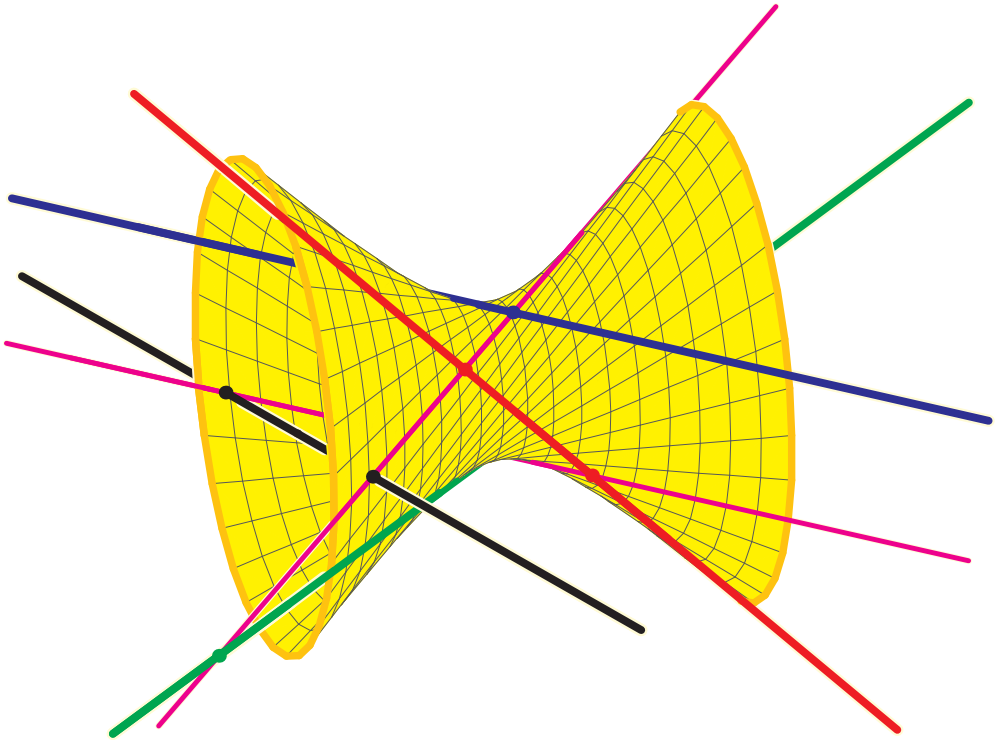
Applicable Algebraic Geometry

Frank Sottile
Thorsten Theobald



January 26, 2015

Contents

Notation and conventions

Global:

$\mathbb{R}$, $\mathbb{C}$, $\mathbb{Q}$	real, complex, rational numbers
Use $\mathbb{N}$, $\mathbb{Z}$	Natural numbers (do we include the 0?), integers
$\subset$	(not necess. strict) subset, ($\subseteq$ has to be avoided)
$\mathbb{K}$	ground field
n, m	natural numbers
$x_1, \ldots, x_n$	usual ring variables (lowercase letters, n variables)
$(a_1, \ldots, a_n)$, b, c	constants from field, sometime points in $\mathbb{K}^n$
X, Y, Z	Varieties
I	ideal
$\mathcal{V}$, $\mathcal{V}_{\mathbb{R}}$	variety, real variety
$\mathcal{I}(S)$	ideal of set S of polynomials
$\mathbb{P}^n(\mathbb{C})$ (?)	n-dim. projective space over $\mathbb{C}$

Chapter Positive polynomials and algebraic certificates:

$\Sigma[x]$	sums of squares in $x = (x_1, \ldots, x_n)$
K	compact feasible set (here: field always $\mathbb{R}$)
C^*	dual of a cone C, $C^* = \{y \in \mathbb{R}^n : \langle x, y\rangle \geq 0 \text{ for all } x \in C\}$,

Chapter Toric varieties

P°	polar polytope, $P^\circ = \{y \in \mathbb{R}^n : \langle x, y\rangle \leq 1 \text{ for all } x \in P\}$ (Frank's concern w.r.t. consistency of inward/outward normal vectors)
$\operatorname{int} P$, $\operatorname{relint} P$	interior, relative interior

Introduction

Recent years have seen many fundamental developments of algebraic geometry. Two wide-reaching, interlocked developments are concerned with *applying* ideas from algebraic geometry in various other disciplines inside and outside mathematics as well as with new and more intense *discrete* viewpoints and discrete revelations of algebraic geometry.

The purpose of this book is to provide a broad but friendly introduction to algebraic geometry, with a particular focus on *combinatorial*, *algorithmic*, and *real* aspects, as well as on having the potential to be *applicable* in various contexts. While the particular choice of topics treated in the text is clearly reflecting the interests of the authors, we think that the alignment to the preceding aspects provides an ubiquitous access for many challenging research directions.

Concerning the underlying ground field, we will both be concerned with algebraically closed fields as well as with the field of real numbers. The real numbers are important in applications, yet real number phenomena in algebraic geometry texts are rarely treated in a manner useful for applications. Our presentation adopts the point of view that one first understands the situation for the complex numbers, and then studies how things hange when restricted to the real points of varieties.

The book includes an introduction to the theory and use of powerful practical methods for solving and studying systems of polynomial equations, methods that have been developed by theoretical mathematicians, but which have yet to be incorporated into the standard toolkit of applied scientists. These include Gröbner bases and resultant-based methods in symbolic computation. Our goal is to facilitate the transfer of technology from theoretical mathematicians to applied scientists, and give an entry points into some applied research directions for the theoretical mathematician.

The emergence of effective computational tools and user-friendly software is changing the possibilities of applications. For example, we may now find explicit numerical solutions to systems of polynomial equations of moderate size—a previously intractable problem. Symbolic computation and numerical homotopy continuation not only provide these tools, but they also give insight into several key notions in algebraic geometry. New fundamental links to optimization and to discrete geometry have provided new tools for effective algebraic-geometric computations, and have given new approaches to learning algebraic geometry. Ideas from these areas will pervade our development of algebraic geometry.

In the following we provide a brief overview on the material covered in the book.

Chapter 1 we introduce to algebraic sets, focussing on the viewpoint of the defining equations, as well as provide some illustrative concrete examples for some key notions of our book. The goal of the chapter is to provide our readers - coming from different backgrounds - with sufficiently knowledge on the key players of the book: polynomials, polynomial equations, ideals, and algebraic varieties. In particular, we will be concerned with affine and projective varieties. Focussing on algebraically closed fields then, the chapter provides the classical dictionary translating algebraic into geometric notions and vice versa. This includes a treatment of classical topics including Hilbert's Nullstellensatz, the coordinate ring, and regular maps.

Chapter 2 provides some important common algorithmic concepts for handling polynomial equations. We discuss the primary concepts of the theory of Gröbner basis which have been introduced by Buchberger in 1965, which provide a way to establish a unique basis of a given polynomial ideal and which have provided a landmark for the development of symbolic computation. Gröbner basis are at the heart of many algorithms on polynomial ideals. We explain how Gröbner bases can be used to solve systems of polynomial equations from a symbolic point of view. In the chapter we also discuss classical as well as some modern aspects of resultants which provide a diffent access towards eliminating variables in polynomial systems and towards solving those systems. The chapter closes with a discussion of deformation techniques for polynomial equations, which have enabled the effective use of numerical techniques in algebraic geometry. These deformations establish important connections between algebraic geometry and polyhedral geometry, which includes the study of toric varieties studied in the later Chapter ...

In Chapter 3 we deepen the treatment of the structure of algebraic varieties. We discuss some topological properties, the decomposition of varieties as well as rational functions on varieties. In order to introduce the key notions of the degree and the dimension of a variety, we use a combinatorial access basedn on the Hilbert function and the Hilbert polynomial.

Chapter 4 develops fundamental ideas and algorithms from real algebraic and semialgebraic geometry. In the applied sciences real solutions are often much more important than complex ones, yet their understanding is often more difficult due to lacking algebraic closedness of the field of real numbers. Our treatment starts from a discussion of univariate polynomials, including the classical method of Sturm sequences to count the number of real zeroes of a univariate polynomial. We then investigate the basic properties of semialgebraic sets, which are defined by polynomial inequalities over the real numbers. Interwoven with the developments in semidefinite programming (as discussed in Chapter 5), within the last two decades it has turned out that semialgebraic sets which arise as linear matrix inequalities (so-called spectrahedra) or as the projections of linear matrix inequalities are particularly useful with regard to computational handling. The chapter also contains eigenvalue-based methods to study the complex and real roots of a zero-dimensional ideals. Effectively applying methods from linear algebra, these methods combine symbolic and numerical techniques to study and compute the solutions of

polynomial systems.

The question to certify (i.e., to provide a witness) for the non-negativity of a multivariate real polynomial has a distinguished history dating back to Minkowski and Hilbert and underlies Hilbert' 17th problem from his famous list of 23 problems from 1900. Statements which certify the emptyness of set defined by real polynomial equations or inequalities can be seen as analogues to Hilbert's Nullstellensatz in the algebraically closed case. Chapter 5 discusses the Positivstellensatz (due to Krivine and Stengle) as well as Putinar's and Schmüdgen's Theorems which provide the theoretical foundation for the striking connections between semialgebraic geometry presented in the subsequent Chapter 6. From the dual point of view, positive polynomials relate to the rich and classical world of moment problems.

Chapter 6 centers around the connection between algebraic geometry and optimization which has become very lively within the last decade. This link is established through polynomial optimization, sums of squares and semidefinite programming. The roots of these modern developments go back to N.Z. Shor (in the 80s) and were substantially advanced by Parrilo and Lasserre (around the year 2000). While some fundamental theorems connecting nonnegative polynomials with sums of squares already date back to Hilbert, the modern development have recognized that sums of squares can be computationally handled much better than nonnegative polynomials and that that general idea can also be effectively applied to rather general constrained polynomial optimization problems. The computational engine behind sums of squares computation is semidefinite programming which can be seen as linear programming over the cone of positive semidefinite mattrices. The chapter both provides theoretical as well as practical issues of sums of squares based relaxation schemes for polynomial optimization.

Chapter ... deals with toric varieties ...

Chapters 9 and 10 provide an access to the field of tropical geometry. Within the last decade, under this term several research directions of various mathematical subdisciplines have fruitfully found together. From a combinatorial viewpoint, tropical geometry can be seen as the geometry of the semiring $(\mathbb{R}, \max, +)$ (respectively $(\mathbb{R}, \min, +)$). Tropical hypersurfaces are polyhedral complexes in Euclidean space. From the algebraic viewpoint, tropical geometry replaces complex toric varieties by linear spaces and complex algebraic varieties by polyhedral complexes. The roots of tropical geometry origin in Bergmans logarithmic limit sets (in the 70s), Viro's patchworking method (in the late 70s), in the Maslov dequantization of real numbers (in the 80s) as well as the use of idempotent semiring in optimization and control theory).

