Appl Math Optim 45:169–184 (2002)
DOI: 10.1007/s00245-001-0034-6

Applied Mathematics & Optimization

Permanently Going Back and Forth between the "Quadratic World" and the "Convexity World" in Optimization

J.-B. Hiriart-Urruty[1] and M. Torki[2]

[1]Université Paul Sabatier,
118 route de Narbonne, 31062 Toulouse Cedex, France
jbhu@cict.fr

[2]Université d'Avignon,
339 chemin des Meinajaries, 84911 Avignon, France
torki@iup.univ-avignon.fr

Abstract. The objective of this work is twofold. Firstly, we propose a review of different results concerning convexity of images by quadratic mappings, putting them in a chronological perspective. Next, we enlighten these results from a geometrical point of view in order to provide new and comprehensive proofs and to immerse them in a more general and abstract context.

Key Words. Quadratic functions, Range convexity, Joint positive definiteness.

AMS Classification. 90C20, 52A20, 15A60.

Introduction

The "quadratic" character and the "convexity" one seem to belong to completely different worlds in mathematics; although they are old and well known, quadratic mappings and convex sets still continue to be objects of active research. It happens there are unexpected but very interesting results of convexity concerning quadratic mappings. One of the goals in this paper is to review the main ones, putting them in a chronological perspective (Sections 1 and 2).

Section 1 deals with the convexity of images by quadratic mappings; we display there results by Dines (1941), Brickman (1961), and Barvinok (1995), amongst others. The first results in this area seem due to Dines and Brickman: Dines showed that, for

any real symmetric matrices A and B, the image set

$$\{(\langle Ax, x\rangle, \langle Bx, x\rangle) \mid x \in \mathbb{R}^n\} \tag{1}$$

is a convex cone in $\mathbb{R}^2$, whatever the dimension n. Brickman established a finer (and trickier) result: for any real symmetric matrices A and B,

$$\{(\langle Ax, x\rangle, \langle Bx, x\rangle) \mid \|x\| = 1\} \tag{2}$$

is a convex compact set in $\mathbb{R}^2$, whenever $n \geq 3$. We will see with the results of Barvinok what the extensions of the two previous results are when m (≥ 3) quadratic functions are involved. These convexity results on images have immediate consequences in fields where quadratic functions appear naturally and play an essential part, for example, in optimization and control theory (more specifically in an infinite-dimensional context in the latter case). For instance, the famous so-called *$\mathcal{S}$-procedure*, which deals with the nonnegativity of a quadratic form on a set described by quadratic inequalities, provides a powerful tool for proving stability of nonlinear control systems. In optimization, the so-called *trust region subproblem*, which consists in minimizing a quadratic function subject to a norm constraint, arises in solving general nonlinear programs. This problem like some other nonconvex quadratic optimization problems enjoys a *hidden convexity property*. All that is closely related and even due to the convexity of images by quadratic mappings.

In Section 2 of the paper, we explore conditions under which there exists linear (or convex) combinations of given matrices A_i's which are positive (semi-)definite. We present there results by Finsler (1936), Calabi (1964), Yuan (1978), and Barvinok (1995), amongst others. As we will see, all these results are related to those displayed in Section 1. As a general rule, in Section 1 as well as in Section 2, the results dealing with *two* quadratic mappings are extended to the case of m (≥ 3) quadratic mappings by substituting *matrices* of specific size to the usual (vector) variables $x \in \mathbb{R}^n$.

The convexity of the image sets in (1) and (2) is essentially a consequence of the special facial structure of the convex cone of symmetric positive semidefinite matrices. In Section 3 we generalize some of the results presented in Sections 1 and 2 by adopting a geomerical viewpoint in an abstract context of euclidean spaces. Indeed, we exploit a characterization of the faces of the intersection of two convex sets in order to prove the convexity of the following set:

$$\{(\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle) \mid x \in \mathcal{F}_m(K)\}, \tag{3}$$

where $a_1, \ldots, a_m$ are elements of the euclidean space $(E, \langle\cdot,\cdot\rangle)$, K is a closed convex pointed cone in E, and $\mathcal{F}_m(K)$ is the union of all faces of K whose dimension is at most m. Then, using a separation argument, we derive a "*facial alternative theorem*" for linear mappings. For the particular case where K is the cone of positive semidefinite matrices, we retrieve Dines's and Yuan's theorems as well as their extensions by Barvinok.

1. Convexity of Images by Quadratic Mappings

We begin by fixing some notations which are used throughout. We work in a finite-dimensional (euclidean) setting, say $\mathbb{R}^n$, equipped with the standard inner product denoted by $\langle\cdot,\cdot\rangle$ and the associated norm denoted by $\|\cdot\|$. However, some remarks on what

happens on quadratic mappings on a Hilbert space (for example, convexity of images is, to a certain extent, easier to obtain than in a finite-dimensional setting), are given.

$\mathcal{S}_n(\mathbb{R})$ denotes the space of real symmetric (n, n)-matrices.

When $A \in \mathcal{S}_n(\mathbb{R})$, $A \succ 0$ (resp. $A \succeq 0$) is the notation used to express that A is positive definite (resp. positive semidefinite). Associated with $A \in \mathcal{S}_n(\mathbb{R})$ is the *quadratic form* q on $\mathbb{R}^n$ defined as $q\colon \mathbb{R}^n \ni x \mapsto q(x) := \langle Ax, x\rangle$. By a quadratic function f on $\mathbb{R}^n$ we mean $f\colon \mathbb{R}^n \ni x \mapsto f(x) := \langle Ax, x\rangle + \langle b, x\rangle + c$ where $A \in \mathcal{S}_n(\mathbb{R})$, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$.

