Generalized Ellipsoids

Amir Ali Ahmadi [*] Abraar Chaudhry [*] Cemil Dibek[†]

Abstract

We introduce a family of symmetric convex bodies called *generalized ellipsoids of degree d* (GE-ds), with ellipsoids corresponding to the case of $d = 0$. Generalized ellipsoids (GEs) retain many geometric, algebraic, and algorithmic properties of ellipsoids. We show that the conditions that the parameters of a GE must satisfy can be checked in strongly polynomial time, and that one can search for GEs of a given degree by solving a semidefinite program whose size grows only linearly with dimension. We give an example of a GE which does not have a second-order cone representation, but show that every GE has a semidefinite representation whose size depends linearly on both its dimension and degree. In terms of expressiveness, we prove that for any integer $m \geq 2$, every symmetric full-dimensional polytope with $2m$ facets and every intersection of m co-centered ellipsoids can be represented exactly as a GE-d with $d \leq 2m - 3$. Using this result, we show that every symmetric convex body can be approximated arbitrarily well by a GE-d and we quantify the quality of the approximation as a function of the degree d. Finally, we present applications of GEs to several areas, such as time-varying portfolio optimization, stability analysis of switched linear systems, robust-to-dynamics optimization, and robust polynomial regression.

Keywords: *ellipsoids, convex bodies, conic optimization, semidefinite representations, polynomial matrices.*

1 Introduction

An *ellipsoid* in Euclidean space $\mathbb{R}^n$ is a set of the type

$$\mathcal{E} = \{x \in \mathbb{R}^n \mid (x - x_0)^T P(x - x_0) \leq 1\}, \tag{1}$$

where P is a (symmetric) positive definite matrix and $x_0 \in \mathbb{R}^n$ is a given vector. Ellipsoids are among the most prominent examples of convex sets in applied and computational mathematics. In optimization, they represent sublevel sets of objective functions of convex quadratic programs, feature in the description of celebrated algorithms such as the ellipsoid method and Dikin's method, and serve as primary examples of uncertainty sets in robust optimization. In control and robotics, they appear as sublevel sets of quadratic Lyapunov functions or in the description of the manipulability set of a robotic system. In convex geometry, they are used to approximate convex bodies with an approximation guarantee established by John's ellipsoid theorem. In probability and statistics, they appear as confidence regions for Gaussian or more generally elliptical distributions. These examples are a small sample among many. We refer the reader to [45, Chap. 1] for some reasons why ellipsoids are so ubiquitous in many areas.

Since ellipsoids are sublevel sets of strictly convex quadratic polynomials, a very natural generalization of ellipsoids would be to consider sublevel sets of strictly convex polynomials of degree higher

[*] Princeton University, Operations Research and Financial Engineering. AAA and AC were partially supported by the MURI award of the AFOSR and the Sloan Fellowship. Emails: aaa@princeton.edu; azc@princeton.edu

[†] Koç University, Turkey, Department of Industrial Engineering. Email: cdibek@ku.edu.tr

than two. However, unless P=NP, the set of convex (or strictly convex) polynomials of degree at least four does not admit a tractable description [8]. This implies that algorithms based on this approach would in general not scale well with increasing dimension.

In this work, we propose a different generalization of ellipsoids to sets that we call *generalized ellipsoids of degree* d (GE-ds), with ellipsoids corresponding to the case of $d = 0$. Generalized ellipsoids (GEs) retain some key geometric and algebraic properties of ellipsoids, such as the properties of being convex and semialgebraic. Importantly, they are also algorithmically tractable to search for and optimize over. The reason for this tractability stems from the fact that our generalization (see Definition 2.1) keeps the defining inequalities of a GE *quadratic* in $x \in \mathbb{R}^n$, while adding a *single* new variable on which these inequalities depend polynomially. Despite the univariate nature of this dependence, we show that GEs can approximate any n-dimensional symmetric convex body to arbitrary accuracy.

1.1 Organization and main contributions

The remainder of this paper is organized as follows. In Section 2, we give the definition of generalized ellipsoids, justify our definition, and provide some examples.

In Section 3, we focus on recognition of GEs and search for GEs. In Section 3.1, we show that the conditions that the parameters of a GE must satisfy can be checked in strongly polynomial time. In particular, we show that one can check if a univariate polynomial matrix is positive semidefinite over an interval (or the real line) in strongly polynomial time. This result may be of independent interest. In Section 3.2, using the fact that certain low-degree sum of squares tests for positive semidefiniteness of polynomial matrices are exact, we show that one can search for GEs of a given degree by solving a single semidefinite program. In fact, the size of this semidefinite program grows only *linearly* with dimension.

In Section 4, we investigate whether GEs can be represented as the feasible set of three increasingly expressive families of tractable conic programs. This is relevant to applications involving optimization over a GE. In Section 4.1, we show that when $d \geq 2$, GE-ds cannot always be described by finitely many convex quadratic constraints. In Section 4.2, we show that when $d \geq 16$, GE-ds do not always have a second-order cone representation. In Section 4.3, we show that every GE has a semidefinite representation whose size depends linearly on both its dimension and degree.

In Section 5, we focus on the expressive power of GEs. We show that for any integer $m \geq 2$, every compact intersection of m "semiellipsoids", and in particular every symmetric full-dimensional polytope with $2m$ facets and every intersection of m co-centered ellipsoids, can be represented exactly as a GE-d with $d \leq 2m - 3$. A technical lemma that goes into this proof shows that for any dimension $m \geq 2$, there is a polynomial curve of degree $2m - 3$ that lies within the unit simplex in $\mathbb{R}^m$ and visits every one of its corners. We believe this statement and our game-theoretic proof of it may be of independent interest. We then show that every symmetric convex body can be approximated arbitrarily well by a GE-d and we quantify the quality of the approximation as a function of the degree d.

In Section 6, we present four applications involving GEs and provide some numerical examples. In Section 6.1, we consider a time-varying extension of the minimum-variance portfolio optimization problem in finance. In Section 6.2, we consider an application in dynamical systems and show that asymptotically stable switched linear systems always admit a GE as an invariant set. In Section 6.3, we show how GEs can provide inner approximations to feasible sets of robust-to-dynamics optimization problems when the dynamical system is uncertain. In Section 6.4, we present an application of GEs to the problem of polynomial regression in statistics when there is uncertainty in the measurements.

Finally, in Section 7, we list a few questions for future research.

2 Definition of GEs

We begin by establishing some notation and terminology. We denote the set of real symmetric $n \times n$ matrices by S^n. We write $M \succeq 0$ if $M \in S^n$ is positive semidefinite (psd) and $M \succ 0$ if M is positive definite (pd). We denote the set of $n \times n$ psd (resp. pd) matrices by S^n_+ (resp. S^n_{++}). The kernel of M is denoted by Ker(M). We refer to a matrix with polynomial entries as a *polynomial matrix*. We can now give the definition of our generalization of ellipsoids.

Definition 2.1. *A set $\mathcal{E}_d \subset \mathbb{R}^n$ is a* generalized ellipsoid of degree d *(GE-d) if it can be written as*

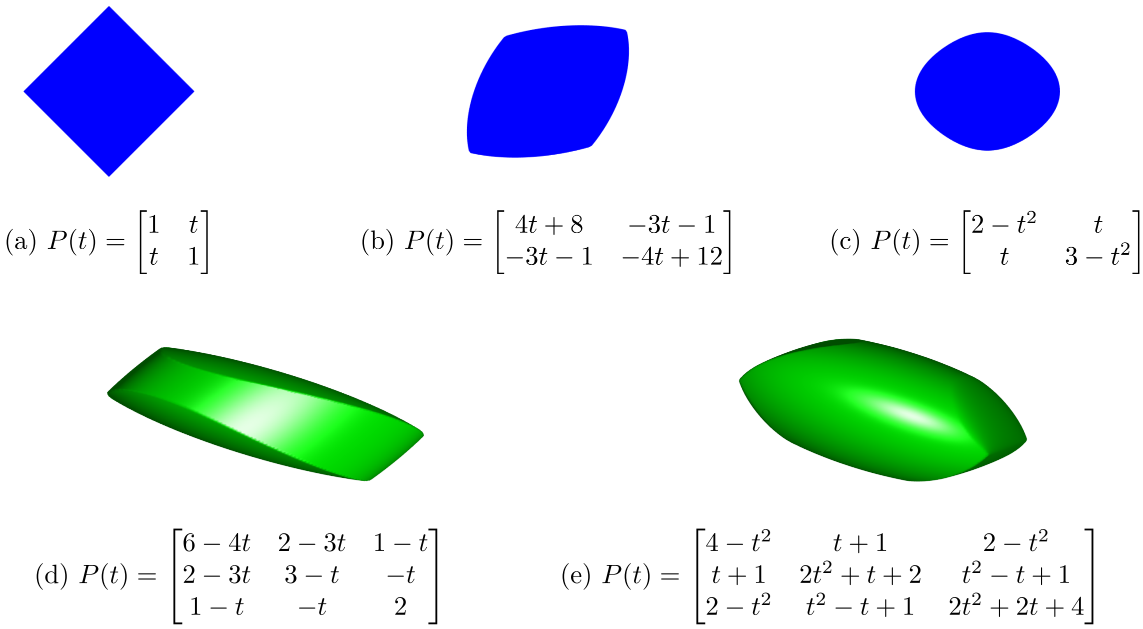
$$\mathcal{E}_d = \{x \in \mathbb{R}^n \mid (x - x_0)^T P(t)(x - x_0) \leq 1 \ \forall t \in [-1, 1]\} \tag{2}$$

for some vector $x_0 \in \mathbb{R}^n$ and some univariate polynomial matrix $P(t)$ of degree (at most) d that satisfies

- $P(t) \succeq 0 \quad \forall t \in [-1, 1]$*, and*
- $\bigcap_{t \in [-1,1]}$ *Ker*$(P(t)) = \{0\}$.

We refer to these two conditions as the "psd condition" and the "kernel condition", respectively. We say that a set is a generalized ellipsoid *(GE) if it is a GE-d for some nonnegative integer d.*

Observe that ellipsoids correspond precisely to GE-0s. Indeed, when $d = 0$, $P(t)$ is a constant matrix, say $P(t) = P$, and we have $P \succ 0$ if and only if $P \succeq 0$ and Ker$(P) = \{0\}$. It is straightforward to check that GEs are symmetric convex bodies (i.e., compact convex sets with non-empty interior that are symmetric around their center). Throughout this paper, without loss of generality, we assume that the center x_0 of our GEs is at the origin. Figure 1 demonstrates a few examples of GEs, together with the corresponding polynomial matrices $P(t)$.



(a) $P(t) = \begin{bmatrix} 1 & t \\ t & 1 \end{bmatrix}$

(b) $P(t) = \begin{bmatrix} 4t+8 & -3t-1 \\ -3t-1 & -4t+12 \end{bmatrix}$

(c) $P(t) = \begin{bmatrix} 2-t^2 & t \\ t & 3-t^2 \end{bmatrix}$

(d) $P(t) = \begin{bmatrix} 6-4t & 2-3t & 1-t \\ 2-3t & 3-t & -t \\ 1-t & -t & 2 \end{bmatrix}$

(e) $P(t) = \begin{bmatrix} 4-t^2 & t+1 & 2-t^2 \\ t+1 & 2t^2+t+2 & t^2-t+1 \\ 2-t^2 & t^2-t+1 & 2t^2+2t+4 \end{bmatrix}$

Figure 1: Some examples of GEs with the corresponding $P(t)$ matrices.

Note that a GE-d has a compact representation in terms of the coefficients of the polynomial matrix $P(t)$. The reason that the matrix $P(t)$ in the definition of GEs depends on a single variable t, and that this dependence is polynomial, is justified by algorithmic tractability purposes (see Section 3 and Section 4). Yet, this family of sets is quite expressive (see Section 5). For now, let us just justify the kernel condition. When generalizing ellipsoids, one may find it more natural to consider positive definiteness of $P(t)$ for all $t \in [-1,1]$, or for some $t \in [-1,1]$. Suppose $P(t) \succeq 0 \ \forall t \in [-1,1]$ and consider the following three candidate conditions:

(i) $P(t) \succ 0 \ \forall t \in [-1,1]$, (ii) $P(t) \succ 0$ for some $t \in [-1,1]$, (iii) $\bigcap_{t\in[-1,1]} \operatorname{Ker}(P(t)) = \{0\}$.

Clearly, (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii). However, (ii) $\not\Rightarrow$ (i), as seen, e.g., by $P(t) = \begin{bmatrix} 1-t & 0 \\ 0 & 1+t \end{bmatrix}$. Similarly, (iii) $\not\Rightarrow$ (ii); as seen, e.g., by

$$P(t) = \begin{bmatrix} (1+t)^2 & (1-t)(1+t) \\ (1-t)(1+t) & (1-t)^2 \end{bmatrix}.$$

Indeed, we have $\operatorname{Ker}(P(1)) \cap \operatorname{Ker}(P(-1)) = \{0\}$, but $P(t)$ is not pd for any $t \in [-1,1]$ as $\det(P(t))$ is identically zero. Hence, our choice of the kernel condition leads to a more inclusive definition.

We also note that a set of the type (2) with $P(t) \succeq 0 \ \forall t \in [-1,1]$ is a convex body if and only if the kernel condition is satisfied. We restate this claim in a lemma below in terms of norms that convex bodies define. Recall that any ellipsoid $\mathcal{E}$ as in (1) defines an ellipsoid (quadratic) norm by

$$||x||_{\mathcal{E}} = \sqrt{x^T P x}.$$

Similarly, for any generalized ellipsoid $\mathcal{E}_d$ as in (2), we can define a *generalized ellipsoid norm of degree* d (GE-d-norm) by

$$||x||_{\mathcal{E}_d} = \max_{t\in[-1,1]} \sqrt{x^T P(t) x}. \tag{3}$$

The proof of the following lemma is straightforward and hence omitted.

Lemma 2.2. *A function $f : \mathbb{R}^n \to \mathbb{R}$ of the type $f(x) = \max_{t\in[-1,1]} \sqrt{x^T P(t) x}$ with $P(t) \succeq 0 \ \forall t \in [-1,1]$ is a norm if and only if $\bigcap_{t\in[-1,1]} \operatorname{Ker}(P(t)) = \{0\}$.*

3 Recognition of GEs and Search for GEs

3.1 Efficient recognition of GEs

In this section, we show that GEs can be recognized in strongly polynomial time[1]; i.e, given the coefficients of the entries of a univariate polynomial matrix $P(t)$, we can check both the kernel condition and the psd condition of Definition 2.1 in strongly polynomial time. In particular, the running time of the recognition procedure is polynomial in both the dimension n and the degree d of the GE-d, and its number of arithmetic operations does not depend on the encoding length of the coefficients of $P(t)$.

[1] We recall that a *strongly polynomial time algorithm* is an algorithm such that (i) it consists of the elementary arithmetic operations: addition, subtraction, comparison, multiplication, and division, (ii) the number of elementary operations depends polynomially on the dimension of the input to the algorithm, and (iii) the encoding length of the numbers occurring during the algorithm is bounded by a polynomial function of the encoding length of the input. We also recall that when we speak of polynomial time or strongly polynomial time algorithms, we are working in the Turing model of computation where the input to the problem consists of rational numbers and hence has finite encoding length. Here, the encoding length could for example be taken as the number of bits required in a binary representation of the input. See [25] for more details. In our case, the input to the problem is the rational coefficients of the entries of $P(t)$ and the dimension of the input is the total number of coefficients which is $\binom{n+1}{2}(d+1)$.