In Chapter 9 we concentrate on the combinatorial viewpoint and on tropical hypersurfaces. We study the dual subdivision and provide polyhedral characterization. The we discuss in detail Kapranov's theorem which – for the case of hypersurfaces – connects the two distinct viewpoints on tropical geometry stated above.

In Chapter 10 introduced to the more general concepts of a tropical variety and of a tropical basis. We discuss the connection to amoebas which are the logarithmic images of algebraic varieties as well as the connection to the problem of counting complex algebraic

curves.

Chapter 11 investigates algebraic geometry problems arising from nonlinear aspects of computational geometry, geometric modeling, and computer vision. These problems often involve few variables, and thus avoid the complexity bottleneck of computational algebraic geometry. In particular, the chapter provides some insights on how the methods from earlier chapters can be applied, by means of some transversal and tangent problems from computational the Steward platform arising in mechanical engineering.

Since the book requires background from several areas we provide brief introductions to background topics in appendices.

Each chapter ends with some exercises as well as with notes in which historical aspects as well as pointers to more specialized literature is given.

Expected background

With its applied focus on algebraic geometry the book is intended for students, researchers as well as practioniers in pure and applied mathematics as well inclinded applied scientists and engineers.

We expect the reader to have some mathematical maturity beyond the undergraduate level, coupled with a standard undergraduate mathematics background, including linear algebra, some abstract algebra and analysis. We also assume a some familiarity with some basic concepts from computational algebra, as developed for example in popular undergraduate textbooks, such as Adams and Loustaunau [1] or by Cox, Little and O'Shea [20].

Since we review computational algebra, our book will also provide a (somewhat steep) introduction into this topic for those not already familiar with it.

Uses of this book

This book can serve as a textbook for a topics course for mathematics graduate students, or more advanced students from outside of mathematics. Parts of it have been used in this manner at Texas A&M University and at Goethe University Frankfurt.

While the first two chapters are rather essential for all the material and there is a slight increase in dependency throughout the book, there are various ways to pass through Chapters 3–10.

We recommend Chapter 3 for a focus on structural-based treatment, Chapters 4–6 (which depend sequentially on each other) for a focus on semialgebraic geometry and optimization, Chapters 7–9 (where in particular Chapters 9 depends on Chapter 8) for a focus on discrete aspects, and Chapter 10 for a glimpse to some geometric applications. See Figure ...

Chapter 1

Varieties

Algebraic geometry uses tools from algebra to study geometric sets called varieties, which are the common zeroes of a collection of polynomials. We develop some basic notions of algebraic geometry, perhaps the most fundamental being the dictionary between algebraic and geometric concepts. The basic objects we introduce and concepts we develop will be used throughout the book. These incude affine varieties, important notions from the algebra-geometry dictionary, and projective varieties. We provide additional algebraic background in the appendices and pointers to other sources of introductions to algebraic geometry in the references provided at the end of the chapter.

1.1 Affine Varieties

Let $\mathbb{K}$ be a field, which for us will almost always be either the complex numbers $\mathbb{C}$, the real numbers $\mathbb{R}$, or the rational numbers $\mathbb{Q}$. These different fields have their individual strengths and weaknesses. The complex numbers are *algebraically closed*; every univariate polynomial has a complex root. Algebraic geometry works best when using an algebraically closed field, and most introductory texts restrict themselves to the complex numbers. However, quite often real number answers are needed in applications. Because of this, we will often consider real varieties and work over $\mathbb{R}$. Symbolic computation provides many useful tools for algebraic geometry, but it requires a field such as $\mathbb{Q}$, which can be represented on a computer. Much of what we do remains true for arbitrary fields, such as the Gaussian rationals $\mathbb{Q}[i]$, or $\mathbb{C}(t)$, the field of rational functions in the variable t, or finite fields. We will at times use this added generality.

Algebraic geometry is fundamentally about the interplay of algebra and geometry, with its most basic objects the ring $\mathbb{K}[x_1, \ldots, x_n]$ of polynomials in indeterminates $x_1, \ldots, x_n$ with coefficients in $\mathbb{K}$, and the space $\mathbb{K}^n$ of n-tuples $a = (a_1, \ldots, a_n)$ of numbers from $\mathbb{K}$. We regard $\mathbb{K}^n$ as the domain of polynomials in $\mathbb{K}[x_1, \ldots, x_n]$, which are then functions from $\mathbb{K}^n \to \mathbb{K}$. We make our main definition.

Definition 1.1.1. An *affine variety* is the set of common zeroes of a collection of polynomials. Given a set $S \subset \mathbb{K}[x_1, \ldots, x_n]$ of polynomials, the affine variety defined by S is the set

$$\mathcal{V}(S) \;:=\; \{a \in \mathbb{K}^n \mid f(a) = 0 \quad \text{for } f \in S\}\,.$$

This is a(n *affine*) *subvariety* of $\mathbb{K}^n$ or simply a *variety* or *algebraic variety*.

If X and Y are varieties with $Y \subset X$, then Y is a *subvariety* of X. In Exercise 2, you will be asked to show that if $S \subset T$, then $\mathcal{V}(S) \supset \mathcal{V}(T)$.

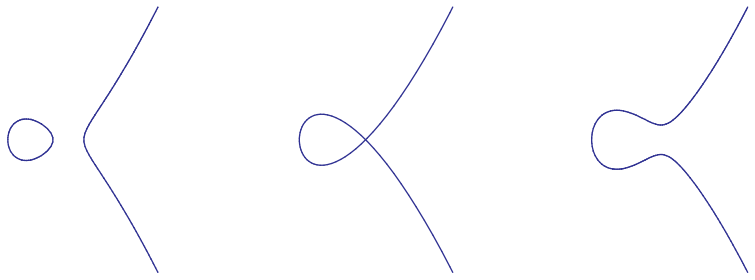
The empty set $\emptyset = \mathcal{V}(1)$ and affine space itself $\mathbb{K}^n = \mathcal{V}(0)$ are varieties. Any linear or affine subspace L of $\mathbb{K}^n$ is a variety. Indeed, an affine subspace L has an equation $Ax = b$, where A is a matrix and b is a vector, and so $L = \mathcal{V}(Ax - b)$ is defined by the linear polynomials which form the rows of the column vector $Ax - b$. An important special case is when $L = \{b\}$ is a point of $\mathbb{K}^n$. Writing $b = (b_1, \ldots, b_n)$, then L is defined by the equations $x_i - b_i = 0$ for $i = 1, \ldots, n$.

Any finite subset $Z \subset \mathbb{K}^1$ is a variety as $Z = \mathcal{V}(f)$, where

$$f \;:=\; \prod_{z \in Z} (x - z)$$

is the monic polynomial with simple zeroes in Z.

A non-constant polynomial $f(x, y)$ in the variables x and y defines a *plane curve* $\mathcal{V}(f) \subset \mathbb{K}^2$. Here are the plane cubic curves $\mathcal{V}(f + \frac{1}{20})$, $\mathcal{V}(f)$, and $\mathcal{V}(f - \frac{1}{20})$, where $f(x, y) := y^2 - x^3 - x^2$.



A *quadric* is a variety defined by a single quadratic polynomial. The smooth quadrics in $\mathbb{K}^2$ are the plane conics (circles, ellipses, parabolas, and hyperbolas in $\mathbb{R}^2$) and the smooth quadrics in $\mathbb{R}^3$ are the spheres, ellipsoids, paraboloids, and hyperboloids. Figure 1.1 shows a hyperbolic paraboloid $\mathcal{V}(xy + z)$ and a hyperboloid of one sheet $\mathcal{V}(x^2 - x + y^2 + yz)$.

These examples, finite subsets of $\mathbb{K}^1$, plane curves, and quadrics, are varieties defined by a single polynomial and are called *hypersurfaces*. Any variety is an intersection of hypersurfaces, one for each polynomial defining the variety.

The set of four points $\{(-2, -1), (-1, 1), (1, -1), (1, 2)\}$ in $\mathbb{K}^2$ is a variety. It is the

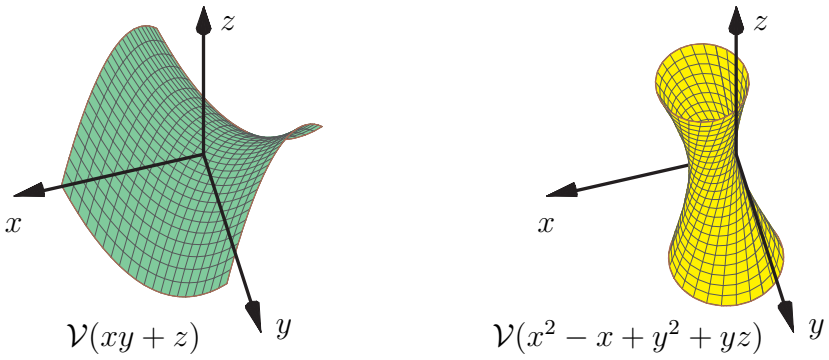
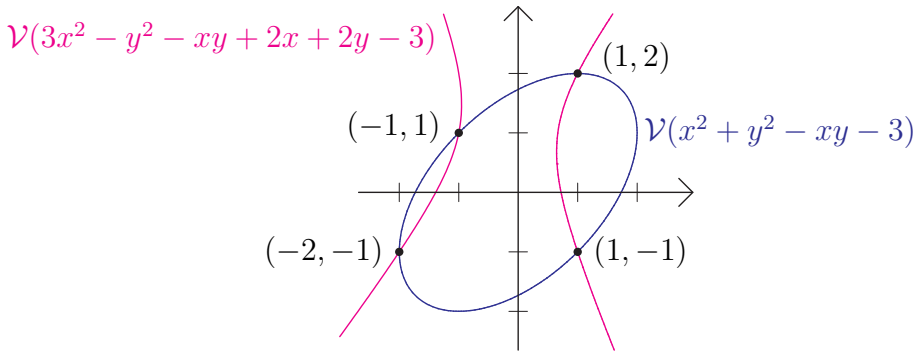


Figure 1.1: Two hyperboloids.

intersection of an ellipse $\mathcal{V}(x^2+y^2-xy-3)$ and a hyperbola $\mathcal{V}(3x^2-y^2-xy+2x+2y-3)$.



The quadrics of Figure 1.1 meet in the variety $\mathcal{V}(xy+z,\ x^2-x+y^2+yz)$, which is shown on the right in Figure 1.2. This intersection is the union of two space curves. One is the

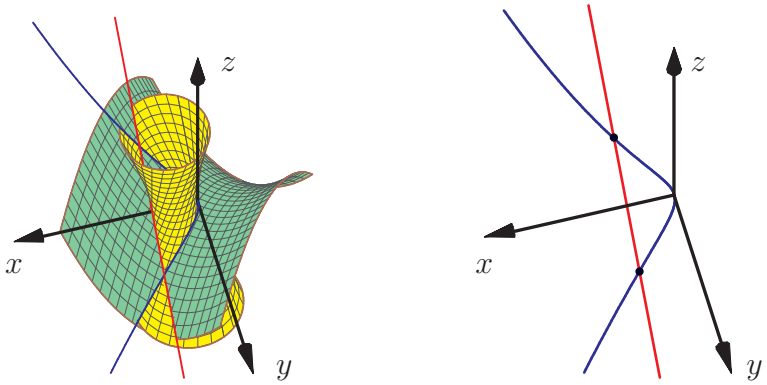


Figure 1.2: Intersection of two quadrics.

line $x = 1, y + z = 0$, while the other is the cubic space curve which has parametrization $t \mapsto (t^2, t, -t^3)$. Observe that the sum of the degrees of these curves, 1 (for the line) and 3 (for the space cubic) is equal to the product $2 \cdot 2$ of the degrees of the quadrics defining the intersection.

