MIT 6.7230 - Algebraic techniques and semidefinite optimization February 7, 2024

Lecture 1

Lecturer: Pablo A. Parrilo *Scribe: Pablo A. Parrilo*

1 Introduction: what is this course about?

In this course we aim to understand the properties, both mathematical and computational, of sets defined by polynomial equations and inequalities. In particular, we want to work towards methods that will enable the solution (either exact or approximate) of optimization problems with feasible sets that are defined in terms of polynomial systems. Needless to say (is it?), many problems in estimation, control, robotics, statistics, machine learning, signal processing, etc., admit simple formulations in terms of polynomial equations and inequalities. However, these formulations can be tremendously difficult to solve, and thus our methods should try to exploit as many structural properties as possible.

The computational aspects of these sets are not fully understood at the moment. In the well-known case of polyhedra, for instance, there is a well defined relationship between the geometrical properties of the set (e.g., the number of facets, or the number of extreme points) and its algebraic representation. Furthermore, polyhedral sets are preserved by natural operations (e.g., projections). None of this will generally be true for (basic) semialgebraic sets, and this causes a very interesting interaction between their geometry and their algebraic descriptions.

2 Topics

To understand better what is going on, we will embark in a journey to learn a wide variety of methods used to approach these problems. Some of our stops along the way will include:

- Linear optimization, second order cones, semidefinite programming
- Algebra: groups, fields, rings
- Univariate polynomials
- Resultants and discriminants
- Hyperbolic polynomials
- Sum of squares
- Ideals, varieties, Groebner bases, Hilbert's Nullstellensatz
- Quantifier elimination
- Real Nullstellensatz
- And much more...

We are interested in computational methods, and want to emphasize efficiency. Throughout, applications will play an important role, both as motivation and illustration of the techniques.

3 Review: convexity

A very important notion in modern optimization is that of *convexity*. To a large extent, modulo some (important) technicalities, there is a huge gap between the theoretical and practical solvability of optimization problems for which the feasible set is convex, versus those where this property fails. Recommended presentations of convex optimization from a modern viewpoint are [BV04, BTN01, BNO03], with [Roc70] being the classical treatment of convex analysis.

Unless specified otherwise, we will work on a finite-dimensional real vector space, which we will identify with $\mathbb{R}^n$. The same will extend to the corresponding dual spaces. Often, we will implicitly use the standard Euclidean inner product, thus identifying $\mathbb{R}^n$ and its dual.

Here are some relevant definitions:

Definition 1 *A set S is* convex *if $x_1, x_2 \in S$ implies $\lambda x_1 + (1-\lambda)x_2 \in S$ for all $0 \leq \lambda \leq 1$.*

The intersection of convex sets is always convex. Given a convex set S, a point $x \in S$ is *extreme* if for any two points x_1, x_2 in S, having $x = \lambda x_1 + (1-\lambda)x_2$ and $\lambda \in (0,1)$ implies that $x_1 = x_2 = x$.

Example 2 *The following are examples of convex sets:*

- *The n-dimensional* hypercube *is defined by $2n$ linear inequalities:*

$$\{x \in \mathbb{R}^n \,:\, -1 \leq x_i \leq 1, \qquad i = 1, \ldots, n\}.$$

  This convex set has 2^n extreme points, namely all those of the form $(\pm 1, \pm 1, \ldots, \pm 1)$.

- *The n-dimensional* Euclidean unit ball *is defined by the inequality $x_1^2 + \cdots + x_n^2 \leq 1$. This set has an infinite number of extreme points, namely all those on the hypersurface $x_1^2 + \cdots + x_n^2 = 1$.*

- *The n-dimensional* crosspolytope *has $2n$ extreme points, namely all those whose coordinates are permutations of $(\pm 1, 0, \ldots, 0)$. It can be defined using 2^n linear inequalities, of the form*

$$\pm x_1 \pm x_2 \pm \cdots \pm x_n \leq 1.$$

All these examples actually correspond to unit balls of different norms (ℓ_∞, ℓ_2, and ℓ_1, respectively). It is easy to show that the unit ball of every norm is always a convex set. Conversely, given any full-dimensional convex set symmetric with respect to the origin, one can define a norm via the corresponding gauge *(or* Minkowski*) functional.*

One of the most important results about convex sets is the *separating hyperplane* theorem.

Theorem 3 *Given two disjoint convex sets S_1, S_2 in $\mathbb{R}^n$, there exists a nontrivial linear functional c and a scalar d such that*

$$\begin{aligned} \langle c, x \rangle &\geq d \qquad \forall x \in S_1 \\ \langle c, x \rangle &\leq d \qquad \forall x \in S_2. \end{aligned}$$

Under certain additional conditions, *strict* separation can be guaranteed. One of the most useful cases is when one of the sets is compact and the other one is closed.

Convex cones. An important class of convex sets are those that are invariant under nonnegative scalings.

Definition 4 *A set $S \subseteq \mathbb{R}^n$ is a* cone *if $\lambda \geq 0$, $x \in S \Rightarrow \lambda x \in S$.*

Definition 5 *The* dual *of a set S is $S^* := \{y \in \mathbb{R}^n : \langle y, x\rangle \geq 0 \quad \forall x \in S\}$.*

Given any set S, its dual S^* is always a closed convex cone. Duality reverses inclusion, that is, $S_1 \subseteq S_2$ implies $S_1^* \supseteq S_2^*$. If S is a closed convex cone, then $S^{**} = S$. Otherwise, S^{**} is the closure of the smallest convex cone that contains S.

A cone $\mathcal{K}$ is *pointed* if $\mathcal{K} \cap (-\mathcal{K}) = \{0\}$, and *solid* if the interior of $\mathcal{K}$ is not empty. A cone that is convex, closed, pointed and solid is called a *proper* cone. The dual set of a proper cone is also a proper cone, called the *dual cone*. An element x is in the interior of the cone K if and only if $\langle x, y\rangle > 0$, $\forall y \in K^*, y \neq 0$.

Example 6 *The* nonnegative orthant *is defined as $\mathbb{R}^n_+ := \{x \in \mathbb{R}^n : x_i \geq 0\}$, and is a proper cone. The nonnegative orthant is self-dual, i.e., we have $(\mathbb{R}^n_+)^* = \mathbb{R}^n_+$.*

A proper cone $\mathcal{K}$ induces a *partial order*[1] on the vector space, via $x \preceq y$ if and only if $y - x \in \mathcal{K}$. We also use $x \prec y$ if $y - x$ is in the interior of $\mathcal{K}$. Important examples of proper cones are the nonnegative orthant, the Lorentz cone, the set of symmetric positive semidefinite matrices, and the set of nonnegative polynomials. We will discuss some of these in more detail later in the lectures and the exercises.

Example 7 *Consider the* second-order cone*, defined by $\{(x_0, x_1, \ldots, x_n) \in \mathbb{R}^{n+1} : \left(\sum_{i=1}^n x_i^2\right)^{\frac{1}{2}} \leq x_0\}$. This is a self-dual proper cone, and is also known as the* ice-cream*, or* Lorentz *cone.*

An interesting physical interpretation of the partial order induced by this cone appears in the theory of special relativity. In this case, the cone can be expressed (after an inconsequential rescaling and reordering) as

$$\{(x, y, z, t) \in \mathbb{R}^4 : x^2 + y^2 + z^2 \leq c^2 t^2, \quad t \geq 0\},$$

where c is a given constant (speed of light). In this case, the vector space is interpreted as the Minkowski spacetime. Given a fixed point x_0, those points x for which $x \succeq x_0$ correspond to the absolute future*, while those for which $x \preceq x_0$ are in the* absolute past*. There are, however, many points that are neither in the absolute future nor in the absolute past (for these, the causal order will depend on the observer).*

Remark 8 *Convexity has two natural definitions. The first one is the one given above, that emphasizes the "internal" aspect, in terms of convex combinations of elements of the set. Alternatively, one can look at the "external" aspect, and define a convex set as the intersection of a (possibly infinite) collection of half-spaces. The possibility of these "dual" descriptions is what enables many of the useful and intriguing properties of convex sets. In the context of convex functions, for instance, these ideas are made concrete through the use of the Legendre-Fenchel transformation.*

[1] A *partial order* is a binary relation $\preceq$ that is reflexive, antisymmetric, and transitive.

4 Review: linear programming

Linear programming (LP) is the problem of minimizing a linear function, subject to finitely many linear inequality constraints. An LP in standard form is written as

$$\min\ c^T x \qquad \text{s.t.} \quad \begin{cases} Ax &= b \\ x &\geq 0. \end{cases} \tag{P}$$

Every LP problem has a corresponding *dual* problem, which in this case is

$$\max\ b^T y \qquad \text{s.t.} \quad c - A^T y \geq 0. \tag{D}$$

Linear programming has many important properties. Among them, we mention the following ones:

Geometry of the feasible set: The feasible sets of linear programs are *polyhedra*. The geometry of polyhedra is quite well understood. In particular, the Minkowski-Weyl theorem (e.g. [BT97], or [Zie95, Section 1.1]) states that every polyhedron P is finitely generated, i.e., it can be written as

$$P = \text{conv}(u_1, \ldots, u_r) + \text{cone}(v_1, \ldots, v_s),$$

where u_i, v_i are the *vertices* and *extreme rays* of P, and the *convex hull* and *conical hull* are given by

$$\text{conv}(u_1, \ldots, u_r) = \left\{ \sum_{i=1}^{r} \lambda_i u_i \;\middle|\; \sum_{i=1}^{r} \lambda_i = 1, \quad \lambda_i \geq 0, \quad i = 1, \ldots, r \right\},$$

and

$$\text{cone}(v_1, \ldots, v_s) = \left\{ \sum_{i=1}^{s} \lambda_i v_i \;\middle|\; \lambda_i \geq 0, \quad i = 1, \ldots, s \right\}.$$

Rational solutions: Unless the problem is unbounded, the optimal solution of a linear programming problem is always achieved at extreme points of the feasible set. Since these correspond to vertices of a polyhedron, the solution can be characterized in terms of a system of linear equations, corresponding to the equations and inequalities that are active at the optimal point. Thus, if the problem description (i.e., the matrices A, b, c) is given by rational numbers, there are always extreme points that are rational and achieve the optimal cost.

Weak duality: For *any* feasible solutions x, y of (P) and (D), respectively, it always holds that:

$$c^T x - b^T y = x^T c - (Ax)^T y = x^T (c - A^T y) \geq 0, \tag{1}$$

where the last inequality follows from the feasibility conditions $x \geq 0$ and $A^T y \leq c$. Thus, from any feasible dual solution one can obtain a lower bound on the value of the primal. Conversely, primal feasible solutions give upper bounds on the value of the dual.

Strong duality: If both primal and dual problems are feasible, then they achieve exactly the same optimal value, and there exist optimal feasible solutions $x^\star, y^\star$ such that $c^T x^\star = b^T y^\star$. This is a consequence of the separation theorems for convex sets.

Complementary slackness: Strong duality, combined with (1), implies that at optimality we must have

$$x_i^\star(c - A^T y^\star)_i = 0 \qquad i = 1, \ldots, n.$$

In other words, there is a correspondence between primal variables and dual inequalities, that says that whenever a primal variable is nonzero, the corresponding dual inequality must be tight. Conversely, if at optimality a dual inequality is not tight, the corresponding primal variable must vanish.

Some of these properties (which ones?) will break down as soon as we leave LP and go the more general case of conic or semidefinite programming. These will cause some difficulties, although with the right assumptions, the resulting theory will closely parallel the LP case.

Remark 9 *Among others, the software codes* `cdd` *(Komei Fukuda,* `https://www.inf.ethz.ch/personal/fukudak/cdd_home/`*),* `lrs` *(David Avis,* `http://cgm.cs.mcgill.ca/~avis/C/lrs.html`*), or* `Normaliz` *(*`https://www.normaliz.uni-osnabrueck.de`*) are very useful for polyhedral computations. In particular, all of them allow to convert an inequality representation of a polyhedron (usually called an H-representation) into extreme points/rays (V-representation), and viceversa. There are also MATLAB or Julia toolboxes for polyhedral computations, like* `MPT3` *(*`https://www.mpt3.org`*) or* `Polyhedra.jl` *(*`https://github.com/JuliaPolyhedra/Polyhedra.jl`*)*

References

[BNO03] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar. *Convex analysis and optimization.* Athena Scientific, Belmont, MA, 2003.

[BT97] D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization.* Athena Scientific, 1997.

[BTN01] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization.* MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[Roc70] R. T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, New Jersey, 1970.

[Zie95] G. M. Ziegler. *Lectures on polytopes*, volume 152 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1995.

MIT 6.7230 - Algebraic techniques and semidefinite optimization February 9, 2024

Lecture 2

Lecturer: Pablo A. Parrilo *Scribe: Pablo A. Parrilo*

Notation: The set of real symmetric $n \times n$ matrices is denoted $\mathcal{S}^n$. A matrix $A \in \mathcal{S}^n$ is called *positive semidefinite* if $x^T Ax \geq 0$ for all $x \in \mathbb{R}^n$, and is called *positive definite* if $x^T Ax > 0$ for all nonzero $x \in \mathbb{R}^n$. The set of positive semidefinite matrices is denoted $\mathcal{S}^n_+$ and the set of positive definite matrices is denoted by $\mathcal{S}^n_{++}$. As we shall prove soon, $\mathcal{S}^n_+$ is a proper cone (i.e., closed, convex, pointed, and solid). We will use the inequality signs "$\preceq$" and "$\succeq$" to denote the partial order induced by $\mathcal{S}^n_+$ (usually called the *Löwner* partial order).

1 Positive semidefinite (PSD) matrices

There are several equivalent conditions for a matrix to be positive (semi)definite. We present below some of the most useful ones:

Proposition 1. *The following statements are equivalent:*

- *The matrix $A \in \mathcal{S}^n$ is positive semidefinite ($A \succeq 0$).*
- *For all $x \in \mathbb{R}^n$, $x^T Ax \geq 0$.*
- *All eigenvalues of A are nonnegative.*
- *All $2^n - 1$ principal minors of A are nonnegative.*
- *There exists a factorization $A = B^T B$.*

For the definite case, we have a similar characterization:

Proposition 2. *The following statements are equivalent:*

- *The matrix $A \in \mathcal{S}^n$ is positive definite ($A \succ 0$).*
- *For all nonzero $x \in \mathbb{R}^n$, $x^T Ax > 0$.*
- *All eigenvalues of A are strictly positive.*
- *All n leading principal minors of A are positive.*
- *There exists a factorization $A = B^T B$, with B square and nonsingular.*

Here are some useful additional facts:

- If T is nonsingular, $A \succ 0 \Leftrightarrow T^T AT \succ 0$.
- Schur complement. The following conditions are equivalent:

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \succ 0 \quad \Leftrightarrow \quad \left\{ \begin{array}{r} A \succ 0 \\ C - B^T A^{-1} B \succ 0 \end{array} \right. \quad \Leftrightarrow \quad \left\{ \begin{array}{r} C \succ 0 \\ A - BC^{-1} B^T \succ 0 \end{array} \right.$$

We now prove the following result:

Theorem 3. *The set $\mathcal{S}_+^n$ of positive semidefinite matrices is a proper cone.*

Proof. Invariance under nonnegative scalings follows directly from the definition, so S_+^n is a cone. By the second statement in Proposition 1, S_+^n is the intersection of infinitely many closed halfspaces, and hence it is both closed and convex. To show pointedness, notice that if there is a symmetric matrix A that belongs to both S_+^n and $-S_+^n$, then $x^T A x$ must vanish for all $x \in \mathbb{R}^n$, thus A must be the zero matrix. Finally, the cone is solid since $I_n + X$ is positive definite for all X provided $\|X\|$ is small enough (e.g., by continuity of eigenvalues). □

We state next some additional facts on the geometry of the cone $\mathcal{S}_+^n$ of positive semidefinite matrices.

- If S_n is equipped with the inner product $\langle X, Y \rangle := X \bullet Y = \mathrm{Tr}(XY)$, then $\mathcal{S}_+^n$ is a self-dual cone.
- The cone $\mathcal{S}_+^n$ is *not* polyhedral, and its extreme rays are the rank one matrices.

2 Semidefinite programming

Semidefinite programming (SDP) is a specific kind of convex optimization problem (e.g., [VB96, Tod01, BV04]), with very appealing numerical properties. An SDP problem corresponds to the optimization of a linear function subject to matrix inequality constraints.

An SDP problem in standard primal form is written as:

$$\begin{array}{lrcl} \text{minimize} & C \bullet X & & \\ \text{subject to} & A_i \bullet X & = & b_i, \quad i = 1, \ldots, m \\ & X & \succeq & 0, \end{array} \tag{1}$$

where $C, A_i \in \mathcal{S}^n$, and $X \bullet Y := \mathrm{Tr}(XY)$. The matrix $X \in \mathcal{S}^n$ is the variable over which the maximization is performed. The inequality in the second line means that the matrix X must be positive semidefinite, i.e., all its eigenvalues should be greater than or equal to zero. The set of feasible solutions, i.e., the set of matrices X that satisfy the constraints, is always a convex set. In the particular case in which $C = 0$, the problem reduces to whether or not the inequality can be satisfied for some matrix X. In this case, the SDP is referred to as a *feasibility problem*. The convexity of SDP has made it possible to develop sophisticated and reliable analytical and numerical methods to solve them.

A very important feature of SDP problems, from both the theoretical and applied viewpoints, is the associated *duality theory*. For every SDP of the form (1) (usually called the *primal problem*), there is another associated SDP, called the *dual problem*, that can be stated as

$$\begin{array}{ll} \text{maximize} & b^T \mathbf{y} \\ \text{subject to} & \sum_{i=1}^m A_i y_i \preceq C, \end{array} \tag{2}$$

where $b = (b_1, \ldots, b_m)$, and the vector $\mathbf{y} = (y_1, \ldots, y_m)$ contains the dual decision variables.

The key relationship between the primal and the dual problem is the fact that feasible solutions of one can be used to bound the values of the other problem. Indeed, let X and $\mathbf{y}$ be any two feasible solutions of the primal and dual problems respectively. We then have the following inequality:

$$C \bullet X - b^T \mathbf{y} = (C - \sum_{i=1}^m A_i y_i) \bullet X \geq 0, \tag{3}$$

where the last inequality follows from the fact that the two terms are positive semidefinite matrices. From (1) and (2) we can see that the left hand side of (3) is just the difference between the objective functions of the primal and dual problems. The inequality in (3) tells us that the value of the primal objective function evaluated at any feasible matrix X is always greater than or equal to the value of the dual objective function at any feasible vector $\mathbf{y}$. This property is known as *weak duality*. Thus, we can use any feasible X to compute an upper bound for the optimum of $b^T\mathbf{y}$, and we can also use any feasible $\mathbf{y}$ to compute a lower bound for the optimum of $C \bullet X$. Furthermore, in the case of feasibility problems (i.e., $C = 0$), the dual problem can be used to certify the nonexistence of solutions to the primal problem. This property will be crucial in our later developments.

2.1 Conic duality

A general formulation, discussed briefly during the previous lecture, that unifies LP and SDP (as well as some other classes of optimization problems) is *conic programming*. We will be more careful than usual here (risking being a bit pedantic) in the definition of the respective spaces and mappings. It does not make much of a difference if we are working on $\mathbb{R}^n$ (since we can identify a space and its dual), but it is "good hygiene" to keep these distinctions in mind, and also useful when dealing with more complicated spaces.

We will start with two real vector spaces, S and T, and a linear mapping $\mathcal{A} : S \to T$. Every real vector space has an associated dual space, which is the vector space of real-valued linear functionals. We will denote these dual spaces by S^* and T^*, respectively, and the pairing between an element of a vector space and one of the dual as $\langle \cdot, \cdot \rangle$ (i.e., $f(x) = \langle f, x \rangle$). The *adjoint mapping* of $\mathcal{A}$ is the unique linear map $\mathcal{A}^* : T^* \to S^*$ defined through the property

$$\langle \mathcal{A}^* y, x \rangle_S = \langle y, \mathcal{A}x \rangle_T \qquad \forall x \in S, y \in T^*.$$

Notice here that the brackets on the left-hand side of the equation represent the pairing in S, and those on the right-hand side correspond to the pairing in T. We can then define the primal-dual pair of (conic) optimization problems:

$$\begin{array}{lll} \text{minimize} & \langle c, x \rangle_S & \\ \text{subject to} & \left\{ \begin{array}{rcl} \mathcal{A}x & = & b \\ x & \in & \mathcal{K} \end{array} \right. \end{array} \qquad\qquad \begin{array}{ll} \text{maximize} & \langle y, b \rangle_T \\ \text{subject to} & c - \mathcal{A}^* y \in \mathcal{K}^*, \end{array}$$

where $b \in T$, $c \in S^*$, $\mathcal{K} \subset S$ is a proper cone, and $\mathcal{K}^* \subset S^*$ is the corresponding dual cone. Notice that exactly the same proof presented earlier works here to show weak duality:

$$\begin{aligned} \langle c, x \rangle_S - \langle y, b \rangle_T &= \langle c, x \rangle_S - \langle y, \mathcal{A}x \rangle_T \\ &= \langle c, x \rangle_S - \langle \mathcal{A}^* y, x \rangle_S \\ &= \langle c - \mathcal{A}^* y, x \rangle_S \\ &\geq 0. \end{aligned}$$

In the usual cases (e.g., LP and SDP), the vector spaces are finite dimensional, and thus isomorphic to their duals. The specific correspondence between these is given through whatever inner product we use.

Among the classes of problems that can be interpreted as particular cases of the general conic formulation we have *linear programs*, *second-order cone programs* (SOCP), and SDP, when we take the cone $\mathcal{K}$ to be the nonnegative orthant $\mathbb{R}^n_+$, the second order cone in n variables, or the PSD cone $\mathcal{S}^n_+$. We have then the following natural inclusion relationship among the different optimization classes.

$$\text{LP} \quad \subseteq \quad \text{SOCP} \quad \subseteq \quad \text{SDP}.$$

2.2 Geometric interpretation: separating hyperplanes

We give here a simple interpretation of duality, in terms of the separating hyperplane theorem. For simplicity, we concentrate on the case of feasibility only, i.e., where we are interested in deciding the existence of a solution x to the equations

$$\mathcal{A}x = b, \qquad x \in \mathcal{K}, \tag{4}$$

where as before $\mathcal{K}$ is a proper cone in the vector space S.

Consider now the image $\mathcal{A}(\mathcal{K})$ of the cone under the linear mapping. Notice that feasibility of (4) is equivalent to the point b being contained on $\mathcal{A}(\mathcal{K})$. We have now two convex sets in T, namely $\mathcal{A}(\mathcal{K})$ and $\{b\}$, and we are interested in knowing whether they intersect or not. *If* these sets satisfy certain properties (for instance, closedness and compactness) then we could apply the separating hyperplane theorem, to produce a linear functional y that will be positive on one set, and negative on the other. In particular, nonnegativity on $\mathcal{A}(\mathcal{K})$ implies

$$\langle y, \mathcal{A}(x)\rangle \geq 0 \ \ \forall x \in \mathcal{K} \qquad \Leftrightarrow \qquad \langle \mathcal{A}^*(y), x\rangle \geq 0 \ \ \forall x \in \mathcal{K} \qquad \Leftrightarrow \qquad \mathcal{A}^*(y) \in \mathcal{K}^*.$$

Thus, under these conditions, if (4) is infeasible, there is a linear functional y satisfying

$$\langle y, b\rangle < 0, \qquad \mathcal{A}^* y \in \mathcal{K}^*.$$

This yields a *certificate* of the infeasibility of the conic system (4).

2.3 Strong duality in SDP

Despite the formal similarities, there are a number of differences between linear programming and general conic programming (and in particular, SDP). Among them, we notice that in SDP optimal solutions may not necessarily exist (even if the optimal value is finite), and there can be a nonzero duality gap.

Nevertheless, we have seen that weak duality always holds for conic programming problems. As opposed to the LP case, *strong* duality can fail in general SDP. A nice example is given in [VB96, p. 65], where both the primal and dual problems are feasible, but their optimal values are different (i.e., there is a nonzero finite duality gap).

Nevertheless, under relatively mild constraint qualifications (Slater's condition, equivalent to the existence of strictly feasible primal and dual solutions) that are usually satisfied in practice, SDP problems have strong duality, and thus zero duality gap.

Theorem 4. *Assume that both the primal and dual problems are* strictly *feasible. Then, both achieve their optimal solutions, and there is no duality gap.*

There are several geometric interpretations of what causes the failure of strong duality for general SDP problems. A good one is based on the fact that the image of a proper cone under a linear transformation is not necessarily a proper cone. This fact seems quite surprising (or even wrong!) the first time one encounters it, but after a little while it starts being quite reasonable. Can you think of an example where this happens? What property will fail?

It should be mentioned that it is possible to formulate a more complicated SDP dual program (called the "Extended Lagrange-Slater Dual" in [Ram97]) for which strong duality always holds. For details, as well as a comparison with the more general "minimal cone" approach, we refer the reader to [Ram97, RTW97].

3 Applications

There have been *many* applications of SDP in a variety of areas of applied mathematics and engineering. We present here just a few, to give a flavor of what is possible. Many more will follow.

3.1 Lyapunov stability and control

Consider a linear difference equation (i.e., a discrete-time linear system) given by

$$x(k+1) = Ax(k), \qquad x(0) = x_0.$$

It is well-known (and easy to prove) that $x(k)$ converges to zero for all initial conditions x_0 iff $|\lambda_i(A)| < 1$, for $i = 1, \ldots, n$.

There is a simple characterization of this spectral radius condition in terms of a quadratic *Lyapunov function* $V(x(k)) = x(k)^T P x(k)$.

Theorem 5. *Given an $n \times n$ real matrix A, the following conditions are equivalent:*

(i) All eigenvalues of A are inside the unit circle, i.e., $|\lambda_i(A)| < 1$ for $i = 1, \ldots, n$.

(ii) There exist a matrix $P \in \mathcal{S}^n$ such that

$$P \succ 0, \qquad A^T P A - P \prec 0.$$

Proof. $(ii) \Rightarrow (i)$: Let $Av = \lambda v$. Then,

$$0 > v^*(A^T P A - P)v = (|\lambda|^2 - 1)\underbrace{v^* P v}_{>0},$$

and therefore $|\lambda| < 1$.
$(i) \Rightarrow (ii)$: Let $P := \sum_{k=0}^{\infty} (A^k)^T Q A^k$, where $Q \succ 0$. The sum converges by the eigenvalue assumption. Then,

$$A^T P A - P = \sum_{k=1}^{\infty} (A^k)^T Q A^k - \sum_{k=0}^{\infty} (A^k)^T Q A^k = -Q \prec 0$$

□

Consider now the case where A is not stable, but we can use linear state feedback, i.e., $A(K) = A + BK$, where K is a fixed matrix. We want to find a matrix K such that $A + BK$ is stable, i.e., all its eigenvalues have absolute value smaller than one.

Use Schur complements to rewrite the condition:

$$\begin{array}{c} (A+BK)^T P (A+BK) - P \prec 0, \qquad P \succ 0 \\ \Updownarrow \\ \left[\begin{array}{cc} P & (A+BK)^T P \\ P(A+BK) & P \end{array} \right] \succ 0 \end{array}$$

This condition is not simultaneously convex in (P, K) (since it is bilinear). However, we can do a congruence transformation with $Q := P^{-1}$, and obtain:

$$\left[\begin{array}{cc} Q & Q(A+BK)^T \\ (A+BK)Q & Q \end{array} \right] \succ 0$$

Now, defining a new variable $Y := KQ$ we have

$$\begin{bmatrix} Q & QA^T + Y^T B^T \\ AQ + BY & Q \end{bmatrix} \succ 0.$$

This problem is now linear in (Q, Y). In fact, it is an SDP problem. After solving it, we can recover the controller K via $K = YQ^{-1}$.

3.2 Theta function

Given a graph $G = (V, E)$, a *stable set* (or *independent set*) is a subset of V with the property that the induced subgraph has no edges. In other words, none of the selected vertices are adjacent to each other.

The *stability number* of a graph, usually denoted by $\alpha(G)$, is the cardinality of the largest stable set. Computing the stability number of a graph is NP-hard. There are many interesting applications of the stable set problem. In particular, they can be used to provide upper bounds on the *Shannon capacity of a graph* [Lov79], a problem of that appears in coding theory (when computing the zero-error capacity of a noisy channel [Sha56]). In fact, this was one of the first appearances of what today is known as SDP.

The Lovász theta function is denoted by $\vartheta(G)$, and is defined as the solution of the SDP :

$$\max\ J \bullet X \qquad \text{s.t.} \begin{cases} \operatorname{Tr}(X) = 1 \\ X_{ij} = 0 \qquad (i,j) \in E \\ X \succeq 0 \end{cases} \tag{5}$$

where J is the matrix with all entries equal to one. The theta function is an upper bound on the stability number, i.e.,

$$\alpha(G) \leq \vartheta(G).$$

The inequality is easy to prove. Consider the indicator vector $\xi(S)$ of any stable set S, and define the matrix $X := \frac{1}{|S|}\xi\xi^T$. Is is easy to see that this X is a feasible solution of the SDP, and it achieves an objective value equal to $|S|$. As a consequence, the inequality above directly follows.

3.3 Euclidean distance matrices

Assume we are given a list of pairwise distances between a finite number of points. Under what conditions can the points be embedded in some finite-dimensional space, and those distances be realized as the *Euclidean* metric between the embedded points? This problem appears in a large number of applications, including distance geometry, computational chemistry, and machine learning.

Concretely, assume we have a list of distances d_{ij}, for $(i, j) \in [1, n] \times [1, n]$. We would like to find points $x_i \in \mathbb{R}^k$ (for some value of k), such that $\|x_i - x_j\| = d_{ij}$ for all i, j. What are necessary and sufficient conditions for such an embedding to exist? In 1935, Schoenberg [Sch35] gave an exact characterization in terms of the semidefiniteness of the matrix of squared distances:

Theorem 6. *The distances d_{ij} can be embedded in a Euclidean space if and only if the $n \times n$ matrix*

$$D := \begin{bmatrix} 0 & d_{12}^2 & d_{13}^2 & \dots & d_{1n}^2 \\ d_{12}^2 & 0 & d_{23}^2 & \dots & d_{2n}^2 \\ d_{13}^2 & d_{23}^2 & 0 & \dots & d_{3n}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{1n}^2 & d_{2n}^2 & d_{3n}^2 & \dots & 0 \end{bmatrix}$$

is negative semidefinite on the subspace orthogonal to the vector $e := (1, 1, \ldots, 1)$.

Proof. We show only the necessity of the condition. Assume an embedding exists, i.e., there are points $x_i \in \mathbb{R}^k$ such that $d_{ij} = \|x_i - x_j\|$. Consider now the Gram matrix G of inner products

$$G := \begin{bmatrix} \langle x_1, x_1\rangle & \langle x_1, x_2\rangle & \ldots & \langle x_1, x_n\rangle \\ \langle x_2, x_1\rangle & \langle x_2, x_2\rangle & \ldots & \langle x_2, x_n\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle x_n, x_1\rangle & \langle x_n, x_2\rangle & \ldots & \langle x_n, x_n\rangle \end{bmatrix} = [x_1, \ldots, x_n]^T [x_1, \ldots, x_n],$$

which is positive semidefinite by construction. Since $D_{ij} = \|x_i - x_j\|^2 = \langle x_i, x_i\rangle + \langle x_j, x_j\rangle - 2\langle x_i, x_j\rangle$, we have

$$D = \operatorname{diag}(G) \cdot e^T + e \cdot \operatorname{diag}(G)^T - 2G,$$

from where the result directly follows. □

Notice that the dimension of the embedding is given by the rank k of the Gram matrix G.

For more on this and related embeddings problems, good starting points are Schoenberg's original paper [Sch35], as well as the book [DL97].

4 Software

Remark 7. *There are many good software codes for semidefinite programming. Among the most well-known, we mention the following ones:*

- *SeDuMi, originally by Jos Sturm, now being maintained by the optimization group at Lehigh:* `http://sedumi.ie.lehigh.edu`
- *SDPT3, by Kim-Chuan Toh, Reha Tütüncü, and Mike Todd.* `http://www.math.nus.edu.sg/~mattohkc/sdpt3.html`
- *SDPA, by the research group of Masakazu Kojima,* `http://sdpa.indsys.chuo-u.ac.jp/sdpa/`
- *CSDP, originally by Brian Borchers, now a COIN-OR project:* `https://projects.coin-or.org/Csdp/`
- *MOSEK, a commercial high performance software for large-scale conic programming (including SDP):* `http://www.mosek.com`.

A very convenient way of using these (and other) SDP solvers under MATLAB is through the YALMIP parser/solver (Johan Löfberg, `http://users.isy.liu.se/johanl/yalmip/`*), or the disciplined convex programming software CVX (Michael Grant and Stephen Boyd,* `http://www.stanford.edu/~boyd/cvx`*. If you use Julia instead, you can use the modeling languages/environments JuMP or Convex.jl; see* `https://www.juliaopt.org`*.*

References

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[DL97] M. M. Deza and M. Laurent. *Geometry of cuts and metrics*, volume 15 of *Algorithms and Combinatorics*. Springer-Verlag, Berlin, 1997.

[Lov79] L. Lovász. On the Shannon capacity of a graph. *IEEE Transactions on Information Theory*, 25(1):1–7, 1979.

[Ram97] M. V. Ramana. An exact duality theory for semidefinite programming and its complexity implications. *Math. Programming*, 77(2, Ser. B):129–162, 1997.

[RTW97] M. V. Ramana, L. Tunçel, and H. Wolkowicz. Strong duality for semidefinite programming. *SIAM J. Optim.*, 7(3):641–662, 1997.

[Sch35] I. J. Schoenberg. Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert". *Ann. of Math. (2)*, 36(3):724–732, 1935.

[Sha56] C. Shannon. The zero error capacity of a noisy channel. *IRE Transactions on Information Theory*, 2(3):8–19, September 1956.

[Tod01] M. Todd. Semidefinite optimization. *Acta Numerica*, 10:515–560, 2001.

[VB96] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, March 1996.

MIT 6.7230 - Algebraic techniques and semidefinite optimization February 14, 2024

Lecture 3

Lecturer: Pablo A. Parrilo *Scribe: Pablo A. Parrilo*

In this lecture we discuss one of the most important applications of semidefinite programming, namely its use in the formulation of convex relaxations of nonconvex optimization problems. We will present the results from several different, but complementary, points of view. These will also serve us as starting points for the generalizations to be presented later in the course.

We discuss first the case of binary quadratic optimization, since in this case the notation is simpler, and perfectly illustrates many of the issues appearing in more complicated problems. Afterwards, a more general formulation containing arbitrary linear and quadratic constraints will be presented.

1 Binary optimization

Binary (or Boolean) quadratic optimization is a classical combinatorial optimization problem. In the version we consider, we want to minimize a quadratic function, where the decision variables can only take the values ± 1. In other words, we are minimizing an (indefinite) quadratic form over the vertices of an n-dimensional hypercube. The problem is formally expressed as:

$$\begin{array}{rl} \text{minimize} & x^T Q x \\ \text{subject to} & x_i \in \{-1, 1\} \end{array} \tag{1}$$

where $Q \in \mathcal{S}^n$. There are many well-known problems that can be naturally written in the form above. Among these, we mention the maximum cut problem (MAXCUT) discussed below, the 0-1 knapsack, the linear quadratic regulator (LQR) control problem with binary inputs, etc.

Notice that we can model the Boolean constraints using quadratic equations, i.e.,

$$x_i \in \{-1, 1\} \quad \Longleftrightarrow \quad x_i^2 - 1 = 0.$$

These n quadratic equations define a finite set, with an exponential number of elements, namely all the n-tuples with entries in $\{-1, 1\}$. There are exactly 2^n points in this set, so a direct enumeration approach to (1) is computationally prohibitive when n is large (already for $n = 30$, we have $2^n \approx 10^9$).

We can thus write the equivalent polynomial formulation:

$$\begin{array}{rl} \text{minimize} & x^T Q x \\ \text{subject to} & x_i^2 = 1 \end{array} \tag{2}$$

We will denote the optimal value and optimal solution of this problem as $f_\star$ and $x_\star$, respectively. It is well-known that the decision version of this problem is *NP-complete* (e.g., [GJ79]). Notice that this is true even if the matrix Q is positive definite (i.e., $Q \succ 0$), since we can always make Q positive definite by adding to it a constant multiple of the identity (this only shifts the objective by a constant).

Example 1 (MAXCUT) *The maximum cut (MAXCUT) problem consists in finding a partition of the nodes of a graph $G = (V, E)$ into two disjoint sets V_1 and V_2 ($V_1 \cap V_2 = \emptyset, V_1 \cup V_2 = V$),*

in such a way to maximize the number of edges that have one endpoint in V_1 and the other in V_2. It has important practical applications, such as optimal circuit layout. The decision version of this problem (does there exist a cut with value greater than or equal to K?) is NP-complete [GJ79].

We can easily rewrite the MAXCUT problem as a binary optimization problem. A standard formulation (for the weighted problem) is the following:

$$\max_{y_i \in \{-1,1\}} \frac{1}{4} \sum_{i,j} w_{ij}(1 - y_i y_j), \tag{3}$$

where w_{ij} is the weight corresponding to the (i, j) edge, and is zero if the nodes i and j are not connected. The constraints $y_i \in \{-1, 1\}$ are equivalent to the quadratic constraints $y_i^2 = 1$.

We can easily convert the MAXCUT formulation into binary quadratic programming. Removing the constant term, and changing the sign, the original problem is clearly equivalent to:

$$\min_{y_i^2 = 1} \sum_{i,j} w_{ij} y_i y_j. \tag{4}$$

Remark Under special assumptions on the matrix Q, the binary optimization problem (1) may be solved in polynomial time. Here are some examples:

- The matrix $-Q$ is low-rank, i.e., $Q = -V^T V$ for some $V \in \mathbb{R}^{k \times n}$, where $k \ll n$ [FFL05].
- The graph associated with the sparsity pattern of Q is planar [Had75], or not contractible to K_5 [Bar83], or has bounded treewidth [BJ94].

2 Semidefinite relaxations

Computing "good" solutions to the binary optimization problem given in (2) is a quite difficult task, so it is of interest to produce accurate bounds on its optimal value. As in all minimization problems, *upper bounds* can be directly obtained from feasible points. In other words, if $x_0 \in \mathbb{R}^n$ has entries equal to ± 1, it always holds that $f_\star \leq x_0^T Q x_0$ (of course, for a poorly chosen x_0, this upper bound may be very loose).

To prove *lower bounds*, we need a different technique. There are several approaches to do this, but as we will see in detail in the next sections, many of them will turn out to be exactly equivalent in the end. Indeed, many of these different approaches will yield a characterization of a lower bound in terms of the following primal-dual pair of semidefinite programming problems:

minimize	$\operatorname{Tr} QX$	maximize	$\operatorname{Tr} \Lambda$	
subject to	$X_{ii} = 1$	subject to	$Q \succeq \Lambda$	(5)
	$X \succeq 0$		Λ diagonal	

In the next sections, we will derive these SDPs several times, in a number of different ways. Let us notice here first that for this primal-dual pair of SDP, strong duality always holds, and both achieve their corresponding optimal solutions (why?).

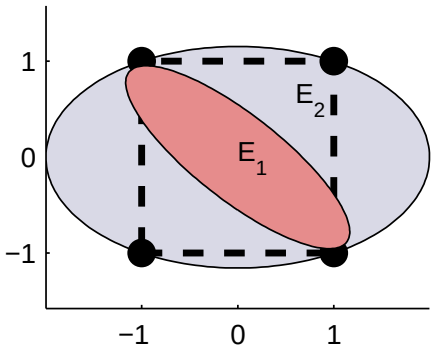

Figure 1: The ellipsoids $\mathcal{E}_1$ and $\mathcal{E}_2$.

2.1 Lagrangian duality

A general approach to obtain lower bounds on the value of (non)convex minimization problems is to use Lagrangian duality. As we have seen, the original Boolean minimization problem can be written as:

$$\begin{array}{ll} \text{minimize} & x^T Q x \\ \text{subject to} & x_i^2 - 1 = 0. \end{array} \tag{6}$$

For notational convenience, let $\Lambda := \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Then, the Lagrangian function can be written as:

$$L(x, \lambda) = x^T Q x - \sum_{i=1}^{n} \lambda_i (x_i^2 - 1) = x^T (Q - \Lambda) x + \mathrm{Tr}\, \Lambda.$$

For the dual function $g(\lambda) := \inf_x L(x, \lambda)$ to be bounded below, we need the implicit constraint that the matrix $Q - \Lambda$ must be positive semidefinite. In this case, the optimal value of x is zero, yielding $g(\lambda) = \mathrm{Tr}\, \Lambda$, and thus we obtain a lower bound on $f_\star$ given by the solution of the SDP:

$$\begin{array}{ll} \text{maximize} & \mathrm{Tr}\Lambda \\ \text{subject to} & Q - \Lambda \succeq 0 \end{array} \tag{7}$$

This is exactly the dual side of the SDP in (5).

2.2 Underestimator of the objective

A different but related interpretation of the SDP relaxation (5) is through the notion of an *underestimator* of the objective function. Indeed, the quadratic function $x^T \Lambda x$ is an "easily optimizable" function that is guaranteed to lie below the desired objective $x^T Q x$. To see this, notice that for any feasible x we have

$$x^T Q x \geq x^T \Lambda x = \sum_{i=1}^{n} \Lambda_{ii} x_i^2 = \mathrm{Tr}\, \Lambda,$$

where

- The first inequality follows from $Q \succeq \Lambda$
- The second equation holds since the matrix Λ is diagonal
- Finally, the third one holds since $x_i \in \{+1, -1\}$

There is also a nice corresponding geometric interpretation. For simplicity, we assume without loss of generality that Q is positive definite. Then, the problem (2) can be interpreted as finding the largest value of γ for which the ellipsoid $\{x \in \mathbb{R}^n | x^T Q x \leq \gamma\}$ does not contain a vertex of the unit hypercube.