Lemma 3.1. *Given a univariate polynomial matrix $P(t)$, one can check in strongly polynomial time whether $\bigcap_{t\in[-1,1]} \operatorname{Ker}(P(t)) = \{0\}$.*

Proof. The kernel condition is equivalent to checking whether there is a vector $x \neq 0$ such that $P(t)x = 0 \ \forall t \in [-1, 1]$. Since a univariate polynomial vanishes on $[-1, 1]$ if and only if all of its coefficients are zero, this condition is equivalent to the existence of a solution $x \neq 0$ to the linear system $Ax = 0$, where the matrix $A \in \mathbb{R}^{n(d+1)\times n}$ is such that Ax consists of the coefficients of the n polynomials in $P(t)x$. To check existence of a nonzero solution, we can equivalentally check whether the rank of A is less than n, which can be done in strongly polynomial time, e.g., by Edmonds' implementation of Gaussian elimination (see, e.g., [25, Corollary 1.4.9.b]). □

Next, we show that checking the psd condition in Definition 2.1 can be done in strongly polynomial time. The following theorem, and the lemma it relies on, may be of independent interest.

Theorem 3.2. *Given a univariate polynomial matrix $P(t)$, one can check in strongly polynomial time whether $P(t) \succeq 0 \ \forall t \in [-1, 1]$.*

To prove this theorem, we first establish the following lemma which generalizes a result in [40].

Lemma 3.3. *Given a univariate polynomial matrix $P(t)$, one can check in strongly polynomial time whether $P(t) \succeq 0 \ \forall t \in \mathbb{R}$.*

Proof. For $k = 1, \ldots, n$ and $t \in \mathbb{R}$, let $P(t)_k$ denote the $k \times k$ leading principal submatrix of $P(t)$ and define the univariate polynomial

$$p_{k,t}(s) = \det(P(t)_k + sI_k),$$

where I_k is the $k \times k$ identity matrix. For $k = 1, \ldots, n$ and $i = 0, \ldots, k$, let $c_{k,i}(t)$ denote the coefficient of s^i in $p_{k,t}(s)$. We first claim that for any $t \in \mathbb{R}$, $P(t) \succeq 0$ if and only if $c_{k,i}(t) \geq 0$ for $k = 1, \ldots, n$ and $i = 0, \ldots, k$.

Fix $t \in \mathbb{R}$ and suppose first that $P(t) \succeq 0$. Fix any $k \in \{1, \ldots, n\}$ and observe that the principal submatrix $P(t)_k$ is also psd. For $j = 1, \ldots, k$, let λ_j denote the (nonnegative) eigenvalues of $P(t)_k$. Then we can write

$$p_{k,t}(s) = \prod_{j=1}^{k} (\lambda_j + s).$$

It is clear from this representation that all coefficients of $p_{k,t}(s)$ are nonnegative.

To see the converse, fix $t \in \mathbb{R}$ again and suppose $c_{k,i}(t) \geq 0$ for $k = 1, \ldots, n$ and $i = 0, \ldots, k$. Since the coefficients of $p_{k,t}(s)$ are nonnegative and not all equal to zero (observe $c_{k,k}(t) = 1$), it follows that $p_{k,t}(s) > 0$ for all $s > 0$. By Sylvester's criterion[2], $P(t) + sI_n$ is positive definite for arbitrarily small $s > 0$. Thus, $P(t)$ is psd. This completes the proof of the claim.

It remains to show that one can test nonnegativity of each function $c_{k,i}(t)$ in strongly polynomial time. For the remainder of the proof, fix k to be an integer between 1 and n and fix i to be an integer between 0 and k. Observe that $c_{k,i}(t)$ is a polynomial in t of degree $d(k - i)$. To obtain the coefficients of $c_{k,i}(t)$, it suffices to evaluate this polynomial at $d(k - i) + 1$ distinct points and then solve a nonsingular linear system for the coefficients of the interpolating polynomial (recall one can solve a nonsingular linear system in strongly polynomial time; see, e.g., [25, Corollary 1.4.9.a]). To evaluate $c_{k,i}(t)$ at a particular t, we need access to the coefficient of s^i in $p_{k,t}(s)$. To compute these coefficients, it similarly suffices to evaluate the polynomial $p_{k,t}(s)$ at $k + 1$ distinct points and then solve an associated linear system. Recall that the determinant of a constant matrix can be computed in strongly polynomial time (see, e.g., [25, Corollary 1.4.9.c]); hence, each evaluation of $p_{k,t}(s)$ can

[2] Recall that Sylvester's criterion states that an $n \times n$ symmetric matrix is positive definite if and only if its n leading principal minors are all positive.

be done in strongly polynomial time via the determinantal definition of $p_{k,t}(s)$. Hence, overall, we can compute the coefficients of $c_{k,i}(t)$ by computing $(d(k-i)+1)(k+1)$ determinants and solving $d(k-i)+2$ linear systems. With the coefficients of $c_{k,i}(t)$ at hand, one can check the nonnegativity of each $c_{k,i}(t)$ in strongly polynomial time via the method of [40, Section 5]. □

Proof of Theorem 3.2. We claim that a polynomial matrix $P(t)$ of degree d is psd for every $t \in [-1,1]$ if and only if the polynomial matrix $Q(t) = (t^2+1)^d P\left(\frac{t^2-1}{t^2+1}\right)$ is psd for every $t \in \mathbb{R}$.

Suppose first that $P(t) \succeq 0 \ \forall t \in [-1,1]$. For every $t \in \mathbb{R}$, we have $\frac{t^2-1}{t^2+1} \in [-1,1)$, and thus, $P\left(\frac{t^2-1}{t^2+1}\right) \succeq 0$. Since $(t^2+1)^d$ is nonnegative, it follows that $Q(t) \succeq 0 \ \forall t \in \mathbb{R}$.

To see the converse, suppose for the sake of contradiction that $P(t) \not\succeq 0$ for some $t \in [-1,1]$. By the closedness of the psd cone, there is some $\hat{t} \in [-1,1)$ such that $P(\hat{t}) \not\succeq 0$. Consider $\bar{t} = \sqrt{\frac{-1-\hat{t}}{-1+\hat{t}}}$. Observe $Q(\bar{t}) = (\frac{-1-\hat{t}}{-1+\hat{t}}+1)^d P(\hat{t})$. Since $\hat{t} \in [-1,1)$, we have $\frac{-1-\hat{t}}{-1+\hat{t}}+1 > 0$. Therefore, $Q(\bar{t}) \not\succeq 0$, a contradiction.

Thus, to check that $P(t)$ is psd for every $t \in [-1,1]$, we can check if $Q(t)$ is psd for every $t \in \mathbb{R}$. It is straightforward to see that the coefficients of $Q(t)$ can be derived from those of $P(t)$ in strongly polynomial time. Hence, the result follows from Lemma 3.3. □

3.2 Efficient search for GEs

In this section, we observe that the set of $n \times n$ polynomial matrices $P(t)$ of degree d that satisfy $P(t) \succeq 0 \ \forall t \in [-1,1]$ has a semidefinite representation of size linear in d (resp. in n) for fixed n (resp. fixed d). As semidefinite programs (SDPs) can be solved in polynomial time to arbitrary accuracy [47], this observation leads to efficient algorithms for applications where one needs to *search* for GEs (see, e.g., Section 6.1 and Section 6.3). It can also reformulate the problem of checking if a given polynomial matrix $P(t)$ satisfies $P(t) \succeq 0 \ \forall t \in [-1,1]$ (i.e., the recognition question of the previous subsection) as a semidefinite programming feasibility problem. However, it is currently unknown whether semidefinite programming feasibility problems can be solved in polynomial time (let alone strongly polynomial time), and this is the justification for our alternative algorithm in the proof of Theorem 3.2.

The semidefinite programming formulation that we present here arises from a connection to sum of squares polynomials. We recall that a polynomial $p : \mathbb{R}^n \to \mathbb{R}$ is *nonnegative* if $p(x) \geq 0$ for all $x \in \mathbb{R}^n$ and a *sum of squares* (sos) if there exist polynomials $q_1(x), \ldots, q_m(x)$ such that $p(x) = \sum_{i=1}^m q_i^2(x)$. A polynomial matrix $Y : \mathbb{R}^n \to S^k$ is said to be an *sos-matrix* if $Y(x) = A(x)^T A(x)$ for some (not necessarily square) polynomial matrix $A(x)$. It is a classical fact in algebra that a univariate polynomial matrix $Y(t)$ is positive semidefinite for all $t \in \mathbb{R}$ if and only if it is an sos-matrix; see, e.g., [49, 50] for some early proofs, [19, Theorem 7.1] for a short constructive proof that specifies the inner dimension in the factorization, and [10] for more context, a decomposition algorithm, and connections to other areas. For our purposes, we need a version of this statement that applies to univariate polynomial matrices that are positive semidefinite over an interval. This variant appears for example in [22, Theorem 2.5] and [37, Theorem 6.11].

Theorem 3.4. *Let $P(t)$ be a symmetric univariate polynomial matrix of degree d. If d is odd, then $P(t) \succeq 0 \ \forall t \in [-1,1]$ if and only if there exist sos-matrices $X_1(t)$ and $X_2(t)$ of degree $d-1$ such that*

$$P(t) = (t+1)X_1(t) + (1-t)X_2(t) \quad \forall t \in \mathbb{R}.$$

Similarly, if d is even, then $P(t) \succeq 0 \ \forall t \in [-1,1]$ if and only if there exist sos-matrices $X_1(t)$ and $X_2(t)$ of degree d and $d-2$, respectively, such that

$$P(t) = X_1(t) + (1-t^2)X_2(t) \quad \forall t \in \mathbb{R}.$$

As we describe next, Theorem 3.4 leads to a semidefinite representation of polynomial matrices that are psd on $[-1,1]$. This is essentially based on the following facts: (i) A polynomial matrix $Y : \mathbb{R}^n \to S^k$ is an sos-matrix if and only if the (scalar-valued) polynomial $y^TY(x)y$ in the variables $(x_1,\ldots,x_n,y_1,\ldots,y_k)$ is a sum of squares; (ii) A polynomial $p : \mathbb{R}^n \to \mathbb{R}$ of degree $2d$ is a sum of squares if and only if there exists a psd matrix Q such that $p(x) = v(x)^TQv(x)$, where $v(x)$ is the vector of all monomials of degree up to d (see, e.g., [20, 39]).

Proposition 3.5. *Let $P(t)$ be a symmetric univariate $n \times n$ polynomial matrix of degree d. For a positive integer d', let*

$$v_{d'}(t,y) := \left(y_1,\ldots,y_n,y_1t,\ldots,y_nt,\ldots,y_1t^{d'},\ldots,y_nt^{d'}\right)^T$$

be the vector of all monomials of the form $y_\ell t^k$ for $\ell = 1,\ldots,n$ and $k = 0,\ldots,d'$.

If d is odd, then $P(t) \succeq 0\ \forall t \in [-1,1]$ if and only if there exist positive semidefinite matrices Q_1, Q_2 of size $(\frac{d+1}{2})n \times (\frac{d+1}{2})n$ that satisfy the following equations ($\forall t \in \mathbb{R}$ and $\forall y \in \mathbb{R}^n$):

- $P(t) = (t+1)X_1(t) + (1-t)X_2(t)$,
- $y^TX_i(t)y = v_{\frac{d-1}{2}}(t,y)^TQ_iv_{\frac{d-1}{2}}(t,y)$ *for* $i = 1,2.$

Similarly, if d is even, then $P(t) \succeq 0\ \forall t \in [-1,1]$ if and only if there exist positive semidefinite matrices Q_1, Q_2 of size $(\frac{d}{2}+1)n \times (\frac{d}{2}+1)n$ and $(\frac{d}{2})n \times (\frac{d}{2})n$, respectively, that satisfy the following equations ($\forall t \in \mathbb{R}$ and $\forall y \in \mathbb{R}^n$):

- $P(t) = X_1(t) + (1-t^2)X_2(t),$
- $y^TX_1(t)y = v_{\frac{d}{2}}(t,y)^TQ_1v_{\frac{d}{2}}(t,y),$
- $y^TX_2(t)y = v_{\frac{d}{2}-1}(t,y)^TQ_2v_{\frac{d}{2}-1}(t,y).$

Since two polynomials are equal everywhere if and only if they have the same coefficients, Proposition 3.5 reduces the task of checking the psd condition in Definition 2.1 to solving an SDP; see, e.g., [4, Proposition 2] for a more explicit representation of this SDP.

Remark 1. We make two remarks regarding computational considerations:

1. The sizes of the matrices Q_1, Q_2 in Proposition 3.5 grow only linearly with d (resp. with n) for fixed n (resp. fixed d). This is in contrast to the SDP hierarchies that arise in a search for convex homogeneous polynomials of degree d in n variables whose sublevel sets can also be considered as a natural generalization of ellipsoids. The size of the semidefinite constraint, even in the first level of this SDP hierarchy (see [3]), grows at the rate $n\binom{n+\frac{d}{2}-2}{\frac{d}{2}-1}$.
2. For implementation purposes, many parsers (e.g., YALMIP [33] or SumOfSquares.jl [48]) directly accept sum of squares constraints on a polynomial or polynomial matrix with unknown coefficients and do the conversion to an SDP automatically. Therefore, for our purposes, one can use these parsers to directly work with the representation in Theorem 3.4. The resulting SDPs can be readily solved, e.g., by general-purpose interior point methods. While not the focus of this paper, we suspect implementation improvements may be possible by exploiting the univariate nature of the sum of squares constraints, for example using techniques from [34, 36, 38, 32, 26].

Finally, we note that by searching over polynomial matrices of degree d that satisfy the psd condition in Definition 2.1, we are searching over the closure of the set of polynomial matrices that satisfy both conditions in that definition. We can always check the kernel condition in Definition 2.1 via Lemma 3.1 as a post-processing step. This type of approach is common in applications of semidefinite and sum of squares programming where the optimization set of interest is not closed.

4 Conic Representation of GEs

In this section, we study whether GEs admit a representation as the feasible set of different families of tractable conic programs. This is relevant for applications where one needs to optimize over a GE. For the sake of clarity, we stress that in contrast to Section 3.2, where the representation was in the space of coefficients of polynomial matrices, the representation questions that we are concerned with in this section are in x space and relate to a fixed GE-d as defined in (2). The results of Section 3.2 show that one can use semidefinite programming to search for polynomial matrices which define a GE, whereas the results of this section are concerned with the problem of optimizing over the set of vectors contained in a given GE.

It is clear that one cannot always optimize over a given GE using linear programming. Indeed, since ellipsoids in dimension two or higher have an infinite number of extreme points, already a GE-0 fails to be polyhedral. In our next three subsections, we ask if increasingly broader classes of convex sets can represent GEs.

4.1 Can GEs be described by finitely many convex quadratic constraints?

Considering the definition of a GE-d in (2) and the fact that a GE-0 is an ellipsoid, a natural question that comes to mind is whether a GE-d can always be described by finitely many convex quadratic constraints. The next proposition shows that this is the case only when $d = 0$ or $d = 1$.