The intersection of the hyperboloid $x^2+(y-\frac{3}{2})^2-z^2 = \frac{1}{4}$ with the sphere $x^2+y^2+z^2 = 4$ is a singular space curve (the figure ∞ on the left sphere in Figure 1.3). If we instead intersect the hyperboloid with the sphere centered at the origin having radius 1.9, then we obtain the smooth quartic space curve drawn on the right sphere in Figure 1.3.

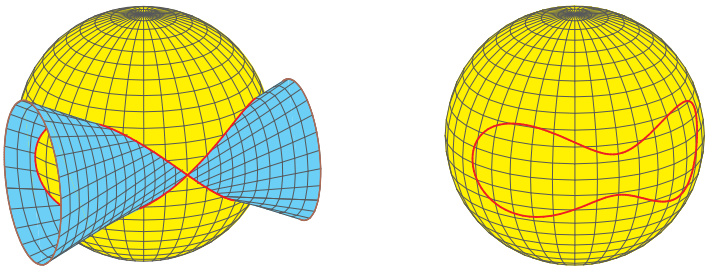


Figure 1.3: Quartics on spheres.

The product $X \times Y$ of two varieties X and Y is again a variety. Indeed, suppose that $X \subset \mathbb{K}^n$ is defined by the polynomials $f_1, \ldots, f_s \in \mathbb{K}[x_1, \ldots, x_n]$ and that $Y \subset \mathbb{K}^m$ is defined by the polynomials $g_1, \ldots, g_t \in \mathbb{K}[y_1, \ldots, y_m]$. Then $X \times Y \subset \mathbb{K}^n \times \mathbb{K}^m = \mathbb{K}^{n+m}$ is defined by the polynomials $f_1, \ldots, f_s, g_1, \ldots, g_t \in \mathbb{K}[x_1, \ldots, x_n, y_1, \ldots, y_m]$. Given a point $x \in X$, the product $\{x\} \times Y$ is a subvariety of $X \times Y$ which may be identified with Y simply by forgetting the coordinate x.

The set $\mathrm{Mat}_{n\times n}$ or $\mathrm{Mat}_{n\times n}(\mathbb{K})$ of $n \times n$ matrices with entries in $\mathbb{K}$ is identified with the affine space $\mathbb{K}^{n^2}$, which may be written $\mathbb{K}^{n\times n}$. An interesting class of varieties are linear algebraic groups, which are algebraic subvarieties of $\mathrm{Mat}_{n\times n}$ that are closed under multiplication and taking inverses. The *special linear group* is the set of matrices with determinant 1,

$$SL_n \;:=\; \{M \in \mathrm{Mat}_{n\times n} \mid \det M = 1\}\,,$$

which is a linear algebraic group. Since the determinant of a matrix in $\mathrm{Mat}_{n\times n}$ is a polynomial in its entries, SL_n is the variety $\mathcal{V}(\det -1)$. We will later show that SL_n is smooth, irreducible, and has dimension $n^2 - 1$. (We must first, of course, define these notions.)

There is a general construction of other linear algebraic groups. Let g^T be the transpose of a matrix $g \in \mathrm{Mat}_{n\times n}$. For a fixed matrix $M \in \mathrm{Mat}_{n\times n}$, set

$$G_M \;:=\; \{g \in SL_n \mid gMg^T = M\}\,.$$

This a linear algebraic group, as the condition $gMg^T = M$ is n^2 polynomial equations in the entries of g, and G_M is closed under matrix multiplication and matrix inversion.

When M is skew-symmetric and invertible, G_M is a *symplectic group*. In this case, n is necessarily even. If we let J_n denote the $n \times n$ matrix with ones on its anti-diagonal, then the matrix

$$\begin{bmatrix} 0 & J_n \\ -J_n & 0 \end{bmatrix}$$

is conjugate to every other invertible skew-symmetric matrix in $\mathrm{Mat}_{2n\times 2n}$. We assume M is this matrix and write Sp_{2n} for the symplectic group.

When M is symmetric and invertible, G_M is a *special orthogonal group*. When $\mathbb{K}$ is algebraically closed, all invertible symmetric matrices are conjugate, and we may assume $M = J_n$. For general fields, there may be many different forms of the special orthogonal group. For instance, when $\mathbb{K} = \mathbb{R}$, let k and l be, respectively, the number of positive and negative eigenvalues of M (these are conjugation invariants of M). Then we obtain the group $SO_{k,l}\mathbb{R}$. We have $SO_{k,l}\mathbb{R} \simeq SO_{l,k}\mathbb{R}$.

Consider the two extreme cases. When $l = 0$, we may take $M = I_n$, and so we obtain the special orthogonal group $SO_{n,0} = SO_n(\mathbb{R})$ of rotation matrices in $\mathbb{R}^n$, which is compact in the usual topology. The other extreme case is when $|k - l| \leq 1$, and we may take $M = J_n$. This gives the split form of the special orthogonal group which is not compact.

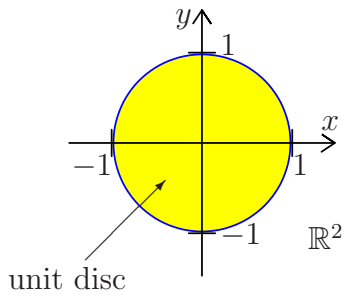
When $n = 2$, consider the two different real groups:

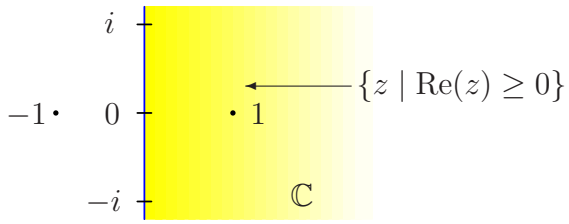
$$\begin{aligned} SO_{2,0}\mathbb{R} &:= \left\{ \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \mid \theta \in S^1 \right\} \\ SO_{1,1}\mathbb{R} &:= \left\{ \begin{bmatrix} a & 0 \\ 0 & a^{-1} \end{bmatrix} \mid a \in \mathbb{R}^\times \right\} \end{aligned}$$

Note that in the Euclidean topology $SO_{2,0}(\mathbb{R})$ is compact, while $SO_{1,1}(\mathbb{R})$ is not. The complex group $SO_2(\mathbb{C})$ is also not compact in the Euclidean topology.

We also point out some subsets of $\mathbb{K}^n$ which are not varieties. The set $\mathbb{Z}$ of integers is not a variety. The only polynomial vanishing at every integer is the zero polynomial, whose variety is all of $\mathbb{K}$. The same is true for any other infinite proper subset of $\mathbb{K}$, for example, the infinite sequence $\{1, \frac{1}{2}, \frac{1}{3}, \dots\}$ is not a subvariety of $\mathbb{K}$.

Other subsets which are not varieties (for the same reasons) include the unit disc in $\mathbb{R}^2$, $\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$ or the complex numbers with positive real part.





Sets like these last two which are defined by inequalities involving real polynomials are called *semi-algebraic*. We will study them in Chapter 4.

Exercises

1. Show that no proper nonempty open subset S of $\mathbb{R}^n$ or $\mathbb{C}^n$ is a variety. Here, we mean open in the usual (Euclidean) topology on $\mathbb{R}^n$ and $\mathbb{C}^n$. (Hint: Consider the Taylor expansion of any polynomial that vanishes identically on S.)

2. Suppose that $S \subset T$ are sets of polynomials in $\mathbb{K}[x_1, \ldots, x_n]$. Show that $\mathcal{V}(S) \supset \mathcal{V}(T)$.

3. Prove that in $\mathbb{K}^2$ we have $\mathcal{V}(y-x^2) = \mathcal{V}(y^3-y^2x^2,\ x^2y-x^4)$.

4. Express the cubic space curve C with parametrization (t, t^2, t^3) in each of the following ways.

   (a) The intersection of a quadric hypersurface and a cubic hypersurface.

   (b) The intersection of two quadrics.

   (c) The intersection of three quadrics.

5. Let $\mathbb{K}^{n\times n}$ be the set of $n \times n$ matrices over $\mathbb{K}$.

   (a) Show that the set $SL(n, \mathbb{K}) \subset \mathbb{K}^{n\times n}$ of matrices with determinant 1 is an algebraic variety.

   (b) Show that the set of singular matrices in $\mathbb{K}^{n\times n}$ is an algebraic variety.

   (c) Show that the set $GL(n, \mathbb{K})$ of invertible matrices is not an algebraic variety in $\mathbb{K}^{n\times n}$. Show that $GL_n(\mathbb{K})$ can be identified with an algebraic subset of $\mathbb{K}^{n^2+1} = \mathbb{K}^{n\times n} \times \mathbb{K}^1$ via a map $GL_n(\mathbb{K}) \to \mathbb{K}^{n^2+1}$.

6. An $n\times n$ matrix with complex entries is *unitary* if its columns are orthonormal under the complex inner product $\langle z, w\rangle = z \cdot \overline{w}^t = \sum_{i=1}^n z_i\overline{w_i}$. Show that the set $\mathbf{U}(n)$ of unitary matrices is not a complex algebraic variety. Show that it can be described as the zero locus of a collection of polynomials with real coefficients in $\mathbb{R}^{2n^2}$, and so it is a real algebraic variety.

7. Let $\mathbb{K}^{m\times n}$ be the set of $m \times n$ matrices over $\mathbb{K}$.

   (a) Show that the set of matrices of rank $\leq r$ is an algebraic variety.

   (b) Show that the set of matrices of rank $= r$ is not an algebraic variety if $r > 0$.

8. (a) Show that the set $\{(t, t^2, t^3) \mid t \in \mathbb{K}\}$ is an algebraic variety in $\mathbb{K}^3$.

   (b) Show that the following sets are not algebraic varieties

   (i) $\{(x, y) \in \mathbb{R}^2 | y = \sin x\}$

   (ii) $\{(\cos t, \sin t, t) \in \mathbb{R}^3 \mid t \in \mathbb{R}\}$

   (iii) $\{(x, e^x) \in \mathbb{R}^2 \mid x \in \mathbb{R}\}$

1.2 The algebraic-geometric dictionary

The strength and richness of algebraic geometry as a subject and source of tools for applications comes from its dual, simultaneously algebraic and geometric, nature. Intuitive geometric concepts are tamed via the precision of algebra while basic algebraic notions are enlivened by their geometric counterparts. The source of this dual nature is a correspondence between algebraic concepts and geometric concepts that we refer to as the algebraic-geometric dictionary.

