A SURVEY OF THE S-LEMMA

IMRE PÓLIK* AND TAMÁS TERLAKY†

Abstract. In this survey we review the many faces of the S-lemma, a result about the correctness of the S-procedure. The basic idea of this widely used method came from control theory but it has important consequences in quadratic and semidefinite optimization, convex geometry and linear algebra as well. These were active research areas, but as there was little interaction between researchers in different areas, their results remained mainly isolated. Here we give a unified analysis of the theory by providing three different proofs for the S-lemma and revealing hidden connections with various areas of mathematics. We prove some new duality results and present applications from control theory, error estimation and computational geometry.

Key words. S-lemma, S-procedure, control theory, nonconvex theorem of alternatives, numerical range, relaxation theory, semidefinite optimization, generalized convexities

AMS subject classifications. 90C20, 90C22, 90C26, 49M20, 34D20, 26B25

DOI.

This review is divided into two parts. The first part gives a general overview of the S-lemma. It starts from the basics, provides three different proofs of the fundamental result, discusses possible extensions and presents some counterexamples. Some illustrative applications from control theory and error estimation are also discussed. This part is written in a way that enables the majority of the SIAM community—including those who are not experts on this topic—to understand the concepts and proofs.

The second part, starting with §5, shows how the basic theory is related to various fields of mathematics: functional analysis, rank-constrained optimization and generalized convexities. This part goes beyond the proofs and demonstrates how the same result was discovered several times throughout the 60-year history of the S-procedure. New duality results and further directions are also presented.

PART I

1. Introduction. In this section we expose the basic question of the S-lemma and provide some historical background.

1.1. Motivation. The fundamental question of the theory of the S-lemma is the following:

When is a quadratic (in)equality a consequence of other quadratic (in)equalities?

If we ask the same question for linear or general convex (in)equalities then the Farkas Lemma ([11], Theorem 1.3.1.) and the Farkas Theorem (Theorem 2.1 in this review, or [72], §6.10) give the answer, respectively. These results essentially state that a concave inequality is a (logical) consequence of some convex inequalities if and only if it is a nonnegative linear combination of those convex inequalities and an identically true inequality. This is important, since it is relatively easy to check if an inequality is a linear combination of some other inequalities.

*McMaster University, Advanced Optimization Lab, Department of Mathematics and Statistics, 1280 Main St W, Hamilton (ON), L8S 4K1, Canada (imre.polik@gmail.com)

†McMaster University, Advanced Optimization Lab, Department of Computing and Software, 1280 Main St W, Hamilton (ON), L8S 4K1, Canada (terlaky@mcmaster.ca)

However, this notion is not guaranteed to work in a general setting, as it is demonstrated by the following example, taken from [11]. Consider the inequalities

$$u^2 \geq 1 \tag{1.1a}$$
$$v^2 \geq 1 \tag{1.1b}$$
$$u \geq 0 \tag{1.1c}$$
$$v \geq 0 \tag{1.1d}$$

then the inequality

$$uv \geq 1 \tag{1.1e}$$

is clearly a consequence of those four inequalities. However, taking those four and the trivially true inequalities (such as $0 > -1$, $u^2 \pm 2uv + v^2 \geq 0$, etc.) we can not combine them in a linear way to obtain this consequence. Thus only a fraction of the logical consequences can be reached using linear combinations. In this survey we discuss when we can apply this approach to quadratic systems. As it is shown by the above simple example, it does not work in general, but there are some special cases when we can answer our question using various methods.

Quadratic inequalities are formed naturally in many areas of theoretical and applied mathematics. Consider the following examples.

Quadratic intersections: The following problem arises in computer graphics, see [48] for a standard reference. Let us take two quadratic surfaces represented by the equations $x^T A_i x + b_i^T x + c_i = 0$, $i = 1, 2$. How can we decide whether the two surfaces intersect without actually computing the intersections? Can we compute the intersection efficiently? How can we do the same for the solid bodies bounded by the surfaces?

Noise estimation: Level sets for the density function of measurements in $\mathbb{R}^n$ with a Gaussian noise are ellipsoids. Given two such measurements each with a fixed noise level, we need to find a bound on the noise of their sum. In other words, we are looking for the smallest ellipsoid that contains the sum of the two ellipsoids.

Trust Region Problem: A broad range of functions can be efficiently approximated locally with quadratic functions. The Trust Region Problem is a quadratic optimization problem with a single quadratic constraint, i.e.,

$$\min\ x^T A x + b^T x + c \tag{1.2a}$$
$$\|x - \hat{x}\|_2 \leq \alpha, \tag{1.2b}$$

where $\alpha, c \in \mathbb{R}$, $b, x, \hat{x} \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$. This problem is easily solvable, methods based on these approximations are widely applied in nonlinear optimization, for more details, see [18].

Quadratically Constrained Quadratic Minimization: Finally, we can replace the norm constraint (1.2b) with a set of general quadratic constraints:

$$\min\ x^T A x + b^T x + c \tag{1.3a}$$
$$x^T A^i x + {b^i}^T x + c^i \leq 0,\ i = 1, \ldots, m. \tag{1.3b}$$

Besides the solvability of this problem it is often necessary to decide whether the problem is feasible (i.e., whether system (1.3b) is solvable). If the problem is not feasible then one might require a certificate that proves infeasibility.

Since this last problem includes integer programming we can not hope for very general results.

The S-procedure is a frequently used technique originally arising from the stability analysis of nonlinear systems. Despite being used in practice quite frequently, its theoretical background is not widely understood. On the other hand, the concept of the S-procedure is well-known in the optimization community, although not under this name.

In short terms, the S-procedure is a relaxation method; it tries to solve a system of quadratic inequalities via a Linear Matrix Inequality (LMI) relaxation. Yakubovich [78] was the first to give sufficient conditions on when this relaxation is exact, i.e., when it is possible to obtain a solution for the original system using the solution of the LMI; this result is the S-lemma. The advantage we gain is the computational cost: while solving a general quadratic system can take an exponential amount of work, LMIs can be solved more efficiently.

1.2. Historical background. The earliest result of this kind is due to Finsler [26], which was later generalized by Hestenes and McShane [35]. In 1937 Finsler proved that if A and B are two symmetric matrices and $x^TBx = 0$ $(x \neq 0)$ implies $x^TAx > 0$ then there exists an $y \in \mathbb{R}$ such that $A + yB$ is positive definite.

On the practical side, this idea was first used probably by Lure'e and Postnikov [52] in the 1940s, but at that time there was no well-founded theory behind the method. The theoretical background was developed some 30 years later by Yakubovich: in the early 70's he proved a theorem known as the S-lemma [78, 80] using an old theoretical result of Dines [22] on the convexity of homogeneous quadratic mappings. The simplicity of the method allowed rapid advances in control theory. Later, Megretsky and Treil [54] extended the results to infinite dimensional spaces giving rise to more general applications. Articles written since then mainly discuss some new applications, not new extensions to the theory.

Yakubovich himself presented some applications [79], which were followed by many others [15], including contemporary ones [29, 50], spanning over a broad range of engineering, financial mathematics and abstract dynamical systems. We will discuss various applications in §4.

Although the result emerged mainly from practice, Yakubovich himself was aware of the theoretical implications [27] of the S-lemma. The theoretical line was then continued by others (see e.g., [15], or recently, [20, 21, 51, 73] but apart from a few exceptions such as [11, 15, 50] or [43] the results did not reach the control community. Moreover, to our best knowledge no thorough study presenting all these approaches has been written so far. The collection of lecture notes by Ben-Tal and Nemirovski [11] contains some of the ideas explained here, and several aspects of this theory have been presented in the context of the LMI relaxation of systems of quadratic inequalities.

The term *S-method* was coined by Aizerman and Gantmacher in their book [1], but later it changed to *S-procedure*. The S-method tries to decide the stability of a system of linear differential equations by constructing a Lyapunov matrix. During the process an auxiliary matrix S (for stability) is introduced. This construction leads to a system of quadratic equations (the Lure'e resolving equations, [52]). If that quadratic system can be solved then a suitable Lyapunov function can be constructed. The term *S-lemma* refers to results stating that such a system can be solved under some conditions; the first such result is due to Yakubovich [78]. In this survey our main interest is the S-lemma, but we will present an example from control theory in §4.

1.3. About this survey. In this paper we show how the S-lemma relates to well known concepts in optimization, relaxation methods and functional analysis. This survey is structured as follows. In §2 we give three independent proofs for the S-lemma, illustrating how the original result is connected to more general theories. In §3 we show some examples and counterexamples, and present other variants of the S-lemma. Applications from control theory and computational geometry are discussed in §4.

In the second part we go deeper and investigate three major topics. First, in §5 we discuss a classical topic, the convexity of the numerical range of a linear operator. Following that, in §6 we present a seemingly different field, rank-constrained optimization. In §6.5 we merge the results of these two fields and show the equivalence of the theories. Finally, in §7 we put the problem in a more general context and show that the S-lemma is a special case of a duality theorem due to Illés and Kassay [40, 41, 42].

Some miscellaneous topics are then discussed in §8. These topics (trust region problems, algebraic geometric connections and complexity issues) can not be presented in full detail due to lack of space. Instead, we briefly summarize their connection to the S-lemma and indicate directions for further research.

Possible future directions and some open questions are discussed in §9.

We made every effort to make the major sections (§5-7) self-contained, i.e., any one of them can be skipped depending on the reader's area of interest. Each of them starts with a motivation part where we describe how the S-lemma is related to the selected area, then we summarize the major theoretical results of the field, and finally, we apply the theory to the problem and conclude the results.

1.4. Notations. Matrices are denoted by capital letters $(A, B, \ldots)$, the j^{th} element of the i^{th} row is A_{ij}. Vectors are denoted by lowercase Roman letters $(u, v, \ldots)$, the i^{th} component of a vector is of the form u_i, while u^i is the i^{th} vector in a list. Vectors are considered to be column vectors, implying that $u^T v$ is the scalar or dot product while uv^T is the so-called outer product of two vectors. Sometimes we will break this convention for typographic reasons, e.g., we will simply use $(1, 4)$ to denote a column vector, and accordingly, $(1, 4)^T(2, 3)$ will denote the scalar product of two such vectors. We try to avoid any ambiguities and use the notations in a clear and concise way. Borrowing the usual Matlab notation we will use $u_{2:n}$ to denote the vector $(u_2, \ldots, u_n)^T$ and $A_{:,1}$ to denote the full first column of A.

Matrices in this review are usually symmetric and positive (semi)definite. Let $\mathbb{S}^n$ be the space of $n \times n$ real symmetric matrices and $\mathbb{PS}^n \subseteq \mathbb{S}^n$ the convex cone of positive semidefinite matrices. In the following, $A \succ 0$ will denote that A is symmetric and positive definite, while $A \succeq 0$ will denote that A is symmetric and positive semidefinite. The notations $A \succ B$ and $A \succeq B$ are interpreted as $A - B \succ 0$ and $A - B \succeq 0$, respectively. Considering the $n \times n$ matrices as vectors in $\mathbb{R}^{n \times n}$ we can use the scalar product for vectors, i.e., the sum of the products of the corresponding elements. Denoting this scalar product of two matrices A and B with $A \bullet B$ we have the following properties, see, e.g., [39]:

1. $A \bullet B = \operatorname{Tr}(AB) = \operatorname{Tr}(BA)$.
2. If $B = bb^T$ is a rank-1 matrix then $A \bullet B = b^T Ab$.
3. If A and B are positive semidefinite matrices then $A \bullet B \geq 0$.

We use the standard symbols $\mathbb{R}^n$ and $\mathbb{C}^n$ to denote the n-dimensional linear space of real and complex vectors, respectively. In addition, we use $\mathbb{R}^n_+$ to denote the set of n-dimensional vectors with nonnegative coordinates, i.e., the nonnegative orthant. The symbol $\mathbb{R}^n_-$ is interpreted accordingly. To avoid confusion with indices, we will

use $\mathbf{i}$ to denote the imaginary unit, i.e., $\mathbf{i}^2 = -1$.

2. Proofs for the basic S-lemma. In this section we present three proofs for the basic S-lemma. We start with the original proof of Yakubovich, then we present a modern proof based on LMIs, and we conclude with an elementary, analytic proof. The key concepts of these proofs will be further investigated and generalized in the remaining sections.

2.1. The two faces of the S-lemma. Now we present the central result of the theory starting from the very basics and showing the main directions of the rest of the survey. The theorems we are going to discuss can be viewed in two ways. As it is illustrated by the example in §4, the original application of the S-lemma is to decide whether a quadratic (in)equality is satisfied over a domain. As these domains are usually defined by quadratic inequalities, the question we are investigating is when a quadratic inequality is a consequence of other quadratic inequalities. This idea can be formalized as follows:

$$g_j(x) \le 0,\ j = 1, \dots, m \stackrel{?}{\Rightarrow} f(x) \ge 0, \tag{2.1}$$

where $x \in \mathbb{R}^n$ and $f, g_j : \mathbb{R}^n \to \mathbb{R}, j = 1, \dots, m$ are quadratic functions.

The alternative approach views the question as a feasibility problem: is there any point in the domain where the inequality in question does not hold? This problem is of the same form as the Farkas Theorem, a fundamental theorem of alternatives in convex analysis, see e.g., [68], Theorem 21.1., [72], §6.10, or [19]:

THEOREM 2.1 (Farkas Theorem). *Let $f, g_1, \dots, g_m : \mathbb{R}^n \to \mathbb{R}$ be convex functions, $\mathcal{C} \subseteq \mathbb{R}^n$ a convex set and let us assume that the Slater condition holds for $g_1, \dots, g_m$, i.e., there exists an $\bar{x} \in \operatorname{rel\,int} \mathcal{C}$ such that $g_j(\bar{x}) < 0$, $j = 1, \dots, m$. The following two statements are equivalent:*

(i) *The system*

$$f(x) < 0 \tag{2.2a}$$
$$g_j(x) \le 0,\ j = 1, \dots, m \tag{2.2b}$$
$$x \in \mathcal{C} \tag{2.2c}$$

is not solvable.

(ii) *There are $y_1, \dots, y_m \ge 0$ such that*

$$f(x) + \sum_{i=1}^{m} y_j g_j(x) \ge 0, \tag{2.3}$$

for all $x \in \mathcal{C}$.

The proof is based on a separation argument using the fact that the set

$$\left\{(u, v_1, \dots, v_m) \in \mathbb{R}^{m+1} : \exists x \in \mathbb{R}^n,\ f(x) < u,\ g_i(x) \le v_i,\ i = 1, \dots, m\right\} \tag{2.4}$$

is convex and that system (2.2) is not solvable if and only of the origin is not in this convex set. The convexity of this set is trivial if all the functions are convex, but in other cases it typically fails to hold.

When writing this review we had to decide which form to use. Since in our opinion the latter one is easier to write and more popular in the optimization community we will present all the results in the form of the Farkas Theorem.

The theorem we present here was first proved by Yakubovich [78, 80] in 1971.

THEOREM 2.2 (S-lemma, Yakubovich, 1971). *Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be quadratic functions and suppose that there is an $\bar{x} \in \mathbb{R}^n$ such that $g(\bar{x}) < 0$. Then the following two statements are equivalent.*

(i) *There is no $x \in \mathbb{R}^n$ such that*

$$f(x) < 0 \tag{2.5a}$$

$$g(x) \leq 0. \tag{2.5b}$$

(ii) *There is a non-negative number $y \geq 0$ such that*

$$f(x) + yg(x) \geq 0, \forall x \in \mathbb{R}^n. \tag{2.6}$$

2.2. The traditional approach. Yakubovich used the following convexity result to prove the S-lemma:

PROPOSITION 2.3 (Dines, 1941, [22]). *If $f, g : \mathbb{R}^n \to \mathbb{R}$ are* homogeneous *quadratic functions then the set $\mathcal{M} = \{(f(x), g(x)) : x \in \mathbb{R}^n\} \subset \mathbb{R}^2$ is convex.*

Proof. We will verify the definition of convexity directly. Let us take two points, $u = (u_f, u_g)$ and $v = (v_f, v_g)$. If these two points and the origin are collinear then obviously the line segment between u and v belongs to $\mathcal{M}$, since the functions are homogeneous. From now on we will assume that these points are not collinear with the origin. Since they belong to $\mathcal{M}$ there are points $x_u, x_v \in \mathbb{R}^n$ such that

$$u_f = f(x_u),\ u_g = g(x_u) \tag{2.7a}$$

$$v_f = f(x_v),\ v_g = g(x_v). \tag{2.7b}$$

We will further assume without loss of generality that

$$v_f u_g - u_f v_g = d^2 > 0. \tag{2.8}$$

Let $\lambda \in (0, 1)$ be a constant. We try to show that there exists an $x_\lambda \in \mathbb{R}^n$ such that

$$(f(x_\lambda), g(x_\lambda)) = (1 - \lambda)u + \lambda v. \tag{2.9}$$

Let us look for x_λ in the form[1]

$$x_\lambda = \rho(x_u \cos\theta + x_v \sin\theta), \tag{2.10}$$

where ρ and θ are real variables. Substituting these to the defining equation of x_λ we get:

$$\rho^2 f(x_u \cos\theta + x_v \sin\theta) = (1 - \lambda)u_f + \lambda v_f \tag{2.11a}$$

$$\rho^2 g(x_u \cos\theta + x_v \sin\theta) = (1 - \lambda)u_g + \lambda v_g. \tag{2.11b}$$

Eliminating ρ^2 from these equations and expressing λ as a function of θ we get

$$\lambda(\theta) = \frac{u_g f(x_u \cos\theta + x_v \sin\theta) - u_f g(x_u \cos\theta + x_v \sin\theta)}{(u_g - v_g) f(x_u \cos\theta + x_v \sin\theta) - (u_f - v_f) g(x_u \cos\theta + x_v \sin\theta)}. \tag{2.12}$$

[1] This is simply a clever parameterization of the plane spanned by x_u and x_v. If we used the form $x_\lambda = px_u + qx_v$, where $p, q \in \mathbb{R}$, then we could reduce the problem to a 2×2 convexity problem, which can be solved by an elementary but tedious analysis. This is the idea of the proof of the S-lemma in [11].

Here the denominator of $\lambda(\theta)$ is a quadratic function of $\cos\theta$ and $\sin\theta$, let us denote it by $T(\cos\theta,\sin\theta)=\alpha\cos^2\theta+\beta\sin^2\theta+2\gamma\cos\theta\sin\theta$. Computing $T(0)$, $T(\pm\pi/2)$, and using (2.7) we get that $\alpha=\beta=d^2>0$, thus $T(\cos\theta,\sin\theta)=d^2+\gamma\sin(2\theta)$. If $\gamma\geq 0$ then $T(\cos\theta,\sin\theta)>0$ for $\theta\in[0,\pi/2]$, and similarly, if $\gamma\leq 0$ then $T(\cos\theta,\sin\theta)>0$ for $\theta\in[-\pi/2,0]$. We assume without loss of generality that it is the former, then $\lambda(\theta)$ is defined on the whole interval $[0,\frac{\pi}{2}]$ and it is continuous as well. Since $\lambda(0)=0$ and $\lambda(\frac{\pi}{2})=1$, we can find a value $\theta_\lambda\in[0,\frac{\pi}{2}]$ such that $\lambda(\theta_\lambda)=\lambda$. Using this θ_λ we get ρ from (2.11) and the desired vector x_λ from (2.10). This completes the proof. □

Yakubovich used this result to prove Theorem 2.2.