Proposition 4.1. *The following are equivalent:*

(i) for every GE-d $\mathcal{E}_d$, there exists a nonnegative integer m and matrices $P_1, \dots P_m \succeq 0$ such that

$$\mathcal{E}_d = \{x \in \mathbb{R}^n \mid x^T P_i x \leq 1 \quad i = 1, \dots, m\};$$

(ii) $d \in \{0, 1\}$.

Proof. (ii) ⇒ (i): A GE-0 is precisely an ellipsoid so the claim is established with $m = 1$. We show that a GE-1 can always be described by two convex quadratic constraints. Let

$$\mathcal{E}_1 = \{x \in \mathbb{R}^n \mid \max_{t \in [-1,1]} x^T P(t) x \leq 1\}$$

be a GE-1. Since $P(t)$ is a univariate polynomial matrix of degree 1, for a given $x \in \mathbb{R}^n$, $x^T P(t) x$ is an affine function in t, and therefore its maximum over $[-1, 1]$ is attained at $t = 1$ or at $t = -1$. Hence,

$$\begin{aligned} \mathcal{E}_1 &= \{x \in \mathbb{R}^n \mid \max_{t \in \{-1,1\}} x^T P(t) x \leq 1\} \\ &= \{x \in \mathbb{R}^n \mid x^T P(1) x \leq 1\} \cap \{x \in \mathbb{R}^n \mid x^T P(-1) x \leq 1\}. \end{aligned}$$

This establishes the claim with $m = 2$.[3]

(i) ⇒ (ii): We show that for $d \geq 2$, there are GE-ds that cannot be described by finitely many convex quadratic constraints. Consider the set given by $\mathcal{E}_2 = \{x \in \mathbb{R}^2 \mid \max_{t \in [-1,1]} x^T P(t) x \leq 1\}$ with

$$P(t) = \begin{bmatrix} 2 - t^2 & t \\ t & 3 - t^2 \end{bmatrix}.$$

This set is depicted in Figure 1c. It is easy to verify that $P(t) \succ 0 \ \forall t \in [-1, 1]$, and therefore $\mathcal{E}_2$ is a valid GE-2. For any given $x \neq 0$, the function $x^T P(t) x = -(x_1^2 + x_2^2)t^2 + 2x_1 x_2 t + 2x_1^2 + 3x_2^2$ is a

[3] In fact, the intersection of any given two co-centered ellipsoids can be represented by a GE-1. This is a special case of Theorem 5.1 in Section 5.

concave quadratic polynomial in t and is maximized at $t^* = \frac{x_1 x_2}{x_1^2+x_2^2}$. Observe that $t^* \in [-1,1]$. It is then easy to verify that

$$\mathcal{E}_2 = \left\{ x \in \mathbb{R}^2 \;\middle|\; \frac{x_1^2 x_2^2}{x_1^2+x_2^2} + 2x_1^2 + 3x_2^2 \leq 1 \right\}.$$

Let $h(x) := \frac{x_1^2 x_2^2}{x_1^2+x_2^2} + 2x_1^2 + 3x_2^2$ and assume for a contradiction that $\mathcal{E}_2$ can be described by finitely many (convex) quadratic constraints, i.e., $\mathcal{E}_2 = \{x \in \mathbb{R}^2 \mid x^T P_i x \leq 1 \quad i = 1, \ldots, m\}$ for some (psd) matrices $P_1, \ldots, P_m$. Let $g(x) := \max_{i=1,\ldots,m} x^T P_i x$. Since $g(x)$ and $h(x)$ have the same 1-sublevel set and the same degree of homogeneity, we have $g(x) = h(x)$ for every $x \in \mathbb{R}^2$. It is straightforward to verify that $h(x_1, 1)$ is not piecewise-quadratic while $g(x_1, 1)$ is. This is a contradiction. □

4.2 Are GEs SOCP-representable?

Since second-order order cone programs (SOCPs) are more expressive than convex quadratically constrained programs, it is natural to ask if GEs are SOCP-representable sets. For $n \geq 1$, recall the definition of the n-dimensional second-order cone:

$$L_n := \left\{ x \in \mathbb{R}^n \;\middle|\; \sqrt{\sum_{i=1}^{n-1} x_i^2} \leq x_n \right\}.$$

A set $\Omega \subseteq \mathbb{R}^n$ is *SOCP-representable* if for some nonnegative integers k, m, a cone $K \subset \mathbb{R}^m$ which is the product of second-order cones, some matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{m \times k}$, and some vector $b \in \mathbb{R}^m$, one can write

$$\Omega = \{x \in \mathbb{R}^n \mid \exists u \in \mathbb{R}^k \text{ s.t. } Ax + Bu + b \in K\}.$$

See [14] for a reference on properties of SOCP-representable sets. Since sets defined by convex quadratic constraints are SOCP-representable, we already know from Proposition 4.1 that GE-0s and GE-1s are SOCP-representable. The next theorem shows that this is not the case for all GE-ds.

Theorem 4.2. *There exists a GE-16 in dimension 9 that is not SOCP-representable.*

Our proof relies on the following result from [23].

Theorem 4.3 (Corollary 1 of [23])**.** *The set of nonnegative univariate polynomials of degree (at most) 4 is not SOCP-representable.*

Proof of Theorem 4.2. Let $\phi(t) := (1, t, t^2, \ldots, t^8)^T$. Define $P(t) := \phi(t)\phi(t)^T$ and

$$\Omega := \{x \in \mathbb{R}^9 \mid x^T P(t) x \leq 1 \quad \forall t \in [-1,1]\}.$$

Observe that Ω is a valid GE as $P(t) \succeq 0$ for all $t \in [-1,1]$ and the kernel condition is satisfied. To see that the latter claim, suppose for the sake of contradiction that there exists a nonzero vector $y \in \mathbb{R}^9$ such that $P(t)y = 0$ for all $t \in [-1,1]$. This would imply that the univariate polynomial $y^T\phi(t)$ vanishes for all $t \in [-1,1]$, which can only happen if all its coefficients are zero, hence contradicting the fact that y was a nonzero vector.

We claim that Ω is not SOCP-representable. Let $\mathbb{R}_d[t]$ denote the set of polynomials with real coefficients in the variable t of degree at most d. We identify a vector $x \in \mathbb{R}^9$ with the degree-8 polynomial $p(t) = x^T\phi(t)$. Using this identification, we can write

$$\Omega = \{p \in \mathbb{R}_8[t] \mid |p(t)| \leq 1 \quad \forall t \in [-1,1]\}.$$

Suppose for the sake of contradiction that Ω was SOCP-representable. Consider the affine map $f : \mathbb{R}_4[t] \to \mathbb{R}_8[t]$ which maps a degree-4 polynomial $q(t)$ to the degree-8 polynomial

$$p(t) = 2(1-t^2)^4 q\left(\frac{2t}{1-t^2}\right) - 1.$$

Since taking the inverse image of a set under an affine map preserves SOCP-representability (see [14, 2.3.D]), it follows that the set

$$f^{-1}(\Omega) = \left\{ q \in \mathbb{R}_4[t] \;\middle|\; \left| 2(1-t^2)^4 q\left(\frac{2t}{1-t^2}\right) - 1 \right| \leq 1 \quad \forall t \in [-1,1] \right\}$$

is SOCP-representable. Note that $t \to \frac{2t}{1-t^2}$ maps the interval $(-1,1)$ to the entire real line. Furthermore, this map possesses an inverse, $s \to \frac{\sqrt{s^2+1}-1}{s}$. Now, for any polynomial $q \in \mathbb{R}_4[t]$, we have

$$\begin{aligned}
&\left| 2(1-t^2)^4 q\left(\frac{2t}{1-t^2}\right) - 1 \right| \leq 1 \quad \forall t \in [-1,1] \\
\Leftrightarrow \quad & 0 \leq (1-t^2)^4 q\left(\frac{2t}{1-t^2}\right) \leq 1 \quad \forall t \in [-1,1] \\
\Leftrightarrow \quad & 0 \leq \left(1 - \left(\frac{\sqrt{s^2+1}-1}{s}\right)^2\right)^4 q(s) \leq 1 \quad \forall s \in \mathbb{R} \\
\Leftrightarrow \quad & 0 \leq q(s) \leq g(s) \quad \forall s \in \mathbb{R},
\end{aligned}$$

where $g(s) := (1 - (\frac{\sqrt{s^2+1}-1}{s})^2)^{-4}$. Hence,

$$f^{-1}(\Omega) = \{q \in \mathbb{R}_4[s] \mid 0 \leq q(s) \leq g(s) \ \forall s \in \mathbb{R}\}.$$

By [14, Proposition 2.3.1], it follows that the set

$$\left\{ (q, M) \in \mathbb{R}_4[s] \times \mathbb{R} \;\middle|\; M > 0, \ 0 \leq \frac{q(s)}{M} \leq g(s) \ \forall s \in \mathbb{R} \right\}$$

is SOCP-representable. Since the map $(q, M) \to q$ is affine, it follows (see, e.g., [14, 2.3.C]) that the set

$$\{q \in \mathbb{R}_4[s] \mid \exists M > 0 \text{ s.t. } 0 \leq q(s) \leq M g(s) \ \forall s \in \mathbb{R}\}$$

is also SOCP-representable.

One can check that $g(s) \geq \max(1, \frac{s^4}{16})$ for all $s \in \mathbb{R}$. Therefore, for all $q \in \mathbb{R}_4[s]$, there exists some $M > 0$ such that $q(s) \leq M g(s)$ for all $s \in \mathbb{R}$. Thus, the above set is equal to the set

$$\{q \in \mathbb{R}_4[s] \mid q(s) \geq 0 \ \forall s \in \mathbb{R}\}.$$

SOCP-representability of this set contradicts Theorem 4.3. □

4.3 Are GEs SDP-representable?

Since semidefinite programs are more expressive than second-order cone programs, it is natural to ask if GEs are SDP-representable sets. In this section, we answer this question in the affirmative. Recall that a set $\Omega \subseteq \mathbb{R}^n$ is *SDP-representable* if

$$\Omega = \left\{ x \in \mathbb{R}^n \;\middle|\; \exists y \in \mathbb{R}^k \text{ s.t. } A_0 + \sum_{i=1}^n x_i A_i + \sum_{i=1}^k y_i B_i \succeq 0 \right\}$$

for some integer $k \geq 0$ and symmetric $m \times m$ matrices $A_0, A_1, \ldots, A_n, B_1, \ldots, B_k$.

Theorem 4.4. *Every GE is an SDP-representable set.*

Proof. We provide two different SDP representations, one which explicitly uses the decomposition in Theorem 3.4, and one which implicitly uses the psd condition of Definition 2.1. The second representation will typically be of smaller size, but the first one can be smaller when the polynomial matrix $P(t)$ has a low-rank decomposition. See Remark 2 for a more precise comparison.

SDP representation 1: Let $\mathcal{E}_d \subset \mathbb{R}^n$ be a GE-d defined by the $n \times n$ polynomial matrix $P(t)$ of degree d. Since $P(t) \succeq 0$ for all $t \in [-1, 1]$, recalling the definition of an sos-matrix, by Theorem 3.4 we can write:

$$P(t) = B(t)^T B(t) + (1 - t^2) C(t)^T C(t)$$

if d is even and

$$P(t) = (t + 1) B(t)^T B(t) + (1 - t) C(t)^T C(t)$$

if d is odd, where $B(t)$ and $C(t)$ are univariate polynomial matrices of respective sizes $r_1 \times n$ and $r_2 \times n$. As shown in [19, Theorem 7.1], there always exists a decomposition such that $r_1, r_2 \leq 2n$. The degrees of $B(t)$ and $C(t)$ are respectively $\frac{d}{2}$ and $\frac{d}{2} - 1$ if d is even, or both equal to $\frac{d-1}{2}$ if d is odd.

Observe that by taking Schur complements, for any $x \in \mathbb{R}^n$ and $t \in [-1, 1]$, we have

$$x^T P(t) x \leq 1 \Leftrightarrow M_x(t) := \begin{bmatrix} I_{r_1} & 0 & B(t)x \\ 0 & (1-t^2) I_{r_2} & (1-t^2) C(t) x \\ x^T B(t)^T & (1-t^2) x^T C(t)^T & 1 \end{bmatrix} \succeq 0$$

for d even, or

$$x^T P(t) x \leq 1 \Leftrightarrow M_x(t) := \begin{bmatrix} (1+t) I_{r_1} & 0 & (1+t) B(t)x \\ 0 & (1-t) I_{r_2} & (1-t) C(t) x \\ (1+t) x^T B(t)^T & (1-t) x^T C(t)^T & 1 \end{bmatrix} \succeq 0$$

for d odd. Then, we can write

$$\mathcal{E}_d = \{x \in \mathbb{R}^n \mid M_x(t) \succeq 0 \ \forall t \in [-1, 1]\}.$$

Observe that in both cases, $M_x(t)$ is a univariate polynomial matrix whose coefficients depend affinely on x. Thus, in view of Proposition 3.5, the constraint that $M_x(t) \succeq 0 \ \forall t \in [-1, 1]$ can be reduced to affine and semidefinite constraints on x. It follows that $\mathcal{E}_d$ is an SDP-representable set.

SDP representation 2: Let $\mathcal{E}_d \subset \mathbb{R}^n$ be a GE-d defined by the $n \times n$ polynomial matrix $P(t)$ of degree d. We claim that $\mathcal{E}_d = \bar{\mathcal{E}}_d$, where

$$\bar{\mathcal{E}}_d := \left\{ x \in \mathbb{R}^n \;\middle|\; \exists X \in S^n \text{ s.t. } \begin{bmatrix} X & x \\ x^T & 1 \end{bmatrix} \succeq 0, \quad \operatorname{Tr}(X P(t)) \leq 1 \ \ \forall t \in [-1, 1] \right\}.$$

The set $\bar{\mathcal{E}}_d$ is clearly SDP-representable since the constraint that $\operatorname{Tr}(XP(t)) \leq 1$ for all $t \in [-1, 1]$ can be reformulated as an SDP constraint in view of Proposition 3.5 (applied to the scalar-valued polynomial $1 - \operatorname{Tr}(XP(t))$).

To see that $\mathcal{E}_d = \bar{\mathcal{E}}_d$, first observe that any $x \in \mathcal{E}_d$ also belongs to $\bar{\mathcal{E}}_d$ since the vector x and the matrix $X = xx^T$ satisfy the constraints of $\bar{\mathcal{E}}_d$. Conversely, for any $x \in \bar{\mathcal{E}}_d$, fix a matrix $X \in S^n$ such that $\begin{bmatrix} X & x \\ x^T & 1 \end{bmatrix} \succeq 0$ and $\operatorname{Tr}(XP(t)) \leq 1$ for all $t \in [-1, 1]$. By taking a Schur complement, the first constraint implies that $X \succeq xx^T$. For all $t \in [-1, 1]$, since $P(t) \succeq 0$, we have

$$x^T P(t) x = \operatorname{Tr}(xx^T P(t)) \leq \operatorname{Tr}(XP(t)) \leq 1,$$

and therefore $x \in \mathcal{E}_d$. □

Remark 2. Following the proof of Theorem 4.4 and Proposition 3.5, one can check that the size of the semidefinite constraints necessary to represent a GE-d in dimension n via the first construction in the proof of Theorem 4.4 is

$$\begin{cases} (\frac{d}{4}+1)(r_1+r_2+1) & d \equiv 0 \mod 4 \\ (\frac{d+3}{4})(r_1+r_2+1) & d \equiv 1 \mod 4 \\ (\frac{d+2}{4}+1)(r_1+r_2+1) & d \equiv 2 \mod 4 \\ (\frac{d+1}{4}+1)(r_1+r_2+1) & d \equiv 3 \mod 4. \end{cases}$$

Given that by [19, Theorem 7.1] we can always take $r_1, r_2 \leq 2n$, the size of the semidefinite constraints grows only linearly with d (resp. with n) for fixed n (resp. fixed d).