We defined varieties $\mathcal{V}(S)$ associated to sets $S \subset \mathbb{K}[x_1, \ldots, x_n]$ of polynomials,

$$\mathcal{V}(S) \;=\; \{x \in \mathbb{K}^n \mid f(x) = 0 \text{ for all } f \in S\}\,.$$

We would like to invert this association. Given a subset Z of $\mathbb{K}^n$, consider the collection of polynomials that vanish on Z,

$$\mathcal{I}(Z) \;:=\; \{f \in \mathbb{K}[x_1, \ldots, x_n] \mid f(z) = 0 \text{ for all } z \in Z\}\,.$$

The map $\mathcal{I}$ reverses inclusions so that $Z \subset Y$ implies $\mathcal{I}(Z) \supset \mathcal{I}(Y)$.

These two inclusion-reversing maps

$$\{\text{Subsets } S \text{ of } \mathbb{K}[x_1, \ldots, x_n]\} \quad \underset{\mathcal{I}}{\overset{\mathcal{V}}{\rightleftarrows}} \quad \{\text{Subsets } Z \text{ of } \mathbb{K}^n\} \tag{1.1}$$

form the basis of the algebra-geometry dictionary of affine algebraic geometry. We will refine this correspondence to make it more precise.

An *ideal* is a subset $I \subset \mathbb{K}[x_1, \ldots, x_n]$ which is closed under addition and under multiplication by polynomials in $\mathbb{K}[x_1, \ldots, x_n]$. If $f, g \in I$ then $f + g \in I$ and if we also have $h \in \mathbb{K}[x_1, \ldots, x_n]$, then $hf \in I$. The *ideal* $\langle S \rangle$ generated by a subset S of $\mathbb{K}[x_1, \ldots, x_n]$ is the smallest ideal containing S. It is the set of all expressions of the form

$$h_1 f_1 + \cdots + h_m f_m$$

where $f_1, \ldots, f_m \in S$ and $h_1, \ldots, h_m \in \mathbb{K}[x_1, \ldots, x_n]$. We work with ideals because if f, g, and h are polynomials and $x \in \mathbb{K}^n$ with $f(x) = g(x) = 0$, then $(f + g)(x) = 0$ and $(hf)(x) = 0$. Thus $\mathcal{V}(S) = \mathcal{V}(\langle S \rangle)$, and so we may restrict $\mathcal{V}$ to the ideals of $\mathbb{K}[x_1, \ldots, x_n]$. In fact, we lose nothing if we restrict the left-hand-side of the correspondence (1.1) to the ideals of $\mathbb{K}[x_1, \ldots, x_n]$.

Lemma 1.2.1. *For any subset S of $\mathbb{K}^n$, $\mathcal{I}(S)$ is an ideal of $\mathbb{K}[x_1, \ldots, x_n]$.*

Proof. Let $f, g \in \mathcal{I}(S)$ be two polynomials which vanish at all points of S. Then $f + g$ vanishes on S, as does hf, where h is any polynomial in $\mathbb{K}[x_1, \ldots, x_n]$. This shows that $\mathcal{I}(S)$ is an ideal of $\mathbb{K}[x_1, \ldots, x_n]$. □

When S is infinite, the variety $\mathcal{V}(S)$ is defined by infinitely many polynomials. Hilbert's Basis Theorem tells us that only finitely many of these polynomials are needed.

Hilbert's Basis Theorem. *Every ideal I of $\mathbb{K}[x_1, \ldots, x_n]$ is finitely generated.*

We will prove this in Chapter 2. Be more specific!

Hilbert's Basis Theorem implies many important finiteness properties of algebraic varieties.

Corollary 1.2.2. *Any variety $Z \subset \mathbb{K}^n$ is the intersection of finitely many hypersurfaces.*

Proof. Let $Z = \mathcal{V}(I)$ be defined by the ideal I. By Hilbert's Basis Theorem, I is finitely generated, say by $f_1, \ldots, f_s$, and so $Z = \mathcal{V}(f_1, \ldots, f_s) = \mathcal{V}(f_1) \cap \cdots \cap \mathcal{V}(f_s)$. □

Example 1.2.3. The ideal of the cubic space curve C of Figure 1.2 with parametrization $(t^2, -t, t^3)$ not only contains the polynomials $xy+z$ and x^2-x+y^2+yz, but also y^2-x, x^2+yz, and y^3+z. Not all of these polynomials are needed to define C as $x^2-x+y^2+yz = (y^2-x)+(x^2+yz)$ and $y^3+z = y(y^2-x)+(xy+z)$. In fact three of the quadrics suffice,

$$\mathcal{I}(C) \;=\; \langle xy+z,\; y^2-x,\; x^2+yz \rangle\,.$$

Lemma 1.2.4. *For any subset Z of $\mathbb{K}^n$, if $X = \mathcal{V}(\mathcal{I}(Z))$ is the variety defined by the ideal $\mathcal{I}(Z)$, then $\mathcal{I}(X) = \mathcal{I}(Z)$ and X is the smallest variety containing Z.*

Proof. Set $X := \mathcal{V}(\mathcal{I}(Z))$. Then $\mathcal{I}(Z) \subset \mathcal{I}(X)$, since if f vanishes on Z, it will vanish on X. However, $Z \subset X$, and so $\mathcal{I}(Z) \supset \mathcal{I}(X)$, and thus $\mathcal{I}(Z) = \mathcal{I}(X)$.

If Y was a variety with $Z \subset Y \subset X$, then $\mathcal{I}(X) \subset \mathcal{I}(Y) \subset \mathcal{I}(Z) = \mathcal{I}(X)$, and so $\mathcal{I}(Y) = \mathcal{I}(X)$. But then we must have $Y = X$ for otherwise $\mathcal{I}(X) \subsetneq \mathcal{I}(Y)$, as is shown in Exercise 3. □

Thus we also lose nothing if we restrict the right-hand-side of the correspondence (1.1) to the subvarieties of $\mathbb{K}^n$. Our correspondence now becomes

$$\{\text{Ideals } I \text{ of } \mathbb{K}[x_1, \ldots, x_n]\} \quad \underset{\mathcal{I}}{\overset{\mathcal{V}}{\rightleftarrows}} \quad \{\text{Subvarieties } X \text{ of } \mathbb{K}^n\}\,. \tag{1.2}$$

This association is not a bijection. In particular, the map $\mathcal{V}$ is not one-to-one and the map $\mathcal{I}$ is not onto. There are several reasons for this.

For example, when $\mathbb{K} = \mathbb{Q}$ and $n = 1$, we have $\emptyset = \mathcal{V}(1) = \mathcal{V}(x^2-2)$. The problem here is that the rational numbers are not algebraically closed and we need to work with a larger field (for example $\mathbb{Q}(\sqrt{2})$) to study $\mathcal{V}(x^2-2)$. When $\mathbb{K} = \mathbb{R}$ and $n = 1$, $\emptyset \neq \mathcal{V}(x^2-2)$, but we have $\emptyset = \mathcal{V}(1) = \mathcal{V}(1+x^2) = \mathcal{V}(1+x^4)$. While the problem here is again that the real numbers are not algebraically closed, we view this as a manifestation of positivity. The two polynomials $1+x^2$ and $1+x^4$ only take positive values. When working over $\mathbb{R}$ (as our interest in applications leads us to do so) positivity of polynomials plays an important role, as we will see in later chapters.

The problem with the map $\mathcal{V}$ is more fundamental than these examples reveal and occurs even when $\mathbb{K} = \mathbb{C}$. When $n = 1$ we have $\{0\} = \mathcal{V}(x) = \mathcal{V}(x^2)$, and when $n = 2$, we invite the reader to check that $\mathcal{V}(y - x^2) = \mathcal{V}(y^2 - yx^2, xy - x^3)$. Note that while $x \notin \langle x^2 \rangle$, we have $x^2 \in \langle x^2 \rangle$. Similarly, $y - x^2 \notin \mathcal{V}(y^2 - yx^2, xy - x^3)$, but

$$(y - x^2)^2 \;=\; y^2 - yx^2 \;-\; x(xy - x^3) \;\in\; \langle y^2 - yx^2, xy - x^3 \rangle \,. \tag{1.3}$$

In both cases, the lack of injectivity of the map $\mathcal{V}$ is because f and f^N have the same set of zeroes, for any positive integer N. For example, if $f_1, \ldots, f_s$ are polynomials, then the two ideals

$$\langle f_1,\, f_2,\, \ldots,\, f_s \rangle \qquad \text{and} \qquad \langle f_1,\, f_2^2,\, f_3^3,\, \ldots,\, f_s^s \rangle$$

both define the same variety, and if $f^N \in \mathcal{I}(Z)$, then $f \in \mathcal{I}(Z)$.

We clarify this point with a definition. An ideal $I \subset \mathbb{K}[x_1, \ldots, x_n]$ is *radical* if whenever $f^N \in I$ for some positive integer N, then $f \in I$. The radical $\sqrt{I}$ of an ideal I of $\mathbb{K}[x_1, \ldots, x_n]$ is

$$\sqrt{I} \;:=\; \{f \in \mathbb{K}[x_1, \ldots, x_n] \mid f^N \in I \,, \text{ for some } N \geq 1\} \,.$$

You will show in Exercise 2 that $\sqrt{I}$ is the smallest radical ideal containing I. For example (1.3) shows that

$$\sqrt{\langle y^2 - yx^2, xy - x^3 \rangle} \;=\; \langle y - x^2 \rangle \,.$$

The reason for this definition is twofold: first, $\mathcal{I}(Z)$ is radical, and second, an ideal I and its radical$\sqrt{I}$ both define the same variety. We record these facts.

Lemma 1.2.5. *For $Z \subset \mathbb{K}^n$, $\mathcal{I}(Z)$ is a radical ideal. If $I \subset \mathbb{K}[x_1, \ldots, x_n]$ is an ideal, then $\mathcal{V}(I) = \mathcal{V}(\sqrt{I})$.*

When $\mathbb{K}$ is algebraically closed, the precise nature of the correspondence (1.2) follows from Hilbert's Nullstellensatz (null=zeroes, stelle=places, satz=theorem), another of Hilbert's foundational results in the 1890's that helped to lay the foundations of algebraic geometry and usher in twentieth century mathematics. We first state a weak form of the Nullstellensatz, which describes the ideals defining the empty set.

Theorem 1.2.6 (Weak Nullstellensatz)**.** *Suppose that $\mathbb{K}$ is algebraically closed. If I is an ideal of $\mathbb{K}[x_1, \ldots, x_n]$ with $\mathcal{V}(I) = \emptyset$, then $I = \mathbb{K}[x_1, \ldots, x_n]$.*

Let $b = (b_1, \ldots, b_n) \in \mathbb{K}^n$. Then $\{b\}$ is defined by the linear polynomials $x_i - b_i$ for $i = 1, \ldots, n$. A polynomial f is equal to the constant $f(b)$ modulo the ideal $\mathfrak{m}_b := \langle x_1 - b_1, \ldots, x_n - b_n \rangle$ generated by these polynomials, thus the quotient ring $\mathbb{K}[x_1, \ldots, x_n]/\mathfrak{m}_b$ is isomorphic to the field $\mathbb{K}$ and so $\mathfrak{m}_b$ is a maximal ideal. In fact, these are the only maximal ideals.

