} y_l (A^l,b^l,c^l) \ : \ \|y\|_1 \leq 1 \ \right\}.$$

with

$$A^l \in M_{n,p}(\mathbf{R}), \quad b^l \in \mathbf{R}^p, \quad c^l \in \mathbf{R}, \quad l = 0, ..., L.$$

Let us focus on $\|y\|_1 \leq 1$. It is the intersection of 2^L hyperspaces which are $(\mp x_1 ... \mp x_L \leq 1)$. These hyperspaces can be defined with 2^{L-1} squared inequalities which are $((+x_1 \mp x_2 ... \mp x_L)^2 \leq 1)$

Note: The $L \times L$ symmetric matrix Q which has range

$$(x_1, ..., x_L)^T Q (x_1, ..., x_L) = (x_1 + .. + x_k - x_{k+1} - .. - x_L)^2$$

can be defined as,

$$Q_{ij} = \left\{ \begin{array}{cc} -1 & i < j, i = k+1, ..., L \ j \neq k+1, ..., L \\ Q_{ji} & j < i \\ 1 & otherwise \end{array} \right\}.$$

Now we know that these inequalities can be represented by a positive semi-definite matrix which are given in the note above.

We also know that summation of these 2^{L-1} positive semi-definite matrix is positive definite matrix, because if any nonzero vector makes one of the matrices zero, then

$$(\mp x_1 ... \mp x_L)^2 = 0 \Rightarrow \mp x_1 ... \mp x_L = 0.$$

At least two elements of the vector should be nonzero. Among these 2^{L-1} matrices, there exists a matrix which makes one of the nonzero elements (not first one) of the vector negative. Therefore we obtain the nonzero value for the new matrix which is already in the system. Then summation of these matrices is positive definite matrix.

Therefore, we can apply Approximate S-Lemma to the problem by making same steps as in (19) and (20), and obtain robust counterpart given in the (2.17) with $K = 2^{L-1}$ and the level of conservativeness which is:

$$\rho := \Big(\, 2 \log \Big(\, 6 \textstyle\sum_{l=1}^{2^{L-1}} \operatorname{rank} Q_l \, \Big) \, \Big)^{\frac{1}{2}}$$

where

$$Q_l = (x_1 \mp x_2 ... \mp x_L)^2 \quad l = 1, ..., 2^{L-1}.$$

Only remaining is rank analysis of matrices. The matrix that has following form,

$$(x_1, ..., x_L)^T Q (x_1, ..., x_L) = (x_1 + .. + x_k - x_{k+1} - .. - x_L)^2$$

can be written as:

$$Q = xy^T$$

where $x \in \mathbf{R}^L$, $y \in \mathbf{R}^L$ with

$$x = (x_1, .., x_k, -x_{k+1}, .., -x_L)$$

$$y = (x_1 + .. + x_k - x_{k+1} - .. - x_L, ..., x_1 + .. + x_k - x_{k+1} - .. - x_L).$$

Therefore we can say that rank Q is one. (see Matrix Analysis book of Horn and Johnson (pp.61)) At the end, we obtain the level of conservativeness which is:

$$\tilde{\rho} := \left(\ 2\log(6)(2^{L-1}) \ \right)^{\frac{1}{2}}.$$

Hence, we found $O(\sqrt{L})$ bound for norm-1 cases by using approximate S-Lemma. Similar bound can also be found by Nesterov in Handbook of Semidefinite Programming book [39] (pp.387). For the problem:

$$P^* = max \left\{ \ \langle Ax, x\rangle \ \ : \ \ \|x\|_1 \leq 1, \ \ x \in \mathbf{R}^L \ \right\}.$$