Finally, a quadratic mapping $q = (q_1, \ldots, q_m)^T\colon \mathbb{R}^n \to \mathbb{R}^m$ is a mapping whose component functions q_i are all quadratic forms.

In the space $M_{r,s}(\mathbb{R})$ of real (r, s)-matrices, we adopt the usual inner product $\langle\langle U, V\rangle\rangle :=$ trace of $U^T V$.

It is worth noticing from now on that if $X \in M_{n,p}(\mathbb{R})$ and $A \in \mathcal{S}_n(\mathbb{R})$, then $\langle\langle AX, X\rangle\rangle = \langle\langle A, XX^T\rangle\rangle$; in particular, for $p = 1$ (X is then the one column matrix associated with $x \in \mathbb{R}^n$), $\langle\langle A, xx^T\rangle\rangle = \langle Ax, x\rangle$.

All our results are exposed in the "real setting," nothing is said concerning theorems of the Toeplitz–Hausdorff type for quadratic forms with complex (n, n)-matrices.

1.1. *The Theorem of Dines* (1941)

The following series of results by Dines [10] seem to be the first ones concerning the convexity of images of sets by *two* quadratic forms.

Let A and B be in $\mathcal{S}_n(\mathbb{R})$. Then

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

is a convex *cone of* $\mathbb{R}^2$ (*denoted as* K). *If, moreover,*

$$\begin{pmatrix} \langle Ax, x\rangle = 0 \\ \textit{and} \\ \langle Bx, x\rangle = 0 \end{pmatrix} \Rightarrow (x = 0), \tag{4}$$

then K is a closed *cone, which is either the whole $\mathbb{R}^2$ or a cone with "angle" $\theta < \pi$.*

1.2. *The Theorem of Brickman* (1961)

The next convexity type result, due to Brickman [6] is better spread than the previous one.

Let A and B be in $\mathcal{S}_n(\mathbb{R})$, and assume $n \geq 3$. Then

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

is a convex *compact subset of* $\mathbb{R}^2$.

When $n = 2$, the image of the unit sphere by two quadratic forms may fail to be convex, we explain why later. Here is a counterexample [6, p. 63]:

$$A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \in \mathcal{S}_2(\mathbb{R});$$

then

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

is the unit circle of $\mathbb{R}^2$.

1.3. *Extension of Dines's Theorem by Barvinok* (1995)

When more than two quadratic forms are involved, say $q_1, \ldots, q_m$ with $m \geq 3$, the image of the whole space by $q_1, \ldots, q_m$ may fail to be convex. Here is a simple counterexample. Let

$$A = \begin{bmatrix} 1 & -1 \\ -1 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & -1 \\ -1 & 1 \end{bmatrix}, \quad \text{and} \quad C = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \in \mathcal{S}_2(\mathbb{R}).$$

Then

$$\{(\langle Ax, x\rangle, \langle Bx, x\rangle, \langle Cx, x\rangle) \mid x \in \mathbb{R}^2\} := \mathcal{R}$$

is not convex since the intersection of $\mathcal{R}$ with the plane $x_3 = 0$ is

$$\{(\alpha, 0, 0) \mid \alpha \geq 0\} \cup \{(0, \beta, 0) \mid \beta \geq 0\}.$$

Under some special relationships between the dimension n and the number m of quadratic forms involved, one can, however, get an upper bound on the number r of elements needed to yield, via convex combinations, the elements of the convex hull of the image $\mathcal{R}$ of the unit sphere by $(q_1, \ldots, q_m)$ [4, Theorem 1.4]. That still shows the limitation due to the fact that "usual" quadratic forms on $\mathbb{R}^n$ are considered.

More general convexity results on images can be obtained if one accepts "relaxing" or "enriching" the space on which quadratic forms are considered, say enlarging $\mathbb{R}^n \equiv M_{n,1}(\mathbb{R})$ to some $M_{n,p}(\mathbb{R})$. The following result by Barvinok [2] can be viewed as an extension of Dines's one in Section 1.

Let $A_1, A_2, \ldots, A_m \in \mathcal{S}_n(\mathbb{R})$, and let $p := \lfloor(\sqrt{8m+1} - 1)/2\rfloor$ ($\lfloor x \rfloor$ stands for the "integer part of x"). Then

$$\{(\langle\langle A_1 X, X\rangle\rangle, \langle\langle A_2 X, X\rangle\rangle, \ldots, \langle\langle A_m X, X\rangle\rangle) \mid X \in M_{n,p}(\mathbb{R})\}$$

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

The definition of p may look mysterious at first glance, some explanation is provided later in Section 3. Note, however, that $m = 2$ yields $p = 1$ (Dines's situation in Section 1), but $m = 3$ yields $p = 2$.

1.4. *Extension of Brickman's Theorem by Poon* (1997)

The next result by Poon [22] extends Brickman's theorem in Section 1.2.

Let $A_1, A_2, \ldots, A_m \in \mathcal{S}_n(\mathbb{R})$, and let the integer p be defined as follows:

$$p := \begin{cases} \left\lfloor \dfrac{\sqrt{8m+1}-1}{2} \right\rfloor & \text{if} \quad \dfrac{n(n+1)}{2} \neq m+1, \qquad (*) \\ \left\lfloor \dfrac{\sqrt{8m+1}-1}{2} \right\rfloor + 1 & \text{if} \quad \dfrac{n(n+1)}{2} = m+1. \qquad (**) \end{cases}$$

Then

$$\{(\langle\langle A_1 X, X\rangle\rangle, \langle\langle A_2 X, X\rangle\rangle, \ldots, \langle\langle A_m X, X\rangle\rangle) \mid X \in M_{n,p}(\mathbb{R}), \|X\|_{\mathrm{F}} = 1\}$$

is a convex *compact subset of $\mathbb{R}^m$ (here $\|\cdot\|_{\mathrm{F}}$ denotes the Schur–Frobenius norm on $M_{n,p}(\mathbb{R})$, that is the one derived from $\langle\langle\cdot,\cdot\rangle\rangle$).*

Observe that when $m = 2$ (two quadratic forms involved), the first case $(*)$ occurs for $n \geq 3$, and this gives rise to $p = 1$ (that is Brickman's case in Section 1.2). Still for $m = 2$, the second case $(**)$ occurs for $n = 2$, which imposes us to take $p = 2$.