Theorem 1.2.7. *Every maximal* $\mathfrak{m}$ *ideal of* $\mathbb{K}[x_1, \ldots, x_n]$ *has the form* $\mathfrak{m}_b$ *for some* $b \in \mathbb{K}^n$.

Proof. We prove this when $\mathbb{K}$ is uncountable field, e.g. $\mathbb{K} = \mathbb{C}$. Then $\mathbb{K}[x_1, \ldots, x_n]/\mathfrak{m}$ is a field, L that contains $\mathbb{K}$ whose dimension as a $\mathbb{K}$-vector space is at most countable (it is spanned by the images of the monomials). Since $\mathbb{K}$ is algebraically closed, we have $L \neq \mathbb{K}$ only if L contains an element that is transcendental over $\mathbb{K}$. But then L contains a subfield isomorphic to the field $\mathbb{K}(t)$ of rational functions in t. Consider the uncountable subset of $\mathbb{K}(t)$,

$$\left\{ \frac{1}{t-a} \mid a \in \mathbb{K} \right\}.$$

We claim that this set is linearly independent. If we had a linear dependency,

$$0 \;=\; \sum_{i=1}^{m} \lambda_i \frac{1}{t-a_i},$$

then we could multiply it by $(t-a_i)$, simplify, and substitute $t = a_i$ to find that $\lambda_i = 0$, for every i. Thus $\mathbb{K}(t)$ has uncountable dimension over $\mathbb{K}$ and so L cannot contain a subfield isomorphic to $\mathbb{K}(t)$.

Thus we conclude that $L = \mathbb{K}$. If $b_i \in \mathbb{K}$ is the image of the variable x_i, then we see that $\mathfrak{m} \supset \mathfrak{m}_b$. As these are maximal ideals, they are in fact equal. □

Proof of the weak Nullstellensatz. We prove the contrapositive, if $I \subsetneq \mathbb{C}[x_1, \ldots, x_n]$ is a proper ideal, then $\mathcal{V}(I) \neq \emptyset$. There is a maximal ideal $\mathfrak{m}_b$ with $b \in \mathbb{K}^n$ of $\mathbb{C}[x_1, \ldots, x_n]$ which contains I. But then

$$\{b\} \;=\; \mathcal{V}(\mathfrak{m}_b) \;\subset\; \mathcal{V}(I),$$

and so $\mathcal{V}(I) \neq \emptyset$. Thus if $\mathcal{V}(I) = \emptyset$, we must have $I = \mathbb{C}[x_1, \ldots, x_n]$, which proves the weak Nullstellensatz. □

The Fundamental Theorem of Algebra states that any nonconstant polynomial $f \in \mathbb{C}[x]$ has a root (a solution to $f(x) = 0$). We recast the weak Nullstellensatz as the multivariate fundamental theorem of algebra.

Theorem 1.2.8 (Multivariate Fundamental Theorem of Algebra)**.** *If the polynomials* $f_1, \ldots, f_m \in \mathbb{C}[x_1, \ldots, x_n]$ *generate a proper ideal of* $\mathbb{C}[x_1, \ldots, x_n]$*, then the system of polynomial equations*

$$f_1(x) \;=\; f_2(x) \;=\; \cdots \;=\; f_m(x) \;=\; 0$$

has a solution in $\mathbb{K}^n$.

We now deduce the strong Nullstellensatz, which we will use to complete the characterization (1.2).

Theorem 1.2.9 (Nullstellensatz)**.** *If $I \subset \mathbb{C}[x_1, \ldots, x_n]$ is an ideal, then $\mathcal{I}(\mathcal{V}(I)) = \sqrt{I}$.*

Proof. Since $\mathcal{V}(I) = \mathcal{V}(\sqrt{I})$, we have $\sqrt{I} \subset \mathcal{I}(\mathcal{V}(I))$. We show the other inclusion. Suppose that we have a polynomial $f \in \mathcal{I}(\mathcal{V}(I))$. Introduce a new variable t. Then the variety $\mathcal{V}(I, tf-1) \subset \mathbb{K}^{n+1}$ defined by I and $tf-1$ is empty. Indeed, if $(a_1, \ldots, a_n, b)$ were a point of this variety, then $(a_1, \ldots, a_n)$ would be a point of $\mathcal{V}(I)$. But then $f(a_1, \ldots, a_n) = 0$, and so the polynomial $tf - 1$ evaluates to 1 (and not 0) at the point $(a_1, \ldots, a_n, b)$.

By the weak Nullstellensatz, $\langle I, tf-1\rangle = \mathbb{C}[x_1, \ldots, x_n, t]$. In particular, $1 \in \langle I, tf-1\rangle$, and so there exist polynomials $f_1, \ldots, f_m \in I$ and $g, g_1, \ldots, g_m \in \mathbb{C}[x_1, \ldots, x_n, t]$ such that

$$1 \;=\; f_1(x)g_1(x,t) + f_2(x)g_2(x,t) + \cdots + f_m(x)g_m(x,t) \;+\; g(x,t)(tf(x)-1)\,.$$

If we apply the substitution $t = \frac{1}{f}$, then the last term with factor $tf - 1$ vanishes and each polynomial $g_i(x,t)$ becomes a rational function in $x_1, \ldots, x_n$ whose denominator is a power of f. Clearing these denominators gives an expression of the form

$$f^N \;=\; f_1(x)G_1(x) + f_2(x)G_2(x) + \cdots + f_m(x)G_m(x)\,,$$

where $G_1, \ldots, G_m \in \mathbb{C}[x_1, \ldots, x_n]$. But this shows that $f \in \sqrt{I}$, and completes the proof of the Nullstellensatz. □

Corollary 1.2.10 (Algebraic-Geometric Dictionary I)**.** *Over any field $\mathbb{K}$, the maps $\mathcal{V}$ and $\mathcal{I}$ give an inclusion reversing correspondence*

$$\{\textit{Radical ideals } I \textit{ of } \mathbb{K}[x_1, \ldots, x_n]\} \quad \underset{\mathcal{I}}{\overset{\mathcal{V}}{\rightleftarrows}} \quad \{\textit{Subvarieties } X \textit{ of } \mathbb{K}^n\} \tag{1.4}$$

with $\mathcal{V}(\mathcal{I}(X)) = X$. When $\mathbb{K}$ is algebraically closed, the maps $\mathcal{V}$ and $\mathcal{I}$ are inverses, and this correspondence is a bijection.

Proof. First, we already observed that $\mathcal{I}$ and $\mathcal{V}$ are reverse inclusions and these maps have the domain and range indicated. Let X be a subvariety of $\mathbb{K}^n$. In Lemma 1.2.4 we showed that $X = \mathcal{V}(\mathcal{I}(X))$. Thus $\mathcal{V}$ is onto and $\mathcal{I}$ is one-to-one.

Now suppose that $\mathbb{K}$ is algebraically closed. By the Nullstellensatz, if I is radical then $\mathcal{I}(\mathcal{V}(I)) = I$, and so $\mathcal{I}$ is onto and $\mathcal{V}$ is one-to-one. This shows that $\mathcal{I}$ and $\mathcal{V}$ are inverse bijections. □

Corollary 1.2.10 is only the beginning of the algebraic-geometric dictionary. Many natural operations on varieties correspond to natural operations on their ideals. The *sum* $I + J$ and *product* $I \cdot J$ of ideals I and J are defined to be

$$\begin{aligned} I + J &:= \{f + g \mid f \in I \quad \text{and} \quad g \in J\} \\ I \cdot J &:= \langle f \cdot g \mid f \in I \quad \text{and} \quad g \in J\rangle\,. \end{aligned}$$

Lemma 1.2.11. *Let I, J be ideals in $\mathbb{K}[x_1, \ldots, x_n]$ and set $X := \mathcal{V}(I)$ and $Y := \mathcal{V}(J)$ to be their corresponding varieties. Then*

1. $\mathcal{V}(I+J) = X \cap Y$,
2. $\mathcal{V}(I \cdot J) = \mathcal{V}(I \cap J) = X \cup Y$,

If $\mathbb{K}$ *is algebraically closed, then we also have*

3. $\mathcal{I}(X \cap Y) = \sqrt{I+J}$, *and*
4. $\mathcal{I}(X \cup Y) = \sqrt{I \cap J} = \sqrt{I \cdot J}$.

Example 1.2.12. It can happen that $I \cdot J \neq I \cap J$. For example, if $I = \langle xy - x^3 \rangle$ and $J = \langle y^2 - x^2 y \rangle$, then $I \cdot J = \langle xy(y-x^2)^2 \rangle$, while $I \cap J = \langle xy(y-x^2) \rangle$.

This correspondence will be further refined in Section 1.3 to include maps between varieties. Because of this correspondence, each geometric concept has a corresponding algebraic concept, and *vice-versa*, when $\mathbb{K}$ is algebraically closed. When $\mathbb{K}$ is not algebraically closed, this correspondence is not exact. In that case we will often use algebra to guide our geometric definitions.

Exercises

1. Verify the claim in the text that the smallest ideal containing a set $S \subset \mathbb{K}[x_1, \ldots, x_n]$ of polynomials consists of all expressions of the form

$$h_1 f_1 + \cdots + h_m f_m$$

where $f_1, \ldots, f_m \in S$ and $h_1, \ldots, h_m \in \mathbb{K}[x_1, \ldots, x_n]$.

2. Let I be an ideal of $\mathbb{K}[x_1, \ldots, x_n]$. Show that

$$\sqrt{I} \ := \ \{f \in \mathbb{K}[x_1, \ldots, x_n] \mid f^N \in I, \text{ for some } N \in \mathbb{N}\}$$

is an ideal, is radical, and is the smallest radical ideal containing I.

3. If $Y \subsetneq X$ are varieties, show that $\mathcal{I}(X) \subsetneq \mathcal{I}(Y)$.

4. Suppose that I and J are radical ideals. Show that $I \cap J$ is also a radical ideal.

5. Give radical ideals I and J for which $I + J$ is not radical.

6. Let I be an ideal in $\mathbb{K}[x_1, \ldots, x_n]$. Prove or find counterexamples to the following statements. Make your assumptions clear.

   (a) If $\mathcal{V}(I) = \mathbb{K}^n$ then $I = \langle 0 \rangle$.

   (b) If $\mathcal{V}(I) = \emptyset$ then $I = \mathbb{K}[x_1, \ldots, x_n]$.

7. Give two algebraic varieties Y and Z such that $\mathcal{I}(Y \cap Z) \neq \mathcal{I}(Y) + \mathcal{I}(Z)$.

8. (a) Let I be an ideal of $\mathbb{K}[x_1, \ldots, x_n]$. Show that if $\mathbb{K}[x_1, \ldots, x_n]/I$ is a finite dimensional $\mathbb{K}$-vector space then $\mathcal{V}(I)$ is a finite set.