Proof. (Yakubovich, 1971, [78]) It is obvious that (ii) implies (i). On the other hand let us assume (i) and try to prove (ii). First let f and g be homogeneous functions, then by Proposition 2.3 the 2D image of $\mathbb{R}^n$ under the mapping (f,g) is convex, and by (i) this image does not intersect the convex cone $\mathcal{C}=\{(u_1,u_2):u_1<0,u_2\leq 0\}\subset\mathbb{R}^2$, thus they can be separated by a line. This means that there are real numbers y_1 and y_2 such that

$$y_1u_1+y_2u_2\leq 0,\ \forall(u_1,u_2)\in\mathcal{C} \tag{2.13a}$$

$$y_1f(x)+y_2g(x)\geq 0,\ \forall x\in\mathbb{R}^n. \tag{2.13b}$$

Taking $(-1,0)\in\mathcal{C}$ we have $y_1\geq 0$ and setting $(-\varepsilon,-1)\in\mathcal{C}$ where ε is arbitrarily small gives $y_2\geq 0$. The case $y_1=0$ can be ruled out by substituting $\bar{x}$ in the second equation, so we have $y_1>0$. Letting $y=y_2/y_1\geq 0$ then satisfies (ii).

Now let f and g be general, not necessarily homogeneous quadratic functions satisfying (i). First let us notice that we can assume $\bar{x}=0$, if this is not the case then let $\bar{g}(x)=g(x+\bar{x})$ be our new function. Let the functions be defined as

$$f(x)=x^TA_fx+b_f^Tx+c_f, \tag{2.14a}$$

$$g(x)=x^TA_gx+b_g^Tx+c_g, \tag{2.14b}$$

then the Slater condition is equivalent to $g(0)=c_g<0$. Let us introduce the homogeneous version of our functions

$$\tilde{f}:\mathbb{R}^{n+1}\to\mathbb{R},\ \tilde{f}(x,\tau)=x^TA_fx+\tau b_f^Tx+\tau^2c_f \tag{2.15a}$$

$$\tilde{g}:\mathbb{R}^{n+1}\to\mathbb{R},\ \tilde{g}(x,\tau)=x^TA_gx+\tau b_g^Tx+\tau^2c_g. \tag{2.15b}$$

Now we prove that the new functions satisfy (i), i.e., there is no $(x,\tau)\in\mathbb{R}^{n+1}$ such that

$$\tilde{f}(x,\tau)<0 \tag{2.16a}$$

$$\tilde{g}(x,\tau)\leq 0. \tag{2.16b}$$

Let us assume on the contrary that there is an $(x,\tau)\in\mathbb{R}^{n+1}$ with these properties. If $\tau\neq 0$ then

$$f(x/\tau)=\tilde{f}(x,\tau)/\tau^2<0, \tag{2.17a}$$

$$g(x/\tau)=\tilde{g}(x,\tau)/\tau^2\leq 0, \tag{2.17b}$$

which contradicts to (i). If $\tau=0$ then $x^TA_fx<0$ and $x^TA_gx\leq 0$, therefore

$$\underbrace{(\lambda x)^TA_f(\lambda x)}_{<0}+\lambda b_f^Tx+c_f<0, \text{ if } |\lambda| \text{ is large enough, and} \tag{2.18a}$$

$$\underbrace{(\lambda x)^TA_g(\lambda x)}_{\leq 0}+\lambda b_g^Tx+\underbrace{c_g}_{<0}<0, \text{ if } \lambda \text{ has the proper sign,} \tag{2.18b}$$

contradicting to (i). This implies that the new system (2.16) is not solvable. Further, taking $(0, 1)$ gives

$$\tilde{g}(0,1) = g(0) < 0, \tag{2.19}$$

therefore the Slater condition is satisfied so we can apply the already proved homogeneous version of the theorem. We get that there exists a $y \geq 0$ such that

$$\tilde{f}(x,\tau) + y\tilde{g}(x,\tau) \geq 0,\ \forall (x,\tau) \in \mathbb{R}^{n+1}, \tag{2.20}$$

and with $\tau = 1$ we get (ii). □

Problems similar to Proposition 2.3 were first investigated by Hausdorff [34] and Toeplitz [74] in the late 1910's in a more general context: the joint numerical range of Hermitian operators. The importance of this simple fact becomes more obvious if we recall that the S-lemma is actually a non-convex theorem of alternatives, an extended version of the Farkas Theorem.

2.3. A modern approach. This proof is similar to the one found in [11], but extends it for the nonhomogeneous case.

The following lemma from [73] plays a crucial role in this theory:

LEMMA 2.4. *Let $G, X \in \mathbb{R}^{n\times n}$ be symmetric matrices X being positive semidefinite and rank r. Then $G\bullet X \leq 0$ ($G\bullet X = 0$) if and only if there are $p^1, \ldots, p^r \in \mathbb{R}^n$ such that*

$$X = \sum_{i=1}^{r} p^i {p^i}^T \ \text{ and } \ G \bullet p^i {p^i}^T = {p^i}^T G p^i \leq 0\ ({p^i}^T G p^i = 0),\ \forall i = 1, \ldots, r. \tag{2.21}$$

Proof. We only prove the first version of the lemma. The proof is based on [73] and is constructive. Consider the following procedure:

Input: X and $G \in \mathbb{R}^{n\times n}$ such that $X \succeq 0$ and $G \bullet X \leq 0$, $\operatorname{rank}(X) = r$.

Step 1: Compute the rank-1 decomposition of X, i.e.,

$$X = \sum_{i=1}^{r} p^i {p^i}^T. \tag{2.22}$$

Step 2: If $({p^1}^T G p^1)({p^i}^T G p^i) \geq 0$ for all $i = 2, \ldots, r$ then return $y = p^1$. Otherwise we have a j such that $({p^1}^T G p^1)({p^j}^T G p^j) < 0$.

Step 3: Since ${p^1}^T G p^1$ and ${p^j}^T G p^j$ have opposite sign we must have an $\alpha \in \mathbb{R}$ such that

$$(p^1 + \alpha p^j)^T G (p^1 + \alpha p^j) = 0. \tag{2.23}$$

In this case we return

$$y = \frac{p^i + \alpha p^j}{\sqrt{1+\alpha^2}}. \tag{2.24}$$

Output: $y \in \mathbb{R}^n$ such that $0 \geq y^T G y \geq G \bullet X$, $X - yy^T$ is positive semidefinite and $\operatorname{rank}(X - yy^T) = r - 1$.

If the procedure stops in Step 2 then ${p^i}^T G p^i$ has the same sign for all $i = 1, \ldots, r$. Since the sum of these terms is negative we have $y^T G y = {p^1}^T G p^1 \leq 0$ and $X - yy^T = \sum_{i=2}^{r} p^i {p^i}^T$ implying that the remaining matrix is positive semidefinite and has rank $r-1$.

If the procedure does not stop in Step 2 then the quadratic equation for α must have two distinct roots, and by definition $0 = y^T G y \geq G \bullet X$. Finally, we can see that

$$X - yy^T = uu^T + \sum_{i \in \{2,3,\ldots,r\} \setminus j} p^i {p^i}^T, \tag{2.25a}$$

where

$$u = \frac{p^j - \alpha p^1}{\sqrt{1+\alpha^2}}. \tag{2.25b}$$

This decomposition ensures that $X - yy^T$ has rank $r-1$ and the procedure is correct. Applying the procedure r times we get the statement of the lemma. □

REMARK 2.5. *The lemma can also be proved using the rank-1 approach presented in the proof of Theorem 6.1.*

Now we can finish the proof of Theorem 2.2.

Proof. (Theorem 2.2, S-lemma) It is obvious that if either of the two systems has a solution then the other one can not have one, so what we have to prove is that at least one system has a solution. Let the functions be given as

$$f(x) = x^T A_f x + b_f^T x + c_f, \tag{2.26a}$$

$$g(x) = x^T A_g x + b_g^T x + c_g, \tag{2.26b}$$

and let us consider the following notation:

$$H_f = \begin{bmatrix} c_f & \frac{1}{2} b_f^T \\ \frac{1}{2} b_f & A_f \end{bmatrix}, \quad H_g = \begin{bmatrix} c_g & \frac{1}{2} b_g^T \\ \frac{1}{2} b_g & A_g \end{bmatrix}. \tag{2.27}$$

Using this notation the first system can be rewritten as:

$$H_f \bullet \begin{bmatrix} 1 & x^T \\ x & xx^T \end{bmatrix} < 0, \quad H_g \bullet \begin{bmatrix} 1 & x^T \\ x & xx^T \end{bmatrix} \leq 0, \quad x \in \mathbb{R}^n. \tag{2.28}$$

Here the rank-1 matrices

$$\begin{bmatrix} 1 & x^T \\ x & xx^T \end{bmatrix} \tag{2.29}$$

are positive semidefinite and symmetric. This can inspire us to look at the following relaxation of (2.28):

$$H_f \bullet Z < 0, \quad H_g \bullet Z \leq 0, \quad Z \succeq 0. \tag{2.30}$$

The key idea of our proof is to show that this relaxation is exact in the sense that problem (2.28) is solvable if and only if the relaxed problem (2.30) is solvable. More specifically, we will prove the following lemma.

LEMMA 2.6. *Using the notations introduced in (2.27)-(2.30), if the relaxed system (2.30) has a solution then it has a rank-1 solution of the form $Z = zz^T$, where the first*

coordinate of z is 1. This gives a solution for (2.28). Moreover, (2.30) has a solution that strictly satisfies all the inequalities including the semidefinite constraint.

Proof. Let Z be a solution of (2.30). Then, since Z is positive semidefinite it can be written as

$$Z = \sum_{j=1}^{r} q^j {q^j}^T, \tag{2.31}$$

where $q^j \in \mathbb{R}^{n+1}$ and $r = \operatorname{rank}(Z)$. Applying Lemma 2.4 we see that it is possible to choose these vectors such that

$$H_g \bullet q^j {q^j}^T = {q^j}^T H_g q^j \leq 0, \quad j = 1, \ldots, r. \tag{2.32}$$

Now, from the strict inequality of (2.30) we can conclude that there is a vector $q = q^j$ for some $1 \leq j \leq r$ such that

$$H_f \bullet q q^T < 0, \tag{2.33}$$

otherwise the sum of these terms could not be negative. It means that $Z = qq^T$ is a rank-1 solution of (2.30). Observe that this result was obtained without using the Slater condition.

If the first coordinate of q is nonzero then $x = q_{2:n+1}/q_1$ gives a solution for (2.28). If this is not the case then let us introduce

$$\tilde{q} = q + \alpha \begin{bmatrix} 1 \\ \bar{x} \end{bmatrix}, \tag{2.34}$$

where $\bar{x}$ is the point satisfying the Slater condition. Notice that

$$H_f \bullet \tilde{q}\tilde{q}^T = \underbrace{H_f \bullet q q^T}_{<0} + 2\alpha H_f \bullet \begin{bmatrix} 1 \\ \bar{x} \end{bmatrix} q^T + \alpha^2 H_f \bullet \begin{bmatrix} 1 & \bar{x}^T \\ \bar{x} & \bar{x}\bar{x}^T \end{bmatrix} < 0 \tag{2.35a}$$

if $|\alpha|$ is small and

$$H_g \bullet \tilde{q}\tilde{q}^T = \underbrace{H_g \bullet q q^T}_{\leq 0} + 2\alpha H_g \bullet \begin{bmatrix} 1 \\ \bar{x} \end{bmatrix} q^T + \alpha^2 \underbrace{H_g \bullet \begin{bmatrix} 1 & \bar{x}^T \\ \bar{x} & \bar{x}\bar{x}^T \end{bmatrix}}_{=g(\bar{x})<0} < 0 \tag{2.35b}$$

if the sign of α is chosen to make the middle term negative. Further, if $\alpha \neq -q_1$ then $\tilde{q}_1 \neq 0$. It is obvious that these conditions can be satisfied simultaneously, i.e., we have $\tilde{q}\tilde{q}^T$ that solves (2.30) and $x = \tilde{q}_{2:n+1}/\tilde{q}_1$ gives a solution for (2.28). Finally, letting

$$\widetilde{Z} = \tilde{q}\tilde{q}^T + \beta I, \tag{2.36}$$

where $\beta \in \mathbb{R}$ and $I \in \mathbb{R}^{(n+1)\times(n+1)}$ is the identity matrix, provides $H_f \bullet \widetilde{Z}$ and $H_g \bullet \widetilde{Z} < 0$ if $|\beta|$ is small enough and $\widetilde{Z} \succ 0$. In other words $\widetilde{Z}$ satisfies the strict version of all the inequalities. This completes the proof of the lemma. □

Now we can easily finish the proof of Theorem 2.2. It follows directly from the Farkas Theorem (see Theorem 2.1) that system (2.30) is solvable if and only if the dual system

$$H_f + y H_g \succeq 0 \tag{2.37a}$$

$$y \geq 0. \tag{2.37b}$$

is not solvable. Now, by Lemma 2.6, the solvability of the original quadratic system (2.28) is equivalent to the solvability of its LMI relaxation (2.30), which—by duality—is equivalent to the non-solvability of the dual system (2.37). This means that there is a $y \geq 0$ such that $f(x) + yg(x) \geq 0$ for all $x \in \mathbb{R}^n$. This completes the proof of the S-lemma. □

Quadratic systems can always be relaxed using linear matrix inequalities, so the key question is when this relaxation is exact. This topic is further discussed in §6.

2.4. An elementary proof. This proof is based on Lemma 2.3 in [82], and it is the most elementary one of the proofs presented in this section. We only prove the homogeneous version, the nonhomogeneous case can be handled similarly. We will use the following lemma:

LEMMA 2.7 (Yuan, [82], 1990). *Let $A, B \in \mathbb{R}^{n\times n}$ be two symmetric matrices and let $\mathcal{F}, \mathcal{G} \subseteq \mathbb{R}^n$ be closed sets such that $\mathcal{F} \cup \mathcal{G} = \mathbb{R}^n$. If*

$$x^T Ax \geq 0, \quad \forall x \in \mathcal{F} \tag{2.38a}$$

$$x^T Bx \geq 0, \quad \forall x \in \mathcal{G} \tag{2.38b}$$

then there is a $t \in [0, 1]$ such that $tA + (1 - t)B$ is positive semidefinite.

Proof. The lemma is trivially true if either of the two sets is empty, so we can assume that both are non-empty. Further, we can assume that both sets are symmetric about the origin, i.e., $\mathcal{F} = -\mathcal{F}$ and $\mathcal{G} = -\mathcal{G}$. Let $\lambda(t)$ be the smallest eigenvalue of $tA + (1 - t)B$. If $\lambda(t) \geq 0$ for some $t \in [0, 1]$ then the lemma is true. Let us assume now that $\lambda(t) < 0$ for all $t \in [0, 1]$. We define the following set:

$$S(t) = \{x : (tA + (1 - t)B)x = \lambda(t)x, \|x\| = 1\}. \tag{2.39}$$

Since $\lambda(t)$ is an eigenvalue, $S(t)$ is not empty and it is closed by continuity, so

$$S(t) \supseteq \left\{x : x = \lim_{k\to\infty} x_k, x_k \in S(t_k), t = \lim_{k\to\infty} t_k\right\}. \tag{2.40}$$

If $x \in S(0)$, then $x^T Bx = \lambda(0) < 0$, thus $x \notin \mathcal{G}$, so $x \in \mathcal{F}$. This shows that $S(0) \subseteq \mathcal{F}$, thus $S(0) \cap \mathcal{F} = S(0) \neq \emptyset$. Let $t_{\max}$ be the largest number in $[0, 1]$ such that $S(t_{\max}) \cap \mathcal{F} \neq \emptyset$, this number exists due to (2.40). If $t_{\max} = 1$ then by the assumptions on $\mathcal{G}$, we have $S(1) \cap \mathcal{G} = S(1) \neq \emptyset$. If $t_{\max} < 1$ then for every $t \in (t_{\max}, 1]$ we have

$$S(t) \cap \mathcal{G} = (S(t) \cap \mathcal{G}) \cup \underbrace{(S(t) \cap \mathcal{F})}_{=\emptyset} = S(t) \neq \emptyset, \tag{2.41}$$

where we used the assumption that $\mathcal{F} \cup \mathcal{G} = \mathbb{R}^n$. Again, using (2.40) we get that

$$S(t_{\max}) \cap \mathcal{G} \neq \emptyset. \tag{2.42}$$

Since $S(t)$ is the intersection of a subspace (the eigenspace of $\lambda(t)$) and the unit ball, it is a unit ball in some dimension, therefore $S(t)$ is either connected, or it is the union of two points, symmetric about the origin.

If $S(t_{\max})$ is a connected ball, then any path connecting a point in $S(t_{\max}) \cap \mathcal{F}$ with a point in $S(t_{\max}) \cap \mathcal{G}$ contains a point in $S(t_{\max}) \cap \mathcal{F} \cap \mathcal{G}$, since both $\mathcal{F}$ and $\mathcal{G}$ are closed. This shows that $S(t_{\max}) \cap \mathcal{F} \cap \mathcal{G} \neq \emptyset$, thus there exists an $x \in \mathcal{F} \cap \mathcal{G}$

such that $x^T\left(tA+(1-t)B\right)x = \lambda(t) < 0$, but then either $x^TAx < 0$ or $x^TBx < 0$, contradicting to (2.38).

If on the other hand $S(t_{\max})$ consists of two points then since $\mathcal{F} = -\mathcal{F}$ and $\mathcal{G} = -\mathcal{G}$, we have $S(t_{\max}) \subseteq \mathcal{F}$ and $S(t_{\max}) \subseteq \mathcal{G}$, thus we reach the same conclusion. This completes the proof of Lemma 2.7. □

Now the S-lemma (Theorem 2.2) can be proved easily. Let A and B be symmetric matrices and assume that the system

$$x^TAx < 0 \tag{2.43a}$$
$$x^TBx \leq 0 \tag{2.43b}$$

is not solvable, but the Slater condition is satisfied, i.e., $\exists \bar{x} \in \mathbb{R}^n : \bar{x}^TB\bar{x} < 0$. Defining the closed sets $\mathcal{F} = \left\{x : x^TBx \leq 0\right\}$ and $\mathcal{G} = \left\{x : x^TBx \geq 0\right\}$ one has $\mathcal{F} \cup \mathcal{G} = \mathbb{R}^n$. By the assumption of nonsolvability we have that $x^TAx \geq 0 : \forall x \in \mathcal{F}$ and $x^TBx \geq 0 : \forall x \in \mathcal{G}$, thus all the conditions of Lemma 2.7 are satisfied and we can conclude that there is a $t \in [0,1]$ such that $tA+(1-t)B$ is positive semidefinite. Now t can not be 0, otherwise B would be positive semidefinite and the Slater condition could not be satisfied. Dividing by t we get that $A + \frac{1-t}{t}B$ is positive semidefinite.