For an indefinite A with its maximal eigenvalue $\lambda_{max}(A)$, Nesterov gave the inequality

$$\lambda_{max}(A) \geq P^* \geq \frac{\lambda_{max}(A)}{L}.$$

For calculating quadratic function $\langle Ax, x\rangle$, take the square of our bound and there is no need to use R_0 of approximate S-Lemma in that problem so coefficient of bound is:

$$\tilde{\rho}^2 := 2\log(2)(2^{L-1}) = L\log 4,$$

and the bound is:

$$P^* \geq \tfrac{SDP}{\tilde{\rho}^2}, \ \ SDP = \min_{\mu} \ (\textstyle\sum_{l=1}^{2^{L-1}} \mu_l \ \ : \ \ \sum_{l=1}^{2^{L-1}} \mu_l Q_l \succeq A, \ \mu \geq 0),$$

that depends on also L. For $\lambda_{max}(A) \geq \frac{SDP}{\log 4}$, even though bound of Nesterov is slightly better than ours, that uses specific character of norm 1 case, both of them depend on L.

It is the end of the example and the chapter. Although important results are found by mathematicians, there are still some open problems. In the next chapter, we talk about some possible future works and conclude the thesis.

Chapter 5

Conclusion

In this study, we dealt with S-procedure and its variants. Since it is a fundamental tool of different fields such as optimal control and robust optimization, both S-procedure and approximate S-Lemma are important for us. 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 to NP complete. Hence to find 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 mathematical world.

For general case, we dealt with corollary of the theorems of Barvinok and Poon to understand their meaning for S-procedure. It also gives an idea of 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 take an assumption of positive definiteness of a linear combination of matrices. This corollary also gives the same result with Polyak's theorem for classical vector case.

In the case of S-procedure, the best result we get 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 what the minimal assumptions are to satisfy the case $m > 2$. This attractive problem is still open and is potentially the future work of many mathematicians.

Then, we turned our interest into approximate S-Lemma. We firstly work on the dyadic case. We find the extended version of conjecture needed. Our claim is for continuous case but the necessary case is the discrete case which is related with the popular subject of today's mathematicians: number theory. Therefore, our lemma is not enough to improve the bound. The conjecture is still waiting for someone to prove or disprove it.

Our second interest on approximate S-Lemma is the conjecture of general case. The bound given by Ben-Tal *et.al.* is $\frac{1}{8n^2}$ and we improve this bound to $\frac{1}{2n}$. These coefficients are given under the logarithm in total bound. Also this bound is related on summation of ranks of matrices. Therefore our improvement is better for small summation of ranks of matrices. Although we improved the bound, we were not able to prove the conjecture given by Ben-Tal *et.al.* which is $\frac{1}{4}$. It is the other possible future work for researchers.

Finally, we give an example about norm 1 and try to solve it with approximate S-Lemma. We get $O(\frac{1}{n})$ bound for this problem which is similar with the bound of Nesterov. People also try to improve this bound. For example, Pinar [31] finds an $O(\frac{1}{logn})$ bound but the relaxed problem he uses is not convex. Therefore, it is an another open area for further research.

Although getting an improvement is difficult in this area, any little success is important for several fields of science. Therefore this subject keeps its attractiveness in the future even though we come to the end of our journey: *cursum perficio.*

Bibliography

[1] Au-Yeung, Y. H., A theorem on a mapping from a sphere to the circle and the simultaneous diagonalisation of two hermitian matrices. *Proc. Amer. Math. Soc.*, 20:545–548, 1969.

[2] Au-Yeung, Y. H., and Poon, Y. T., A remark on the convexity and positive definiteness concerning hermitian matrices. *Southeast Asian Bull. Math.*, 3:85–92, 1979.

[3] Barvinok, A. I., Problems of distance geometry and convex properties of quadratic maps. *Discrete Comput. Geom.*, 13:189–202, 1995.

[4] Ben-Tal, A., Conic and robust optimization. Technical report, Israel Institute of Technology, Technion, July 2002.

[5] Ben-Tal, A., Goryashko, A., Guslitzer, E., and Nemirovski, A., Adjustable robust solutions to uncertain linear programs. *Math. Prog.*, 99:351–376, 2004.