2. Existence of a Linear (or Convex) Combination of $A_1, A_2, \ldots, A_m$ Which Is Positive (Semi-)Definite

The first results we present below in Sections 2.1 and 2.2, concerning the existence of a linear (or convex) combination of A and B which is positive definite (or semidefinite), have been derived independently of any convexity result on images by quadratic forms. They are, however, directly linked to such convexity results (in Sections 1.1 and 1.2) as we will see. This is even clearer in the case where more than two quadratic forms are involved.

2.1. *The Theorem of Finsler* (1936) *and Calabi* (1964)

The following result was proved by Finsler [11], and rediscovered by Calabi [7].

Let A and B be in $\mathcal{S}_n(\mathbb{R})$, and assume $n \geq 3$. Then the following are equivalent:

$$\text{(i)} \quad \begin{pmatrix} \langle Ax, x\rangle = 0 \\ \textit{and} \\ \langle Bx, x\rangle = 0 \end{pmatrix} \Rightarrow (x = 0). \qquad (5)$$

(ii) *There exists $\mu_1, \mu_2 \in \mathbb{R}$ such that $\mu_1 A + \mu_2 B \succ 0$.* (6)

The implication (5) $\Rightarrow$ (6) does not necessarily hold true when $n = 2$; a counterexample is the one already considered in Section 1.2:

$$q_1(x_1, x_2) := x_1^2 - x_2^2 \qquad \text{associated with} \quad A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

$$q_2(x_1, x_2) := 2x_1x_2 \qquad \text{associated with} \quad B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix};$$

then (5) holds true but there is no way to have

$$\mu_1 A + \mu_2 B = \begin{bmatrix} \mu_1 & \mu_2 \\ \mu_2 & -\mu_1 \end{bmatrix}$$

positive definite.

There is another statement, a more classical one, which is known to be related to (5)–(6); it is expressed in terms of simultaneous diagonalization of A and B via congruence. Recall that a collection $\{A_1, A_2, \ldots, A_m\}$ of matrices in $\mathcal{S}_n(\mathbb{R})$ is said to be *simultaneously diagonalizable via congruence*, if there exists a nonsingular matrix P such that each of the $P^T A_i P$ is diagonal. Simultaneous diagonalization via congruence corresponds to transforming the quadratic forms q_i associated with the A_i's into linear combinations of squares by a single linear change of variables; it is a more accessible property than the (usual) diagonalization via similarity. It is known that (6) *implies*

A and B are simultaneously diagonalizable via congruence. (7)

See Section 7.6 of [15] for example. We note also that the implication (5) $\Rightarrow$ (7) is given on pp. 272–280 of [12]; the proof, Milnor's one, clearly shows that the hypothesis $n \geq 3$ on the dimension of the underlying space is essential. A very good account of Finsler–Calabi type results, including the historical developments, remains the survey paper [24].

We propose below a simple proof of the theorem of Finsler and Calabi by using the convexity result of Brickman exposed in Section 1.2.

Let A and B be in $\mathcal{S}_n(\mathbb{R})$ and assume $n \geq 3$. We know from Brickman's theorem that

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

is a convex compact set in $\mathbb{R}^2$. What (5) expresses is that $0 \notin \mathcal{R}$. Thus, referring to a classical argument in Convex analysis, we can "separate" 0 from $\mathcal{R}$: there is a line strictly separating $\{0\}$ from $\mathcal{R}$, i.e., there exists $(\mu_1, \mu_2) \in \mathbb{R}^2$ and $r \in \mathbb{R}$ such that

$$\langle (\mu_1, \mu_2), (u, v)\rangle > r > 0 \qquad \text{for all} \quad (u, v) \in \mathcal{R}. \tag{8}$$

Now, (8) is nothing more than

$$\mu_1\langle Ax, x\rangle + \mu_2\langle Bx, x\rangle > 0 \qquad \text{for all} \quad x \in \mathbb{R}^n, \quad \|x\| = 1, \tag{9}$$

that is (6).

Among various attempts to extend the above stated theorems, we mention two recent contributions:

- A *vectorial* version of the Finsler–Calabi theorem [13, Theorem 2].
- A result analogous to the equivalence (5)–(6) for three quadratic forms by Polyak [21]; the main result [21, Theorem 2.1] can be formulated as follows:

Let A, B, and C be in $\mathcal{S}_n(\mathbb{R})$, and assume $n \geq 3$. Then the two next assertions are equivalent:

(i) $$\begin{pmatrix} \langle Ax, x\rangle = 0 \\ \langle Bx, x\rangle = 0 \\ \quad \textit{and} \\ \langle Cx, x\rangle = 0 \end{pmatrix} \Rightarrow (x = 0), \tag{10}$$

and $\mathcal{K} := \{(\langle Ax, x\rangle, \langle Bx, x\rangle, \langle Cx, x\rangle) \mid x \in \mathbb{R}^n\}$ is a pointed closed convex cone ($\mathcal{K}$ pointed means $\mathcal{K} \cap (-\mathcal{K}) = \{0\}$).

(ii) *There exist $\mu_1, \mu_2, \mu_3 \in \mathbb{R}$ such that $\mu_1 A + \mu_2 B + \mu_3 C \succ 0$.* (11)

Moreover, this superb paper [21] contains various applications in optimization and control theory of the ongoing convexity results on images by quadratic forms or functions.