Remark 2.8. *In the proof of Lemma 2.7 we used little about the quadratic nature of the functions, this gives an incentive to try to extend this lemma for more general functions.*

3. Special results and counterexamples. In this section we present some related results and counterexamples.

3.1. Other variants. For the sake of completeness we enumerate other useful forms of the basic S-lemma. One can get these results by modifying the original proof slightly. For references see [43, 51].

Proposition 3.1 (S-lemma with equality). *Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be quadratic functions where g is assumed to be strictly concave (or strictly convex) and let us assume a stronger form of the Slater condition, namely g takes both positive and negative values. Then the following two statements are equivalent:*

(i) *The system*

$$f(x) < 0 \tag{3.1a}$$
$$g(x) = 0 \tag{3.1b}$$

is not solvable.

(ii) *There exists a multiplier $y \in \mathbb{R}$ such that*

$$f(x) + yg(x) \geq 0, \quad \forall x \in \mathbb{R}^n. \tag{3.2}$$

In the presence of two non-strict inequalities we have the following result.

Proposition 3.2 (S-lemma with non-strict inequalities). *Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be homogeneous quadratic functions. The following two statements are equivalent.*

(i) *The system*

$$f(x) \leq 0 \tag{3.3a}$$
$$g(x) \leq 0 \tag{3.3b}$$

is not solvable.

(ii) *There exist nonnegative multipliers* $y_1, y_2 \geq 0$ *such that*

$$y_1 f(x) + y_2 g(x) > 0, \quad \forall x \in \mathbb{R}^n \setminus \{0\}. \tag{3.4}$$

If we assume the Slater condition for one of the functions then we can make the corresponding multiplier positive.

3.2. General results. In this section we present some known results on how the solvability of the system

$$f(x) < 0 \tag{3.5a}$$

$$g_i(x) \leq 0, \quad i = 1, \ldots, m \tag{3.5b}$$

$$x \in \mathbb{R}^n \tag{3.5c}$$

and the existence of a dual vector $y = (y_1, \ldots, y_m) \geq 0$ such that

$$f(x) + \sum_{i=1}^{m} y_i g_i(x) \geq 0 \quad \forall x \in \mathbb{R}^n \tag{3.6}$$

are related to each other. We will assume that

$$f(x) = x^T A x + b^T x + c \tag{3.7a}$$

$$g_i(x) = x^T B_i x + {p^i}^T x + q^i, \quad i = 1, \ldots, m \tag{3.7b}$$

where A and B_i are symmetric but not necessarily positive semidefinite matrices. Unfortunately, the S-lemma is not true in this general setting. It fails to hold even if we restrict ourselves to $m = 2$ and assume the Slater condition. This does not prevent us from studying the idea of the S-procedure even when the procedure is theoretically not exact. It is trivial that the two systems in the S-lemma can not be solved simultaneously, so if we are lucky enough to find a solution for the second system then we can be sure that the first system is not solvable. However, the non-solvability of the second system does not always guarantee the solvability of the first one.

First let us present what can be found in the literature about this general setting. Our main sources are [11] and [73]. We will outline the proofs as necessary.

Let us start with a general result that contains the S-lemma as a special case.

THEOREM 3.3. *Consider the systems (3.5)-(3.6) and let us assume that the functions* f *and* $g_i, i = 1, \ldots, m$ *are all homogeneous and* $m \leq n$. *If system (3.5) is not solvable then there exist a nonnegative vector* $y = (y_1, \ldots, y_m) \geq 0$ *and an* $n - m + 1$ *dimensional subspace* $V^{n-m+1} \subseteq \mathbb{R}^n$ *such that*

$$f(x) + \sum_{i=1}^{m} y_i g_i(x) \geq 0 \quad \forall x \in V^{n-m+1}. \tag{3.8}$$

In other words, the matrix

$$A + \sum_{i=1}^{m} y_i B_i \tag{3.9}$$

has at least $n - m + 1$ *nonnegative eigenvalues (counted with multiplicities) and the above subspace is spanned by the corresponding eigenvectors.*

REMARK 3.4. *If $m = 1$ this gives the usual S-lemma.*

The proof of this theorem is based on differential geometric arguments, see §4.10 in [11]. Despite the relative strength of the theorem, it is not straightforward to apply it in practice. One such way is to exploit the possible structure of the linear combination and rule out a certain number of negative eigenvalues.

Besides this general result we have only very special ones.

PROPOSITION 3.5. *If A and B_i, $i = 1, \ldots, m$ are all*

- *diagonal matrices, i.e., $f(x)$ and $g_i(x)$ are all weighted sums of squares, or*
- *linear combinations of two fixed matrices, i.e.,* $\operatorname{rank}(A, B_1, \ldots, B_m) \leq 2$,

then the system

$$f(x) < 0 \tag{3.10a}$$

$$g_i(x) \leq 0, \quad i = 1, \ldots, m \tag{3.10b}$$

$$x \in \mathbb{R}^n \tag{3.10c}$$

is not solvable if and only if there exists a nonnegative vector $y = (y_1, \ldots, y_m) \geq 0$ such that

$$f(x) + \sum_{i=1}^{m} y_i g_i(x) \geq 0 \quad \forall\, x \in \mathbb{R}^n. \tag{3.10d}$$

In other words, the matrix

$$A + \sum_{i=1}^{m} y_i B_i \tag{3.10e}$$

is positive semidefinite.

The first part of this proposition can be proved easily using the substitution $z_i = x_i^2$ and applying the Farkas Lemma. The second part is a new result, we will prove it in §5.4, see Theorem 5.24.

If we want to incorporate more quadratic constraints, we need extra conditions.

PROPOSITION 3.6 ($m = 2$, $n \geq 2$). *Let f, g_1 and g_2 be homogeneous quadratic functions and assume that there is an $\bar{x} \in \mathbb{R}^n$ such that $g_1(\bar{x})$, $g_2(\bar{x}) < 0$ (Slater condition). If either*

- *$m = 2$, $n \geq 3$ and there is a positive definite linear combination of A, B_1 and B_2, or*
- *$m = n = 2$ and there is a positive definite linear combination of B_1 and B_2,*

then the following two statements are equivalent:

(i) *The system*

$$f(x) < 0 \tag{3.11a}$$

$$g_i(x) \leq 0, \quad i = 1, 2 \tag{3.11b}$$

$$x \in \mathbb{R}^n \tag{3.11c}$$

is not solvable.

(ii) *There are nonnegative numbers y_1 and y_2 such that*

$$f(x) + y_1 g_1(x) + y_2 g_2(x) \geq 0 \quad \forall x \in \mathbb{R}^n. \tag{3.12}$$

REMARK 3.7. *The condition $n \geq 3$ in the first part is necessary, see the counterexample for $n = 2$ in §3.3.2.*

REMARK 3.8. *An equivalent condition on when some matrices have a positive definite linear combination is given in [23]. If $n \geq 3$ then the property that two symmetric matrices have a positive definite linear combination is equivalent to the nonexistence of a common root of the quadratic forms, see [26]. Further, in this case the matrices are simultaneously diagonalizable by a real congruence, see [5, 61, 76]. In the general case, symmetric matrices $A_1, \ldots, A_m$ have a positive definite linear combination if and only if $A_i \bullet S = 0, i = 1, \ldots, m$ implies that S is indefinite. This result is a trivial corollary of the duality theory of convex optimization, but was only rediscovered around the middle of the 20th century, see [23].*

One additional step is to include linear constraints. First, some equalities:

PROPOSITION 3.9. *Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be quadratic functions, $H \in \mathbb{R}^{m\times n}$ and $r \in \mathbb{R}^m$. Assume that there exists an $\bar{x} \in \mathbb{R}^n$ such that $H\bar{x} = r$ and $g(\bar{x}) < 0$. The following two statements are equivalent:*

(i) *The system*

$$f(x) < 0 \tag{3.13a}$$
$$g(x) \leq 0 \tag{3.13b}$$
$$Hx = r \tag{3.13c}$$
$$x \in \mathbb{R}^n \tag{3.13d}$$

is not solvable.

(ii) *There is a nonnegative number y such that*

$$f(x) + yg(x) \geq 0 \quad \forall x \in \mathbb{R}^n,\ Hx = r. \tag{3.13e}$$

With some convexity assumption we can include a linear inequality:

PROPOSITION 3.10. *Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be quadratic functions, $h \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$. Assume that $g(x)$ is convex and there exists an $\bar{x} \in \mathbb{R}^n$ such that $h^T\bar{x} < \alpha$ and $g(\bar{x}) < 0$. The following two statements are equivalent:*

(i) *The system*

$$f(x) < 0 \tag{3.14a}$$
$$g(x) \leq 0 \tag{3.14b}$$
$$h^T x \leq \alpha \tag{3.14c}$$
$$x \in \mathbb{R}^n \tag{3.14d}$$

is not solvable.

(ii) *There is a nonnegative multiplier $y \geq 0$ and a vector $(u_0, u) \in \mathbb{R}^{n+1}$ such that*

$$f(x) + yg(x) + (u^T x - u_0)(h^T x - \alpha) \geq 0 \quad \forall x \in \mathbb{R}^n, \tag{3.15a}$$
$$u^T x \geq 0 \quad \forall x \in \mathbb{R}^n : x^T A_g x \leq 0,\ b_g^T x \leq 0 \tag{3.15b}$$
$$u^T x - u_0 \geq 0 \quad \forall x \in \mathbb{R}^n : g(x) \leq 0,\ c_g + b_g^T x \leq 0, \tag{3.15c}$$

where $g(x) = x^T A_g x + b_g^T x + c_g$.

For a proof see [73]. We can see that including even one linear inequality is difficult, and the dual problem (3.15) is not any easier than the primal problem (3.14).

3.3. Counterexamples. In this section we present some counterexamples.

3.3.1. More inequalities. The generalization to the case $m \geq 3$ is practically hopeless as it is illustrated by the following example taken from [11]. Consider the matrices

$$A = \begin{pmatrix} 1 & 1.1 & 1.1 \\ 1.1 & 1 & 1.1 \\ 1.1 & 1.1 & 1 \end{pmatrix}, \quad B_1 = \begin{pmatrix} -2.1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{3.16a}$$

$$B_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -2.1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad B_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2.1 \end{pmatrix}. \tag{3.16b}$$

LEMMA 3.11. *There is an $\bar{x}$ such that $\bar{x}^T B_i \bar{x} < 0$, $i = 1, 2, 3$ (the Slater condition holds). Further, the system*

$$x^T A x < 0 \tag{3.17a}$$

$$x^T B_i x \leq 0, \ i = 1, 2, 3 \tag{3.17b}$$

is not solvable in $\mathbb{R}^3$.

Proof. The Slater condition is satisfied with $\bar{x} = (1, 1, 1)$. Let us now look for a solution in the form $x = (x_1, x_2, x_3)$. If $x_3 = 0$ then by the last three inequalities we can conclude that $x_1 = x_2 = 0$, which does not satisfy the first inequality. Since all the functions are homogeneous, and since x and $-x$ are essentially the same solution we can assume that $x_3 = 1$, thus we can reduce the dimension of the problem. Now all four of the inequalities define some quadratic areas in $\mathbb{R}^2$. Instead of a formal and

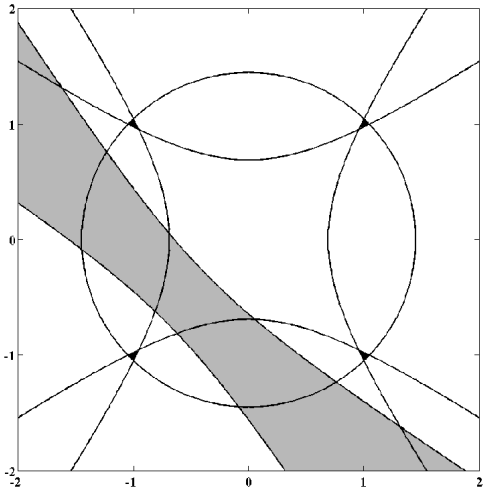

FIG. 3.1. *The quadratic regions defined by system (3.16) with $x_3 = 1$.*

tedious proof we simply plot these areas in Fig. 3.1. The grey area represents the set of points (x_1, x_2) where $x = (x_1, x_2, 1)$ satisfies $x^T A x < 0$, while the four black corners are the feasible set for the remaining three inequalities. Intuitively, the last

three inequalities are satisfied for values close to ± 1. However, it is easy to see that such values can not satisfy the first inequality. □

LEMMA 3.12. *There are no nonnegative multipliers y_1, y_2, y_3 for which $A+y_1B_1+y_2B_2+y_3B_3 \succeq 0$.*

Proof. Consider the matrix

$$X = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}, \tag{3.18}$$

which is positive semidefinite with eigenvalues 0 and 3. Moreover,

$$A \bullet X = -0.6 < 0 \tag{3.19a}$$

$$B_i \bullet X = -0.2 \leq 0. \tag{3.19b}$$

Now for any nonnegative linear combination of the matrices, we have

$$(A + y_1B_1 + y_2B_2 + y_3B_3) \bullet X = -0.6 - 0.2(y_1 + y_2 + y_3) < 0, \tag{3.20}$$

therefore $A + y_1B_1 + y_2B_2 + y_3B_3$ can never be positive semidefinite, since the scalar product of positive semidefinite matrices is nonnegative. □

These two lemmas show that neither of the alternatives is true in the general theorem.

3.3.2. The $n = 2$ case. Discussing the $m = 2$ case (Proposition 3.6) we noted that the result fails to hold if $n = 2$. Here is a counterexample taken from [11] to demonstrate this.

Let us consider the following three matrices:

$$A = \begin{pmatrix} \lambda\mu & 0.5(\mu - \lambda) \\ 0.5(\mu - \lambda) & -1 \end{pmatrix} \tag{3.21a}$$

$$B = \begin{pmatrix} -\mu\nu & -0.5(\mu - \nu) \\ -0.5(\mu - \nu) & 1 \end{pmatrix} \tag{3.21b}$$

$$C = \begin{pmatrix} -\lambda^2 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.21c}$$

We will verify the following claims:

PROPOSITION 3.13. *Let $\lambda = 1.1$, $\mu = 0.818$ and $\nu = 1.344$.*

(i) *There is a positive definite linear combination of A, B and C.*

(ii) *There is a vector $\bar{x}$ such that $\bar{x}^T B\bar{x} < 0$ and $\bar{x}^T C\bar{x} < 0$.*

(iii) *The quadratic system*

$$x^T Ax < 0 \tag{3.22a}$$

$$x^T Bx \leq 0 \tag{3.22b}$$

$$x^T Cx \leq 0 \tag{3.22c}$$

is not solvable.

(iv) *There is no $y_1, y_2 \geq 0$ such that*

$$A + y_1B + y_2C \succeq 0. \tag{3.23}$$

Proof.

(i) The linear combination

$$-1.15A - 0.005B - C = \begin{pmatrix} 0.18072696 & 0.160835 \\ 0.160835 & 0.1450 \end{pmatrix} \tag{3.24}$$

is positive definite as it is seen by the diagonal elements and the determinant.

(ii) Let $\bar{x} = (1,0)^T$ then $\bar{x}^T B \bar{x} < 0$ and $\bar{x}^T C \bar{x} < 0$.

(iii) Let us exploit the special structure of the matrices. If we are looking for a solution $x = (x_1, x_2) \in \mathbb{R}^2$ then we get

$$x^T A x = \lambda\mu x_1^2 - x_2^2 + (\mu - \lambda) x_1 x_2 = (\lambda x_1 + x_2)(\mu x_1 - x_2) \tag{3.25a}$$

$$x^T B x = -\mu\nu x_1^2 + x_2^2 - (\mu - \nu) x_1 x_2 = (\nu x_1 + x_2)(-\mu x_1 + x_2) \tag{3.25b}$$

$$x^T C x = -\lambda^2 x_1^2 + x_2^2 = (-\lambda x_1 + x_2)(\lambda x_1 + x_2). \tag{3.25c}$$

Now, in order to satisfy (3.22) we need to solve one of the following two systems corresponding to which terms are negative and positive:

$$\lambda x_1 + x_2 > 0 \tag{3.26a}$$

$$\mu x_1 - x_2 < 0 \tag{3.26b}$$

$$\nu x_1 + x_2 \leq 0 \tag{3.26c}$$

$$-\lambda x_1 + x_2 \leq 0 \tag{3.26d}$$

or

$$\lambda x_1 + x_2 < 0 \tag{3.27a}$$

$$\mu x_1 - x_2 > 0 \tag{3.27b}$$

$$\nu x_1 + x_2 \geq 0 \tag{3.27c}$$

$$-\lambda x_1 + x_2 \geq 0. \tag{3.27d}$$

It is easy to check that with the values specified in the statement both of these systems are inconsistent, therefore (3.22) is not solvable.

(iv) The proof of this part is similar to the proof of Lemma 3.12. The matrix

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{3.28}$$

satisfies

$$A \bullet X < 0 \tag{3.29a}$$

$$B \bullet X \leq 0 \tag{3.29b}$$

$$C \bullet X \leq 0, \tag{3.29c}$$

thus no nonnegative linear combination $A + y_1 B + y_2 C$ of A, B and C is positive semidefinite.

This completes the proof of the lemma. □

Let us examine briefly why the S-lemma fails to hold for this example. We have already shown in Proposition 3.6 that if $n = m = 2$ then in order for the result to hold we need to assume that a certain linear combination of B and C is positive definite. However, taking the positive definite matrix

$$X = \begin{pmatrix} 1 & \frac{\lambda^2 - \mu\nu}{\mu - \nu} \\ \frac{\lambda^2 - \mu\nu}{\mu - \nu} & \lambda^2 \end{pmatrix} \tag{3.30}$$

yields

$$B \bullet X = 0 \tag{3.31a}$$

$$C \bullet X = 0, \tag{3.31b}$$

therefore no linear combination of B and C can be positive definite.