Similarly, one can check that the size of the semidefinite constraints necessary to represent a GE-d in dimension n via the second construction in the proof of Theorem 4.4 is

$$\begin{cases} \max\left\{n+1, \frac{d}{2}+1\right\} & d \text{ even} \\ \max\left\{n+1, \frac{d+1}{2}\right\} & d \text{ odd.} \end{cases}$$

Thus, the size of the semidefinite constraints grows only linearly with $\max\{n, d\}$.

By contrast, to our knowledge, sublevel sets of n-variate homogeneous convex polynomials of degree d (which would also serve as a natural generalization ellipsoids) are not known to have a semidefinite representation [27]. If one instead works with "sos-convex" polynomials (a stronger condition), then the sublevel sets do have a semidefinite representation (see, e.g., [27, 31]), but the size of the semidefinite constraint would be $\binom{n+\frac{d}{2}}{\frac{d}{2}}$.

Remark 3. Note that the second SDP representation in the proof of Theorem 4.4 is directly in terms of the polynomial matrix $P(t)$. The first SDP representation, however, is in terms of polynomial matrices $B(t)$ and $C(t)$ that appear in the decomposition of the polynomial matrix $P(t)$. In some situations (e.g., the application in Section 6.4, the construction in the proof of Theorem 4.2, or when dealing with factor models), the matrix $P(t)$ appears already in decomposed form. In situations where this is not the case, there are multiple ways of obtaining polynomial matrices $B(t)$ and $C(t)$ from the polynomial matrix $P(t)$.

One option is to solve the SDP in Proposition 3.5 to find psd matrices Q_1 and Q_2 in a sum of squares decomposition of $y^T X_i(t) y$, for $i = 1, 2$, associated with the sos-matrices $X_1(t), X_2(t)$ that appear in Theorem 3.4. By performing a matrix decomposition of the form $Q_i = L_i^T L_i$, $i = 1, 2$, we can readily get an expression for $B(t)$ and $C(t)$; see, e.g., [44, Lemma 1]. We note that there are alternative ways of factoring univariate sos-matrices; e.g., by using the algorithm proposed in [10], or by following the proof of [19, Theorem 7.1].

Another option is to directly find a decomposition of the polynomial matrix $Q(t) := (t^2+1)^d P(\frac{t^2-1}{t^2+1})$ as $R(t)^T R(t)$ for a polynomial matrix $R(t)$. Such a decomposition must exist (since $Q(t) \succeq 0$ for all $t \in \mathbb{R}$) and can be found by methods mentioned in the previous paragraph. One can then convert the decomposition of $Q(t)$ into a suitable decomposition of $P(t)$, e.g., by following the proof of [37, Theorem 6.11].

4.3.1 Distance between two GEs

As an application of Theorem 4.4, we show here how one can compute the distance between two GEs. The problem of computing distances between two convex sets has applications in many areas, for example robotics, computer-aided design, and computer graphics [24, 6].

As a concrete example, consider the following two GEs

$$\mathcal{E}^y = \{x \in \mathbb{R}^n \mid (x - c_y)^T P_y(t)(x - c_y) \leq 1 \ \forall t \in [-1, 1]\},$$

$$\mathcal{E}^z = \{x \in \mathbb{R}^n \mid (x - c_z)^T P_z(t)(x - c_z) \leq 1 \ \forall t \in [-1, 1]\},$$

where

$$P_y(t) = \begin{bmatrix} 6-4t & 2-3t & 1-t \\ 2-3t & 3-t & -t \\ 1-t & -t & 2 \end{bmatrix}, \quad P_z(t) = \begin{bmatrix} 2-t^2 & t & 0 \\ t & 3-t^2 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad c_y = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad c_z = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}.$$

The distance between $\mathcal{E}^y$ and $\mathcal{E}^z$ is equal to

$$\begin{array}{ll} \min\limits_{y,z} & ||y-z||_2 \\ \text{s.t.} & y \in \mathcal{E}^y \\ & z \in \mathcal{E}^z \end{array} \quad = \quad \begin{array}{ll} \min\limits_{y,z} & ||y-z||_2 \\ \text{s.t.} & (y-c_y)^T P_y(t)(y-c_y) \leq 1 \ \forall t \in [-1,1] \\ & (z-c_z)^T P_z(t)(z-c_z) \leq 1 \ \forall t \in [-1,1]. \end{array}$$

As $\mathcal{E}^y$ is a GE-1, from the proof of Proposition 4.1, we have $y + c_y \in \mathcal{E}^y$ if and only if

$$y^T \begin{bmatrix} 10 & 5 & 2 \\ 5 & 4 & 1 \\ 2 & 1 & 2 \end{bmatrix} y \leq 1 \quad \text{and} \quad y^T \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} y \leq 1. \tag{4}$$

Since $P_z(t) \succeq 0$ for all $t \in [-1, 1]$ and its degree is even, by Theorem 3.4, we can find a representation of the form

$$P_z(t) = B(t)^T B(t) + (1 - t^2) C(t)^T C(t).$$

We can take for example

$$B(t) = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & -1 \\ \frac{t}{\sqrt{2}} & \sqrt{2} & 0 \end{bmatrix} \quad \text{and} \quad C(t) = \begin{bmatrix} 0 & 1 & 0 \\ \sqrt{\frac{3}{2}} & 0 & 0 \end{bmatrix}.$$

Following the first construction in the proof of Theorem 4.4, we have that $z + c_z \in \mathcal{E}^z$ if and only if

$$\begin{bmatrix} I_3 & 0 & \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & -1 \\ \frac{t}{\sqrt{2}} & \sqrt{2} & 0 \end{bmatrix} z \\ 0 & (1-t^2) I_2 & (1-t^2) \begin{bmatrix} 0 & 1 & 0 \\ \sqrt{\frac{3}{2}} & 0 & 0 \end{bmatrix} z \\ z^T \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & -1 \\ \frac{t}{\sqrt{2}} & \sqrt{2} & 0 \end{bmatrix}^T & (1-t^2) z^T \begin{bmatrix} 0 & 1 & 0 \\ \sqrt{\frac{3}{2}} & 0 & 0 \end{bmatrix}^T & 1 \end{bmatrix} \succeq 0 \quad \forall t \in [-1, 1]. \tag{5}$$

Thus, the problem of computing the distance between $\mathcal{E}_y$ and $\mathcal{E}_z$ can be reformulated as

$$\begin{array}{ll} \min\limits_{y,z} & \|(y + c_y) - (z + c_z)\| \\ \text{s.t.} & (4), (5). \end{array} \tag{6}$$

Alternatively, by the second construction in the proof of Theorem 4.4, we have that $z + c_z \in \mathcal{E}^z$ if and only if there is a matrix $Z \in S^3$ such that

$$\begin{bmatrix} Z & z \\ z^T & 1 \end{bmatrix} \succeq 0, \qquad \text{Tr}(Z P_z(t)) \leq 1 \ \ \forall t \in [-1, 1]. \tag{7}$$

Thus, the problem of computing the distance between $\mathcal{E}_y$ and $\mathcal{E}_z$ can be reformulated as

$$\begin{aligned} \min_{y,z,Z} \quad & \|(y + c_y) - (z + c_z)\| \\ \text{s.t.} \quad & (4), (7). \end{aligned} \tag{8}$$

In view of Proposition 3.5, both (6) and (8) are semidefinite programming problems. By solving either numerically, we find the distance between $\mathcal{E}_y$ and $\mathcal{E}_z$ to be 0.4635 to four digits of accuracy. The two GEs and a line segment connecting points in each set that attain this minimum distance are plotted in Figure 2.

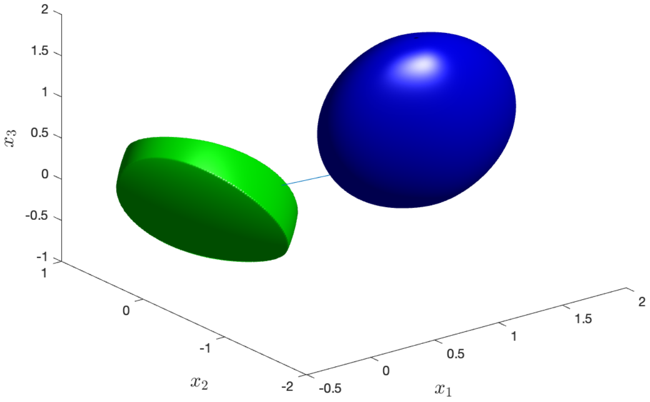


Figure 2: The distance between the GEs $\mathcal{E}^y$ and $\mathcal{E}^z$ in Section 4.3.1.

5 How Expressive are GEs?

We have already seen that GEs can express some nontrivial convex sets; e.g., the set of univariate polynomials $p : \mathbb{R} \to \mathbb{R}$ such that $|p(t)| \leq 1$ for $t \in [-1, 1]$ (see the proof of Theorem 4.2). In this section, we show that every symmetric full-dimensional polytope and every finite intersection of co-centered ellipsoids can be represented exactly as a GE (Corollary 5.6, following from Theorem 5.1). We then use this result to show that every convex body can be approximated arbitrarily well by a GE. We also quantify the quality of the approximation as a function of the degree of the GE (Theorem 5.7).

Let us begin with a definition. We call a set $T \subseteq \mathbb{R}^n$ a *semiellipsoid* if it can be written as $T = \{x \in \mathbb{R}^n \mid x^T P x \leq 1\}$ for some positive semidefinite matrix P. Note that a compact semiellipsoid is always a GE-0.

Theorem 5.1. *For every integer $m \geq 2$, a compact intersection of m semiellipsoids is a GE-d with $d \leq 2m - 3$.*

The proof of Theorem 5.1 relies on the following lemma, which could be of independent interest. The lemma states that in any dimension m, there exists a polynomial curve that stays within the unit simplex in $\mathbb{R}^m$ and visits every corner.

Lemma 5.2 (Polynomial tour of the simplex)**.** *For every integer $m \geq 2$, there exist univariate polynomials $p_1, \ldots, p_m$ of degree at most $2m - 3$ satisfying*

1. $p_i(t) \geq 0 \quad \forall t \in [-1, 1], \quad i = 1, \ldots, m,$
2. $\sum_{i=1}^m p_i(t) = 1 \quad \forall t \in [-1, 1]$, *and*
3. *for every $i = 1, \ldots, m, \exists t_i \in [-1, 1]$ such that $p_i(t_i) = 1$.*

The proof of this lemma utilizes a nonconstructive argument inspired by game theory and may be of independent interest as a proof technique. We note that all but two of the m polynomials can be taken to have degree at most $2m-4$, but the degree bound of $2m-3$ in Lemma 5.2 is the lowest possible (see Lemma 5.5). We first recall the following result from game theory.

Theorem 5.3 (Debreu, Glicksberg, Fan; see, e.g., [21])**.** *Consider a game with N players indexed by $i = 1, \dots, N$. Suppose that for each i, player i chooses an action a_i from a nonempty, compact, and convex set $A_i \subseteq \mathbb{R}^M$ and receives a payoff of $u_i(a_1, \dots, a_N)$. If for each i, the function u_i is continuous in $a_1, \dots, a_N$ and quasiconcave in a_i (over A_i), then the game possesses a pure-strategy Nash equilibrium; i.e., there exist actions $\bar{a}_1 \in A_1, \dots, \bar{a}_N \in A_N$, such that for every $i = 1, \dots, N$, we have*

$$u_i(\bar{a}_1, \dots, \bar{a}_N) = \max_{a_i \in A_i} u_i(a_i, \bar{a}_{-i}), \tag{9}$$

where the index $-i$ represents all players besides player i.

Proof of Lemma 5.2. We set up a game whose pure-strategy Nash equilibria correspond to roots of polynomials that satisfy the three conditions in the lemma. Consider the following game with players indexed by $i = 1, \dots, m$ and action sets $A_i \subseteq [-1, 1]$. Players 1 and m must respectively play -1 and 1 (i.e., $A_1 = \{-1\}$ and $A_m = \{1\}$), and their payoffs are taken to be 0. For $i = 2, \dots, m-1$, player i chooses an action $a_i \in [-1, 1]$ and receives a payoff of

$$u_i(a) = -f(a_{i-1} - a_i) - f(a_i - a_{i+1}) + (1 - a_i^2) \prod_{1<j<i} f(a_i - a_j) \prod_{i<j<m} f(a_j - a_i), \tag{10}$$

where $a := (a_1, \dots, a_m)$ and the function f is defined as

$$f(x) = \begin{cases} 0 & x < 0 \\ x^2 & x \geq 0 \end{cases}.$$

The first two terms in the payoff function incentivize player i to take an action a_i that belongs to the interval (a_{i-1}, a_{i+1}). The third term essentially incentivizes player i to increase the geometric mean of the deviations between her action and the actions of the others. This third term will soon be used when we construct the polynomials desired by the lemma. As an example, a plot of $u_3(a_3, a_{-3})$ is shown in Figure 3 for the case $m = 5$ and where a_{-3} corresponds to the actions of players $1, 2, 4, 5$ at a pure-strategy Nash equilibrium.

The functions u_i are continuous in a, since the function f is continuous. Next, we claim that for each $i = 2, \dots, m-1$, the function u_i is quasiconcave in a_i. Fix some i and a_{-i}, and consider u_i as a function of only a_i. Consider three cases:

- $a_{i-1} \geq a_{i+1}$: In this case, the second derivative of u_i with respect to a_i is -4 in the interval (a_{i+1}, a_{i-1}) and -2 outside of the interval $[a_{i+1}, a_{i-1}]$. Thus, u_i is strictly concave with respect to a_i.
- $a_{i-1} < a_{i+1}$ and $\max_{j<i} a_j \geq \min_{j>i} a_j$: In this case, u_i is equal to $-f(a_{i-1} - a_i) - f(a_i - a_{i+1})$. Given that $-f$ is concave, it follows that u_i is concave.
- $a_{i-1} < a_{i+1}$ and $\max_{j<i} a_j < \min_{j>i} a_j$: In this case, we first claim that u_i is quasiconcave when $a_i \in [\max_{j<i} a_j, \min_{j>i} a_j]$ where u_i is equal to $(1 - a_i^2) \prod_{j \notin \{1,i,m\}} (a_i - a_j)^2$. Since this polynomial is real-rooted, by interlacing, it must have a root between any two roots of its derivative. Thus, in this interval, u_i is either quasiconvex or quasiconcave in a_i (see, e.g., [18, Section 3.4.2]). Since u_i

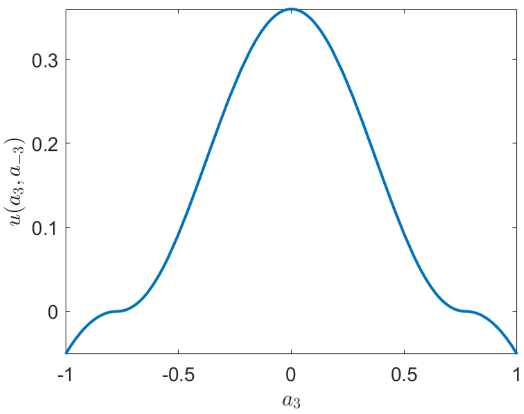


Figure 3: The payoff function $u_3(a_3, a_{-3})$ of the third player for the case $m = 5$ and where $a_{-3} = (a_1, a_2, a_4, a_5) = (-1, -\sqrt{0.6}, \sqrt{0.6}, 1)$ corresponds to the actions of the other players at a pure-strategy Nash equilibrium for the game appearing in the proof of Lemma 5.2.