   (b) Let $J = \langle xy, yz, xz \rangle$ be an ideal in $\mathbb{K}[x, y, z]$. Find the generators of $\mathcal{I}(\mathcal{V}(J))$. Show that J cannot be generated by two polynomials in $\mathbb{K}[x, y, z]$. Describe $V(I)$ where $I = \langle xy, xz - yz \rangle$. Show that $\sqrt{I} = J$.

9. Let $f, g \in \mathbb{K}[x, y]$ be polynomials without a common factor. Use Exercise 8(a) to show that $\mathcal{V}(f) \cap \mathcal{V}(g)$ is a finite set.

10. Prove that there are three points p, q, and r in $\mathbb{K}^2$ such that

$$\sqrt{\langle x^2 - 2xy^4 + y^6, y^3 - y \rangle} \;=\; I(\{p\}) \cap I(\{q\}) \cap I(\{r\})\,.$$

Show directly that the ideal $\langle x^2 - 2xy^4 + y^6, y^3 - y \rangle$ is not radical.

1.3 The algebrac-geometric dictionary II

We strengthen the algebra-geometry dictionary of Section 1.2 in two ways. We first replace affine space $\mathbb{K}^n$ by an affine variety X and the polynomial ring by the ring $\mathbb{K}[X]$ of regular functions on X and establish a correspondence between subvarieties of X and radical ideals of $\mathbb{K}[X]$. Next, we establish a correspondence between regular maps of varities and homomorphisms of their coordinate rings.

Let $X \subset \mathbb{K}^n$ be an affine variety and suppose that $\mathbb{K}$ is infinite. Any polynomial function $f \in \mathbb{K}[x_1, \ldots, x_n]$ restricts to give a *regular function* on X, $f\colon X \to \mathbb{K}$. We may add and multiply regular functions, and the set of all regular functions on X forms a ring, $\mathbb{K}[X]$, called the *coordinate ring* of the affine variety X or the ring of regular functions on X. The coordinate ring of an affine variety X is a basic invariant of X, which we will show is in fact equivalent to X itself.

The restriction of polynomial functions on $\mathbb{K}^n$ to regular functions on X defines a surjective ring homomorphism $\mathbb{K}[x_1, \ldots, x_n] \twoheadrightarrow \mathbb{K}[X]$. The kernel of this restriction homomorphism is the set of polynomials that vanish identically on X, that is, the ideal $\mathcal{I}(X)$ of X. Under the correspondence between ideals, quotient rings, and homomorphisms, this restriction map gives an isomorphism between $\mathbb{K}[X]$ and the quotient ring $\mathbb{K}[x_1, \ldots, x_n]/\mathcal{I}(X)$.

Example 1.3.1. The coordinate ring of the parabola $y = x^2$ is $\mathbb{K}[x, y]/\langle y - x^2\rangle$, which is isomorphic to $\mathbb{K}[x]$, the coordinate ring of $\mathbb{K}^1$. To see this, observe that substituting x^2 for y rewrites and polynomial $f(x, y)$ as a polynomial $g(x)$ in x alone, and $y - x^2$ divides the difference $f(x, y) - g(x)$.

Parabola Cuspidal Cubic

On the other hand, the coordinate ring of the cuspidal cubic $y^2 = x^3$ is $\mathbb{K}[x, y]/\langle y^2 - x^3\rangle$. This ring is not isomorphic to $\mathbb{K}[x, y]/\langle y - x^2\rangle$. Indeed, the element $y^2 = x^3$ has two distinct factorizations into indecomposable elements, while polynomials $f(x)$ in one variable always factor uniquely.

Let X be a variety. Its coordinate ring $\mathbb{K}[X] = \mathbb{K}[x_1, \ldots, x_n]/\mathcal{I}(X)$ is finitely generated by the images of the variables x_i. Since $\mathcal{I}(X)$ is radical, Exercise 4 implies that this quotient ring has no nilpotent elements (elements f such that $f^M = 0$ for some M). Such a ring with no nilpotents is called *reduced*. When $\mathbb{K}$ is algebraically closed, these two properties characterize coordinate rings of algebraic varieties.

Theorem 1.3.2. *Suppose that $\mathbb{K}$ is algebraically closed. Then a $\mathbb{K}$-algebra R is the coordinate ring of an affine variety if and only if R is finitely generated and reduced.*

Proof. We need only show that a finitely generated reduced $\mathbb{K}$-algebra R is the coordinate ring of some affine variety. Suppose that the reduced $\mathbb{K}$-algebra R has generators $r_1, \ldots, r_n$. Then there is a surjective ring homomorphism

$$\varphi \,:\, \mathbb{K}[x_1, \ldots, x_n] \;\longrightarrow\!\!\!\!\!\rightarrow\; R$$

given by $x_i \mapsto r_i$. Let $I \subset \mathbb{K}[x_1, \ldots, x_n]$ be the kernel of φ. This identifies R with $\mathbb{K}[x_1, \ldots, x_n]/I$. Since R is reduced, we see that I is radical.

As $\mathbb{K}$ is algebraically closed, the algebraic-geometric dictionary of Corollary 1.2.10 shows that $I = \mathcal{I}(\mathcal{V}(I))$ and so $R \simeq \mathbb{K}[x_1, \ldots, x_n]/I \simeq \mathbb{K}[\mathcal{V}(I)]$. □

A different choice $s_1, \ldots, s_m$ of generators for R in this proof will give a different affine variety with the same coordinate ring R. One goal of this section is to understand this apparent ambiguity.

Example 1.3.3. The finitely generated $\mathbb{K}$-algebra $R := \mathbb{K}[t]$ is the coordinate ring of the affine line $\mathbb{K}$. Note that if we set $x := t + 1$ and $y := t^2 + 3t$, these generate R. As $y = x^2 + x - 2$, this choice of generators realizes R as $\mathbb{K}[x, y]/\langle y - x^2 - x + 2\rangle$, which is the coordinate ring of a parabola.

Among the coordinate rings $\mathbb{K}[X]$ of affine varieties are the polynomial algebras $\mathbb{K}[x_1, \ldots, x_n]$. Many properties of polynomial algebras, including the algebraic-geometric dictionary of Corollary 1.2.10 and the Hilbert Theorems hold for these coordinate rings $\mathbb{K}[X]$.

Given regular functions $f_1, \ldots, f_m \in \mathbb{K}[X]$ on an affine variety $X \subset \mathbb{K}^n$, their set of common zeroes

$$\mathcal{V}(f_1, \ldots, f_m) \;:=\; \{x \in X \mid f_1(x) = \cdots = f_m(x) = 0\}\,,$$

is a subvariety of X. To see this, let $F_1, \ldots, F_m \in \mathbb{K}[x_1, \ldots, x_n]$ be polynomials which restrict to the functions $f_1, \ldots, f_m$ on X. Then

$$\mathcal{V}(f_1, \ldots, f_m) \;=\; X \,\cap\, \mathcal{V}(F_1, \ldots, F_m)\,,$$

and we recall that intersecrtions of varieties are again varieties. As in Section 1.2, we may extend this notation and define $\mathcal{V}(I)$ for an ideal I of $\mathbb{K}[X]$. If $Y \subset X$ is a subvariety of X, then $\mathcal{I}(X) \subset \mathcal{I}(Y)$ and so $\mathcal{I}(Y)/\mathcal{I}(X)$ is an ideal in the coordinate ring $\mathbb{K}[X] = \mathbb{K}[\mathbb{K}^n]/\mathcal{I}(X)$ of X. Write $\mathcal{I}(Y) \subset \mathbb{K}[X]$ for the ideal of Y in $\mathbb{K}[X]$.

Both Hilbert's Basis Theorem and Hilbert's Nullstellensätze have analogs for affine varieties X and their coordinate rings $\mathbb{K}[X]$. These consequences of the original Hilbert Theorems follow from the surjection $\mathbb{K}[x_1, \ldots, x_n] \twoheadrightarrow \mathbb{K}[X]$ and corresponding inclusion $X \hookrightarrow \mathbb{K}^n$.

Theorem 1.3.4 (Hilbert Theorems for $\mathbb{K}[X]$)**.** *Let X be an affine variety. Then*

1. *Any ideal of $\mathbb{K}[X]$ is finitely generated.*

2. *If Y is a subvariety of X then $\mathcal{I}(Y) \subset \mathbb{K}[X]$ is a radical ideal.*

3. *Suppose that $\mathbb{K}$ is algebraically closed. An ideal I of $\mathbb{K}[X]$ defines the empty set if and only if $I = \mathbb{K}[X]$.*

As in Section 1.2 we obtain a version of the algebraic-geometric dictionary between subvarieties of an affine variety X and radical ideals of $\mathbb{K}[X]$. The proofs are nearly the same, so we leave them to the reader. For this, you will need to recall that ideals of a quotient ring R/I all have the form J/I, where J is an ideal of R which contains I.

Theorem 1.3.5. *Let X be an affine variety. Then the maps $\mathcal{V}$ and $\mathcal{I}$ give an inclusion reversing correspondence*

$$\{\text{Radical ideals } I \text{ of } \mathbb{K}[X]\} \quad \underset{\mathcal{I}}{\overset{\mathcal{V}}{\rightleftarrows}} \quad \{\text{Subvarieties } Y \text{ of } X\} \tag{1.5}$$

with $\mathcal{I}$ injective and $\mathcal{V}$ surjective. When $\mathbb{K}$ is algebraically closed, the maps $\mathcal{V}$ and $\mathcal{I}$ are inverse bijections.

We do not just study varieties, but also the maps between them.

Definition 1.3.6. A list $f_1, \ldots, f_m \in \mathbb{K}[X]$ of regular functions on an affine variety X defines a function

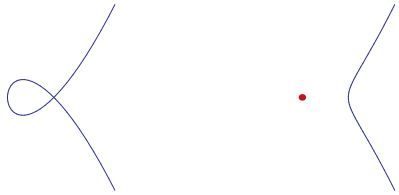
$$\begin{array}{rccl} \varphi : & X & \longrightarrow & \mathbb{K}^m \\ & x & \longmapsto & (f_1(x), f_2(x), \ldots, f_m(x)) \,, \end{array}$$

which we call a *regular map*.

Example 1.3.7. The elements $t^2, t, -t^3 \in \mathbb{K}[t]$ define the map $\mathbb{K}^1 \to \mathbb{K}^3$ whose image is the cubic curve of Figure 1.2.

The elements t^2, t^3 of $\mathbb{K}[t]$ define a map $\mathbb{K}^1 \to \mathbb{K}^2$ whose image is the cuspidal cubic that we saw earlier.

Let $x = t^2 - 1$ and $y = t^3 - t$, which are elements of $\mathbb{K}[t]$. These define a map $\mathbb{K}^1 \to \mathbb{K}^2$ whose image is the nodal cubic curve $\mathcal{V}(y^2 - (x^3 + x^2))$ on the left below. If we instead take $x = t^2 + 1$ and $y = t^3 + t$, then we get a different map $\mathbb{K}^1 \to \mathbb{K}^2$ whose image is the curve $\mathcal{V}(y^2 - (x^3 - x^2))$ on the right below.