[6] Ben-Tal, A., and Nemirovski, A., Robust convex optimization. *Math. Oper. Res.*, 23:769–805, 1998.

[7] Ben-Tal, A., and Nemirovski, A., *Lectures on Modern Convex Optimization: Analysis, Algorithms and Engineering Applications*. SIAM-MPS, Philadelphia, 2000.

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

[9] Bohnenblust, H. F., Joint positiveness of matrices. Unpublished manuscript.

[10] Boyd, S., El-Ghaoui, L., Feron, E., and Balakhrishnan, V., *Linear Matrix Inequalities in Systems and Control Theory*. SIAM, Philadelphia, 1994.

[11] Brickman, L., On the fields of values of a matrix. *Proc. Amer. Math. Soc.*, 12:61–66, 1961.

[12] Calabi, E., Linear systems of real quadratic forms. *Proc. Amer. Math.*, 15:844–846, 1964.

[13] Dines, L. L., On the mapping of quadratic forms. *Bull. Amer. Math. Soc.*, 47:494–498, 1941.

[14] Dines, L. L., On the mapping of n quadratic forms. *Bull. Amer. Math. Soc.*, 48:467–471, 1942.

[15] Dines, L. L., On linear combinations of quadratic forms. *Bull. Amer. Math. Soc.*, 49:388–393, 1943.

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

[17] El-Ghaoui, L., Oustry, F., and Lebret, H., Robust solutions to uncertain semidefinite programs. *SIAM J. Optim.*, 9:33–52, 1998.

[18] Finsler, P., Über das Vorkommen definiter und semidefiniter Formen in Scharen quadratischer Formen. *Comment. Math. Helv.*, 9:188–192, 1936/37.

[19] Fradkov, A. L., Duality theorems for certain nonconvex extremum problems. *Siberian Math. J.*, 14:247–264, 1973.

[20] Hausdorff, F., Der Wertvorrat einer Bilinearform. *Math. Z.*, 3:314–316, 1919.

[21] Hestenes, M. R., and McShane, E. J., A theorem on quadratic forms and its application in the calculus of variations. *Trans. Amer. Math. Soc.*, 40:501–512, 1940.

[22] Hiriart-Urruty, J. B., and Torki, M., Permanently going back and forth between the quadratic world and the convexity world in optimization. *Appl. Math. Optim.*, 45:169–184, 2002.

[23] Horn, R. A., and Johnson, C. R., *Matrix analysis*. Cambridge University Press, New York, 1990.

[24] John, F., A note on the maximum principle for elliptic differential equations. *Bull. Amer. Math. Soc.*, 44:268–271, 1938.

[25] Klerk, E. de, *Aspects of Semidefinite Programming*. Kluwer Academic Publishers, Dordrecht, 2002.

[26] Kühne, R., Über eine Klasse J-selbstadjungierter Operatoren. *Math. Ann.*, 154:56–69, 1964.

[27] Luo, Z. Q., Sturm, J. F., and Zhang, S., Multivariate nonnegative quadratic mappings. Technical report SEEM2003-02, The Chinese University of Hong Kong, January 2003.

[28] Lur'e, A. I., and Postnikov, V. N., On the theory of stability of control systems. *Applied Mathematics and Mechanics*, 8(3), 1944. in Russian.

[29] Megretsky, A., and Treil, S., Power distribution inequalities in optimization and robustness of uncertain systems. *Math. Systems Estimation Control*, 3:301–319, 1993.

[30] Nemirovski, A., Five lectures on modern convex optimization. Technical report, Israel Institute of Technology, Technion, August 2002.

[31] Pinar, M. Ç., Private communication.

[32] Polyak, B. T., Convexity of quadratic transformations and its use in control and optimization. *J. Optim. Theory Appl.*, 99:553–583, 1998.

[33] Poon, Y. T., Generalized numerical ranges, joint positive definiteness and multiple eigenvalues. *Proc. Amer. Math. Soc.*, 125:1625–1634, 1997.