We end this section by noting that the simultaneous diagonalization via congruence of $A_1, A_2 \ldots, A_m \in \mathcal{S}_n(\mathbb{R})$ makes it possible to generalize, to some extent, the aforementioned results (due to the fact that the image set $\mathcal{K}$ in that case is a polyhedral closed convex cone). This property of the A_i's was already set as an assumption in [16]. However, the following problem remains posed:

Question 1. Find sensible and "palpable" conditions on $A_1, A_2, \ldots, A_m$ ensuring they are simultaneously diagonalizable via congruence.

2.2. *The Theorem of Yuan* (1990)

Condition (5) can be formulated in a variational fashion as follows:

$$\max\{|\langle Ax, x\rangle|, |\langle Bx, x\rangle|\} > 0 \qquad \text{for all} \quad x \neq 0 \quad \text{in } \mathbb{R}^n. \tag{12}$$

Yuan's theorem [26] that we present below is just a "unilateral" version of the equivalence (5)–(6).

Let A and B be in $\mathcal{S}_n(\mathbb{R})$. Then the following are equivalent:

(i) $\max\{\langle Ax, x\rangle, \langle Bx, x\rangle\} \geq 0 \qquad$ *for all* $\quad x \in \mathbb{R}^n$

(*resp.* > 0 *for all* $x \neq 0$ *in* $\mathbb{R}^n$). (13)

(ii) *There exists $\mu_1 \geq 0, \mu_2 \geq 0, \mu_1 + \mu_2 = 1$ such that $\mu_1 A + \mu_2 B \succeq 0$*

(*resp.* $\succ 0$). (14)

Again here, starting from Dines's convexity result displayed in Section 1, one can derive Yuan's theorem via a separation argument. What (13) says is that the (convex) image set $\mathcal{K} := \{(\langle Ax, x\rangle, \langle Bx, x\rangle) \mid x \in \mathbb{R}^n\}$ does not meet the open convex cone $\mathcal{N} := \{(\alpha, \beta) \in \mathbb{R}^2 \mid \alpha < 0 \text{ and } \beta < 0\}$. There therefore exists a line passing through the origin just separating them, i.e., there exists a nonnull $(\mu_1, \mu_2) \in \mathbb{R}^2$ such that

$$\langle(\mu_1, \mu_2), (u, v)\rangle \geq 0 \qquad \text{for all} \quad (u, v) \in \mathcal{K}, \tag{15}$$

$$\langle(\mu_1, \mu_2), (\alpha, \beta)\rangle \geq 0 \qquad \text{for all} \quad (\alpha, \beta) \in \mathcal{N}. \tag{16}$$

Then (15) yields that $\mu_1 A + \mu_2 B$ is positive semidefinite, while (16) ensures that $\mu_1 \geq 0$ and $\mu_2 \geq 0$.

Yuan's result does not hold true if three matrices A, B, and C in $\mathcal{S}_n(\mathbb{R})$ are involved; counterexamples are given in papers [9] and [8], where further analysis of results à la Yuan is developed.

We end this section by posing two open questions.

Question 2. Let $A_1, A_2, \ldots, A_m \in \mathcal{S}_n(\mathbb{R})$. How to express equivalently, in terms of the A_i's, the following assertion:

$$\begin{pmatrix} \langle A_1 x, x\rangle = 0 \\ \langle A_2 x, x\rangle = 0 \\ \vdots \\ \langle A_m x, x\rangle = 0 \end{pmatrix} \Rightarrow (x = 0)? \tag{17}$$

To have a linear combination of the A_i's positive definite indeed secures (17), but it is too strong a sufficient condition, by far.

The "unilateral" version of the question above is the next one.

Question 3. Let $A_1, A_2, \ldots, A_m \in \mathcal{S}_n(\mathbb{R})$. How to express equivalently, in terms of the A_i's, the following situation:

$$\max\{\langle A_1 x, x\rangle, \langle A_2 x, x\rangle, \ldots, \langle A_m x, x\rangle\} \geq 0 \qquad \text{for all} \quad x \in \mathbb{R}^n? \tag{18}$$

It is interesting to note that (18) is related to the field of necessary/sufficient conditions in nonsmooth optimization: the nonsmooth function $x \mapsto f(x) := \max\{\langle A_i x, x\rangle, i = 1, \ldots, m\}$ is globally minimized at $\bar{x} = 0$. However, expressing some first or generalized second-order necessary condition for minimality for f at $\bar{x} = 0$ does not give any interesting information about the A_i's.

To have a convex combination of the A_i's positive semidefinite is obviously a sufficient condition for (18), but not a necessary one.

2.3. *Extension of Section* 2.2*, or Corollary to Barvinok's Theorem in Section* 1.3

As one can easily imagine now, the convexity result in Section 1.3 yields, via separation techniques from convex analysis, the next statement.

Let $A_1, A_2, \ldots, A_m \in \mathcal{S}_n(\mathbb{R})$, *and let* $p := \lfloor(\sqrt{8m+1}-1)/2\rfloor$ ($m = 2$ *gives* $p = 1$, $m = 3$ *gives* $p = 2$, *etc*.). *Then the following are equivalent*:

(i) $\max\{\langle\langle A_1X, X\rangle\rangle, \langle\langle A_2X, X\rangle\rangle, \ldots, \langle\langle A_mX, X\rangle\rangle\} \geq 0$ *for all* $X \in M_{n,p}(\mathbb{R})$

(*resp.* > 0 *for all* $X \neq 0$ *in* $M_{n,p}(\mathbb{R})$). (19)