Consider now the two ellipsoids in $\mathbb{R}^n$ defined by:

$$\mathcal{E}_1 = \{x \in \mathbb{R}^n \,|\, x^T Q x \leq \mathrm{Tr}\Lambda\}$$
$$\mathcal{E}_2 = \{x \in \mathbb{R}^n \,|\, x^T \Lambda x \leq \mathrm{Tr}\Lambda\}.$$

The principal axes of ellipsoid $\mathcal{E}_2$ are aligned with the coordinates axes (since Λ is diagonal), and furthermore its boundary contains all the vertices of the unit hypercube. Also, it is easy to see that the condition $Q \succeq \Lambda$ implies $\mathcal{E}_1 \subseteq \mathcal{E}_2$.

With these facts, it is easy to understand the related problem that the SDP relaxation is solving: dilating $\mathcal{E}_1$ as much as possible, while ensuring the existence of another ellipsoid $\mathcal{E}_2$ with coordinate-aligned axes and touching the hypercube in all 2^n vertices; see Figure 1 for an illustration.

2.3 Probabilistic interpretation

The standard semidefinite relaxation described above can also be motivated via a probabilistic argument. For this, assume that rather than choosing the optimal x in a deterministic fashion, we want to find instead a probability distribution that will yield "good" solutions on average. For symmetry reasons, we can always restrict ourselves to distributions with zero mean. The objective value then becomes

$$\mathbf{E}[x^T Q x] = \mathbf{E}[\mathrm{Tr}\, Q x x^T] = \mathrm{Tr}\, Q \mathbf{E}[x x^T] = \mathrm{Tr}\, Q X, \tag{8}$$

where X is the covariance matrix of the distribution (which is necessarily positive semidefinite). For the constraints, we may require that the solutions we generate be feasible on expectation, thus having:

$$\mathbf{E}[x_i^2] = X_{ii} = 1. \tag{9}$$

Maximizing the expected value of the expected cost (8), under the constraint (9) yields the primal side of the SDP relaxation presented in (5).

2.4 Lifting and rank relaxation

We present yet another derivation of the SDP relaxations, this time focused on the primal side. Recall the original formulation of the optimization problem (2). Define now $X := xx^T$. By construction, the matrix $X \in \mathcal{S}^n$ satisfies $X \succeq 0$, $X_{ii} = x_i^2 = 1$, and has *rank one*. Conversely, any matrix X with

$$X \succeq 0, \qquad X_{ii} = 1, \qquad \mathrm{rank}\, X = 1$$

necessarily has the form $X = xx^T$ for some ± 1 vector x (why?). Furthermore, by the cyclic property of the trace, we can express the objective function directly in terms of the matrix X, via:

$$x^T Q x = \mathrm{Tr}\, x^T Q x = \mathrm{Tr}\, Q x x^T = \mathrm{Tr}\, Q X.$$

As a consequence, the original problem (2) can be exactly rewritten as:

$$\begin{array}{rl} \text{minimize} & \mathrm{Tr}\, Q X \\ \text{subject to} & X_{ii} = 1, \quad X \succeq 0 \\ & \mathrm{rank}(X) = 1. \end{array}$$

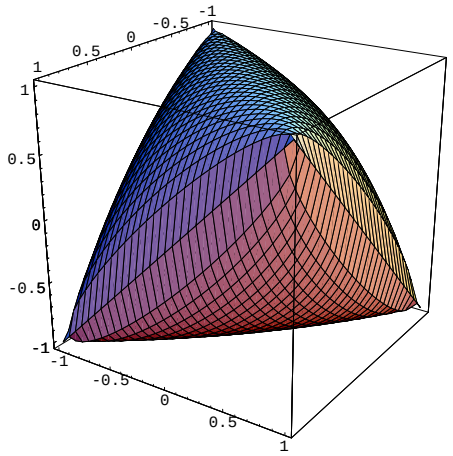

Figure 2: Three-dimensional elliptope. This is the set of 3×3 positive semidefinite matrices with unit diagonal, i.e., the set $\left\{ (x,y,z) : \begin{bmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{bmatrix} \succeq 0 \right\}$.

This is almost an SDP problem (all the constraints are either linear or conic), except for the rank one constraint on X. Since this is a minimization problem, a lower bound on the solution can be obtained by dropping the (nonconvex) rank constraint, which enlarges the feasible set.

A useful interpretation is in terms of a nonlinear *lifting* to a higher dimensional space. Indeed, rather than solving the original problem in terms of the n-dimensional vector x, we are instead solving for the $n \times n$ matrix X, effectively converting the problem from $\mathbb{R}^n$ to $\mathcal{S}^n$ (which has dimension $\binom{n+1}{2}$).

Observe that this line of reasoning immediately shows that if we find an optimal solution X of the SDP (5) that has rank one, then we have solved the original problem. Indeed, in this case the upper and lower bounds on the solution coincide.

As a graphical illustration, in Figure 2 we depict the set of 3×3 positive semidefinite matrices of unit diagonal. The rank one matrices correspond to the four "vertices" of this convex set, and are in (two-to-one) correspondence with the eight 3-vectors with ± 1 entries.

In general, it is not the case that the optimal solution of the SDP relaxation will be rank one. However, as we will see in the next section, it is possible to use *rounding schemes* to obtain "nearby" rank one solutions. Furthermore, in some cases, it is possible to do so while obtaining some approximation guarantees on the quality of the rounded solutions.

3 Bounds: Goemans-Williamson, Nesterov, and Grothendieck-Krivine

So far, our use of the SDP relaxation (5) has been limited to providing only *a posteriori* bounds on the optimal solution of the original minimization problem. However, two desirable features are missing:

- Approximation guarantees: is it possible to prove general properties on the quality of the bounds obtained by SDP?

- Feasible solutions: can we use the SDP relaxations to provide not just bounds, but actual feasible points with good (or optimal) values of the objective?

As we will see, it turns out that both questions can be answered in the positive. As it has been shown by Goemans and Williamson [GW95] in the MAXCUT case, and Nesterov, Grothendieck and Krivine in more general settings, we can actually achieve both of these objectives by randomly "rounding" in an appropriate manner the solution X of this relaxation. We discuss these results below.

3.1 Goemans-Williamson rounding

In their celebrated MAXCUT paper, Goemans and Williamson developed the following randomized method for finding a "good" feasible cut from the solution of the SDP.

- Factorize X as $X = V^T V$, where $V = [v_1 \dots v_n] \in \mathbb{R}^{r \times n}$, where r is the rank of X.
- Then $X_{ij} = v_i^T v_j$, and since $X_{ii} = 1$ this factorization gives n vectors v_i on the unit sphere in $\mathbb{R}^r$
- Instead of assigning either 1 or -1 to each variable, we have assigned to each x_i a point on the unit sphere in $\mathbb{R}^r$.
- Now, choose a random hyperplane in $\mathbb{R}^r$, and assign to each variable x_i either a $+1$ or a -1, depending on which side of the hyperplane the point v_i lies.

It turns out that this procedure gives a solution that, on average, is quite close to the value of the SDP bound. We will compute the expected value of the rounded solution in a slightly different form from the original G-W argument, but one that will be helpful later.

The random hyperplane can be characterized by its normal vector p, which is chosen to be uniformly distributed on the unit sphere (e.g., by suitably normalizing a standard multivariate Gaussian random variable). Then, according to the description above, the rounded solution is given by $x_i = \text{sign}(p^T v_i)$. The expected value of this solution can then be written as:

$$\mathbf{E}_p[x^T Q x] = \sum_{ij} Q_{ij} \mathbf{E}_p[x_i x_j] = \sum_{ij} Q_{ij} \mathbf{E}_p[\text{sign}(p^T v_i)\,\text{sign}(p^T v_j)].$$

We can easily compute the value of this expectation. Consider the plane spanned by v_i and v_j, and let θ_{ij} be the angle between these two vectors. Then, it is easy to see that the desired expectation is equal to the probability that both points are on the same side of the hyperplane, minus the probability that they are on different sides. These probabilities are $1 - \frac{\theta_{ij}}{\pi}$ and $\frac{\theta_{ij}}{\pi}$, respectively. Thus, the expected value of the rounded solution is exactly:

$$\sum_{ij} Q_{ij} \left(1 - \frac{2\theta_{ij}}{\pi}\right) = \sum_{ij} Q_{ij} \left(1 - \frac{2}{\pi} \arccos(v_i^T v_j)\right) = \frac{2}{\pi} \sum_{ij} Q_{ij} \arcsin X_{ij}. \tag{10}$$

Notice that the expression is of course well-defined, since if X is PSD and has unit diagonal, all its entries are bounded in absolute value by 1. This result exactly characterizes the expected value of the rounding procedure, as a function of the optimal solution of the SDP. We would like, however, to directly relate this quantity to the optimal solution of the original optimization problem. For this, we will need additional assumptions on the matrix Q. We discuss next three of the most important results in this direction.

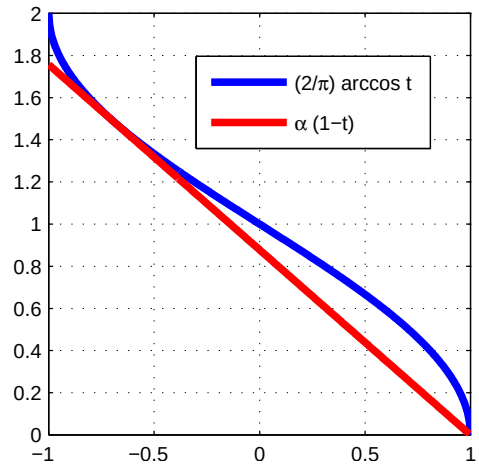

Figure 3: Bound on the inverse cosine function, for $\alpha \approx 0.878$.

3.2 MAXCUT bound

Recall from (3) that for the MAXCUT problem, the objective function does not only include the quadratic part, but there is actually a constant term:

$$\frac{1}{4}\sum_{ij} w_{ij}\,(1 - y_i y_j).$$

The expected value of the cut is then:

$$c_{\text{sdp-expected}} = \frac{1}{4}\sum_{ij} w_{ij}\,\left(1 - \frac{2}{\pi}\arcsin X_{ij}\right) = \frac{1}{4}\cdot\frac{2}{\pi}\sum_{ij} w_{ij}\arccos X_{ij},$$

where we have used the identity $\arcsin t + \arccos t = \frac{\pi}{2}$. On the other hand, the optimal solution of the primal SDP gives an upper bound on the cut capacity equal to:

$$c_{\text{sdp-upper-bound}} = \frac{1}{4}\sum_{ij} w_{ij}(1 - X_{ij}).$$

To relate these two quantities, we look for a constant α such that

$$\alpha\,(1-t) \le \frac{2}{\pi}\arccos(t) \qquad \text{for all } t \in [-1, 1]$$

The best possible (i.e., largest) such constant is $\alpha \approx 0.878$; see Figure 3. So we have

$$c_{\text{sdp-upper-bound}} \le \frac{1}{\alpha}\cdot\frac{1}{4}\cdot\frac{2}{\pi}\sum_{ij} w_{ij}\arccos X_{ij} = \frac{1}{\alpha}\,c_{\text{sdp-expected}}$$

Notice that here we have used the nonnegativity of the weights (i.e., $w_{ij} \ge 0$). Thus, so far we have the following inequalities:

- $c_{\text{sdp-upper-bound}} \le \frac{1}{\alpha}\,c_{\text{sdp-expected}}$
- Also clearly $c_{\text{sdp-expected}} \le c_{\max}$
- And $c_{\max} \le c_{\text{sdp-upper-bound}}$

Putting it all together, we can sandwich the value of the relaxation as follows:

$$\boxed{\alpha \cdot c_{\text{sdp-upper-bound}} \le c_{\text{sdp-expected}} \le c_{\max} \le c_{\text{sdp-upper-bound}}.}$$

3.3 Nesterov's $\frac{2}{\pi}$ result

A result by Nesterov generalizes the MAXCUT bound described above, but for a larger class of problems. The original formulation is for the case of binary *maximization*, and applies to the case when the matrix Q is *positive semidefinite*. Since the problem is homogeneous, the optimal value is guaranteed to be nonnegative.

As we have seen, the expected value of the solution after randomized rounding is given by (10). Since X is positive semidefinite, it follows from the nonnegativity of the Taylor series of $\arcsin(t) - t$ and the Schur product theorem that

$$\arcsin[X] \succeq X,$$

where the arcsin function is applied componentwise. This inequality can be combined with (10) to give the bounds:

$$\boxed{\frac{2}{\pi} \cdot f_{\text{sdp-upper-bound}} \leq f_{\text{sdp-expected}} \leq f_{\max} \leq f_{\text{sdp-upper-bound}},}$$

where $2/\pi \approx 0.636$. For more details, see [BTN01, Section 4.3.4]. Among others, the paper [Meg01] presents several new results, as well as a review of many of the available approximation schemes.

3.4 Grothendieck-Krivine

This case corresponds to the situation where the matrix Q has a bipartite structure, and has been analyzed in [AN04, Meg01]. We assume that the matrix Q has a structure

$$Q = \frac{1}{2} \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}.$$

An equivalent formulation is in terms of a *bilinear* optimization problem

$$\text{maximize}\, p^T A q,$$

where $A \in \mathbb{R}^{n \times m}$ and p, q are in $\{+1, -1\}^n$ and $\{+1, -1\}^m$, respectively.

This problem has a long history in operator theory and functional analysis, and has been first analyzed (in a quite different form) by Grothendieck. In fact, the optimal ratio between the bilinear binary optimization problem and its semidefinite relaxation is called the *Grothendieck constant*, and its exact value is unknown at this time. The argument below is essentially due to Krivine [Kri79], and provides an upper bound to the Grothendieck constant.

Since there are no assumptions on the entries of A, we cannot directly apply the techniques above to prove a bound on the quality of hyperplane rounding. The basic strategy in Krivine's approach is the following: instead of using hyperplane rounding directly on the solution X of the SDP relaxation, we will apply first a particular componentwise transformation, to obtain a matrix Y, and then apply hyperplane rounding to Y. The reason is that this will notably simplify the computation of the expected value of the cut.

To do this, we will use a "block" version of the argument used in Nesterov's proof:

Lemma 2 *Let $f, g : \mathbb{R} \to \mathbb{R}$ be functions such that both $f + g$ and $f - g$ have nonnegative Taylor coefficients. Let*

$$X = \begin{bmatrix} X_{11} & X_{12} \\ X_{12}^T & X_{22} \end{bmatrix}, \quad Y = \begin{bmatrix} f[X_{11}] & g[X_{12}] \\ g[X_{12}^T] & f[X_{22}] \end{bmatrix}. \tag{11}$$

Then, $X \succeq 0$ implies $Y \succeq 0$.

Now, the result follows from a specific choice of f and g. Let

$$f(t) = \sinh(c\pi t/2), \qquad g(t) = \sin(c\pi t/2).$$

where $c = \frac{2}{\pi}\sinh^{-1}(1) = \frac{2}{\pi}\log(1+\sqrt{2}) \approx 0.5611$ is chosen so $f(1) = 1$. Since

$$\sinh(t) = \sum_{k=0}^{\infty} \frac{t^{2k+1}}{(2k+1)!}, \qquad \sin(t) = \sum_{k=0}^{\infty} (-1)^k \frac{t^{2k+1}}{(2k+1)!},$$

both $f+g$ and $f-g$ have nonnegative Taylor expansions.

Let X be the optimal solution of the SDP relaxation, and define Y as in (11). Notice that the matrix Y satisfies $Y \succeq 0$ and $Y_{ii} = 1$. We can therefore apply hyperplane rounding to it, to obtain a vector y. Computing the expected value of this solution, we have:

$$\mathbf{E}[y^T Q y] = \frac{2}{\pi} \cdot Q \bullet \arcsin Y = \frac{2}{\pi} \cdot A \bullet (c\pi X_{12}/2) = c \cdot A \bullet X_{12}$$

and therefore this gives us a randomized algorithm with expected value c times the value of the SDP relaxation:

$$\boxed{c \cdot f_{\text{sdp-upper-bound}} = f_{\text{sdp-expected}} \leq f_{\max} \leq f_{\text{sdp-upper-bound}}.}$$

Recent results [BMMN13] show that the value c given above that quantifies the gap between the original problem and its SDP relaxation is not the best possible one. The "true" value of the Grothendieck constant is still unknown at this time.

4 Linearly constrained problems

In this section we extend the earlier results, to general quadratic optimization problems under linear and quadratic constraints. For notational simplicity, we write the constraints in homogeneous form, i.e., in terms of the vector $x = \begin{bmatrix} 1 & y^T \end{bmatrix}^T$.

The general primal form of the SDP optimization problems we are concerned with is

$$\begin{aligned} \text{minimize} \quad & x^T Q x \\ \text{subject to} \quad & x^T A_i x \geq 0 \\ & Bx \geq 0 \\ & x = \begin{bmatrix} 1 \\ y \end{bmatrix} \end{aligned}$$

The corresponding primal and dual SDP relaxations are given by

$$\begin{aligned} \text{minimize} \quad & Q \bullet X \\ \text{subject to} \quad & e_1^T X e_1 = 1 \\ & A_i \bullet X \geq 0 \\ & BXe_1 \geq 0 \\ & BXB^T \geq 0 \\ & X \succeq 0 \end{aligned} \qquad \begin{aligned} \text{maximize} \quad & \gamma \\ \text{subject to} \quad & Q \succeq \gamma\, e_1 e_1^T + \textstyle\sum_i \lambda_i A_i + \cdots \\ & \qquad + e_1 \mu^T B + B^T \mu e_1^T + B^T \Theta B \\ & \lambda_i \geq 0 \\ & \mu \geq 0 \\ & \Theta \geq 0 \\ & \Theta_{ii} = 0 \end{aligned} \tag{12}$$

Here e_1 is the n-vector with a 1 on the first component, and all the rest being zero. The dual variables λ_i can be interpreted as Lagrange multipliers associated to the quadratic constraints of the primal problem, while the nonnegative symmetric matrix Θ corresponds to pairwise products of the linear constraints.

References

[AN04] N. Alon and A. Naor. Approximating the cut-norm via Grothendieck's inequality. In *Proceedings of the thirty-sixth annual ACM Symposium on Theory of Computing*, pages 72–80. ACM, 2004.

[Bar83] F. Barahona. The max-cut problem on graphs not contractible to K_5. *Operations Research Letters*, 2(3):107–111, 1983.

[BJ94] H. L. Bodlaender and K. Jansen. On the complexity of the maximum cut problem. In *STACS 94*, pages 769–780. Springer, 1994.

[BMMN13] M. Braverman, K. Makarychev, Y. Makarychev, and A. Naor. The Grothendieck constant is strictly smaller than Krivine's bound. *Forum of Mathematics, Pi*, 1:e4, 2013.

[BTN01] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization*. MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001.

[FFL05] J.-A. Ferrez, K. Fukuda, and T. M. Liebling. Solving the fixed rank convex quadratic maximization in binary variables by a parallel zonotope construction algorithm. *European Journal of Operational Research*, 166(1):35–50, 2005.

[GJ79] M. R. Garey and D. S. Johnson. *Computers and Intractability: A guide to the theory of NP-completeness*. W. H. Freeman and Company, 1979.

[GW95] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, 1995.

[Had75] F. Hadlock. Finding a maximum cut of a planar graph in polynomial time. *SIAM Journal on Computing*, 4(3):221–225, 1975.

[Kri79] J.L. Krivine. Constantes de Grothendieck et fonctions de type positif sur les spheres. *Adv. Math*, 31:16–30, 1979.

[Meg01] A. Megretski. Relaxations of quadratic programs in operator theory and system analysis. In *Systems, approximation, singular integral operators, and related topics (Bordeaux, 2000)*, volume 129 of *Oper. Theory Adv. Appl.*, pages 365–392. Birkhäuser, Basel, 2001.

MIT 6.7230 - Algebraic techniques and semidefinite optimization February 21, 2024

Lecture 4

Lecturer: Pablo A. Parrilo *Scribe: Pablo A. Parrilo*

In this lecture we will review some basic elements of abstract algebra. We also introduce and begin studying the main objects of our considerations, multivariate polynomials.

1 Review: groups, rings, fields

We present here standard background material on abstract algebra. Most of the definitions are from [Lan71, CLO97, DF91, BCR98].

Definition 1 *A* group *consists of a set G and a binary operation "$\cdot$" defined on G, for which the following conditions are satisfied:*

1. *Associative: $(a \cdot b) \cdot c = a \cdot (b \cdot c)$, for all $a, b, c \in G$.*
2. *Identity: There exist $1 \in G$ such that $a \cdot 1 = 1 \cdot a = a$, for all $a \in G$.*
3. *Inverse: Given $a \in G$, there exists $b \in G$ such that $a \cdot b = b \cdot a = 1$.*

For example, the integers $\mathbb{Z}$ form a group under addition, but not under multiplication. Another example is the set $GL(n, \mathbb{R})$ of real nonsingular $n \times n$ matrices, under matrix multiplication.

If we drop the condition on the existence of an inverse, we obtain a *monoid*. Note that a monoid always has at least one element, the identity. As an example, given a set S, then the set of all strings of elements of S is a monoid, where the monoid operation is string concatenation and the identity is the empty string λ. Another example is given by $\mathbb{N}_0$, with the operation being addition (in this case, the identity is the zero). Monoids are also known as *semigroups with identity*.

In a group we only have one binary operation ("multiplication"). We will introduce another operation ("addition"), and study the structure that results from their interaction.

Definition 2 *A* commutative ring *(with identity) consists of a set k and two binary operations "$\cdot$" and "$+$", defined on k, for which the following conditions are satisfied:*

1. *Associative: $(a + b) + c = a + (b + c)$ and $(a \cdot b) \cdot c = a \cdot (b \cdot c)$, for all $a, b, c \in k$.*
2. *Commutative: $a + b = b + a$ and $a \cdot b = b \cdot a$, for all $a, b \in k$.*
3. *Distributive: $a \cdot (b + c) = a \cdot b + a \cdot c$, for all $a, b, c \in k$.*
4. *Identities: There exist $0, 1 \in k$ such that $a + 0 = a \cdot 1 = a$, for all $a \in k$.*
5. *Additive inverse: Given $a \in k$, there exists $b \in k$ such that $a + b = 0$.*

A simple example of a ring are the integers $\mathbb{Z}$ under the usual operations. After formally introducing polynomials, we will see a few more examples of rings.

If we add a requirement for the existence of multiplicative inverses, we obtain *fields*.

Definition 3 *A* field *consists of a set k and two binary operations "$\cdot$" and "$+$", defined on k, for which the following conditions are satisfied:*

1. *Associative:* $(a+b)+c = a+(b+c)$ *and* $(a \cdot b) \cdot c = a \cdot (b \cdot c)$, *for all* $a, b, c \in k$.

2. *Commutative:* $a + b = b + a$ *and* $a \cdot b = b \cdot a$, *for all* $a, b \in k$.

3. *Distributive:* $a \cdot (b + c) = a \cdot b + a \cdot c$, *for all* $a, b, c \in k$.

4. *Identities: There exist* $0, 1 \in k$, *where* $0 \neq 1$, *such that* $a + 0 = a \cdot 1 = a$, *for all* $a \in k$.

5. *Additive inverse: Given* $a \in k$, *there exists* $b \in k$ *such that* $a + b = 0$.

6. *Multiplicative inverse: Given* $a \in k, a \neq 0$, *there exists* $c \in k$ *such that* $a \cdot c = 1$.

Any field is obviously a commutative ring. Some commonly used fields are the rationals $\mathbb{Q}$, the reals $\mathbb{R}$ and the complex numbers $\mathbb{C}$. There are also Galois or finite fields (the set k has a finite number of elements), such as $\mathbb{Z}_p$, the set of integers modulo p, where p is a prime. Another important field is given by $k(x_1, \ldots, x_n)$, the set of *rational functions* with coefficients in the field k, with the natural operations.

2 Polynomials and ideals

Consider a given field k, and let $x_1, \ldots, x_n$ be indeterminates. We can then define *polynomials*.

Definition 4 *A* polynomial f *in* $x_1, \ldots, x_n$ *with coefficients in a field* k *is a finite linear combination of monomials:*

$$f = \sum_{\alpha} c_\alpha x^\alpha = \sum_{\alpha} c_\alpha x_1^{\alpha_1} \ldots x_n^{\alpha_n}, \qquad c_\alpha \in k, \tag{1}$$

where the sum is over a finite number of n*-tuples* $\alpha = (\alpha_1, \ldots, \alpha_n)$, $\alpha_i \in \mathbb{N}_0$. *The set of all polynomials in* $x_1, \ldots, x_n$ *with coefficients in* k *is denoted* $k[x_1, \ldots, x_n]$.

It follows from the previous definitions that $k[x_1, \ldots, x_n]$, i.e., the set of polynomials in n variables with coefficients in k, is a commutative ring with identity. We also notice that it is possible (and sometimes, convenient) to define polynomials where the coefficients belong to a ring with identity, not necessarily to a field.

Definition 5 *A* form *is a polynomial where all the monomials have the same degree* $d := \sum_i \alpha_i$. *In this case, the polynomial is* homogeneous *of degree* d, *since it satisfies* $f(\lambda x_1, \ldots, \lambda x_n) = \lambda^d f(x_1, \ldots, x_n)$.

A polynomial in n variables of degree d has $\binom{n+d}{d}$ coefficients. Since there is a natural bijection between n-variate forms and $(n-1)$-variate polynomials via homogenization, it then follows that a form in n variables of degree d has $\binom{n+d-1}{d}$ coefficients.

A commutative ring is called an *integral domain* if it has no *zero divisors*, i.e. $a \neq 0, b \neq 0 \Rightarrow a \cdot b \neq 0$. Every field is also an integral domain (why?). Two examples of rings that are not integral domains are the set of matrices $\mathbb{R}^{n \times n}$, and the set of integers modulo n, when n is a composite number (with the usual operations). If k is an integral domain, then so is $k[x_1, \ldots, x_n]$.

Remark 6 *Another important example of a ring (in this case, non-commutative) appears in systems and control theory, through the ring* $\mathcal{M}(s)$ *of stable proper rational functions. This is the set of matrices (of fixed dimension) whose entries are rational functions of* s *(i.e., in the field* $\mathbb{C}(s)$*), are bounded at infinity, and have all poles in the strict left-half plane. In this algebraic setting (usually called "coprime factorization approach"), the question of finding a stabilizing controller is exactly equivalent to the solvability of a Diophantine equation* $ax + by = 1$.

	Formally real	Not formally real
Algebraically closed	———	$\mathbb{C}$
Not algebraically closed	$\mathbb{R}, \mathbb{Q}$	finite fields $\mathbb{F}_{p^k}$

Table 1: Examples of fields.

2.1 Algebraically closed and formally real fields

A very important property of a univariate polynomial p is the existence of a *root*, i.e., an element x_0 for which $p(x_0) = 0$. Depending on the solvability of these equations, we can characterize a particular nice class of fields.

Definition 7 *A field k is* algebraically closed *if every nonconstant polynomial in $k[x]$ has a root in k.*

If a field is algebraically closed, then it has an infinite number of elements (why?). What can we say about the most usual fields, $\mathbb{C}$ and $\mathbb{R}$? The Fundamental Theorem of Algebra ("every univariate polynomial has at least one complex root") shows that $\mathbb{C}$ is an algebraically closed field.

However, this is clearly *not* the case of $\mathbb{R}$, since for instance the polynomial x^2+1 does not have any real root. The lack of algebraic closure of $\mathbb{R}$ is one of the main sources of complications when dealing with systems of polynomial equations and inequalities. To deal with the case when the base field is not algebraically closed, the *Artin-Schreier* theory of *formally real fields* was introduced.

The starting point is one of the intrinsic properties of $\mathbb{R}$:

$$\sum_{i=1}^{n} x_i^2 = 0 \quad \Longrightarrow \quad x_1 = \cdots = x_n = 0. \tag{2}$$

A field will be called *formally real* if it satisfies the above condition (clearly, $\mathbb{R}$ and $\mathbb{Q}$ are formally real, but $\mathbb{C}$ is not). As we can see from the definition, the theory of formally real fields has very strong connections with sums of squares, a notion that will reappear in several forms later in the course. For example, an alternative (but equivalent) statement of (2) is to say that a field is formally real if and only if the element -1 is not a sum of squares.

The relationships between these concepts, as well as a few examples, are presented in Table 2.1. Notice that if a field is algebraically closed, then it cannot be formally real, since we have that $(\sqrt{-1})^2 + 1^2 = 0$ (and $\sqrt{-1}$ is in the field).

A related important notion is that of an *ordered* field:

Definition 8 *A field k is said to be* ordered *if a relation $>$ is defined on k, that satisfies*

1. *If $a, b \in k$, then either $a > b$ or $a = b$ or $b > a$.*
2. *If $a > b$, $c \in k$, $c > 0$ then $ac > bc$.*
3. *If $a > b$, $c \in k$, then $a + c > b + c$.*

A crucial result relating these two notions is the following:

Lemma 9 *A field can be ordered if and only if it is formally real.*

For a field to be ordered (or equivalently, formally real), it necessarily must have an infinite number of elements. This is somewhat unfortunate, since this rules out several modular methods for dealing with real solutions to polynomial inequalities.

2.2 Ideals

We consider next *ideals*, which are subrings with an "absorbent" property:

Definition 10 *Let R be a commutative ring. A subset $I \subset R$ is an* ideal *if it satisfies:*

1. $0 \in I$.
2. *If $a, b \in I$, then $a + b \in I$.*
3. *If $a \in I$ and $b \in R$, then $a \cdot b \in I$.*

A simple example of an ideal is the set of even integers, considered as a subset of the integer ring $\mathbb{Z}$. Another important example is the set of nilpotent elements of a ring, i.e., those $x \in R$ for which there exists a positive integer k such that $x^k = 0$. Also, notice that if the ideal I contains the multiplicative identity 1, then $I = R$.

To introduce another important example of ideals, we need to define the concept of an algebraic variety as the zero set of a set of polynomial equations:

Definition 11 *Let k be a field, and let $f_1, \ldots, f_s$ be polynomials in $k[x_1, \ldots, x_n]$. Let the set $\mathbf{V}$ be*

$$\mathbf{V}(f_1, \ldots, f_s) = \{(a_1, \ldots, a_n) \in k^n : f_i(a_1, \ldots, a_n) = 0 \qquad \forall 1 \leq i \leq s\}.$$

We call $\mathbf{V}(f_1, \ldots, f_s)$ the affine variety *defined by $f_1, \ldots, f_s$.*

Then, the set of polynomials that vanish in a given variety, i.e.,

$$\mathbf{I}(V) = \{f \in k[x_1, \ldots, x_n] : f(a_1, \ldots, a_n) = 0 \qquad \forall (a_1, \ldots, a_n) \in V\},$$

is an ideal, called the *ideal of V*.

By Hilbert's Basis Theorem [CLO97], $k[x_1, \ldots, x_n]$ is a *Noetherian* ring, i.e., every ideal $I \subset k[x_1, \ldots, x_n]$ is finitely generated. In other words, there always exists a finite set $f_1, \ldots, f_s \in k[x_1, \ldots, x_n]$ such that for every $f \in I$, we can find $g_i \in k[x_1, \ldots, x_n]$ that verify $f = \sum_{i=1}^{s} g_i f_i$.

We also define the *radical* of an ideal:

Definition 12 *Let $I \subset k[x_1, \ldots, x_n]$ be an ideal. The* radical *of I, denoted $\sqrt{I}$, is the set*

$$\{f \mid f^k \in I \text{ for some integer } k \geq 1\}.$$

It is clear that $I \subset \sqrt{I}$, and it can be shown that $\sqrt{I}$ is also a polynomial ideal. A very important result, that we will see later in some detail, is the following:

Theorem 13 (Hilbert's Nullstellensatz) *If I is a polynomial ideal, then $\mathbf{I}(\mathbf{V}(I)) = \sqrt{I}$.*

2.3 Associative algebras

Another important notion, that we will encounter at least twice later in the course, is that of an *associative algebra*.

Definition 14 *An* associative algebra $\mathcal{A}$ *over $\mathbb{C}$ is a vector space with a $\mathbb{C}$-bilinear operation $\cdot : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$ that satisfies*

$$x \cdot (y \cdot z) = (x \cdot y) \cdot z, \qquad \forall x, y, z \in \mathcal{A}.$$

In general, associative algebras do not need to be commutative (i.e., $x \cdot y = y \cdot x$). However, that is an important special case, with many interesting properties. We list below several examples of finite dimensional associative algebras.

- Full matrix algebra $\mathbb{C}^{n \times n}$, standard product.
- The subalgebra of square matrices with equal row and column sums.
- The diagonal, lower triangular, or circulant matrices.
- The n-dimensional algebra generated by a single $n \times n$ matrix.
- The incidence algebra of a partially ordered finite set.
- The Clifford algebra, that generalizes the reals, complex, quaternions, ...
- The group algebra: formal $\mathbb{C}$-linear combination of group elements.
- Polynomial multiplication modulo a zero dimensional ideal.
- The Bose-Mesner algebra of an association scheme.

We will discuss the last three in more detail later in the course.

3 Questions about polynomials

There are many natural questions that we may want to answer about polynomials, even in the univariate case. Among them, we mention:

- When does a univariate polynomial have *only* real roots?
- What conditions must it satisfy for *all* roots to be real?
- When does a polynomial satisfy $p(x) \geq 0$ for all x?

We will answer many of these next week.

References

[BCR98] J. Bochnak, M. Coste, and M-F. Roy. *Real Algebraic Geometry*. Springer, 1998.

[CLO97] D. A. Cox, J. B. Little, and D. O'Shea. *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer, 1997.

[DF91] D. S. Dummit and R. M. Foote. *Abstract algebra*. Prentice Hall Inc., Englewood Cliffs, NJ, 1991.

[Lan71] S. Lang. *Algebra*. Addison-Wesley, 1971.

MIT 6.7230 - Algebraic techniques and semidefinite optimization February 23, 2024

Lecture 5

Lecturer: Pablo A. Parrilo *Scribe: Pablo A. Parrilo*

In this lecture we study univariate polynomials, particularly questions regarding the existence of real roots and nonnegativity conditions. For instance:

- When does a univariate polynomial have *only* real roots?
- What conditions must it satisfy for *all* roots to be real?
- When is a polynomial nonnegative, i.e., it satisfies $p(x) \geq 0$ for all $x \in \mathbb{R}$?

1 Univariate polynomials

A univariate polynomial $p(x) \in \mathbb{R}[x]$ of degree n has the form:

$$p(x) = p_n x^n + p_{n-1} x^{n-1} + \cdots + p_1 x + p_0, \tag{1}$$

where the coefficients p_k are real. We normally assume $p_n \neq 0$, and occasionally we will normalize it to $p_n = 1$, in which case we say that $p(x)$ is *monic*.

As we have seen, the field $\mathbb{C}$ of complex numbers is algebraically closed:

Theorem 1 (Fundamental theorem of algebra)**.** *Every nonzero univariate polynomial of degree n has exactly n complex roots (counted with multiplicity). Furthermore, we have the unique factorization*

$$p(x) = p_n \prod_{k=1}^{n} (x - x_k),$$

where $x_k \in \mathbb{C}$ are the roots *of $p(x)$.*

If all the coefficients p_k are real and x_k is a root, then so is its complex conjugate x_k^*. In other words, all complex roots appear in complex conjugate pairs. In particular, if the degree n is odd, then there is always at least one real root.

2 Counting real roots

How many *real roots* does a polynomial have? There are many options, ranging from all roots being real (e.g., $(x-1)(x-2)\ldots(x-n)$), to all roots being complex (e.g., $x^{2d}+1$). We will give a couple of different characterizations of the location of the roots of a polynomial, both of them in terms of some associated symmetric matrices.

2.1 The companion matrix

A very well-known relationship between univariate polynomials and matrices is given through the so-called companion matrix.

Definition 2. *The* companion matrix $\mathcal{C}_p$ *associated with the polynomial* $p(x)$ *in (1) is the* $n \times n$ *real matrix*

$$\mathcal{C}_p := \begin{bmatrix} 0 & 0 & \cdots & 0 & -p_0/p_n \\ 1 & 0 & \cdots & 0 & -p_1/p_n \\ 0 & 1 & \cdots & 0 & -p_2/p_n \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -p_{n-1}/p_n \end{bmatrix}.$$

Lemma 3. *The characteristic polynomial of* $\mathcal{C}_p$ *is (up to a constant) equal to* $p(x)$*. Formally,* $\det(xI - \mathcal{C}_p) = \frac{1}{p_n} p(x)$*.*

From this lemma, it directly follows that the eigenvalues of $\mathcal{C}_p$ are exactly equal to the roots x_i of $p(x)$, including multiple roots the appropriate number of times. In other words, if we want to obtain the roots of a polynomial, we can do this by computing instead the eigenvalues of the associated (nonsymmetric) companion matrix. In fact, that is exactly the way that MATLAB computes roots of polynomials; see the source file `roots.m`.

Left and right eigenvectors The companion matrix $\mathcal{C}_p$ is diagonalizable if and only the polynomial $p(x)$ has no multiple roots. What are the corresponding diagonalizing matrices (equivalently, the right and left eigenvectors)?

Define the $n \times n$ *Vandermonde matrix*

$$V = \begin{bmatrix} 1 & x_1 & \ldots & x_1^{n-1} \\ 1 & x_2 & \ldots & x_2^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^{n-1} \end{bmatrix} \tag{2}$$

where $x_1, \ldots, x_n \in \mathbb{C}$. It can be shown that the matrix V is nonsingular if and only if all the roots x_i are distinct. We have then the identity

$$V \cdot \mathcal{C}_p = \operatorname{diag}[x_1, \ldots, x_n] \cdot V, \tag{3}$$

and thus the left eigenvectors of $\mathcal{C}_p$ are the rows of the Vandermonde matrix.

The right eigenvectors are of course given by the columns of V^{-1}, as can be easily seen by left- and right-multiplying (3) by this inverse. A natural interpretation of this dual basis (i.e., the columns of V^{-1}) is in terms of the *Lagrange interpolating polynomials* of the points x_i. These are a set of n univariate polynomials that satisfy the property $L_j(x_i) = \delta_{ij}$, where δ is the Kronecker delta. It is easy to verify that the columns of V^{-1} are the coefficients (in the monomial basis) of the corresponding Lagrange interpolating polynomials.

Example 4. *Consider the polynomial* $p(x) = (x-1)(x-2)(x-5)$*. Its companion matrix is*

$$\mathcal{C}_p = \begin{bmatrix} 0 & 0 & 10 \\ 1 & 0 & -17 \\ 0 & 1 & 8 \end{bmatrix},$$

and it is diagonalizable since p has simple roots. Ordering the roots as $\{1, 2, 5\}$*, the corresponding Vandermonde matrix and its inverse are:*

$$V = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 5 & 25 \end{bmatrix}, \qquad V^{-1} = \frac{1}{12} \begin{bmatrix} 30 & -20 & 2 \\ -21 & 24 & -3 \\ 3 & -4 & 1 \end{bmatrix}.$$

From the columns of V^{-1}, we can read the coefficients of the Lagrange interpolating polynomials; e.g., $L_1(x) = (30 - 21x + 3x^2)/12 = (x-2)(x-5)/4$.

Symmetric functions of roots For any $A \in \mathbb{C}^{n\times n}$, we always have $\mathrm{Tr}A = \sum_{i=1}^n \lambda_i(A)$, and $\lambda_i(A^k) = \lambda_i(A)^k$. Therefore, it follows that $\sum_{i=1}^n x_i^k = \mathrm{Tr}\,[\mathcal{C}_p^k]$. As a consequence of linearity, we have that if $q(x) = \sum_{j=0}^m q_j x^j$ is a univariate polynomial,

$$\sum_{i=1}^n q(x_i) = \sum_{i=1}^n \sum_{j=0}^m q_j x_i^j = \sum_{j=0}^m q_j \mathrm{Tr}[\mathcal{C}_p^j] = \mathrm{Tr}[\sum_{j=0}^m q_j \mathcal{C}_p^j] = \mathrm{Tr}\,[q(\mathcal{C}_p)], \tag{4}$$

where the expression $q(\mathcal{C}_p)$ indicates the evaluation of the polynomial $q(x)$ on the companion matrix of $p(x)$. Note that if p is monic, then the final expression in (4) is a polynomial in the coefficients of p. This is an identity that we will use several times in the sequel.

Remark 5. *Our presentation of the companion matrix has been somewhat unmotivated, other than noticing that "it just works." After presenting some additional material on Gröbner bases, we will revisit this construction, where we will give a natural interpretation of $\mathcal{C}_p$ as representing a well-defined linear operator in the quotient ring $\mathbb{R}[x]/\langle p(x)\rangle$. This will enable a very appealing extension of many results about companion matrices to multivariate polynomials, in the case where the underlying system has only a finite number of solutions (i.e., a "zero dimensional ideal"). For instance, the generalization of the diagonalizability of the companion matrix $\mathcal{C}_p$ when $p(x)$ has only simple roots will be the fact that the multiplication algebra associated with a zero-dimensional ideal is semisimple if and only if the ideal is radical.*

2.2 Inertia and signature

Definition 6. *Consider a symmetric matrix A. The* inertia *of A, denoted $\mathcal{I}(A)$, is the integer triple (n_+, n_0, n_-), where n_+, n_0, n_- are the number of positive, zero, and negative eigenvalues, respectively. The* signature *of A is equal to the number of positive eigenvalues minus the number of negative eigenvalues, i.e., the integer $n_+ - n_-$.*

Notice that, with the notation above, the rank of A is equal to $n_+ + n_-$. A symmetric positive definite $n \times n$ matrix has inertia $(n, 0, 0)$, while a positive semidefinite one has $(n-k, k, 0)$ for some $k \geq 0$. The inertia is an important invariant of a quadratic form, since it holds that $\mathcal{I}(A) = \mathcal{I}(T^*AT)$, where T is nonsingular. This invariance of the inertia of a matrix under congruence transformations is known as *Sylvester's law of inertia*; see for instance [HJ95].

2.3 The Hermite form

While the companion matrix is quite useful, we will present now a different characterization of the roots of a polynomial. Among others, an advantage of this formulation is the fact that we will be using *symmetric* matrices.