In the curve on the right, the image of $\mathbb{R}^1$ is the arc, while the isolated or *solitary point* is the image of the points $\pm\sqrt{-1}$.

Suppose that X is an affine variety and we have a regular map $\varphi\colon X \to \mathbb{K}^m$ given by regular functions $f_1, \ldots, f_m \in \mathbb{K}[X]$. A polynomial $g \in \mathbb{K}[x_1, \ldots, x_m]$ *pulls back* along φ to give the regular function $\varphi^* g$, which is defined by

$$\varphi^* g \;:=\; g(f_1, \ldots, f_m)\,.$$

This element of the coordinate ring $\mathbb{K}[X]$ of X is the usual pull back of a function. For $x \in X$ we have

$$(\varphi^* g)(x) \;=\; g(\varphi(x)) \;=\; g(f_1(x), \ldots, f_m(x))\,.$$

The resulting map $\varphi^*\colon \mathbb{K}[x_1, \ldots, x_m] \to \mathbb{K}[X]$ is a homomorphism of $\mathbb{K}$-algebras. Conversely, given a homomorphism $\psi\colon \mathbb{K}[x_1, \ldots, x_m] \to \mathbb{K}[X]$ of $\mathbb{K}$-algebras, if we set $f_i := \psi(x_i)$, then $f_1, \ldots, f_m \in \mathbb{K}[X]$ define a regular map φ with $\varphi^* = \psi$.

We have just shown the following basic fact.

Lemma 1.3.8. *The association $\varphi \mapsto \varphi^*$ defines a bijection*

$$\left\{ \begin{array}{c} \textit{Regular maps} \\ \varphi\colon X \to \mathbb{K}^m \end{array} \right\} \quad \longleftrightarrow \quad \left\{ \begin{array}{c} \mathbb{K}\textit{-algebra homomorphisms} \\ \psi\colon \mathbb{K}[x_1, \ldots, x_m] \to \mathbb{K}[X] \end{array} \right\}$$

In the examples that we gave, the image $\varphi(X)$ of X under φ was contained in a subvariety. This is always the case.

Lemma 1.3.9. *Let X be an affine variety, $\varphi\colon X \to \mathbb{K}^m$ a regular map, and $Y \subset \mathbb{K}^m$ a subvariety. Then $\varphi(X) \subset Y$ if and only if $\mathcal{I}(Y) \subset \ker \varphi^*$.*

In particular, $\mathcal{V}(\ker \varphi^*)$ is the smallest subvariety of $\mathbb{K}^m$ that contains the image $\varphi(X)$ of X under φ.

Proof. First suppose that $\varphi(X) \subset Y$. If $f \in \mathcal{I}(Y)$ then f vanishes on Y and hence on $\varphi(X)$. But then $\varphi^* f$ is the zero function, and so $\mathcal{I}(Y) \subset \ker \varphi^*$.

For the other direction, suppose that $\mathcal{I}(Y) \subset \ker \varphi^*$ and let $x \in X$. If $f \in \mathcal{I}(Y)$, then $\varphi^* f = 0$ and so $0 = \varphi^* f(x) = f(\varphi(x))$. This implies that $\varphi(x) \in Y$, and so we conclude that $\varphi(X) \subset Y$. □

Definition 1.3.10. Affine varieties X and Y are *isomorphic* if there are regular maps $\varphi\colon X \to Y$ and $\psi\colon Y \to X$ such that both $\varphi \circ \psi$ and $\psi \circ \varphi$ are the identity maps on Y and X, repsectively. In this case, we say that φ and ψ are isomorphisms.

Corollary 1.3.11. *Let X be an affine variety, $\varphi\colon X \to \mathbb{K}^m$ a regular map, and $Y \subset \mathbb{K}^m$ a subvariety. Then*

(1) $\ker \varphi^*$ *is a radical ideal.*

(2) $\mathcal{V}(\ker \varphi^*)$ *is the smallest affine variety containing $\varphi(X)$.*

(3) *If* $\varphi\colon X \to Y$*, then* $\varphi^*\colon \mathbb{K}[\mathbb{K}^m] \to \mathbb{K}[X]$ *factors through* $\mathbb{K}[Y]$ *inducing a homomorphism* $\mathbb{K}[Y] \to \mathbb{K}[X]$.

(4) φ *is an isomorphism of varieties if and only if* φ^* *is an isomorphism of* $\mathbb{K}$*-algebras.*

We write φ^* for the induced map $\mathbb{K}[Y] \to \mathbb{K}[X]$ of part (4).

Proof. For (1), suppose that $f^N \in \ker \varphi^*$, so that $0 = \varphi^*(f^N) = (\varphi^*(f))^N$. Since $\mathbb{K}[X]$ has no nilpotent elements, we conclude that $\varphi^*(f) = 0$ and so $f \in \ker \varphi^*$.

Suppose that Y is an affine variety containing $\varphi(X)$. By Lemma 1.3.9, $\mathcal{I}(Y) \subset \ker \varphi^*$ and so $\mathcal{V}(\ker \varphi^*) \subset Y$. Statement (2) follows as we also have $X \subset \mathcal{V}(\ker \varphi^*)$.

For (3), we have $\mathcal{I}(Y) \subset \ker \varphi^*$ and so the map $\varphi^*\colon \mathbb{K}[x_1, \ldots, x_m] \to \mathbb{K}[X]$ factors through the quotient map $\mathbb{K}[x_1, \ldots, x_m] \twoheadrightarrow \mathbb{K}[x_1, \ldots, x_m]/\mathcal{I}(Y) = \mathbb{K}[Y]$.

Statement (4) is immediate from the definitions. □

Thus we may refine the correspondence of Lemma 1.3.8. Let X and Y be affine varieties. Then the association $\varphi \mapsto \varphi^*$ gives a bijective correspondence

$$\left\{ \begin{array}{c} \text{Regular} \\ \text{maps} \\ \varphi\colon X \to Y \end{array} \right\} \quad \longleftrightarrow \quad \left\{ \begin{array}{c} \mathbb{K}\text{-algebra homomorphisms} \\ \psi\colon \mathbb{K}[Y] \to \mathbb{K}[X] \end{array} \right\}.$$

This map $X \mapsto \mathbb{K}[X]$ from affine varieties to finitely generated reduced $\mathbb{K}$-algebras not only sends objects to objects, but it induces an isomorphism on maps between objects (reversing their direction however). In mathematics, such an association is called a *contravariant equivalence of categories*. The point of this equivalence is that an affine variety and its coordinate ring are different packages for the same information. Each one determines and is determined by the other. Whether we study algebra or geometry, we are studying the same thing.

The prototypical example of a contravariant equivalence of categories comes from linear algebra. To a finite-dimensional vector space V, we may associate its dual space V^*. Given a linear transformation $L\colon V \to W$, its adjoint is a map $L^*\colon W^* \to V^*$. Since $(V^*)^* = V$ and $(L^*)^* = L$, this association is a bijection on the objects (finite-dimensional vector spaces) and a bijection on linear maps linear maps from V to W.

Exercises

1. Give a proof of Theorem 1.3.4.

2. Let $V = \mathcal{V}(y - x^2) \subset \mathbb{K}^2$ and $W = \mathcal{V}(xy - 1) \subset \mathbb{K}^2$. Show that

$$\begin{aligned} \mathbb{K}[V] &:= \mathbb{K}[x, y]/\mathcal{I}(V) \cong \mathbb{K}[t] \\ \mathbb{K}[W] &:= \mathbb{K}[x, y]/\mathcal{I}(W) \cong \mathbb{K}[t, t^{-1}] \end{aligned}$$

Conclude that the hyperbola $V(xy - 1)$ is not isomorphic to the affine line.

3. Suppose that $\mathbb{K}$ is an infinite field. Show that $f \in \mathbb{K}[x_1, \ldots, x_n]$ defines the zero function $f\colon \mathbb{K}^n \to \mathbb{K}$ if and only if f is the zero polynomial. (Hint: One direction is easy, and for the other, consider first the case when $n = 1$ and then use induction.)

4. Let $I \subset \mathbb{K}[x_1, \ldots, x_n]$ be an ideal. Show that the factor ring $\mathbb{K}[x_1, \ldots, x_n]/I$ has nilpotent elements if and only if I is not a radical ideal.

1.4 Projective varieties

Projective space and projective varieties are undoubtedly the most important objects in algebraic geometry. We motivate projective space with an example.

Consider the intersection of the parabola $y = x^2$ in the affine plane $\mathbb{K}^2$ with a line, $\ell := \mathcal{V}(ay + bx + c)$. Solving these implied equations gives

$$ax^2 + bx + c \;=\; 0 \qquad \text{and} \qquad y \;=\; x^2\,.$$

There are several cases to consider.

(i) $a \neq 0$ and $b^2 - 4ac > 0$. Then ℓ meets the parabola in two distinct real points.

(i′) $a \neq 0$ and $b^2 - 4ac < 0$. While ℓ does not appear to meet the parabola, that is because we have drawn the real picture, and ℓ meets it in two complex conjugate points.

When $\mathbb{K}$ is algebraically closed, then cases (i) and (i′) coalesce to the case of $a \neq 0$ and $b^2-4ac \neq 0$. These two points of intersection are predicted by Bézout's Theorem in the plane (Theorem 2.3.15).

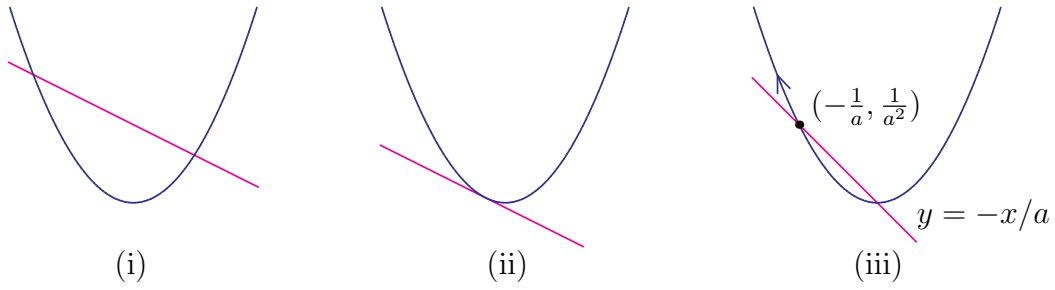
(ii) $a \neq 0$ but $b^2-4ac = 0$. Then ℓ is tangent to the parabola and we solve the equations to get

$$a(x - \tfrac{b}{2a})^2 \;=\; 0 \qquad \text{and} \qquad y \;=\; x^2\,.$$

Thus there is one solution, $(\frac{b}{2a}, \frac{b^2}{4a^2})$. As $x = \frac{b}{2a}$ is a root of multiplicity 2 in the first equation, it is reasonable to say that this one solution to our geometric problem occurs with multiplicity 2.

(iii) $a = 0$. There is a single, unique solution, $x = -c/b$ and $y = c^2/b^2$.