[34] Sommerville, D. M. Y., *An Introduction to the Geometry of n Dimensions*. Dover Publications, New York, 1958.

[35] Sturm, J. F., and Zhang, S., On cones of nonnegative quadratic functions. Technical Report SEEM2001-01, The Chinese University of Hong Kong, March 2001. To appear in Mathematics of Operations Research.

[36] Taussky, O., *Positive-definite matrices, in Inequalities (O. Shisha, Ed.)*, pages 309–319. Academic, New York, 1967.

[37] Toeplitz, O., Das algebraische Analogen zu einem Satze von Fejér. *Math. Z.*, 2:187–197, 1918.

[38] Uhlig, F., A recurring theorem about pairs of quadratic forms and extension: a survey. *Linear Algebra Appl.*, 25:219–237, 1979.

[39] Wolkowicz, H., Saigal, R., and Vandenberghe, L., Edited by, *Handbook of semidefinite programming: theory, algorithms, and applications*. Kluwer Academic Publishers, Massachusetts, 2000.

[40] Yakubovich, V. A., The S-procedure in nonlinear control theory. *Vestnik Leningr. Univ.*, 4:73–93, 1977. in Russian — 1971, No.1, 62–77.

[41] Yuan, Y., On a subproblem of trust region algorithms for constrained optimization. *Math. Programming*, 47:53–63, 1990.

Appendix A

Application of S-Lemma on Robust Optimization

Theorem 29 *For $A^l \in M_{n,p}(\mathbf{R})$, $b^l \in \mathbf{R}^p$, $c^l \in \mathbf{R}$, $l = 0, ..., L$ a vector $x \in \mathbf{R}^p$ is a solution of*

$$x^T A^T A x \leq 2b^T x + c \quad \forall (A, b, c) \in U_{simple}, \tag{A.1}$$

where

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

if and only if for some $\lambda \in \mathbf{R}$ the pair (x, λ) is a solution of the following linear matrix inequality (LMI):

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

Proof: Using uncertain set, (2.4) can be written as:

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

Taking $\tau \leq 1$,

$$-x^T[A^0\tau+\sum_{l=1}^{L} y_l A^l]^T[A^0\tau+\sum_{l=1}^{L} y_l A^l]x+2\tau[b^0\tau+\sum_{l=1}^{L} y_l b^l]^T x+\tau[c^0\tau+\sum_{l=1}^{L} y_l c^l] \geq 0 \ \ \forall((\tau,y) : \|y\|^2 \leq \tau^2).$$

Clearly, If $\tau^2 - \|y\|^2 \geq 0$ 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

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

which is the same as

$$(\tau, y^T)\left[\begin{pmatrix} c^0+2x^Tb^0 & \frac{1}{2}c^1+x^Tb^1 & . & . & \frac{1}{2}c^L+x^Tb^L \\ \frac{1}{2}c^1+x^Tb^1 & & & & \\ . & & & & \\ . & & & & \\ \frac{1}{2}c^L+x^Tb^L & & & & \end{pmatrix} - (A^0x, A^1x, .., A^Lx)\begin{pmatrix} A^0x \\ A^1x \\ . \\ . \\ A^Lx \end{pmatrix}\right]\begin{pmatrix} \tau \\ y \end{pmatrix}$$

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

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

Lemma 30 *(Schur complement) Suppose A, B, C are respectively $n \times n$, $n \times p$, and $p \times p$ matrices. Also C is positive definite and A is symmetric. Let*

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

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

$$A - BC^{-1}B^T.$$

The following are equivalent:

- *M is positive (semi)definite*
- *$A - BC^{-1}B^T$ is positive (semi)definite.*

Proof: The proof can be seen from de Klerk's book [25], (pp.236).