Let $q(x)$ be a fixed auxiliary polynomial. Consider the following $n \times n$ symmetric Hankel matrix $H_q(p)$ with entries defined by

$$[H_q(p)]_{jk} = \sum_{i=1}^n q(x_i) x_i^{j+k-2}. \tag{5}$$

Like every symmetric matrix, $H_q(p)$ defines an associated quadratic form via

$$\begin{aligned} f^T H_q(p) f &= \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \end{bmatrix}^T \begin{bmatrix} \sum_{i=1}^n q(x_i) & \sum_{i=1}^n q(x_i)x_i & \cdots & \sum_{i=1}^n q(x_i)x_i^{n-1} \\ \sum_{i=1}^n q(x_i)x_i & \sum_{i=1}^n q(x_i)x_i^2 & \cdots & \sum_{i=1}^n q(x_i)x_i^n \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n q(x_i)x_i^{n-1} & \sum_{i=1}^n q(x_i)x_i^n & \cdots & \sum_{i=1}^n q(x_i)x_i^{2n-2} \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_{n-1} \end{bmatrix} \\ &= \sum_{i=1}^n q(x_i)(f_0 + f_1 x_i + \cdots + f_{n-1}x_i^{n-1})^2 \\ &= \mathrm{Tr}[(qf^2)(\mathcal{C}_p)]. \end{aligned}$$

Although not immediately obvious from the definition (5), the expression above shows that when $p(x)$ is monic, the entries of $H_q(p)$ are actually polynomials in the coefficients of $p(x)$. Notice that we have used (4) in the derivation of the last step.

Recall that a Vandermonde matrix defines a linear transformation mapping the coefficients of a degree $n-1$ polynomial f to its values $\big[f(x_1), \ldots, f(x_n)\big]$. Since this transformation is invertible, given any $y \in \mathbb{R}^n$ there always exists an f of degree $n-1$ such that $f(x_i) = y_i$ (i.e., there is always an interpolating polynomial). From expression (5), we have the factorization

$$H_q(p) = V^T \,\mathrm{diag}[q(x_1), \ldots, q(x_n)]\, V.$$

This is almost a congruence transformation, except that there are complex entries in V if some of the x_i are complex. However, this can be easily resolved, to obtain the theorem below.

Theorem 7. *The signature of $H_q(p)$ is equal to the number of real roots x_j of p for which $q(x_j) > 0$, minus the number of real roots for which $q(x_j) < 0$.*

Proof. For simplicity, we assume all roots are distinct (this is easy to change, at the expense of slightly more complicated notation). We have then

$$\begin{aligned} f^T H_q(p) f &= \sum_{j=1}^n q(x_j)(f_0 + f_1 x_j + \cdots + f_{n-1}x_j^{n-1})^2 \\ &= \sum_{x_j \in \mathbb{R}} q(x_j) f(x_j)^2 + \sum_{x_j, x_j^* \in \mathbb{C}\setminus\mathbb{R}} q(x_j) f(x_j)^2 + q(x_j^*) f(x_j^*)^2 \\ &= \sum_{x_j \in \mathbb{R}} q(x_j) f(x_j)^2 + 2 \sum_{x_j, x_j^* \in \mathbb{C}\setminus\mathbb{R}} \begin{bmatrix} \Re f(x_j) \\ \Im f(x_j) \end{bmatrix}^T \begin{bmatrix} \Re q(x_j) & -\Im q(x_j) \\ -\Im q(x_j) & -\Re q(x_j) \end{bmatrix} \begin{bmatrix} \Re f(x_j) \\ \Im f(x_j) \end{bmatrix}. \end{aligned}$$

Notice that an expression of the type $f(x_i)$ is a linear form in $[f_0, \ldots, f_{n-1}]$. Because of the assumption that all the roots x_j are distinct, the linear forms $\{f(x_j)\}_{j=1,\ldots,n}$ are linearly independent (the corresponding Vandermonde matrix is nonsingular), and thus so are $\{f(x_j)\}_{x_j \in \mathbb{R}} \cup \{\Re f(x_j), \Im f(x_j)\}_{x_j \in \mathbb{C}\setminus\mathbb{R}}$. Therefore, the expression above gives a congruence transformation of $H_q(p)$, and we can obtain its signature by adding the signatures of the scalar elements $q(x_j)$ and the 2×2 blocks. The signature of the 2×2 blocks is always zero (they have zero trace), and thus the result follows. □

In particular, notice that if we want to count the number of real roots, we can just use $q(x) = 1$.

The matrix corresponding to this quadratic form (called the *Hermite form*) is:

$$H_1(p) = V^T V = \begin{bmatrix} s_0 & s_1 & \cdots & s_{n-1} \\ s_1 & s_2 & \cdots & s_n \\ \vdots & \vdots & \ddots & \vdots \\ s_{n-1} & s_n & \cdots & s_{2n-2} \end{bmatrix}, \qquad s_k = \sum_{j=1}^{n} x_j^k.$$

The s_k are known as the *power sums* and can be computed using (4) (although there are much more efficients ways, such as the Newton identities). When $p(x)$ is monic, the s_k are polynomials of degree k in the coefficients of $p(x)$.

Example 8. *Consider the monic cubic polynomial*

$$p(x) = x^3 + p_2 x_2 + p_1 x + p_0.$$

Then, the first five power sums are:

$$\begin{aligned} s_0 &= 3 \\ s_1 &= -p_2 \\ s_2 &= p_2^2 - 2p_1 \\ s_3 &= -p_2^3 + 3p_1 p_2 - 3p_0 \\ s_4 &= p_2^4 - 4p_1 p_2^2 + 2p_1^2 + 4p_0 p_2. \end{aligned}$$

Lemma 9. *The signature of $H_1(p)$ is equal to the number of real roots. The rank of $H_1(p)$ is equal to the number of distinct complex roots of $p(x)$.*

Corollary 10. *If $p(x)$ has odd degree, there is always at least one real root.*

Example 11. *Consider $p(x) = x^3 + 2x^2 + 3x + 4$. The corresponding Hermite matrix is:*

$$H(p) = \begin{bmatrix} 3 & -2 & -2 \\ -2 & -2 & -2 \\ -2 & -2 & 18 \end{bmatrix}$$

This matrix has one negative and two positive eigenvalues, all distinct (i.e., its inertia is $(2, 0, 1)$). Thus, $p(x)$ has three simple roots, and exactly one of them is real.

Sylvester's law of inertia guarantees that this result is actually coordinate independent.

3 Nonnegativity

An important property of a polynomial is whether it only takes nonnegative values. As we will see, this is of interest in a wide variety of applications.

Definition 12. *A univariate polynomial $p(x)$ is* positive semidefinite *or* nonnegative *if $p(x) \geq 0$ for all real values of x.*

Clearly, if $p(x)$ is nonnegative, then its degree must be an even number. The set of nonnegative polynomials has very interesting properties. Perhaps the most appealing one for our purposes is the following:

Theorem 13. *Consider the set P_n of nonnegative univariate polynomials of degree less than or equal to n (n is even). Then, identifying a polynomial with its $n+1$ coefficients $(p_n, \ldots, p_0)$, the set P_n is a proper cone (i.e., closed, convex, pointed, solid) in $\mathbb{R}^{n+1}$.*

An equivalent condition for the (nonconstant) univariate polynomial (1) to be strictly positive, is that $p(x_0) > 0$ for some x_0, and it that has no real roots. Thus, we can use Theorem 7 to write explicit conditions for a polynomial $p(x)$ to be nonnegative in terms of the signature of the associated Hermite matrix $H_1(p)$.

4 Sum of squares

Definition 14. *A univariate polynomial $p(x)$ is a* sum of squares *(SOS) if there exist $q_1, \ldots, q_m \in \mathbb{R}[x]$ such that*

$$p(x) = \sum_{k=1}^{m} q_k^2(x).$$

If a polynomial $p(x)$ is a sum of squares, then it obviously satisfies $p(x) \geq 0$ for all $x \in \mathbb{R}$. Thus, a SOS condition is a sufficient condition for global nonnegativity.

Interestingly, in the univariate case, the converse is also true:

Theorem 15. *A univariate polynomial is nonnegative if and only if it is a sum of squares.*

Proof. ($\Leftarrow$) Obvious. If $p(x) = \sum_k q_k^2(x)$ then $p(x) \geq 0$.

($\Rightarrow$) Since $p(x)$ is univariate, we can factorize it as

$$p(x) = p_n \prod_j (x - r_j)^{n_j} \prod_k (x - a_k + ib_k)^{m_k} (x - a_k - ib_k)^{m_k},$$

where r_j and $a_k \pm ib_k$ are the real and complex roots, respectively, of multiplicities n_j and m_k. Because $p(x)$ is nonnegative, then $p_n > 0$ and the multiplicies of the real roots are even, i.e., $n_j = 2s_j$.

Notice that $(x - a + ib)(x - a - ib) = (x - a)^2 + b^2$. Then, we can write

$$p(x) = p_n \prod_j (x - r_j)^{2s_j} \prod_k \left((x - a_k)^2 + b_k^2\right)^{m_k},$$

Since products of sums of squares are sums of squares, and all the factors in the expression above are SOS, it follows that $p(x)$ is SOS.

Furthermore, the two-squares identity $(\alpha^2 + \beta^2)(\gamma^2 + \delta^2) = (\alpha\gamma - \beta\delta)^2 + (\alpha\delta + \beta\gamma)^2$ allows us to combine every partial product as a sum of only two squares. □

Notice that the proof shows that if $p(x)$ is SOS, then there exists a representation $p(x) = q_1^2(x) + q_2^2(x)$.

As we will see very soon, we can decide whether a univariate polynomial is a sum of squares (equivalently, if it is nonnegative) by solving a semidefinite optimization problem.

5 Positive semidefinite matrices

Recall from Lecture 2 the (apparent) disparity between the stated conditions for a matrix to be positive definite versus the semidefinite case. In the former, we could use a test (Sylvester's criterion) that required the calculation of only n minors, while for the semidefinite case apparently we needed a much larger number, $2^n - 1$.

If the matrix X is positive definite, Sylvester's criterion requires the positivity of the leading principal minors, i.e.,

$$\det X_{1,1} > 0, \quad \det X_{12,12} > 0, \quad \ldots, \quad \det X > 0.$$

For positive semidefiniteness, it is not enough to replace strict positivity with the nonstrict inequality; a simple counterexample is the matrix

$$\begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix},$$

for which the leading minors vanish, but is not PSD. As mentioned, an alternative approach is given by the following classical result:

Lemma 16. *Let $A \in \mathcal{S}^n$ be a symmetric matrix. Then $A \succeq 0$ if and only if all $2^n - 1$ principal minors of A are nonnegative.*

Although the condition above requires the nonnegativity of $2^n - 1$ expressions, it is possible to do the same by checking only n inequalities:

Theorem 17 (e.g. [HJ95, p. 403])**.** *A real $n \times n$ symmetric matrix A is positive semidefinite if and only if all the coefficients of its characteristic polynomial $p(\lambda) = \det(\lambda I - A) = \lambda^n + p_{n-1}\lambda^{n-1} + \cdots + p_1\lambda + p_0$ alternate in sign, i.e., they satisfy $p_i(-1)^{n-i} \geq 0$.*

We prove this below, since we will use a slightly more general version of this result when discussing hyperbolic polynomials. Note that in the $n = 2$ case, Theorem 17 is the familiar result that $A \in \mathcal{S}^2$ is positive semidefinite if and only if $\det A \geq 0$ and $\mathrm{Tr} A \geq 0$.

Lemma 18. *Consider a monic univariate polynomial $p(t) = t^n + \sum_{k=0}^{n-1} p_k t^k$, that has only real roots. Then, all roots are nonpositive if and only if all coefficients are nonnegative (i.e., $p_k \geq 0, k = 0, \ldots, n-1$).*

Proof. Since all roots of $p(t)$ are real, this can be obtained from a direct application of Descartes' rules of signs; see e.g. [BPR03]. For completeness, we present here a direct proof.

If all roots t_i are nonpositive ($t_i \leq 0$), from the factorization

$$p(t) = \prod_{k=1}^{n} (t - t_i)$$

it follows directly that all coefficients p_k are nonnegative.

For the other direction, from the nonnegativity of the coefficients it follows that $p(0) \geq 0$ and $p(t)$ is nondecreasing. If there exists a $t_i > 0$ such that $p(t_i) = 0$, then the polynomial must vanish in the interval $[0, t_i]$, which is impossible since it is monic and hence nonzero. □

Definition 19. *A set $S \subset \mathbb{R}^n$ is* basic closed semialgebraic *if it can be written as*

$$S = \{x \in \mathbb{R}^n \mid f_i(x) \geq 0, \quad h_j(x) = 0\}$$

for some finite set of polynomials $\{f_i, h_j\}$.

Theorem 20. *Both the primal and dual feasible sets of a semidefinite program are basic closed semialgebraic.*

Proof. The condition $X \succeq 0$ is equivalent to n nonstrict polynomial inequalities in the entries of X. This can be conveniently shown applying Lemma 18 to the characteristic polynomial of $-X$, i.e.,

$$p(\lambda) = \det(\lambda I + X) = \lambda^n + \sum_{k=0}^{n-1} p_k(X)\lambda^k.$$

where the $p_k(X)$ are homogeneous polynomials of degree $n-k$ in the entries of X. For instance, we have $p_0(X) = \det X$, and $p_{n-1}(X) = \mathrm{Tr}X$.

Since X is symmetric, all its eigenvalues are real, and thus $p(\lambda)$ has only real roots. Positive semidefiniteness of X is equivalent to $p(\lambda)$ having no roots that are strictly positive. It then follows than the two following statements are equivalent:

$$X \succeq 0 \quad \Leftrightarrow \quad p_k(X) \geq 0 \quad k = 0, \ldots, n-1.$$

□

Remark 21. *These inequalities correspond to the elementary symmetric functions e_k evaluated at the eigenvalues of the matrix X.*

As we will see in subsequent lectures, the same inequalities will reappear when we consider a class of optimization problems known as *hyperbolic programs*.

References

[BPR03] S. Basu, R. Pollack, and M.-F. Roy. *Algorithms in real algebraic geometry*, volume 10 of *Algorithms and Computation in Mathematics*. Springer-Verlag, Berlin, 2003.

[HJ95] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1995.

MIT 6.7230 - Algebraic techniques and semidefinite optimization March 1, 2024

Lecture 6

Lecturer: Pablo A. Parrilo *Scribe: ???*

Last week we learned about explicit conditions to determine the number of real roots of a univariate polynomial. Today we will expand on these themes, and study two mathematical objects of fundamental importance: the *resultant* of two polynomials, and the closely related *discriminant*.

The resultant will be used to decide whether two univariate polynomials have common roots, while the discriminant will give information about the existence of multiple roots. Furthermore, we will see the intimate connections between discriminants and the boundary of the cone of nonnegative polynomials.

Besides the properties described above, a direct consequence of their definitions, there are many other interesting applications of resultants and discriminant. We describe a few of them below, and we will encounter them again in later lectures, when studying elimination theory and the construction of cylindrical algebraic decompositions. For much more information about resultants and discriminants, particularly their generalizations to the sparse and multipolynomial case, we refer the reader to the very readable introductory article [Stu98] and the books [CLO97, GKZ94].

1 Resultants

Consider two polynomials $p(x)$ and $q(x)$, of degree n, m, respectively. We want to obtain an easily checkable criterion to determine whether they have a common root, that is, there exists an $x_0 \in \mathbb{C}$ for which $p(x_0) = q(x_0) = 0$. There are several approaches, seemingly different at first sight, for constructing such a criterion:

- **Sylvester matrix**: If $p(x_0) = q(x_0) = 0$, then we can write the following $(n+m) \times (n+m)$ linear system:

$$\begin{bmatrix} p_n & p_{n-1} & \cdots & p_1 & p_0 & & & & \\ & p_n & & \ddots & \ddots & \ddots & & & \\ \vdots & & & & & & & & \\ & & & & & & p_1 & p_0 & \\ & & & & & & p_2 & p_1 & p_0 \\ q_m & q_{m-1} & \cdots & q_0 & & & & & \\ & q_m & & \ddots & \ddots & & & & \\ & & \ddots & & & & & & \\ & & & & & & q_1 & q_0 & \\ & & & & & & q_2 & q_1 & q_0 \end{bmatrix} \begin{bmatrix} x_0^{n+m-1} \\ x_0^{n+m-2} \\ \vdots \\ x_0^n \\ x_0^{n-1} \\ \vdots \\ x_0 \\ 1 \end{bmatrix} = \begin{bmatrix} p(x_0)x_0^{m-1} \\ p(x_0)x_0^{m-2} \\ \vdots \\ p(x_0)x_0 \\ p(x_0) \\ q(x_0)x_0^{n-1} \\ q(x_0)x_0^{n-2} \\ \vdots \\ q(x_0)x_0 \\ q(x_0) \end{bmatrix} = 0.$$

  This implies that the matrix on the left-hand side, called the *Sylvester matrix* $\mathrm{Syl}_x(p,q)$ associated to p and q, is singular and thus its determinant must vanish. It is not too difficult to show that the converse is also true; if $\det \mathrm{Syl}_x(p,q) = 0$, then there exists a vector in the kernel of $\mathrm{Syl}_x(p,q)$ of the form shown in the equation above, and thus a common root x_0.

- **Root products and companion matrices**: Let α_j, β_k be the roots of $p(x)$ and $q(x)$, respectively. By construction, the expression

$$\prod_{j=1}^{n}\prod_{k=1}^{m}(\alpha_j - \beta_k)$$

vanishes if and only if there exists a root of p that is equal to a root of q. Although the computation of this product seems to require explicit access to the roots, this can be avoided. Multiplying by a convenient normalization factor, we have:

$$\begin{aligned} p_n^m q_m^n \prod_{j=1}^{n}\prod_{k=1}^{m}(\alpha_j - \beta_k) &= p_n^m \prod_{j=1}^{n} q(\alpha_j) = p_n^m \det q(\mathcal{C}_p) \\ &= (-1)^{nm} q_m^n \prod_{k=1}^{m} p(\beta_k) = (-1)^{nm} q_m^n \det p(\mathcal{C}_q) \end{aligned} \tag{1}$$

- **Kronecker products**: Using a well-known connection to Kronecker products, we can also write (1) as

$$p_n^m q_m^n \det(\mathcal{C}_p \otimes I_m - I_n \otimes \mathcal{C}_q).$$

- **Bézout matrix:** Given $p(x)$ and $q(x)$ as before, consider the bivariate function

$$B(s,t) := \frac{p(s)q(t) - p(t)q(s)}{s-t}.$$

It is easy to see that this is actually a polynomial in the variables s, t, and is invariant under the interchange $s \leftrightarrow t$. Let $d := \max(n, m)$, and $\mathrm{Bez}_x(p, q)$ be the symmetric $d \times d$ matrix that represents this polynomial in the standard monomial basis, i.e.,

$$B(s,t) = \begin{bmatrix} 1 \\ s \\ \vdots \\ s^{d-1} \end{bmatrix}^T \mathrm{Bez}_x(p,q) \begin{bmatrix} 1 \\ t \\ \vdots \\ t^{d-1} \end{bmatrix}.$$

The Bézout matrix is singular if and only p and q have a common root.

Notice the differences with the Sylvester matrix: while that approach requires a non-symmetric $(n+m) \times (n+m)$ matrix depending linearly on the coefficients, in the Bézout approach the matrix is smaller and symmetric, but with entries that depend bilinearly on the p_i, q_i.

If can be shown that all these constructions are equivalent. They define exactly the same polynomial, called the *resultant* of p and q, denoted as $\mathrm{Res}_x(p, q)$:

$$\begin{aligned} \mathrm{Res}_x(p,q) &= \det \mathrm{Syl}_x(p,q) \\ &= p_n^m \det q(\mathcal{C}_p) \\ &= (-1)^{nm} q_m^n \det p(\mathcal{C}_q) \\ &= p_n^m q_m^n \det(\mathcal{C}_p \otimes I_m - I_n \otimes \mathcal{C}_q) \\ &= \frac{(-1)^{\binom{n}{2}}}{p_n^{n-m}} \det \mathrm{Bez}_x(p,q). \end{aligned}$$

The resultant is a homogeneous multivariate polynomial, with integer coefficients, and of degree $n+m$ in the $n+m+2$ variables p_j, q_k. It vanishes if and only if the polynomials p and q have a common root. Notice that the definition is not symmetric in its two arguments, $\mathrm{Res}_x(p,q) = (-1)^{nm}\mathrm{Res}(q,p)$ (of course, this does not matter in checking whether it is zero).

Remark 1. *To compute the resultant of two polynomials $p(x)$ and $q(x)$ in Maple, you can use the command* `resultant(p,q,x)`. *In Mathematica, use instead* `Resultant[p,q,x]`.

2 Discriminants

As we have seen, the resultant allows us to write an easily checkable condition for the simultaneous vanishing of two univariate polynomials. Can we use the resultant to produce a condition for a polynomial to have a double root? Recall that if a polynomial $p(x)$ has a double root at x_0 (which can be real or complex), then its derivative $p'(x)$ also vanishes at x_0. Thus, we can check for the existence of a root of multiplicity two (or higher) by computing the resultant betweeen a polynomial and its derivative.

Definition 2. *The* discriminant *of a univariate polynomial $p(x)$ is defined as*

$$\mathrm{Dis}_x(p) := (-1)^{\binom{n}{2}} \frac{1}{p_n} \mathrm{Res}_x\left(p(x), \frac{dp(x)}{dx}\right).$$

Similarly to what we did in the resultant case, the discriminant can also be obtained by writing a natural condition in terms of the roots α_i of $p(x)$:

$$\mathrm{Dis}_x(p) = p_n^{2n-2} \prod_{j<k} (\alpha_j - \alpha_k)^2.$$

If $p(x)$ has degree n, its discriminant is a homogeneous polynomial of degree $2n-2$ in its $n+1$ coefficients $p_n, \ldots, p_0$.

Example 3. *Consider the quadratic univariate polynomial $p(x) = ax^2 + bx + c$. Its discriminant is:*

$$\mathrm{Dis}_x(p) = -\frac{1}{a}\mathrm{Res}_x(ax^2+bx+c, 2ax+b) = b^2 - 4ac.$$

For the cubic polynomial $p(x) = ax^3 + bx^2 + cx + d$ we have

$$\mathrm{Dis}_x(p) = -27a^2d^2 + 18adcb + b^2c^2 - 4b^3d - 4ac^3.$$

3 Applications

3.1 Polynomial equations

One of the most natural applications of resultants is in the solution of polynomial equations in two variables. For this, consider a polynomial system

$$p(x,y) = 0, \qquad q(x,y) = 0, \tag{2}$$

with only a finite number of solutions (which is generically the case). Consider a fixed value of y_0, and the two univariate polynomials $p(x,y_0), q(x,y_0)$. If y_0 corresponds to the y-component of

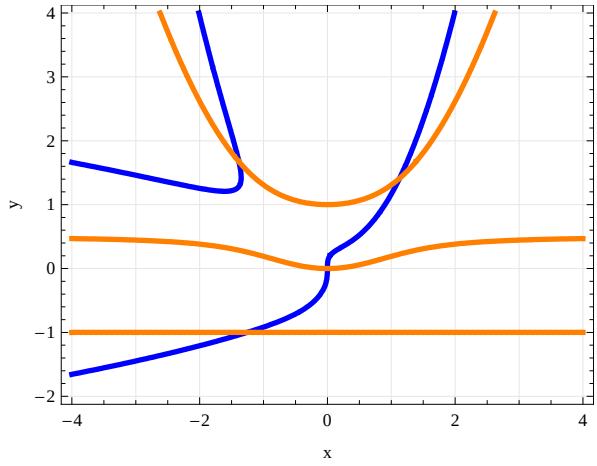

Figure 1: Zero sets of the polynomials $p(x, y)$ and $q(x, y)$ in Example 4.

a root, then these two univariate polynomials clearly have a common root, hence their resultant vanishes.

Therefore, to solve (2), we can compute $\mathrm{Res}_x(p, q)$, which is a univariate polynomial in y. Solving this univariate polynomial, we obtain a finite number of points y_i. Backsubstituting in p (or q), we obtain the corresponding values of x_i. Naturally, the same construction can be used by computing first the univariate polynomial in x given by $\mathrm{Res}_y(p, q)$.

Example 4. *Let* $p(x, y) = 2xy + 3y^3 - 2x^3 - x - 3x^2y^2$, *and* $q(x, y) = 2x^2y^2 - 4y^3 - x^2 + 4y + x^2y$. *The corresponding zero sets are shown in Figure 1. The resultant (with respect to the* x *variable) is*

$$\mathrm{Res}_x(p, q) = y(y + 1)^3(72y^8 - 252y^7 + 270y^6 - 145y^5 + 192y^4 - 160y^3 + 28y + 4).$$

One particular root of this polynomial is $y_\star \approx 1.6727$, *with the corresponding value of* $x_\star \approx -1.3853$.

3.2 Implicitization of plane rational curves

Consider a plane curve parametrized by rational functions, i.e.,

$$x(t) = \frac{p_1(t)}{q_1(t)}, \qquad y(t) = \frac{p_2(t)}{q_2(t)}.$$

What is the implicit equation of the curve, i.e., what constraint $h(x, y) = 0$ must the points $(x, y) \in \mathbb{R}^2$ that lie on the curve satisfy? The corresponding equation can be easily obtained by computing a resultant to eliminate the parametrizing variable t, i.e.,

$$h(x, y) = \mathrm{Res}_t(q_1(t) \cdot x - p_1(t), q_2(t) \cdot y - p_2(t)).$$

Example 5. *Consider the curve described by the parametrization.*

$$x(t) = \frac{t(1 + t^2)}{1 + t^4}, \qquad y(t) = \frac{t(1 - t^2)}{1 + t^4}. \tag{3}$$

Its implicit equation can be computed by the resultant:

$$\mathrm{Res}_t((1 + t^4)x - t(1 + t^2), (1 + t^4)y - t(1 - t^2)) = 4y^4 + 8y^2x^2 + 4x^4 + 4y^2 - 4x^2.$$

Remark 6. *The inverse problem (given an implicit polynomial equation for a curve, find a rational parametrization) is not always solvable. In fact, there is a full characterization of when this is possible, in terms of a topological invariant of the curve called the* genus *(the rationally parametrizable curves are exactly those of genus zero). For example, the curve* $x^4 + y^4 = 1$ *does* not *admit a polynomial parametrization, since it has genus 3.*

3.3 Eigenvalue distribution of random matrices

This section is based on the results in [RE08]. The eigenvalues of a random symmetric matrix belonging to a given ensemble can be characterized in terms of the asymptotic eigenvalue distribution $F(x)$ (e.g., the semi-circle law, Marčenko-Pastur, etc). Often, rather than the actual distribution, it is more convenient to use instead some other equivalent object, such as its moment generating function, Stieltjes transform, R-transform, etc. For many ensembles of interest, these auxiliary transforms $\tilde{F}(z)$ are algebraic functions, in the sense that they satisfy an equation of the form $\psi(\tilde{F}(z), z) = 0$, where $\psi(s,t)$ is a bivariate polynomial, and furthermore they can all be derived from each other. As a consequence, to each given random ensemble of this class we can associate a bivariate polynomial that uniquely describes the limiting eigenvalue distribution.

A natural question arises: given two matrices M_1, M_2, belonging to random ensembles with associated polynomials $\psi_1(s,t)$ and $\psi_2(s,t)$, what can be said about the eigenvalue distribution of the sum $M_1 + M_2$ (or the product $M_1 M_2$)? Voiculescu has shown that under a certain natural independence condition ("freeness"), the R-transform of the sum is the sum of the individual transforms (this is somewhat akin to the well-known fact that the pdf of the sum of independent random variables is the convolution of the individual pdfs, or the additivity of the moment generating function). Under the freeness condition, the bivariate polynomial associated with the ensemble $M_3 = M_1 + M_2$ can be computed from the individual polynomials ψ_1, ψ_2 via:

$$\psi_3(s,t) = \mathrm{Res}_u(\psi_1(s-u,t), \psi_2(u,t)).$$

Similar expressions are also possible for the product $M_1 M_2$, also in terms of resultants. This allows the computation of the spectra of arbitrary random ensembles, that can be built from individual "building blocks" with known eigenvalue distributions.

We cannot provide a full description here of this area, and the very interesting connections with "free probability." We refer the reader to [RE08] for a more complete account.

4 The set of nonnegative polynomials

One of the main reasons why nonnegativity conditions about polynomials are difficult is because these sets can have a quite complicated structure, even though they are always convex.

Recall from last lecture that we have defined $P_n \subset \mathbb{R}^{n+1}$ as the set of nonnegative polynomials of degree n. It is easy to see that if $p(x)$ lies on the boundary of the set P_n, then it must have a real root, of multiplicity at least two. Indeed, if there is no real root, then $p(x)$ is in the strict interior of P (small enough perturbations will not create a root), and if it has a simple real root it clearly cannot be nonnegative.

Thus, on the boundary of P_n, the discriminant of $p(x)$ must necessarily vanish. However, it turns out that $\mathrm{Dis}_x(p)$ does not vanish *only* on the boundary, but it also vanishes at points inside the set. Why is this?

Example 7. *Consider the univariate polynomial* $p(x) = x^4 + 2ax^2 + b$*. For what values of* a, b *does it hold that* $p(x) \geq 0 \, \forall x \in \mathbb{R}$*? Since the leading term* x^4 *has even degree and is strictly positive,*

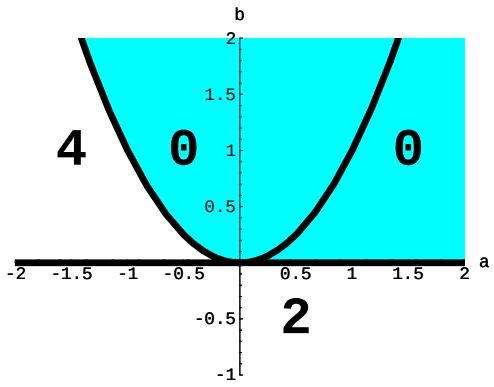

Figure 2: The shaded region corresponds to the values of (a, b) for which the polynomial $x^4 + 2ax^2 + b$ is nonnegative. The numbers indicate how many real roots the polynomial has.

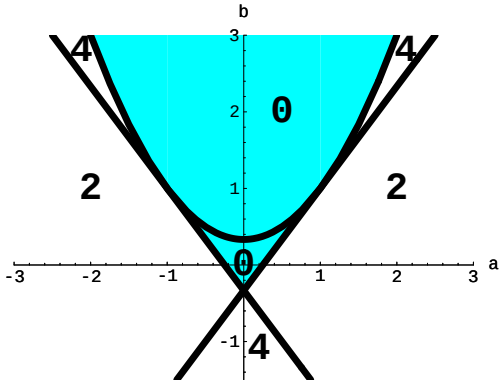

Figure 3: Region of nonnegativity of the polynomial $x^4 + 4ax^3 + 6bx^2 + 4ax + 1$, and number of real roots.

$p(x)$ is strictly positive if and only if it has no real roots. The discriminant of $p(x)$ is equal to $256\, b\, (a^2 - b)^2$. The set of (a, b) for which $p(x)$ is nonnegative is shown in Figure 2.

Here is a slightly different example, showing the same phenomenon.

Example 8. *Consider now the polynomial $p(x) = x^4 + 4ax^3 + 6bx^2 + 4ax + 1$. Its discriminant, in factored form, is equal to $256(1 + 3b + 4a)(1 + 3b - 4a)(1 + 2a^2 - 3b)^2$. The corresponding nonnegativity region and number of real roots are presented in Figure 3.*

As we can see, this creates some difficulties. For instance, even though we have a perfectly valid analytic expression for the boundary of the set, we cannot get a good sense of "how far we are" from the boundary by looking at the absolute value of the discriminant.

From the mathematical viewpoint, there are a couple of (unrelated?) reasons with these sets cannot be directly handled by "standard" optimization, at least if we want to keep the polynomial structure. One has to do with its algebraic structure, and the other one with convexity, and in particular its facial structure.

Lemma 9 (e.g., [And03])**.** *The set described in Figure 2 is not basic closed semialgebraic.*

Remark 10. *Notice that the convex sets described in Figures 2 and 3 both have an uncommon feature. They both have proper faces that are not* exposed*, i.e., they cannot be isolated by a supporting*

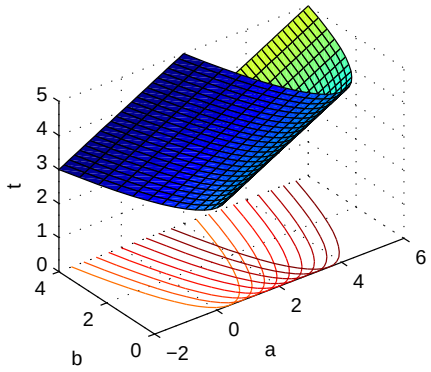

Figure 4: A three-dimensional convex set, described by one quadratic and one linear inequality, whose projection on the (a, b) plane is equal to the set in Figure 2.

hyperplane[1]*. Indeed, in Figure 2 the origin* $(0, 0)$ *is a non-exposed zero-dimensional face, while in Figure 3 the point* $(1, 1)$ *has the same property. A non-exposed face is a known obstruction for a convex set to be the feasible set of a semidefinite program, see [RG95].*

Even though these sets have these complicating features, it turns out that we can often provide some "good" representations. These are normally given as a *projection* from higher dimensional spaces, where the object "upstairs" is much more smooth and well-behaved. For instance, as a graphical illustration, in Figure 4 we can see the three-dimensional convex set $\{(a, b, t) \in \mathbb{R}^3 : b \geq (a - t)^2,\ t \geq 0\}$, whose projection on the plane (a, b) is exactly the set discussed in Example 7 and Figure 2.

The presence of "extraneous" components of the discriminant inside the feasible set is an important roadblock for the availability of "easily computable" barrier functions. Indeed, every polynomial that vanishes on the boundary of the set P_n must necessarily have the discriminant as a factor. This is an striking difference with the case of the case of the nonnegative orthant or the PSD cone, where the standard barriers are given (up to a logarithm) by products of the linear constraints or a determinant (which are polynomials). The way out of this problem is to produce non-polynomial barrier functions, either by partial minimization from a higher-dimensional barrier (i.e., projection) or other constructions such as the "universal" barrier function introduced by Nesterov and Nemirovski [NN94].

In this direction, there have been several research efforts that aim at directly characterizing barrier functions for the set of nonnegative polynomials (or related modifications). Among them, we mention the work of Kao and Megretski [KM02] and Faybusovich [Fay02], both of which produce barriers that rely on the computation of one or more integral expressions. Given the fact that these integrals must be computed numerically, there is no clear consensus yet on how useful this approach is in practical optimization problems.

References

[And03] C. Andradas. Characterization and description of basic semialgebraic sets. In *Algorithmic and quantitative real algebraic geometry (Piscataway, NJ, 2001)*, volume 60 of *DIMACS*

[1] A *face* of a convex set S is a convex subset $F \subseteq S$, with the property that $x, y \in S, \frac{1}{2}(x + y) \in F \Rightarrow x, y \in F$. A face F is *exposed* if it can be written as $F = S \cap H$, where H is a supporting hyperplane of S.

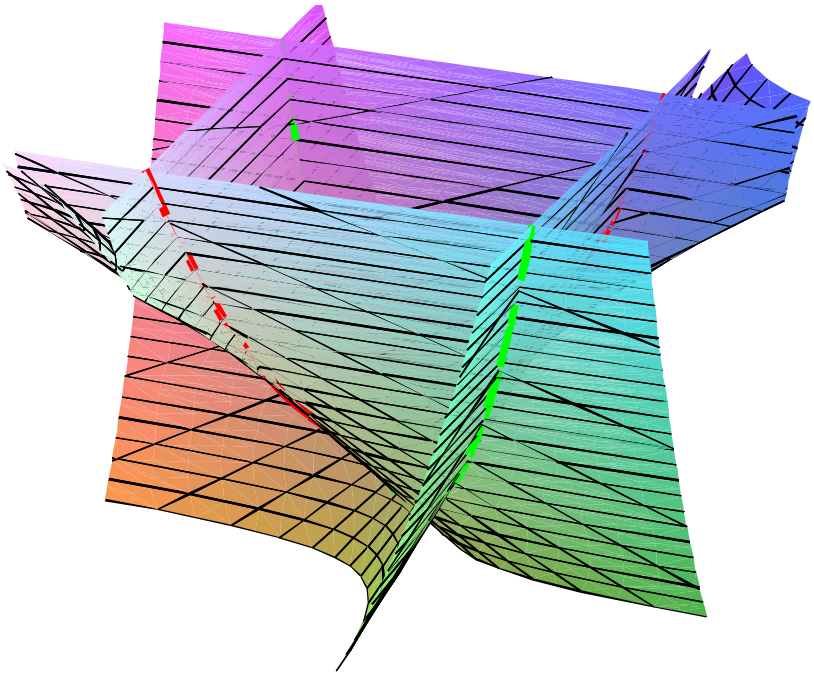

Figure 5: The discriminant of the polynomial $x^4 + 4ax^3 + 6bx^2 + 4cx + 1$. The convex set inside the "bowl" corresponds to the region of nonnegativity. There is an additional one-dimensional component inside the set.

Ser. Discrete Math. Theoret. Comput. Sci., pages 1–12. Amer. Math. Soc., Providence, RI, 2003.

[CLO97] D. A. Cox, J. B. Little, and D. O'Shea. *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer, 1997.

[Fay02] L. Faybusovich. Self-concordant barriers for cones generated by Chebyshev systems. *SIAM J. Optim.*, 12(3):770–781, 2002.

[GKZ94] I. M. Gel'fand, M. Kapranov, and A. Zelevinsky. *Discriminants, Resultants, and Multidimensional Determinants*. Birkhäuser, 1994.

[KM02] C. Y. Kao and A. Megretski. A new barrier function for IQC optimization problems. In *American Control Conference*, 2002.

[NN94] Y. E. Nesterov and A. Nemirovski. *Interior point polynomial methods in convex programming*, volume 13 of *Studies in Applied Mathematics*. SIAM, Philadelphia, PA, 1994.

[RE08] N. Raj Rao and Alan Edelman. The polynomial method for random matrices. *Found. Comput. Math.*, 8(6):649–702, 2008.

[RG95] M. Ramana and A. J. Goldman. Some geometric results in semidefinite programming. *J. Global Optim.*, 7(1):33–50, 1995.

[Stu98] B. Sturmfels. Introduction to resultants. In *Applications of computational algebraic geometry (San Diego, CA, 1997)*, volume 53 of *Proc. Sympos. Appl. Math.*, pages 25–39. Amer. Math. Soc., Providence, RI, 1998.

MIT 6.7230 - Algebraic techniques and semidefinite optimization March 6, 2024

Lecture 7

Lecturer: Pablo A. Parrilo *Scribe: ???*

In this lecture we introduce a special class of multivariate polynomials, called *hyperbolic*. These polynomials were originally studied in the context of partial differential equations. As we will see, they have many surprising properties, and are intimately linked with convex optimization problems that have an algebraic structure. A few good references about the use of hyperbolic polynomials in optimization are [Gül97, BGLS01, Ren06].

1 Hyperbolic polynomials

Consider a homogeneous multivariate polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]$ of degree d. Here *homogeneous of degree d* means that the sum of degrees of each monomial is constant and equal to d, i.e.,

$$p(x) = \sum_{|\alpha|=d} c_\alpha x^\alpha,$$

where $\alpha = (\alpha_1, \ldots, \alpha_n), \alpha_i \in \mathbb{N} \cup \{0\}$, and $|\alpha| = \alpha_1 + \cdots + \alpha_n$. A homogeneous polynomial satisfies $p(tw) = t^d p(w)$ for all real t and vectors $w \in \mathbb{R}^n$. We denote the set of such polynomials by $\mathcal{H}_n(d)$. By identifying a polynomial with its vector of coefficients, we can consider $\mathcal{H}_n(d)$ as a vector space of dimension $\binom{n+d-1}{d}$.

Definition 1. *Let e be a fixed vector in $\mathbb{R}^n$. A polynomial $p \in \mathcal{H}_n(d)$ is* hyperbolic with respect to e *if $p(e) > 0$ and, for all vectors $x \in \mathbb{R}^n$, the univariate polynomial $t \mapsto p(x - te)$ has only real roots.*

A natural geometric interpretation is the following: consider the hypersurface in $\mathbb{R}^n$ given by $p(x) = 0$. Then, hyperbolicity is equivalent to the condition that every line in $\mathbb{R}^n$ parallel to the direction e intersects this hypersurface at exactly d points (counting multiplicities), where d is the degree of the polynomial.

Example 2. *The polynomial $x_1 x_2 \cdots x_n$ is hyperbolic with respect to the vector $e = (1, 1, \ldots, 1)$, since the univariate polynomial $t \mapsto (x_1 - t)(x_2 - t) \cdots (x_n - t)$ has roots $x_1, x_2, \ldots, x_n$.*

Hyperbolic polynomials enjoy a very surprising property, that connects in an unexpected way algebra with convex analysis. Given a hyperbolic polynomial $p(x)$, consider the set defined as:

$$\Lambda_{++} := \{x \in \mathbb{R}^n : p(x - te) = 0 \;\Rightarrow\; t > 0\}.$$

Geometrically, this condition says that if we start at the point $x \in \mathbb{R}^n$, and slide along a line in the direction parallel to e, then we will never encounter the hypersurface $p(x) = 0$, while if we move in the opposite direction, we will cross it exactly d times. Figure 1 illustrates a particular hyperbolicity cone.

It is immediate from homogeneity and the definition above that $\lambda > 0$, $x \in \Lambda_{++} \Rightarrow \lambda x \in \Lambda_{++}$. Thus, we call Λ_{++} the *hyperbolicity cone* associated to p, and denote its closure by Λ_+. As we will see shortly, it turns out that these cones are actually *convex cones*. We prove this following the arguments in Renegar [Ren06]; the original results are due to Gårding [Går59].

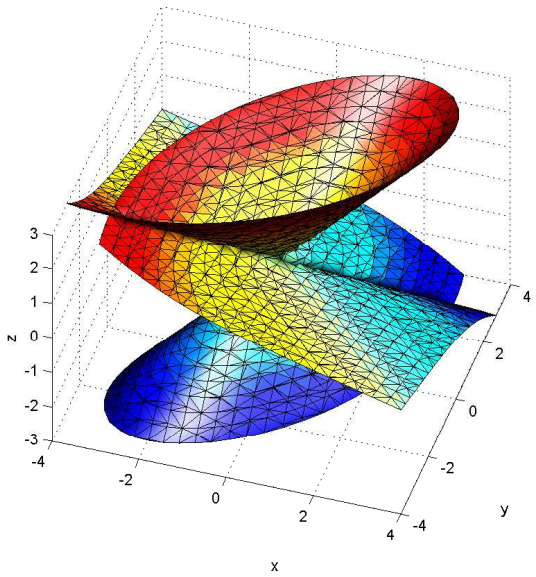

Figure 1: Hyperbolicity cone corresponding to the polynomial $p(x, y, z) = 4xyz + xz^2 + yz^2 + 2z^3 - x^3 - 3zx^2 - y^3 - 3zy^2$. This polynomial is hyperbolic with respect to $(0, 0, 1)$.