4. Practical applications.

4.1. Stability analysis. The first example is taken from [43].

Let us consider the following dynamical system

$$\dot{x} = Ax + Bw, \quad x(0) = x_0 \tag{4.1a}$$

$$v = Cx \tag{4.1b}$$

with a so-called sector constraint

$$\sigma(v, w) = (\beta v - w)^T (w - \alpha v) \geq 0, \tag{4.1c}$$

where $\alpha < \beta$ are real numbers. We would like to use the basic tool of Lyapunov functions [15]: for the quadratic stability of the system it is necessary and sufficient to have a symmetric matrix P such that $V(x) = x^T P x$ is a Lyapunov function, i.e.,

$$\dot{V}(x) = 2x^T P(Ax + Bw) < 0, \quad \forall (x, w) \neq 0 \text{ s.t. } \sigma(Cx, w) \geq 0. \tag{4.2}$$

Introducing the quadratic forms

$$\sigma_0(x, w) = \begin{bmatrix} x \\ w \end{bmatrix}^T \begin{bmatrix} A^T P + PA & PB \\ B^T P & 0 \end{bmatrix} \begin{bmatrix} x \\ w \end{bmatrix} \tag{4.3a}$$

$$\sigma_1(x, w) = 2\sigma(Cx, w) = \begin{bmatrix} x \\ w \end{bmatrix}^T \begin{bmatrix} -2\beta\alpha C^T C & (\beta + \alpha) C^T \\ (\beta + \alpha) & -2 \end{bmatrix} \begin{bmatrix} x \\ w \end{bmatrix} \tag{4.3b}$$

the Lyapunov condition can be written as

$$\sigma_0(x, w) < 0, \quad \forall (x, w) \neq 0 \text{ s.t. } \sigma_1(x, w) \geq 0, \tag{4.4}$$

or in other words, we have to decide the solvability of the quadratic system

$$\sigma_0(x, w) \geq 0 \tag{4.5a}$$

$$\sigma_1(x, w) \geq 0. \tag{4.5b}$$

Using $\alpha < \beta$ we can see that the strict version of the second inequality can be satisfied. Based on a suitable form of the S-lemma (Proposition 3.2) we get that the non-solvability of this system is equivalent to the existence of $y \geq 0$ such that

$$\sigma_0(x, w) + y\sigma_1(x, w) < 0 \ \forall (x, w) \neq (0, 0). \tag{4.6}$$

Now $y = 0$ would imply that σ_0 is negative definite and would contradict to the non-solvability of (4.5). Thus we can divide with y and use P to denote P/y. Finally, we can state the criterion using the LMI formulation. We get the following theorem:

THEOREM 4.1 (Circle criterion, [15]). *A necessary and sufficient condition for the quadratic stability of the system*

$$\dot{x} = Ax + Bw, \quad x(0) = x_0 \tag{4.7a}$$

$$v = Cx \tag{4.7b}$$

$$\sigma(v, w) = (\beta v - w)^T (w - \alpha v) \geq 0, \tag{4.7c}$$

where $\alpha < \beta$, is the existence of a symmetric matrix P such that

$$\begin{bmatrix} A^TP + PA - 2\beta\alpha C^TC & PB + (\beta+\alpha)C^T \\ B^TP + (\beta+\alpha)C & -2 \end{bmatrix} \prec 0. \tag{4.7d}$$

4.2. Sum of two ellipsoids. This example is taken from [63].

Let $E_i = E(a^i, A_i) \subseteq \mathbb{R}^n$, $n \geq 2$ be an ellipsoid with center a^i and shape A_i, i.e.,

$$E(a^i, A_i) = \left\{ x \in \mathbb{R}^n : (x - a^i)^T A_i (x - a^i) \leq 1 \right\}. \tag{4.8}$$

The sum of two such ellipsoids is the usual Minkowski sum:

$$Q = E_1 + E_2 = \left\{ x^1 + x^2 : x^1 \in E_1,\, x^2 \in E_2 \right\}. \tag{4.9}$$

We are looking for a general description of all the ellipsoids that contain this sum. This is an important question in error estimation. Errors on measurements are usually modeled by Gaussian distributions. If we are looking for the error of the sum of two measurements then we need to solve this problem.

First notice that since

$$Q = a^1 + a^2 + E(0, A_1) + E(0, A_2) \tag{4.10}$$

we can assume that all the ellipsoids are centered at the origin. Our target object is an ellipsoid $E(0, A_0)$ such that

$$E(0, A_0) \supset E(0, A_1) + E(0, A_2). \tag{4.11}$$

This condition can be stated equivalently using the notation $x = (x^1, x^2) \in \mathbb{R}^{2n}$. The ellipsoid E_0 contains the sum $E_1 + E_2$ if and only if the following system is not solvable:

$$x^T \begin{pmatrix} -A_0 & -A_0 \\ -A_0 & -A_0 \end{pmatrix} x < 1 \tag{4.12a}$$

$$x^T \begin{pmatrix} A_1 & 0 \\ 0 & 0 \end{pmatrix} x \leq 1 \tag{4.12b}$$

$$x^T \begin{pmatrix} 0 & 0 \\ 0 & A_2 \end{pmatrix} x \leq 1. \tag{4.12c}$$

Since the matrix

$$\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} \tag{4.13}$$

is positive definite and $n \geq 2$, we can apply a slightly modified version of Proposition 3.6 to obtain the following characterization:

THEOREM 4.2. *An ellipsoid $E(0, A_0)$ contains the Minkowski sum of the ellipsoids $E(0, A_1)$ and $E(0, A_2)$ if and only if there exist nonnegative numbers y_1 and y_2 such that $y_1 + y_2 \leq 1$ and the matrix*

$$\begin{pmatrix} -A_0 + y_1 A_1 & -A_0 \\ -A_0 & -A_0 + y_2 A_2 \end{pmatrix} \tag{4.14}$$

is positive semidefinite.

This condition can be validated in polynomial time. Further, we can use this result to build an algorithm to minimize the maximum eigenvalue of A_0. A similar argument can be repeated for the intersection of two ellipsoids, see [63].

Several other applications can be found in the literature, such as distance geometry [9], portfolio management [3, 4, 29], statistics [38], signal processing [50], or control and stability problems [11, 15, 53, 56], just to mention a few.

PART II

The results and examples presented in the first part show how diverse areas are contributing to the theory of the S-lemma. Now we give a summary of these connections. In each section we first present how the S-lemma is related to the specific theory then we summarize the relevant results of the field and finally we discuss the consequences of the results and draw the conclusions.

5. Convexity of the joint numerical range. In this section we investigate the theory behind the first proof for the S-lemma, see §2.2.

5.1. Motivation. Recall that the key step in Yakubovich's proof was to use Dines's result about the convexity of the set

$$\left\{(x^TAx, x^TBx) : x \in \mathbb{R}^n\right\} \subseteq \mathbb{R}^2. \tag{5.1}$$

The separation idea we used can be extended to more inequalities. Let us assume that the system

$$x^TAx < 0 \tag{5.2a}$$

$$x^TB_ix \leq 0, \quad i = 1, \ldots, m \tag{5.2b}$$

is not solvable and assume that the Slater condition is satisfied, i.e., there exists an $\bar{x} \in \mathbb{R}^n$ such that $\bar{x}^TB_i\bar{x} < 0$ for all $i = 1, \ldots, m$. If the set

$$H_{\mathbb{R}}(A, B_1, \ldots, B_m) = \left\{(x^TAx, x^TB_1x, \ldots, x^TB_mx) : x \in \mathbb{R}^n\right\} \subseteq \mathbb{R}^{m+1} \tag{5.3}$$

is convex then an equivalent characterization of the non-solvability of (5.2) is that

$$H_{\mathbb{R}}(A, B_1, \ldots, B_m) \cap \mathcal{C} \tag{5.4}$$

is empty, where

$$\mathcal{C} = \{(u_0, u) : u_0 < 0, u \leq 0\} \subset \mathbb{R}^{m+1} \tag{5.5}$$

is a convex cone. A well known basic fact in convex analysis (see [68, 72]) is that disjoint convex sets can be separated by a hyperplane, i.e., there exist $(y_0, y) \in \mathbb{R}^{m+1} \setminus (0, 0)$ such that

$$y_0u_0 + \sum_{i=1}^{m} y_iu_i \geq 0, \ \forall (u_0, u) \in H_{\mathbb{R}} \tag{5.6a}$$

$$y_0u_0 + \sum_{i=1}^{m} y_iu_i \leq 0, \ \forall (u_0, u) \in \mathcal{C}. \tag{5.6b}$$

Since $(-1,0) \in \mathcal{C}$ we get $y_0 \geq 0$, and using $(-\varepsilon, -e^i) \in \mathcal{C}$ where $e^i \in \mathbb{R}^m$ is the i^{th} unit vector we get $y \geq 0$.

Using the Slater condition we have a $(\bar{u}_0, \bar{u}) \in H_{\mathbb{R}}$ where $\bar{u} < 0$. If $y = 0$ then, since all the coefficients can not be zero, we have $y_0 > 0$. On the other hand, if $y \neq 0$ then using the Slater-point we have

$$y_0 \bar{u}_0 + \underbrace{\sum_{i=1}^{m} y_i \bar{u}_i}_{<0} \geq 0 \tag{5.7}$$

implying that $y_0 > 0$. After dividing by y_0 we get the desired coefficients. We have thus proved that the convexity of $H_{\mathbb{R}}(A, B_1, \ldots, B_m)$ implies the validity of the S-lemma.

5.2. Theoretical results. First we present results over the field of real numbers, then we generalize the concept to complex numbers.

5.2.1. Results over real numbers. The key question is the following: How can we guarantee the convexity of $H_{\mathbb{R}}(A, B_1, \ldots, B_m)$?

Most of the results (see [7, 32, 49]) on the convexity of the numerical range investigate the question over the complex field. However, we need convexity results for $H_{\mathbb{R}}(A, B_1, \ldots, B_m)$. The first such result has already been mentioned. It is credited to Dines [22] and dates back to 1941.

THEOREM 5.1 (Dines, [22], 1941). *If $f, g : \mathbb{R}^n \to \mathbb{R}$ are* homogeneous *quadratic functions then the set $\mathcal{M} = \{(f(x), g(x)) : x \in \mathbb{R}^n\} \subset \mathbb{R}^2$ is convex.*

An analogous result for three quadratic forms was proved by Polyak [63] in 1998, using a theorem by Brickman [17].

THEOREM 5.2 (Polyak, [63], 1998). *If $n \geq 3$ and $f, g, h : \mathbb{R}^n \to \mathbb{R}$ are* homogeneous *quadratic functions such that there exists a positive definite linear combination of them, then the set $\mathcal{M} = \{(f(x), g(x), h(x)) : x \in \mathbb{R}^n\} \subset \mathbb{R}^3$ is convex.*

Another interesting source is Ramana's article [67] from 1995. He characterizes the quadratic transformations $F : \mathbb{R}^n \to \mathbb{R}^m$ for which the set $F(\mathbb{R}^n)$ is convex. He calls such maps Image Convex (ICON) and establishes the following theorem:

THEOREM 5.3 (Ramana, [67], 1995). *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a constant-free ($F(0) = 0$) quadratic map. Then F is ICON if and only if $F(\mathbb{R}^n) = F^Q(\mathbb{R}^n) + F(\mathbb{R}^n)$ where $F^Q(x) = \frac{F(x)+F(-x)}{2}$ is the quadratic part of F, and the sum is the Minkowski-sum.*

If F is homogeneous, as it is so in our case, then the equivalent condition reduces to $F(\mathbb{R}^n) = F(\mathbb{R}^n) + F(\mathbb{R}^n)$, which is trivial.[2] Further, Ramana proves that the identification of ICON maps is NP-hard.

Similarly, Ramana investigates those quadratic maps, under which the image of every linear subspace is convex. He calls these maps Linear Image Convex (LICON). Obviously, all LICON maps are ICON maps, thus equivalent conditions for the LICON property provide sufficient conditions for the ICON property. He establishes the following equivalent condition:

THEOREM 5.4 (Ramana, 1995, [67]). *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a constant-free ($F(0) = 0$) quadratic map, then F is LICON if and only if at least one of the following conditions holds:*

(i) *F is of the form $(u^T x + a)Ax$, or*

[2] Using the terms of §7 this means that $F(\mathbb{R}^n)$ is König-linear.

(ii) $\chi(F) \leq 1$, *or*
(iii) F *is homogeneous and*

$$\operatorname{rank}\{F(x), F(y), F(x+y)\} \leq 2, \forall x, y \in \mathbb{R}^n, \tag{5.8}$$

where $\chi(F)$ *is the polynomial rank*[3] *of* F.

REMARK 5.5. *This theorem can be exploited to check the LICON property in polynomial time. This shows that the recognition of LICON maps is much simpler than that of the ICON maps.*

A similar problem is the convexity of the image of the unit sphere. Since we are dealing with homogeneous quadratic functions, the image of the whole space is the cone spanned by the image of the unit sphere, thus if the image of the unit sphere is convex, then so is the image of the whole space. The corresponding theorem for the real case was proved by Brickman in 1961:

THEOREM 5.6 (Brickman, [17], 1961). *Let* $f, g : \mathbb{R}^n \to \mathbb{R}$ *be homogeneous quadratic functions. If* $n \geq 3$ *then the set*

$$W_{f,g} = \{(f(x), g(x)) : x \in \mathbb{R}^n, \|x\| = 1\} \subset \mathbb{R}^2 \tag{5.11}$$

is convex.

REMARK 5.7. *If* $n = 2$ *then this set is not necessarily convex. Let us take* $x = (x_1, x_2)$ *and define* $f(x) = x_1^2 - x_2^2$ *and* $g(x) = 2x_1x_2$. *These functions satisfy* $f(x)^2 + g(x)^2 = 1$, *thus the image is the unit circle line, which is not convex.*

Polyak used this result to prove his earlier mentioned theorem (see Theorem 5.2). The original proof of this theorem uses advanced differential geometric arguments. The following elementary proof is by H. Pépin from [60].[4] It only uses the characterization of quadratic surfaces in $\mathbb{R}^3$.

Proof. The proof is based on the following two simple lemmas.

LEMMA 5.8. *Let* $P : \mathbb{R}^2 \to \mathbb{R}^2$ *be an affine map,* $f, g : \mathbb{R}^n \to \mathbb{R}$ *homogeneous quadratic functions, then there are homogeneous quadratic functions* $\tilde{f}, \tilde{g} : \mathbb{R}^n \to \mathbb{R}$ *such that* $P(W_{f,g}) = W_{\tilde{f},\tilde{g}}$.

Proof. Let P be of the form $P(u, v) = (a_1u + b_1v + c_1, a_2u + b_2v + c_2)$, then $\tilde{f}(x) = a_1f(x) + b_1g(x) + c_1\|x\|^2$ and $\tilde{g}(x) = a_2f(x) + b_2g(x) + c_2\|x\|^2$ satisfy the requirements. □

LEMMA 5.9. *Let* $V \subseteq \mathbb{R}^n$ *be a subspace with* $\dim(V) \geq 3$. *If there are two points* $x, y \in V$ *such that* $f(x) = f(y) = 0$ *and* $g(x)g(y) < 0$ *then there is a third point* $z \in V$ *for which* $\|z\| = 1$ *and* $f(z) = g(z) = 0$.

Proof. We can assume without loss of generality that $\dim(V) = 3$. Let us define the following cone:

$$\mathcal{C} = \{x \in V : f(x) = 0\}, \tag{5.12}$$

[3] Given a polynomial map of the form

$$F(x) = \sum_{\alpha \in \mathcal{A}} x^\alpha v_\alpha, \tag{5.9}$$

where $\{x^\alpha : \alpha \in \mathcal{A}\}$ is the set of monomials appearing in F and $v_\alpha \in \mathbb{R}^m$, the polynomial rank of F is defined as

$$\chi(F) = \operatorname{rank}\{v_\alpha : \alpha \in \mathcal{A}\}. \tag{5.10}$$

The condition $\chi(F) \leq 1$ requires that all the vector-coefficients of the terms x_i^2, x_ix_j and x_i are constant multiplies of each other.

[4] We thank Jean-Baptiste Hiriart-Urruty for this reference.

then $x, y \in \mathcal{C}$, and they must be linearly independent since $g(x)g(y) < 0$. As $\mathcal{C}$ is a three dimensional homogeneous quadratic surface, and it is not a trivial cone of one point or one direction, it can either be a second-order cone, a plane, a union of two planes or the whole subspace V. In any case, the set $\mathcal{C} \setminus \{0\}$ is either connected or it consists of two centrally symmetric connected components. Now taking $-y$ instead of y we can assume that x and y belong to the same connected component of $\mathcal{C}$ and by the continuity of $g(x)$ we have a point u in the component satisfying $g(u) = 0$. Finally, $z = u/\|u\|$ satisfies all the requirements of the lemma. □

Using these two lemmas we can finish the proof easily. Let $V \subseteq \mathbb{R}^n$ be a subspace with $\dim(V) \geq 3$, $f, g : V \to \mathbb{R}$ homogeneous quadratic functions, and assume that $W_{f,g}$ is not only the origin. Let a and b be two distinct points in $W_{f,g}$ and let c be a point on the open line segment between a and b. Let x and y be the pre-images of a and b, respectively. Let us choose an affine bijection $P : \mathbb{R}^2 \to \mathbb{R}^2$ for which $P(c) = (0, 0)$, $P(a) = (0, 1)$; then there is a $\beta < 0$ such that $P(b) = (0, \beta)$. Applying Lemma 5.8 we get functions $\tilde{f}$ and $\tilde{g}$ such that $P(W_{f,g}) = W_{\tilde{f},\tilde{g}}$. Now we have $\tilde{f}(x) = 0$, $\tilde{g}(x) = 1$, $\tilde{f}(y) = 0$ and $\tilde{g}(y) = \beta < 0$. Applying Lemma 5.9 we get a point $z \in V$ such that $\|z\| = 1$ and $\tilde{f}(z) = \tilde{g}(z) = 0$. This means that $(0, 0) \in W_{\tilde{f},\tilde{g}} = P(W_{f,g})$ and therefore $c = P^{-1}(0, 0) \in W_{f,g}$. This completes the proof of Theorem 5.6. □

More recently, Polyak proved a local version of these theorems:

THEOREM 5.10 (Polyak, [64], 2001). *Let $x \in \mathbb{R}^n$ and $f(x) = (f_1(x), \ldots, f_m(x))$, where*

$$f_i(x) = \frac{1}{2} x^T A_i x + {a^i}^T x,\ i = 1, \ldots, m \tag{5.13}$$

are quadratic functions. Let A be an $n \times m$ matrix with columns a^i and let us define

$$L = \sqrt{\sum_{i=1}^{m} \|A_i\|^2} \tag{5.14a}$$

$$\nu = \sigma_{\min}(A) = \sqrt{\lambda_{\min}(A^T A)}, \tag{5.14b}$$

where $\|A_i\|$ is the operator norm of A_i, and $\sigma_{\min}(A)$ denotes the smallest singular value of A. If $\varepsilon < \nu/(2L)$ then the image $\{f(x) : x \in \mathbb{R}^n,\ \|x\| \leq \varepsilon\}$ is a convex set in $\mathbb{R}^m$.

The statement remains true if we take a small ellipsoid instead of the ball. In that case the norm constraint becomes $x^T H x < \varepsilon$, where H is some positive definite matrix. Notice that the theorem says nothing if all the functions are homogeneous, since in that case $A = 0$.

Polyak proved his result for much more general nonlinear functions in Hilbert spaces, but the general idea is the same: if f is "close" to its linear approximation then it will preserve convexity (as its linear approximation would also do so).

Finally, let us mention a negative result. Under some structural assumptions it is impossible that the image of the real unit surface is convex. Let $A = (A_1, \ldots, A_m)$ be an m-tuple of $n \times n$ real symmetric matrices, and let $1 \leq k \leq n$. The set

$$W_k^{\mathbb{R}}(A) = \left\{ (\mathrm{Tr}(X^T A_1 X), \ldots, \mathrm{Tr}(X^T A_m X)) : X \in \mathbb{R}^{n \times k},\ X^T X = I_k \right\} \tag{5.15}$$

is called the real k^{th} joint numerical range of A (see [49]). This object is closely connected to the quadratic system presented in §6.4.