Appendix B

Approximate S-Lemma and Its Proof

Lemma 31 *(Approximate S-Lemma). Let $R, R_0, R_1, ..., R_k$ be symmetric $n \times n$ matrices such that*

$$R_1, ..., R_k \succeq 0, \tag{B.1}$$

and assume that

$$\exists \lambda_0, \lambda_1, ..., \lambda_k \geq 0 \; s.t. \; \sum_{k=0}^{K} \lambda_k R_k \succ 0. \tag{B.2}$$

Consider the following quadratically constrained quadratic program,

$$QCQ = \max_{y \varepsilon R^n} \{ \; y^T R y \;\; : \;\; y^T R_0 y \leq r_0, y^T R_k y \leq 1, k = 1, ..., K \; \} \tag{B.3}$$

and the semidefinite optimization problem

$$SDP = \min_{\mu_0, \mu_1, ..., \mu_K} \{ \; r_0 \mu_0 + \textstyle\sum_{k=1}^{K} \mu_k \;\; : \;\; \sum_{k=0}^{K} \mu_k R_k \succeq R, \mu \geq 0 \; \}. \tag{B.4}$$

Then

(i) If problem (B.3) is feasible, then problem (B.4) is bounded below and

$$SDP \geq QCQ. \tag{B.5}$$

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

$$y_*^T R y_* = SDP, \tag{B.6}$$

$$y_*^T R_0 y_* \leq r_0, \tag{B.7}$$

$$y_*^T R_k y_* \leq \tilde{\rho}^2, \ k = 1, ..., K, \tag{B.8}$$

where

$$\tilde{\rho} := (\ 2log(\ 6 \sum_{k=1}^K rank\, R_k\ )\ )^{\frac{1}{2}}, \tag{B.9}$$

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

$$\tilde{\rho} := (\ 2log(\ 16n^2 \sum_{k=1}^K rank\, R_k\ )\ )^{\frac{1}{2}} \tag{B.10}$$

otherwise.

(ii) If

$$r_0 > 0, \tag{B.11}$$

then (B.3) is feasible, problem (B.4) is solvable, and

$$0 \leq QCQ \leq SDP \leq \tilde{\rho}^2 QCQ. \tag{B.12}$$

Proof: Notice that problem (B.4) is the SDP dual of

$$RQCQ = \max_{X \succeq 0} \{\ TrRX \ \ : \ \ TrR_0X \leq r_0, TrR_kX \leq 1, k = 1, ..., K\ \}. \tag{B.13}$$

The latter problem is the standard SDP-relaxation of the QCQ-problem (B.3), so we have

$$RQCQ \geq QCQ. \tag{B.14}$$

In part (i) of the lemma, (B.3) is assumed to be feasible, hence (B.13) is feasible as well, and hence, by weak duality, between problem (B.4) and its dual (B.13), problem (B.4) is bounded below. Now assumption (B.2) ensures that (B.4) is strictly feasible, thus from SDP-duality theory, problem (B.13) is solvable and

$$SDP = RQCQ. \tag{B.15}$$

By (B.14) and (B.15), $SDP \geq QCQ$, which completes the proof of the first part of claim (i) in the lemma. To prove the second part, we first simplify the system (B.6)-(B.8). Letting X_* denote an optimal solution of problem (B.13), we introduce

$$\hat{R} = X_*^{\frac{1}{2}} R X_*^{\frac{1}{2}}. \tag{B.16}$$

Let

$$\hat{R} = U\tilde{R}U^T \quad (U^TU = I, \tilde{R} = diag(r_1, ..., r_n)) \tag{B.17}$$

be the eigenvalue decomposition of $\hat{R}$. Choosing

$$y_* = X_*^{\frac{1}{2}} U u, \quad u \in \mathbf{R}^n, \tag{B.18}$$

we have

$$y_*^T R y_* = u^T U^T X_*^{\frac{1}{2}} R X_*^{\frac{1}{2}} U u = u^T U^T \hat{R} U u = u^T \tilde{R} u = \sum_{i=1}^n r_i u_i^2.$$