Lemma 3. *The hyperbolicity cone Λ_{++} is the connected component of $p(x) > 0$ that includes e.*

Example 4. *The hyperbolicity cone Λ_{++} associated with the polynomial $x_1 x_2 \cdots x_n$ discussed in Example 2 is the open positive orthant $\{x \in \mathbb{R}^n \,|\, x_i > 0\}$.*

The first step is to show that we can replace e with any vector in the hyperbolicity cone.

Lemma 5. *If $p(x)$ is hyperbolic with respect to e, then it is also hyperbolic with respect to every direction $v \in \Lambda_{++}$. Furthermore, the hyperbolicity cones are the same.*

Proof. By Lemma 3 we have $p(v) > 0$. We need to show that for every $x \in \mathbb{R}^n$, the polynomial $\beta \mapsto p(\beta v + x)$ has only real roots if $v \in \Lambda_{++}$.

Let $\alpha > 0$ be fixed, and consider the polynomial $\beta \mapsto p(\alpha ie + \beta v + \gamma x)$, where i is the imaginary unit. We claim that if $\gamma \geq 0$, this polynomial has only roots in the lower half-plane. Let's look at the $\gamma = 0$ case first. It is clear that $\beta \mapsto p(\alpha ie + \beta v)$ cannot have a root at $\beta = 0$, since $p(\alpha ie) = (\alpha i)^d p(e) \neq 0$. If $\beta \neq 0$, we can write

$$p(\alpha ie + \beta v) = 0 \quad \Leftrightarrow \quad p(\alpha\beta^{-1} ie + v) = 0 \quad \Rightarrow \quad \alpha\beta^{-1} i < 0 \quad \Rightarrow \quad \beta \in i\mathbb{R}_-,$$

and thus the roots of this polynomial are on the strict negative imaginary axis (we have used $v \in \Lambda_{++}$ in the second implication). If by increasing γ there is ever a root in the upper half-plane, then there must exist a $\gamma_\star$ for which $\beta \mapsto p(\alpha ie + \beta v + \gamma_\star x)$ has a real root $\beta_\star$, and thus $p(\alpha ie + \beta_\star v + \gamma_\star x) = 0$. However, this contradicts hyperbolicity, since $\beta_\star v + \gamma_\star x \in \mathbb{R}^n$. Thus, for all $\gamma \geq 0$, the roots of $\beta \mapsto p(\alpha ie + \beta v + \gamma x)$ are in the lower half-plane.

The conclusion above was true for any $\alpha > 0$. Letting $\alpha \to 0$, by continuity of the roots we have that the polynomial $\beta \mapsto p(\beta v + \gamma x)$ must also have its roots in the lower closed half-plane. However, since it is a polynomial with real coefficients (and therefore its roots always appear in complex-conjugate pairs), then all the roots must actually be real. Taking now $\gamma = 1$, we have that $\beta \mapsto p(\beta v + x)$ has real roots for all x, or equivalently, p is hyperbolic in the direction v. □

The following result shows that this set is actually convex:

Theorem 6 ([Går59])**.** *The hyperbolicity cone Λ_{++} is convex.*

Proof. We want to show that $u, v \in \Lambda_{++}$, $\beta, \gamma > 0$ implies that $\beta u + \gamma v \in \Lambda_{++}$. The previous result implies that it is enough to show hyperbolicity of p with respect to v (instead of e), i.e., to analyze the polynomial $t \mapsto p(x - tv)$. Notice that the roots of $t \mapsto p(\beta u + \gamma v - tv)$ are just a nonnegative affine scaling of the roots of $t \mapsto p(u - tv)$, since

$$p(u - t_\star v) = 0 \qquad \Leftrightarrow \qquad p(\beta u + \gamma v - (\beta t_\star + \gamma)v) = 0,$$

and $u \in \Lambda_{++}$ this implies that $t_\star > 0$, hence $\beta t_\star + \gamma > 0$. As a consequence, $\beta u + \gamma v \in \Lambda_{++}$. □

Hyperbolic polynomials are of interest in convex optimization, because they unify in a quite appealing way many facts about the most important tractable classes: linear, second order, and semidefinite programming.

Example 7 (SOCP)**.** *Let $p(x) = x_{n+1}^2 - \sum_{k=1}^n x_k^2$. This is a homogeneous quadratic polynomial, hyperbolic in the direction $e = (0, \ldots, 0, 1)$, since*

$$p(x - te) = (x_{n+1} - t)^2 - \sum_{k=1}^n x_k^2 = t^2 - 2tx_{n+1} + \left(x_{n+1}^2 - \sum_{k=1}^n x_k^2 \right),$$

and the discriminant of this quadratic equation is equal to

$$4x_{n+1}^2 - 4\left(x_{n+1}^2 - \sum_{k=1}^n x_k^2 \right) = 4 \sum_{k=1}^n x_k^2,$$

which is always nonnegative, so the polynomial $t \mapsto p(x - te)$ has only real roots. The corresponding hyperbolicity cone is the Lorentz or second order cone given by

$$\Lambda_+ = \left\{ x \in \mathbb{R}^{n+1} \mid x_{n+1} \geq 0, \quad \sum_{k=1}^n x_k^2 \leq x_{n+1}^2 \right\}.$$

Example 8 (SDP)**.** *Consider the homogeneous polynomial*

$$p(x) = \det(x_1 A_1 + \cdots + x_n A_n),$$

where $A_i \in \mathcal{S}^d$ are given symmetric matrices, with $A_1 \succ 0$. The polynomial $p(x)$ is homogeneous of degree d. Letting $e = (1, 0, \ldots, 0)$, we have

$$p(x - te) = \det\left(\sum_{k=1}^n x_k A_k - tA_1 \right) = \det A_1 \cdot \det\left(\sum_{k=1}^n x_k A_1^{-\frac{1}{2}} A_k A_1^{-\frac{1}{2}} - tI \right),$$

and as a consequence the roots of $p(x - te)$ are always real since they are the eigenvalues of a symmetric matrix. Thus, $p(x)$ is hyperbolic with respect to e. The corresponding hyperbolicity cone is

$$\Lambda_{++} = \{ x \in \mathbb{R}_n \mid x_1 A_1 + \cdots + x_n A_n \succ 0 \}.$$

Thus, by Lemma 5, $p(x)$ is hyperbolic with respect to every $x \in \Lambda_{++}$.

Based on the results discussed earlier regarding the number of real roots of a univariate polynomial, we have the following lemma.

Lemma 9. *The polynomial $p(x)$ is hyperbolic with respect to e if and only if the Hermite matrix $H_1(p) \in \mathcal{S}^n[x]$ is positive semidefinite for all $x \in \mathbb{R}^n$.*

As we will see later in the course, this observation will allow us to give an exact characterization in terms of semidefinite programming of the hyperbolicity of trivariate polynomials [Par].

Lemma 10. *The hyperbolicity cone Λ_+ is basic closed semialgebraic, i.e., it can be described by unquantified polynomial inequalities.*

The two following results are of importance in optimization and the formulation of interior-point methods.

Theorem 11 ([Ren06])**.** *A hyperbolic cone Λ_+ is facially exposed.*

Theorem 12 ([Gül97])**.** *The function $-\log p(x)$ is a logarithmically homogeneous self-concordant barrier*[1] *for the hyperbolicity cone Λ_{++}, with barrier parameter equal to d.*

One of the main open issues regarding hyperbolic cones is about their generality. As Example 8 shows, the cone associated with a semidefinite program is a hyperbolic cone. An open question (known as the generalized Lax conjecture) is whether the converse holds, more specifically, whether every hyperbolic cone is a "slice" of the semidefinite cone, i.e., it can be represented as the intersection of an affine subspace and $\mathcal{S}^n_+$. As we will see in the next lecture, a few special cases of the conjecture have been settled recently.

In recent years, there have been an increasing number of appearances of hyperbolic polynomials in challenging questions in combinatorics and optimization. Among them, we mention the relationships with matroid theory explored in [COSW04], Gurvits' slick proof of Van der Waerden conjecture on permanents [Gur06] and the recent proof of Weaver's reformulation of the Kadison-Singer problem by Marcus-Spielman-Srivastava [MSS15].

2 SDP representability

Recall that in the previous lecture, we encountered a class of convex sets in $\mathbb{R}^2$ that lacked certain desirable properties (namely, being basic semialgebraic, and facially exposed). As we will see, hyperbolic polynomials will play a fundamental role in the characterization of the properties a set in $\mathbb{R}^2$ must satisfy for it to be the feasible set of a semidefinite program.

References

[BGLS01] H. H. Bauschke, O. Güler, A. S. Lewis, and H. S. Sendov. Hyperbolic polynomials and convex analysis. *Canad. J. Math.*, 53(3):470–488, 2001.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[1] A function $f : \mathbb{R} \to \mathbb{R}$ is *self-concordant* if it satisfies $f''(x) \geq |\frac{1}{2} f'''(x)|^{\frac{2}{3}}$. A function $f : \mathbb{R}^n \to \mathbb{R}$ is self-concordant if the univariate function obtained when restricting to any line is self-concordant. Self-concordance implies convexity, and is a crucial property in the analysis of the polynomial-time global convergence of Newton's method; see [NN94] or [BV04, Section 9.6] for more details.

[COSW04] Y.B. Choe, J.G. Oxley, A.D. Sokal, and D.G. Wagner. Homogeneous multivariate polynomials with the half-plane property. *Advances in Applied Mathematics*, 32(1-2):88–187, 2004.

[Går59] L. Gårding. An inequality for hyperbolic polynomials. *J. Math. Mech.*, 8:957–965, 1959.

[Gül97] O. Güler. Hyperbolic polynomials and interior point methods for convex programming. *Math. Oper. Res.*, 22(2):350–377, 1997.

[Gur06] L. Gurvits. Hyperbolic polynomials approach to Van der Waerden/Schrijver-Valiant like conjectures: sharper bounds, simpler proofs and algorithmic applications. In *STOC'06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 417–426, New York, 2006. ACM.

[MSS15] A. Marcus, D. A. Spielman, and N. Srivastava. Interlacing families II: Mixed characteristic polynomials and the Kadison-Singer problem. *Annals of Mathematics*, pages 327–350, 2015.

[NN94] Y. E. Nesterov and A. Nemirovski. *Interior point polynomial methods in convex programming*, volume 13 of *Studies in Applied Mathematics*. SIAM, Philadelphia, PA, 1994.

[Par] P. A. Parrilo. Hyperbolic polynomials and SOS matrices. Manuscript in preparation, 2007.

[Ren06] J. Renegar. Hyperbolic programs, and their derivative relaxations. *Found. Comput. Math.*, 6(1):59–79, 2006.

MIT 6.7230 - Algebraic techniques and semidefinite optimization March 6, 2024

Lecture 8

Lecturer: Pablo A. Parrilo *Scribe: ???*

1 SDP representability

A few lectures ago, when discussing the set of nonnegative polynomials, we encountered convex sets in $\mathbb{R}^2$ that lacked certain desirable properties (namely, being basic semialgebraic, and facially exposed). As we will see, hyperbolic polynomials will play a fundamental role in the characterization of the properties a set in $\mathbb{R}^2$ must satisfy for it to be the feasible set of a semidefinite program.

2 Convex sets in $\mathbb{R}^2$

In this lecture we will study conditions that a set $S \subset \mathbb{R}^2$ must satisfy for it to be *semidefinite representable*, i.e., to admit a characterization of the type

$$\{(x, y) \in \mathbb{R}^2 \mid I + xB + yC \succeq 0\}, \tag{1}$$

where $B, C \in \mathcal{S}^d$. Notice that we have assumed (without loss of generality) that $0 \in \operatorname{int} S$, and normalized the first matrix in the matrix pencil to be an identity matrix (this can always be achieved by left- and right-multiplying by an appropriate factor).

Remark 1. *We should not confuse the notion of semidefinite representability described above, with the much more general* lifted SDP representability*, that allows the representation of the original set as a* projection *of a higher-dimensional SDP set. In other words, here we are not allowed to use additional variables.*

Clearly, from (1), we have the following necessary conditions for SDP representability:

- **Closed:** Every set of the form (1) is closed, in the standard topology.
- **Convex:** Every set of the form (1) is necessarily convex, since it is (the projection of) the intersection of an affine subspace and the convex set of PSD matrices. Of course, this is also easy to prove directly.
- **Basic semialgebraic:** As we have discussed, the boundary of the set (1) is defined by d unquantified polynomial inequalities of degree at most equal to d. In fact, the interior of this set exactly corresponds to the connected component of $\det(I + xB + yC) > 0$ that contains the origin.

There is a less obvious additional condition, which we have also seen already:

- **Exposed faces:** Every convex set of the form (1) has proper faces that are *exposed*. In other words, every face F must have a representation as $F = S \cap H$, where H is a supporting hyperplane of the convex set S.

A natural question, then, is the following: are the conditions listed above *sufficient* for SDP representability? If a set $S \subset \mathbb{R}^2$ satisfies these four conditions, do there always exist matrices B, C, for which the set (1) is exactly equal to S? To ask a concrete question: does the set in Figure 1 admit an SDP representation? Before settling this issue, let us discuss first an apparently different question, involving hyperbolic polynomials.

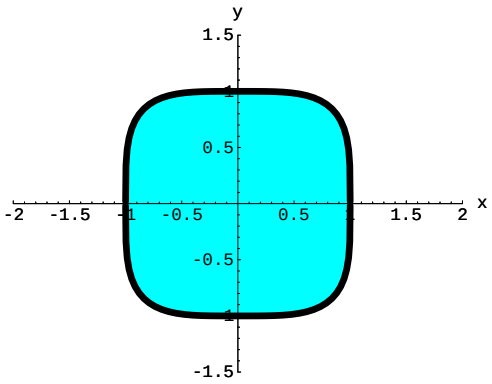

Figure 1: Convex set defined by $x^4 + y^4 \leq 1$.

3 Hyperbolicity and the Lax conjecture

Recall from the previous lecture that a hyperbolic polynomial is a homogeneous polynomial $p(x)$ of degree d, with the property that when restricted to lines parallel to a particular direction e, the resulting univariate polynomial has all its d roots real.

Furthermore, we have also seen that every polynomial of the form

$$p(x) = \det(x_1 A_1 + \cdots + x_n A_n), \tag{2}$$

where $A_i \in \mathcal{S}^d$ and $A_1 \succ 0$, is hyperbolic with respect to the $(1, 0, \ldots, 0)$ direction.

A 1958 conjecture by Peter Lax [Lax58], asks whether the converse is true in the case $n = 3$ (i.e., trivariate polynomials). In other words, is it true that for every hyperbolic polynomial $p(x)$ in three variables of degree d, there exist three symmetric matrices $\{A_1, A_2, A_3\} \subset \mathcal{S}^d$ for which (2) holds?

As a first step towards answering this question, let us verify that this at least makes sense in terms of dimension counting. As we have seen, the dimension of the set of hyperbolic polynomials in three variables ($n = 3$) and degree d is equal to $\binom{n+d-1}{d} = \binom{d+2}{2}$. On the other hand, for a polynomial of the form (2), by an appropriate similarity transform we can always assume without loss of generality $A_1 = a_0 I_d$, and $A_2 = \operatorname{diag}(a_1, \ldots, a_d)$. The total number of parameters is then $1 + d + \binom{d+1}{2}$, which is exactly equal to $\binom{d+2}{2}$. Of course, this by itself does not prove the result, but it shows that it is certainly possible.

4 Relating SDP-representable sets and hyperbolic polynomials

As we will see shortly, these two apparently different problems are in fact one and the same. Before showing this, let us consider one additional necessary condition for a set in $\mathbb{R}^2$ to be SDP-representable. For later reference, we first define the following notion:

Definition 2. *A polynomial $p \in \mathbb{R}[x]$ is a* real zero polynomial *if for every $x \in \mathbb{R}^n$, $p(tx) = 0$ implies that t is real.*

Recall that the boundary of a set described by (1) is determined by the zero set of the polynomial $\det(I + xB + yC)$. Consider now any line passing through the origin, i.e., of the form $(x, y) = (\beta t, \gamma t)$. We have then

$$\det[I + (\beta B + \gamma C)t] = 0,$$

and this univariate polynomial in t has exactly d real roots (namely, the negative inverse of the eigenvalues of $\beta B + \gamma C$). In terms of the notation just introduced, the polynomial defined by $\det(I + xB + yC)$ is a real zero polynomial. Equivalently, for every set of the form (1), it is always the case that every line through the origin intersects (the Zariski closure[1] of) the boundary of the set exactly d times.

In the preceding, our starting point was directly a determinantal representation as in (1). It can be shown (see [HV07]) that if we start directly from a given set that admits an SDP representation, we can precisely characterize a unique minimal polynomial that defines the boundary of the set.

Hence, this gives us an additional necessary condition ([HV07]) for SDP representability:

- **Rigid convexity:** Consider a set in $\mathbb{R}^2$, with the origin in the interior. Every line that passes through the origin must intersect the polynomial defining the boundary exactly d times (counting multiplicities, and points at infinity), where d is the degree of the boundary polynomial.

This additional requirement is quite strong, and immediately allows us to discard sets for which the previous conditions were satisfied.

Example 3. *Consider the set described by $x^4 + y^4 \leq 1$; see Figure 1. It clearly satisfies the first four necessary conditions. However, it we consider any line through the origin, it will intersect the defining polynomial only two times, instead of the four required by the rigid convexity condition. Thus, this set is not rigidly convex, and hence does not admit a (non-lifted) semidefinite representation.*

5 Characterization

It should be apparent that the rigid convexity condition looks very similar to the hyperbolicity property of a polynomial. In fact, they are exactly the *same* condition, provided we redefine things accordingly [LPR05]. As we will see, this equivalence will make explicit the connection between the Helton & Vinnikov characterization of SDP-representable sets and the Lax conjecture described earlier.

Theorem 4 ([LPR05])**.** *If $p \in \mathbb{R}[x, y, z]$ is a polynomial of degree d, hyperbolic with respect to $e = (0, 0, 1)$ and that satisfies $p(e) = 1$, then the polynomial in $\mathbb{R}[x, y]$ defined by $q(x, y) = p(x, y, 1)$ is a real zero polynomial of degree no more than d, and satisfying $q(0, 0) = 1$.*

Conversely, if $q \in \mathbb{R}[x, y]$ is a real zero polynomial of degree d satisfying $q(0, 0) = 1$, then the polynomial defined by

$$p(x, y, z) = z^d q\left(\frac{x}{z}, \frac{y}{z}\right)$$

is a hyperbolic polynomial of degree d with respect to $e = (0, 0, 1)$, and $p(e) = 1$.

In their paper [HV07], Helton and Vinnikov proved that the rigid convexity condition fully characterizes the plane sets that are semidefinite representable.

Theorem 5 ([HV07])**.** *If $p(x, y)$ is a real zero polynomial of degree d with $p(0) > 0$, then the closure of the connected component of $p(x) > 0$ containing the origin admits a representation as in (1).*

[1] The Zariski topology on $\mathbb{C}^n$ can be defined in terms of its closed sets, which are the algebraic varieties, i.e., the vanishing set of a finite set of polynomial equations. The Zariski topology is a very weak topology, and is quite different from the usual topology in $\mathbb{C}^n$. For instance, the Zariski closure of the open interval $(0, 1)$ is equal to $\mathbb{C}$. The Zariski topology is not Hausdorff, i.e., distinct points do not always have disjoint neighborhoods.

For hyperbolic cones, we have shown earlier that the specific hyperbolicity direction e does not matter too much (as long as it belong to the hyperbolicity cone). Similarly, it can be shown that when checking the real zero condition we can choose any point in the interior of the set, not necessarily the origin.

Combining these two results, the truth of the Lax conjecture follows:

Theorem 6. *Every hyperbolic polynomial in three variables admits a determinantal representation of the type (2). If coordinates are chosen so that $e = (1, 0, 0)$, then we can choose $A_1 = I$.*

An interesting issue concerns the possibility of a constructive approach. In other words, given a hyperbolic polynomial in three variables, how to effectively obtain matrices A_i that give a determinantal representation? While "explicit" formulae for these matrices are given in [HV07] in terms of objects that are quite complicated to compute (namely, theta functions of Jacobian varieties), it may perhaps be the case that a more elementary formulation exists. For details about the state of the art of the computation of these representations, please see the recent work [PSV12]. In the homework exercises, we will explore two important special cases for which relatively straightforward constructions are possible.

5.1 Example

As an illustration, consider the convex set shown in Figure 3, which corresponds to the "oval" of the elliptic curve given by $3 + x - x^3 - 3x^2 - 2y^2 = 0$. This set satisfies the real zero condition, since every line that passes through a point in the interior of the set intersects the polynomial defining the boundary at exactly three points (if the lines are vertical, then the corresponding intersections are at infinity).

Homogenizing this polynomial, we obtain $p(x, y, z) = 3z^3 + xz^2 - x^3 - 3x^2z - 2y^2z$; the corresponding zero set is given in Figure 3. As we can see (and proved earlier), the section corresponding to the plane $z = 1$ is exactly the zero set of the original polynomial. Furthermore, lines parallel to the hyperbolicity direction e are projectively mapped into lines in this plane that go through the origin. Hence, the number of intersections (and thus, real roots) is preserved.

The theorem presented above promises the existence of a semidefinite representation. In this case, one such representation is:

$$\begin{bmatrix} x+1 & 0 & y \\ 0 & 2 & -x-1 \\ y & -x-1 & 2 \end{bmatrix} \succeq 0, \tag{3}$$

with the corresponding determinantal representation of the hyperbolic polynomial being:

$$p(x, y, z) = \det \begin{bmatrix} x+z & 0 & y \\ 0 & 2z & -x-z \\ y & -x-z & 2z \end{bmatrix}. \tag{4}$$

References

[HV07] J. W. Helton and V. Vinnikov. Linear matrix inequality representation of sets. *Comm. Pure Appl. Math.*, 60(5):654–674, 2007.

[Lax58] P. D. Lax. Differential equations, difference equations and matrix theory. *Comm. Pure Appl. Math.*, 11:175–194, 1958.

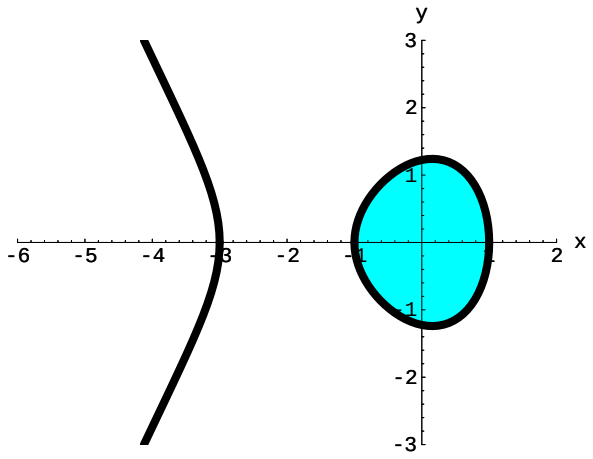

Figure 2: Convex set defined by $\{3+x-x^3-3x^2-2y^2 \geq 0, x \geq -1\}$. A semidefinite representation is given in (3).

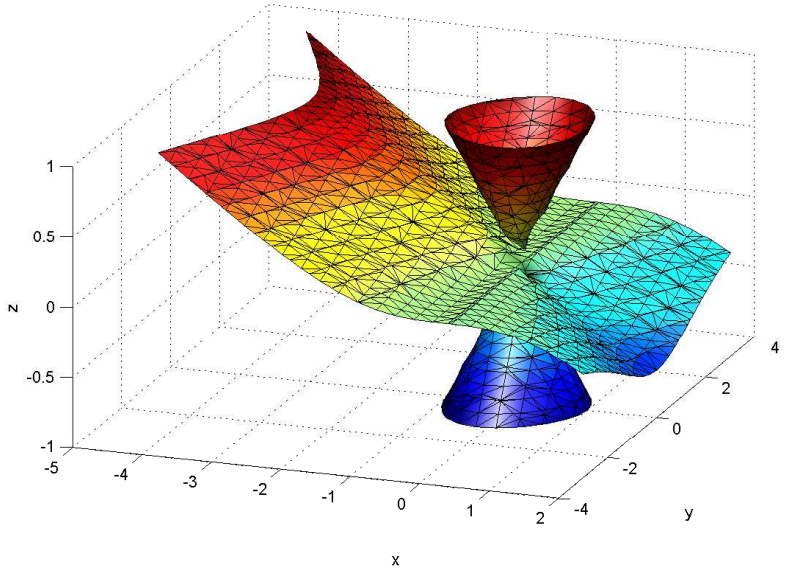

Figure 3: The polynomial $3z^3+xz^2-x^3-3x^2z-2y^2z=0$ and corresponding hyperbolicity cone.

[LPR05] A. S. Lewis, P. A. Parrilo, and M. V. Ramana. The Lax conjecture is true. *Proc. Amer. Math. Soc.*, 133(9):2495–2499, 2005.

[PSV12] D. Plaumann, B. Sturmfels, and C. Vinzant. Computing linear matrix representations of Helton-Vinnikov curves. In *Mathematical Methods in Systems, Optimization, and Control*, pages 259–277. Springer, 2012.

MIT 6.7230 - Algebraic techniques and semidefinite optimization March 8, 2024

Lecture 7

Lecturer: Pablo A. Parrilo *Scribe: ???*

In this lecture, we study first a relatively simple type of polynomial equations, namely *binomial equations*. As we will see, in this case there exists a quite efficient solution method. We define next an important geometric and combinatorial object associated with every multivariate polynomial, called the *Newton polytope*. Finally, we put together these two notions in the formulation of a family of bounds on the number of solutions of systems of polynomial equations. Our presentation of the material here is inspired by [Stu02, Chapter 3] and [CLO98].

1 Binomial equations

We introduce in this section a particular kind of polynomial equations, that have nice computational properties. A *binomial* system of polynomial equations is one where each equation has only two terms. We also assume that the system has only a finite number of complex solutions, i.e., the solution set is a finite set of points in $\mathbb{C}^n$. We are interested in determining the exact number of solutions, and in efficient computational procedures for solving the system.

Let's start with an example. Consider the binomial system given by

$$\begin{aligned} 8x^2y^3 - 1 &= 0 \\ 2x^3y^2 - yx &= 0. \end{aligned} \tag{1}$$

If we assume that the solutions satisfy $x \neq 0, y \neq 0$, then we can put these equations in the more symmetric form

$$\begin{aligned} 8x^2y^3 &= 1 \\ 2x^2y &= 1. \end{aligned} \tag{2}$$

Now, by dividing the first equation by the second one, we obtain $4y^2 = 1$, which has two solutions ($y = \frac{1}{2}$ and $y = -\frac{1}{2}$). Substituting into the resulting equations for every value of y we have two corresponding values of x, so the system has a big total of four complex solutions.

Let's try to understand in a bit more detail the manipulations we were performing here. For this, let's define the integer matrix

$$B = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}$$

corresponding to the exponents in (2). Notice that when we divided the two equations, that is equivalent to an elementary row operation in the matrix B, namely subtracting the second row of B from the first one. Thus, the operations we have done can be understood as the matrix multiplication $UB = C$, where

$$U = \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix}, \qquad C = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}.$$

The fact that the matrix C is lower triangular, is what allows us to start solving the system for y, and then backsolving for the other variable.

It is not too difficult to understand from this example how to generalize this. Let $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$, and consider a system of binomial equations in n variables, where we are interested in computing

(or bounding the number of) solutions in $(\mathbb{C}^*)^n$. We can always put the system in the normalized form in (2). Notice that, in general, the entries of the integer B could be either positive or negative (i.e., we write polynomials in x_i and x_i^{-1}, which is fine since $x_i \neq 0$).

Then, a well-known result in integer linear algebra (the Hermite normal form of an integer matrix) guarantees the existence of a unimodular matrix $U \in SL_n(\mathbb{Z})$ (an integer matrix, with determinant equal to ± 1), such that $C = UB$ is a lower triangular matrix. We can then use this expression to obtain values for the last variable, and backsolve to obtain all solutions.

How can we determine the number of solutions from this factorization? When backsubstituting using C, at each step we have to solve an equation of the type $x_i^{c_{ii}} = d_i$, and thus the current number of possible solutions is multiplied by $|c_{ii}|$. Therefore, the total number of solutions in $(\mathbb{C}^*)^n$ will then be equal to $|\det(C)| = |\det(U)\det(B)| = |\det(B)|$.

Remark 1. *To compute the Hermite normal form of an integer matrix in Maple, you can use the command* `ihermite`*. In Mathematica, use instead* `HermiteDecomposition`*.*

Example 2. *Consider the binomial polynomial system*

$$y^3 + 3xz = 0, \quad x^2y - 7z^4 = 0, \quad 1 - 2xy^2 = 0.$$

The corresponding matrix of exponents B is

$$B = \begin{bmatrix} -1 & 3 & -1 \\ 2 & 1 & -4 \\ -1 & -2 & 0 \end{bmatrix}$$

and its associated Hermite normal form factorization $C = UB$ is

$$\begin{bmatrix} 1 & 0 & 5 \\ 0 & 1 & 9 \\ 0 & 0 & 23 \end{bmatrix} = \begin{bmatrix} -1 & -1 & -2 \\ -1 & -2 & -3 \\ -3 & -5 & -7 \end{bmatrix} \begin{bmatrix} -1 & 3 & -1 \\ 2 & 1 & -4 \\ -1 & -2 & 0 \end{bmatrix}.$$

Since $|\det(B)| = 23$, this polynomial system has 23 solutions on $(\mathbb{C}^*)^3$.

2 Newton polytopes

Many of the polynomial systems that appear in practice are far from being "generic," but rather present a number of structural features that, when properly exploited, allow for much more efficient computational techniques. This is quite similar to the situation in numerical linear algebra, where there is a big difference in performance between algorithms that take into account the sparsity structure of a matrix and those that do not. For matrices, the standard notion of sparsity is relatively straightforward, and relates mostly to the number of nonzero coefficients. In computational algebra, however, there exists a much more refined notion of sparsity that refers not only to the number of zero coefficients of a polynomial, but also to the underlying combinatorial structure.

This notion of sparsity for multivariate polynomials is usually presented in terms of the *Newton polytope* of a polynomial, defined below.

Definition 3. *Consider a multivariate polynomial $p(x_1, \ldots, x_n) = \sum_\alpha c_\alpha x^\alpha$. The* Newton polytope *of p, denoted by New(f), is defined as the convex hull of the set of exponents α, considered as vectors in $\mathbb{R}^n$.*

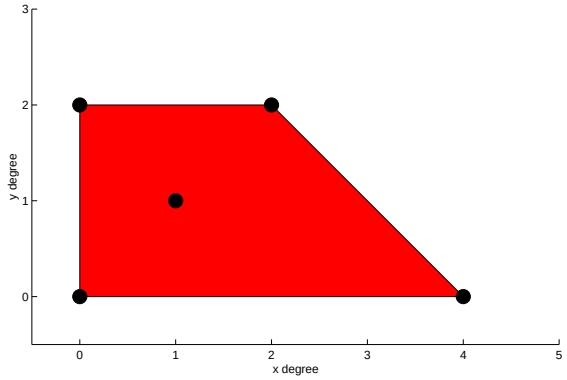

Figure 1: Newton polytope of the polynomial $p(x, y) = 5 - xy - x^2y^2 + 3y^2 + x^4$.

Thus, the Newton polytope of a polynomial always has integer extreme points, given by a subset of the exponents of the polynomial.

Example 4. *Consider the polynomial* $p(x, y) = 5 - xy - x^2y^2 + 3y^2 + x^4$. *Its Newton polytope* $New(f)$, *displayed in Figure 1, is the convex hull of the points* $(0, 0), (1, 1), (2, 2), (0, 2), (4, 0)$.

Example 5. *Consider the polynomial* $p(x, y) = 1 - x^2 + xy + 4y^4$. *Its Newton polytope* $New(p)$ *is the triangle in* $\mathbb{R}^2$ *with vertices* $\{(0, 0), (2, 0), (0, 4)\}$.

Newton polytopes are an essential tool when considering polynomial arithmetic because of the following fundamental identity:

$$\text{New}(g \cdot h) = \text{New}(g) + \text{New}(h),$$

where $+$ denotes the Minkowski addition of polytopes.

Example 6. *Let* $p(a, b, c, d) = (a^4 + 1)(b^4 + 1)(c^4 + 1)(d^4 + 1) + 2a + 3b + 4c + 5d$. *Its Newton polytope is the hypercube in* $\mathbb{R}^4$ *of side length equal to* 4, *and with opposing vertices at* $(0, 0, 0, 0)$ *and* $(4, 4, 4, 4)$.

It is a general theme in computational algebra that the complexity of many problems involving polynomials is directly related to some measure of the size of the corresponding Newton polytopes. We discuss an example below, in terms of the number of solutions of polynomial equations. We will encounter Newton polytopes again later in the course, when discussing the semidefinite characterization of polynomials that are sums of squares.

3 The Bézout and BKK bounds

Consider a system of two polynomial equations, $p(x, y) = 0$, $q(x, y) = 0$. As we have seen in previous lectures, we can solve this by computing the resultant of the polynomials p and q with respect to either variable, and then factorizing the corresponding univariate polynomial. If the degree of the polynomials is d_1 and d_2, respectively, then the degree of the resultant is bounded by $d_1 \cdot d_2$, and thus the number of zeros of the system is at most this number.

However, when the polynomials p and q are sparse (in the sense defined earlier) then the number of solutions can be much smaller. For instance, the system

$$\begin{aligned} a + bx + cy + dy^2 &= 0 \\ ex + fy + gxy &= 0 \end{aligned} \tag{3}$$

has, for a generic choice of the coefficients $\{a, \ldots, g\}$, exactly three complex roots, while the bound based on the individuals degrees (usually called the Bézout bound) will give a total of $2 \times 2 = 4$. As we will see, much sharper bounds can be obtained by considering the Newton polytopes of the individual equations.

To introduce the main theorem, we need to introduce the following concept, that generalizes the notion of volume of a polytope, to a collection of them.

Definition 7. *Consider polytopes $P_1, \ldots, P_n \subset \mathbb{R}^n$, nonnegative scalars $\lambda_1, \ldots, \lambda_n$, and let $V(\lambda) = \mathrm{Vol}(\lambda_1 P_1 + \cdots + \lambda_n P_n)$. It can be shown that $V(\lambda)$ is a homogeneous polynomial of degree n. The* mixed volume $MV(P_1, \ldots, P_n)$ *is the coefficient of this polynomial, corresponding to the monomial $\lambda_1 \lambda_2 \cdots \lambda_n$.*

Although not obvious from its definition, the mixed volume is always a nonnegative number. It further satisfies a number of very interesting properties, such as the Aleksandrov-Fenchel inequality[1]. Although computing the mixed volume is difficult in general, in certain cases it can be approximated via convex optimization methods with strong relationships to hyperbolic polynomials [Gur09].

One of the main result in this area, with different versions due to Bernstein, Kouchnirenko, and Khovanskii, relates the number of solutions of a sparse polynomial system with the mixed volume of the Newton polytopes of the individual equations. Formally, we have

Theorem 8 (BKK bound)**.** *The number of solutions in $(\mathbb{C}^*)^n$ of a sparse polynomial system of n equations and n unknowns is less than or equal to the mixed volume of the n Newton polytopes. If the coefficients are "generic" enough, then the upper bound is achieved.*

The basic idea behind the derivation of the theorem is to introduce an additional parameter t in the equations, in such a way that for $t = 1$ we have the original system, while for $t = 0$ the system is binomial, which as we have seen can be solved in an efficient manner. This process is usually called a *toric deformation*, and is somewhat similar in spirit to the homotopies used in interior point methods. To make our words a bit more precise, an important fact is that we will not deform to just one binomial system, but actually to a collection of them, given by what is called a *mixed subdivision* of the sum of Newton polytopes. The important fact is that the sum of the number of roots of all these binomial systems is exactly equal to the mixed volume of the collection of polytopes.

Example 9. *Consider the univariate polynomial*

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_m x^m,$$

where $n \geq m$. It is clear that the Newton polytope is the line segment with endpoints in n and m. The mixed volume (in this case, just the volume) is equal to $n - m$. Thus, the BKK bound for this polynomial is equal to $n - m$, which is clearly exact for generic choices of the coefficients.

[1] The Aleksandrov-Fenchel inequality, valid for convex bodies $C_i \subset \mathbb{R}^n$, states that for $1 \leq k \leq n$,

$$MV(C_1, C_2, \ldots, C_n)^k \geq \prod_{j=1}^{k} MV(\underbrace{C_j, \ldots, C_j}_{k \text{ times}}, C_{k+1}, \ldots, C_n).$$

An important consequence of Aleksandrov-Fenchel is the (convex) Brunn-Minkowski inequality: if S and T are convex bodies in $\mathbb{R}^n$ and $0 \leq \lambda \leq 1$, then $\mathrm{Vol}((1-\lambda)S + \lambda T)^{\frac{1}{n}} \geq (1-\lambda)\mathrm{Vol}(S)^{\frac{1}{n}} + \lambda \mathrm{Vol}(T)^{\frac{1}{n}}$; see [Gar02]. Notice that if S and T are ellipsoids, this yields the concavity of the n-th root of the determinant on positive semidefinite matrices.

Example 10. *Let us consider again the example discussed in (1). The Newton polytope of the first polynomial is the line segment with endpoints* $(0,0)$ *and* $(2,3)$*, while the second one has endpoints* $(1,1)$ *and* $(3,2)$*. If we denote these by* P_1 *and* P_2*, it is easy to see that*

$$\mathrm{Vol}(\lambda_1 P_1 + \lambda_2 P_2) = 4\lambda_1\lambda_2,$$

and thus the mixed volume of (P_1, P_2) *is equal to* 4*, which is the number of solutions of (1).*

Example 11. *Consider now the example in equation (3). The Newton polytope of the first polynomial is the triangle with vertices* $\{(0,0),(1,0),(0,2)\}$*, and the second one is the triangle with vertices* $\{(1,0),(0,1),(1,1)\}$*. It's not hard to show (how?) that*

$$\mathrm{Vol}(\lambda_1 P_1 + \lambda_2 P_2) = \lambda_1^2 + \frac{1}{2}\lambda_2^2 + 3\lambda_1\lambda_2,$$

and thus $MV(P_1, P_2) = 3$*, which as we have seen, is the number of solutions of (3) when the coefficients are "generic."*

4 Application: Nash equilibria

We can use the results described, to give a bound on the number of isolated Nash equilibria of a game. For simplicity, consider the three-player case, where each player has two pure strategies. We are interested here only in totally mixed equilibria, i.e., those where the players randomize among all their pure strategies with nonzero probability (if this is not the case, then by eliminating the never played strategies we can reduce the game to the totally mixed case). Thus, the mixed strategies can be parametrized in terms of three variables $a, b, c \in (0,1)$, representing the probabilities with which they play their different strategies.

It can be shown that the Nash equilibrium condition result in a polynomial system of the structure

$$\begin{aligned} p_{11}bc + p_{12}b + p_{13}c + p_{14} &= 0 \\ p_{21}ca + p_{22}c + p_{23}a + p_{24} &= 0 \\ p_{31}ab + p_{32}a + p_{33}b + p_{34} &= 0, \end{aligned} \tag{4}$$

where the coefficients p_{ij} are explicit linear functions of the payoffs. The mixed volume of the Newton polytopes of these three equations is equal to 2, so the maximum number of totally mixed Nash equilibria that a three-player, two-strategy game can have is equal to two.

The same argument can be generalized to the case of n players, obtaining the following result:

Theorem 12 ([MM97],[Stu02, p.82])**.** *The maximum number of isolated totally mixed Nash equilibria for an* n*-person game where each player has two pure strategies is equal to the mixed volume of the* n *facets of the* n *cube.*

This mixed volume can be computed explicitly, and is equal to the number of *derangements* (fixed-point free permutations) of a set with n elements. This number is also the *permanent*[2] of the matrix $E_n - I_n$, where E_n is the all-ones matrix. It can be shown that this number is the closest integer to $n!/e$.

There are extensions of this result to the case of graphical games; see [Stu02] and the references therein for details.

[2]The permanent of a square matrix $A \in \mathbb{R}^{n\times n}$ is defined as $\mathrm{per}(A) := \sum_{\sigma\in\Pi_n}\prod_{i=1}^{n} a_{i,\sigma(i)}$, where Π_n is the set of all permutations in n elements. The formula is quite similar to that of the determinant (except that the signs of all terms are always positive). In contrast to the determinant, which can easily be obtained in polynomial time via Gaussian elimination, it is believed that the permanent is hard to compute (in fact, it is #P-hard).

References

[CLO98] D. A. Cox, J. B. Little, and D. O'Shea. *Using Algebraic Geometry*, volume 185 of *Graduate Texts in Mathematics*. Springer-Verlag, 1998.

[Gar02] R. Gardner. The Brunn-Minkowski inequality. *Bulletin of the American Mathematical Society*, 39(3):355–405, 2002.

[Gur09] L. Gurvits. A polynomial-time algorithm to approximate the mixed volume within a simply exponential factor. *Discrete & Computational Geometry*, 41(4):533–555, 2009.

[MM97] R.D. McKelvey and A. McLennan. The maximal number of regular totally mixed Nash equilibria. *Journal of Economic Theory*, 72(2):411–425, 1997.

[Stu02] B. Sturmfels. *Solving Systems of Polynomial Equations*. AMS, Providence, R.I., 2002.

MIT 6.7230 - Algebraic techniques and semidefinite optimization March 13, 2024

Lecture 8

Lecturer: Pablo A. Parrilo *Scribe: ???*

In this lecture we begin our study of one of the main themes of the course, namely the relationships between polynomials that are sums of squares and semidefinite programming.