(ii) *There exists* $\mu_1, \ldots, \mu_m \geq 0$, $\mu_1 + \cdots + \mu_m = 1$ *such that*

$$\sum_{i=1}^{m} \mu_i A_i \succeq 0 \qquad (\textit{resp.} \succ 0). \tag{20}$$

2.4. *Extension of Section* 2.1, *or Corollary to the Result in Section* 1.4, *by Bohnenblust* [5]

Let $A_1, A_2, \ldots, A_m \in \mathcal{S}_n(\mathbb{R})$, *and let*

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

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

(i) $$\begin{pmatrix} \langle\langle A_1X, X\rangle\rangle = 0, \\ \langle\langle A_2X, X\rangle\rangle = 0, \\ \vdots \quad X \in M_{n,p}(\mathbb{R}) \\ \langle\langle A_mX, X\rangle\rangle = 0, \end{pmatrix} \Rightarrow (X = 0). \tag{21}$$

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

$$\sum_{i=1}^{m} \mu_i A_1 \succ 0. \tag{22}$$

Remark 1. It is natural to try to extend results of Sections 1 or 2 to the case where quadratic functions are defined on a (general) Hilbert space. The main motivation for that comes from control theory where many questions can be formulated in abstract form as the problem of minimizing a quadratic function on a closed convex subset of a Hilbert space (usually described as quadratic inequality constraints). Among the various results à la Dines–Brickman (see [25], [20], [18], [19], [1]), we single out the following typical one by Matveev [18], [19]:

Let $A_1, A_2, \ldots, A_m$ *be self-adjoint continuous linear mappings on a* (*real*) *Hilbert space* $(H, \langle\cdot,\cdot\rangle)$, *let* $q_1, \ldots, q_m$ *denote the associated continuous quadratic forms on* H (*i.e.*, $q_i(x) := \langle A_ix, x\rangle$ *for all* $x \in H$). *If, for any* $(\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$, *the maximal and the*

minimal points of the spectrum of $\lambda_1 A_1 + \cdots + \lambda_m A_m$ are not isolated eigenvalues of finite geometric multiplicity, then the range $\mathcal{R}$ of the unit sphere of H under the quadratic mapping $q := (q_1, \ldots, q_m)$ is almost convex, i.e., there exists a convex set C of $\mathbb{R}^m$ such that $C \subset \mathcal{R} \subset \bar{C}$.

3. Some General Results in a Euclidean Space Context

In this section we generalize some of the results presented in Sections 1 and 2 by adopting a geometrical viewpoint in an abstract context of euclidean spaces.

Our general setting here is: *a euclidean space $(E, \langle\cdot,\cdot\rangle)$ and a closed convex cone K in E*. We keep in mind, however, the guiding example where $E = \mathcal{S}_n(\mathbb{R})$ is endowed with the inner product $\langle\langle\cdot,\cdot\rangle\rangle$, and $K = \{A \in \mathcal{S}_n(\mathbb{R}) \mid A \succeq 0\}$.

We begin with some definitions and technical results.

- $B \subset K$ is called a *basis* for K when the following holds true: for all $x \in K\backslash\{0\}$, there exists a unique pair $(\lambda > 0, y \in B)$ such that $x = \lambda y$.
- The (positive) *polar cone* of K is defined as follows:

$$K^+ := \{x \in E \mid \langle x, d\rangle \geq 0 \text{ for all } d \in K\}.$$

Proposition 1. *The following statements are equivalent*:

(i) *K possesses a compact basis.*
(ii) *The interior of K^+ is nonempty.*
(iii) *K is pointed* (*i.e.*, $K \cap (-K) = \{0\}$).

The next theorem, the so-called "faces of intersection theorem" by Dubins and Klee (see p. 116 of [23] for example), serves as our main technical tool.

Theorem 2. *Let C_1 and C_2 be two closed convex sets in E. Then F is a face of $C_1 \cap C_2$ if and only if there exists a face F_1 of C_1 and a face F_2 of C_2 such that $F = F_1 \cap F_2$. Moreover, F_1 and F_2 can be chosen such that*

$$\text{Aff } F = \text{Aff } F_1 \cap \text{Aff } F_2.$$

In that case

$$\text{codim } F \leq \text{codim } F_1 + \text{codim } F_2.$$

(Aff F *stands for the affine hull of* F, *and* codim F *is the codimension of* F.)

More specifically, Theorem 2 will be used when $F = \{\bar{x}\}$, $\bar{x}$ is an extreme point of $C_1 \cap C_2$, and C_2 is an affine subspace or halfspace of E.

3.1. *Convexity of Images of Faces of K by Linear Mappings*

Our first result in this section is as follows.

Theorem 3. *Assume K is pointed and consider $a_1, a_2, \ldots, a_m \in E$. Then*

$$\{(\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle) \mid x \in K\} = \{(\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle) \mid x \in \mathcal{F}_m(K)\}, \tag{23}$$

where $\mathcal{F}_m(K)$ denotes the union of all faces of K whose dimension is $\leq m$. The same result holds true if a convex compact subset C of E is substituted for K.

As an immediate consequence of Theorem 3, we get the convexity of the images of the faces of K (or C) whose dimension is $\leq m$ by linear mappings $x \in E \mapsto (\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle) \in \mathbb{R}^m$.

Corollary 1. *Under the assumptions of Theorem* 3, *the image set*

$$\{(\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle) \mid x \in \mathcal{F}_m(K)\},$$

is a convex cone of $\mathbb{R}^m$. Similarly, if C is a convex compact subset of E, then

$$\{(\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle) \mid x \in \mathcal{F}_m(C)\}$$

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

Proof. Let $(\alpha_1, \ldots, \alpha_m) \in \{(\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle) \mid x \in K\}$ and consider the following affine subspace:

$$V := \{x \in E \mid \langle a_i, x\rangle = \alpha_i \text{ for all } i = 1, \ldots, m\}. \tag{24}$$

Since $K \cap V$ is nonempty (by definition of V from $(\alpha_1, \ldots, \alpha_m)$), we consider $x_0 \in K \cap V$. Let us prove that $K \cap V$ possesses extreme points.