> vanishes at $\max\limits_{j<i} a_j$ and $\min\limits_{j>i} a_j$ and is positive in between, it must be quasiconcave over the interval $[\max\limits_{j<i} a_j, \min\limits_{j>i} a_j]$. To extend the quasiconcavity argument to all of $[-1, 1]$, observe that u_i vanishes in the intervals $[a_{i-1}, \max\limits_{j<i} a_j]$ and $[\min\limits_{j>i} a_j, a_{i+1}]$. As u_i is nonpositive outside of the interval (a_{i-1}, a_{i+1}), our previous argument implies that for any $\alpha > 0$, the α-superlevel set of u_i is convex. In addition, for any $\alpha \leq 0$, we have $\{a_i \in \mathbb{R} \mid u_i(a_i, a_{-i}) \geq \alpha\} = [a_{i-1} - \sqrt{-\alpha}, a_{i+1} + \sqrt{-\alpha}]$. Thus, u_i is quasiconcave in a_i.

By Theorem 5.3 (applied with $N = m, M = 1$), there exist actions $\bar{a}_1, \dots, \bar{a}_m \in [-1, 1]$, such that for every $i = 1, \dots, m$ we have

$$u_i(\bar{a}_1, \dots, \bar{a}_m) = \max_{a_i \in [-1,1]} u_i(a_i, \bar{a}_{-i}). \tag{11}$$

Fix such actions $\bar{a} := (\bar{a}_1, \dots, \bar{a}_m)$. We next claim that $\bar{a}_1 < \bar{a}_2 < \cdots < \bar{a}_m$. First observe that by the definition of the payoff function u_i in (10) and in view of (11), for every $i = 2, \dots, m-1$, we have

- if $\bar{a}_{i-1} > \bar{a}_{i+1}$, then $\bar{a}_i \in (\bar{a}_{i+1}, \bar{a}_{i-1})$,
- if $\bar{a}_{i-1} \leq \bar{a}_{i+1}$, then $\bar{a}_i \in [\bar{a}_{i-1}, \bar{a}_{i+1}]$,
  - if we further have $\max\limits_{j<i} \bar{a}_j = \bar{a}_{i-1} < \bar{a}_{i+1} = \min\limits_{j>i} \bar{a}_j$, then $\bar{a}_i \in (\bar{a}_{i-1}, \bar{a}_{i+1})$.

We first show that $\bar{a}_1 \leq \bar{a}_2 \leq \cdots \leq \bar{a}_m$ by induction. Since $\bar{a}_1 = -1$ and $\bar{a}_2 \in [-1, 1]$, we have $\bar{a}_1 \leq \bar{a}_2$. Now assume by induction that $\bar{a}_{i-1} \leq \bar{a}_i$ for some $i < m$. Suppose for the sake of contradiction that $\bar{a}_i > \bar{a}_{i+1}$. If $\bar{a}_{i-1} > \bar{a}_{i+1}$, we have $\bar{a}_i \in (\bar{a}_{i+1}, \bar{a}_{i-1})$ and in particular $\bar{a}_i < \bar{a}_{i-1}$, contradicting the inductive hypothesis. If $\bar{a}_{i-1} \leq \bar{a}_{i+1}$, we have $\bar{a}_i \in [\bar{a}_{i-1}, \bar{a}_{i+1}]$ and in particular, $\bar{a}_i \leq \bar{a}_{i+1}$, a contradiction. Thus, we have $\bar{a}_i \leq \bar{a}_{i+1}$ and by induction $\bar{a}_1 \leq \cdots \leq \bar{a}_m$.

Suppose for the sake of contradiction that (at least) two actions are equal. Let $k \in \{1, \dots, m\}$ be the smallest index corresponding to an action in a set of equal actions of maximum cardinality. Either $\bar{a}_k \neq -1$ or $\bar{a}_k \neq 1$. Assume without loss of generality that $\bar{a}_k \neq -1$. Since $\bar{a}_1 = -1$, we must have $k > 1$. By minimality of k, we must have $\bar{a}_{k-1} < \bar{a}_k$. Since we have $\bar{a}_1 \leq \bar{a}_2 \leq \cdots \leq \bar{a}_m$, we have that $\max\limits_{j<k} \bar{a}_j = \bar{a}_{k-1}$ and $\bar{a}_{k+1} = \min\limits_{j>k} \bar{a}_j$. Since we further have $\bar{a}_{k-1} < \bar{a}_k = \bar{a}_{k+1}$, by the last (sub)-bullet above, we must have that $\bar{a}_k \in (\bar{a}_{k-1}, \bar{a}_{k+1})$, and in particular $\bar{a}_k < \bar{a}_{k+1}$, contradicting the fact that the equal action set has size at least two. Thus, $\bar{a}_1 < \cdots < \bar{a}_m$.

Going back to the statement of the lemma, we define, for $i = 1, \ldots, m$, $t_i = \bar{a}_i$ and

$$p_i(t) = \frac{q_i(t)}{q_i(t_i)},$$

where

$$q_1(t) = (1-t) \prod_{j \notin \{1,m\}} (t - t_j)^2, \qquad q_m(t) = (1+t) \prod_{j \notin \{1,m\}} (t - t_j)^2,$$

$$q_i(t) = (1-t^2) \prod_{j \notin \{1,i,m\}} (t - t_j)^2 \qquad i = 2, \ldots, m-1.$$

Note that since no two members of $\{\bar{a}_1, \ldots, \bar{a}_m\}$ coincide, we have $q_i(t_i) \neq 0$ for $i = 1, \ldots, m$, and therefore the polynomials $p_i(t)$ are well defined. With this construction, property 3 of Lemma 5.2 is immediate from the definition of $p_i(t)$ as $p_i(t_i) = 1$. Property 1 is also straightforward to check since each $p_i(t)$ is a positive scaling of a product of squares multiplied by either $(1-t)$,$(1+t)$, or $(1-t^2)$ which are all nonnegative for $t \in [-1, 1]$.

It remains to verify property 2. Fix an arbitrary index $i \in \{2, \ldots, m-1\}$. First observe that since $\bar{a}_1 \leq \bar{a}_2 \leq \cdots \leq \bar{a}_m$, for $a_i \in (\bar{a}_{i-1}, \bar{a}_{i+1})$, the payoff function $u_i(a_i, \bar{a}_{-i}) = q_i(a_i)$. Then by (11), the point $t_i = \bar{a}_i$ is a local maximum of $q_i(t)$ and thus also of $p_i(t)$. Combined with the fact that $p_i(t_i) = 1$, it follows that the point t_i is a double root of $1 - p_i(t)$. Additionally, for $k \in \{1, \ldots, i-1, i+1, \ldots, m\}$, t_i is a double root of $p_k(t)$ by construction. Hence, t_i is a double root of $1 - \sum_{\ell=1}^m p_\ell(t)$. The polynomial $1 - \sum_{\ell=1}^m p_\ell(t)$ also has roots at -1 and 1. Thus, counting with multiplicity, $1 - \sum_{\ell=1}^m p_\ell(t)$ has $2m-2$ roots. However, since for $\ell = 1, \ldots, m$, each $p_\ell(t)$ has degree at most $2m-3$, the degree of $1 - \sum_{\ell=1}^m p_\ell(t)$ is at most $2m-3$. Since a degree-d nonzero polynomial has at most d roots, it follows that $1 - \sum_{\ell=1}^m p_\ell(t)$ is identically zero. Therefore, $\sum_{\ell=1}^m p_\ell(t) = 1$ for all t. □

Figure 4 demonstrates polynomials $p_i(t)$ that satisfy the three conditions of Lemma 5.2 for the case $m = 5$. These polynomials were found by a numerical search for a Nash equilibrium for the game set up in the proof of Lemma 5.2. Interestingly, this equilibrium corresponds to the roots of the Legendre polynomial of degree 3. The next lemma proves that the roots of Legendre polynomials (see, e.g., [1, Chapter 22] for a definition) always provide a Nash equilibrium to our game.[4]

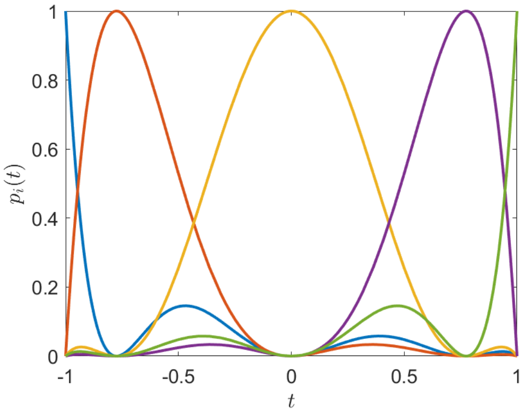


Figure 4: Polynomials $p_i(t)$ satisfying the three conditions of Lemma 5.2 for $m = 5$.

[4] In view of the explicit construction of a Nash equilibrium in Lemma 5.4, we do not actually need to invoke Theorem 5.3 for our purposes. However, we include Theorem 5.3 to highlight that the game theory approach could potentially be used more generally to prove existence of polynomials with certain desired properties even when a Nash equilibrium cannot be constructed explicitly.

Lemma 5.4. *Let $\bar{a}_1 = -1$, $\bar{a}_m = 1$, and let $\bar{a}_2, \dots, \bar{a}_{m-1}$ be the roots of the Legendre polynomial of degree $m-2$, arranged in ascending order. Then the actions $\bar{a}_1, \dots, \bar{a}_m$ form a pure-strategy Nash equilibrium of the game defined in the proof of Lemma 5.2.*

Proof. Since the action sets of players 1 and m are singletons, it is trivial that (9) holds for $i = 1, m$. Fix an arbitrary index $i \in \{2, \dots, m-1\}$. From the definition of the payoff function u_i in (10), $u_i(a_i, \bar{a}_{-i})$ is only positive in the interval $(\bar{a}_{i-1}, \bar{a}_{i+1})$, where it is equal to the polynomial function $(1 - a_i^2) \prod_{j \notin \{1,i,m\}} (a_i - \bar{a}_j)^2$. Thus,

$$\max_{a_i \in [-1,1]} u_i(a_i, \bar{a}_{-i}) = \max_{a_i \in (\bar{a}_{i-1}, \bar{a}_{i+1})} u_i(a_i, \bar{a}_{-i}) = \max_{a_i \in (\bar{a}_{i-1}, \bar{a}_{i+1})} (1 - a_i^2) \prod_{j \notin \{1,i,m\}} (a_i - \bar{a}_j)^2.$$

This polynomial is zero at $\bar{a}_{i-1}$ and $\bar{a}_{i+1}$ and positive in the interval $(\bar{a}_{i-1}, \bar{a}_{i+1})$. As argued in the proof of Lemma 5.2, the derivative of this polynomial has at most one root over $(\bar{a}_{i-1}, \bar{a}_{i+1})$. Therefore, any root of the derivative in this interval attains the maximum. Thus, in order to show that

$$u_i(\bar{a}_i, \bar{a}_{-i}) = \max_{a_i \in (\bar{a}_{i-1}, \bar{a}_{i+1})} (1 - a_i^2) \prod_{j \notin \{1,i,m\}} (a_i - \bar{a}_j)^2,$$

if suffices to show that

$$\frac{d}{da_i} \left((1 - a_i^2) \prod_{j \notin \{1,i,m\}} (a_i - \bar{a}_j)^2 \right) \Bigg|_{a_i = \bar{a}_i} = 0.$$

To this end, we recall that due to the properties of the Legendre polynomials, the polynomial function $\ell_{m-2}(t) := \prod_{i=2}^{m-1} (t - \bar{a}_i)$ satisfies the Legendre differential equation (see, e.g., [1, 22.6.13]):

$$(1 - t^2) \ell''_{m-2}(t) - 2t \ell'_{m-2}(t) + (m-2)(m-1) \ell_{m-2}(t) = 0 \quad \forall t.$$

One can then check that

$$\begin{aligned} &\frac{d}{da_i} \left((1 - a_i^2) \prod_{j \notin \{1,i,m\}} (a_i - \bar{a}_j)^2 \right) \Bigg|_{a_i = \bar{a}_i} \\ &= \left(\prod_{j \notin \{1,i,m\}} (\bar{a}_i - \bar{a}_j) \right) \Big((1 - \bar{a}_i^2) \ell''_{m-2}(\bar{a}_i) - 2\bar{a}_i \ell'_{m-2}(\bar{a}_i) + (m-2)(m-1) \ell_{m-2}(\bar{a}_i) \Big) = 0. \end{aligned}$$ □

Lemma 5.5. *For $m \geq 2$, let $\{p_1, \dots, p_m\}$ be any set of polynomials satisfying the three properties of Lemma 5.2. Then at least one of these polynomials must have degree at least $2m - 3$.*

Proof. The claim is trivial for $m = 2$, so assume $m \geq 3$. Let $t_1, \dots, t_m \in [-1, 1]$ be any set of points such that $p_i(t_i) = 1$ for $i = 1, \dots, m$. The properties of Lemma 5.2 imply that the points $t_1, \dots, t_m$ are distinct and that $p_i(t_j) = 0$ for $i \neq j$. If -1 or 1 are in the set $\{t_1, \dots, t_m\}$, fix i such that $t_i \in \{-1, 1\}$; otherwise choose i arbitrarily. If both -1 and 1 are in the set $\{t_1, \dots, t_m\}$, fix j such that $\{t_i, t_j\} = \{-1, 1\}$; otherwise choose any $j \neq i$. For each $k \notin \{i, j\}$, since $p_i(t_k) = 0$, $p_i(t) \geq 0$ for $t \in [-1, 1]$, and $t_k \notin \{-1, 1\}$, t_k must be a root of $p_i(t)$ of multiplicity at least two. Furthermore, since t_j is a root of $p_i(t)$, counting with multiplicities, $p_i(t)$ must have at least $2(m-2) + 1 = 2m - 3$ roots. □

We now shift our focus back to GEs and show that they can represent any compact finite intersection of semiellipsoids.