1 Nonnegativity and sums of squares

Recall from a previous lecture the definition of a polynomial being a sum of squares.

Definition 1. *A univariate polynomial $p(x)$ is a* sum of squares *(SOS) if there exist $q_1, \ldots, q_m \in \mathbb{R}[x]$ such that*

$$p(x) = \sum_{k=1}^{m} q_k^2(x). \tag{1}$$

If a polynomial $p(x)$ is a sum of squares, then it obviously satisfies $p(x) \geq 0$ for all $x \in \mathbb{R}$. Thus, a SOS condition is a sufficient condition for global nonnegativity.

As we have seen, in the univariate case, the converse is also true:

Theorem 2. *A univariate polynomial is nonnegative if and only if it is a sum of squares.*

There is a very direct link between sum of squares conditions on polynomials and semidefinite programming. We study first the univariate case.

2 Sums of squares and semidefinite programming

Consider a polynomial $p(x)$ of degree $2d$ that is a sum of squares, i.e., it can be written as in (1). Notice that the degree of the polynomials q_k is at most equal to d, since the highest term of each q_k^2 is positive, and thus there cannot be any cancellation in the highest power of x. Then, we can write

$$\begin{bmatrix} q_1(x) \\ q_2(x) \\ \vdots \\ q_m(x) \end{bmatrix} = V \begin{bmatrix} 1 \\ x \\ \vdots \\ x^d \end{bmatrix}, \tag{2}$$

where $V \in \mathbb{R}^{m \times (d+1)}$, and its kth row contains the coefficients of the polynomial q_k. For future reference, let $[x]_d$ be the vector in the right-hand side of (2). Consider now the matrix $Q = V^T V$. We then have $p(x) = \sum_{k=1}^m q_k^2(x) = (V[x]_d)^T (V[x]_d) = [x]_d^T V^T V [x]_d = [x]_d^T Q [x]_d$.

Conversely, assume there exists a symmetric positive definite Q, for which $p(x) = [x]_d^T Q [x]_d$. Then, by factorizing $Q = V^T V$ (e.g., via Cholesky, or square root factorization), we arrive at a SOS decomposition of p.

We formally express this in the following lemma, that gives a direct relation between positive semidefinite matrices and a sum of squares condition.

Lemma 3. *Let $p(x)$ be a univariate polynomial of degree $2d$. Then, $p(x)$ is nonnegative (or SOS) if and only if there exists $Q \in \mathcal{S}_+^{d+1}$ that satisfies*

$$p(x) = [x]_d^T Q [x]_d. \tag{3}$$

Indexing the rows and columns of Q by $\{0, \ldots, d\}$, we have:

$$[x]_d^T Q [x]_d = \sum_{j=0}^{d} \sum_{k=0}^{d} Q_{jk} x^{j+k} = \sum_{i=0}^{2d} \left(\sum_{j+k=i} Q_{jk} \right) x^i$$

Thus, for this expression to be equal to $p(x)$, it should be the case that

$$p_i = \sum_{j+k=i} Q_{jk}, \qquad i = 0, \ldots, 2d. \tag{4}$$

This is a system of $2d+1$ linear equations between the entries of Q and the coefficients of $p(x)$. Thus, since Q is simultaneously constrained to be positive semidefinite, and to belong to a particular affine subspace, a SOS condition is exactly equivalent to a semidefinite programming problem.

Lemma 4. *A polynomial $p(x) = \sum_{i=0}^{2d} p_i x^i$ is a sum of squares if and only if there exists $Q \in \mathcal{S}_+^{d+1}$ satisfying (4). This is a semidefinite programming problem.*

Example 5. *Consider the univariate polynomial*

$$p(x) = x^4 + 4x^3 + 6x^2 + 4x + 5,$$

for which we want to find an sos decomposition. Proceeding as described earlier, we consider the expression:

$$\begin{aligned} p(x) &= \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}^T \begin{bmatrix} q_{00} & q_{01} & q_{02} \\ q_{01} & q_{11} & q_{12} \\ q_{02} & q_{12} & q_{22} \end{bmatrix} \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix} \\ &= q_{22}x^4 + 2q_{12}x^3 + (q_{11} + 2q_{02})x^2 + 2q_{01}x + q_{00}. \end{aligned}$$

Matching coefficients, we obtain the linear equality constraints:

$$\begin{aligned} x^4: &\quad 1 = q_{22}, \\ x^3: &\quad 4 = 2q_{12}, \\ x^2: &\quad 6 = q_{11} + 2q_{02}, \\ x: &\quad 4 = 2q_{01}, \\ 1: &\quad 5 = q_{00}. \end{aligned}$$

We need to find a positive semidefinite matrix Q, that satisfies these linear equations (i.e., solve a semidefinite program). In this case, the SDP is feasible, and we can obtain a solution given by:

$$Q = \begin{bmatrix} 5 & 2 & 0 \\ 2 & 6 & 2 \\ 0 & 2 & 1 \end{bmatrix} = V^T V, \qquad V = \begin{bmatrix} 0 & 2 & 1 \\ \sqrt{2} & \sqrt{2} & 0 \\ \sqrt{3} & 0 & 0 \end{bmatrix},$$

which yields the sum of squares decomposition

$$p(x) = (x^2 + 2x)^2 + 2(x+1)^2 + 3.$$

Remark 6. *When writing down the SDP associated to the SOS decomposition of a polynomial, as we did in the previous section, there is an implicit choice of bases for two vector spaces: one for the space of polynomials $\mathbb{R}_n[x]$, and one for the dual space (to write the constraints). Indeed, in our formulation, the polynomial $p(x)$ was expressed as a quadratic form on the vector space $\mathbb{R}_n[x]$, represented by the matrix Q with respect to the monomial basis; see* (3)*. Similarly, the constraints* (4) *correspond to the coefficients of $p(x) = [x]_d^T Q[x]_d$ with respect to the monomial basis. While these choices are perhaps "canonical," there are several alternative bases that can be used instead, and these can have quite different algebraic and numerical properties – we may explore some of these issues in the homework problems.*

3 Applications and extensions

We discuss first a few applications of the SDP characterization of nonnegative polynomials, followed by several extensions.

3.1 Optimization

Our first application concerns the global optimization of a univariate polynomial $p(x)$. Rather than focusing on computing an $x_\star$ for which $p(x_\star)$ is as small as possible, we attempt first to obtain a good (or the best) lower bound on its optimal value. It is easy to see that a number γ is a global lower bound of a polynomial $p(x)$, if and only if the polynomial $p(x) - \gamma$ is nonnegative, i.e.,

$$p(x) \geq \gamma \quad \forall x \in \mathbb{R} \qquad \Longleftrightarrow \qquad p(x) - \gamma \geq 0 \quad \forall x \in \mathbb{R}.$$

Notice that the polynomial $p(x) - \gamma$ has coefficients that depend affinely on γ. Consider now the optimization problem defined by

$$\max \gamma \qquad \text{s.t.} \qquad p(x) - \gamma \text{ is SOS.}$$

It should be clear that this is a *convex* problem, since the feasible set is defined by an infinite number of linear inequalities. Its optimal solution $\gamma_\star$ is equal to the global minimum of the polynomial, $p(x_\star)$. Furthermore, using Lemma 4, we can easily write this as a semidefinite programming problem. We can thus obtain the global minimum of a univariate polynomial, by solving an SDP problem. Notice also that at optimality, we have $0 = p(x_\star) - \gamma_\star = \sum_{k=1}^m q_k^2(x_\star)$, and thus all the q_k simultaneously vanish at $x_\star$, which gives a way of computing the optimal solution $x_\star$. As we shall see later, we can also obtain this solution directly from the dual problem, by using complementary slackness.

Notice that even though $p(x)$ may be highly nonconvex, we are nevertheless effectively computing its global minimum.

3.2 Nonnegativity on intervals

We have seen how to characterize a univariate polynomial that is nonnegative on $(-\infty, \infty)$ in terms of SDP conditions. But what if we are interested in polynomials that are nonnegative only in an interval (either finite, or semi-infinite)? As explained below, we can use very similar ideas, and two classical characterizations, usually associated to the names Pólya-Szegö, Fekete, or Markov-Lukacs. The basic results are the following:

Theorem 7. *The polynomial $p(x)$ is nonnegative on $[0, \infty)$, if and only if it can be written as*

$$p(x) = s(x) + x \cdot t(x),$$

where $s(x), t(x)$ are SOS. If $\deg(p) = 2d$, then we have $\deg(s) \leq 2d$, $\deg(t) \leq 2d-2$, while if $\deg(p) = 2d+1$, then $\deg(s) \leq 2d$, $\deg(t) \leq 2d$.

Theorem 8. *Let $a < b$. Then, $p(x)$ is nonnegative on $[a, b]$, if and only if it can be written as*

$$\begin{cases} p(x) = s(x) + (x-a) \cdot (b-x) \cdot t(x), & \text{if } \deg(p) \text{ is even} \\ p(x) = (x-a) \cdot s(x) + (b-x) \cdot t(x), & \text{if } \deg(p) \text{ is odd} \end{cases}$$

where $s(x), t(x)$ are SOS. In the first case, we have $\deg(p) = 2d$, and $\deg(s) \leq 2d$, $\deg(t) \leq 2d-2$. In the second, $\deg(p) = 2d+1$, and $\deg(s) \leq 2d$, $\deg(t) \leq 2d$.

Notice that in both of these results, the "easy" direction of the characterization is obvious.

3.3 Rational functions

What happens if we want to minimize a univariate rational function, rather than a polynomial? Consider a rational function given as a quotient of polynomials $p(x)/q(x)$, where $q(x)$ is strictly positive (why?). Then, we have

$$\frac{p(x)}{q(x)} \geq \gamma \qquad \Leftrightarrow \qquad p(x) - \gamma\, q(x) \geq 0,$$

and therefore we can find the global minimum of the rational function by solving

$$\max \gamma \qquad \text{s.t.} \quad p(x) - \gamma\, q(x) \text{ is SOS.}$$

The constrained case (i.e., over finite or semi-infinite intervals) are very similar, and can be formulated using the results in Section 3.2. The details are left for the exercises.

4 Multivariate polynomials

If a polynomial $p(x)$ is a sum of squares, it is always true that $p(x) \geq 0$. However, for polynomials in more than one variable, it is no longer true that nonnegativity is equivalent to a sum of squares condition. In fact, for polynomials of degree greater than or equal to four, deciding polynomial nonnegativity is an NP-hard problem (as a function of the number of variables).

More than a century ago, David Hilbert showed that equality between the set of nonnegative and SOS polynomials holds only in the following three cases:

- Univariate polynomials (i.e., $n = 1$)
- Quadratic polynomials ($2d = 2$)
- Bivariate quartics ($n = 2$, $2d = 4$)

For all other cases, there always exist nonnegative polynomials that are *not* sums of squares. A classical counterexample is the bivariate sextic ($n = 2$, $2d = 6$) due to Motzkin, given by (in dehomogenized form)

$$M(x, y) = x^4y^2 + x^2y^4 + 1 - 3x^2y^2.$$

This polynomial is nonnegative, but is not a sum of squares. We will prove both facts later. An excellent account of much of the classical work in this area has been provided by Bruce Reznick in [Rez00].

4.1 SDP formulation

Essentially the same construction we have seen in Lemma 4 applies to the multivariate case. In this case, we consider polynomials in n variables of degree $2d$. In the dense case, i.e., when the polynomial is not sparse, the number of coefficients is equal to $\binom{n+2d}{2d}$. If we let $p(x) = \sum_\alpha p_\alpha x^\alpha$, and indexing the matrix Q by the $\binom{n+d}{d}$ monomials in n variables of degree d, we have the SDP conditions on $Q \in \mathcal{S}_+^{\binom{n+d}{d}}$:

$$p_\alpha = \sum_{\beta+\gamma=\alpha} Q_{\beta\gamma}, \qquad\qquad Q \succeq 0. \tag{5}$$

We have exactly $\binom{n+2d}{2d}$ linear equations, one per each coefficient of $p(x)$. As before, these conditions are affine conditions relating the entries of Q and the coefficients of $p(x)$. Thus, we can decide membership to, or optimize over, the set of SOS polynomials by solving an SDP.

4.2 Using the Newton polytope

Recall that we have defined in a previous lecture the *Newton polytope* of a polynomial $p(x) \in \mathbb{R}[x_1, \ldots, x_n]$ as the convex hull of the set of exponents appearing in p. This allowed us to introduce a notion of sparseness for a polynomial, related to the size of its Newton polytope. Sparsity (in this algebraic sense) allows a notable reduction in the computational cost of checking sum of squares conditions of multivariate polynomials. The reason is the following theorem due to Reznick:

Theorem 9 ([Rez78], Theorem 1)**.** *If $p(x) = \sum q_i(x)^2$, then $New(q_i) \subseteq \frac{1}{2} New(p)$.*

In other words, this theorem allows us, without loss of generality, to restrict the set of monomials appearing in the representation (5) to those in the Newton polytope of p, scaled by a factor of $\frac{1}{2}$. This reduces the size of the corresponding matrix Q, thus simplifying the SDP problem.

Example 10. *Consider the following polynomial:*

$$p = (w^4 + 1)(x^4 + 1)(y^4 + 1)(z^4 + 1) + 2w + 3x + 4y + 5z.$$

The polynomial p has degree $2d = 16$, and four independent variables ($n = 4$). A naive approach, along the lines described earlier, would require a matrix Q of size $\binom{n+d}{d} = 495$. However, the Newton polytope of p is easily seen to be the four dimensional hypercube with vertices in $(0, 0, 0, 0)$ and $(4, 4, 4, 4)$. Therefore, the polynomials q_i in the SOS decomposition of p will have at most $3^4 = 81$ distinct monomials, and as a consequence the full decomposition can be computed by solving a much smaller SDP.

5 Duality and density

In the next lecture, we will revisit the sum of squares construction, but emphasizing this time the dual side, and its appealing measure-theoretic interpretation. We will also review some recent results on the relative density of the cones of nonnegative polynomials and SOS.

References

[Rez78] B. Reznick. Extremal PSD forms with few terms. *Duke Mathematical Journal*, 45(2):363–374, 1978.

[Rez00] B. Reznick. Some concrete aspects of Hilbert's 17th problem. In *Contemporary Mathematics*, volume 253, pages 251–272. American Mathematical Society, 2000.

MIT 6.7230 - Algebraic techniques and semidefinite optimization March 15, 2024

Lecture 9

Lecturer: Pablo A. Parrilo *Scribe: ???*

In this lecture we continue our study of SOS polynomials. After presenting a couple of applications, we discuss the dual side, and provide a natural probabilistic interpretation of the corresponding problem. We further introduce a natural geometric description, in terms of approximations to the convex hull of a certain algebraic variety.

1 Applications of sum of squares

1.1 Lyapunov functions

Expressing conditions for a polynomial to be a sum-of-squares as an SDP is very useful, since we can use the SOS property as a convenient "replacement" for polynomial nonnegativity. In the dynamical systems context, recent work has applied the sum-of-squares approach to the problem of finding a Lyapunov function for nonlinear systems [Par00, PP02].

This approach enables the search over affinely parametrized polynomial or rational Lyapunov functions for systems with dynamics of the form

$$\dot{x}_i(t) = f_i(x(t)) \qquad \text{for all } i = 1, \ldots, n \tag{1}$$

where the functions f_i are polynomials or rational functions. Recall that for a system to be globally asymptotically stable, it is sufficient to prove the existence of a Lyapunov function that satisfies

$$V(x) > 0, \qquad \dot{V}(x) = \left(\frac{\partial V}{\partial x}\right)^T f(x) < 0$$

for all $x \in \mathbb{R}^n \setminus \{0\}$, where without loss of generality we have assumed that the system (1) has an equilibrium at the origin (see, e.g., [Kha92]). Then the condition that the Lyapunov function be positive, and that its Lie derivative be negative, are both directly imposed as sum-of-squares constraints in terms of the coefficients of the Lyapunov function.

As an example, consider the following system:

$$\begin{aligned} \dot{x} &= -x + (1+x)y \\ \dot{y} &= -(1+x)x. \end{aligned} \tag{2}$$

It is known that this system has no quadratic Lyapunov function. However, using SOSTOOLS [PPP05] we easily find a quartic polynomial Lyapunov function, which after rounding (for purely cosmetic reasons) is given by

$$V(x,y) = 6x^2 - 2xy + 8y^2 - 2y^3 + 3x^4 + 6x^2y^2 + 3y^4.$$

The corresponding phase diagram, showing both the trajectories and the level sets of the Lyapunov function V, is given in Figure 1. It can be readily verified that both $V(x,y)$ and $(-\dot{V}(x,y))$ are SOS, since

$$V = \begin{bmatrix} x \\ y \\ x^2 \\ xy \\ y^2 \end{bmatrix}^T \begin{bmatrix} 6 & -1 & 0 & 0 & 0 \\ -1 & 8 & 0 & 0 & -1 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 6 & 0 \\ 0 & -1 & 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \\ x^2 \\ xy \\ y^2 \end{bmatrix}, \qquad -\dot{V} = \begin{bmatrix} x \\ y \\ x^2 \\ xy \end{bmatrix}^T \begin{bmatrix} 10 & 1 & -1 & 1 \\ 1 & 2 & 1 & -2 \\ -1 & 1 & 12 & 0 \\ 1 & -2 & 0 & 6 \end{bmatrix} \begin{bmatrix} x \\ y \\ x^2 \\ xy \end{bmatrix},$$

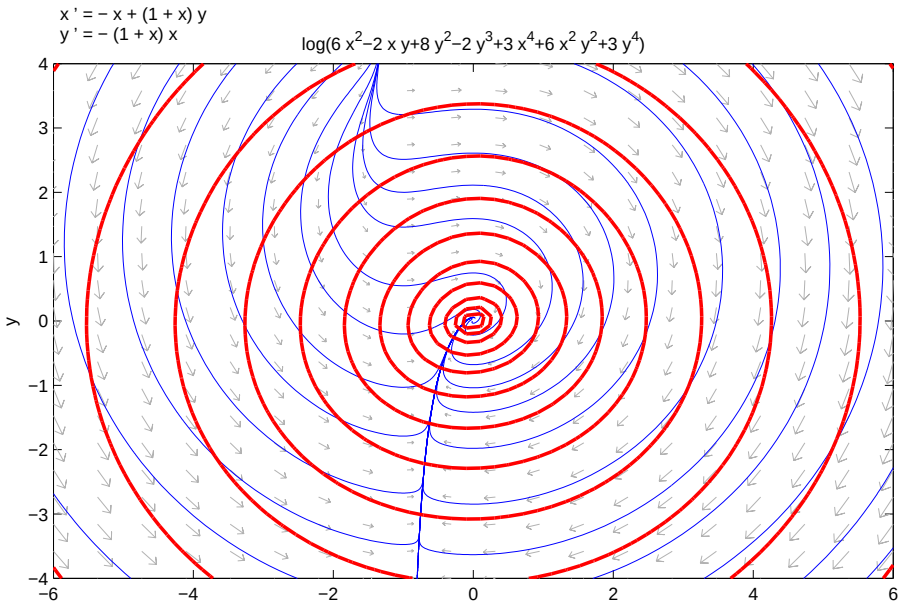

Figure 1: Trajectories of the dynamical system (2) and level sets of the SOS Lyapunov function $V(x, y)$.

and the matrices in the expression above are positive definite. Similar approaches may also be used for finding Lyapunov functionals for certain classes of hybrid systems.

Remark 1. *In stark constrast to the linear case (where quadratic Lyapunov always exist), polynomial dynamical systems that are globally asymptotically stable may not admit polynomial Lyapunov functions; see for instance [AKP11] for a simple counterexample. Nevertheless, the basic technique described above can be easily modified to deal with stability over compact regions – in this case (essentially, by the Stone-Weierstrass approximation theorem) polynomial Lyapunov functions that prove stability always exist.*

1.2 Entangled states in quantum mechanics

The state of a finite-dimensional quantum system can be described in terms of a positive semidefinite Hermitian matrix, called the *density matrix*. An important property of a bipartite quantum state ρ is whether or not it is *separable*, which means that it can be written as a convex combination of tensor products of rank one matrices, i.e.,

$$\rho = \sum_i p_i \, (x_i x_i^T) \otimes (y_i y_i^T), \qquad p_i \geq 0, \quad \sum_i p_i = 1,$$

where for simplicity we have restricted ρ, x_i, y_i to be real. Here $x_i \in \mathbb{R}^{n_1}$, $y_i \in \mathbb{R}^{n_2}$, and $\rho \in \mathcal{S}_+^{n_1 n_2}$. If the state is not separable, then it is said to be *entangled*.

A question of interest is the following: Given the density matrix ρ of a quantum state, how to recognize whether the state is entangled or not? How can we certify that the state is entangled? It has been shown by Gurvits that in general this is an NP-hard question [Gur03].

A natural mathematical object to study in this context is the set of *positive maps*, i.e., the linear operators $\Lambda : \mathcal{S}^{n_1} \to \mathcal{S}^{n_2}$ that map positive semidefinite matrices into positive semidefinite

matrices. Notice that to any such Λ, we can associate a unique "observable" $L \in \mathcal{S}^{n_1 n_2}$, that satisfies $y^T \Lambda(xx^T)y = (x \otimes y)^T L(x \otimes y)$. Furthermore, if Λ is a positive map, then the pairing between the observable L and any separable state will always give a nonnegative number, since

$$\begin{aligned}\langle L, \rho \rangle &= \operatorname{Tr} L \cdot (\sum_i p_i \, (x_i x_i^T) \otimes (y_i y_i^T)) = \sum_i p_i \operatorname{Tr} L \cdot (x_i \otimes y_i) \cdot (x_i \otimes y_i)^T \\ &= \sum_i p_i \, (x_i \otimes y_i)^T L (x_i \otimes y_i) = \sum_i p_i \, y_i^T \Lambda(x_i x_i^T) y_i \geq 0.\end{aligned}$$

In other words, every positive map yields a *separating hyperplane* for the convex set of separable states. It can further be shown that this is in fact a complete characterization (and thus, these sets are dual to each other).

The set of positive maps can be exactly characterized in terms of a multivariate polynomial nonnegativity condition, since the map $\Lambda : \mathcal{S}^{n_1} \to \mathcal{S}^{n_2}$ is positive if and only if the polynomial $p(x, y) = y^T \Lambda(xx^T)y$ is nonnegative for all x, y (why?). Replacing nonnegativity with sum of squares based conditions, we can obtain a family of efficiently computable criteria that certify entanglement.

For more background and details about this problem, see [DPS02, DPS04] and the references therein.

2 Dual side: moments

Consider a nonnegative measure μ on $\mathbb{R}$ (or if you prefer, a real-valued random variable X). We can then define the *moments*, which are the expectation of powers of X.

$$\mu_k := \mathbf{E}[X^k] = \int x^k d\mu \tag{3}$$

What constraints, if any, should the μ_k satisfy? Is is true that for any set of numbers $\mu_0, \mu_1, \ldots, \mu_k$, there always exists a nonnegative measure having exactly these moments? This is the classical (truncated) moment problem [Akh65].

It should be apparent that some conditions are required. For instance, consider (3) for an even value of k. Since the measure μ is nonnegative, it is clear that in this case we have $\mu_k \geq 0$.

However, that's clearly not enough, and more restrictions should hold. A simple one can be derived by recalling the relationship between the first and second moments and the variance of a random variable, i.e., $\operatorname{var}(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = \mu_2 - \mu_1^2$. Since the variance is always nonnegative, we should have $\mu_2 - \mu_1^2 \geq 0$.

How to systematically derive conditions of this kind? Notice that the previous inequality can be obtained by noticing that for all a, b,

$$0 \leq \mathbf{E}[(a + bX)^2] = a^2 \mathbf{E}[1] + 2ab\mathbf{E}[X] + b^2 \mathbf{E}[X^2] = \begin{bmatrix} a \\ b \end{bmatrix}^T \begin{bmatrix} 1 & \mu_1 \\ \mu_1 & \mu_2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix},$$

which implies that the 2×2 matrix above must be positive semidefinite. Interestingly, the variance inequality obtained earlier corresponds to the determinant of this matrix.

Exactly the same procedure can be done for higher-order moments. Proceeding this way, by considering expectations of squares of higher-order polynomials, we have

$$0 \leq \mathbf{E}[(c_0 + c_1 X + \cdots + c_d X^d)^2] = \sum_{j=0}^{d} \sum_{k=0}^{d} c_j c_k \mathbf{E}[X^{j+k}]$$

This is a quadratic form in the variables $(c_0, c_1, \ldots, c_d)$, and thus the higher order moments must always satisfy the semidefinite condition on a Hankel matrix given by

$$\begin{bmatrix} 1 & \mu_1 & \mu_2 & \cdots & \mu_d \\ \mu_1 & \mu_2 & \mu_3 & \cdots & \mu_{d+1} \\ \mu_2 & \mu_3 & \mu_4 & \cdots & \mu_{d+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_d & \mu_{d+1} & \mu_{d+2} & \cdots & \mu_{2d} \end{bmatrix} \succeq 0. \tag{4}$$

Notice that the diagonal elements correspond to even-order moments, which should obviously be nonnegative.

As we will see below, this condition is "almost" necessary and sufficient in the univariate case. In the multivariate case, however, there will be more serious problems (just like for polynomial nonnegativity vs. sums of squares).

Remark 2. *For unbounded intervals, the SDP conditions characterize the* closure *of the set of moments, but not necessarily the whole set. As an example, consider the set of moments given by* $\mu = (1, 0, 0, 0, 1)$*, corresponding to the Hankel matrix*

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Although the matrix above is PSD, it is not hard to see that there is no nonnegative measure corresponding to those moments. However, the parametrized atomic measure given by

$$\mu_\varepsilon = \frac{\varepsilon^4}{2} \cdot \delta(x + \frac{1}{\varepsilon}) + (1 - \varepsilon^4) \cdot \delta(x) + \frac{\varepsilon^4}{2} \cdot \delta(x - \frac{1}{\varepsilon})$$

has as first five moments $(1, 0, \varepsilon^2, 0, 1)$*, and thus as* $\varepsilon \to 0$ *the corresponding Hankel matrix is the one given above.*

2.1 Nonnegative measures on intervals

Just like we did for the case of polynomials nonnegative on intervals, we can similarly obtain necessary and sufficient characterizations for moments of measures supported on intervals. For simplicity, we present below only one particular case, corresponding to the interval $[-1, 1]$.

Lemma 3. *There exists a nonnegative measure supported on* $[-1, 1]$ *with moments* $(\mu_0, \mu_1, \ldots, \mu_{2d+1})$ *if and only if*

$$\begin{bmatrix} \mu_0 & \mu_1 & \mu_2 & \cdots & \mu_d \\ \mu_1 & \mu_2 & \mu_3 & \cdots & \mu_{d+1} \\ \mu_2 & \mu_3 & \mu_4 & \cdots & \mu_{d+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_d & \mu_{d+1} & \mu_{d+2} & \cdots & \mu_{2d} \end{bmatrix} \pm \begin{bmatrix} \mu_1 & \mu_2 & \mu_3 & \cdots & \mu_{d+1} \\ \mu_2 & \mu_3 & \mu_4 & \cdots & \mu_{d+2} \\ \mu_3 & \mu_4 & \mu_5 & \cdots & \mu_{d+3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_{d+1} & \mu_{d+2} & \mu_{d+3} & \cdots & \mu_{2d+1} \end{bmatrix} \succeq 0. \tag{5}$$

The necessity of this condition is clear, since it follows from consideration of the quadratic form (in the c_i):

$$0 \leq \mathbf{E}\left[(1 \pm X)(c_0 + c_1 X + \cdots + c_d X^d)^2\right] = \sum_{j=0}^{d} \sum_{k=0}^{d} (\mu_{j+k} \pm \mu_{j+k+1}) c_j c_k,$$

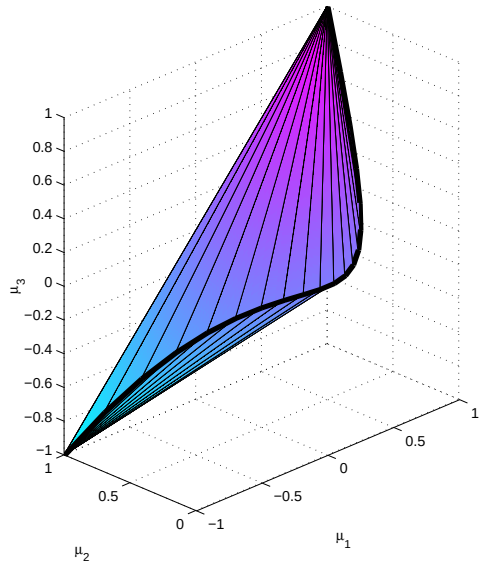

Figure 2: Set of valid moments (μ_1, μ_2, μ_3) of a probability measure on $[-1, 1]$. This is the convex hull of the moment curve (t, t^2, t^3), for $-1 \leq t \leq 1$. An explicit SDP representation is given in (5).

where the first inequality follows since $1 \pm X$ is always nonnegative, since X is supported on $[-1, 1]$. Notice the similarities (in fact, the duality) with the conditions for polynomial nonnegativity discussed in a previous lecture.

2.2 Convex hull of the moment curve

An appealing geometric interpretation of the set of valid moments is in terms of the so-called *moment curve*, which is the parametric curve in $\mathbb{R}^{d+1}$ given by $t \mapsto (1, t, t^2, \ldots, t^d)$. Indeed, it is easy to see that every point on the curve can be associated to a Dirac measure where all the probability is concentrated on a given point, since if the pdf of X is $\delta(x - a)$, then

$$\mu_k = \mathbf{E}[X^k] = a^k.$$

Thus, every finite (or infinite) measure on the interval corresponds to a point in the convex hull. In Figure 2 we present an illustration of the set of valid moments, for the case $d = 3$.

Remark 4. *Consider again the situation described in Remark 2. By restricting to even measures (i.e., $\mu(x) = \mu(-x)$), we can see that the moment curve corresponds to the half-parabola $C = \{(a, b) : a \geq 0, a^2 = b\}$. The point $(0, 1)$ is in the closure of* conv(C)*, but it is not in* conv(C) *(since the convex hull is not closed). Notice that this situation cannot happen if the measure is supported on a compact set, since in this case the set of valid moments is always compact.*

3 Bridging the gap

What to do in the cases where the set of nonnegative polynomials is no longer equal to the SOS ones? As we will see in much more detail later, it turns out that we can approximate *any* semialgebraic problem (including the simple case of a single polynomial being nonnegative) by sum of squares techniques.

As a preview, and a hint at some of the possibilities, let's consider how to prove nonnegativity of a particular polynomial which is not a sum of squares. Recall that the Motzkin polynomial was defined as:

$$M(x,y) = x^4y^2 + x^2y^4 + 1 - 3x^2y^2.$$

and is a nonnegative polynomial that is not SOS. We can try multiplying it by another polynomial which is known to be positive, and check whether the resulting product is SOS. For instance, for the Motzkin example, multiplying by the factor $(x^2 + y^2)$ we can find the decomposition

$$(x^2 + y^2) \cdot M(x,y) = y^2(1 - x^2)^2 + x^2(1 - y^2)^2 + x^2y^2(x^2 + y^2 - 2)^2,$$

which clearly certifies that $M(x,y) \geq 0$.

More details will follow...

References

[Akh65] N. I. Akhiezer. *The classical moment problem.* Hafner Publishing Company, New York, 1965.

[AKP11] A. A. Ahmadi, M. Krstic, and P. A. Parrilo. A globally asymptotically stable polynomial vector field with no polynomial Lyapunov function. In *IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, pages 7579–7580. IEEE, 2011.

[DPS02] A. C. Doherty, P. A. Parrilo, and F. M. Spedalieri. Distinguishing separable and entangled states. *Physical Review Letters*, 88(18), 2002.

[DPS04] A. C. Doherty, P. A. Parrilo, and F. M. Spedalieri. Complete family of separability criteria. *Physical Review A*, 69:022308, 2004.

[Gur03] L. Gurvits. Classical deterministic complexity of Edmonds' problem and quantum entanglement. In *STOC '03: Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 10–19, New York, NY, USA, 2003. ACM.

[Kha92] H. Khalil. *Nonlinear Systems.* Macmillan Publishing Company, 1992.

[Par00] P. A. Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization.* PhD thesis, California Institute of Technology, May 2000.

[PP02] A. Papachristodoulou and S. Prajna. On the construction of Lyapunov functions using the sum of squares decomposition. In *Proceedings of the 41th IEEE Conference on Decision and Control*, 2002.

[PPP05] S. Prajna, A. Papachristodoulou, and P. A. Parrilo. *SOSTOOLS: Sum of squares optimization toolbox for MATLAB*, 2002-05. Available from `http://www.cds.caltech.edu/sostools` and `http://www.mit.edu/~parrilo/sostools`.

MIT 6.7230 - Algebraic techniques and semidefinite optimization March 20, 2024

Lecture 10

Lecturer: Pablo A. Parrilo *Scribe: ???*

In previous lectures, we described necessary conditions for the existence of a nonnegative measure with given moments. In the univariate case, these conditions were also sufficient. We revisit first a classical algorithm to effectively obtain this measure.

1 Recovering a measure from moments

We review next a classical method for producing a univariate atomic measure with a given set of moments (e.g., [ST43, Dev86]). Other similar variations of this method are commonly used in signal processing, e.g., Pisarenko's harmonic decomposition method, where we are interested in producing a superposition of sinusoids with a given covariance matrix. This technique (or essentially similar ones) is known under a variety of names, such as Prony's method, or the Vandermonde decomposition of a Hankel matrix.

Consider the set of moments $(\mu_0, \mu_1, \ldots, \mu_{2n-1})$ for which we want to find an associated nonnegative measure, supported on the real line. In general, there are infinitely many measures that will exact match those moments.

One possible approach, on which we will not elaborate here, is to pick the measure with *maximum entropy* that matches the given moments. This corresponds to the Gaussian case (if moments up to second order are given) or exponential families (in the general case).

The approach we follow here is to compute a *discrete* (finitely supported) measure, of the form $\sum_{i=1}^{n} w_i \delta(x - x_i)$. For this, consider the linear system

$$\begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_{n-1} \\ \mu_1 & \mu_2 & \cdots & \mu_n \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n-1} & \mu_n & \cdots & \mu_{2n-2} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{bmatrix} = - \begin{bmatrix} \mu_n \\ \mu_{n+1} \\ \vdots \\ \mu_{2n-1} \end{bmatrix}. \tag{1}$$

The Hankel matrix on the left-hand side of this equation is the one that appeared earlier as a sufficient condition for the moments to represent a nonnegative measure. The linear system in (1) has a unique solution if the matrix is positive definite. In this case, we let x_i be the roots of the univariate polynomial

$$x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0 = 0.$$

These roots are all real and distinct (why?). Given the locations x_i, we can then obtain the corresponding weights w_i by solving the nonsingular Vandermonde system given by

$$\sum_{i=1}^{n} w_i x_i^j = \mu_j \qquad (0 \le j \le n-1).$$

In the exercises, you will have to prove that this method actually works (i.e., the x_i are real and distinct, the w_i are nonnegative, and the moments are the correct ones).

Example 1. *Let's find a nonnegative measure whose first six moments are given by* $(1, 1, 2, 1, 6, 1)$. *The solution of the linear system (1) yields the polynomial*

$$x^3 - 4x^2 - 9x + 16 = 0,$$

whose roots are -2.4265, 1.2816, *and* 5.1449. *The corresponding weights are* 0.0772, 0.9216, *and* 0.0012, *respectively.*

Example 2. *We outline here a "stylized" application of these results. Consider a time-domain signal that is the sum of* k *Dirac functions, i.e.,* $f(x) := \sum_{i=1}^{k} w_i \delta(x - x_i)$, *where the* $2k$ *parameters* w_i, x_i *are unknown. By the results above, it is enough to obtain* $2k$ *linear functionals on the signal (namely, the moments* $\mu_i := \int x^i f(x) dx$*) to fully recover it from the measurements. Indeed, the signal can always be exactly reconstructed from these* $2k$ *moments, by using the algorithm described above. Notice that the nonnegativity assumption on the weights* w_i *is not critical, and can easily be removed.*

More realistic, but essentially similar results can be obtained by considering signals that are sums of (possibly damped) sinusoids of different frequencies. This viewpoint has a number of interesting connections with error-correcting codes (in particular, interpolation-based codes such as Reed-Solomon), as well as the recent "compressed sensing" results.

Remark 3. *As described, the measure recovery method described always works correctly, provided the computations are done in exact arithmetic. In most practical applications, it is necessary or convenient to use floating-point computations. Furthermore, in many settings such as optimization the moment information may be noisy, and therefore the matrices may contain some (hopefully small) perturbations from their nominal values. For these reasons, it is of interest to understand sensitivity issues, both at the level of what is intrinsic about the problem (conditioning), and about the specific algorithm used (numerical stability).*

As described, the technique described above can run into numerical difficulties. On the conditioning side, it is well-known that from the numerical viewpoint, the monomial basis (with respect to which we are taking moments) is a "bad" basis for the space of polynomials. On the numerical stability side, the algorithm above does a number of inefficient calculations, such as explicitly computing the coefficients c_i *of the polynomial corresponding to the support of the measure. A better approach involves directly computing the nodes* x_i *as the generalized eigenvalues of a matrix pencil. Some of these issues will be explored in more detail in the exercises.*

2 A probabilistic interpretation

We also mention here an appealing probabilistic interpretation of the dual (5), commonly used in integer and quadratic programming or game theory, and developed by Lasserre in the polynomial optimization case [Las01]. Consider as before the problem of minimizing a polynomial. Now, rather than looking directly for the minimizer $x_\star$ in $\mathbb{R}^n$, let's "relax" our notion of solution to allow for probabilities densities μ on $\mathbb{R}^n$, and replace the objective function by its natural generalization $\mathbf{E}_\mu[p(x)] = \int p(x) d\mu$. Since

$$\min_\mu \mathbf{E}_\mu[p(x)] \leq \mathbf{E}_{\delta_{x_\star}}[p(x)] = p(x_\star),$$

it clearly holds that the new objective is never larger than the original one, since we are making the feasible set bigger.

This change makes the problem convex (trivially so), although infinite-dimensional. To produce a finite dimensional approximation (which may or may not be exact), we rewrite the objective function in terms of the moments of the measure μ, and write valid semidefinite contraints for the moments μ_k.

3 Duality and complementary slackness

What is the relationship between this classical method and semidefinite programming duality? Recall our approach to minimizing a polynomial $p(x)$ by computing

$$\max \gamma \qquad \text{s.t.} \qquad p(x) - \gamma \quad \text{is SOS.} \tag{2}$$

The corresponding dual is

$$\min_{L} L[p] \qquad \text{s.t.} \qquad \begin{cases} L[q^2] \geq 0 \quad \forall q \in \mathbb{R}[x]_d, \\ L[1] = 1, \end{cases} \tag{3}$$

where $L[\cdot]$ is an element of the dual space $\mathbb{R}[x]^*_{2d}$ (a "Riesz functional" or "pseudoexpectation"). Here, L is constrained to be nonnegative on squares, and normalized. By choosing appropriate bases of $\mathbb{R}[x]$ and $\mathbb{R}[x]^*$, the constraints can be expressed as a standard SDP; see below for details.

On the primal side, the SDP formulation of the univariate optimization problem is given by

$$\max \gamma \qquad \text{s.t.} \qquad \begin{cases} Q_{00} + \gamma = p_0 \\ \displaystyle\sum_{j,k:j+k=i} Q_{jk} = p_i \qquad i = 1, \ldots, 2d \\ Q \succeq 0 \end{cases} \tag{4}$$

and its dual

$$\min \sum_{i=0}^{2d} p_i \mu_i \qquad \text{s.t.} \quad M(\mu) := \begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_d \\ \mu_1 & \mu_2 & \cdots & \mu_{d+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_d & \mu_{d+1} & \cdots & \mu_{2d} \end{bmatrix} \succeq 0, \quad \mu_0 = 1. \tag{5}$$

Notice the direct relationship between equations (2)-(4) and (3)-(5). Indeed, to obtain the corresponding SDPs, on the primal side we use a monomial basis and the representation $p(x) - \gamma = [x]_d^T Q [x]_d$, while on the dual we have $L[q] = \sum_{k=0}^{2d} c_k \mu_k$, where $q(x) = \sum_{k=0}^{2d} c_k x^k$.

When is the relaxation exact? If this relaxation is exact (i.e., the optimal γ is equal to the optimal value of the polynomial) then at optimality, we necessarily have $p(x_\star) - \gamma_\star = \sum_i g_i^2(x_\star)$. This implies that all the g_i must vanish at the optimal point. We can thus obtain the optimal value by looking at the roots of the polynomials $g_i(x)$.

However, it turns out that if we are simultaneously solving the primal and the dual SDPs (as most modern interior point solvers) this is unnecessary, since from complementary slackness we can extract almost all the information needed. In particular, notice that if we have

$$p(x) - \gamma = [x]_d^T Q [x]_d = 0$$

then necessarily $Q \cdot [x]_d = 0$.

At optimality, complementarity slackness holds, i.e., the product of the primal and dual matrices vanishes. We have then $M(\mu) \cdot Q = 0$. Assume that the leading $k \times k$ submatrix of $M(\mu)$ is nonsingular. Then, the procedure described in Section 1 gives a k-atomic measure, with support in the minimizers of $p(x)$. Generically, this matrix $M(\mu)$ will be rank one, which will correspond to the case of a unique optimal solution.

Remark 4. *Unlike the univariate case, a multivariate polynomial that is bounded below may not achieve its minimum. A well-known example is $p(x, y) = x^2 + (1 - xy)^2$, which clearly satisfies $p(x, y) \geq 0$. Since $p(x, y) = 0$ would imply $x = 0$ and $1 - xy = 0$ (which is impossible), this value cannot be achieved. However, we can get arbitrarily close, since $p(\epsilon, 1/\epsilon) = \epsilon^2$, for any $\epsilon > 0$.*

4 Multivariate case

We have seen previously that in the multivariate case, it is no longer the case that nonnegative polynomials are always sums of squares. The corresponding result on the dual side is that the set of valid moments is no longer described by the "obvious" semidefinite constraints, obtained by considering the expected value of squares (even if we require strict positivity).