Since K is assumed pointed, the interior of K^+ is nonempty (Proposition 1). Whenever a_0 is chosen in the interior of K^+, it is easy to check that

$$B := K \cap \{x \in E \mid \langle a_0, x\rangle = 1\}$$

is a compact basis for K.

Let $\sigma := \langle a_0, x_0\rangle (\sigma > 0)$. The set

$$C := K \cap V \cap \{x \in E \mid \langle a_0, x\rangle \leq 2\sigma\}$$

is clearly closed convex and nonempty (it contains x_0). It is also bounded; indeed

$$\begin{aligned} C &= V \cap \{\lambda y\colon\ \lambda \geq 0, y \in B, \langle a_0, \lambda y\rangle \leq 2\sigma\} \\ &= V \cap \{\lambda y\colon\ 0 \leq \lambda \leq 2\sigma, y \in B\} \qquad (\langle a_0, y\rangle = 1 \text{ whenever } y \in B), \end{aligned}$$

whence the boundedness of C follows from that of B.

C does have extreme points, but we claim that there is at least one extreme point $\bar{x}$ in C satisfying $\langle a_0, \bar{x}\rangle < 2\sigma$. If not, all the extreme points in C (and therefore the whole of C) would lie in

$$K \cap V \cap \{x \in E \mid \langle a_0, x\rangle = 2\sigma\},$$

which is not possible since $x_0 \in C$ while $\langle a_0, x_0\rangle = \sigma < 2\sigma$.

We thus take such an extreme point $\bar{x}$ of C. Write C as $C_1 \cap C_2$ where $C_1 := K \cap V$, $C_2 := \{x \in E \mid \langle a_0, x\rangle \leq 2\sigma\}$, and consider the face $F := \{\bar{x}\}$ built up from $\bar{x}$. According to the "faces of intersection theorem" (Theorem 2), there exists a face F_1 of $K \cap V$, a face F_2 of $\{x \in E \mid \langle a_0, x\rangle \leq 2\sigma\}$ such that

$$\{\bar{x}\} = F_1 \cap F_2$$

and

$$\dim E = \operatorname{codim} F \leq \operatorname{codim} F_1 + \operatorname{codim} F_2. \tag{25}$$

Now, since $\langle a_0, \bar{x}\rangle < 2\sigma$, the called up face F_2 is nothing more than $C_2 = \{x \in E \mid \langle a_0, x\rangle \leq 2\sigma\}$ itself; whence $\operatorname{codim} F_2 = 0$. As a consequence, it comes from the inequality (25) that $\operatorname{codim} F_1 = \dim E$, which means that $\bar{x}$ is an extreme point of $C_1 = K \cap V$.

We now apply again the "faces of intersection theorem" to $C_1 = K \cap V$; there therefore exists a face F_K of K, a face F_V of V such that

$$\{\bar{x}\} = F_V \cap F_K$$

and

$$\dim E = \operatorname{codim}\{\bar{x}\} \leq \operatorname{codim} F_V + \operatorname{codim} F_K. \tag{26}$$

Since V is the affine subspace of E defined in (24), $F_V = V$ and $\operatorname{codim} F_V \leq m$. It then follows from (26) that $\operatorname{codim} F_K \geq \dim E - m$, that is to say $\dim F_K \leq m$. Thus

$$\{\bar{x}\} = V \cap F_K \text{ is nonempty;}$$

in other words, V intersects a face of K whose dimension is $\leq m$. We finally have proved that

$$(\alpha_1 \ldots, \alpha_m) \in \{(\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle) \mid x \in \mathcal{F}_m(K)\},$$

i.e.,

$$\{(\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle) \mid x \in K\} \subset \{(\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle) \mid x \in \mathcal{F}_m(K)\}.$$

Since the converse inclusion is trivial, the announced equality is proved. □

3.2. *Facial Alternative Theorems*

The standard alternative theorem based upon results in convex analysis expresses the equivalence of the two following statements:

(i) $\max\{(\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle)\} \geq 0$ for all $x \in K$. (27)

(ii) There exists $\mu_1, \ldots, \mu_m \geq 0$, $\mu_1 + \cdots + \mu_m = 1$, such that

$$\sum_{i=1}^{m} \mu_i a_i \in K^+. \tag{28}$$

Therefore, as a by-product of the convexity result in Theorem 3, we get a first *"facial alternative theorem"* (*linear case*).

Theorem 4. *Assume K is pointed and consider $a_1, \ldots, a_m \in E$. Then the following are equivalent*:

(i) $\max\{(\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle)\} \geq 0 \qquad$ *for all* $\quad x \in \mathcal{F}_m(K)$

(*resp.* > 0 *for all* $x \neq 0 \in \mathcal{F}_m(K)$). (29)

(ii) *There exists* $\mu_1, \ldots, \mu_m \geq 0$, $\mu_1 + \cdots + \mu_m = 1$ *such that*

$$\sum_{i=1}^{m} \mu_i a_i \in K^+ \qquad (\textit{resp.} \in \operatorname{int}(K^+)). \tag{30}$$

Remark 2. The above alternative theorem remains valid if one substitutes $\mathcal{F}_{\bar{m}}(K)$ for $\mathcal{F}_m(K)$ (for the $\geq$ inequality), where $\bar{m}$ denotes the rank of the family $\{a_1, \ldots, a_m\}$ (thus $\bar{m} \leq m$). This is not true for the strict inequality. In this case, one can substitute $\mathcal{F}_{\underline{m}}(K)$ for $\mathcal{F}_m(K)$ where $\underline{m} := \min\{\bar{m} + 1, m\} (\leq m)$.