Proof of Theorem 5.1. Let $T \subset \mathbb{R}^n$ be a compact finite intersection of m semiellipsoids. Then there exist matrices $P_1, \ldots, P_m \in S_+^n$ such that $T = \bigcap_{i=1}^m \{x \in \mathbb{R}^n \mid x^T P_i x \leq 1\}$. Let $\{p_1, \ldots, p_m\}$ be polynomials of degree at most $2m - 3$ satisfying the three properties in Lemma 5.2. Define the polynomial matrix $P(t)$ and the set $\bar{T}$ as follows:

$$P(t) = \sum_{i=1}^m p_i(t) P_i \quad \text{and} \quad \bar{T} = \bigcap_{t \in [-1,1]} \{x \in \mathbb{R}^n \mid x^T P(t) x \leq 1\}.$$

We claim that $\bar{T} = T$ and that $\bar{T}$ is a GE-d (with $d \leq 2m - 3$). To see that $\bar{T} \subseteq T$, recall first that, by Lemma 5.2, for each $i = 1, \ldots, m$, $\exists t_i \in [-1, 1]$ such that $p_i(t_i) = 1$. Then we have

$$\bar{T} = \bigcap_{t \in [-1,1]} \{x \in \mathbb{R}^n \mid x^T P(t) x \leq 1\} \subseteq \bigcap_{i=1}^m \{x \in \mathbb{R}^n \mid x^T P(t_i) x \leq 1\} = \bigcap_{i=1}^m \{x \in \mathbb{R}^n \mid x^T P_i x \leq 1\} = T.$$

To see that $T \subseteq \bar{T}$, take $x \in T$ and $t \in [-1, 1]$. Recall by Lemma 5.2 that $\sum_{i=1}^m p_i(t) = 1$. Then we have

$$x^T P(t) x = \sum_{i=1}^m p_i(t) x^T P_i x \leq \max_{i=1,\ldots,m} x^T P_i x \leq 1.$$

Thus, $x \in \bar{T}$. Finally, we note that $\bar{T}$ is a valid GE-d. The psd condition holds since for $i = 1, \ldots, m$, $p_i(t) \geq 0$ for all $t \in [-1, 1]$, and $P_i \succeq 0$. The kernel condition holds since $\bar{T}$ is compact (as T is compact). □

Corollary 5.6. *Every symmetric full-dimensional polytope and every finite intersection of co-centered ellipsoids is a GE.*

Proof. The claim for a finite intersection of co-centered ellipsoids is immediate from Theorem 5.1. Let $T \subset \mathbb{R}^n$ be a symmetric full-dimensional polytope. By translation, we may assume T is symmetric around the origin; from this and full-dimensionality, it follows that T contains the origin in its interior. Thus, we may write

$$T = \{x \in \mathbb{R}^n \mid |a_i^T x| \leq 1 \quad i = 1, \ldots, m\}$$

for some vectors $a_1, \ldots, a_m \in \mathbb{R}^n$. Therefore, the description of T can be rewritten as

$$T = \bigcap_{i=1}^m \{x \in \mathbb{R}^n \mid x^T (a_i a_i^T) x \leq 1\}.$$

As T is evidently a (compact) finite intersection of semiellipsoids, the claim follows from Theorem 5.1. □

We now show that every symmetric convex body can be approximated arbitrarily well by a GE.

Theorem 5.7. *There exists $\varepsilon_0 > 0$ such that for any $\varepsilon \in (0, \varepsilon_0)$ and for any symmetric convex body $C \subset \mathbb{R}^n$, there is a GE-d $\mathcal{E}_d$, with $d \leq 2\left(\frac{1}{2\sqrt{\varepsilon}} \log(\frac{1}{\varepsilon})\right)^n - 3$, satisfying*

$$\mathcal{E}_d \subseteq C \subseteq (1 + \varepsilon)\mathcal{E}_d. \tag{12}$$

Proof. Let $C^o := \{y \in \mathbb{R}^n \mid \langle x, y \rangle \leq 1\}$ be the polar dual of C. Since C^o is a convex body, by [13, Corollary 1.2], there exists $\varepsilon_0 > 0$ ($\varepsilon_0(\frac{1}{2})$ in the language of that paper) such that for any $\varepsilon \in (0, \varepsilon_0)$, there exists a symmetric (full-dimensional) polytope $T \subset \mathbb{R}^n$ with at most $\left(\frac{1}{2\sqrt{\varepsilon}} \log(\frac{1}{\varepsilon})\right)^n$ vertices that satisfies $T \subseteq C^o \subseteq (1 + \varepsilon)T$. By duality, we have $\frac{1}{1+\varepsilon} T^o \subseteq C \subseteq T^o$. Then, the set $\mathcal{E}_d := \frac{1}{1+\varepsilon} T^o$ satisfies (12). Since $\mathcal{E}_d$ is a scaling of the polar dual of T, it is symmetric, full-dimensional, and the number of its facets is at most $\left(\frac{1}{2\sqrt{\varepsilon}} \log(\frac{1}{\varepsilon})\right)^n$ (see, e.g., [12, Chap. 4]). By Corollary 5.6, $\mathcal{E}_d$ is a GE-d with $d \leq 2\left(\frac{1}{2\sqrt{\varepsilon}} \log(\frac{1}{\varepsilon})\right)^n - 3$. □

6 Applications

In this section, we present four potential applications involving GEs.

6.1 Minimum-variance portfolio optimization with time-varying covariance

In its simplest form, the minimum-variance portfolio optimization problem in finance takes the form

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & x^T \Sigma x \\ \text{s.t.} \quad & \mathbf{1}^T x = 1, \quad x \geq 0, \end{aligned}$$

where, for $i = 1, \ldots, n$, the i^{th} entry of the portfolio $x \in \mathbb{R}^n$ determines the fraction of our wealth that we invest in asset i. Here, $\mathbf{1}$ denotes the vector of all ones, the nonnegativity constraint on x is entrywise, and $\Sigma \in S^n_+$ is the covariance matrix of the underlying asset returns[5] over a fixed time period and is assumed to be known.[6]

We consider a generalization of this problem where we commit to a portfolio at the start of a period (say at time $t = -1$), but allow ourselves to liquidate the portfolio at any time within a given horizon (say at anytime $t \in [-1, 1]$). In this setting, the covariance matrix is no longer constant and depends on the liquidation time. In other words, there is a function $\Sigma : [-1, 1] \to S^n_+$, such that $\Sigma(t)$ is the covariance matrix of the asset returns at time t. Then to find a portfolio that minimizes the worst-case variance of the returns over all possible liquidation times, we must solve the problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \max_{t \in [-1,1]} \quad x^T \Sigma(t) x \\ \text{s.t.} \quad & \mathbf{1}^T x = 1, \quad x \geq 0. \end{aligned}$$

It is unreasonable to assume access to the function $\Sigma(t)$. Instead, we assume we have noisy measurements $\Sigma_1, \ldots, \Sigma_m \in S^n$ of this function at times $t_1, \ldots, t_m \in [-1, 1]$ during similar past periods. Given this data, one might consider minimizing the worst-case variance with respect to the measurements. This corresponds to solving

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \max_{i=1,\ldots,m} \quad x^T \hat{\Sigma}_i x \\ \text{s.t.} \quad & \mathbf{1}^T x = 1, \quad x \geq 0, \end{aligned} \tag{13}$$

where $\hat{\Sigma}_i$ is the nearest (e.g., in Frobenius distance) psd matrix to Σ_i.

As a model-based alternative, we propose to fit a polynomial matrix $P(t)$ to the measurements $\Sigma_1, \ldots, \Sigma_m$ by solving

$$\begin{aligned} \min_{P \in S^n_d[t]} \quad & \sum_{i=1}^{m} \| P(t_i) - \Sigma_i \|_F^2 \\ \text{s.t.} \quad & P(t) \succeq 0 \quad \forall t \in [-1, 1], \end{aligned} \tag{14}$$

where $S^n_d[t]$ denotes the set of symmetric $n \times n$ univariate polynomial matrices of degree (at most) d and $\| \cdot \|_F$ denotes the Frobenius norm. Note that the constraint in (14) ensures that our model

[5] The *return* of asset i over a period is given by $\frac{p^i_{\text{end}} - p^i_{\text{beg}}}{p^i_{\text{beg}}}$, where p^i_{beg} (resp. p^i_{end}) is the price of the asset at the beginning (resp. the end) of the period.

[6] In the Markowitz variant of this problem, one has an additional linear constraint that imposes a lower bound on the expected return of the portfolio. Such a constraint can be easily incorporated into our framework. However, our focus here is on the variance of the portfolio since we want to highlight how time-varying versions of convex quadratic programs can give rise to GEs.

produces a valid covariance matrix at all times. In view of Proposition 3.5, this constrained regression problem can be formulated as an SDP.

Let $P^*(t)$ be an optimal solution to (14). To find our portfolio, we propose to solve the following problem:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \max_{t \in [-1,1]} \quad x^T P^*(t) x \\ \text{s.t.} \quad & \mathbf{1}^T x = 1, \quad x \geq 0. \end{aligned} \tag{15}$$

This problem searches for a portfolio which has minimum GE-d-norm defined by $P^*(t)$ (see (3) in Section 2 for a definition). By Theorem 4.4, problem (15) can be formulated as an SDP.

6.1.1 Numerical example

As a numerical example, we consider a universe of $n = 10$ assets such that the covariance matrix of their returns at time t is given by the (non-polynomial) function

$$\Sigma(t) = 6\sin(t+1)A_1A_1^T + 2(1-t^2)A_2A_2^T + (t+1)^2 A_3A_3^T, \tag{16}$$

where the entries of the matrices $A_1, A_2, A_3 \in \mathbb{R}^{10\times 2}$ were generated independently and according to the standard Gaussian distribution. Note that $\Sigma(t) \succeq 0$ for $t \in [-1,1]$ and that $\Sigma(-1) = 0$ as there is no uncertainty in the return at the beginning of the period. As described before, we do not assume access to the function $\Sigma(t)$, but instead to $m = 500$ noisy measurements $\Sigma_1, \ldots, \Sigma_{500}$ of it at equally spaced times $t_1, \ldots, t_{500}$ in the interval $[-1,1]$. More specifically, we let $\Sigma_i = \Sigma(t_i) + Z_i$, where Z_i is a 10×10 symmetric matrix with upper triangular entries drawn independently from a Gaussian with mean zero and standard deviation 30.

In Figure 5, we compare the variance of four different portfolios with respect to the true covariance matrix $\Sigma(t)$ in (16). The curves correspond to $x_*^T \Sigma(t) x_*$, where $x_* \in \{x_*^{(13)}, x_*^{\text{GE}-0}, x_*^{\text{GE}-1}, x_*^{\text{GE}-2}\}$. Here, $x_*^{(13)}$ is optimal to (13), and $x_*^{\text{GE}-0}, x_*^{\text{GE}-1}, x_*^{\text{GE}-2}$ are optimal to (15) with $d = 0, 1, 2$, respectively. Note that to find the latter three portfolios, we first solve (14) for $d = 0, 1, 2$ to obtain the optimal matrix $P^*(t)$ that goes as input to (15). We observe that the GE-based portfolios have lower variance throughout time than the portfolio coming from the solution to (13). In this example, the improvement seems to saturate at degree $d = 2$. The worst-case variances of the four portfolios (with respect to the true covariance matrix $\Sigma(t)$) are reported in Table 1.

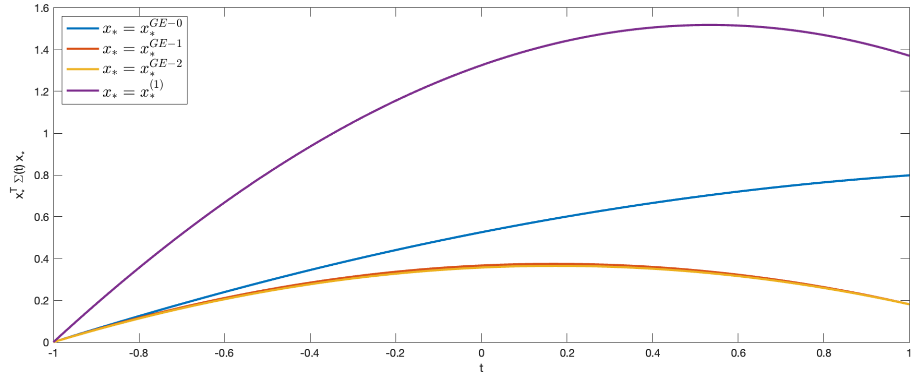


Figure 5: Comparing the variance of the four portfolios discussed in Section 6.1.1 with respect to the true covariance matrix $\Sigma(t)$.

	$x_* = x_*^{(13)}$	$x_* = x_*^{\mathrm{GE}-0}$	$x_* = x_*^{\mathrm{GE}-1}$	$x_* = x_*^{\mathrm{GE}-2}$
$\max\limits_{t\in[-1,1]} x_*^T\Sigma(t)x_*$	1.5179	0.7981	0.3745	0.3649

Table 1: The worst-case variance $\max\limits_{t\in[-1,1]} x_*^T\Sigma(t)x_*$ for four different portfolios x_*.

6.2 Joint spectral radius and stability of switched linear systems

In this section, we show that asymptotically stable switched linear systems always admit a GE as an invariant set. Equivalently, we show that if the joint spectral radius of a set of matrices is less than one, then there must exist a "contracting" GE-*d*-norm.

Recall that the *spectral radius* $\rho(A)$ of a matrix $A \in \mathbb{R}^{n\times n}$ is defined as

$$\rho(A) = \lim_{k\to\infty} ||A^k||^{1/k},$$

where $||\cdot||$ is any matrix norm. This quantity coincides with the maximum of the absolute values of the eigenvalues of A. The discrete-time linear dynamical system $x_{k+1} = Ax_k$, where $x_k \in \mathbb{R}^n$ is the state of the system at time $k \in \mathbb{N}$, is said to be *asymptotically stable* if for any starting state $x_0 \in \mathbb{R}^n$, $x_k \to 0$ as $k \to \infty$. It is straightforward to establish that a linear system is asymptotically stable if and only if $\rho(A) < 1$.

The *joint spectral radius* $\rho(\mathcal{A})$ of a set of matrices $\mathcal{A} := \{A_1, \ldots, A_m\} \subseteq \mathbb{R}^{n\times n}$ is defined as

$$\rho(\mathcal{A}) := \lim_{k\to\infty} \max_{\sigma\in\{1,\ldots,m\}^k} ||A_{\sigma_k}\ldots A_{\sigma_1}||^{1/k},$$

where $||\cdot||$ is any matrix norm [41]. Note that when $\mathcal{A}$ contains a single matrix, the definition of the joint spectral radius (JSR) simplifies to that of the spectral radius. However, computing the JSR is significantly more challenging than computing the spectral radius; for example the problem of testing if $\rho(\mathcal{A}) \le 1$ is undecidable already when $m = 2$ [17, 15]. The JSR has a close connection to stability of a discrete-time *switched* linear system, i.e., a dynamical system of the type $x_{k+1} = A_k x_k$, where the matrix $A_k \in \mathbb{R}^{n\times n}$ can vary arbitrarily in each iteration within the set $\mathcal{A}$. We say that a switched linear system is asymptotically stable if for any starting state $x_0 \in \mathbb{R}^n$ and any sequence of products of matrices in $\mathcal{A}$, $x_k \to 0$ as $k \to \infty$. One can show that this property holds if and only if $\rho(\mathcal{A}) < 1$; see, e.g., [30]. Therefore, much research has focused on providing conditions that guarantee the JSR is less than one. Theorem 6.2 below shows that GEs can always provide such a condition. This theorem can be seen as a direct generalization of the following classical result in linear systems theory.

Theorem 6.1 (see, e.g., Theorem 8.4 in [28]). *For a matrix $A \in \mathbb{R}^{n\times n}$, we have $\rho(A) < 1$ if and only if there exists a contracting quadratic norm; i.e., a function $V : \mathbb{R}^n \to \mathbb{R}$ of the form $V(x) = \sqrt{x^TQx}$, with $Q \succ 0$, such that $V(Ax) < V(x)$ $\forall x \neq 0$.*

Geometrically, the above theorem implies that if a linear system is asymptotically stable, then there is an ellipsoid (given by any sublevel set of the quadratic norm) that is invariant under the trajectories, i.e., trajectories starting in this ellipsoid remain in the ellipsoid for all time. It is well known that the existence of such an ellipsoid is no longer necessary for asymptotic stability of switched linear systems involving at least two matrices. The following theorem implies that the existence of an invariant GE is however a necessary condition. The theorem is stated in the language of GE-*d*-norms (recall the definition from (3) in Section 2).