Example 5 ("Dual Motzkin")**.** *Consider the existence of a probability measure on $\mathbb{R}^2$, that satisfies the moment constraints:*

$$\begin{aligned} E[1] = E[X^4Y^2] = E[X^2Y^4] &= 1, \\ E[X^2Y^2] &= 2, \\ E[XY] = E[XY^2] = E[X^2Y] = E[X^2Y^3] = E[X^3Y^2] = E[X^3Y^3] &= 0. \end{aligned} \tag{6}$$

The "obvious" nonnegativity constraints are satisfied, since

$$E[(a + bXY + cXY^2 + dX^2Y)^2] = a^2 + 2b^2 + c^2 + d^2 \geq 0.$$

However, it turns out that these conditions are only necessary, but not sufficient. This can be seen by computing the expectation of the Motzkin polynomial (which is nonnegative), since in this case we have

$$E[X^4Y^2 + X^2Y^4 + 1 - 3X^2Y^2] = 1 + 1 + 1 - 6 = -3,$$

thus proving that no nonnegative measure with the given moments can exist.

5 Density results

Recent results by Blekherman [Ble06] give quantitative bounds on the relative density of the cone of sum of squares versus the cone of nonnegative polynomials. Concretely, in [Ble06] it is proved that a suitably normalized section of the cone of positive polynomials $\tilde{P}_{n,2d}$ satisfies

$$c_1 n^{-\frac{1}{2}} \leq \left(\frac{\operatorname{Vol} \tilde{P}_{n,2d}}{\operatorname{Vol} B_M} \right)^{\frac{1}{D_M}} \leq c_2 n^{-\frac{1}{2}},$$

while the corresponding expression for the section of the cone of sum of squares $\tilde{\Sigma}_{n,2d}$ is

$$c_3 n^{-\frac{d}{2}} \leq \left(\frac{\operatorname{Vol} \tilde{\Sigma}_{n,2d}}{\operatorname{Vol} B_M} \right)^{\frac{1}{D_M}} \leq c_4 n^{-\frac{d}{2}},$$

where c_1, c_2, c_3, c_4 depend on d only (explicit expressions are available), $D_M = \binom{n+2d}{2d} - 1$, and B_M is the unit ball in $\mathbb{R}^{D_M}$.

These expressions show that for fixed d, as $n \to \infty$ the volume of the set of sum of squares becomes vanishingly small when compared to the nonnegative polynomials.

`Show the values of the actual bounds, for reasonable dimensions`	ToDo

References

[Ble06] G. Blekherman. There are significantly more nonegative polynomials than sums of squares. *Israel Journal of Mathematics*, 153(1):355–380, 2006.

[Dev86] L. Devroye. *Nonuniform random variate generation*. Springer-Verlag, New York, 1986.

[Las01] J. B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM J. Optim.*, 11(3):796–817, 2001.

[ST43] J.A. Shohat and J.D. Tamarkin. *The Problem of Moments*. American Mathematical Society Mathematical surveys, vol. II. American Mathematical Society, New York, 1943.

MIT 6.7230 - Algebraic techniques and semidefinite optimization April 3, 2024

Lecture 11

Lecturer: Pablo A. Parrilo *Scribe: ???*

Today we introduce the first basic elements of algebraic geometry, namely ideals and varieties over the complex numbers. This dual viewpoint (ideals for the algebra, varieties for the geometry) is enormously powerful, and will help us later in the development of methods for solving polynomial equations. We also present the notion of quotient rings, which are very natural when considering functions defined on algebraic varieties (e.g., in polynomial optimization problems with equality constraints). Finally, we begin our study of Groebner bases, by defining the notion of term orders. A superb introduction to algebraic geometry, emphasizing the computational aspects, is the textbook of Cox, Little, and O'Shea [CLO97]. Another recommended reference is the introductory-level book of Hassett [Has07].

1 Polynomial ideals

For notational simplicity, we use $\mathbb{C}[\mathbf{x}]$ to denote the polynomial ring in n variables $\mathbb{C}[x_1, \ldots, x_n]$. Specializing the general definition of an ideal to a polynomial ring, we have the following:

Definition 1. *A subset $I \subset \mathbb{C}[\mathbf{x}]$ is an* ideal *if it satisfies:*

1. $0 \in I$.
2. *If $a, b \in I$, then $a + b \in I$.*
3. *If $a \in I$ and $b \in \mathbb{C}[\mathbf{x}]$, then $a \cdot b \in I$.*

The two most important examples of polynomial ideals for our purposes are the following:

- The set of polynomials that vanish in a given set $S \subset \mathbb{C}^n$, i.e.,

$$\mathbf{I}(S) := \{f \in \mathbb{C}[\mathbf{x}] : f(a_1, \ldots, a_n) = 0 \qquad \forall (a_1, \ldots, a_n) \in S\},$$

  is an ideal, called the *vanishing ideal* of S.

- The ideal generated by a finite set of polynomials $\{f_1, \ldots, f_s\}$, defined as

$$\langle f_1, \ldots, f_s \rangle := \{f \mid f = g_1 f_1 + \cdots + g_s f_s, \quad g_i \in \mathbb{C}[\mathbf{x}]\}. \tag{1}$$

An ideal is *finitely generated* if it can be written as in (1) for some finite set of polynomials $\{f_1, \ldots, f_s\}$. An ideal is called *principal* if it can be generated by a single polynomial. The intersection of two ideals is again an ideal. What about the union of ideals?

Example 2. *In the univariate case (i.e., the polynomial ring is $\mathbb{C}[x]$), every ideal is principal.*

One of the most important facts about polynomial ideals is Hilbert's finiteness theorem:

Theorem 3 (Hilbert Basis Theorem)**.** *Every polynomial ideal in $\mathbb{C}[\mathbf{x}]$ is finitely generated.*

We will present a proof of this after learning about Groebner bases.
From the computational viewpoint, two very natural questions about ideals are the following:

- Given a polynomial $p(x)$, how to decide if it belongs to a given ideal?
- How to find a "convenient" representation of an ideal? What does "convenient" mean?

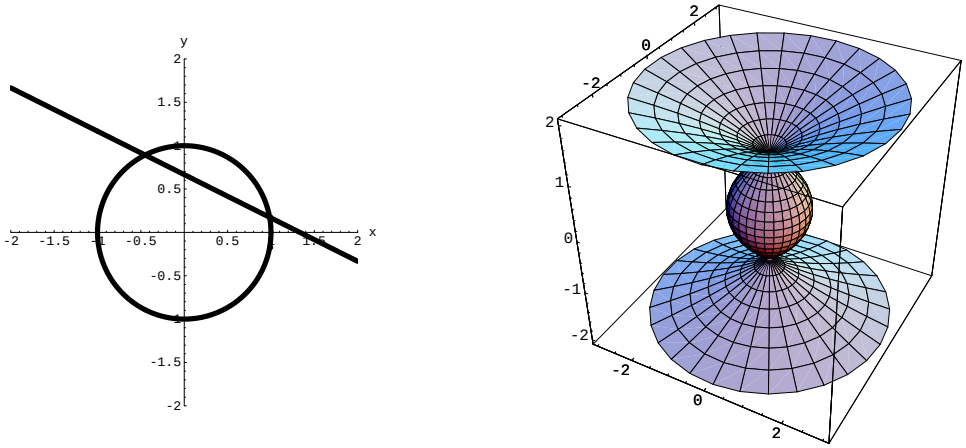

Figure 1: Two algebraic varieties. The one on the left is defined by the equation $(x^2+y^2-1)(3x+6y-4)=0$. The one on the right is a quartic surface, defined by $1-x^2-y^2-2z^2+z^4=0$.

2 Algebraic varieties

An (affine) algebraic variety is the zero set of a finite collection of polynomials (see formal definition below). The word "affine" here means that we are working in the standard affine space, as opposed to projective space, where we identify $x, y \in \mathbb{C}^n$ if $x = \lambda y$ for some $\lambda \neq 0$.

Definition 4. *Let* $\{f_1, \ldots, f_s\}$ *be a finite set of polynomials in* $\mathbb{C}[\mathbf{x}]$. *Let* $\mathbf{V}$ *be*

$$\mathbf{V}(f_1, \ldots, f_s) := \{(a_1, \ldots, a_n) \in \mathbb{C}^n : f_i(a_1, \ldots, a_n) = 0 \qquad 1 \leq i \leq s\}.$$

We call $\mathbf{V}(f_1, \ldots, f_s)$ *the* affine variety *defined by* $f_1, \ldots, f_s$.

A simple example of a variety is a (complex) affine subspace, that corresponds to the vanishing of a finite collection of affine polynomials. A few additional examples of varieties are shown in Figure 1.

It is not too hard to show that *finite* unions and intersections of algebraic varieties are again algebraic varieties. But what about the infinite case? For infinite unions, simple counterexamples (e.g., the set of integers) show that in general these are not algebraic varieties. For infinite intersections, however, the answer is more interesting...

Lemma 5. *The (arbitrary) intersection of algebraic varieties is an algebraic variety.*

To see this, notice that for any index set $\mathcal{S}$, we have $\cap_{k\in\mathcal{S}}\mathbf{V}(I_k) = \mathbf{V}(\sum_{k\in\mathcal{S}} I_k)$. Then, using Hilbert's basis theorem, the ideal $\sum_{k\in\mathcal{S}} I_k$ is finitely generated, and thus the intersection is indeed an algebraic variety.

Recall the following standard definition:

Definition 6. *A* topology *is a collection* $\mathcal{T}$ *of subsets of a set* S *satisfying the following properties:*

1. *The empty set* $\emptyset$ *and* S *are in* $\mathcal{T}$.
2. *The intersection of a finite collection of sets from* $\mathcal{T}$ *is again in* $\mathcal{T}$.
3. *The union of any collection of sets from* $\mathcal{T}$ *is again in* $\mathcal{T}$.

A subset of V is open *if it is in $\mathcal{T}$. A subset of V is* closed *if its complement (in S) is open.*

Notice that finite unions of closed sets are closed, and so are arbitrary intersections. Thus, it follows from our earlier discussion that one can define a topology on $\mathbb{C}^n$ (known as the *Zariski topology*) where the closed sets are the algebraic varieties. The Zariski topology has many interesting properties (sometimes counterintuitive), and we will explore it in more detail in the exercises.

Perhaps the most natural question about algebraic varieties is the following:

- Given a variety V, how to decide if it is nonempty?

Let's start connecting ideals and varieties. Consider a finite set of polynomials $\{f_1, \ldots, f_s\}$. We already know how to generate an ideal, namely $\langle f_1, \ldots, f_s \rangle$. However, we can also look at the corresponding variety $\mathbf{V}(f_1, \ldots, f_s)$. Since this variety is a subset of $\mathbb{C}^n$, we can form the corresponding vanishing ideal, $\mathbf{I}(\mathbf{V}(f_1, \ldots, f_s))$. How do these two ideals related to each other? Is it always the case that

$$\langle f_1, \ldots, f_s \rangle = \mathbf{I}(\mathbf{V}(f_1, \ldots, f_s)),$$

and if it is not, what are the reasons? The answer to these questions (and more) will be given by another famous result by Hilbert, known as the Nullstellensatz.

3 Quotient rings

Whenever we have an ideal in a ring, we can immediately define a notion of equivalence classes, where we identify two elements in the ring if and only if their difference is in the ideal.

Example 7. *Recall that a simple example of an ideal in the ring $\mathbb{Z}$ was the set of even integers. By identifying two integers if their difference is even, we partition $\mathbb{Z}$ into two equivalence classes, namely the even and the odd numbers. More generally, if the ideal is given by the integer multiples of a given number m, then $\mathbb{Z}$ can be partitioned into m equivalence classes.*

We can do this for the polynomial ring $\mathbb{C}[\mathbf{x}]$, and any ideal I.

Definition 8. *Let $I \subset \mathbb{C}[\mathbf{x}]$ be an ideal, and let $f, g \in \mathbb{C}[\mathbf{x}]$. We say f and g are* congruent *modulo I, written*

$$f \equiv g \mod I,$$

if $f - g \in I$.

It is easy to show that this is an equivalence relation, i.e., it is reflexive, symmetric, and transitive. Thus, this partitions $\mathbb{C}[\mathbf{x}]$ into equivalence classes, where two polynomials are "the same" if their difference belongs to the ideal. This allows us to define the quotient ring:

Definition 9. *The* quotient $\mathbb{C}[\mathbf{x}]/I$ *is the set of equivalence classes for congruence modulo I.*

The quotient $\mathbb{C}[\mathbf{x}]/I$ inherits the ring structure of $\mathbb{C}[\mathbf{x}]$, with the natural operations. Thus, with these operations now defined between equivalence classes, $\mathbb{C}[\mathbf{x}]/I$ becomes a ring, known as the *quotient ring*.

Quotient rings are particularly useful when considering a polynomial function $p(x)$ over the algebraic variety defined by $g_i(x) = 0$. Notice that if we define the ideal $I = \langle g_i \rangle$, then *any* polynomial q that is congruent with p modulo I takes exactly the same values in the variety.

4 Monomial orderings

In order to begin studying "nice" bases for ideals, we need a way of ordering monomials. In the univariate case, this is straightforward, since we can define $x^a \succ x^b$ as being true if and only if $a > b$. In the multivariate case, there are a lot more options.

We also want the ordering structure to be consistent with polynomial multiplication. This is formalized in the following definition.

Definition 10. *A* monomial ordering *on* $\mathbb{C}[\mathbf{x}]$ *is a binary relation* $\succ$ *on* $\mathbb{Z}_+^n$ *(i.e., the monomial exponents), such that:*

1. *The relation* $\succ$ *is a total ordering.*
2. *If* $\alpha \succ \beta$*, and* $\gamma \in \mathbb{Z}_+^n$*, then* $\alpha + \gamma \succ \beta + \gamma$*.*
3. *The relation* $\succ$ *is a well-ordering (every nonempty subset has a smallest element).*

One of the simplest examples of a monomial ordering is the *lexicographic* ordering, where $\alpha \succ_{\text{lex}} \beta$ if the left-most nonzero entry of $\alpha - \beta$ is positive. We will see a few other examples of monomial orderings in the next lecture.

References

[CLO97] D. A. Cox, J. B. Little, and D. O'Shea. *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra.* Springer, 1997.

[Has07] B. Hassett. *Introduction to algebraic geometry.* Cambridge University Press, 2007.

MIT 6.7230 - Algebraic techniques and semidefinite optimization April 5, 2024

Lecture 12

Lecturer: Pablo A. Parrilo *Scribe: ???*

After a brief review of monomial orderings, we develop the basic ideas of Groebner bases, followed by examples and applications. For background and much more additional material, we recommend the textbook of Cox, Little, and O'Shea [CLO97]. Other good, more specialized references are [AL94, BW93, KR00].

1 Monomial orderings

Recall from last lecture the notion of a monomial ordering:

Definition 1. *A* monomial ordering *on* $\mathbb{C}[\mathbf{x}]$ *is a binary relation* $\succ$ *on* $\mathbb{Z}_+^n$ *(i.e., the monomial exponents), such that:*

1. *The relation* $\succ$ *is a total ordering.*
2. *If* $\alpha \succ \beta$*, and* $\gamma \in \mathbb{Z}_+^n$*, then* $\alpha + \gamma \succ \beta + \gamma$*.*
3. *The relation* $\succ$ *is a well-ordering (every nonempty subset has a smallest element).*

There are several term orderings of interest in computational algebra. Among them, we mention:

- Lexicographic ("dictionary"). Here $\alpha \succ_{\text{lex}} \beta$ if the left-most nonzero entry of $\alpha - \beta$ is positive. Notice that a particular order of the variables is assumed, and by changing this, we obtain $n!$ nonequivalent lexicographic orderings.
- Graded lexicographic. Sort first by total degree, then lexicographic, i.e., $\alpha \succ_{\text{grlex}} \beta$ if $|\alpha| > |\beta|$, or if $|\alpha| = |\beta|$ and $\alpha \succ_{\text{lex}} \beta$.
- Graded reverse lexicographic. Here $\alpha \succ_{\text{grevlex}} \beta$ if $|\alpha| > |\beta|$, or if $|\alpha| = |\beta|$ and the right-most nonzero entry of $\alpha - \beta$ is negative. This ordering, although somewhat nonintuitive, has some desirable computational properties.
- General matrix orderings. Described by a weight matrix $W \in \mathbb{R}^{k \times n} (k \leq n)$, where $\alpha \succ_W \beta$ if $(W\alpha) \succ_{\text{lex}} (W\beta)$. For W to correspond to a monomial ordering as defined, the first nonzero entry on each column must be positive.

It turns out that every monomial ordering can be described by an associated matrix W, i.e., every monomial ordering is a matrix ordering. What are the matrices corresponding to the first three orderings described?

Example 2. *Consider the polynomial ring* $\mathbb{C}[x, y]$*. In the lexicographic ordering (*$\prec_{\text{lex}}$*) discussed, we have:*

$$1 \prec y \prec y^2 \prec \cdots \prec x \prec xy \prec xy^2 \prec \cdots \prec x^2 \prec x^2 y \prec x^2 y^2 \prec \cdots,$$

while for the other two orderings ($\prec_{\text{grlex}}$ *and* $\prec_{\text{grevlex}}$*), which in the special case of two variables coincide, we have:*

$$1 \prec y \prec x \prec y^2 \prec xy \prec x^2 \prec y^3 \prec xy^2 \prec x^2 y \prec x^3 \prec \cdots.$$

Picture comparing different orderings	ToDo

Example 3. *Consider the monomials $\alpha = x^3y^2z^8$ and $\beta = x^2y^9z^2$. If the variables are ordered as (x, y, z), we have*

$$\alpha \succ_{\text{lex}} \beta, \qquad \alpha \succ_{\text{grlex}} \beta, \qquad \alpha \prec_{\text{grevlex}} \beta.$$

Notice that $x \succ y \succ z$ for all three orderings.

2 Groebner bases

2.1 Monomial ideals

Before studying general ideals, it is convenient to introduce first a special class, known as *monomial ideals*.

Definition 4. *A* monomial ideal *is a polynomial ideal that can be generated by monomials.*

What are the possible monomials that belong to a given monomial ideal? Since $x^\alpha \in I \Rightarrow x^{\alpha+\beta} \in I$ for $\beta \geq 0$, we have that these sets are "closed upwards."

Picture of monomial ideals	ToDo

Furthermore, a polynomial belongs to a monomial ideal I if and only if it all its terms are in I.

Theorem 5 (Dickson's lemma)**.** *Every monomial ideal is finitely generated.*

We consider next a special monomial ideal, associated to every polynomial ideal I. From now on, we assume a fixed monomial ordering (e.g., graded reverse lexicographic), and denote by $in(f)$ the "largest" (or *leading*) monomial appearing in the polynomial $f \neq 0$.

Definition 6. *Consider an ideal $I \subset \mathbb{C}[x]$, and a fixed monomial ordering. The* initial ideal *of I, denoted $in(I)$, is the monomial ideal generated by the leading monomials of all the elements in I, i.e.,*

$$in(I) := \langle in(f) : f \in I \setminus \{0\} \rangle.$$

A monomial x^α is called standard*, if it does not belong to the initial ideal $in(I)$.*

2.2 Groebner bases

Given an ideal $I = \langle f_1, \ldots, f_s \rangle$, we can construct two monomial ideals associated with it. On the one hand, we have the initial ideal $in(I)$, previously defined. However, we can also consider the monomial ideal generated by the leading monomials of the generators, i.e., $\langle in(f_1), \ldots, in(f_s) \rangle$. Although we always have $\langle in(f_1), \ldots, in(f_s) \rangle \subset in(I)$, in general these two monomial ideals are distinct.

Example 7. *Consider the ideal $I = \langle x^3-1, x^2+1 \rangle$. Since $1 = \frac{1}{2}(x-1)(x^3-1) - \frac{1}{2}(x^2-x-1)(x^2+1)$, we have $1 \in I$, and thus $in(I) = I = \mathbb{C}[x]$. On the other hand, $1 \notin \langle x^3, x^2 \rangle$.*

However, it may be possible to produce a set of generators for which these two ideals are the same. This is exactly the notion of a *Groebner basis*.

Definition 8. *Consider the polynomial ring $\mathbb{C}[x]$, with a fixed monomial ordering, and an ideal I. A finite set of polynomials $\{g_1, \ldots, g_s\} \subset I$ is a* Groebner basis *of I if the initial ideal of I is generated by the leading terms of the g_i, i.e.,*

$$in(I) = \langle in(g_1), \ldots, in(g_s) \rangle. \tag{1}$$

Theorem 9. *Every ideal I has a Groebner basis G. Furthermore, $I = \langle g_1, \ldots, g_s \rangle$.*

The previous theorem essentially establishes Hilbert's finiteness result, and gives an explicit characterization of a finite generating set for the ideal I. Furthermore, since there are explicit algorithms to compute Groebner bases, this is a constructive version of this theorem.

Even though the monomial ordering is fixed, Groebner bases as defined are not unique (why?). This can be easily fixed, by refining the concept to the so-called *reduced* Groebner bases, which are uniquely defined.

There are several possible algorithms to effectively compute Groebner bases. The traditional one is *Buchberger's algorithm*, developed by Bruno Buchberger around 1965, and many variants have been proposed since. There are also several newer methods, based on sparse linear algebra, that in some instances can significantly outperform the Buchberger approach. Good specialized programs for Groebner bases calculations (and much more) are CoCoA[CoC], Macaulay2 [GS] and Singular [GPS05].

Example 10. *Consider the ideal $I = \langle x^2 + y^2 - 2, x^2 - y^2 \rangle$. The associated variety is finite, equal to the four points $(\pm 1, \pm 1)$. It is easy to see that the given generators are* not *a Groebner basis of I (why?). Computing a Groebner basis (e.g., using the lexicographic order) yields $gb(I) = \{x^2 - 1, y^2 - 1\}$. The standard monomials are $\{1, x, y, xy\}$.*

2.3 Quotients and normal forms

Recall that given an ideal $I \subset \mathbb{C}[x]$, we defined the quotient ring $\mathbb{C}[x]/I$ as the set of equivalence classes modulo the ideal. For computational purposes, we want a "good" representation of these classes, and in particular, a way to provide a "unique representative" to every polynomial. This can in fact be easily done once we have computed a Groebner basis. To each polynomial $p \in \mathbb{C}[x]$, we can associate a unique "normal form", defined below.

Lemma 11. *Let G be a Groebner basis of the ideal $I \subset \mathbb{C}[x]$. Given any $p \in \mathbb{C}[x]$, there exists a unique polynomial $\bar{p}$, called the* normal form *of p, such that*

1. *The polynomials p and $\bar{p}$ are congruent mod I, i.e., $p - \bar{p} \in I$.*
2. *Only standard monomials appear in $\bar{p}$.*

Notice that we have $p = q_1 g_1 + \cdots + q_s g_s + \bar{p}$. Thus, the normal form can be interpreted as the "remainder" after a division-like process by the generators g_i. The key property (1) guarantees that this remainder is uniquely defined.

As a consequence of this, we can solve the ideal membership problem: to check if a polynomial $p(x)$ is in a given ideal I, compute a Groebner basis G of I, and check if the normal form of $p(x)$ is the zero polynomial, i.e., $p \in I \Leftrightarrow \bar{p} = 0$.

3 Applications and examples

Groebner bases enable the algorithmic solution of many problems in computational algebraic geometry. We discuss some these below.

- Ideal membership. As we have seen, given an ideal I and a polynomial p, we can check if $p \in I$ by computing the normal form of p.
- Consistency of polynomial equations. Consider a finite set of polynomial equations $\{f_i = 0\}$, and let $I = \langle f_i \rangle$ be the corresponding ideal. By the Nullstellensatz, the given equations are infeasible if and only if $\{1\}$ is the reduced Groebner basis of I.
- Radical membership. Consider an ideal $I = \langle f_1, \ldots, f_s \rangle \subset \mathbb{C}[x]$, and a polynomial p, for which we want to check whether $p \in \sqrt{I}$. Since $\sqrt{I}$ is also an ideal, we could compute a Groebner basis for it, and then reduce the problem to the previous one. However, it is often more efficient to instead use the following result ("Rabinowitsch's trick" [Rab30]):

  $$p \in \sqrt{I} \qquad \Leftrightarrow \qquad 1 \in \langle f_1, \ldots, f_s, 1 - yp \rangle,$$

  where y is a (new) additional variable.
- Elimination. For notational simplicity, consider an ideal $I \subset \mathbb{C}[x, y, z]$. Suppose that we want to compute all the polynomials in I, that *do not* depend on the variable z, i.e., $I \cap \mathbb{C}[x, y]$. Geometrically, this elimination of variables corresponds to (the Zariski closure of) the projection of the corresponding variety into (x, y). This intersection (or projection) can be easily obtained, by computing a Groebner basis G of I with respect to a lexicographic (or elimination) ordering. The corresponding ideal is then generated by $G \cap \mathbb{C}[x, y]$.

4 Implementations

Groebner bases techniques have been implemented in many software packages. We show below a simple example in both Maple and Macaulay2. Maple is a general numeric and symbolic computer algebra package, easily available from MIT. For specific algebraic-geometric computations, the specialized software Macaulay2 [GS] may be a more flexible choice.

```
> with(Groebner):
> tod := plex(z,y,x);
                              tod := plex(z, y, x)

> Id := {x*y-z,y*z-x,z*x-y};
                      Id := {y z - x, x z - y, x y - z}

> G := Basis(Id,tod);
                        3       2          2    2
                G := [x  - x, x  y - y, -x  + y , -x y + z]

> NormalForm(x*y*z+4*x*z^4,G,tod);
                                    2
                                   x  + 4 x
```

After loading the Groebner package, we define a pure lexicographic order, where $x \prec y \prec z$. The third and fourth lines define an ideal I by its generators, and compute a Groebner basis with respect to the specified term order. The leading monomials of the elements of this basis are $\{x^3, x^2y, y^2, z\}$, respectively. Finally, the last line computes the normal form (i.e., a canonical representative in $\mathbb{C}[x]/I$) of the polynomial $xyz + 4xz^4$. Notice that only standard monomials appear in the normal form (i.e., they are not divisible by the leading monomials of the basis).

Here are the same calculations again, but using Macaulay2 instead:

```
i1 : R = QQ[z,y,x, MonomialOrder => Lex]

o1 = R

o1 : PolynomialRing

i2 : I = ideal(x*y-z,y*z-x,z*x-y)

o2 = ideal (- z + y*x, z*y - x, z*x - y)

o2 : Ideal of R

i3 : gens gb I

o3 = | x3-x yx2-y y2-x2 z-yx |

              1       4
o3 : Matrix R  <--- R

i4 : (x*y*z+4*x*z^4)  % I

      2
o4 = x  + 4x

o4 : R
```

5 Zero-dimensional ideals

In practice, we are often interested in polynomial systems that have only a finite number of solutions (the "zero-dimensional" case), and many interesting things happen in this case. Among other properties, the quotient ring $\mathbb{C}[x]/I$ is now a finite dimensional vector space, with its dimension being equal to the number of standard monomials. Furthermore, Groebner bases can be used to fully reduce their solution to a classical eigenvalue problem, generalizing the "companion matrix" notion from the univariate case. All this, and much more, next time...

References

[AL94] W.W. Adams and P. Loustaunau. *An introduction to Gröbner bases*, volume 3 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1994.

[BW93] T. Becker and V. Weispfenning. *Gröbner bases*, volume 141 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1993.

[CLO97] D. A. Cox, J. B. Little, and D. O'Shea. *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer, 1997.

[CoC] CoCoATeam. CoCoA: a system for doing Computations in Commutative Algebra. Available at http://cocoa.dima.unige.it.

[GPS05] G.-M. Greuel, G. Pfister, and H. Schönemann. SINGULAR 3.0. A Computer Algebra System for Polynomial Computations, Centre for Computer Algebra, University of Kaiserslautern, 2005. http://www.singular.uni-kl.de.

[GS] D.R. Grayson and M. E. Stillman. Macaulay 2, a software system for research in algebraic geometry. Available at http://www.math.uiuc.edu/Macaulay2/.

[KR00] M. Kreuzer and L. Robbiano. *Computational commutative algebra. 1*. Springer-Verlag, Berlin, 2000.

[Rab30] J. L. Rabinowitsch. Zum Hilbertschen Nullstellensatz. *Mathematische Annalen*, 102(1):520–520, 1930.

MIT 6.7230 - Algebraic techniques and semidefinite optimization April 10, 2024

Lecture 13

Lecturer: Pablo A. Parrilo *Scribe: ???*

Today we will see a few more examples and applications of Groebner bases, and we will develop in detail the zero-dimensional case.

1 Zero-dimensional ideals

In practice, we are often interested in polynomial systems that have only a finite number of solutions (the "zero-dimensional" case), and as we will see, many interesting things happen in this case.

Definition 1. *An ideal I is* zero-dimensional *if the associated variety $V(I)$ is a finite set.*

Given a system of polynomial equations, how to decide if it has a finite number of solutions (i.e., if the corresponding ideal is zero-dimensional)? If the system has finitely many solutions, then the quotient ring $\mathbb{C}[x]/I$ is finite-dimensional, and the number of roots (counted with multiplicity) is equal to the dimension of this vector space (why?). Furthermore, this is also equal to the number of *standard monomials*, i.e., the monomials "under the staircase."

When do we have a finite number of standard monomials? We can state a simple criterion for this in terms of a Groebner basis.

Lemma 2. *Let G be a Groebner basis of the ideal $I \subset \mathbb{C}[x_1, \ldots, x_n]$. The ideal I is zero-dimensional if and only if for each i ($1 \leq i \leq n$), there exists an element in the Groebner basis whose initial term is a pure power of x_i.*

Among other important consequences, when I is a zero-dimensional ideal the quotient ring $\mathbb{C}[x]/I$ is a *finite dimensional* vector space, with its dimension being equal to the number of *standard monomials*. These are the monomials that are not in the initial ideal $in(I)$ (i.e., the monomials "under the staircase"). In fact, the condition in the lemma can be easily interpreted as ensuring that the number of standard monomials remains finite.

In the zero-dimensional case we can produce *two* natural bases of the coordinate ring $\mathbb{C}[x]/I$ (for simplicity, we assume no multiplicities). The first one, as already explained, is given by the standard monomials (and thus, functions on the variety can be represented in terms of linear combinations of standard monomials). The other basis is perhaps more natural, and simply corresponds to the values that a function takes on the points of the variety. The existence of these two bases (and the coordinate change between them) will allow us to use linear algebra techniques to solve polynomial equations.

Polynomial systems via eigenvalues In the zero-dimensional case, we can use Groebner bases to reduce a zero-dimensional polynomial system to a standard eigenvalue problem, generalizing the "companion matrix" notion from the univariate case. We sketch this below.

Recall that in this case, the quotient $\mathbb{C}[x]/I$ is a finite dimensional vector space. The main idea is to consider the homomorphisms given by the n linear maps $M_{x_i} : \mathbb{C}[x]/I \to \mathbb{C}[x]/I$, $f \mapsto \widehat{(x_i f)}$ (that is, multiplication by the coordinate variables, followed by normal form). Choosing as a basis the set of standard monomials, we can effectively compute a matrix representation of these linear maps. This defines n matrices M_{x_i}, that commute with each other (why?).

Assume for simplicity that all the roots have single multiplicity. Since all the M_{x_i} commute, they can be simultaneously diagonalized by a single matrix V, and the kth diagonal entry of $VM_{x_i}V^{-1}$ contains the ith coordinate of the kth solution, for $1 \leq k \leq \#\{V(I)\}$.

(In general, we can block-diagonalize this commutative algebra, splitting into its semisimple and nilpotent components. The nilpotent part is trivial if and only if the ideal is radical.)

To understand these ideas a bit better, let's recall the univariate case.

Example 3. *Consider the ring $\mathbb{C}[x]$ of polynomials in a single variable x, and an ideal $I \subset \mathbb{C}[x]$. Since every ideal in this ring is principal, I can be generated by a single polynomial $p(x) = p_n x^n + \cdots + p_1 x + p_0$. Then, we can write $I = \langle p(x) \rangle$, and $\{p(x)\}$ is a Groebner basis for the ideal (why?). The quotient $\mathbb{C}[x]/I$ is an n-dimensional vector space, with a suitable basis given by the standard monomials $\{1, x, \ldots, x^{n-1}\}$.*

Consider as before the linear map $M_x : \mathbb{C}[x]/I \to \mathbb{C}[x]/I$. The matrix representation of this linear map in the given basis is given by

$$\begin{bmatrix} 0 & 0 & 0 & \cdots & -p_0/p_n \\ 1 & 0 & 0 & \cdots & -p_1/p_n \\ 0 & 1 & 0 & \cdots & -p_2/p_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -p_{n-1}/p_n \end{bmatrix},$$

which is the standard companion matrix $\mathcal{C}_p$ associated with $p(x)$. Its eigenvalues are exactly the roots of $p(x)$.

We present next a multivariate example.

Example 4. *Consider the ideal $I \subset \mathbb{C}[x, y, z]$ given by*

$$I = \langle xy - z, yz - x, zx - y \rangle.$$

Choosing a term ordering (e.g., lexicographic, where $x \prec y \prec z$), we obtain the Groebner basis

$$G = \{x^3 - x, yx^2 - y, y^2 - x^2, z - yx\}.$$

We can directly see from this that I is zero-dimensional (why?). A basis for the quotient space is given by $\{1, x, x^2, y, yx\}$. Considering the maps M_x, M_y, and M_z, we have that their corresponding matrix representations are given by

$$M_x = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \qquad M_y = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}, \qquad M_z = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

It can be verified that these three matrices commute. A simultaneous diagonalizing transformation is given by the matrix:

$$V = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & 1 \end{bmatrix}, \qquad V^{-1} = \frac{1}{4} \begin{bmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & -1 & -1 \\ -4 & 1 & 1 & 1 & 1 \\ 0 & 1 & -1 & 1 & -1 \\ 0 & 1 & -1 & -1 & 1 \end{bmatrix}.$$

The corresponding transformed matrices are:

$$VM_xV^{-1} = \mathrm{diag}(0, 1, 1, -1, -1)$$
$$VM_yV^{-1} = \mathrm{diag}(0, 1, -1, 1, -1),$$
$$VM_zV^{-1} = \mathrm{diag}(0, 1, -1, -1, 1)$$

from where the coordinates of the five roots can be read.

In the general (radical) case, the matrix V is a generalized Vandermonde matrix, with rows indexed by roots (points in the variety) and columns indexed by the standard monomials. The V_{ij} entry contains the j-th monomial evaluated at the ith root. Since $VV^{-1} = I$, we can also interpret the jth column of V^{-1} as giving the coefficients of a Lagrange interpolating polynomial $p_j(x)$, that vanishes at all the points in the variety, except at r_j, where it takes the value 1 (i.e., $p_j(r_k) = \delta_{jk}$).

```
Generalize Hermite form, etc
```
ToDo

Remark 5. *In practice, a better alternative to a full diagonalization (which is in general numerically unstable) is a Schur-like approach, where we find a unitary matrix U that simultaneously triangularizes the matrices in the commuting family; see [CGT97] for details.*

2 Hilbert series

Consider an ideal $I \subset \mathbb{C}[x]$ and the corresponding quotient ring $\mathbb{C}[x]/I$. We have seen that, once a particular Groebner basis is chosen, we could associate to every element of $\mathbb{C}[x]/I$ a unique representative, namely a $\mathbb{C}$-linear combination of *standard monomials*, obtained as the remainder after division with the corresponding Groebner basis. We are interested in studying, for every integer k, the dimension of the vector space of remainders of degree less than or equal to k. Expressed in a simpler way, we want to know how many standard monomials of degree k there are, for any given k.

Rather than studying this for different values of k separately, it is convenient to collect (or bundle) all these numbers together in a single object (this general technique is usually called "generating function"). The *Hilbert series* of I, denoted $H_I(t)$, is then defined as the generating function of the dimension of the space of residues of degree k, i.e.,

$$H_I(t) = \sum_{k=0}^{\infty} \dim(\mathbb{C}[x]/I \cap P_{n,k}) \cdot t^k, \tag{1}$$

where $P_{n,k}$ denotes the set of homogeneous polynomials is n variables of degree k.

Notice that, if the ideal is zero-dimensional, the corresponding Hilbert series is actually a finite sum, and thus a polynomial. The number of solutions is then equal to $H_I(1)$.

Example 6. *For the ideal I in Example 4, the corresponding Hilbert function is $H_I(t) = 1+2t+2t^2$.*

To compute the Hilbert series, we use the fact that for *graded* orderings (i.e., those for which $|\alpha| < |\beta|$ implies $\alpha \prec \beta$), the Hilbert series of I and of the initial ideal $in(I)$ are the same. For the monomial ideal $in(I)$, the Hilbert function can be easily obtained via inclusion-exclusion. Thus, all the relevant algebraic and geometric properties of the ideal (e.g., the number of roots, dimension, etc) can be obtained from a Groebner basis via the Hilbert function.

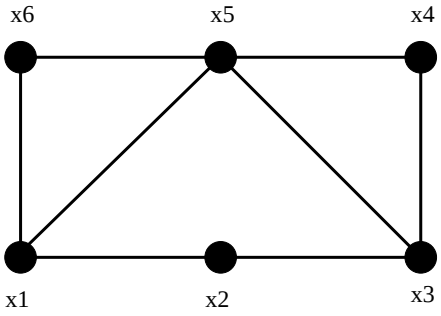

Figure 1: A six-node graph.

3 Examples

3.1 Graph ideals

Consider a graph $G = (V, E)$, and define the associated *edge ideal* $I_G = \langle x_i x_j : (i, j) \in E \rangle$. Notice that I_G is a monomial ideal. For instance, for the graph in Figure 1, the corresponding ideal is given by:

$$I_G := \langle x_1x_2, x_2x_3, x_3x_4, x_4x_5, x_5x_6, x_1x_6, x_1x_5, x_3x_5 \rangle.$$

One of the motivations for studying this kind of ideals is that many graph-theoretic properties (e.g., bipartiteness, acyclicity, connectedness, etc) can be understood in terms of purely algebraic properties of the corresponding ideal. This enables the extension and generalization of these notions to much more abstract settings (e.g., simplicial complexes, resolutions, etc).

For our purposes here, rather than studying I_G directly, we will instead study the ideal obtained when restricting to zero-one solutions[1]. For this, consider the ideal I_b defined as

$$I_b := \langle x_1^2 - x_1, \ldots, x_n^2 - x_n \rangle. \tag{2}$$

Clearly, this is a zero-dimensional radical ideal, with the corresponding variety having 2^n distinct points, namely $\{0, 1\}^n$. Its corresponding Hilbert series is $H_{I_b}(t) = (1 + t)^n = \sum_{k=0}^{n} \binom{n}{k} t^k$.

Since we want to study the intersection of the corresponding varieties, we must consider the *sum* of the ideals, i.e., the ideal $I := I_G + I_b$. It can be shown that the given set of generators (i.e., the ones corresponding to the edges, and the quadratic relations in (2)) are always a Groebner basis of the corresponding ideal. What are the standard monomials? How can they be interpreted in terms of the graph?

The *Hilbert function* of the ideal I can be obtained from the Groebner basis. In this case, the corresponding Hilbert function is given by

$$H_I(t) = 1 + 6t + 7t^2 + t^3,$$

and we can read from the coefficient of t^k the number of stable sets of size k. In particular, the degree of the Hilbert function (which is actually a polynomial, since the ideal is zero-dimensional) indicates the size of the maximum stable set, which is equal to three in this example (for the subset $\{x_2, x_4, x_6\}$).

Remark 7. *An important generalization of the edge ideals of a graph is the* Stanley-Reisner *ideal associated to a simplicial complex; see e.g. [Sta04, MS04] for details and applications.*

[1]There are more efficient ways of doing this, that would not require adding generators. We adopt this approach to keep the discussion relatively straightforward.

3.2 Integer programming

Another interesting application of Groebner bases deals with integer programming. For more details, see the papers [CT91, ST97, TW97].

Consider the integer programming problem

$$\min c^T \mathbf{x} \qquad \text{s.t.} \quad \begin{cases} A\mathbf{x} = b \\ \mathbf{x} \geq 0 \\ \mathbf{x} \in \mathbb{Z}^n \end{cases} \tag{3}$$

where $A \in \mathbb{Z}^{m \times n}$, $b \in \mathbb{Z}^m$, and $c \in \mathbb{Z}^n$. For simplicity, we assume that $A, c \geq 0$, and that we know a feasible solution $\mathbf{x}_0$. These assumptions can be removed.

The main idea to solve (3) will be to interpret the nonnegative integer decision variables $\mathbf{x}$ as the *exponents* of a monomial. The integer program will be modeled in terms of a *binomial ideal*, i.e., where the generating polynomials have only two nonzero terms. The associated Groebner basis can then be interpreted in terms of a *integer test set*, i.e., a set of lattice reduction operations under which the objective function improves monotonically, and that produce the optimal solution.

`Complete`	ToDo

Example 8. *Consider the problem data given by*

$$A = \begin{bmatrix} 4 & 5 & 6 & 1 \\ 1 & 2 & 7 & 3 \end{bmatrix}, \qquad b = \begin{bmatrix} 750 \\ 980 \end{bmatrix}, \qquad c^T = \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix}.$$

An initial feasible solution is given by $\mathbf{x}_0 = [0, 30, 80, 120]^T$. *Notice that given a feasible* $\mathbf{x}_0$ *we can compute the right-hand side b (so we don't really need b). We will work on the ring* $\mathbb{C}[z_1, z_2, w_1, w_2, w_3, w_4]$. *Thus, we need to compute a Groebner basis G of the binomial ideal*

$$\langle z_1^4 z_2 - w_1, z_1^5 z_2^2 - w_2, z_1^6 z_2^7 - w_3, z_1 z_2^3 - w_4 \rangle,$$

for a term ordering that combines elimination of the z_i *with the weight vector c. To obtain the solution, we compute the normal form of the monomial given by the initial feasible point, i.e.,* $w_2^{30} w_3^{80} w_4^{120}$. *This reduction process yields the result* $w_2^8 w_3^{106} w_4^{74}$, *and thus the optimal solution is* $[0, 8, 106, 74]$. *The corresponding costs of the initial and optimal solutions are* $c^T \mathbf{x}_0 = 780$ *and* $c^T \mathbf{x}_{opt} = 630$.