THEOREM 5.11 (Li, [49], 2000). *Assume that the matrices $A_1, \dots, A_m$ are linearly independent. If $m > k(n-k)+1$ then $W_k^{\mathbb{R}}(A)$ is not convex. Further, if the identity matrix is not a linear combination of $A_1, \dots, A_m$ and $m > k(n-k)$ then $W_k^{\mathbb{R}}(A)$ is not convex.*

It is interesting to contrast this theorem with Brickman's result (Theorem 5.6). Brickman proved that if $n \geq 3$ then $W_1(A)$ is convex for $m = 2$. Further generalizations are blocked by Li's negative result: to prove a similar convexity theorem for $m = 3$ it is necessary (but not sufficient!) to assume $n \geq 4$, or the existence of a positive definite linear combination of the matrices and $n \geq 3$. It is surprising, though, that the convexity of the joint numerical range is a structural property, i.e., for certain values of m and n the image is convex for all possible linearly independent families of matrices, while for other values convexity is not possible at all.

If the image of the unit sphere is not convex then one might wonder how much it is nonconvex. One way to tell this is through the description of the convex hull. The Carathéodory Theorem [68] states that every point in the convex hull of an m-dimensional set can be written as the convex combination of $m+1$ points from the set. In our case we have much stronger results:

THEOREM 5.12 (Poon, [65], 1994). *Let $A = (A_1, \dots, A_m)$ be an m-tuple of $n \times n$ real symmetric matrices, and let $W_1^{\mathbb{R}}(A)$ denote the joint numerical range defined earlier, i.e., $W_1^{\mathbb{R}}(A)$ is the quadratic image of the n-dimensional real unit sphere. Then every point in the convex hull of $W_1^{\mathbb{R}}(A)$ can be expressed as the convex combination of at most $k^{\mathbb{R}}(m,n)$ points from $W_1^{\mathbb{R}}(A)$, where*

$$k^{\mathbb{R}}(m,n) = \min\left\{n,\ \left\lfloor \frac{\sqrt{8m+1}-1}{2} \right\rfloor + \delta_{\frac{n(n+1)}{2},m}\right\}, \tag{5.16}$$

and $\delta_{a,b}$ is the Kronecker symbol, i.e., $\delta_{a,b} = 1$ or 0 depending on whether $a = b$ or $a \neq b$, respectively.

5.2.2. Results over complex numbers. Similar questions were first investigated by Hausdorff [34] and Toeplitz [74] who proved that if $A, B_1, \dots, B_m$ are complex Hermitian matrices then the set

$$H_{\mathbb{C}}(A, B_1, \dots, B_m) = \{(z^*Az, z^*B_1z, \dots, z^*B_mz) : z \in \mathbb{C}^n, \|z\| = 1\} \subseteq \mathbb{R}^{m+1}, \tag{5.17}$$

where z^* is the complex conjugate-transpose, is convex if $m = 1$. This object is called the *joint numerical range* of the matrices. Later, Au-Yeung and Poon [6] proved that if $n \geq 3$ then the joint numerical range of three Hermitian matrices is also convex.

A general differential geometric characterization of the cases when the joint numerical range is convex can be found in [32]. This is the strongest known theorem for the general case.

THEOREM 5.13 (Gutkin et al., [32], 2002). *Let $A_1, \dots A_m$ be $n \times n$ complex Hermitian matrices. If the multiplicity of the largest eigenvalue of $\sum_{i=1}^m \mu_i A_i$ is the same for all $\mu_1, \dots \mu_m$, where $\sum_{i=1}^m \mu_i^2 = 1$, and the union of the eigenspaces corresponding to the largest eigenvalue is not the whole $\mathbb{C}^n$ then the set*

$$\{(z^*A_1z, \dots, z^*A_mz) : z \in \mathbb{C}^n,\ z^*z = 1\} \subseteq \mathbb{R}^m \tag{5.18}$$

is convex.

REMARK 5.14. *The second condition is redundant unless $m = n+1$ and the multiplicity of the largest eigenvalue is $n/2$. Moreover, if $m \geq 4$ and the conditions*

of the theorem fail for some matrices but the image is still convex, then there is an arbitrarily small perturbation of $A_1, \dots A_m$ that destroys convexity. Thus, the above condition is almost necessary. For more details and proofs see [32].

Now let us have the complex analogue of Theorem 5.11:

THEOREM 5.15 (Li, [49], 2000). *Let $A = (A_1, \dots, A_m)$ be an m-tuple of $n \times n$ complex Hermitian matrices, and let $1 \le k \le n$. The set*

$$W_k^{\mathbb{C}}(A) = \left\{ (\operatorname{Tr}(X^* A_1 X), \dots, \operatorname{Tr}(X^* A_m X)) : X \in \mathbb{C}^{n\times k},\ X^* X = I_k \right\} \tag{5.19}$$

is the complex k^{th} joint numerical range of A (see [49]). Assume that the matrices $A_1, \dots, A_m$ are linearly independent. If $m > 2k(n-k)+1$ then $W_k^{\mathbb{C}}(A)$ is not convex. Further, if the identity matrix is not a linear combination of $A_1, \dots, A_m$ and $m > 2k(n-k)$ then $W_k^{\mathbb{C}}(A)$ is not convex.

Finally, we can state the complex counterpart of Theorem 5.12.

THEOREM 5.16 (Poon, [65], 1994). *Let $A = (A_1, \dots, A_m)$ be an m-tuple of $n \times n$ complex Hermitian matrices, and let $W_1^{\mathbb{C}}(A)$ denote the joint numerical range defined earlier, i.e., $W_1^{\mathbb{C}}(A)$ is the quadratic image of the n-dimensional complex unit sphere. Every point in the convex hull of $W_1^{\mathbb{C}}(A)$ can be expressed as the convex combination of at most $k^{\mathbb{C}}(m,n)$ points from $W_1^{\mathbb{C}}(A)$, where*

$$k^{\mathbb{C}}(m,n) = \min\left\{ n,\ \lfloor \sqrt{m} \rfloor + \delta_{n^2,m+1} \right\}. \tag{5.20}$$

One might wonder why the complex case is more deeply developed than the real one. The reason for this is twofold. Firstly, from a purely differential geometric point of view the complex field has a much nicer structure, which allows for more advanced proof techniques. Secondly, as we will argue later in §6.5, the problem is structurally simpler over the complex field.

Recently, Faybusovich [25] put all these results in a unified context and provided general proofs using Jordan algebras.

5.3. Implications. It is straightforward to apply these results to our case. Whenever a collection of matrices satisfies the convexity property, we can characterize the solvability of the corresponding quadratic system with a simple LMI. Moreover, we can extend the results to more general cases.

5.3.1. Results over real numbers. Let us start with the results over real numbers. As we have already seen in the first proof, Dines's result (Theorem 5.1) gives rise to the basic homogeneous S-lemma. The generalization for three inequalities (Proposition 3.6) comes from Polyak's convexity result (Theorem 5.2). The norm constrained results will give us something new. Using Theorem 5.6 with a simple separation idea we get the following theorem:

THEOREM 5.17. *Let $n \ge 3$, $A, B \in \mathbb{R}^{n\times n}$ real symmetric matrices. Assume further that there exists a Slater point $\bar{x} \in \mathbb{R}^n$ such that $\|\bar{x}\| = 1$ and $\bar{x}^T B \bar{x} < \beta$. The following two statements are equivalent:*

(i) *The system*

$$x^T A x < \alpha \tag{5.21a}$$

$$x^T B x \le \beta \tag{5.21b}$$

$$\|x\| = 1 \tag{5.21c}$$

is not solvable.

(ii) *There is a nonnegative multiplier y such that*

$$x^T Ax - \alpha + y(x^T Bx - \beta) \geq 0,\ \forall x \in \mathbb{R}^n,\ \|x\| = 1, \tag{5.22a}$$

or equivalently

$$A - \alpha I + y(B - \beta I) \succeq 0. \tag{5.22b}$$

The latter condition is an LMI, thus it can be verified in essentially polynomial time.[5] This result, however, can not be extended to general nonhomogeneous functions.

Polyak's local convexity result (Theorem 5.10) can be used to prove the following local duality result:

THEOREM 5.18. *Let $x \in \mathbb{R}^n$ and $f(x) = (f_1(x), \ldots, f_m(x))$, where*

$$f_i(x) = \frac{x^T A_i x}{2} + {a^i}^T x \tag{5.23}$$

are quadratic functions. If the vectors a^i, $i = 1, \ldots, m$ are linearly independent then there exists an $\bar{\varepsilon} > 0$ such that for all $\varepsilon < \bar{\varepsilon}$ the following two statements are equivalent:

(i) *The system*

$$f_i(x) \leq \alpha_i,\ i = 1, \ldots, m \tag{5.24a}$$

$$\|x\| \leq \varepsilon \tag{5.24b}$$

is not solvable.

(ii) *There exists a vector of nonnegative multipliers $y_1, \ldots, y_m$ (not all of them are zero) such that*

$$\sum_{i=1}^{m} y_i (f_i(x) - \alpha_i) \geq 0,\ \forall x \in \mathbb{R}^n,\ \|x\| \leq \varepsilon. \tag{5.25}$$

Proof. If the vectors a^i, $i = 1, \ldots, m$ are linearly independent then the smallest singular value of $A = (a^1 | \ldots | a^m)$ is positive, therefore from Theorem 5.10 the set

$$\{(f_1(x), \ldots, f_m(x)) : x \in \mathbb{R}^n,\ \|x\| < \varepsilon\} \subset \mathbb{R}^m \tag{5.26}$$

is convex for any $\varepsilon < \bar{\varepsilon} = \nu/(2L)$, where ν and L are defined in Theorem 5.10. After this we can apply the usual separation idea to finish the proof, see e.g., §2.2. □

These kind of results usually require the convexity of the functions (see, [68, 72]). Here, however, we do not need to impose convexity on the functions. It is important to note, that unlike other results presented so far, Theorem 5.18 uses quantitative information of the problem data.

Now we can put together the pieces. If the image is convex then we can use the separation argument presented in §2.2 to get the dual statement. If the image is nonconvex then we can use Theorem 5.12 to characterize how much the image is not convex. This idea yields the following result.

THEOREM 5.19. *If the system*

$$x^T A_i x \leq \alpha_i,\ i = 1, \ldots, m \tag{5.27a}$$

$$\|x\| = 1 \tag{5.27b}$$

is not solvable then one of the following two statements is true:

[5] For a complete discussion on the complexity issues see [11], §6.6.

(i) *There are nonnegative multipliers $y_1, \ldots, y_m$ (not all of them are zero) such that the matrix*

$$\sum_{i=1}^{m} y_i (A_i - \alpha_i I) \tag{5.28}$$

is positive semidefinite.

(ii) *There exist at most $k^{\mathbb{R}}(m,n)$ (see Theorem 5.12) vectors $x_1, \ldots, x_k \in \mathbb{R}^n$ such that*

$$\sum_{j=1}^{k^{\mathbb{R}}(m,n)} x_j^T A_i x_j \leq \alpha_i, \; \forall i = 1, \ldots, m. \tag{5.29}$$

REMARK 5.20. *System (5.29) will be discussed later in §6.4 in connection with rank constrained optimization.*

Finally, let us see what we can derive from Ramana's result about ICON maps (Theorem 5.3). His characterization of ICON maps is different in nature from the previous equivalent conditions. Ramana's conditions are more dependent on the actual data in the matrices. However, the result about LICON maps serves two goals. Firstly, it gives a polynomial time verifiable sufficient condition for the ICON property. Secondly, if $F : \mathbb{R}^n \to \mathbb{R}^m$ is a constant-free LICON map then we can include a set of linear constraints in the system. The following theorem is obtained:

THEOREM 5.21. *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a constant-free ($F(0) = 0$) quadratic map and $M \in \mathbb{R}^{n \times k}$ any matrix. If F is a LICON map (see Theorem 5.4 for an equivalent condition of this) then the following two statements are equivalent:*

(i) *The system*

$$F_i(x) \leq \alpha_i, \quad i = 1, \ldots, m \tag{5.30a}$$

$$Mx = 0 \tag{5.30b}$$

is not solvable.

(ii) *There is a nonzero vector $y = (y_1, \ldots, y_m) \in \mathbb{R}^m_+ \setminus \{0\}$ of nonnegative multipliers such that*

$$y^T (F(x) - \alpha) \geq 0, \quad \forall x : Mx = 0. \tag{5.31}$$

REMARK 5.22. *A special case of this result has also been presented in Proposition 3.9.*

5.3.2. Results over the complex field. The theorems in §5.2.2 can be applied in two ways. First we can easily prove the complex counterparts of all the theorems in the previous subsection. This way we can characterize the solvability of the system:

$$z^* A_i z \leq \alpha_i, \; i = 1, \ldots, m \tag{5.32a}$$

$$z^* z = 1, \tag{5.32b}$$

where A_i, $i = 1, \ldots, m$ are $n \times n$ complex Hermitian matrices. However, as our main topic is the solvability of real quadratic systems, the enumeration of all these results is out of the scope of this paper. We just briefly mention the strongest duality result here because we will refer to it later. It is an easy exercise to derive all the other results.

THEOREM 5.23. *Let $n \geq 3$, $A_i \in \mathbb{C}^n$, $i = 1, 2, 3$. The following two statements are equivalent:*

(i) *The system*

$$z^* A_i z = 0, \quad i = 1, 2, 3 \tag{5.33a}$$

$$z \neq 0 \tag{5.33b}$$

is not solvable.

(ii) *There are multipliers* y_1, y_2, y_3 *such that*

$$\sum_{i=1}^{3} y_i A_i \succ 0. \tag{5.34}$$

On the other hand the complex results have important consequences for real systems, too. Let A be a real symmetric matrix. If $z = x + \mathbf{i}y \in \mathbb{C}^n$ is a complex vector, then

$$z^* A z = (x - \mathbf{i}y)^T A (x + \mathbf{i}y) = x^T A x + y^T A y. \tag{5.35}$$

What we can get this way is a real quadratic system

$$x^T A_i x + y^T A_i y \leq \alpha_i,\ i = 1, \ldots, m \tag{5.36a}$$

$$\|x\|^2 + \|y\|^2 = 1. \tag{5.36b}$$

These equations are in strong relation with rank constrained optimization and will be discussed in more detail in §6.5.

5.4. Further extensions. The particular strength of this approach is that it can be applied to more general constraints. We can characterize the solvability of the system

$$(x^T A_1 x, \ldots, x^T A_m x) \in \mathcal{C} \tag{5.37a}$$

$$x \in \mathbb{R}^n \tag{5.37b}$$

$$(\|x\| = 1), \tag{5.37c}$$

where $\mathcal{C} \subset \mathbb{R}^m$ is a convex and possibly closed set. We can include the norm constraint, if we want to use the results about the joint numerical range. If this system is not solvable and the set of possible LHS vectors is convex then that set can be separated from $\mathcal{C}$ by a hyperplane. The actual form of these duality results depends on the form of $\mathcal{C}$. In the results derived so far $\mathcal{C}$ was the cone of nonnegative vectors. In what follows we present some examples when $\mathcal{C}$ is a polyhedron or a Lorentz cone. Further generalizations to balls, cubes, cones, etc., are also possible.

Let $\mathcal{C} \subset \mathbb{R}^m$ be a nonempty polyhedron defined by the inequalities $Mu \leq h$, where $M \in \mathbb{R}^{l \times m}$, $u \in \mathbb{R}^m$ and $h \in \mathbb{R}^l$. Consider the following system:

$$\left(x^T A_1 x, \ldots, x^T A_m x\right) = u \tag{5.38a}$$

$$Mu \leq h. \tag{5.38b}$$

We can prove the following theorem:

THEOREM 5.24. *Let* $A_1, A_2 \in \mathbb{R}^{n \times n}$ *be real symmetric matrices,* $x \in \mathbb{R}^n$, $M \in \mathbb{R}^{l \times 2}$, $h \in \mathbb{R}^l$. *Let us assume that the polyhedron* $\left\{u \in \mathbb{R}^2 : Mu \leq h\right\}$ *is not empty. The following two statements are equivalent:*

(i) *The quadratic system*

$$M_{i1}x^TA_1x + M_{i2}x^TA_2x \le h_i,\ i = 1,\dots,l \tag{5.39a}$$

$$x \in \mathbb{R}^n \tag{5.39b}$$

is not solvable.

(ii) *There exists a vector of nonnegative multipliers* $y = (y_1,\dots,y_m)$ *such that*

$$\sum_{i=1}^{l} y_i\left(M_{i1}A_1 + M_{i2}A_2\right) \succeq 0 \tag{5.40a}$$

$$y^Th < 0 \tag{5.40b}$$

$$y \ge 0. \tag{5.40c}$$

Proof. First let us assume that we have a vector y satisfying (5.40), then multiplying the inequalities in (5.39) by the corresponding multiplier and taking the sum we get that for any solution of (5.39)

$$0 \le x^T \underbrace{\left(\sum_{i=1}^{l} y_i\left(M_{i1}A_1 + M_{i2}A_2\right)\right)}_{\succeq 0} x \le y^Th < 0, \tag{5.41}$$

which is a contradiction.

Let us assume now that (5.39) is not solvable. In this case the image

$$\left\{(x^TA_1x, x^TA_2x) : x \in \mathbb{R}^n\right\} \tag{5.42}$$

and the polyhedron $\left\{u \in \mathbb{R}^2 : Mu \le h\right\}$ are nonempty, disjoint, convex sets, therefore they can be separated by a hyperplane, i.e., there exist multipliers z_1, z_2 such that

$$z_1x^TA_1x + z_2x^TA_2x \ge 0,\ \forall x \in \mathbb{R}^n \tag{5.43a}$$

$$z^Tu < 0,\ \forall u : Mu \le h. \tag{5.43b}$$

The first inequality states that a linear combination of the matrices is positive semidefinite, while the second one can be written in equivalent form using the well known Farkas Lemma (see, e.g., [19, 68, 72]). After these substitutions we get the statement of the theorem. □

REMARK 5.25. *There is another way to look at this theorem. All the inequalities in system (5.39) are of the form* $x^TB_ix \le h_i$ *where* $B_i = M_{i1}A_1 + M_{i2}A_2$*. In these terms what we proved is that the S-lemma remains true for the multi-inequality case, provided that all the matrices in the system are linear combinations of two matrices, see Prop. 3.5. This explains and generalizes an observation of Sturm and Zhang in §6 of [73].*

For a second example let us use Theorem 5.2 and assume that $\mathcal{C}$ is the three dimensional Lorentz cone, i.e., $\mathcal{C} = \left\{(u,v,w) \in \mathbb{R}^3 : u^2 \ge v^2 + w^2, u \ge 0\right\}$. This set is a closed, convex cone. We can define the dual cone of $\mathcal{C}$:

$$\mathcal{C}^* = \left\{(p,q,r) \in \mathbb{R}^3 : pu + qv + rw \ge 0, \forall (u,v,w) \in \mathcal{C}\right\}. \tag{5.44}$$

In what follows we will use the fact that the Lorentz cone is self dual, i.e., $\mathcal{C} = \mathcal{C}^*$, see [11] for details. We obtain the following theorem:

THEOREM 5.26. *Let* $n \ge 3$*,* $A, B, C \in \mathbb{R}^{n\times n}$ *be symmetric matrices and assume that they have a positive definite linear combination. The following two statements are equivalent:*

(i) *The system*

$$\left(x^T Ax\right)^2 \geq \left(x^T Bx\right)^2 + \left(x^T Cx\right)^2 \tag{5.45a}$$

$$x^T Ax \geq 0 \tag{5.45b}$$

$$x \neq 0 \tag{5.45c}$$

is not solvable.