Theorem 6.2. *For a set of matrices $\mathcal{A} := \{A_1, \ldots, A_m\} \subseteq \mathbb{R}^{n\times n}$, we have $\rho(\mathcal{A}) < 1$ if and only if there exists a contracting GE-d-norm; i.e., a function $V : \mathbb{R}^n \to \mathbb{R}$ of the form $V(x) = \max_{t\in[-1,1]} \sqrt{x^TP(t)x}$,* *where $P(t)$ is a univariate polynomial matrix of degree d satisfying the psd and the kernel conditions in Definition 2.1, such that $V(Ax) < V(x)$ $\forall x \neq 0$ and $\forall A \in \mathcal{A}$.*

The "if" direction of Theorem 6.2 follows from Lyapunov's global stability theorem; see, e.g., [7, Section 1] and references therein (in the language of that paper, our GE-d-norm is a "common" or "simultaneous" Lyapunov function). The "only if" direction of Theorem 6.2 can be shown by first invoking a nonconstructive converse Lyapunov theorem which states that if $\rho(\mathcal{A}) < 1$, then there exists a contracting norm; see, e.g., [41] or [30, page 24]. This abstract norm however can be approximated arbitrarily well by a GE-d-norm. This is a consequence of our Theorem 5.7, which proves that any symmetric convex body can be approximated arbitrarily well by a GE-d. The "only if" direction of Theorem 6.2 also follows from Theorem 6.3, which provides a quantitative version of the statement. This theorem generalizes the main result of [9, 16] (which corresponds to $l = 1$) from ellipsoids to generalized ellipsoids.

Theorem 6.3. *Let $\mathcal{A} := \{A_1, \dots, A_m\} \subseteq \mathbb{R}^{n\times n}$. For a positive integer ℓ, if $\rho(\mathcal{A}) < \frac{1}{\sqrt[2\ell]{n}}$, then there exists a contracting GE-d-norm with $d \leq \max\{2m^{\ell-1} - 3, 0\}$.*

Proof. Suppose $\rho(\mathcal{A}) < \frac{1}{\sqrt[2\ell]{n}}$. Then it follows from [7, Theorem 6.1] that there exist $m^{\ell-1}$ matrices $P_1, \dots, P_{m^{\ell-1}} \in S^n_{++}$ such that the function $W(x) = \max_{i\in\{1,\dots,m^{\ell-1}\}} \sqrt{x^T P_i x}$ satisfies $W(Ax) < W(x)$, $\forall x \neq 0$ and $\forall A \in \mathcal{A}$. If $m = 1$ or $\ell = 1$, the GE-0-norm $V(x) = \sqrt{x^T P_1 x}$ is evidently contracting. Now assume $m, \ell \geq 2$. It follows from Corollary 5.6 that there exists a polynomial matrix $P(t)$ of degree $d \leq 2m^{\ell-1} - 3$ such that the GE-d-norm $V(x) = \max_{t\in[-1,1]} \sqrt{x^T P(t) x} = W(x)$ for all $x \in \mathbb{R}^n$. The claim follows. □

6.2.1 An example

Consider the set of matrices $\mathcal{A}_\gamma = \{\gamma A_1, \gamma A_2\} \subseteq \mathbb{R}^{2\times 2}$ parameterized by a scalar $\gamma \geq 0$ and with

$$A_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \ A_2 = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix}.$$

These matrices that have been studied e.g. in [9] to demonstrate that the JSR can be less than one without existence of a contracting quadratic norm (recall the definition from the statement of Theorem 6.1). Indeed, one can show that $\rho(\mathcal{A}_\gamma) < 1$ for any $\gamma < 1$, while a contracting quadratic norm exists only when $\gamma < \frac{1}{\sqrt{2}}$. By contrast, we observe that a contracting GE-1-norm exists for any $\gamma < 1$.

Let

$$P(t) = \frac{1}{2} \begin{bmatrix} 1-t & 0 \\ 0 & 1+t \end{bmatrix}.$$

It is straightforward to verify that $P(t) \succeq 0$ $\forall t \in [-1, 1]$ and that $\bigcap_{t\in[-1,1]} \mathrm{Ker}(P(t)) = \{0\}$. Let

$$V(x) := \max_{t\in[-1,1]} \sqrt{x^T P(t) x} = \max_{t\in[-1,1]} \sqrt{\frac{1-t}{2} x_1^2 + \frac{1+t}{2} x_2^2} = \max\{|x_1|, |x_2|\}$$

be the associated GE-1-norm. We have $V(\gamma A_1 x) = \gamma|x_1|$ and $V(\gamma A_2 x) = \gamma|x_2|$. It follows that $V(\gamma A_i x) < V(x)$, $\forall \gamma < 1$, $\forall x \neq 0$, and for $i = 1, 2$.

6.3 Robust-to-dynamics optimization

A robust-to-dynamics optimization (RDO) problem is an optimization problem of the form

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & x, g(x), g(g(x)), \ldots \in \Omega, \end{aligned} \tag{17}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^n$, and $g : \mathbb{R}^n \to \mathbb{R}^n$ is a map that represents a dynamical system $x_{k+1} = g(x_k)$ with $k = 0, 1, 2, \ldots$ denoting the index of time. In words, the goal of the RDO problem is to optimize f over the set $\mathcal{S} \subseteq \Omega$ of initial conditions that forever remain in Ω under g. We refer the reader to [5] for more context and to [2] for applications of this problem to safe learning.

In [5], algorithms that provide tractable inner and outer approximations to the feasible set $\mathcal{S}$ of (17) are provided for certain subclasses of the RDO problem. A particular focus is on the case where Ω is a polyhedron and g is a linear map. More specifically, in this setting, $\Omega = \{x \in \mathbb{R}^n \mid Hx \leq 1\}$ where $H \in \mathbb{R}^{m\times n}$ is a given matrix and $g(x) = Ax$ where $A \in \mathbb{R}^{n\times n}$ is *stable*[7], i.e. has spectral radius less than one. Note that any polytope with the origin in its interior can be written in the form of Ω. We refer the reader to [5, Section 2.1.1] (and also [2, Proposition 16]) to see why the assumptions that Ω contains the origin in its interior and that A is stable are made. These assumptions are only slightly stronger than the natural requirement that $\mathcal{S}$ is not a measure-zero set.

In this section, we extend this setting to the case where the matrix A is unknown, but must belong to the convex hull of two given matrices $\hat{A}$ and $\check{A}$. The input to our problem is a matrix $H \in \mathbb{R}^{m\times n}$ representing the polytope $\Omega = \{x \in \mathbb{R}^n \mid Hx \leq 1\}$ and two matrices $\hat{A}, \check{A} \in \mathbb{R}^{n\times n}$. Given this input, we wish to characterize the set

$$\mathcal{S} := \{x \in \mathbb{R}^n \mid HA^k x \leq 1 \quad k = 0, 1, \ldots, \quad \forall A \in \operatorname{conv}(\hat{A}, \check{A})\}. \tag{18}$$

The approach that we present will also work in the more general setting where the matrix A is only known to belong to a given polynomial curve in matrix space. As is done in [5], outer approximations to $\mathcal{S}$ can be obtained by truncating the infinite time horizon, and in our case sampling from $\operatorname{conv}(\hat{A}, \check{A})$; we carry out an outer approximation of this form in the numerical example below. As is the case in [5], finding inner approximations to $\mathcal{S}$ are more challenging however. The following theorem shows how one can inner approximate $\mathcal{S}$ with a GE via semidefinite programming.

Theorem 6.4. *Let $\mathcal{S}$ be as in (18). If $\mathcal{E}_d$ is a GE-d defined by a polynomial matrix $P(t)$ of degree d which satisfies the constraints*

$$\begin{aligned} & P(t) - A(t)^T P(t) A(t) \succeq 0 \quad \forall t \in [-1, 1] \\ & P(t) \succeq h_i h_i^T \quad \forall t \in [-1, 1] \quad i = 1, \ldots, m, \end{aligned} \tag{19}$$

where $A(t) := \frac{1+t}{2}\hat{A} + \frac{1-t}{2}\check{A}$ and h_i^T is the i^{th} row of H, then $\mathcal{E}_d \subseteq \mathcal{S}$.

Proof. For each $t \in [-1, 1]$, define $\mathcal{E}_d(t) := \{x \in \mathbb{R}^n \mid x^T P(t) x \leq 1\}$. By the first constraint in (19), we have $x \in \mathcal{E}_d(t) \Rightarrow A(t)x \in \mathcal{E}_d(t)$. From this, it follows that $x \in \mathcal{E}_d(t) \Rightarrow A^k(t)x \in \mathcal{E}_d(t)$ for all $k \geq 0$. By the second constraint in (19), we have $x \in \mathcal{E}_d(t) \Rightarrow Hx \leq 1$. Then we have

$$x \in \mathcal{E}_d = \bigcap_{t\in[-1,1]} \mathcal{E}_d(t) \Rightarrow HA^k(t)x \leq 1 \quad \forall t \in [-1, 1] \quad k = 0, 1, \ldots,$$

which gives the desired result since the set $\{A(t) \mid t \in [-1, 1]\} = \operatorname{conv}(\hat{A}, \check{A})$. □

[7] A terminology more consistent with Section 6.2 would have been “asymptotically stable”, but in this section we drop the word asymptotic for simplicity.

We note that if Ω is compact and if any matrix in $\mathrm{conv}(\hat{A}, \check{A})$ has spectral radius more than one, then the set $\mathcal{S}$ in (18) will have measure zero [2, Proposition 16]. To avoid this situation, similarly to what is done in [5], we work with the assumption that all matrices in $\mathrm{conv}(\hat{A}, \check{A})$ are stable. Under this assumption, the following lemma ensures that there is always a suitable polynomial matrix $P(t)$ which satisfies the constraints of (19). The second constraint in (19) is not mentioned in this lemma since it can always be satisfied simply by scaling up the matrix $P(t)$. In the language of dynamical systems, this lemma states that if all matrices in the convex hull are stable, then there must exist a polynomially-varying quadratic Lyapunov function $x^T P(t) x$ for the associated linear dynamical systems.

Lemma 6.5 (Special case of Lemma 24 of [2])**.** *For two matrices $\hat{A}, \check{A} \in \mathbb{R}^{n\times n}$, every matrix in the set conv$(\hat{A}, \check{A})$ is stable if and only if there exists a polynomial matrix $P : \mathbb{R} \to S^n$ such that*

1. $P(t) \succ 0 \quad \forall t \in [-1, 1]$,
2. $P(t) - A(t)^T P(t) A(t) \succ 0 \quad \forall t \in [-1, 1]$,

where $A(t) := \frac{1+t}{2}\hat{A} + \frac{1-t}{2}\check{A}$.

We have just shown that the following optimization problem

$$\begin{aligned}
\min_{P \in S_d^n[t],\, \gamma \in \mathbb{R}} \quad & \gamma \\
\text{s.t.} \quad & P(t) \preceq \gamma I \quad \forall t \in [-1,1] \\
& P(t) \succeq 0 \quad \forall t \in [-1,1] \\
& P(t) - A(t)^T P(t) A(t) \succeq 0 \quad \forall t \in [-1,1] \\
& P(t) \succeq h_i h_i^T \quad \forall t \in [-1,1] \quad i = 1, \dots, m,
\end{aligned} \tag{20}$$

where $A(t) := \frac{1+t}{2}\hat{A} + \frac{1-t}{2}\check{A}$, is feasible for a polynomial matrix $P(t)$ of sufficiently large degree. Note that (20) is a semidefinite program; see Section 3.2. The GE-d associated with the polynomial matrix $P(t)$ is an inner approximation to the set $\mathcal{S}$ defined in (18). The objective and the first constraint in this SDP are maximizing the radius of a ball contained in this GE-d.

6.3.1 Numerical example

As an example, we seek to characterize the set $\mathcal{S}$ as defined in (18) with the following input:

$$H = \begin{bmatrix} -1 & 0 \\ 1 & 0 \\ 0 & -1 \\ 0 & 1 \end{bmatrix}, \quad \hat{A} = \begin{bmatrix} -0.9 & 0.6 \\ -1.6 & 1.1 \end{bmatrix}, \quad \check{A} = \begin{bmatrix} 1.1 & 0.6 \\ -1.6 & -0.9 \end{bmatrix}.$$

We solve the SDP in (20) to find a polynomial matrix $P(t)$ of degree d. For $d = 0$, the problem is infeasible. For $d = 1, 2$, the problem is feasible and we denote the GEs defined by the optimal solutions by $\mathcal{E}_1, \mathcal{E}_2$, respectively. In Figure 6, we plot the sets $\Omega, \mathcal{E}_1, \mathcal{E}_2$, as well as the set

$$S_{10} := \left\{ x \in \mathbb{R}^n \mid H A^k(t) x \le 1 \quad \forall t \in \left\{-1, -\frac{9}{10}, \dots, \frac{9}{10}, 1\right\} \quad k = 0, 1, \dots 10 \right\},$$

which is clearly an outer approximation of the set $\mathcal{S}$. Since $\mathcal{S}$ must be sandwiched between the sets $\mathcal{E}_2$ and S_{10}, we can conclude that $\mathcal{E}_2$ provides a good inner approximation to $\mathcal{S}$.

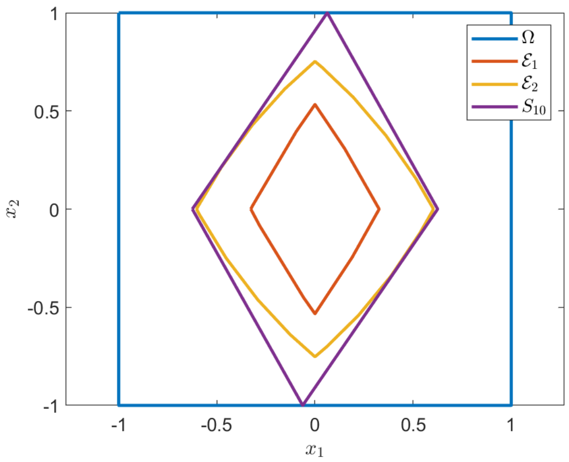


Figure 6: The sets associated with the numerical example in Section 6.3.1. The GE-2 denoted by $\mathcal{E}_2$ provides a good inner approximation of the set $\mathcal{S}$ defined in (18).

6.4 Polynomial regression robust to a shift

A classic task in statistics is to fit a polynomial function $p : \mathbb{R} \to \mathbb{R}$ to observations (x_i, y_i) for $i = 1, \ldots, m$. A standard approach to find a polynomial fit of degree (at most) d is that of least-squares polynomial regression, which solves the problem

$$\min_{c \in \mathbb{R}^{d+1}} \|\Phi(x)c - y\|^2 , \tag{21}$$

where $c \in \mathbb{R}^{d+1}$ is the vector of coefficients of p, the vectors $x, y \in \mathbb{R}^m$ have their i^{th} entry equal to x_i, y_i, respectively, and $\Phi : \mathbb{R}^m \to \mathbb{R}^{m \times (d+1)}$ is the polynomial matrix with the i^{th} row of $\Phi(z)$ equal to $(1, z_i, z_i^2, \ldots, z_i^d)$. In some applications, due to metrological limitations, one may have some error in measuring the points x_i. In particular, one may wish to find a function which provides a good fit to the observations even if the points x_i were slightly shifted. In this section, we describe how GEs arise when solving this problem.