Remark 3. The above result is sharper than the "boundary-type alternative theorem" proposed by Crouzeix et al. [9].

As a corollary of Theorem 4 we have the following "facial alternative theorem" in an *affine* form.

Corollary 2. *Assume K is pointed and consider $a_1, \ldots, a_m \in E$ and $(c_1, \ldots, c_m) \in \mathbb{R}^m \backslash \{(0, \ldots, 0)\}$. Then the following are equivalent*:

(i) $\max\{\langle a_1, x\rangle + c_1, \ldots, \langle a_m, x\rangle + c_m)\} \geq 0 \qquad$ *for all* $\quad x \in \mathcal{F}_m(K)$

(*resp.* > 0 *for all* $x \neq 0 \in \mathcal{F}_m(K)$). (31)

(ii) *There exists* $\mu_1, \ldots, \mu_m \geq 0$, $\mu_1 + \cdots + \mu_m = 1$ *such that*

$$\sum_{i=1}^{m} \mu_i a_i \in K^+ \quad \textit{and} \quad \sum_{i=1}^{m} \mu_i c_i \geq 0 \qquad (\textit{resp.} > 0). \tag{32}$$

Another "facial alternative theorem," the *convex form*, follows by linearization from Theorem 4.

Theorem 5. *Assume K is pointed and consider convex functions $f_1, f_2, \ldots, f_m \colon E \to \mathbb{R}$. We assume that, for all $i = 1, \ldots, m$, f_i is differentiable at 0 and satisfies $f_i(0) = 0$. Then the following are equivalent*:

(i) $\max\{f_1(x), \ldots, f_m(x)\} \geq 0$ *for all* $x \in \mathcal{F}_m(K)$. (33)

(ii) *There exists* $\mu_1, \ldots, \mu_m \geq 0$, $\mu_1 + \cdots + \mu_m = 1$ *such that*

$$\left(\sum_{i=1}^{m} \mu_i f_i\right)(x) \geq 0 \qquad \textit{for all} \quad x \in K. \tag{34}$$

Proof. The first-order necessary and sufficient condition for optimality in convex optimization [14, Chapter VII, Section 1] allows us to reformulate (ii) in the following equivalent form:

(ii)′ There exists $\mu_1, \ldots, \mu_m \geq 0$, $\mu_1 + \cdots + \mu_m = 1$ such that

$$\sum_{i=1}^{m} \mu_i \nabla f_i(0) \in K^+.$$

According to Theorem 5, condition (ii)′ above is equivalent to:

(i)′ $\max\{\langle \nabla f_1(0), x\rangle, \ldots, \langle \nabla f_m(0), x\rangle\} \geq 0$ for all $x \in \mathcal{F}_m(K)$.

Now, using elementary properties of nonsmooth convex functions like

$$\max\{f_1(x), \ldots, f_m(x)\} \geq \max\{\langle \nabla f_1(0), x\rangle, \ldots, \langle \nabla f_m(0), x\rangle\}$$

(remember that $f_i(0) = 0$ for all $i = 1, \ldots, m$),

$$\max\{\langle \nabla f_1(0), x\rangle, \ldots, \langle \nabla f_m(0), x\rangle\} = \inf_{t>0} \frac{\max\{f_1(tx), \ldots, f_m(tx)\}}{t}$$

(see Chapter VI, Section 4.4, of [14]), and the conical structure of $\mathcal{F}_m(K)$, the equivalence between (i)′ and (i) is derived. □

3.3. *Facial Solutions to Simultaneous Linear Equations*

Still under the assumptions of Theorem 4 we have the following variant of Theorem 3:

$$\{(\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle) \mid 0 \neq x \in \mathcal{F}_{m+1}(K)\}$$
$$= \{(\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle) \mid 0 \neq x \in K\}. \tag{35}$$

Proof. First, since K is pointed, it possesses a compact basis B (by Proposition 1). Thus, we can write

$$K \setminus \{0\} = \{\lambda y \mid \lambda > 0, y \in B\}.$$

As a consequence,

$$\{(\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle) \mid 0 \neq x \in K\} = \mathbb{R}_+^* \{(\langle a_1, y\rangle, \ldots, \langle a_m, y\rangle) \mid y \in B\}.$$

Now, by Theorem 3 (B is a convex compact subset of E)

$$\{(\langle a_1, y\rangle, \ldots, \langle a_m, y\rangle) \mid y \in B\} = \{(\langle a_1, y\rangle, \ldots, \langle a_m, y\rangle) \mid y \in \mathcal{F}_m(B)\}.$$

To conclude, it suffices to observe that

$$\mathcal{F}_{m+1}(K)\backslash\{0\} = \{\lambda y \mid \lambda > 0, y \in \mathcal{F}_m(B)\}. \qquad \square$$

It is important to note that this result is not true, as a general rule, if one substitutes m for $m+1$; counterexamples can be found in $E = \mathbb{R}^3$.

As a consequence, we have the following statement characterizing the nonexistence of nonnull facial solutions to a system of linear equations.

Theorem 6. *Assume K is pointed and consider $a_1, \ldots, a_m \in E$. Then the following are equivalent*:

(i) $(\langle a_1, x\rangle, \ldots, \langle a_m, x\rangle) \neq (0, \ldots, 0)$ *for all* $0 \neq x \in \mathcal{F}_{m+1}(K)$. (36)

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

$$\sum_{i=1}^{m} \mu_i a_i \in \operatorname{int}(K^+). \tag{37}$$

Applications. We turn our attention back to our guiding framework: $E = \mathcal{S}_n(\mathbb{R})$, endowed with the usual inner product $\langle\langle \cdot, \cdot \rangle\rangle$; $K = \{A \in \mathcal{S}_n(\mathbb{R}) \mid A \succeq 0\}$. K is a pointed closed convex cone, it is its own (positive) polar cone ($K = K^+$). The facial structure of this cone is also well known: faces of K are closed convex cones of dimension

$$0, 1, 3, \ldots, \frac{p(p+1)}{2}, \frac{(p+1)(p+2)}{2}, \ldots, \frac{n(n+1)}{2}.$$

The apex is the only extreme point (face of dimension 0), while the whole cone K is the only face of full dimension $n(n+1)/2$.