We should remark that there are more efficients ways of implementing this than the one described. Also, although this basic method cannot currently compete with specialized techniques used in integer programming for most problems, there are some particular cases where it is very efficient, mostly related with the solution of parametric problems. Additionally, these techniques have been used to prove the existence of polynomial-time algorithms for certain subclasses of integer programs; see e.g. [DLHK13] and the references therein.

References

[CGT97] R. M. Corless, P. M. Gianni, and B. M. Trager. A reordered Schur factorization method for zero-dimensional polynomial systems with multiple roots. In *ISSAC '97: Proceedings of the 1997 international symposium on Symbolic and algebraic computation*, pages 133–140, New York, NY, USA, 1997.

[CT91] P. Conti and C. Traverso. Buchberger algorithm and integer programming. In *Applied algebra, algebraic algorithms and error-correcting codes (New Orleans, LA, 1991)*, volume 539 of *Lecture Notes in Comput. Sci.*, pages 130–139. Springer, Berlin, 1991.

[DLHK13] J. A. De Loera, R. Hemmecke, and M. Köppe. *Algebraic and geometric ideas in the theory of discrete optimization*, volume 14. SIAM, 2013.

[MS04] E. Miller and B. Sturmfels. *Combinatorial commutative algebra*, volume 227. Springer Science & Business Media, 2004.

[ST97] B. Sturmfels and R. Thomas. Variation of cost functions in integer programming. *Math. Programming*, 77(3, Ser. A):357–387, 1997.

[Sta04] R. P. Stanley. *Combinatorics and commutative algebra*, volume 41. Springer Science & Business Media, 2 edition, 2004.

[TW97] R. Thomas and R. Weismantel. Truncated Gröbner bases for integer programming. *Appl. Algebra Engrg. Comm. Comput.*, 8(4):241–256, 1997.

MIT 6.7230 - Algebraic techniques and semidefinite optimization April 12, 2024

Lecture 14

Lecturer: Pablo A. Parrilo *Scribe: ???*

1 Generalizing the Hermite matrix

Recall the basic construction of the Hermite matrix $H_q(p)$ in the univariate case, whose signature gave important information on the signs of the polynomial $q(x)$ on the real roots of $p(x)$.

In a very similar way to the extension of the companion matrix to the multivariate case, we can construct an analogue of the Hermite form for general zero-dimensional ideals. The basic idea is again to consider the zero-dimensional ideal $I \subset \mathbb{R}[x_1, \ldots, x_n]$, and an associated basis of the quotient ring $B = \{x^{\alpha_1}, \ldots, x^{\alpha_m}\}$, where the elements of B are standard monomials.

For simplicity, we assume first that I is radical. In this case, the corresponding finite variety is given by m distinct points, i.e., $V(I) = \{r_1, \ldots, r_m\} \subset \mathbb{C}^n$. Notice that by the definition of the multiplication matrices M_{x_i}, we have $\sum_{i=1}^m r_i^\beta = \mathrm{Tr}[M_{x_1}^{\beta_1} \cdots M_{x_n}^{\beta_n}]$. Thus, in a similar way as we did in the univariate case, for any polynomial $q = \sum_\beta q_\beta x^\beta$ we have

$$\sum_{i=1}^m q(r_i) = \mathrm{Tr}[q(M_{x_1}, \ldots, M_{x_n})]. \tag{1}$$

Once again, this implies that if we have access to matrix representations $M_{x_1}, \ldots, M_{x_n}$, then we can explicitly compute the sum of q at all roots, by evaluating the trace of the polynomial q at the matrices M_{x_i}. Notice also that, if both q and the generators of the ideal have rational coefficients, then the expression above is also a rational number (even if the roots are not).

Example 1. *Consider the system in Example 4 of the previous lecture, and the polynomial* $p(x, y, z) = (x+y+z)^2$. *To evaluate the sum of the values that this polynomial takes on the variety, we compute:*

$$p(M_x, M_y, M_z) = \mathrm{Tr}\,(M_x + M_y + M_z)^2 = \mathrm{Tr} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 2 & 3 & 2 & 2 & 2 \\ 3 & 2 & 3 & 2 & 2 \\ 2 & 2 & 2 & 3 & 2 \\ 2 & 2 & 2 & 2 & 3 \end{bmatrix} = 12.$$

As expected, the squares of the sum of the coordinates of each of the five roots are $\{0, 9, 1, 1, 1\}$, *with the total sum being equal to 12.*

Given any $q \in \mathbb{R}[x_1, \ldots, x_n]$, we can then define a Hermite-like matrix $H_q(I)$ as

$$[H_q(I)]_{jk} := \sum_{i=1}^m q(r_i) r_i^{\alpha_j + \alpha_k}. \tag{2}$$

Notice that the rows and columns of $H_q(I)$ are indexed by standard monomials.

Consider now a vector $f = [f_1, \ldots, f_m]^T$, and the quadratic form

$$
\begin{aligned}
f^T H_q(I) f &:= \sum_{j,k=1}^{m} \sum_{i=1}^{m} q(r_i)(f_j r_i^{\alpha_j})(f_k r_i^{\alpha_k}) \\
&= \sum_{i=1}^{m} q(r_i)(f_1 r_i^{\alpha_1} + \cdots + f_m r_i^{\alpha_m})^2 \\
&= \mathrm{Tr}[(qf^2)(M_{x_1}, \ldots, M_{x_n})].
\end{aligned} \tag{3}
$$

As we see, the matrix $H_q(I)$ is a specific representation, in a basis given by standard monomials, of a quadratic form $H_q : \mathbb{C}[x]/I \to \mathbb{C}$, with $H_q : f \mapsto \sum_{i=1}^{m} (qf^2)(r_i)$. The expressions in (3) allow us to explicitly compute a matrix representation of this quadratic map. (What is the other "natural" representation of this map?)

The following theorem then generalizes the results of the univariate case, and enables, among other things, to do root counting.

Theorem 2. *The signature of the matrix $H_q(I)$ is equal to the number of real points r_i in $V(I)$ for which $q(r_i) > 0$, minus the number of real points for which $q(r_i) < 0$.*

Corollary 3. *Consider a zero dimensional ideal I. The signature of the matrix $H_1(I)$ is equal to the number of real roots, i.e., $|V(I) \cap \mathbb{R}^n|$.*

In the general (non-radical) case, we would take the property (3) as the definition of $H_q(I)$, instead of (2). Also, in Theorem 2, multiple real zeros are counted only once.

2 Parametric versions

One of the most appealing properties of Groebner-based eigenvalue methods is that they allow us to extend many of the results to the *parametric* case, i.e., when we are interested in obtaining all solutions of a polynomial system as a function of some additional parameters η_i.

Consider for simplicity the case of a single parameter η, and a polynomial system defined by $p_i(x, \eta) = 0$. In order to solve this for any fixed η, we need to compute a Groebner basis of the corresponding ideal. However, when η changes, it is possible that the resulting set of polynomials is no longer a GB. A way of fixing this inconvenience is to compute instead a *comprehensive Groebner basis*, which is a set of polynomials with the the property that it remains a Groebner basis of I for all possible specializations of the parameters. Using the corresponding monomials as a basis for the quotient space, we can give an eigenvalue characterization of the solutions for all values of η.

3 Sums of squares on quotient rings

We describe next a natural modification of the standard sos methods, that will allow us to compute sum of squares decompositions on quotient rings. This can be done by using essentially the same SDP techniques as in the standard case. Since we will need to do effective computations on the quotient, we assume that a Gröbner basis $\mathcal{G} = \{b_1, \ldots, b_k\}$ of the polynomial ideal I is available.

The method will be basically the same as in the standard case (expressing the polynomial as a quadratic form on a vector of monomials and writing linear equations to obtain a semidefinite program), but with two main differences:

- Instead of indexing the rows and columns of the matrix Q in the semidefinite program by the usual monomials, we use *standard* monomials corresponding to the Gröbner basis $\mathcal{G}$ of the

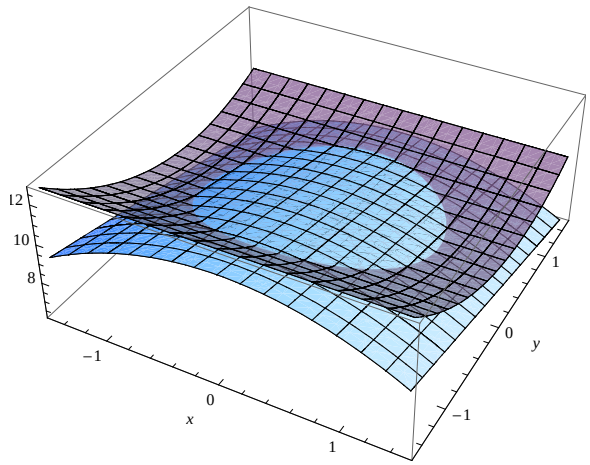

Figure 1: The polynomials $p = 10 - x^2 - y$ and $(3 - \frac{y}{6})^2 + \frac{35}{36}\, y^2$ take exactly the same values on the unit circle $x^2 + y^2 = 1$. Thus, p is nonnegative on the circle.

ideal I. These are the monomials that are not divisible by any leading term of the polynomials b_i in the Gröbner basis.

- When equating the left- and right-hand sides to form linear equations defining the subspace of valid Gram matrices, all operations are performed in the quotient ring, i.e., we rewrite the terms in *normal form* after multiplication.

Rather than giving a formal description, it is more transparent to explain the methodology via a simple example:

Example 4. *Consider the problem of deciding if the polynomial $p := 10 - x^2 - y$ is nonnegative on the variety defined by $f := x^2 + y^2 - 1 = 0$ (the unit circle). We will check whether p is a sum of squares in $\mathbb{R}[x, y]/I$, where I is the ideal $I = \langle f \rangle$. Since the ideal I is principal (generated by a single polynomial), we already have a Gröbner basis, which is simply $\mathcal{G} = \{f\}$. We use a graded lexicographic monomial ordering, where $x \prec y$. The corresponding set of standard monomials is then $\mathcal{B} = \{1, x, y, x^2, xy, x^3, x^2y, \ldots\}$.*

To formulate the corresponding semidefinite program, we pick a partial basis of the quotient ring (i.e., a subset of monomials in $\mathcal{B}$). In this example, we take only $\{1, x, y\}$, and as before, we write p as a quadratic form in these monomials:

$$\begin{aligned} 10 - x^2 - y &= \begin{bmatrix} 1 \\ x \\ y \end{bmatrix}^T \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{12} & q_{22} & q_{23} \\ q_{13} & q_{23} & q_{33} \end{bmatrix} \begin{bmatrix} 1 \\ x \\ y \end{bmatrix} \\ &= q_{11} + q_{22}x^2 + q_{33}y^2 + 2q_{12}x + 2q_{13}y + 2q_{23}xy \\ &\equiv (q_{11} + q_{33}) + (q_{22} - q_{33})x^2 + 2q_{12}x + 2q_{13}y + 2q_{23}xy \qquad \text{mod } I, \end{aligned}$$

where in the last line, we used reduction modulo the ideal to rewrite some terms as linear combinations of standard monomials only (e.g., the term $q_{33}y^2$ is replaced by $q_{33} - q_{33}x^2$). Matching coefficients between left and right, we obtain the linear equations that define the semidefinite program. Solving it, we have for this example

$$Q = \begin{bmatrix} 9 & 0 & -\frac{1}{2} \\ 0 & 0 & 0 \\ -\frac{1}{2} & 0 & 1 \end{bmatrix} = L^T L, \qquad L = \frac{1}{\sqrt{2}} \begin{bmatrix} 3 & 0 & -\frac{1}{6} \\ 0 & 0 & \frac{\sqrt{35}}{6} \end{bmatrix},$$

and therefore

$$10 - x^2 - y \equiv \left(3 - \frac{y}{6}\right)^2 + \frac{35}{36}\, y^2 \qquad mod\ I,$$

which shows that p is indeed a sum of squares on $\mathbb{R}[x, y]/I$. A simple geometric interpretation is shown in Figure 1. As expected, by the condition above, p coincides with an sos polynomial on the variety, and thus it is obviously nonnegative on that set.

Remark 5. *Despite the similarities between the "standard" case of sum of squares on the polynomial ring $\mathbb{R}[x]$ vs. the quotient ring $\mathbb{R}[x]/I$, there are a few important differences. A key distinction is related to computational complexity issues. Consider an sos decomposition $p(x) = \sum_i q_i(x)^2$. When working on $\mathbb{R}[x]$, we can always bound a priori the degree of the polynomials q_i in terms of the degree of p (namely,* $\deg(q_i) \leq \frac{1}{2} deg(p)$*). This is* not *true when working on a quotient ring, since monomials can "wrap around" when computing normal forms. This is the reason why when working on $\mathbb{R}[x]/I$ we typically have some freedom in choosing a finite set of standard monomials to index the matrix Q (unless it is feasible to include all of them).*

In fact, since for the ideal $I = \langle x_1^2 - 1, \ldots, x_n^2 - 1\rangle$ every polynomial nonnegative on $V(I)$ is a sum of squares on $\mathbb{R}[x]/I$ (see below), it directly follows that, in the general case, deciding whether a polynomial is sum of squares modulo I is NP-hard.

Nonnegative polynomials on a finite variety are SOS mod I For simplicity, we assume throughout that the ideal I is radical. Then, if a polynomial is nonnegative on a finite variety, it is a sum of squares on the quotient ring; see [Par02].

Theorem 6. *Let $f(x)$ be nonnegative on $\{x \in \mathbb{R}^n | h_i(x) = 0\}$. If the ideal $I = \langle h_1, \ldots, h_m\rangle$ is radical, then $f(x)$ is a sum of squares in the quotient ring $\mathbb{R}[x]/I$, i.e.. there exist polynomials q_i, λ_i, such that*

$$f(x) = \sum_i q_i^2(x) + \sum_{i=1}^{m} \lambda_i(x) h_i(x).$$

Remark 7. *The assumption that I is radical (or a suitable local modification) is necessary when $f(x)$ is nonnegative but not strictly positive. For instance, the polynomial $f = x$ is nonnegative on the variety defined by the (non-radical) ideal $\langle x^2\rangle$, although no decomposition of the form $x = s_0(x) + \lambda(x)x^2$ (where s_0 is SOS), can possibly exist.*

References

[Par02] P. A. Parrilo. An explicit construction of distinguished representations of polynomials nonnegative over finite sets. Technical Report IfA Technical Report AUT02-02. Available from `https://www.mit.edu/~parrilo`, ETH Zürich, 2002.

MIT 6.7230 - Algebraic techniques and semidefinite optimization April 17, 2024

Lecture 15

Lecturer: Pablo A. Parrilo *Scribe: ???*

One of our main goals in this course is to achieve a better understanding of the techniques available for polynomial systems over the real field. Today we discuss how to certify infeasibility for polynomial equations over the reals, and contrast these approaches with well-known results in linear algebra, linear programming, and complex algebraic geometry.

We will discuss the possible convergence of these schemes in the general case later in the course, concentrating today on an elementary proof of the finite convergence in the zero-dimensional case [Par02].

1 Infeasibility of real polynomial equations

Based on what we have learned in the past weeks, we have a quite satisfactory answer to the question of when a system of polynomial equations has solutions over the complex field. Indeed, as we have seen, given a system of polynomial equations $\{h_i(x) = 0,\, i = 1, \ldots, m\}$, we can form the associated ideal $I = \langle h_1, \ldots, h_m \rangle$. By the Nullstellensatz, the associated complex variety $V(I)$ (i.e., the solution set $\{x \in \mathbb{C}^n \,|\, h_i(x) = 0\}$) will be empty if and only if $I = \mathbb{C}[x]$, or equivalently, $1 \in I$. Computationally, this condition can be checked by computing a reduced Groebner basis of I (with respect to any term ordering), which will be equal to $\{1\}$ if this holds.

What happens, however, when we are interested in *real* solutions, and not just complex ones? Or, if not only we have equations, but also inequalities? Consider, for instance, the basic semialgebraic set given by

$$S = \{x \in \mathbb{R}^n \mid f_i(x) \geq 0, \quad h_i(x) = 0\}. \tag{1}$$

How to decide whether the set S is empty? Can we give a Groebner-like criterion to demonstrate the infeasibility of this system of equations? Even worse, do we even know that this question can be decided algorithmically[1]?

Fortunately for us, a famous result, the Tarski-Seidenberg theorem, guarantees the algorithmic solvability of this problem (in fact, of the much larger class of problems that includes quantified polynomial inequalities). We will discuss this powerful approach in more detail later, when presenting cylindrical algebraic decomposition (CAD) techniques, concentrating instead in a more direct way of tackling the feasibility problem.

2 Certificates

`Discuss certificates: NP/co-NP, Linear algebra, LP, Nullstellensatz, P-satz` ToDo

[1] There are certainly similar-looking problems that are *not* decidable. A famous one is the solvability of polynomial equations over the integers. This is Hilbert's 10th problem, solved in 1970 by Matiyasevich; see [Dav73] for a full account of the solution and historical remarks. This result implies, in particular, the nonexistence of an algorithm to solve integer quadratic programming; see [Jer73]. A nice survey on undecidability questions for polynomial equations is [Poo08].

3 The zero-dimensional case

What happens in the case where the equations $h_i(x) = 0$ in the system (1) define a zero dimensional ideal? It should be intuitively obvious that, in some sense, a finite certificate of the emptiness of S exists. Indeed, *if* we had access to all the roots, of which there are a finite number, by evaluating the remaining inequalities we could decide feasibility or infeasibility. As we will see, we can actually "encode" this process in a set of polynomials, that prove the existence of these certificates.

Theorem 1. *Consider the set S in (1), and assume the ideal $I = \langle h_1, \ldots, h_m \rangle$ is radical. Then, S is empty if and only if there exists a decomposition*

$$-1 = s_0(x) + \sum_{i=1} s_i(x) f_i(x) + \sum_{i=1} \lambda_i(x) h_i(x).$$

where the s_i are sums of squares.

Notice that we can equivalently write

$$-1 \equiv s_0(x) + \sum_{i=1} s_i(x) f_i(x) \qquad \mod I.$$

It should be clear that one direction of the implication is obvious (which one?). For the converse, notice that since the ideal is radical, this is the same as

$$-1 = s_0(v_j) + \sum_{i=1} s_i(v_j) f_i(v_j) \quad \text{for all } v_j \in V(I).$$

Thus, we can construct polynomials s_0 and s_i by defining them pointwise at every point v_j of the variety, with the sos conditions replaced by pointwise nonnegativity of the s_i. As we did earlier in the unconstrained case, we can then construct sos polynomials that interpolate all these values.

4 Optimization

Since optimization can be interpreted as a parametrized family of feasibility problems, we can directly apply these results towards optimization of polynomial or rational functions. For instance, we have the following result:

Theorem 2. *Let $p(x)$ be nonnegative on $S = \{x \in \mathbb{R}^n \mid f_i(x) \geq 0,\ h_i(x) = 0\}$, and assume that the ideal $I = \langle h_1, \ldots, h_m \rangle$ is radical. Consider the optimization problem*

$$\max \gamma \qquad s.t. \quad p(x) - \gamma = s_0(x) + \sum_{i=1} s_i(x) f_i(x) + \sum_{i=1} \lambda_i(x) h_i(x).$$

where the s_i are sums of squares, and the decision variables are γ and the coefficients of the polynomials $s_i(x)$, $\lambda_i(x)$. Then, the optimal value of γ is equal to the minimum of $p(x)$ over S.

Notice that this is a sum of squares program, since all the constraints are linear and/or sum of squares constraints.

Remark 3. *The assumption that I is radical (or a suitable local modification) is necessary for finite convergence when $p(x)$ is nonnegative but not strictly positive. For instance, the polynomial $p(x) := x$ is nonnegative on the variety defined by the (non-radical) ideal $\langle x^2 \rangle$, although no decomposition of the form $x = s_0(x) + \lambda(x)x^2$ (where s_0 is SOS), can possibly exist. Notice, however, that $x + \varepsilon^2 \equiv (\frac{x}{2\varepsilon} + \varepsilon)^2$ for* all $\varepsilon \neq 0$*, and thus we can prove bounds that are arbitrarily close to the optimal value.*

The results naturally extend to the general case, with minimal modifications (particularly, in the compact case). Details will follow in the next lectures...

References

[Dav73] M. Davis. Hilbert's tenth problem is unsolvable. *Amer. Math. Monthly*, 80:233–269, 1973.

[Jer73] R.G. Jeroslow. There cannot be any algorithm for integer programming with quadratic constraints. *Operations Res.*, 21:221–224, 1973.

[Par02] P. A. Parrilo. An explicit construction of distinguished representations of polynomials nonnegative over finite sets. Technical Report IfA Technical Report AUT02-02. Available from `https://www.mit.edu/~parrilo`, ETH Zürich, 2002.

[Poo08] B. Poonen. Undecidability in number theory. *Notices of the AMS*, 55(3):344–350, 2008.

MIT 6.7230 - Algebraic techniques and semidefinite optimization April 19, 2024

Lecture 16

Lecturer: Pablo A. Parrilo *Scribe: ???*

Quantifier elimination (QE) is a very powerful procedure for problems involving first-order formulas over real fields. The cylindrical algebraic decomposition (CAD) is a technique for "efficient" implementation of QE, that effectively reduces a seemingly infinite problem into a finite (but potentially large) instance. For much more information about QE and CAD (including a reprint of Tarski's original 1930 work), we recommend the book [CJ98].

1 Quantifier elimination

A quantifier-free formula is an expression consisting of polynomial equations ($f(x) = 0$) and inequalities ($f(x) \geq 0$) combined using the Boolean operators $\neg$ (negation), $\wedge$ (and), $\vee$ (or), and $\Rightarrow$ (implies). We often also allow strict inequalities $f(x) > 0$ and inequations $f(x) \neq 0$, since these are just shorthands for particular boolean combinations of equations and inequalities.

In general, a *formula* (in prenex form) is an expression in the variables $x = (x_1, \ldots, x_n)$ of the type:

$$(Q_1 x_1)...(Q_s x_s) \quad \mathcal{F}(f_1(x), \ldots, f_r(x)) \tag{1}$$

where each Q_i is one of the quantifiers $\forall$ (for all) and $\exists$ (there exists). Furthermore, $\mathcal{F}(f_1(x), ..., f_r(x))$ is assumed to be a quantifier-free formula. If there is a quantifier corresponding to the variable x_i, we say that x_i is *quantified*, or *free* otherwise.

Example 1. *The following are valid formulas*

$$(\forall x)\left[(x \geq 0) \Rightarrow (x^2 + ax + b \geq 0)\right]$$
$$(\forall x)(\exists y)\left[x > y^2\right]$$
$$(\forall \delta)(\exists \epsilon)\left[(\epsilon^2 + \delta^2 \leq 1) \vee (\epsilon \neq 0)\right] \Rightarrow [\delta < 1].$$

The first formula has two free variables (since the variables a and b are unquantified), while for the other two all variables are quantified.

We will interpret the symbols in a formula as taking only real values. Notice that a formula without free variables (usualled called a *closed* formula or a *sentence*) is either true or false. For instance, the last two expressions in Example 1 are sentences, with the first one being false and the second being true. Notice also that the truth value may depend on the order of the quantifiers.

Tarski showed that for every formula including quantifiers there is always an equivalent quantifier free formula. Obtaining the latter from the former is called quantifier elimination.

Theorem 2 (Tarski-Seidenberg)**.** *For every first-order formula over the real field there exists an equivalent quantifier-free formula. Furthermore, there is an explicit algorithm to compute this quantifier-free formula.*

The Tarski-Seidenberg theorem is an extremely powerful result, since it provides a complete characterization and algorithmic technique for an extremely large collection of problems involving polynomials. Unfortunately, there are very serious computational barriers to the efficient practical implementation of these ideas, since the resulting algorithms have extremely poor scaling properties,

with respect to the number of variables (towers of exponentials). Newer methods, such as the (partial) cylindrical algebraic decomposition (CAD) technique due to Collins and described below, or the critical point method, are by comparison much better. Nevertheless, they still behave exponentially (or worse) in terms of the number of variables (and likely this is required, modulo complexity-theoretic conjectures).

2 Tarski-Seidenberg

Example 3. *Consider the quantified first-order formula:*

$$(\forall x)(\forall y)\left[(x^2 + ay^2 \leq 1) \Rightarrow (ax^2 - a^2xy + 2 \geq 0)\right]. \tag{2}$$

This formula is equivalent to the quantifier free expression:

$$(a \geq 0) \wedge (a^3 - 8a - 16 \leq 0),$$

which defines the interval $[0, a_\star]$*, where* $a_\star \approx 3.538$*. Thus, the original expression (2) is true only for* $a \in [0, a_\star]$*.*

2.1 Geometric interpretation

The Tarski-Seidenberg theorem allows us to understand what happens when we apply certain operations to semialgebraic sets. For instance, many natural constructions such as set closure, convex hulls, conic hulls, projections, etc. can easily be described in terms of first-order formulas. This implies that, by eliminating these quantifiers, "simpler" quantifier-free descriptions of these sets can be obtained. In particular, an important geometric interpretation of the Tarski-Seidenberg theorem is the following:

Theorem 4. *The projection of a semialgebraic set is semialgebraic.*

Recall that a spectrahedron is a basic semialgebraic set. Linear projections of spectrahedra are not necessarily basic semialgebraic (recall the Examples in Lecture 5), but they are always semialgebraic sets.

2.2 Applications

Static output feedback An early application of Tarski-Seidenberg in control theory was the "solution" of the static output feedback stabilization problem in [ABJ75]. Given matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, we want to find a matrix $K \in \mathbb{R}^{m\times p}$ such that the matrix $A + BKC$ is Hurwitz, i.e., all its eigenvalues are in the left-hand plane. Since the existence of such a matrix can be easily expressed as a formula in first order logic[1], the decidability and existence of an effective (but not efficient) algorithm immediately follows.

Simultaneous stabilization A very interesting result by Blondel [Blo94, BG93] shows that the simultaneous stabilization of three linear time-invariant systems is *not* decidable (and thus, cannot be semialgebraic). Notice however that the Tarski-Seidenberg theorem implies that for any given bound on the degree of the controller, the problem is decidable.

[1]For instance, $(\exists K)(\forall x)(\forall \lambda)\left[(A + BKC)x = \lambda x \vee x \neq 0\right] \Rightarrow [\Re(\lambda) \leq 0]$. Notice that we are being a bit sloppy with notation, since for a fully real formulation, we should split x and λ into real and imaginary parts. There are many other equivalent expressions, using for instance a Lyapunov equation, or the Routh array.

Game theory Nash equilibria, as well as almost all other game-theoretic solution concepts, can be expressed as first-order formulas over the reals. For suitable classes of games, and under the right conditions, (e.g., finite games, polynomial payoffs, stochastic games, etc.) this implies that equilibrium sets are semialgebraic; see e.g. [SSZ91].

3 Cylindrical Algebraic Decomposition (CAD)

There are a few approaches for effective implementation of the QE procedure. One of the most well-known, which is also relatively easy to understand, is the cylindrical algebraic decomposition (CAD) due to Collins [Col75]. We describe the basic elements of this approach below. We remark that much better algorithms (in the theoretical complexity sense) are known; see for instance the article by Renegar [Ren91] (also reprinted in [CJ98]) or [BPR03]. In particular, for CAD the number of operations usually scales in a doubly exponential fashion with the number of variables, while the newer methods are doubly exponential in the number of *quantifier alternations*.

3.0.1 Description

Given a set P of multivariate polynomials in n variables, a CAD is a special partition of $\mathbb{R}^n$ into components, called *cells*, over which all the polynomials have constant signs. The algorithm for computing a CAD also provides a point in each cell, called *sample point*, which can be used to determine the sign of the polynomials in the cell.

A cell is called *cylindrical* if it has the form $S \times \mathbb{R}^k$, for some $k \leq n$. A decomposition of $\mathbb{R}^n$ is a CAD if all polynomials have constant sign on each cell, and all cells are cylindrical.

The CAD associated to the formula (1) depends only on its quantifier-free part $\mathcal{F}(f_1(x), \ldots, f_r(x))$. Since all possible truth values of the formula are in correspondence with the values at the sample points, we can use the CAD to evaluate its truth value, and to perform quantifier elimination.

The basic CAD construction consists of two steps: *projection* and *lifting* (plus an additional third one, if formula construction is desired).

In the first projection phase, we compute successive sets of polynomials in $n-1, n-2, \ldots, 1$ variables. The main idea is, given an input set of polynomials, to compute at each step a new set of polynomials obtained by eliminating one variable at a time. In general, the elimination order does matter and a good choice leads to lower computational complexity.

The second phase (lifting) constructs a decomposition of $\mathbb{R}$, at the lowest level of projection, after all but one variable have been eliminated. This decomposition of $\mathbb{R}$ is successively extended to a decomposition of $\mathbb{R}^n$.

The basic operations necessary in the construction of CADs are (sub)resultants and (sub)discriminants.

`Complete`	ToDo

An pretty complete implementation of (an improved version of) the CAD method for quantifier elimination is the software package QEPCAD [Bro03]. The software Mathematica has a restricted implementation of QE, via the `Resolve[]` command. The theorem prover Z3 [dMB08] from Microsoft Research also implements some limited cases of QE for the NRA fragment (Polynomial Real Arithmetic).

References

[ABJ75] B. D. O. Anderson, N. K. Bose, and E. I. Jury. Output feedback stabilization and related problems—solution via decision methods. *IEEE Transactions on Automatic Control*, 20:53–66, 1975.

[BG93] V. Blondel and M. Gevers. Simultaneous stabilizability of three linear systems is rationally undecidable. *Mathematics of Control, Signals, and Systems*, 6(2):135–145, 1993.

[Blo94] V. Blondel. *Simultaneous stabilization of linear systems*, volume 191 of *Lecture Notes in Control and Information Sciences*. Springer-Verlag London Ltd., London, 1994.

[BPR03] S. Basu, R. Pollack, and M.-F. Roy. *Algorithms in real algebraic geometry*, volume 10 of *Algorithms and Computation in Mathematics*. Springer-Verlag, Berlin, 2003.

[Bro03] C.W. Brown. *QEPCAD - Quantifier Elimination by Partial Cylindrical Algebraic Decomposition*, 2003. Available from `https://www.usna.edu/CS/qepcadweb/B/QEPCAD.html`.

[CJ98] B. F. Caviness and J. R. Johnson, editors. *Quantifier elimination and cylindrical algebraic decomposition*, Texts and Monographs in Symbolic Computation, Vienna, 1998. Springer-Verlag.

[Col75] G. E. Collins. Quantifier elimination for real closed fields by cylindrical algebraic decomposition. In *Automata theory and formal languages (Second GI Conf., Kaiserslautern, 1975)*, pages 134–183. Lecture Notes in Comput. Sci., Vol. 33. Springer, Berlin, 1975.

[dMB08] Leonardo Mendonça de Moura and Nikolaj S. Bjørner. Z3: an efficient SMT solver. In C. R. Ramakrishnan and Jakob Rehof, editors, *Tools and Algorithms for the Construction and Analysis of Systems, 14th International Conference, TACAS*, volume 4963 of *Lecture Notes in Computer Science*, pages 337–340. Springer, 2008.

[Ren91] J. Renegar. Recent progress on the complexity of the decision problem for the reals. In *Discrete and computational geometry (New Brunswick, NJ, 1989/1990)*, volume 6 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 287–308. Amer. Math. Soc., Providence, RI, 1991.

[SSZ91] S.H. Schanuel, L.K. Simon, and W.R. Zame. The algebraic geometry of games and the tracing procedure. *Game equilibrium models*, 2:9–43, 1991.

MIT 6.7230 - Algebraic techniques and semidefinite optimization April 21, 2023

Lecture 17

Lecturer: Pablo A. Parrilo *Scribe: ???*

Today we continue with some additional aspects of quantifier elimination. We will then recall the Positivstellensatz and its relations with semidefinite programming. After introducing copositive matrices, we present Pólya's theorem on positive forms on the simplex, and the associated relaxations. Finally, we conclude with an important result due to Schmüdgen about representation of positive polynomials on compact sets.

1 Certificates

Quantifier elimination and decision methods are extremely powerful, since in principle they can handle formulas with arbitrary (finite) quantifier alternations and general semialgebraic expressions. Contrast this with the case of the Psatz, which applies only to the case of existential quantifiers on conjunctions of polynomial equations/inequalities (emptiness of a basic semialgebraic set). Nevertheless, a very important practical advantage of Psatz techniques is that they provide *certificates* of infeasibility, that can be checked in a completely independent fashion, regardless of what process was used to obtain them. For quantifier elimination, typically the only way of certifying that the answer is correct is by ensuring the correctness of the design and implementation of the QE method itself.

2 Psatz revisited

Recall the statement of the Positivstellensatz.

Theorem 1 (Positivstellensatz)**.** *Consider the set* $S = \{x \in \mathbb{R}^n \mid f_i(x) \geq 0,\, h_i(x) = 0\}$*. Then,*

$$S = \emptyset \quad \Leftrightarrow \quad \exists f, h \in \mathbb{R}[x] \; s.t. \begin{cases} f + h &= -1 \\ f &\in \mathbf{cone}\{f_1, \ldots, f_s\} \\ h &\in \mathbf{ideal}\{h_1, \ldots, h_t\} \end{cases}$$

Once again, since the conditions on the polynomials f, h are convex and affine, respectively, by restricting their degree to be less than or equal to a given bound d we have a finite-dimensional semidefinite programming problem.

2.1 Hilbert 17th problem

As we have seen, in the general case nonnegative multivariate polynomials can fail to be a sum of squares (the Motzkin polynomial being the classical counterxample). As part of his famous list of twenty-three problems that he presented at the International Congress of Mathematicians in 1900, David Hilbert asked the following[1]:

[1] This text was obtained from `http://mathcs.clarku.edu/~djoyce/hilbert/`, and corresponds to Newson's translation of Hilbert's original German address. In that website you will also find links to the current status of the problems, as well as the original German text.

> **17. Expression of definite forms by squares.** A rational integral function or form in any number of variables with real coefficient such that it becomes negative for no real values of these variables, is said to be definite. The system of all definite forms is invariant with respect to the operations of addition and multiplication, but the quotient of two definite forms in case it should be an integral function of the variables is also a definite form. The square of any form is evidently always a definite form. But since, as I have shown, not every definite form can be compounded by addition from squares of forms, the question arises which I have answered affirmatively for ternary forms whether every definite form may not be expressed as a quotient of sums of squares of forms. At the same time it is desirable, for certain questions as to the possibility of certain geometrical constructions, to know whether the coefficients of the forms to be used in the expression may always be taken from the realm of rationality given by the coefficients of the form represented.

In other words, can we write every nonnegative polynomial as a sum of squares of *rational functions*? As we we show next, this is a rather direct consequence of the Psatz. Of course, it should be clear (and goes without saying) that we are (badly) inverting the historical order! In fact, much of the motivation for the development of real algebra came from Hilbert's question.

How can we use the Psatz to prove that a polynomial $p(x)$ is nonnegative? Clearly, p is nonnegative if and only if the set $\{x \in \mathbb{R}^n \,|\, p(x) < 0\}$ is empty. Since our version of the Psatz does not allow for strict inequalities (there are slightly more general, though equivalent, formulations that do), we'll need a useful trick discussed earlier ("Rabinowitch's trick"). Introducing a new variable z, the nonnegativity of $p(x)$ is equivalent to the emptiness of the set described by

$$-p(x) \geq 0, \qquad 1 - zp(x) = 0.$$

The Psatz can be used to show that this holds if and only if there exist polynomials $s_0, s_1, t \in \mathbb{R}[x, z]$ such that

$$s_0(x, z) - s_1(x, z) \cdot p + t(x, z) \cdot (1 - zp) = -1,$$

where s_0, s_1 are sums of squares. Replace now $z \to 1/p(x)$, and multiply by p^{2k} (where k is sufficiently large) to obtain

$$\tilde{s}_0(x) - \tilde{s}_1(x) \cdot p(x) = -p(x)^{2k},$$

where $\tilde{s}_0, \tilde{s}_1$ are sums of squares in $\mathbb{R}[x]$. Solving now for p, we have:

$$p(x) = \frac{\tilde{s}_0(x) + p(x)^{2k}}{\tilde{s}_1(x)} = \frac{\tilde{s}_1(x)(\tilde{s}_0(x) + p(x)^{2k})}{\tilde{s}_1^2(x)},$$

and since the numerator is a sum of squares, it follows that $p(x)$ is indeed a sum of squares of rational functions.

3 Copositive matrices and Pólya's theorem

An interesting class of matrices are the *copositive matrices*, which are those for which the associated quadratic form is nonnegative on the nonnegative orthant.

Definition 2. *A matrix $M \in \mathcal{S}^n$ is* copositive *if it satisfies*

$$x^T M x \geq 0, \qquad \textit{for all} \quad x_i \geq 0.$$

As opposed to positive semidefiniteness, which can be checked in polynomial time, the recognition problem for copositive matrices is an NP-hard problem [MK87]. The set of copositive matrices is a proper cone, which we will call $\mathcal{C}$. By the remark above, checking membership to the cone $\mathcal{C}$ is a difficult problem, even though it is convex. Its dual cone $\mathcal{C}^*$ also has a nice characterization, since it corresponds to the set of *completely positive* matrices:

Definition 3. *A matrix $W \in \mathcal{S}^n$ is* completely positive *if it is the sum of outer products of nonnegative vectors, i.e.,*

$$W = \sum_i^m x_i x_i^T, \qquad x_i \geq 0.$$

Alternatively, the matrix W factors as $W = FF^T$, where F is a nonnegative matrix (i.e., $F = [x_1, \ldots, x_m] \in \mathbb{R}^{m \times n}$).

Let $\mathcal{B} = \mathcal{C}^*$ be the set of copositive matrices. A natural sufficient condition for a matrix M to be copositive is if it can be expressed as the sum of a positive semidefinite matrix and a nonnegative matrix, i.e.,

$$M = P + N, \qquad P \succeq 0, \quad N_{ij} \geq 0.$$

This gives the containments

$$\mathcal{C} \supseteq \mathcal{S}_+^n + \mathcal{P}_+^n, \qquad \mathcal{B} \subseteq \mathcal{S}_+^n \cap \mathcal{P}_+^n,$$

where $\mathcal{P}_+^n \approxeq \mathbb{R}_+^{\binom{n+1}{2}}$ is the (self-dual) cone of nonnegative matrices. The containments are strict for $n \geq 5$; specific counterexamples will be discussed in the homework set. It should be clear that these conditions can be checked via SDP.

A good reference on completely positive matrices is [BSM03].

Applications There are many interesting applications of copositive and completely positive matrices. Among others, we mention:

- Consider a graph G, with A being its the adjacency matrix. The stability number α of the graph G is equal to the cardinality of its largest stable set. By a result of Motzkin and Straus, it is known that it can be obtained as:

  $$\frac{1}{\alpha(G)} = \min_{x_i \geq 0, \sum_i x_i = 1} x^T (I + A) x$$

  This implies that $\alpha(G) \leq \gamma$ if and only if the matrix $\gamma \cdot (I + A) - ee^T$ is copositive.

- In the analysis of linear dynamical systems with piecewise affine dynamics, it is often convenient to use piecewise-quadratic Lyapunov functions. In this case, we need to verify positivity conditions of an indefinite quadratic on a polyhedron. To make this precise, consider an affine dynamical system $\dot{x} = Ax + b$, a polyhedron $\mathcal{S}$ and a Lyapunov function $V(x)$ defined by:

  $$\mathcal{S} := \left\{ x \in \mathbb{R}^n \,|\, L \begin{bmatrix} x \\ 1 \end{bmatrix} \geq 0 \right\}, \qquad V(x) = \begin{bmatrix} x \\ 1 \end{bmatrix}^T P \begin{bmatrix} x \\ 1 \end{bmatrix}.$$

  Then, conditions for V and $-\dot{V}$ to be nonnegative on the set are:

  $$P \succeq L^T C_1 L, \qquad P \begin{bmatrix} A & b \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} A & b \\ 0 & 0 \end{bmatrix}^T P \preceq -L^T C_2 L,$$

  with C_1, C_2 copositive.

- Another interesting application of copositive matrices is in the performance analysis of queueing networks; see e.g. [KM96]. Modulo some (important) details, the basic idea is to use a quadratic function $x^T Mx$ as a Lyapunov function, where the matrix M is copositive and x represents the lengths of the queues.

Pólya's theorem and copositive hierarchies Matrix copositivity can be easily interpreted in terms of polynomial nonnegativitiy. Indeed, it exactly corresponds to the condition that the polynomial $p_M(z_1, \ldots, z_n) := \mathbf{z}^T M \mathbf{z}$ be nonnegative, where $\mathbf{z} := [z_1^2, \ldots, z_n^2]^T$. Geometrically, this means that the copositive cone is an affine slice of the cone of nonnegative quartic forms.

It can be shown that the natural SOS relaxation of the nonnegativity of $p_M(\mathbf{z})$ yields the $P + N$ condition described earlier. We can strengthen this result to produce a hierarchy of SDP-representable cones that approximate $\mathcal{B}$ and $\mathcal{C}$. To do this, we use a well-known result by Pólya on positive forms on the simplex:

Theorem 4 (Pólya)**.** *Consider a homogeneous polynomial in n variables of degree d, that is strictly positive on the unit simplex $\Delta_n := \{x \in \mathbb{R}^n \mid x_i \geq 0, \sum_{i=1}^n x_i = 1\}$. Then, for large enough k, the polynomial $(x_1 + \cdots + x_n)^k p(x)$ has nonnegative coefficients.*

It is possible to formulate a natural hierarchy of sufficient conditions for a matrix to be copositive, by considering a sum of squares condition on the polynomial $(\mathbf{z}^T\mathbf{z})^k(\mathbf{z}^T M\mathbf{z}) = (z_1^2 + \cdots + z_n^2)^k p_M(z_1, \ldots, z_n)$. Completeness of this hierarchy follows directly from Pólya's theorem [Par00].

Furthermore, there are interesting connections between Pólya's result and a foundational theorem in probability known as de Finetti's exchangeability theorem. We explore some of these links in the homework problems.

4 Positive polynomials

The Positivstellensatz allows us to obtain certificates of the emptiness of a basic semialgebraic set, explicitly given by polynomials.

What if we want to apply this for optimization? As we have seen, it is relatively straightforward to convert an optimization problem to a family of feasibility problems, by considering the sublevel sets, i.e., the sets $\{x \in \mathbb{R}^n \mid f(x) \leq \gamma\}$.