More concretely, to find a polynomial that fits the observations well even if the points x_i are shifted by up to ε units to the left or right, one can write

$$\min_{c \in \mathbb{R}^{d+1}} \max_{t \in [-1,1]} \|\Phi(x + \varepsilon t)c - y\|^2 . \tag{22}$$

This problem can be reformulated as

$$\begin{aligned} \min_{c \in \mathbb{R}^{d+1}, \gamma \in \mathbb{R}} \quad & \gamma \\ \text{s.t.} \quad & \begin{bmatrix} c \\ 1 \end{bmatrix}^T P(t) \begin{bmatrix} c \\ 1 \end{bmatrix} \leq \gamma \quad \forall t \in [-1, 1], \end{aligned} \tag{23}$$

for

$$P(t) = \begin{bmatrix} \Phi(x + \varepsilon t)^T \Phi(x + \varepsilon t) & -\Phi(x + \varepsilon t)^T y \\ -y^T \Phi(x + \varepsilon t) & y^T y \end{bmatrix} .$$

Note that $P(t) \succeq 0$ for all $t \in [-1, 1]$ and it is straightforward to check that $P(t)$ satisfies the kernel condition under the mild assumption that there are a pair of observations (x_i, y_i) and (x_j, y_j) in the dataset such that $x_i \neq x_j$ and $y_i \neq y_j$. Therefore, problem (22) corresponds to finding a coefficient vector c such that the appended vector $[c, 1]^T$ is minimal with respect to the GE-2d-norm defined by $P(t)$ (see (3) in Section 2). By Theorem 4.4, this problem can be reformulated as an SDP.

6.4.1 Numerical example

For our experiment, we take $m = 10$, the points x_i to be uniformly spaced between -1 and 1, and $y_i = f(x_i)$ where f is the Runge function $f(x) := \frac{1}{1+25x^2}$. We fit a degree-9 polynomial to these observations both with the standard least-squares approach and with the shift-robust approach. We take the shift tolerance ε to be equal to 0.05. In Figure 7, we plot the observations (x_i, y_i) as well as $(x_i \pm 0.05, y_i)$. We also plot the polynomials corresponding to the solutions of (21) and (22), labelled as p_{LS} and p_{GE}, respectively. We can calculate the worst-case errors for these polynomials:

$$\max_{t\in[-1,1]} \|p_{LS}(x + 0.05t) - y\|^2 = 0.8086,$$
$$\max_{t\in[-1,1]} \|p_{GE}(x + 0.05t) - y\|^2 = 0.0447.$$

Here we see that p_{GE} is significantly more robust to a small shift of the points and is overall much smoother than p_{LS}.

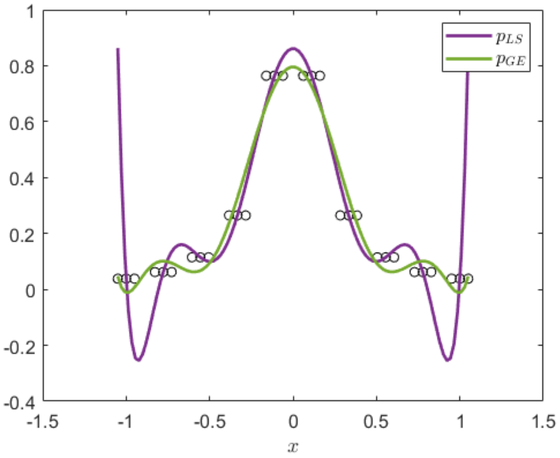


Figure 7: Observations and fitted polynomials associated with the numerical example in Section 6.4.1.

7 Future Research Directions

We conclude with a few questions for future research. Our first two questions concern extensions of some results from Section 5.

- We showed that every symmetric convex body can be approximated arbitrarily well by a GE. Our approximation factor relies on results on polytopic approximation of convex bodies. A related question is: how well can one approximate an n-dimensional symmetric convex body with a finite intersection of m (co-centered) ellipsoids, with m growing potentially with n? Since we have shown that a GE-d, with $d = 2m-3$, can exactly represent an intersection of m co-centered ellipsoids (see Corollary 5.6 and Theorem 5.1), progress on this question can potentially improve our approximation factor in Theorem 5.7.

- Let us call a set $\Omega \subseteq \mathbb{R}^n$ *GE-d-representable* if for some nonnegative integers k, m, a GE-d $\mathcal{E}_d \subset \mathbb{R}^m$, some matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{m\times k}$, and some vector $b \in \mathbb{R}^m$, one can write

$$\Omega = \{x \in \mathbb{R}^n \mid \exists u \in \mathbb{R}^k \text{ s.t. } Ax + Bu + b \in \mathcal{E}_d\}.$$

We say that a set is *GE-representable* if it is GE-d-representable for some nonnegative integer d. It would be interesting to study the expressiveness of GE-representable sets; in particular, do GE-representable sets fall in between SOCP and SDP-representable sets?

Finally, we highlight some results (among many) about ellipsoids which we believe might be interesting to extend to generalized ellipsoids.

- Motivated by problems in subspace identification and factor analysis, the ellipsoid fitting conjecture [42, 43] concerns the maximum number of independent standard Gaussian vectors in $\mathbb{R}^n$ such that with high probability, there exists an ellipsoid (i.e., a GE-0), passing through them. Recently, great progress has been made on this problem which resolves the conjecture up to a constant [46, 11, 29]. How does this maximum number change when one replaces a GE-0 with a GE-d for a fixed value of d?
- In [35], it is shown that the standard SDP relaxation for the (nonconvex) problem of maximizing an arbitrary homogeneous quadratic function over the intersection of m ellipsoids provides an approximation ratio of $\frac{1}{2\log(2m^2)}$. Can one derive a similar result for maximization of quadratic functions over a GE-d and obtain an approximation ratio in terms of d? Note that one cannot directly apply the result of [35] since for $d \geq 2$, a GE-d can be the intersection of an infinite number of ellipsoids.

Acknowledgements

We are grateful to Pablo Parrilo for insightful discussions around Lemma 5.4.

References

[1] M. Abramowitz, I. Stegun, "Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables", *Dover Publications, New York*, (1974).

[2] A.A. Ahmadi, A. Chaudhry, V. Sindhwani, S. Tu, "Safely learning dynamical systems", preprint, `https://arxiv.org/abs/2305.12284`, (2024).

[3] A.A. Ahmadi, E. De Klerk, G. Hall, "Polynomial norms", *SIAM Journal on Optimization*, 29(1) (2019), 399–422.

[4] A.A. Ahmadi, B. El Khadir, "Time-varying semidefinite programs", *Mathematics of Operations Research*, 46(3) (2021), 1054–1080.

[5] A.A. Ahmadi, O. Günlük, "Robust-to-dynamics optimization", To appear in *Mathematics of Operations Research*, `https://doi.org/10.1287/moor.2023.0116`, (2024).

[6] A.A. Ahmadi, G. Hall, A. Makadia, V. Sindhwani, "Geometry of 3D environments and sum of squares polynomials", In *Robotics: Science and Systems*, (2017).

[7] A.A. Ahmadi, R.M. Jungers, P.A. Parrilo, M. Roozbehani, "Joint spectral radius and path-complete graph Lyapunov functions", *SIAM Journal on Control and Optimization*, 52(1) (2014), 687–717.

[8] A.A. Ahmadi, A. Olshevsky, P.A. Parrilo, J.N. Tsitsiklis, "NP-hardness of deciding convexity of quartic polynomials and related problems", *Mathematical Programming*, 137(1-2) (2013), 453–476.

[9] T. Ando, M.-H. Shih, “Simultaneous contractibility”, *SIAM Journal on Matrix Analysis and Applications*, 19(2) (1998), 487–498.

[10] E. Aylward, S. Itani, P.A. Parrilo, “Explicit SOS decompositions of univariate polynomial matrices and the Kalman-Yakubovich-Popov lemma”, In *Proceedings of the 46th IEEE Conference on Decision and Control*, (2007), 5660–5665.

[11] A.S. Bandeira, A. Maillard, S. Mendelson, E. Paquette, “Fitting an ellipsoid to a quadratic number of random points”, preprint, `https://arxiv.org/abs/2307.01181`, (2023).

[12] A. Barvinok, “A Course in Convexity”, Graduate Studies in Mathematics, Volume 54, *American Mathematical Society*, 2002.

[13] A. Barvinok, “Thrifty approximations of convex bodies by polytopes”, In *International Mathematics Research Notices*, (2014), 4341–4356.

[14] A. Ben-Tal, A. Nemirovski, “Lecture Notes on Modern Convex Optimization”, *Society for Industrial and Applied Mathematics*, 2001.

[15] V.D. Blondel, V. Canterini, “Undecidable problems for probabilistic automata of fixed dimension”, *Theory of Computing Systems*, 36 (2003), 231–245.

[16] V.D. Blondel, Y. Nesterov, J.Theys, “On the accuracy of the ellipsoid norm approximation of the joint spectral radius”, *Linear Algebra and its Applications*, 394 (2005), 91–107.

[17] V.D. Blondel, J.N. Tsitsiklis, “The boundedness of all products of a pair of matrices is undecidable”, *Systems and Control Letters*, 41 (2000), 135–140.

[18] S. Boyd, L. Vandenberghe, “Convex Optimization”, *Cambridge University Press, New York*, (2004).

[19] M.D. Choi, T.Y. Lam, B. Reznick, “Real zeros of positive semidefinite forms. I”, *Mathematische Zeitschrift*, 171 (1) (1980), 1–26.

[20] M.D. Choi, T.Y. Lam, B. Reznick, “Sums of squares of real polynomials”, In *Proceedings of Symposia in Pure Mathematics*, 58 (1995), 103–126.

[21] G. Debreu, “A social equilibrium existence theorem”, In *Proceedings of the National Academy of Sciences of the United States of America*, 38(10) (1952), 886–893.

[22] H. Dette, W.J. Studden, “Matrix measures, moment spaces and Favard’s theorem for the interval $[0, 1]$ and $[0, \infty)$”, *Linear Algebra and its Applications*, 345(1-3) (2002), 169–193.

[23] H. Fawzi, “On representing the positive semidefinite cone using the second-order cone”, *Mathematical Programming*, 175 (2019), 109–118.

[24] E.G. Gilbert, D.W. Johnson, S.S. Keerthi, “A fast procedure for computing the distance between complex objects in three-dimensional space”, *IEEE Journal on Robotics and Automation*, 4(2) (1988), 193–203.

[25] M. Grötschel, L. Lovász, A. Schrijver, “Geometric Algorithms and Combinatorial Optimization”, *Springer-Verlag, Berlin*, (1988).

[26] M.T. Harris, P.A. Parrilo, “Improved nonnegativity testing in the Bernstein basis via geometric means”, preprint, `https://arxiv.org/abs/2309.10675`, (2023).

[27] J.W. Helton, J. Nie, “Semidefinite representation of convex sets”, *Mathematical Programming*, 122(1) (2010), 21–64.

[28] J.P. Hespanha, “Linear Systems Theory”, *Princeton University Press*, 2009.

[29] J.T. Hsieh, P.K. Kothari, A. Potechin, J. Xu, “Ellipsoid fitting up to a constant”, In *50th International Colloquium on Automata, Languages, and Programming (ICALP 2023), Leibniz International Proceedings in Informatics (LIPIcs)*, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 261 (2023), 78:1–78:20.

[30] R. Jungers, “The Joint Spectral Radius: Theory and Applications”, Lecture Notes in Control and Information Sciences 385, *Springer-Verlag, Berlin*, (2009).

[31] J.B. Lasserre, “Convexity in semialgebraic geometry and polynomial optimization”, *SIAM Journal on Optimization*, 19(4) (2009), 1995–2014.

[32] B. Legat, C. Yuan, P.A. Parrilo, “Low-rank univariate sum of squares has no spurious local minima”, *SIAM Journal on Optimization*, 33(3) (2023), 2041–2061.

[33] J. Löfberg, “YALMIP: A toolbox for modeling and optimization in MATLAB”, In *Proceedings of the IEEE International Conference on Robotics and Automation*, (2004), 284–289.

[34] J. Löfberg, P.A. Parrilo, “From coefficients to samples: A new approach to SOS optimization”, In *Proceedings of the 43rd IEEE Conference on Decision and Control*, (2004), 3154–3159.

[35] A. Nemirovski, C. Roos, T. Terlaky, “On maximization of quadratic form over intersection of ellipsoids with common center”, *Mathematical Programming*, 86 (1999), 463–473.

[36] D. Papp, “Semi-infinite programming using high-degree polynomial interpolants and semidefinite programming”, *SIAM Journal on Optimization*, 27(3) (2017), 1858–1879.

[37] D. Papp, F. Alizadeh, “Semidefinite characterization of sum-of-squares cones in algebras”, *SIAM Journal on Optimization*, 23(3) (2013), 1398–1423.

[38] D. Papp, S. Yildiz, “Sum-of-squares optimization without semidefinite programming”, *SIAM Journal on Optimization*, 29(1) (2019), 822–851.

[39] P.A. Parrilo, “Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization”, Ph.D. Thesis, California Institute of Technology, (2000).

[40] P. Raghavendra, N. Ryder, N. Srivastava, “Real stability testing”, In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017), Leibniz International Proceedings in Informatics (LIPIcs)*, Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 67 (2017), 5:1–5:15.

[41] G.C. Rota, W.G. Strang, “A note on the joint spectral radius”, *Indag. Math.*, 22 (1960), 379–381.

[42] J. Saunderson, V. Chandrasekaran, P.A. Parrilo, A.S. Willsky, “Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting”, *SIAM Journal on Matrix Analysis and Applications*, 33(4) (2012), 1395–1416.

[43] J. Saunderson, P.A. Parrilo, A.S. Willsky, “Diagonal and low-rank decompositions and fitting ellipsoids to random point”, In *Proceedings of the 52nd IEEE Conference on Decision and Control*, (2013), 6031–6036.

[44] C. Scherer, C. Hol, “Matrix sum-of-squares relaxations for robust semi-definite programs”, *Mathematical Programming*, 107 (2006), 189–211.

[45] M.J. Todd, “Minimum-Volume Ellipsoids: Theory and Algorithms”, *Society for Industrial and Applied Mathematics, Philadelphia, PA, USA*, (2016).

[46] M. Tulsiani, J. Wu, “Ellipsoid fitting up to constant via empirical covariance estimation”, preprint, `https://arxiv.org/abs/2307.10941`, (2023).

[47] L. Vandenberghe, S. Boyd, “Semidefinite programming”, *SIAM Review*, 38(1) (1996), 49–95.

[48] T. Weisser, B. Legat, C. Coey, L. Kapelevich, V.J. Pablo, “Polynomial and moment optimization in Julia and JuMP”, `https://pretalx.com/juliacon2019/talk/QZBKAU/`, (2019).

[49] V.A. Yakubovich, “Factorization of symmetric matrix polynomials”, *Dokl. Akad. Nauk SSSR*, 194(3) (1970), 532–535.

[50] D. Youla, “On the factorization of rational matrices”, *IRE Transactions on Information Theory*, 7(3) (1961), 172–189.