Also

$$SDP = RQCQ = TrRX_* = Tr\hat{R} = Tr\tilde{R} = \sum_{i=1}^n r_i,$$

and thus (B.6) is equivalent to

$$\sum_{i=1}^n r_i u_i^2 = \sum_{i=1}^n r_i.$$

Now defining

$$\hat{R}_k = X_*^{\frac{1}{2}} R_k X_*^{\frac{1}{2}}, \quad \tilde{R}_k = U^T \hat{R} U, \quad k = 0, 1, ..., K,$$

and using (B.18), we obtain

$$y_*^T R_k y_* = u^T U^T X_*^{\frac{1}{2}} R_k X_*^{\frac{1}{2}} U u = u^T U^T \hat{R} U u = u^T \tilde{R} u. \tag{B.19}$$

Since X_* solves $RQCQ$,

$$r_0 \geq TrR_0X_* = Tr\hat{R}_0 = Tr\tilde{R}_0 \tag{B.20}$$

and

$$1 \geq TrR_kX_* = Tr\hat{R}_k = Tr\tilde{R}_k, \quad k = 0, 1, ..., K. \tag{B.21}$$

From (B.19) and (B.20) we see that (B.7) holds if

$$u^T \tilde{R}_0 u \leq Tr\tilde{R}_0,$$

from (B.19) and (B.21), relation (B.8) holds if

$$u^T \tilde{R}_k u \leq \tilde{\rho}^2 Tr\tilde{R}_k, \quad k = 1, ..., K.$$

We conclude that if there exist a $\overline{u}$ satisfying

$$\sum_{i=1}^{n} r_i \overline{u}_i^2 = \sum_{i=1}^{n} r_i, \tag{B.22}$$

$$\overline{u}^T \tilde{R}_0 \overline{u} \leq Tr\tilde{R}_0, \tag{B.23}$$

$$\overline{u}^T \tilde{R}_k \overline{u} \leq \tilde{\rho}^2 Tr\tilde{R}_k, \quad k = 1, ..., K, \tag{B.24}$$

then $y_* = X_*^{\frac{1}{2}} U \overline{u}$ satisfies (B.6)-(B.8). Note that (B.22) is automatically satisfied if $\overline{u}$ is a ± 1-vector. Thus it suffices to show that (B.23) and (B.24) can be satisfied by a ± 1-vector $\overline{u}$.

Let us pretend for a moment that the vector $\overline{u}$ is a random ± 1-vector such that $Pr(\overline{u}_i = 1) = Pr(\overline{u}_i = -1) = \frac{1}{2}$ for each i. Let B denote the event that $\overline{u}$ satisfies (B.23) and C_k the event that $\overline{u}^T \tilde{R}_k \overline{u} \leq \tilde{\rho}^2 Tr\tilde{R}_k$, and $C = \cap_k C_k$, i.e., C denotes the event that $\overline{u}$ satisfies (B.24). Then we only need to show that

$$Pr(B \cap C) > 0. \tag{B.25}$$

(In the original paper, there is a notation error for $\cap, \cup$. We give here a corrected version of it. It does not affect the proof). Since

$$B \subseteq (B \cap C) \cup C^c,$$

we have

$$Pr(B) \leq Pr((B \cap C) \cup C^c) \leq Pr(B \cap C) + Pr(C^c),$$

$$Pr(B \cap C) \geq Pr(B) - Pr(C^c) = Pr(B) - Pr((\cap_k C_k)^c),$$

$$= Pr(B) - Pr(\cup_k C_k^c) \geq Pr(B) - \sum_{k=1}^{K} Pr(C_k^c).$$

Hence (B.25) will certainly hold for some $p_0 > 0$,

$$Pr(B) > p_0. \tag{B.26}$$

Therefore, we have

$$\sum_{k=1}^{K} Pr(C_k^c) \leq p_0. \tag{B.27}$$

At this point, we need extra lemmas from the paper of Ben-Tal *et.al* [8] (pp. 551–554) to complete the proof which are:

Lemma 32 *Let $x = \{x_1, , , , x_n\}$ and $\xi = \{\xi_1, , , , \xi_n\} \in \mathbf{R}^n$. If $\|x\|_2 = 1$ and the coordinates ξ_i of ξ are independently identically distributed random variables with*

$$Pr(\xi_i = 1) = Pr(\xi_i = -1) = 1/2 \tag{B.28}$$

then one has,

$$Pr(|\xi^T x| \leq 1) \geq 1/3. \tag{B.29}$$

Lemma 33 *Let rank $B = k$, $B \succeq 0$, and $\xi = \{\xi_1, , , , \xi_n\} \in \mathbf{R}^n$. The coordinates ξ_i of ξ are independently identically distributed random variables with*

$$Pr(\xi_i = 1) = Pr(\xi_i = -1) = 1/2 \tag{B.30}$$

then

$$Pr(\xi^T B \xi \geq \alpha Tr B) \leq 2ke^{-\frac{\alpha}{2}}, \ \ \forall \alpha > 0. \tag{B.31}$$

Lemma 34 *Let B denote a symmetric $n \times n$ matrix and $\xi = \{\xi_1, , , , \xi_n\} \in \mathbf{R}^n$. The coordinates ξ_i of ξ are independently identically distributed random variables with*

$$Pr(\xi_i = 1) = Pr(\xi_i = -1) = 1/2 \tag{B.32}$$

then one has

$$Pr(\xi^T B\xi \leq TrB) > \frac{1}{8n^2}. \tag{B.33}$$

We first consider the case in which R_0 is dyadic. Then $\hat{R}_0$ and $\tilde{R}_0$ are also dyadic, and hence we may write, for a suitable vector b,

$$\tilde{R}_0 = bb^T.$$

Then condition (B.23) is equivalent to

$$Pr(|u^T x| \leq 1) > p_0, \tag{B.34}$$

where x is the unit vector $\frac{b}{\|b\|}$. By Lemma 32 this inequality certainly holds if $p_0 = \frac{1}{3}$. On the other hand, by Lemma 33, for each k,

$$Pr(C_k^c) = Pr(\overline{u}^T \tilde{R}_k \overline{u} > \tilde{\rho}^2 Tr\tilde{R}_k) \leq 2(\text{rank } \tilde{R}_k)e^{-\frac{\tilde{\rho}^2}{2}}.$$

Since rank $\tilde{R}_k \leq$ rank R_k, we obtain

$$\sum_{k=1}^{K} Pr(C_k^c) \leq 2e^{-\frac{\tilde{\rho}^2}{2}} \sum_{k=1}^{K} \text{rank } R_k,$$

and so inequalities (B.26) and (B.27) will hold if $p_0 = \frac{1}{3}$ and $\tilde{\rho}$ is such that

$$2e^{-\frac{\tilde{\rho}^2}{2}} \sum_{k=1}^{K} \text{rank } R_k = \frac{1}{3} = p_0. \tag{B.35}$$

One may easily verify that the value of $\tilde{\rho}$ as given by (B.9) is indeed the solution of (B.35). Thus the proof is complete for the case where R_0 is dyadic.

We finally consider the general case, where R_0 is an arbitrary symmetric matrix. Then we apply Lemma 34, which gives that (B.26) holds for $p_0 = \frac{1}{8n^2}$. Then solving $\tilde{\rho}$ from (B.35) with this value of p_0 we get the value given in (B.10).

To complete the proof of the lemma we only need to prove $SDP \leq \tilde{\rho}^2 QCQ$, the last inequality in (B.12). For this, let y_* satisfy (B.6)-(B.8). Then, since $\tilde{\rho} > 1$, the vector

$$\overline{y} = \frac{y_*}{\tilde{\rho}}$$

is feasible for problem (B.3). Therefore using (B.6),

$$QCQ \geq \overline{y}^T R \overline{y} = \frac{1}{\tilde{\rho}^2} y_*^T R y_* = \frac{SDP}{\tilde{\rho}^2},$$

and hence the proof is complete.