(ii) *There exist multipliers* y_1, y_2, y_3 *such that*

$$y_1 A + y_2 B + y_3 C \prec 0 \tag{5.46a}$$

$$y_1^2 \geq y_2^2 + y_3^2 \tag{5.46b}$$

$$y_1 \geq 0. \tag{5.46c}$$

Proof. First let us assume that (5.45) is solvable and the multipliers in system (5.46) exist. Then for these solutions we have

$$y_1 x^T Ax + y_2 x^T Bx + y_3 x^T Cx < 0. \tag{5.47}$$

On the other hand both (y_1, y_2, y_3) and $(x^T Ax, x^T Bx, x^T Cx)$ are in a three dimensional Lorentz cone, and since the Lorentz cone is self dual, we have

$$y_1 x^T Ax + y_2 x^T Bx + y_3 x^T Cx \geq 0 \tag{5.48}$$

contradicting to (5.47).

Now assume that system (5.45) is not solvable. This means that

$$\left\{(x^T Ax, x^T Bx, x^T Cx) : x \in \mathbb{R}^n\right\} \cap \mathcal{C} = \{0\} \tag{5.49}$$

and since both sets are convex (the first one by Theorem 5.2, the other one by definition), they can be separated by a hyperplane going through the origin, i.e., there exist y_1, y_2, y_3 such that

$$y_1 u + y_2 v + y_3 w \geq 0, \quad \forall u, v, w : u^2 \geq v^2 + w^2,\ u \geq 0 \tag{5.50a}$$

$$y_1 x^T Ax + y_2 x^T Bx + y_3 x^T Cx < 0, \quad \forall x \neq 0. \tag{5.50b}$$

The first equation requires that (y_1, y_2, y_3) be in the dual cone of the Lorentz cone, but since the Lorentz cone is self dual this is equivalent to (y_1, y_2, y_3) being in a Lorentz cone. This shows that we have a solution for system (5.46). □

Similar theorems can be proved using other special forms of set $\mathcal{C}$. To our best knowledge these results have not been stated explicitly yet.

6. Rank constrained LMI. Looking at the second proof (see §2.3) of the S-lemma we can see that the crucial step is to show that the LMI relaxation of the system of quadratic inequalities is exact. This idea leads to the concept of rank-constrained optimization.

6.1. Motivation. Consider the homogeneous case

$$x^T Ax < 0 \tag{6.1a}$$

$$x^T B_i x \leq 0, \quad i = 1, \ldots, m \tag{6.1b}$$

and assume that the Slater condition is satisfied, i.e., there exists an $\bar{x} \in \mathbb{R}^n$ such that $\bar{x}^T B_i \bar{x} < 0$ for all $i = 1, \ldots, m$. Using the standard notation introduced in §2.3, this system is equivalent to the following LMI:

$$A \bullet Z < 0 \tag{6.2a}$$

$$B_i \bullet Z \leq 0, \quad i = 1, \ldots, m \tag{6.2b}$$

$$Z \succeq 0, \quad \operatorname{rank}(Z) = 1. \tag{6.2c}$$

After relaxing the condition on the rank

$$A \bullet Z < 0 \tag{6.3a}$$

$$B_i \bullet Z \leq 0, \quad i = 1, \ldots, m \tag{6.3b}$$

$$Z \succeq 0 \tag{6.3c}$$

we have to establish the following:

1. Prove the equivalence of the solvability of (6.3) and (6.2).
2. Prove that the Slater condition holds for (6.3).

The latter one is simple, let us take $Z = \bar{x}\bar{x}^T + \alpha I$ where $\bar{x}$ is the Slater-point of (6.1) and I is the identity matrix. If $\alpha > 0$ is small enough then all the linear constraints are satisfied while Z is positive definite. The former one is a more difficult problem and there are only few general results in that area. This is the subject of the next section.

After proving these two statements, we can apply the Farkas Theorem (see Theorem 2.1) to derive the dual equivalent of system (6.3):

$$A + \sum_{i=1}^{m} y_i B_i \succeq 0 \tag{6.4a}$$

$$y \geq 0. \tag{6.4b}$$

This is exactly the second part of the S-lemma. Putting the parts together we get that if the solvability of (6.3) implies the existence of a rank-1 solution then the S-procedure is exact. Let us examine when this can happen.

6.2. Theoretical results. Finding low rank solutions of linear matrix inequalities is a relatively new research field. The earliest general result is due to Pataki [57, 58], which was later discovered in other contexts, too.

THEOREM 6.1 (Pataki, [57, 58], 1994). *If $\mathcal{A} \subseteq \mathbb{S}^n$ is an affine subspace such that the intersection $\mathbb{PS}^n \cap \mathcal{A}$ is non-empty and $\dim(\mathcal{A}) \geq \binom{n+1}{2} - \binom{r+2}{2} + 1$ then there is a matrix $X \in \mathbb{PS}^n \cap \mathcal{A}$ such that $\operatorname{rank}(X) \leq r$.*

Proof. There are many possible ways to prove the theorem but the key observation is that any extreme point[6] of the intersection will have a sufficiently low rank. This also helps one to find such a matrix. The theorem is intuitively plausible: in order to have low rank matrices we have to intersect $\mathbb{PS}^n$ with a high dimensional subspace $\mathcal{A}$.

Let $X \in \mathbb{PS}^n \cap \mathcal{A}$ be an extreme point of the intersection, then we can assume without loss of generality that

$$X = \begin{pmatrix} X_{11} & 0 \\ 0 & 0 \end{pmatrix}, \tag{6.5}$$

[6] An extreme point of a convex set is a point from the set that is not an interior point of any line segment from the set. The intersection $\mathbb{PS}^n \cap \mathcal{A}$, as it does not contain a line, has extreme points. For more details on these issues, see [68, 72].

where X_{11} is positive definite. If $\operatorname{rank}(X) \le r$ then we have the requested low rank matrix, so let us assume that $\operatorname{rank}(X) \ge r+1$. As the dimension of $(r+1)\times(r+1)$ symmetric matrices is $\binom{r+2}{2}$ and X_{11} is constrained by at most $\binom{r+2}{2}-1$ linear equalities we have a nonzero matrix Y such that

$$Y \bullet A = 0 \quad \forall A \in \mathcal{A} \tag{6.6a}$$

$$Y = \begin{pmatrix} Y_{11} & 0 \\ 0 & 0 \end{pmatrix} \tag{6.6b}$$

$$Y_{11} \neq 0. \tag{6.6c}$$

Now $X \pm \varepsilon Y \in \mathbb{PS}^n \cap \mathcal{A}$ for small values of ε, thus X is not an extreme point of the intersection, contradicting to our assumption. This proves that X is of sufficiently low rank. □

REMARK 6.2. *The bound is sharp in the sense that if $r < n$ then one can find a subspace $\mathcal{A} \subseteq \mathbb{S}^n$ such that $\mathbb{PS}^n \cap \mathcal{A} \neq \emptyset$,* $\dim(\mathcal{A}) = \binom{n+1}{2} - \binom{r+2}{2}$ *and for every matrix $X \in \mathbb{PS}^n \cap \mathcal{A}$ we have* $\operatorname{rank}(X) > r$.

If $\mathcal{A}$ is nontrivial then the conditions of the theorem are obviously satisfied for $r = n-1$, implying the following simple corollary:

COROLLARY 6.3. *For any nontrivial (i.e., at least one dimensional) affine subspace $\mathcal{A}$, the intersection $\mathbb{PS}^n \cap \mathcal{A}$ (provided that it is not empty) contains a matrix X such that* $\operatorname{rank}(X) \le n-1$.

The result in Theorem 6.1 was generalized by Barvinok [9]:

THEOREM 6.4 (Barvinok, [9], 2001). *Let $r > 0$, $n \ge r+2$ and $\mathcal{A} \subseteq \mathbb{S}^n$ be an affine subspace such that the intersection $\mathbb{PS}^n \cap \mathcal{A}$ is non-empty,* bounded *and* $\dim(\mathcal{A}) = \binom{n+1}{2} - \binom{r+2}{2}$. *Then there is a matrix $X \in \mathbb{PS}^n \cap \mathcal{A}$ such that* $\operatorname{rank}(X) \le r$.

Barvinok's proof uses a differential geometric argument based on the structure of the cone of positive semidefinite matrices. There is a crucial difference between the two theorems. In the Pataki Theorem any extremal point of the intersection always has a sufficiently low rank, while in the Barvinok Theorem we can only guarantee that there is low rank extremal point. This has important algorithmic consequences: while it is easy to find a suitable low rank matrix for the Pataki Theorem, there is no known algorithm to find such a matrix for the Barvinok Theorem.

Barvinok was aware that this result is not new, but he is the first to state it and to provide a direct proof. He relates his result to a theorem of Au-Yeung and Poon [6] on the image of the sphere under a quadratic map. That result is discussed in §5.

Finally, both of the theorems in this section are extended for general symmetric matrices in [25] using Jordan algebraic techniques.

6.3. Implications. Now let us see what we can obtain using the above theorems. In view of the relaxation argument in §6.1, we are interested in rank-1 solutions, i.e., we use the theorems with $r = 1$.

Let us assume that system (6.3) is solvable, then we have a matrix Z such that

$$A \bullet Z = w_0 < 0 \tag{6.7a}$$

$$B_i \bullet Z = w_i \le 0, \quad i = 1, \dots, m \tag{6.7b}$$

$$Z \succeq 0. \tag{6.7c}$$

Here each equality corresponds to a hyperplane. Let $\mathcal{A}$ be the intersection of these hyperplanes, then

$$\dim(\mathcal{A}) \ge \binom{n+1}{2} - m - 1. \tag{6.8}$$

In order to apply Theorem 6.1 with $r = 1$ we need to have

$$\dim(\mathcal{A}) \geq \binom{n+1}{2} - 2, \tag{6.9}$$

so we must have $m \leq 1$, i.e., we can have at most one non-strict inequality. Then Theorem 6.1 guarantees the existence of a rank-1 solution to (6.3), thus the Pataki Theorem implies the basic S-lemma.

Notice that this way we proved slightly more than the existence of a rank-1 solution. Namely, we proved that the rank-1 matrix xx^T can be chosen such that

$$A \bullet xx^T = w_0 < 0 \tag{6.10a}$$

$$B_i \bullet xx^T = w_i \leq 0, \quad i = 1, \ldots, m \tag{6.10b}$$

$$xx^T \succeq 0, \tag{6.10c}$$

where w_i, $i = 0, \ldots, m$ are the same numbers as in system (6.7).

Now let us try to apply the Barvinok Theorem. Using the same setup and a similar argument as before, and setting $r = 1$, we have $m = 2$, so we can have two non-strict inequalities. However, there is an extra condition to satisfy, namely the solution set of

$$A \bullet Z = w_0 \tag{6.11a}$$

$$B_i \bullet Z = w_i, \quad i = 1, 2 \tag{6.11b}$$

$$Z \succeq 0 \tag{6.11c}$$

must be bounded and since $n \geq r + 2$ we must have $n \geq 3$.

Unfortunately the boundedness does not always hold, we need some extra condition to force it. It is well-known from convex analysis (see [68, 72]) that a convex set is unbounded if and only if it has a nontrivial recession direction, or, in other words, if it contains a half-line. Applying this to our case we get that the solution set of system (6.11) is bounded if and only if the following system is not solvable:

$$A \bullet Z = 0 \tag{6.12a}$$

$$B_i \bullet Z = 0,\ i = 1, 2 \tag{6.12b}$$

$$Z \succeq 0 \tag{6.12c}$$

$$Z \neq 0. \tag{6.12d}$$

Let us note that this system is independent of w, therefore the boundedness of the solution set of (6.11) does not depend on the actual choice of w_0 and w. Either all of them are bounded or all are unbounded (provided they are not empty).

Since this is an LMI it is easy to characterize its solvability. Namely, by the duality theory of LMIs (see [11], Proposition 2.4.2.), system (6.12) is not solvable if and only if there are real numbers λ_0, λ_1 and λ_2 such that $\lambda_0 A + \lambda_1 B_1 + \lambda_2 B_2$ is positive definite. This way we proved Proposition 3.6, i.e., the S-lemma with three inequalities.

6.4. Higher rank solutions. So far we have applied our theorems to the $r = 1$ case. However, the higher rank results also have some consequences for quadratic systems. Consider the following system:

$$\sum_{j=1}^{r} {x^j}^T A x^j < 0 \tag{6.13a}$$

$$\sum_{j=1}^{r} {x^j}^T B_i x^j \leq 0,\ i = 1, \ldots, m, \tag{6.13b}$$

where $x^j \in \mathbb{R}^n$, $j = 1, \ldots, r$. Introducing

$$X = \sum_{j=1}^{r} {x^j}^T x^j \tag{6.14}$$

and using the $\bullet$ notation this system can be written as:

$$A \bullet X < 0 \tag{6.15a}$$

$$B_i \bullet X \leq 0,\ i = 1, \ldots, m \tag{6.15b}$$

$$X \succeq 0 \tag{6.15c}$$

$$\operatorname{rank}(X) \leq r, \tag{6.15d}$$

since a positive semidefinite matrix X can always be decomposed as the sum of $\operatorname{rank}(X)$ positive semidefinite rank-1 matrices. Now relaxing the rank constraint we get an LMI identical to (6.3):

$$A \bullet X < 0 \tag{6.16a}$$

$$B_i \bullet X \leq 0,\ i = 1, \ldots, m \tag{6.16b}$$

$$X \succeq 0. \tag{6.16c}$$

Applying Theorems 6.1 and 6.4, and using a similar argument we obtain the following theorem:

THEOREM 6.5. *Let $A, B_1, \ldots, B_m \in \mathbb{R}^{n \times n}$ be symmetric matrices such that there are vectors $\bar{x}^1, \ldots, \bar{x}^r \in \mathbb{R}^n$ for which the Slater condition is satisfied, i.e.,*

$$\sum_{j=1}^{r} {\bar{x}^j}^T B_i \bar{x}^j < 0,\ i = 1, \ldots, m. \tag{6.17}$$

If

1. $m \leq \binom{r+2}{2} - 2$, *or*
2. $m = \binom{r+2}{2} - 1$, $n \geq r + 2$ *and there is a positive definite linear combination of* $A, B_1, \ldots, B_m$,

then the following two statements are equivalent:

(i) *The quadratic system*

$$\sum_{j=1}^{r} {x^j}^T A x^j < 0 \tag{6.18a}$$

$$\sum_{j=1}^{r} {x^j}^T B_i x^j \leq 0,\ i = 1, \ldots, m \tag{6.18b}$$

is not solvable.

(ii) *The LMI*

$$A + \sum_{i=1}^{m} y_i B_i \succeq 0 \tag{6.19a}$$

$$y_1, \ldots, y_m \geq 0 \tag{6.19b}$$

is solvable.

Now let $z = x^1 + \mathbf{i}x^2 \in \mathbb{C}^n$ be a complex vector, then for any real symmetric matrix A we have $z^*Az = {x^1}^T Ax^1 + {x^2}^T Ax^2$, therefore the above theorems can be used to decide the solvability of the following complex quadratic system, where A and $B_i, (i = 1, \ldots, m)$ are real symmetric matrices:

$$z^*Az < 0 \tag{6.20a}$$

$$z^*B_iz \leq 0,\ i = 1, \ldots, m. \tag{6.20b}$$

The value of m can be at most 3 without any further assumptions, or 4, if there is a positive definite linear combination of the matrices.

6.5. Rank constraints and convexity. We have seen that one can have more inequalities in the S-lemma in the complex case than in the real case. The reason for this is obvious from the previous section. Real solutions come from rank-1 solutions of system (6.3), while for complex solutions it is enough to have a rank-2 solution. This observation sheds some light on the convexity issues discussed in the previous section, particularly, it answers the question why the joint numerical range over the complex field possesses a much nicer structure and has a more straightforward characterization.

There is an interesting connection between the rank constraints and the convexity of the joint numerical range.

Recall Dines' result (Proposition 2.3): $\left\{(x^TAx, x^TBx) : x \in \mathbb{R}^n\right\}$ is convex. Let us take two points in the image, (u_1, u_2) and (v_1, v_2), and let $0 < \lambda < 1$. Now we have to find a point $x \in \mathbb{R}^n$ such that $x^TAx = \lambda u_1 + (1-\lambda)v_1$ and $x^TBx = \lambda u_2 + (1-\lambda)v_2$. Using the notations of this section we are looking for a rank-1 solution of the following system:

$$A \bullet X = \lambda u_1 + (1 - \lambda)v_1 \tag{6.21a}$$

$$B \bullet X = \lambda u_2 + (1 - \lambda)v_2. \tag{6.21b}$$

Using Theorem 6.1 we can see that this system always has a rank-1 solution, $X = xx^T$, thus we proved the convexity of the set $\left\{(x^TAx, x^TBx) : x \in \mathbb{R}^n\right\}$. The same argument can be applied to the $m = 2, n \geq 3$ case, then the Barvinok Theorem (Theorem 6.4) will yield Polyak's convexity result (Theorem 5.2). The connection works in both ways, from the convexity of the joint numerical range we can deduce rank-constraints for real or complex LMIs.[7] This is particularly useful in the complex case, since the theory of the complex numerical ranges has been investigated thoroughly, while there are no results for low-rank solutions of complex LMIs. On the other hand, there are very few results about the convexity of the image of the real space under higher rank quadratic maps.

Finally, we can contrast the Pataki Theorem for low rank matrices (Theorem 6.1) and Poon's Theorem on the number of terms in the convex combinations (Theorem 5.12). Although they are stated in different contexts they have some applications in common. Consider the system

$$A_i \bullet X = b_i, \quad i = 1, \ldots, m \tag{6.22a}$$

$$X \succeq 0, \tag{6.22b}$$

where $A_i, X \in \mathbb{R}^{n\times n}$ and assume that the system is solvable. Using Theorem 6.1 we get that there is a solution X such that

$$\operatorname{rank}(X) \leq \left\lfloor \frac{\sqrt{8m+1} - 1}{2} \right\rfloor = R_1(m, n), \tag{6.23}$$

[7] The authors wish to thank Gábor Pataki for the idea of the above argument.

and of course $\operatorname{rank}(X) \leq n$ also holds.