In the case of constrained problems, however, using the fully general Psatz would yield conditions that are not linear in the unknown parameter γ (because we need products between the constraints), and this presents a difficulty to the direct use of SDP. Notice nevertheless, that the problem is certainly an SDP for any fixed value of γ, and it thus quasiconvex (which is almost as good, except for the fact that we cannot use "standard" SDP solvers to solve it directly, but rather rely on methods such as bisection).

A possible approach, however, is to use certain "distinguished" representations of nonnegative polynomials over semialgebraic sets. Typically, these require some mild assumptions, such as compactness. A good example, which we will discuss later, is the celebrated theorem by Schmüdgen:

Theorem 5 ([Sch91])**.** *If $p(x)$ is strictly positive on the set $K = \{x \in \mathbb{R}^n \mid f_i(x) \geq 0\}$, and K is compact, then $p(x) \in \mathbf{cone}\{f_1, \ldots, f_s\}$.*

In the next lecture we will describe the basic elements of Schmüdgen's proof. His approach combines both algebraic tools (using the Positivstellensatz to prove the boundedness of certain operators) and functional analysis (spectral measures of commuting families of operators and the

Hahn-Banach theorem). We will also describe some alternative versions due to Putinar, as well as a related purely functional-analytic result due to Megretski.

For a comprehensive treatment and additional references, we mention [BCR98, Mar00, PD01] among others.

References

[BCR98] J. Bochnak, M. Coste, and M-F. Roy. *Real Algebraic Geometry*. Springer, 1998.

[BSM03] A. Berman and N. Shaked-Monderer. *Completely positive matrices*. World Scientific, 2003.

[KM96] P. R. Kumar and S. P. Meyn. Duality and linear programs for stability and performance analysis of queuing networks and scheduling policies. *IEEE Trans. Automat. Control*, 41(1):4–17, 1996.

[Mar00] M. Marshall. *Positive polynomials and sums of squares*. Dottorato de Ricerca in Matematica. Dept. di Mat., Univ. Pisa, 2000.

[MJ81] D.H. Martin and D.H. Jacobson. Copositive matrices and definiteness of quadratic forms subject to homogeneous linear inequality constraints. *Linear Algebra and its Applications*, 35:227–258, 1981.

[MK87] K. G. Murty and S. N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39:117–129, 1987.

[Par00] P. A. Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, California Institute of Technology, May 2000.

[PD01] A. Prestel and C. N. Delzell. *Positive polynomials: from Hilbert's 17th problem to real algebra*. Springer Monographs in Mathematics. Springer, 2001.

[Sch91] K. Schmüdgen. The K-moment problem for compact semialgebraic sets. *Math. Ann.*, 289:203–206, 1991.

MIT 6.7230 - Algebraic techniques and semidefinite optimization May 1st, 2024

Lecture 18

Lecturer: Pablo A. Parrilo *Scribe: ???*

In this lecture we introduce Schmüdgen's theorem about the K-moment problem (or equivalently, on the representation of positive polynomials) and describe the basic elements in his proof. This approach combines both algebraic tools (using the Positivstellensatz to prove the boundedness of certain operators) and functional analysis (spectral measures of commuting families of operators and the Hahn-Banach theorem). We will also describe some alternative versions due to Putinar, as well as a related purely functional-analytic result due to Megretski.

For a comprehensive treatment and additional references, we mention [BCR98, Mar00, PD01] among others.

1 Representations of positive polynomials

As we have seen, the Positivstellensatz allows us to obtain certificates of the emptiness of a basic semialgebraic set, explicitly given by polynomials. When looking for bounded degree certificates, this provides a natural hierarchy of SDP-based conditions [Par00, Par03].

What if we want to apply this for the particular case of optimization? As we have seen, it is relatively straightforward to convert a polynomial optimization problem to a one-parameter family of feasibility problems, by considering the sublevel sets, i.e., the sets $\{x \in \mathbb{R}^n \mid f(x) \leq \gamma\}$.

In the general case of constrained problems, however, using the full power of the Psatz will yield conditions that are not linear in the unknown parameter γ (because we need products between the constraints and objective function), and in principle, this presents a difficulty to the direct use of SDP. Notice nevertheless, that the problem is certainly an SDP for any fixed value of γ, and is thus quasiconvex (which is almost as good, except for the fact that we cannot use "standard" SDP solvers to solve it directly, but rather rely on methods such as bisection).

Of course, we can always produce specific families of certificates that are linear in γ, and use them for optimization (e.g., like we did in the copositivity case). However, in general it is unclear whether the desired family is "complete," in the sense that we will be able to prove arbitrarily good bounds on the optimal value as the degree of the polynomials grows to infinity.

1.1 Schmüdgen's theorem

In 1991, Schmüdgen presented a characterization of the moment sequences of measures supported on a compact semialgebraic K (the K-moment problem). As in the one-dimensional case we studied earlier the question is, given an (infinite) sequence of moments, decide whether it actually corresponds to a nonnegative measure with support on a given set K.

His solution combined both real algebraic methods (the Psatz), with some functional analytic tools (reproducing kernel Hilbert spaces, bounded operators, and the spectral theorem).

This characterization of moment sequences can be used, in turn, to produce an explicit description of the set of strictly positive polynomials on a compact semialgebraic set:

Theorem 1 ([Sch91])**.** *If $p(x)$ is strictly positive on $K = \{x \in \mathbb{R}^n \mid f_i(x) \geq 0\}$, and K is compact, then $p(x) \in \mathbf{cone}\{f_1, \ldots, f_m\}$.*

expand	ToDo

There are several interesting ideas in the proof; a coarse description follows. The first step is to use the Positivstellensatz to produce an algebraic certificate of the compactness of the set K. Then the given moment sequence (which is a positive definite function on the semigroup of monomials) is used to construct a particular pre-Hilbert space and its completion (namely, the associated reproducing kernel Hilbert space). In this Hilbert space, we consider linear operators T_{x_i} given by multiplication by the coordinate variables, and use the algebraic certificate of compactness to prove that these are bounded. Now, the T_{x_i} are a finite collection of pairwise commuting, bounded, self-adjoint operators, and thus there exists a spectral measure for the family, from which a measure, only supported in K, can be extracted. Finally, a Hahn-Banach (separating hyperplane) argument is used to prove the final result.

1.2 Putinar's approach

The theorem in the previous section requires (in principle) all $2^m - 1$ squarefree products of constraints[1]. Putinar [Put93] presented a modified formulation (under stronger assumptions) for which the representation is *linear* in the constraints. We introduce the following concept:

Definition 2. *Let* $\{f_1, \ldots, f_m\} \subset \mathbb{R}[x]$. *The* preprime *generated by the* f_i*, and denoted by* $\mathbf{preprime}\{f_1, \ldots, f_m\}$ *is the set of all polynomials of the form* $s_0 + s_1 f_1 + \cdots + s_m f_m$*, where all the* s_i *are sums of squares.*

Notice that $\mathbf{preprime}\{f_i\} \subset \mathbf{cone}\{f_i\}$, and that every element of either set takes only nonnegative values on $\{x \in \mathbb{R}^n, f_i(x) \geq 0\}$.

Theorem 3 ([Put93])**.** *Consider a set* $K = \{x \in \mathbb{R}^n \mid f_i(x) \geq 0\}$*, such that there exists a* $q \in \mathbf{preprime}\{f_1, \ldots, f_m\}$ *with* $\{x \in \mathbb{R}^n, q(x) \geq 0\}$ *compact (this implies that* K *itself is compact). Then,* $p(x) > 0$ *on* K *if and only if* $p(x) \in \mathbf{preprime}\{f_1, \ldots, f_m\}$.

Notice that here, the polynomial q serves as an algebraic certificate of the compactness of K, so in this case the Psatz is not needed.

Putinar's theorem was used by Lasserre to define a hierarchy of semidefinite relaxations for polynomial optimization, based on the dual moment interpretation [Las01].

1.3 An elementary argument for the n-sphere

Complete	ToDo

1.4 Tradeoffs

In principle (and often, in practice) there is a tradeoff between how "expressive" our family of certificates is, the quality of the resulting bounds, and the complexity of finding proofs.

On one extreme, the most general method is the Psatz, as it encapsulates pretty much every possible "algebraic deduction," and will certainly provide the strongest bounds, since it includes the other techniques as special cases. For optimization, Schmüdgen's theorem provides the advantages

[1] Recall that in practice, this may not be a issue at all, since the restriction on the degree of the certificates imposes a strict limit on how many products can be included.

of a linear representation, although (possibly) at the cost of having a large number of products between the constraints. Finally, the Putinar approach has a reduced number of constraints (and thus, SOS multipliers), although the obtained bounds can potentially be much weaker than the previous ones.

In the end, the decision concerning what approach to use should be dictated by the available computational resources, i.e., the size of the SDPs that we can solve in a reasonable time. It is not difficult to produce examples with significant gaps between the corresponding bounds; see for instance [Ste96] for a particularly simple example, that is trivial for the Psatz, but for which either the Schmüdgen or Putinar representations need large degree refutations.

Example 4. *In [Ste96], Stengle presented an interesting example to assess the computational requirements of Schmüdgen's theorem. His concrete example was to find a representation certifying the nonnegativity of $f(x) := 1 - x^2$ over $g(x) := (1 - x^2)^3 \geq 0$.*

The Positivstellansatz gives a very simple certificate of this property, or equivalently, the emptiness of the set $\{g(x) \geq 0, -f(x) \geq 0, zf(x) - 1 = 0\}$ (where we have used, as before, Rabinowitch's trick). Indeed, we have the identity:

$$z^4 \cdot (-f) \cdot g + (zf - 1) \cdot (z^3 f^3 + z^2 f^2 + zf + 1) = -1.$$

Using a simple argument, Stengle proved in [Ste96], that no representation of the form (1) exists when $\gamma = 0$.

$$(1 - x^2) + \gamma = Q(x) + P(x)(1 - x^2)^3, \tag{1}$$

where $Q(x), P(x)$ are sums of squares. To see this, evaluate this expression at $x = \pm 1$.

Furthermore, he has shown that $\gamma \to 0$, the degrees of P, Q satisfying the identity necessarily have to go to infinity, and provided the bounds $O(\gamma^{-\frac{1}{2}}) \leq deg(P) \leq O(\gamma^{-\frac{1}{2}} \log \frac{1}{\gamma})$.

As an interesting aside, it can be shown that the optimal solution of this problem can be exactly computed:

Theorem 5. *Let the degree of $P(x)$ be equal to $4N$. Then, the optimal solution that minimizes γ in (1) has:*

$$\gamma_N^* = \frac{1}{(2N+2)^2 - 1}, \quad P(x) = p(x)^2, \quad Q(x) = q(x)^2$$

where

$$\begin{aligned} p(x) &= 2(N+1)\ {}_2F_1(-N, N+2\,;\, \frac{1}{2}; x^2) \\ q(x) &= \frac{1}{\gamma_N^*}\, x\ {}_2F_1(-N-1, N+1\,;\, \tfrac{3}{2}; x^2) \end{aligned}$$

and ${}_2F_1(a, b; c, x)$ is the standard Gauss hypergeometric function [AS64, Chapter 15].

1.5 Trigonometric case

Megretski [Meg03] analyzed the case of trigonometric polynomials, and gave a simple argument for the existence of certain distinguished representations. We introduce the following notation: let $\mathbb{T}_n = \{z \in \mathbb{C}^n, |z_i| = 1\}$ be the n-dimensional torus, P_n is the set of multivariate Laurent polynomials, and $RP_n \subset P_n$ are the Laurent polynomials that are real-valued on $\mathbb{T}_n$.

Theorem 6 ([Meg03]). *Let $\{F, Q_1, \ldots, Q_m\} \subset RP_n$, such that $F(z) > 0$ for all $z \in \mathbb{T}_n$ satisfying $Q_1(z) = \cdots = Q_m(z) = 0$. Then there exist $V_1, \ldots, V_r \in P_n$, $H_1, \ldots, H_m \in RP_n$, such that*

$$F(z) = \sum_{i=1}^{r} |V_i(z)|^2 + \sum_{j=1}^{m} H_j(z) Q_i(z).$$

By splitting into real and imaginary part, this corresponds to a special kind of (standard) polynomials, and a compact semialgebraic set (so in principle, any of the previous theorems would apply). However, this result exploits the complex structure for a more concise representation.

This theorem deals only with the equality case (no inequalities), and the feasible set is compact (since so it $\mathbb{T}^n$). It essentially states that a positive polynomial is a sum of squares modulo the ideal generated by the Q_i. Recall we have proved similar results in the zero-dimensional case, and this theorem naturally generalizes these.

Megretski's proof is purely functional-analytic, the main tools being Bochner's theorem and Hahn-Banach. If $(G, +)$ is an Abelian group, a function $\phi : G \to \mathbb{C}$ is *positive definite* if the matrix $[\phi(g_j - g_k)]_{j,k=1}^n$ is Hermitian positive semidefinite for all finite subsets $\{g_1, \ldots, g_n\} \subseteq G$. Bochner's theorem, an important result in harmonic analysis, characterizes positive definite functions in terms of the nonnegativity of their Fourier transform:

Theorem 7 (Bochner). *A function $\phi : G \to \mathbb{C}$ is positive definite if and only if there exists a nonnegative measure μ on the dual group $\hat{G}$ such that*

$$\phi(g) = \int_{\hat{G}} \rho(g) d\mu(\rho).$$

Recall that the elements of the dual group $\hat{G}$ are the *characters* (i.e., the group homomorphisms $\rho : G \to \mathbb{T}$, satisfying $\rho(g_1 + g_2) = \rho(g_1)\rho(g_2)$). Thus, Bochner's theorem says that every positive definite function on an Abelian group is a convex combination of characters. In our case, the group $G = \mathbb{Z}^n$ can be identified with the set of monomials, and the corresponding dual group is $\hat{G} = \mathbb{T}_n$.

In simplified terms, one reason why trigonometric (or Laurent) polynomials are somewhat "easier" than the general case is because in this case monomials form a *group*, as opposed to the *semigroup* structure of regular monomials. For the group case, the corresponding theory is the classical harmonic analysis on abelian groups (e.g., [Rud90]); while for semigroups there is the newer, but well-developed characterizations of positive definite functions on (Abelian) semigroups; see for instance [BCR84].

We also mention that there are "purely algebraic" versions of these theorems, that do not use functional analytic ideas (e.g., [Mar00]). Roughly, the role played by the compactness of K in proving the boundedness of the operators T_{x_i} is replaced with a property called *Archimedeanity* of the corresponding preorder.

References

[AS64] M. Abramowitz and I.A. Stegun, editors. *Handbook of Mathematical Functions.* Dover, 1964.

[BCR84] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic analysis on semigroups*, volume 100 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1984.

[BCR98] J. Bochnak, M. Coste, and M-F. Roy. *Real Algebraic Geometry.* Springer, 1998.

[Las01] J. B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM J. Optim.*, 11(3):796–817, 2001.

[Mar00] M. Marshall. *Positive polynomials and sums of squares.* Dottorato de Ricerca in Matematica. Dept. di Mat., Univ. Pisa, 2000.

[Meg03] A. Megretski. Positivity of trigonometric polynomials. In *Proceedings of the 42^{th} IEEE Conference on Decision and Control*, pages 3814–3817, 2003.

[Par00] P. A. Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization.* PhD thesis, California Institute of Technology, May 2000. Available at `http://resolver.caltech.edu/CaltechETD:etd-05062004-055516`.

[Par03] P. A. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Math. Prog.*, 96(2, Ser. B):293–320, 2003.

[PD01] A. Prestel and C. N. Delzell. *Positive polynomials: from Hilbert's 17th problem to real algebra.* Springer Monographs in Mathematics. Springer, 2001.

[Put93] M. Putinar. Positive polynomials on compact semi-algebraic sets. *Indiana Univ. Math. J.*, 42(3):969–984, 1993.

[Rud90] W. Rudin. *Fourier analysis on groups.* Wiley Classics Library. John Wiley & Sons Inc., New York, 1990.

[Sch91] K. Schmüdgen. The K-moment problem for compact semialgebraic sets. *Math. Ann.*, 289:203–206, 1991.

[Ste96] G. Stengle. Complexity estimates for the Schmüdgen Positivstellensatz. *J. Complexity*, 12(2):167–174, 1996.

MIT 6.7230 - Algebraic techniques and semidefinite optimization May 3, 2024

Lecture 19

Lecturer: Pablo A. Parrilo *Scribe: ???*

In this lecture we study techniques to exploit symmetries that may be present in semidefinite programming problems, particularly those arising from sum of squares decompositions [GP04]. For this, we present the basic elements of the representation theory of finite groups. There are many possible applications of these ideas in different fields; for the case of Markov chains, see [BDPX05]. The celebrated Delsarte linear programming upper bound for codes (and generalizations by Levenshtein, McEliece, etc., [DL98]) can be understood as a natural symmetry reduction of the SDP relaxations based on the Lovász theta function; see e.g. [Sch79].

1 Groups and their representations

The representation theory of finite groups is a classical topic; good descriptions are given in [FS92, Ser77]. We concentrate here on the case of finite groups; extensions to compact groups are relatively straightforward.

Definition 1. *A* group *consists of a set G and a binary operation "$\cdot$" defined on G, for which the following conditions are satisfied:*

1. *Associative: $(a \cdot b) \cdot c = a \cdot (b \cdot c)$, for all $a, b, c \in G$.*
2. *Identity: There exist $1 \in G$ such that $a \cdot 1 = 1 \cdot a = a$, for all $a \in G$.*
3. *Inverse: Given $a \in G$, there exists $b \in G$ such that $a \cdot b = b \cdot a = 1$.*

We consider a finite group G, and an n-dimensional vector space V. We also consider the associated (infinite) group $GL(V)$ of nonsingular linear transformations of V, which we can interpret as the set of invertible $n \times n$ matrices. A *linear representation* of the group G is a homomorphism $\rho : G \to GL(V)$. In other words, we have a mapping from the group into linear transformations of V, that respects the group structure, i.e.

$$\rho(st) = \rho(s)\rho(t) \quad \forall s, t \in G.$$

The *dimension* of the representation ρ is the dimension of V.

Example 2. *Let $\rho(g) = 1$ for all $g \in G$. This is the* trivial *representation of the group. It is a one-dimensional representation.*

Example 3. *For a more interesting example, consider the symmetric group S_n, and the "natural" representation $\rho : S_n \to GL(\mathbb{C}^n)$, where $\rho(g)$ is a permutation matrix. For instance, for the group of permutations of two elements, $S_2 = \{e, g\}$, where $g^2 = e$, we have*

$$\rho(e) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \rho(g) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

The representation given in Example 3 has an interesting property. The set of matrices $\{\rho(e), \rho(g)\}$ have common invariant subspaces (other than the trivial ones, namely $\{\mathbf{0}\}$ and $\mathbb{C}^2$). Indeed, we can easily verify that the (orthogonal) one-dimensional subspaces given by (t, t) and $(t, -t)$ are invariant under the action of these matrices. Therefore, the restriction of ρ to those subspaces also gives representations of the group G. In this case, the one corresponding to the subspace (t, t) is "equivalent" (in a well-defined sense) to the trivial representation described in Example 2. The other subspace $(t, -t)$ gives the one-dimensional *alternating* representation of S_2, namely $\rho_A(e) = 1, \rho_A(g) = -1$. Thus, the representation ρ decomposes as $\rho = \rho_T \oplus \rho_A$, a direct sum of the trivial and the alternating representations.

As we will see, the same ideas extend to arbitrary finite groups.

Definition 4. *An* irreducible representation *of a group is a linear representation with no nontrivial invariant subspaces.*

Definition 5. *Two representations $\rho_1 : G \to GL(V_1)$ and $\rho_2 : G \to GL(V_2)$ are* equivalent *if there exists a vector space isomorphism $T : V_1 \to V_2$ such that*

$$\rho_1(g) = T^{-1}\rho_2(g)T \qquad \forall g \in G.$$

Theorem 6. *Every finite group G has a finite number of nonequivalent complex irreducible representations ρ_i, of dimension d_i. The relation $\sum_i d_i^2 = |G|$ holds.*

Example 7. *Consider the group S_3 (permutations in three elements). This group is generated by the two permutations $s : 123 \to 213$ and $c : 123 \to 312$ ("swap" and "cycle"), and has six elements $\{e, s, c, c^2, cs, sc\}$. Notice that $c^3 = e, s^2 = e$, and $s = csc$.*

The group S_3 has three irreducible representations, two one-dimensional, and one two-dimensional (so $1^2 + 1^2 + 2^2 = |S_3| = 6$). These are:

$$\begin{aligned} &\rho_T(s) = 1, && \rho_T(c) = 1 \\ &\rho_A(s) = -1, && \rho_A(c) = 1 \\ &\rho_S(s) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, && \rho_S(c) = \begin{bmatrix} \omega & 0 \\ 0 & \omega^2 \end{bmatrix} \end{aligned}$$

where $\omega = e^{\frac{2\pi}{3}i}$ is a cube root of 1. Notice that it is enough to specify a representation on the generators of the group.

Example 8. *Consider the cyclic group C_n (cyclic permutations of n elements). This group has a single generator c, which satisfies $c^n = e$. Explicitly, the group elements are $\{e, c, c^2, \ldots, c^{n-1}\}$.*

The group C_n has n (complex) irreducible representations $\rho_0, \ldots, \rho_{n-1}$, given by $\rho_k(c) = \omega_k$, where $\omega_k = e^{k\frac{2\pi}{n}j}$ for $k = 0, \ldots, n-1$. All representations are one-dimensional, satisfying $1^2 + \cdots + 1^2 = |C_n| = n$.

2 Symmetry and convexity

Optimization problems that invariant under the action of a symmetry group appear quite often in applications. Whenever these problems are convex, many simplifications are possible (in fact, even for nonconvex problems it is possible to exploit symmetry).

A key property of symmetric *convex* sets is the fact that the "group average" $\frac{1}{|G|}\sum_{g\in G}\sigma(g)x$ always belongs to the set.

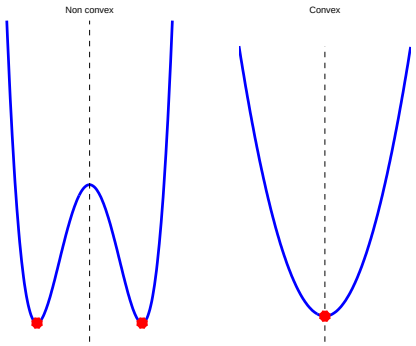

Figure 1: Two symmetric optimization problems, one non-convex and the other convex. For the latter, optimal solutions always lie on the fixed-point subspace.

Therefore, in convex optimization we can always restrict the solution to the fixed-point subspace

$$\mathcal{F} := \{x | \sigma(g)x = x, \quad \forall g \in G\}.$$

In other words, for convex problems, no "symmetry-breaking" is ever necessary.

As another interpretation, that will prove useful later, the "natural" decision variables of a symmetric optimization problem are the *orbits*, not the points themselves. Thus, we may look for solutions in the quotient space.

2.1 Invariant SDPs

We consider a general SDP, described in geometric form. If $\mathcal{L}$ is an affine subspace of $\mathcal{S}^n$, and $C, X \in \mathcal{S}^n$, an SDP is given by:

$$\min \langle C, X \rangle \qquad \text{s.t.} \quad X \in \mathcal{X} := \mathcal{L} \cap \mathcal{S}^n_+.$$

Definition 9. *Given a finite group G, and associated representation $\sigma : G \to GL(\mathcal{S}^n)$, a* σ-invariant SDP *is one where both the feasible set and the cost function are invariant under the group action, i.e.,*

$$\langle C, X \rangle = \langle C, \sigma(g)X \rangle, \qquad \forall g \in G, \qquad\qquad X \in \mathcal{X} \Rightarrow \sigma(g)X \in \mathcal{X} \qquad \forall g \in G$$

Example 10. *Consider the SDP given by*

$$\min a + c, \qquad s.t. \quad \begin{bmatrix} a & b \\ b & c \end{bmatrix} \succeq 0,$$

which is invariant under the Z_2 action:

$$\begin{bmatrix} X_{11} & X_{12} \\ X_{12} & X_{22} \end{bmatrix} \to \begin{bmatrix} X_{22} & -X_{12} \\ -X_{12} & X_{11} \end{bmatrix}.$$

Usually in SDP, the group acts on $\mathcal{S}^n$ through a congruence transformation, i.e., $\sigma(g)M = \rho(g)^T M \rho(g)$, where ρ is a representation of G on $\mathbb{C}^n$. In this case, the restriction to the fixed-point subspace takes the form:

$$\sigma(g)M = M \qquad \Longrightarrow \qquad \rho(g)M - M\rho(g) = 0, \quad \forall g \in G. \tag{1}$$

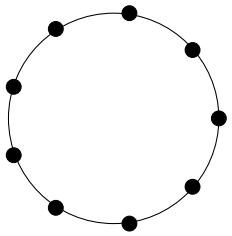

Figure 2: The cyclic graph C_n in n vertices (here, $n = 9$).

The Schur lemma of representation theory exactly characterizes the matrices that commute with a group action.

As a consequence of an important structural result (Schur's lemma), it turns out that every representation can be written in terms of a finite number of primitive blocks, the *irreducible representations* of a group.

Theorem 11. *Every group representation ρ decomposes as a direct sum of irreducible representations:*

$$\rho = m_1\vartheta_1 \oplus m_2\vartheta_2 \oplus \cdots \oplus m_N\vartheta_N$$

where $m_1, \ldots, m_N$ are the multiplicities.

This decomposition induces an isotypic decomposition of the space

$$\mathbb{C}^n = V_1 \oplus \cdots \oplus V_N, \quad V_i = V_{i1} \oplus \cdots \oplus V_{in_i}.$$

In the symmetry-adapted basis, the matrices in the SDP have a block diagonal form:

$$(I_{m_1} \otimes M_1) \oplus \ldots \oplus (I_{m_N} \otimes M_N)$$

In terms of our symmetry-reduced SDPs, this means that not only the SDP block-diagonalizes, but there is also the possibility that many blocks are identical.

2.2 Example: symmetric graphs

Consider the MAXCUT problem on the cycle graph C_n with n vertices (see Figure 2). It is easy to see that the optimal cut has cost equal to n or $n-1$, depending on whether n is even or odd, respectively. What would the SDP relaxation yield in this case? If A is the adjacency matrix of the graph, then the SDP relaxations have essentially the form

$$\begin{array}{rl} \text{minimize} & \operatorname{Tr} AX \\ \text{s.t.} & X_{ii} = 1 \\ & X \succeq 0 \end{array} \qquad\qquad \begin{array}{rl} \text{maximize} & \operatorname{Tr} \Lambda \\ \text{s.t.} & A \succeq \Lambda \\ & \Lambda \text{ diagonal} \end{array} \tag{2}$$

By the symmetry of the graph, the matrix A is *circulant*, i.e., $A_{ij} = a_{i-j \bmod n}$.

We focus now on the dual form. It should be clear that the cyclic symmetry of the graph induces a cyclic symmetry in the SDP, i.e., if $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is a feasible solution, then $\tilde{\Lambda} = \operatorname{diag}(\lambda_n, \lambda_1, \lambda_2, \ldots, \lambda_{n-1})$ is also feasible and achieves the same objective value. Thus, by averaging over the cyclic group, we can always restrict D to be a multiple of the identity matrix,

i.e., $\Lambda = \lambda I$. Furthermore, the constraint $A \succeq \lambda I$ can be block-diagonalized via the Fourier matrix (i.e., the irreducible representations of the cyclic group), yielding:

$$A \succeq \lambda I \qquad \Leftrightarrow \qquad 2\cos\frac{k\pi}{n} \geq \lambda \qquad k = 0, \ldots, n-1.$$

From this, the optimal solution of the relaxation can be directly computed, yielding the exact expressions for the upper bound on the size of the cut

$$mc(C_n) \leq SDP(C_n) = \begin{cases} n & n \text{ even} \\ n\cos^2\frac{\pi}{2n} & n \text{ odd.} \end{cases}$$

Although this example is extremely simple, exactly the same techniques can be applied to much more complicated problems; see for instance [PP04, dKMP$^+$06, Sch05, BV08] for some recent examples.

2.3 Example: even polynomials

Another (but illustrative) example of symmetry reduction is the case of SOS decompositions of even polynomials. Consider a polynomial $p(x)$ that is *even*, i.e., it satisfies $p(x) = p(-x)$. Does this symmetry help in making the computations more efficient?

2.4 Benefits

In the case of semidefinite programming, there are many benefits to exploiting symmetry:

- Replace one big SDP with smaller, coupled problems.
- Instead of checking if a big matrix is PSD, we use one copy of each repeated block (constraint aggregation).
- Eliminates multiple eigenvalues (numerical difficulties).
- For groups, the coordinate change depends only on the group, and not on the problem data.
- Can be used as a general preprocessing scheme. The coordinate change T is unitary, so well-conditioned.

As we will see in the next section, this approach can be extended to more general algebras that do not necessarily arise from groups.

2.5 Sum of squares

In the case of SDPs arising from sum of squares decompositions, a parallel theory can be developed by considering the symmetry-induced decomposition of the full polynomial ring $\mathbb{R}[x]$. Since a complete explanation involves some elements of invariant theory, we omit the details here; see [GP04] for the full story.

Example 12. *Consider the (non-convex) quartic trivariate polynomial:*

$$p(x, y, z) = x^4 + y^4 + z^4 - 4xyz + x + y + z.$$

This polynomial is invariant under all permutations of $\{x, y, z\}$ (the full symmetric group S_3). The global minimum of p is $p_\star \approx -2.1129$, and is achieved at the orbit of global minimizers:

$$(0.988, -1.102, -1.102)\,,\ (-1.102, 0.988, -1.102)\,,\ (-1.102, -1.102, 0.988).$$

For this polynomial, it holds that $p_{sos} = p_\star$.

We show now how to compute p_{sos} by exploiting symmetry. Since $p(x, y, z)$ has $n = 3$ variables, degree $2d = 4$, and a full Newton polytope, its standard sos formulation is indexed by all $\binom{n+d}{d} = \binom{5}{2} = 10$ monomials of degree 2, i.e.,

$$p(x,y,z) - \gamma = \begin{bmatrix} 1 \\ x \\ y \\ z \\ x^2 \\ y^2 \\ z^2 \\ yz \\ xz \\ xy \end{bmatrix}^T \begin{bmatrix} q_{00} & q_{01} & q_{02} & q_{03} & q_{04} & q_{05} & q_{06} & q_{07} & q_{08} & q_{09} \\ q_{01} & q_{11} & q_{12} & q_{13} & q_{14} & q_{15} & q_{16} & q_{17} & q_{18} & q_{19} \\ q_{02} & q_{12} & q_{22} & q_{23} & q_{24} & q_{25} & q_{26} & q_{27} & q_{28} & q_{29} \\ q_{03} & q_{13} & q_{23} & q_{33} & q_{34} & q_{35} & q_{36} & q_{37} & q_{38} & q_{39} \\ q_{04} & q_{14} & q_{24} & q_{34} & q_{44} & q_{45} & q_{46} & q_{47} & q_{48} & q_{49} \\ q_{05} & q_{15} & q_{25} & q_{35} & q_{45} & q_{55} & q_{56} & q_{57} & q_{58} & q_{59} \\ q_{06} & q_{16} & q_{26} & q_{36} & q_{46} & q_{56} & q_{66} & q_{67} & q_{68} & q_{69} \\ q_{07} & q_{17} & q_{27} & q_{37} & q_{47} & q_{57} & q_{67} & q_{77} & q_{78} & q_{79} \\ q_{08} & q_{18} & q_{28} & q_{38} & q_{48} & q_{58} & q_{68} & q_{78} & q_{88} & q_{89} \\ q_{09} & q_{19} & q_{28} & q_{39} & q_{49} & q_{59} & q_{69} & q_{79} & q_{89} & q_{99} \end{bmatrix} \begin{bmatrix} 1 \\ x \\ y \\ z \\ x^2 \\ y^2 \\ z^2 \\ yz \\ xz \\ xy \end{bmatrix},$$

where the matrix Q above will be constrained to be positive semidefinite. Recall that p is invariant under all permutation of the variables (the full symmetric group S_3). Thus, we can constrain the matrix Q to be in the fixed-point subspace, i.e., it should satisfy $Q = \rho(g)^T Q \rho(g)$, where $g \in G$ and $\rho : G \to GL(\mathbb{R}^{10})$ is the induced representation on the vector of monomials that arises from permuting the variables (x, y, z). Solving the equations that define the fixed-point subspace, we find that the Gram matrix must have the structure

$$\widehat{Q} = \begin{bmatrix} r_0 & r_1 & r_1 & r_1 & r_2 & r_2 & r_2 & r_3 & r_3 & r_3 \\ r_1 & r_4 & r_5 & r_5 & r_6 & r_7 & r_7 & r_8 & r_9 & r_9 \\ r_1 & r_5 & r_4 & r_5 & r_7 & r_6 & r_7 & r_9 & r_8 & r_9 \\ r_1 & r_5 & r_5 & r_4 & r_7 & r_7 & r_6 & r_9 & r_9 & r_8 \\ r_2 & r_6 & r_7 & r_7 & r_{10} & r_{11} & r_{11} & r_{12} & r_{13} & r_{13} \\ r_2 & r_7 & r_6 & r_7 & r_{11} & r_{10} & r_{11} & r_{13} & r_{12} & r_{13} \\ r_2 & r_7 & r_7 & r_6 & r_{11} & r_{11} & r_{10} & r_{13} & r_{13} & r_{12} \\ r_3 & r_8 & r_9 & r_9 & r_{12} & r_{13} & r_{13} & r_{14} & r_{15} & r_{15} \\ r_3 & r_9 & r_8 & r_9 & r_{13} & r_{12} & r_{13} & r_{15} & r_{14} & r_{15} \\ r_3 & r_9 & r_9 & r_8 & r_{13} & r_{13} & r_{12} & r_{15} & r_{15} & r_{14} \end{bmatrix}. \tag{3}$$

Notice that the fixed-point subspace is 16-dimensional, as opposed to the $\binom{11}{2} = 55$ degrees of freedom in the original matrix.

We can now, however, give a nicer description of this subspace. Consider the coordinate transformation (a symmetry-adapted basis) of the form $X \mapsto T^T X T$, where the orthogonal matrix T is given by

$$T = \mathit{BlockDiag}(1, R, R, R) \cdot \Pi, \qquad R = \begin{bmatrix} \alpha & \alpha & \alpha \\ \alpha & \beta & \gamma \\ \alpha & \gamma & \beta \end{bmatrix},$$

where $\alpha = 1/\sqrt{3}$, $\beta = (3 - \sqrt{3})/6$, $\gamma = -(3 + \sqrt{3})/6$, and Π is the permutation matrix satisfying $\Pi^T[x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9] = [x_0, x_1, x_4, x_7, x_2, x_5, x_8, x_3, x_6, x_9]$. It can be verified that

under this tranformation, the matrix in (3) *takes now the form*

$$T^T \widehat{Q} T = \mathit{BlockDiag}(Q_1, Q_2, Q_2),$$

where

$$Q_1 = \begin{bmatrix} r_0 & \sqrt{3}r_1 & \sqrt{3}r_2 & \sqrt{3}r_3 \\ \sqrt{3}r_1 & r_4 + 2r_5 & r_6 + 2r_7 & r_8 + 2r_9 \\ \sqrt{3}r_2 & r_6 + 2r_7 & r_{10} + 2r_{11} & r_{12} + 2r_{13} \\ \sqrt{3}r_3 & r_8 + 2r_9 & r_{12} + 2r_{13} & r_{14} + 2r_{15} \end{bmatrix},$$

$$Q_2 = \begin{bmatrix} r_4 - r_5 & r_6 - r_7 & r_8 - r_9 \\ r_6 - r_7 & r_{10} - r_{11} & r_{12} - r_{13} \\ r_8 - r_9 & r_{12} - r_{13} & r_{14} - r_{15} \end{bmatrix}.$$

Notice that the 10×10 *matrix has split in three blocks, one of size* 4×4*, and two identical blocks of size* 3×3*. Also, all entries are otherwise linearly independent (in fact, we have the dimension count* $\binom{5}{2} + \binom{4}{2} = 10 + 6 = 16$*, the number of free parameters in* (3)*).*

Since $\widehat{Q} \succeq 0$ *if and only if* $T^T \widehat{Q} T \succeq 0$*, this implies that instead of solving a semidefinite programming problem with a positivity constraint on a* 10×10 *matrix, we have now a* 4×4 *and a* 3×3 *matrix instead (clearly, we only need one copy of the two identical* 3×3 *blocks), which is a lot simpler.*

3 Algebra decomposition

An alternative (and somewhat more general) approach can be obtained by focusing instead on the *associative algebra* generated by the matrices in a semidefinite program.

Definition 13. *An* associative algebra $\mathcal{A}$ *over* $\mathbb{C}$ *is a vector space with a* $\mathbb{C}$*-bilinear operation* $\cdot : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$ *that satisfies*

$$x \cdot (y \cdot z) = (x \cdot y) \cdot z, \qquad \forall x, y, z \in \mathcal{A}.$$

In general, associative algebras do not need to be commutative (i.e., $x \cdot y = y \cdot x$). However, that is an important special case, with many interesting properties. Important examples of finite dimensional associative algebras are:

- Full matrix algebra $\mathbb{C}^{n \times n}$, standard product.
- The subalgebra of square matrices with equal row and column sums.
- The n-dimensional algebra generated by a single $n \times n$ matrix.
- The group algebra: formal $\mathbb{C}$-linear combination of group elements.
- Polynomial multiplication modulo a zero dimensional ideal.
- The Bose-Mesner algebra of an association scheme.

We have already encountered some of these, when studying the companion matrix and its generalizations to the multivariate case. A particularly interesting class of algebras (for a variety of reasons) are the *semisimple* algebras.

Definition 14. *The* radical *of an associative algebra* $\mathcal{A}$*, denoted rad*$(\mathcal{A})$*, is the intersection of all maximal left ideals of* $\mathcal{A}$*.*

Definition 15. *An associative algebra $\mathcal{A}$ is* semisimple *if $Rad(\mathcal{A}) = 0$.*

For a semidefinite programming problem in standard (dual) form

$$\max b^T y \qquad \text{s.t.} \quad A_0 - \sum_{i=1}^{m} A_i y_i \succeq 0,$$

we consider the algebra generated by the A_i.

Theorem 16. *Let $\{A_0, \ldots, A_m\}$ be given* symmetric *matrices, and $\mathcal{A}$ the generated associative algebra. Then, $\mathcal{A}$ is a semisimple algebra.*

Semisimple algebras have a very nice structure, since they are essentially the direct sum of much simpler algebras.

Theorem 17 (Wedderburn)**.** *Every finite dimensional semisimple associative algebra over $\mathbb{C}$ can be decomposed as a direct sum*

$$\mathcal{A} = \mathcal{A}_1 \oplus \mathcal{A}_2 \oplus \ldots \oplus \mathcal{A}_k.$$

Each $\mathcal{A}_i$ is isomorphic to a simple full matrix algebra.

Example 18. *A well-known example is the (commutative) algebra of circulant matrices, i.e., those of the form*

$$A = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ a_4 & a_1 & a_2 & a_3 \\ a_3 & a_4 & a_1 & a_2 \\ a_2 & a_3 & a_4 & a_1 \end{bmatrix}.$$

Circulant matrices are ubiquitous in many applications, such as signal processing. It is well-known that there exists a fixed *unitary coordinate change (the $n \times n$ discrete Fourier transform matrix with entries $F_{jk} = \frac{1}{\sqrt{n}}\omega^{jk}$ where ω is an n-root of unity) under which* all *matrices A are diagonal (with distinct scalar blocks). For instance, for the example above with $n = 4$, we have*

$$F = \frac{1}{\sqrt{4}} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{bmatrix}, \qquad F^* A F = \operatorname{diag}\left(\begin{bmatrix} a_1 + a_2 + a_3 + a_4 \\ a_1 - ia_2 - a_3 + ia_4 \\ a_1 - a_2 + a_3 - a_4 \\ a_1 + ia_2 - a_3 - ia_4 \end{bmatrix} \right).$$

Remark 19. *In general, any associative algebra is the direct sum of its radical and a semisimple algebra. For the n-dimensional algebra generated by a single matrix $A \in \mathbb{C}^{n \times n}$, we have that $A = S + N$, where S is diagonalizable, N is nilpotent, and $SN = NS$. Thus, this statement is essentially equivalent to the existence of the Jordan decomposition.*

References

[BDPX05] S. Boyd, P. Diaconis, P. A. Parrilo, and L. Xiao. Symmetry analysis of reversible Markov chains. *Internet Math.*, 2(1):31–71, 2005.

[BV08] C. Bachoc and F. Vallentin. New upper bounds for kissing numbers from semidefinite programming. *J. Amer. Math. Soc*, 21(3):909–924, 2008.

[dKMP+06] E. de Klerk, J. Maharry, D.V. Pasechnik, R.B. Richter, and G. Salazar. Improved bounds for the crossing numbers of $K_{m,n}$ and K_n. *SIAM Journal on Discrete Mathematics*, 20:189, 2006.

[DL98] P. Delsarte and VI Levenshtein. Association schemes and coding theory. *IEEE Transactions on Information Theory*, 44(6):2477–2504, 1998.

[FS92] A. Fässler and E. Stiefel. *Group Theoretical Methods and Their Applications*. Birkhäuser, 1992.

[GP04] K. Gatermann and P. A. Parrilo. Symmetry groups, semidefinite programs, and sums of squares. *Journal of Pure and Applied Algebra*, 192(1-3):95–128, 2004.

[PP04] P. A. Parrilo and R. Peretz. An inequality for circle packings proved by semidefinite programming. *Discrete and Computational Geometry*, 31(3):357–367, 2004.

[Sch79] A. Schrijver. A comparison of the Delsarte and Lovász bounds. *IEEE Transactions on Information Theory*, 25(4), 1979.

[Sch05] A. Schrijver. New code upper bounds from the Terwilliger algebra and semidefinite programming. *IEEE Transactions on Information Theory*, 51(8):2859–2866, 2005.

[Ser77] J.-P. Serre. *Linear Representations of Finite Groups*. Springer-Verlag, 1977.