Determining $\mathcal{F}_m(K)$ is rather easy here (see, for instance, Corollary 6.1 of [17]):

$$\mathcal{F}_m(K) = \{XX^T \mid X \in M_{n,p}(\mathbb{R})\}, \tag{38}$$

where p is the smallest integer satisfying $(p+1)(p+2)/2 > m$.

As, for example,

$$\mathcal{F}_1(K) = \mathcal{F}_2(K) = \{xx^T \mid x \in \mathcal{M}_{n,1} \equiv \mathbb{R}^n\}$$

(union of the apex 0 and all the extreme rays of K; there is no face of dimension 2 in K),

$$\mathcal{F}_3(K) = \{XX^T \mid X \in \mathcal{M}_{n,2}\}.$$

The smallest p satisfying $(p+1)(p+2)/2 > m$ turns out to be exactly $p = \lfloor(\sqrt{8m+1}-1)/2\rfloor$, such as introduced and used in Sections 1.3 and 2.3.

As we know that $\langle\langle A, XX^T\rangle\rangle$ with $X \in M_{n,p}(\mathbb{R})$ is nothing more than $\langle\langle AX, X\rangle\rangle$, we are able to "close the loop" with the results à la Barvinok: Corollary 1 goes with the result displayed in Section 1.3, Theorem 4 with the one displayed in Section 2.3, etc.

Acknowledgment

We thank B. T. Polyak (Moscow) for several references he pointed out to us.

References

1. A. V. Arutyunov and V. N. Rozova, Regular zeros of a quadratic mapping and local controllability of nonlinear systems, Differential Equations, 35 (1999), 723–728.
2. A. I. Barvinok, Problems of distance geometry and convex properties of quadratic maps, Discrete Comput. Geom., 13 (1995), 189–202.
3. A. I. Barvinok, Convexity, Duality and Optimization, Lectures Notes, University of Michigan, 1998.
4. A. I. Barvinok, On convex properties of the quadratic image of the sphere, Preprint, University of Michigan, 1999.
5. H. F. Bohnenblust, Joint positiveness of matrices, Unpublished manuscript.
6. L. Brickman, On the fields of values of a matrix, Proc. Amer. Math. Soc., 12 (1961), 61–66.
7. E. Calabi, Linear systems of real quadratic forms, Proc. Amer. Math. Soc., 15 (1964), 844–846.
8. X. Chen and Y. Yuan, A note on quadratic forms, Math. Programming, 86 (1999), 187–197.
9. J.-P. Crouzeix, J. E. Martinez-Legaz, and A. Seeger, An alternative theorem for quadratic forms and extensions, Linear Algebra Appl., 215 (1995), 121–134.
10. L. L. Dines, On the mapping of quadratic forms, Bull. Amer. Math. Soc., 47 (1941), 494–498.
11. P. Finsler, Über das Vorkommen definiter und semidefiniter Formen in Scharen quadratischer Formen, Comment. Math. Helv. 9 (1936/37), 188–192.
12. W. Greub, Linear Algebra, 1st edn., Springer-Verlag, 1958; Heidelberger Taschenbücker, Bd. 179, 1976.
13. C. Hamburger, Two extensions to Finsler's recurring theorem, Appl. Math Optim., 40 (1999), 183–190.
14. J.-B. Hiriart-Urruty and C. Lemarechal, Convex Analysis and Minimzation Algorithms I, Grundleheren der mathematischen Wissenschaften, vol. 305, Springer-Verlag, Berlin, 1993.
15. R. Horn and C. R. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, 1985.
16. D. H. Jacobson, A generalization of Finsler's theorem for quadratic inequalities and equalities, Quaestiones Math., 1 (1976), 19–28.
17. A. S. Lewis, Eigenvalue-constrained faces, Linear Algebra Appl., 269 (1998), 159–181.
18. A. Matveev, Lagrange duality in nonconvex Optimization theory and modifications of the Toeplitz–Hausdorff theorem, St Petersburg Math. J., 7 (1996), 787–815.
19. A. S. Matveev, On the convexity of the ranges of quadratic mappings, St Petersburg Math. J., 10 (1999), 343–372.
20. A. Megretsky and S. Treil, Power distribution inequalities in Optimization and robustness of uncertain systems, Math. Systems Estimation Control, 3 (1993), 301–319.
21. B. T. Polyak, Convexity of quadratic transformations and its use in Control and Optimization, J. Optim. Theory Appl., 99 (1998), 553–583.
22. Y. T. Poon, Generalized numerical ranges, joint positive definitness and multiple eigenvalues, Proc. Amer. Math. Soc., 125 (1997), 1625–1634.
23. J. Stoer and C. Witzgall, Convexity and Optimization in Finite Dimension I, Springer-Verlag, New York, 1970.
24. F. Uhlig, A recurring theorem about pairs of quadratic forms and extension: a survey, Linear Algebra Appl., 25 (1979), 219–237.
25. V. A. Yakubovich, Nonconvex optimization problem: the infinite-horizon linear-quadratic control problem with quadratic constraints, Systems Control Lett., 19 (1992), 13–22.
26. Y. Yuan, On a subproblem of trust region algorithms for constrained optimization, Math. Programming, 47 (1990), 53–63.

Accepted 27 *July* 2001. *Online publication* 7 *January* 2002.