Now we try to get a similar bound from Theorem 5.12. Let X be a solution for system (6.22) and let us consider its rank-1 decomposition:

$$X = \sum_{j=1}^{r} \lambda^j x^j {x^j}^T, \tag{6.24}$$

where $\left\|x^j\right\| = 1$, and since X is positive semidefinite, we have $\lambda^j \geq 0$ for $j = 1, \ldots, m$. Let us define the following scaled quantities:

$$\lambda = \sum_{j=1}^{r} \lambda^j \tag{6.25a}$$

$$\bar{x}^j = \frac{x^j}{\lambda} \tag{6.25b}$$

$$\bar{X} = \frac{X}{\lambda} = \sum_{j=1}^{r} \frac{\lambda^j}{\lambda} x^j {x^j}^T, \tag{6.25c}$$

where the last line shows that $\bar{X}$ is a convex combination of the rank-1 matrices $x^j {x^j}^T$. Now let $\mathcal{A}$ denote the m-tuple $(A_1, \ldots, A_m)$ and let us use the notations of Theorem 5.12. By the definition of the image set we have

$$\left(A_1 \bullet \bar{x}^j {\bar{x}^j}^T, \ldots, A_m \bullet \bar{x}^j {\bar{x}^j}^T\right) \in W_1^{\mathbb{R}}(\mathcal{A}), \quad \forall j = 1, \ldots, r, \tag{6.26}$$

and using the decomposition result on $\bar{X}$ we get that

$$\left(A_1 \bullet \bar{X}, \ldots, A_m \bullet \bar{X}\right) \in \operatorname{conv}\left(W_1^{\mathbb{R}}(\mathcal{A})\right), \quad \forall j = 1, \ldots, r. \tag{6.27}$$

We can use Theorem 5.12 to deduce that there is a matrix $\widetilde{X}$ such that

$$A_i \bullet \widetilde{X} = A_i \bullet \bar{X}, \quad \forall j = 1, \ldots, m \tag{6.28a}$$

$$\widetilde{X} \succeq 0 \tag{6.28b}$$

$$\operatorname{rank}\left(\widetilde{X}\right) \leq \min\left\{n, \left\lfloor \frac{\sqrt{8m+1}-1}{2} \right\rfloor + \delta_{\frac{n(n+1)}{2},m}\right\} = R_2(m,n), \tag{6.28c}$$

and the matrix $\lambda\widetilde{X}$ solves system (6.22). Finally, if $m = \frac{n(n+1)}{2}$ then $R_2(m,n) = n$.

This shows that the two bounds are identical.

7. Generalized convexities. "Hidden convexity" (see [12]) seems to play an important role in the S-lemma. Although we made no convexity assumptions, the image of the quadratic map in §2.2 turned out to be convex. In this section we shed some more light on this interesting phenomenon.

7.1. Motivation. The fact that the S-lemma can be viewed as a non-convex generalization of the Farkas Theorem inspires us to look for a more general sense of convexity, which includes both the classical convex and the quadratic case. The convexity results in §5 show that even though the functions are not convex they describe a convex object, thus the problems admit some hidden convexity.

7.2. Theoretical results. There are many different convexity notions in the literature. Let us assume that X is a nonempty set, Y is a topological vector space over the real numbers with dual space Y^*, i.e., Y^* contains all the linear functions that map from Y to $\mathbb{R}$. If K is a cone in Y then let K^* denote its dual cone,

$$K^* = \{y^* \in Y^* : y^*(y) \geq 0, \forall y \in K\}. \tag{7.1}$$

Let us assume that K is a convex, closed, pointed[8] cone with nonempty interior. The partial orderings $\succeq_K$ and $\succ_K$ over Y are defined as follows:

$$y_1 \succeq_K y_2 \Leftrightarrow y_1 - y_2 \in K \tag{7.2a}$$
$$y_1 \succ_K y_2 \Leftrightarrow y_1 - y_2 \in \operatorname{int}(K). \tag{7.2b}$$

Let us note that the functions in this section are not necessarily quadratic. Using these notations we can define the following convexity notions:

DEFINITION 7.1. *A function $f : X \to Y$ is called*

Ky Fan convex, *if for every $x_1, x_2 \in X$ and $\lambda \in [0,1]$ there exists an $x_3 \in X$ such that*

$$f(x_3) \preceq_K (1-\lambda)f(x_1) + \lambda f(x_2). \tag{7.3}$$

König convex, *if for every $x_1, x_2 \in X$ there exists an $x_3 \in X$ such that*

$$f(x_3) \preceq_K \frac{1}{2}f(x_1) + \frac{1}{2}f(x_2). \tag{7.4}$$

K-convexlike, *if there exists a $\lambda \in (0,1)$ such that for every $x_1, x_2 \in X$ there exists an $x_3 \in X$ such that*

$$f(x_3) \preceq_K (1-\lambda)f(x_1) + \lambda f(x_2). \tag{7.5}$$

Further, in order to deal with both equalities and inequalities we can introduce the mixed versions of the above definitions.

DEFINITION 7.2. *Let Y and Z be locally convex[9] topological vector spaces and let X be any set. The function pair $(f, g) : X \to Y \times Z$ is called*

Ky Fan convex-linear, *if for each $x_1, x_2 \in X$ and $\lambda \in [0,1]$ there exists an $x_3 \in X$ such that*

$$f(x_3) \preceq_K (1-\lambda)f(x_1) + \lambda f(x_2) \tag{7.6a}$$
$$g(x_3) \;=\; (1-\lambda)g(x_1) + \lambda g(x_2), \tag{7.6b}$$

König convex-linear, *if for each $x_1, x_2 \in X$ there exists $x_3 \in X$ such that*

$$2f(x_3) \preceq_K f(x_1) + f(x_2) \tag{7.7a}$$
$$2g(x_3) \;=\; g(x_1) + g(x_2). \tag{7.7b}$$

The generality of these definitions is obvious if we notice that X can be any set, without any topology. Notice, e.g., that any continuous function mapping a compact set into $\mathbb{R}$ is König convex, since it has a minimum.

[8] A cone K is pointed if it does not contain a line.

[9] A topological vector space is locally convex if every point has a neighborhood basis consisting of open convex sets. Every normed space is locally convex.

Special cases of these definitions were first introduced by Ky Fan [24] and König [47]. Obviously, all Ky Fan convex functions are König convex, and all König convex functions are K-convexlike. However, there seems to be a large gap between the two extremities in these definitions. This gap is not that large, as it is shown in the next proposition (see [16]).

PROPOSITION 7.3. *If f is K-convexlike then the set of λ's satisfying the definition of Ky Fan convexity is dense in $[0,1]$.*

COROLLARY 7.4. *If f is continuous (or at least lower semicontinuous) then the convexity notions in Definition 7.1 coincide.*

Illés and his colleagues proved several versions of the Farkas Theorem for these convexities [40, 41, 42].

THEOREM 7.5 (Illés, Joó and Kassay, 1992, [40]). *Let $f : X \to Y$ be König convex, where Y is a locally convex space. If there is no $x \in X$ such that $f(x) \prec_K 0$ then there exists a $y^* \in Y^* \setminus \{0\}$ such that*

$$y^*(f(x)) \geq 0, \forall x \in X. \tag{7.8}$$

For the following theorem let Y_1 and Y_2 be two topological vector spaces over the reals, $K_1 \subseteq Y_1$ and $K_2 \subseteq Y_2$ are convex cones with vertices O_{Y_1} and O_{Y_2} such that $\operatorname{int} K_1 \neq \emptyset$. Let $f : X \to Y_1$ and $g : X \to Y_2$ be two functions and define $Y = Y_1 \times Y_2$, $K = K_1 \times K_2$ and $F = (f, g) : X \to Y$.

THEOREM 7.6 (Illés and Kassay, 1999, [42]). *Suppose $F = (f, g) : X \to Y$ is K-convexlike and the set $F(X) + K$ has nonempty interior. The following assertions hold:*

(i) *If there is no $x \in X$ such that*

$$f(x) \prec_K 0 \tag{7.9a}$$

$$g(x) \preceq_K 0 \tag{7.9b}$$

then there exist $y_1^* \in K_1^*$ and $y_2^* \in K_2^*$ (not both are the origin) such that

$$y_1^*(f(x)) + y_2^*(g(x)) \geq 0, \forall x \in X. \tag{7.10}$$

(ii) *If there exists an $y_1^* \in K_1^* \setminus \{O_{Y_1^*}\}$ and $y_2^* \in K_2^*$ such that (7.10) holds then system (7.9) is not solvable.*

The following theorem deals with systems containing equality constraints.

THEOREM 7.7 (Illés and Kassay, 1994, [41]). *Let Y and Z be locally convex topological vector spaces and let X be any set. Let $(f, g) : X \to Y \times Z$ be Ky Fan convex-linear with K and define*

$$M = \{(f(x) + v, g(x)) : x \in X,\ v \in K\} \subset Y \times Z. \tag{7.11}$$

If $\operatorname{int} M \neq \emptyset$ and there is no $x \in X$ such that

$$\begin{aligned} f(x) &\prec_K 0 \\ g(x) &= 0 \end{aligned}$$

then there exists $y^* \in K^*$ and $z^* \in Z^*$ (not both are the origin) such that

$$y^*(f(x)) + z^*(g(x)) \geq 0, \forall x \in X. \tag{7.12}$$

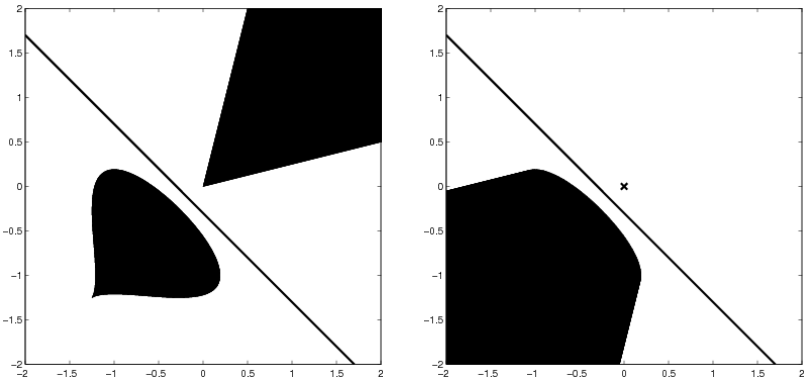

FIG. 7.1. *Separating a convex cone from a nonconvex set (left). The Minkowski sum of the set and the negative cone is separable from the origin (right).*

The proof of the above theorems relies on the following simple observation (see Fig. 7.1). In order to separate a convex cone from a set, the set does not have to be convex. It is enough to be convex on the side "facing the cone." More precisely, if K is a convex cone disjoint from set $\mathcal{C}$, and $\mathcal{C} + (-K)$ is convex then K and $\mathcal{C}$ can be separated by a hyperplane. Another view of this idea is that $K \cap \mathcal{C} = \emptyset$ if and only if $0 \notin \mathcal{C} + (-K)$. Separating this latter set from the origin is equivalent to separating K and $\mathcal{C}$.

For a summary on different convexity concepts see [2, 16] and the references therein.

7.3. Implications. It is straightforward to apply these theorems to our problems, we only need to verify the convexity assumptions. Notice however, that if we restrict ourselves to homogeneous quadratic functions then the above three notions coincide. Thus, e.g., in order to apply the results we need to prove that a pair of quadratic functions is König convex, i.e., for any $x_1, x_2 \in \mathbb{R}^n$ there exists an $x_3 \in \mathbb{R}^n$ such that

$$x_3^T A x_3 \leq \frac{1}{2} x_1^T A x_1 + \frac{1}{2} x_2^T A x_2 \tag{7.13a}$$

$$x_3^T B x_3 \leq \frac{1}{2} x_1^T B x_1 + \frac{1}{2} x_2^T B x_2. \tag{7.13b}$$

In fact, in Proposition 2.3 we proved that these two equations can be satisfied with equality, therefore a pair of quadratic functions is both König convex and König concave, also called König linear. This result seems to be new in the context of generalized convexities.

The three theorems presented in the previous section yield the following results:

THEOREM 7.8. *Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be homogeneous quadratic functions. The following two statements are equivalent:*

(i) *The system*

$$f(x) < 0 \tag{7.14a}$$

$$g(x) < 0 \tag{7.14b}$$

is not solvable.

(ii) *There exist nonnegative multipliers y_1 and y_2 (not both of them are zero) such that*

$$y_1 f(x) + y_2 g(x) \geq 0, \forall x \in \mathbb{R}^n. \tag{7.15}$$

This result is a variant of the S-lemma with two strict inequalities. Notice that the second statement immediately implies the first one, without any assumption on the functions.

THEOREM 7.9. *Suppose $f, g : \mathbb{R}^n \to \mathbb{R}$ are homogeneous quadratic functions. The following assertions hold:*

(i) *If there is no $x \in \mathbb{R}^n$ such that*

$$f(x) < 0 \tag{7.16a}$$
$$g(x) \leq 0, \tag{7.16b}$$

then there exist nonnegative multipliers y_1 and y_2 (not both are the origin) such that

$$y_1 f(x) + y_2 g(x) \geq 0, \forall x \in \mathbb{R}^n. \tag{7.17}$$

(ii) *If there exists a $y_1 > 0$ and $y_2 \geq 0$ such that (7.17) holds then there is no solution for (7.16).*

The gap in this theorem comes from that fact that we did not assume the Slater condition. If we further assume that there exists an $\bar{x} \in \mathbb{R}^n$ with $g(\bar{x}) < 0$ then the nonsolvability of (7.16) automatically implies the existence of multipliers in (ii).

Finally, let us see what we can get if we allow equality constraints.

THEOREM 7.10. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to Y$ be homogeneous quadratic functions, where Y is one of $\{0\}$, $\mathbb{R}$, $\mathbb{R}_+$ or $\mathbb{R}_-$. If there is no $x \in \mathbb{R}^n$ such that*

$$f(x) < 0 \tag{7.18a}$$
$$g(x) = 0 \tag{7.18b}$$

then there exist $y_1 \geq 0$ and $y_2 \in Y^*$ (not both are the origin) such that

$$y_1 f(x) + y_2 g(x) \geq 0, \forall x \in \mathbb{R}^n. \tag{7.19}$$

In fact, depending on the set Y, we obtained four theorems.

If $Y = \{0\}$ (i.e., $g(x) \equiv 0$) then the solvability of (7.18) is equivalent to the solvability of $f(x) < 0$.

If $Y = \mathbb{R}$, i.e., $g(x)$ takes both positive and negative values, then $Y^* = \{0\}$, thus $y_2 = 0$ and consequently $y_1 > 0$. This means that $f(x) \geq 0$ for all $x \in \mathbb{R}^n$. This result (at least for the special homogeneous case) is stronger than the one found in [73], because we showed that the multiplier of $g(x)$ is actually 0. In other words, if $g(x)$ takes both positive and negative values then system (7.18) is not solvable if and only if $f(x) < 0$ is not solvable, i.e., $f(x)$ is positive semidefinite.

If $Y = \mathbb{R}_+$ (or $Y = \mathbb{R}_-$) then $Y^* = \mathbb{R}_+$ ($Y^* = \mathbb{R}_-$) and $g(x) = 0$ is equivalent to $g(x) \leq 0$ ($g(x) \geq 0$), therefore the problem is reduced to Theorem 7.9.

8. Miscellaneous topics. In the last section before the summary we discuss some miscellaneous topics. These are all related to the S-lemma or, more generally, to systems of quadratic equations. Our goal with this section is not to give a full-detailed review of all these topics, rather to show the connections and raise some further questions.

8.1. Trust region problems. Trust region algorithms [18] are among the most successful methods to solve unconstrained nonlinear optimization problems. At each iteration of the algorithm we minimize a quadratic approximation of the objective function over a ball, called the trust region. Thus, given the current iterate $\hat{x}$, the trust region subproblem is

$$\min x^T H x + b^T x \tag{8.1a}$$
$$\|x - \hat{x}\|_2 \leq \alpha. \tag{8.1b}$$

In a slightly more general setting we can replace the (8.1b) ball-constraint with an ellipsoidal constraint:

$$\min x^T H x + b^T x \tag{8.2a}$$
$$(x - \hat{x})^T A (x - \hat{x}) \leq \alpha, \tag{8.2b}$$

where A is a symmetric positive semidefinite matrix, or, more generally, we can allow any matrix A. This way we get indefinite trust region subproblems [36, 37, 59, 71, 82]. These problems still can be solved in polynomial time, and that is what makes them suitable for building an algorithm. The reason behind the polynomial solvability is basically the S-lemma: a solution $\tilde{x}$ is optimal, if and only if the following system is not solvable:

$$x^T H x + b^T x < \tilde{x}^T H \tilde{x} + b^T \tilde{x} \tag{8.3a}$$
$$(x - \hat{x})^T A (x - \hat{x}) \leq \alpha. \tag{8.3b}$$

Using the S-lemma (Theorem 2.2) we can conclude that the optimality of $\tilde{x}$ is further equivalent to the solvability of the following system:

$$x^T H x + b^T x - \tilde{x}^T H \tilde{x} - b^T \tilde{x} + y \left((x - \hat{x})^T A (x - \hat{x}) - \alpha \right) \geq 0, \ \forall x \in \mathbb{R}^n \tag{8.4a}$$
$$y \geq 0, \tag{8.4b}$$

that can be homogenized and written as an LMI. Thus, using a simple bisection scheme we can solve the indefinite trust region subproblem to any given precision in polynomial time.

Moreover, we can build more complex trust regions, i.e., minimize a quadratic function over the intersection of two ellipsoids, or use one ellipsoidal and one general (possibly indefinite) constraint. Exact details and properties of these algorithms are to be developed.

8.2. SDP relaxation of QCQP. A broad range of optimization problems can be written as quadratically constrained quadratic programs. Recently, quite a number of articles considered the solvability of these problems. The research has two main branches. On one hand one might consider the cases when the SDP relaxation of general quadratic problems is exact [28, 44, 81], or, if the relaxation is not exact, then we can ask for bounds on the quality of the approximation [20, 55, 62, 75, 77]. Further, Kojima and Tunçel [45] propose a successive SDP relaxation scheme to solve these problems.

The exact relaxation results of this area are quite different in nature from the results discussed so far. The sufficient conditions for the validity of the S-lemma are usually structural conditions, i.e., they do not involve explicit data in the matrices. On the other hand, the sufficient conditions in the SDP relaxation theory are more

dependent on the problem data. The reason for this might be purely practical: as we presented so far, structural conditions allow only for a relatively small number of equations and inequalities.

The amount of results makes it impossible to include any of them here, the reader is thus referred to the references mentioned in the first paragraph.

8.3. Algebraic geometry. So far little was said about the algebraic nature of the problem, the discussion was more geometric and analytic. In this section we briefly review the algebraic connections.

The S-lemma is a surprising result from the point of view of algebraic geometry. Recall Hilbert's classical theorem, which can be found in any advanced algebra textbook, see, e.g., [10, 14, 30]:

THEOREM 8.1 (Nullstellensatz). *Let $p_1, \dots, p_m : \mathbb{C}^n \to \mathbb{C}$ be complex polynomials. The following two statements are equivalent:*

(i) *The system*

$$p_i(z) = 0, \quad i = 1, \dots, m \tag{8.5a}$$

$$z \in \mathbb{C}^n \tag{8.5b}$$

is not solvable.

(ii) *There exist complex polynomials $y_1(z), \dots, y_m(z)$ such that*

$$\sum_{i=1}^{m} y_i(z) p_i(z) \equiv -1. \tag{8.6}$$

This theorem holds only for the complex case. For real polynomials we have the following theorem:

THEOREM 8.2 (Real Nullstellensatz). *Let $p_1, \dots, p_m : \mathbb{R}^n \to \mathbb{R}$ be real polynomials. The following two statements are equivalent:*

(i) *The system*

$$p_i(x) = 0, \quad i = 1, \dots, m \tag{8.7a}$$

$$x \in \mathbb{R}^n \tag{8.7b}$$

is not solvable.

(ii) *There exist real polynomials $y_1(x), \dots, y_m(x)$ and $s_1(x), \dots, s_k(x)$ such that*

$$\sum_{j=1}^{k} s_j^2(x) + \sum_{i=1}^{m} y_i(x) p_i(x) \equiv -1. \tag{8.8}$$

If we know an *a priori* bound on the degree of the multipliers $y_1(x), \dots, y_m(x)$ (and the degree and number of $s_1(x), \dots, s_k(x)$ in the real case) then we can find them easily by equating the coefficients and solving a large linear system. Since any combinatorial optimization problem can be written as a system of polynomial equations, such bounds must be exponential in the number of variables. The best available degree bounds are summarized in the following theorem (for simplicity we are dealing with the complex case only).

THEOREM 8.3 (Effective Nullstellensatz). *Let $p_1, \dots, p_m : \mathbb{C}^n \to \mathbb{C}$ be complex polynomials with $\max_i \deg p_i = d$. If there exist complex polynomials $y_1(z), \dots, y_m(z)$ such that*

$$\sum_{i=1}^{m} y_i(z) p_i(z) \equiv -1, \tag{8.9}$$

then these polynomials can be chosen to satisfy

$$\max_i \deg y_i \leq \begin{cases} d^n & \text{if } d \geq 3 \text{ (see [46])} \\ 2^{\min\{m,n\}} & \text{if } d = 2 \text{ (see [70]).} \end{cases} \tag{8.10}$$

These bounds are essentially sharp, better estimates use some additional information about the polynomials, such as the geometric degree [69], the sparsity [70] or the height [33], just to mention a few. It would be interesting to specialize those bounds for quadratic systems and also to derive similar results for the real case. More details can also be found in [13].

In the simple case of two complex quadratic polynomials Theorem 8.3 gives a rather weak corollary (compared with Theorem 5.23). In this regard it is surprising that under the conditions of the S-lemma, the multiplier polynomials can be chosen to be constants. This also suggests further directions for the research on degree bounds in the Nullstellensatz.

8.4. Computational complexity. Let $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$ be quadratic functions. Using more advanced techniques (see [8, 31]) one can decide the solvability of a system

$$f_i(x) = 0, \quad i = 1, \ldots, m \tag{8.11a}$$

$$\|x\|_2 = 1 \tag{8.11b}$$

$$x \in \mathbb{R}^n, \tag{8.11c}$$

and a solution can also be obtained. These problems include most combinatorial optimization problems thus in general we can not hope for a polynomial time algorithm. However, there are some special cases. If $m = 1$ then solving system (8.11) is equivalent to determining the definiteness of $f_1(x)$, which can be done in polynomial time. Further, the S-lemma gives rise to a polynomial time algorithm if $m = 2$.

In [31] Grigoriev and Pasechnik propose an algorithm to solve system (8.11). The algorithm is polynomial in n and exponential in m. Whence, if m is fixed, or at least bounded, then (8.11) can be solved in polynomial time.

9. Summary. We have seen through several different approaches and examples that Yakubovich's S-lemma is only the tip of the iceberg. The theories that lead to this result are much more general and the S-lemma has a wealth of applications in various areas of applied mathematics. We demonstrated that the generalization for more inequalities is not practical, as the conditions we need to assume are more and more complex. In fact, the minimal conditions under which the S-lemma holds with more inequalities can be obtained easily from the characterization theorems for convex images in §5. This answers a question posed as open in [21]. The results and counterexamples collected in this paper suggest that the basic theory is well established but some more general topics deserve investigation.

9.1. Future research. To finish this survey we present some problems and projects that are worth some further research.

SOCO relaxations: Remaining in the quadratic world, second order conic (SOCO) relaxations offer an alternative way to treat quadratic systems. This procedure is similar to semidefinite relaxation, which was discussed in §2.3 and §6. Currently, there are not many results on when a SOCO relaxation of a quadratic system is exact.

Real numerical range: A characterization theorem for the convexity of the joint numerical range over the real numbers (similar to Theorem 5.13) could be developed.

Polynomial systems: The question arises whether it is possible to extend the concept of the S-lemma to polynomial or even more general functions. Since any polynomial system can be written equivalently as a quadratic system with more equations and variables, this case essentially reduces to the quadratic case, but finding these results and providing suitable applications remains a challenging task.

Convex images: The convexity of the image of a set under some transformation is crucial to this theory. However, there are hardly any results about the convexity of the image of sets under a nonlinear transformation, Polyak's local convexity result (Theorem 5.10) is one of them. In [67] Ramana poses the same question for general maps, but then only deals with the quadratic case. This is an open and mainly untouched research field. Generalized convexities (see §7) offer another framework for this research.

Algebraic methods: Leaving the concept of the S-lemma and concentrating on more general related results we find that polynomial and sum of squares (SOS) optimization (see [66]) have recently drawn a lot of interest both in the optimization and in the control community. This research pointed the attention of optimizers towards algebraic geometric results, e.g., the Nullstellensatz (see §8.3). In our opinion, since polynomials are well-studied in advanced algebra, more research should be carried out in this direction. On the other hand, the example of the S-lemma shows that under some conditions stronger degree bounds are possible in the effective Nullstellensatz.

Applications: As of today, there are no practical applications of the S-lemma with more than two inequalities. This might be due to the fact that nonconvex quadratic systems are usually believed to be difficult and people often use alternative formulations instead. After all, the fact that the S-lemma allows for polynomial time solvability (see §8.4) of certain quadratic systems should be promoted. Similarly, one can look for applications of Theorem 5.18 or Theorem 5.26.

Acknowledgment. The authors were supported by the NSERC Discovery Grant #5-48923, the Canada Research Chair Program and a MITACS project. The authors would like to thank Etienne de Klerk, Jean-Baptiste Hiriart-Urruty, Dima Pasechnik and Gábor Pataki for their comments and suggestions during the development of this paper.

We also thank the anonymous referees and the editor, Michael Overton, for their corrections and their constructive approach throughout the refereeing process.

REFERENCES

[1] M. A. Aizerman and F. R. Gantmacher, *Absolute Stability of Regulator Systems*, Holden-Day Series in Information Systems, Holden-Day, Inc, San Francisco, 1964. Originally published as Absolutnaya Ustoichivost' Reguliruyemykh Sistem, by The Academy of Sciences of the USSR, Moscow, 1963.

[2] A. Aleman, *On some generalizations of convex sets and convex functions*, L'Analyse Numérique et la Théorie de l'Approximation, 14 (1985), pp. 1–6.

[3] M. Anitescu, *Degenerate nonlinear programming with a quadratic growth condition*, SIAM Journal on Optimization, 10 (2000), pp. 1116–1135.

[4] ———, *A superlinearly convergent sequential quadratically constrained quadratic program-*

ming algorithm for degenerate nonlinear programming, SIAM Journal on Optimization, 12 (2002), pp. 949–978.

[5] Y.-H. Au-Yeung, *A theorem on a mapping from a sphere to the circle and the simultaneous diagonalization of two hermitian matrices*, Proceedings of the AMS, 20 (1969), pp. 545–548.

[6] Y.-H. Au-Yeung and Y. T. Poon, *A remark on the convexity and positive definiteness concerning Hermitian matrices*, Southeast Asian Bulletin of Mathematics, 3 (1979), pp. 85–92.

[7] Y.-H. Au-Yeung and N.-K. Tsing, *An extension of the Hausdorff-Toeplitz theorem on the numerical range*, Proceedings of the AMS, 89 (1983), pp. 215–218.

[8] A. I. Barvinok, *Feasibility testing for systems of real quadratic equations*, Discrete & Computational Geometry, 10 (1993), pp. 1–13.

[9] ———, *A remark on the rank of positive semidefinite matrices subject to affine constraints*, Discrete & Computational Geometry, 25 (2001), pp. 23–31.

[10] S. Basu, R. Pollack, and M.-F. Roy, *Algorithms in Real Algebraic Geometry*, Springer-Verlag, 2003.

[11] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*, MPS-SIAM Series on Optimization, SIAM, Philadelphia, PA, 2001.

[12] A. Ben-Tal and M. Teboulle, *Hidden convexity in some nonconvex quadratically constrained quadratic programming*, Math. Programming, 72 (1996), pp. 51–63.

[13] C. Berenstein and D. Struppa, *Recent improvements in the complexity of the effective Nullstellensatz*, Linear Algebra and Its Applications, 157 (1991), pp. 203–215.

[14] J. Bochnak, M. Coste, and M.-F. Roy, *Real Algebraic Geometry*, Springer-Verlag, Berlin, 1998.

[15] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory*, SIAM Studies in Applied Mathematics, SIAM, Philadelphia, PA, 1994.

[16] W. W. Breckner and G. Kassay, *A systemization of convexity concepts for sets and functions*, Journal of Convex Analysis, 4 (1997), pp. 109–127.

[17] L. Brickman, *On the field of values of a matrix*, Proceedings of the AMS, 12 (1961), pp. 61–66.

[18] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust-Region Methods*, MPS-SIAM Series on Optimization, SIAM, Philadelphia, PA, 2000.

[19] E. de Klerk, C. Roos, and T. Terlaky, *Nonlinear Optimization*. Lecture notes, Delft University of Technology, Delft, The Netherlands, 2003.

[20] K. Derinkuyu and M. C. Pinar, *An improved approximate S-lemma*. Working paper, Bilkent University, Ankara, Turkey, January 2005.

[21] ———, *On the S-procedure and some variants*, Mathematical Methods of OR, (2005). To appear.

[22] L. L. Dines, *On the mapping of quadratic forms*, Bulletin of the AMS, 47 (1941), pp. 494–498.

[23] ———, *On linear combinations of quadratic forms*, Bulletin of the AMS, 49 (1943), pp. 388–393.

[24] K. Fan, *Minimax theorems*, Proceedings of the National Academy of Sciences, (1953), pp. 42–47.

[25] L. Faybusovich, *Jordan-algebraic approach to convexity theorems for quadratic mappings*, SIAM Journal on Optimization, 17 (2006), pp. 558–576.

[26] P. Finsler, *Über das Vorkommen definiter und semidefiniter Formen in Scharen quadratischer Formen*, Commentaria Mathematicae Helvetia, 9 (1937), pp. 188–192.

[27] A. L. Fradkov and V. A. Yakubovich, *The S-procedure and a duality relation in nonconvex problems of quadratic programming*, Vestnik Leningrad University, 5 (1979), pp. 101–109. Originally in Russian in 1973.

[28] T. Fujie and M. Kojima, *Semidefinite programming relaxation for nonconvex quadratic programs*, Journal of Global Optimization, 10 (1997), pp. 367–380.

[29] D. Goldfarb and G. Iyengar, *Robust portfolio selection problems*, Mathematics of OR, 28 (2003), pp. 1–38.

[30] P. Griffiths and J. Harris, *Principles of Algebraic Geometry*, Wiley Classics Library, John Wiley & Sons, 1994.

[31] D. Grigoriev and D. V. Pasechnik, *Polynomial-time computing over quadratic maps I: sampling in real algebraic sets*, Computational Complexity, 14 (2005), pp. 20–52.

[32] E. Gutkin, E. A. Jonckheere, and M. Karow, *Convexity of the joint numerical range: topological and differential geometric viewpoints*, Linear Algebra and Its Applications, 376 (2004), pp. 143–171.

[33] K. Hagele, J. E. Morais, L. M. Pardo, and M. Sombra, *On the intrinsic complexity of the arithmetic Nullstellensatz*, Journal of Pure and Applied Algebra, 146 (2000), pp. 103–183.

[34] F. Hausdorff, *Der Wertvorrat einer Bilinearform*, Math. Zentralblatt, 3 (1919), pp. 314–316.
[35] M. R. Hestenes and E. J. McShane, *A theorem on quadratic forms and its application in the calculus of variations*, Transactions of the AMS, 47 (1940), pp. 501–512.
[36] J.-B. Hiriart-Urruty, *Conditions for global optimality 2*, Journal of Global Optimization, 13 (1998), pp. 349–367.
[37] ———, *Global optimality conditions in maximizing a convex quadratic function under convex quadratic constraints*, Journal of Global Optimization, 21 (2001), pp. 445–455.
[38] A. E. Hoerl and R. W. Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 12 (1970), pp. 55–67.
[39] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, 1991.
[40] T. Illés, I. Joó, and G. Kassay, *On a nonconvex Farkas theorem and its application in optimization theory*, Report 1992-03, Eötvös University, Budapest, Hungary, 1992.
[41] T. Illés and G. Kassay, *Farkas type theorems for generalized convexities*, Pure Mathematics and Applications, 5 (1994), pp. 225–239.
[42] ———, *Theorems of the alternative and optimality conditions for convexlike and general convexlike programming*, Journal of Optimization Theory and Applications, 101 (1999), pp. 243–257.
[43] U. T. Jönsson, *A lecture on the S-procedure*. Division of Optimization and Systems Theory, Royal Institute of Technology, Stockholm, Sweden, May 2001.
[44] S. Kim and M. Kojima, *Exact solutions of some nonconvex quadratic optimization problems via SDP and SOCP relaxations*, Computational Optimization and Applications, 26 (2003), pp. 143–154.
[45] M. Kojima and L. Tunçel, *On the finite convergence of successive SDP relaxation methods*, European Journal of Operations Research, 143 (2002), pp. 325–341.
[46] J. Kollár, *Sharp effective Nullstellensatz*, Journal of the AMS, 1 (1988), pp. 963–975.
[47] H. König, *Über das von Neumannsche Minimax-Theorem*, Archiv der Mathematik, 19 (1968), pp. 482–487.
[48] J. Levin, *Mathematical models for determining the intersections of quadric surfaces*, Computer Graphics and Image Processing, 11 (1979), pp. 73–87.
[49] C.-K. Li and Y.-T. Poon, *Convexity of the joint numerical range*, SIAM Journal on Matrix Analysis and Applications, 21 (2000), pp. 668–678.
[50] Z.-Q. Luo, *Applications of convex optimization in signal processing and digital communication*, Math. Programming (Series B), 97 (2003), pp. 177–207.
[51] Z.-Q. Luo, J. F. Sturm, and S. Zhang, *Multivariate nonnegative quadratic mappings*, SIAM Journal on Optimization, 14 (2004), pp. 1140–1162.
[52] A. I. Lur'e and V. N. Postnikov, *On the theory of stability of control systems*, Prikl. Mat. i Mekh., 8 (1944), pp. 3–13.
[53] A. Megretsky, *S-procedure in optimal non-stochastic filtering*. Unpublished manuscript, without year.
[54] A. Megretsky and S. Treil, *Power distribution in optimization and robustness of uncertain systems*, Journal of Mathematical Systems, Estimation, and Control, 3 (1993), pp. 301–319.
[55] A. Nemirovski, C. Roos, and T. Terlaky, *On maximization of quadratic forms over intersection of ellipsoids with common center*, Math. Programming, 86 (1999), pp. 463–473.
[56] R. Ortega, *An energy amplification condition for decentralized adaptive stabilization*, IEEE Transactions on Automatic Control, 41 (1996), pp. 285–288.
[57] G. Pataki, *On the facial structure of cone-LP's and semidefinite programs*, Management Science Research Report MSRR-#595, Graduate School of Industrial Administration, Carnegie Mellon University, 1994.
[58] ———, *On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues*, Mathematics of Operations Research, 23 (1998), pp. 339–358.
[59] J.-M. Peng and Y.-X. Yuan, *Optimality conditions for the minimization of a quadratic with two quadratic constraints*, SIAM Journal on Optimization, 7 (1997), pp. 579–594.
[60] H. Pépin, Revue de la filière Mathématiques (RMS), 3 (2004), pp. 171–172. Answer to a problem posed by J.-B. Hiriart-Urruty. In French.
[61] E. Pesonen, *Über die Sprektraldarstellung quadratischer Formen in linearen Räumen mit indefiniter Metrik*, vol. 227 of Annales Academiae Scientiarum Fennicae Series A. I. Mathematica, 1956.
[62] M. C. Pinar and M. Teboulle, *On semidefinite bounds for maximization of a non-convex quadratic objective over the ℓ_1 unit ball*, RAIRO - Operations Research, (2006). To appear.
[63] B. T. Polyak, *Convexity of quadratic transformations and its use in control and optimization*, Journal of Optimization Theory and Applications, 99 (1998), pp. 553–583.

[64] ———, *Convexity of nonlinear image of a small ball with applications to optimization*, Set-Valued Analysis, 9 (2001), pp. 159–168.
[65] Y. T. Poon, *On the convex hull of the multiform numerical range*, Linear and Multilinear Algebra, 37 (1994), pp. 221–223.
[66] A. Prestel and C. N. Delzell, *Positive Polynomials*, Springer-Verlag, Berlin, 2001.
[67] M. Ramana and A. J. Goldman, *Quadratic maps with convex images*, Report 36-94, Rutgers Center for Operations Research, Rutgers, The State University of New Jersey, 1994.
[68] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, N. J., 1970.
[69] M. Sombra, *Bounds for the Hilbert function of polynomial ideals and for the degrees in the Nullstellensatz*, Journal of Pure and Applied Algebra, 117&118 (1991), pp. 565–559.
[70] ———, *A sparse effective Nullstellensatz*, Advances in Applied Mathematics, 22 (1999), pp. 271–295.
[71] R. J. Stern and H. Wolkowicz, *Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations*, SIAM Journal on Optimization, 5 (1995), pp. 286–313.
[72] J. Stoer and C. Witzgall, *Convexity and Optimization in Finite Dimensions*, vol. I, Springer-Verlag, Heidelberg, 1970.
[73] J. F. Sturm and S. Zhang, *On cones of nonnegative quadratic functions*, Mathematics of Operations Research, 28 (2003), pp. 246–267.
[74] O. Toeplitz, *Das algebraische Analogon zu einem Satze von Fejér*, Math. Zentralblatt, 2 (1918), pp. 187–197.
[75] P. Tseng, *Further results on approximating nonconvex quadratic optimization by semidefinite programming relaxation*, SIAM Journal on Optimization, 14 (2003), pp. 268–283.
[76] K. Weierstrass, *Zur Theorie der bilinearen und quadratischen Formen*, Monatsberichte der Königl. Preuss. Akademie der Wissenschaften zu Berlin, (1868), pp. 310–338.
[77] D. Xu and S. Zhang, *Approximation bounds for quadratic maximization with semidefinite programming relaxation*, Technical Report SEEM 2003-01, Department of Systems Engineering & Engineering Management, The Chinese University of Hong Kong, 2003.
[78] V. A. Yakubovich, *S-procedure in nonlinear control theory*, Vestnik Leningrad University, 1 (1971), pp. 62–77. In Russian.
[79] ———, *Minimization of quadratic functionals under quadratic constraints and the necessity of a frequency condition in the quadratic criterion for absolute stability of nonlinear control systems*, Soviet Math. Doklady, 14 (1973), pp. 593–597.
[80] ———, *S-procedure in nonlinear control theory*, Vestnik Leningrad University, 4 (1977), pp. 73–93. English translation.
[81] Y. Ye and S. Zhang, *New results on quadratic minimization*, SIAM Journal on Optimization, 14 (2003), pp. 245–267.
[82] Y.-X. Yuan, *On a subproblem of trust region algorithms for constrained optimization*, Math. Programming, 47 (1990), pp. 53–63.