Suppose now that $c = 0$ and let $b = 1$. For $a \neq 0$, there are two solutions $(0, 0)$ and $(-\frac{1}{a}, \frac{1}{a^2})$. In the limit as $a \to 0$, the second solution disappears off to infinity.



One purpose of projective space is to prevent this last phenomenon from occurring.

Definition 1.4.1. The set of all 1-dimensional linear subspaces of $\mathbb{K}^{n+1}$ is called *n-dimensional projective space* and written $\mathbb{P}^n$ or $\mathbb{P}^n_{\mathbb{K}}$. If V is a finite-dimensional vector space, then $\mathbb{P}(V)$ is the set of all 1-dimensional linear subspaces of V. Note that $\mathbb{P}(V) \simeq \mathbb{P}^{\dim V - 1}$, but there are no preferred coordinates for $\mathbb{P}(V)$.

Example 1.4.2. The projective line $\mathbb{P}^1$ is the set of lines through the origin in $\mathbb{K}^2$. When $\mathbb{K} = \mathbb{R}$, we see that the line $x = ay$ through the origin intersects the circle $\mathcal{V}(x^2 + (y-1)^2 - 1)$ in the origin and in the point $(2a/(1+a^2), 2/(1+a^2))$, as shown in Figure 1.4. Identifying the x-axis with the origin and the lines $x = ay$ with this point of intersection gives a one-to-one map from $\mathbb{P}^1_{\mathbb{R}}$ to the circle, where the origin becomes the point at infinity.

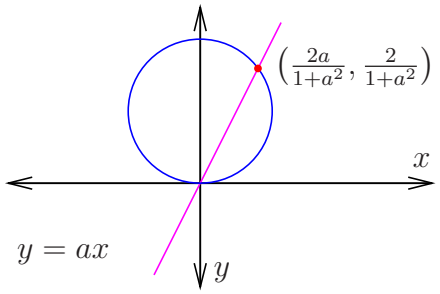


Figure 1.4: Lines through the origin meet the circle in a second point.

This definition of $\mathbb{P}^n$ leads to a system of global homogeneous coordinates for $\mathbb{P}^n$. We may represent a point, ℓ, of $\mathbb{P}^n$ by the coordinates $[a_0, a_1, \ldots, a_n]$ of any non-zero vector lying on the one-dimensional linear subspace $\ell \subset \mathbb{K}^{n+1}$. These coordinates are not unique. If $\lambda \neq 0$, then $[a_0, a_1, \ldots, a_n]$ and $[\lambda a_0, \lambda a_1, \ldots, \lambda a_n]$ both represent the same point. This non-uniqueness is the reason that we use rectangular brackets $[\ldots]$ in our notation for these *homogeneous coordinates*. Some authors prefer the notation $[a_0 : a_1 : \cdots : a_n]$.

Example 1.4.3. When $\mathbb{K} = \mathbb{R}$, note that a 1-dimensional subspace of $\mathbb{R}^{n+1}$ meets the sphere S^n in two antipodal points, v and $-v$. This identifies real projective space $\mathbb{P}^n_{\mathbb{R}}$ with the quotient $S^n/\{\pm 1\}$, showing that $\mathbb{P}^n_{\mathbb{R}}$ is a compact manifold in the usual topology.

Suppose that $\mathbb{K} = \mathbb{C}$. Given a point $a \in \mathbb{P}^n_{\mathbb{C}}$, after scaling, we may assume that $|a_0|^2 + |a_1|^2 + \cdots + |a_n|^2 = 1$. Identifying $\mathbb{C}$ with $\mathbb{R}^2$, this is the set of points a on the $2n+1$-sphere $S^{2n+1} \subset \mathbb{R}^{2n+2}$. If $[a_0, \ldots, a_n] = [b_0, \ldots, b_n]$ with $a, b \in S^{2n+1}$, then there is some $\zeta \in S^1$, the unit circle in $\mathbb{C}$, such that $a_i = \zeta b_i$. This identifies $\mathbb{P}^n_{\mathbb{C}}$ with the quotient of S^{2n+1}/S^1, showing that $\mathbb{P}^n_{\mathbb{C}}$ is a compact manifold. Since $\mathbb{P}^n_{\mathbb{R}} \subset \mathbb{P}^n_{\mathbb{C}}$, we again see that $\mathbb{P}^n_{\mathbb{R}}$ is compact.

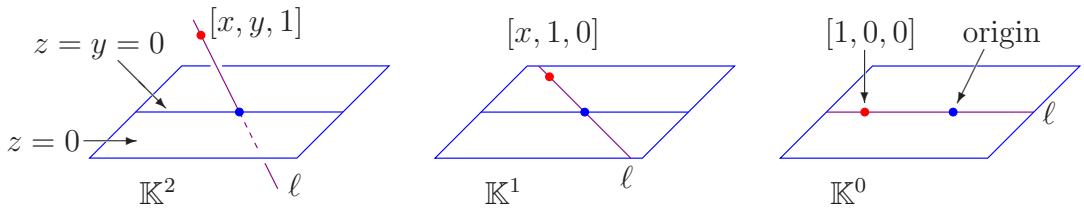
Homogeneous coordinates of a point are not unique. Uniqueness may be restored, but at the price of non-uniformity. Let $A_i \subset \mathbb{P}^n$ be the set of points $[a_0, a_1, \ldots, a_n]$ in projective space $\mathbb{P}^n$ with $a_i \neq 0$, but $a_{i+1} = \cdots = a_n = 0$. Given a point $a \in A_i$, we may

divide by its ith coordinate to get a representative of the form $[a_0, \ldots, a_{i-1}, 1, 0, \ldots, 0]$. These i numbers $(a_0, \ldots, a_{i-1})$ provide coordinates for A_i, identifying it with the affine space $\mathbb{K}^i$. This decomposes projective space $\mathbb{P}^n$ into a disjoint union of $n+1$ affine spaces

$$\mathbb{P}^n \;=\; \mathbb{K}^n \sqcup \cdots \sqcup \mathbb{K}^1 \sqcup \mathbb{K}^0 \,.$$

When a variety admits a decomposition as a disjoint union of affine spaces, we say that it is *paved by affine spaces*. Many important varieties admit such a decomposition.

It is instructive to look at this closely for $\mathbb{P}^2$. Below, we show the possible positions of a one-dimensional linear subspace $\ell \subset \mathbb{K}^3$ with respect to the x, y-plane $z = 0$, the x-axis $z = y = 0$, and the origin in $\mathbb{K}^3$.



There is also a scheme for local coordinates on projective space.

1. For $i = 0, \ldots, n$, let U_i be the set of points $a \in \mathbb{P}^n$ in projective space whose ith coordinate is non-zero. Dividing by this ith coordinate, we obtain a representative of the point having the form

   $$[a_0, \ldots, a_{i-1}, 1, a_{i+1}, \ldots, a_n] \,.$$

   The n coordinates $(a_0, \ldots, a_{i-1}, \; a_{i+1}, \ldots, a_n)$ determine this point, identifying U_i with affine n-space, $\mathbb{K}^n$. Every point of $\mathbb{P}^n$ lies in some U_i,

   $$\mathbb{P}^n \;=\; U_0 \cup U_1 \cup \cdots \cup U_n \,.$$

   When $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$, these U_i are coordinate charts for $\mathbb{P}^n$ as a manifold.

   For any field $\mathbb{K}$, these affine sets U_i provide coordinate charts for $\mathbb{P}^n$.

2. We give a coordinate-free description of these affine charts. Let $\Lambda\colon \mathbb{K}^{n+1} \to \mathbb{K}$ be a linear map, and let $H \subset \mathbb{K}^{n+1}$ be the set of points x where $\Lambda(x) = 1$. Then $H \simeq \mathbb{K}^n$, and the map

   $$H \ni x \;\longmapsto\; [x] \in \mathbb{P}^n$$

   identifies H with the complement $U_\Lambda = \mathbb{P}^n - \mathcal{V}(\Lambda)$ of the points where Λ vanishes.

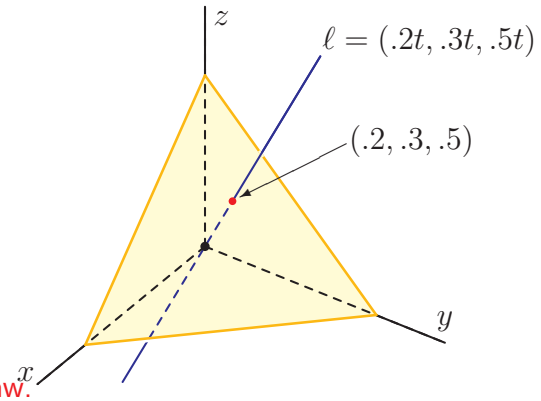
Example 1.4.4 (Probability simplex)**.** This more general description of affine charts leads to the beginning of an important application of algebraic geometry to statistics.

Here $\mathbb{K} = \mathbb{R}$, the real numbers and we set $\Lambda(x) := x_0 + \cdots + x_n$. If we consider those points x where $\Lambda(x) = 1$ which have positive coordinates, we obtain the *probability simplex*

$$\Delta \;:=\; \{(p_0, p_1, \ldots, p_n) \in \mathbb{R}^{n+1}_+ \mid p_0 + p_1 + \cdots + p_n = 1\}\,,$$

where $\mathbb{R}^{n+1}_+$ is the *positive orthant*, the points of $\mathbb{R}^{n+1}$ with nonnegative coordinates. Here p_i represents the probability of an event i occurring, and the condition $p_0 + \cdots + p_n = 1$ reflects that every event does occur.

Here is a picture when $n = 2$.



Rotate the picture and redraw.

We wish to extend the definitions and structures of affine algebraic varieties to projective space. One problem arises immediately: given a polynomial $f \in \mathbb{K}[x_0, \ldots, x_n]$ and a point $a \in \mathbb{P}^n$, we cannot in general define $f(a) \in \mathbb{K}$. To see why this is the case, for each non negative integer d, let f_d be the sum of the terms of f of degree d.[1] We call f_d the dth *homogeneous component* of f. If $[a_0, \ldots, a_n]$ and $[\lambda a_0, \ldots, \lambda a_n]$ are two representatives of $a \in \mathbb{P}^n$, and f has degree m, then

$$f(\lambda a_0, \ldots, \lambda a_n) \;=\; f_0(a_0, \ldots, a_n) + \lambda f_1(a_0, \ldots, a_n) + \cdots + \lambda^m f_m(a_0, \ldots, a_n)\,, \qquad (1.6)$$

since we can factor λ^d from every monomial $(\lambda x)^\alpha$ of degree d. Thus $f(a)$ is a well-defined number only if the polynomial (1.6) in λ is constant. That is, if and only if

$$f_i(a_0, \ldots, a_n) \;=\; 0 \qquad i = 1, \ldots, \deg(f)\,.$$

In particular, a polynomial f vanishes at a point $a \in \mathbb{P}^n$ if and only if every homogeneous component f_d of f vanishes at a. A polynomial f is *homogeneous* of degree d when $f = f_d$. We also use the term *homogeneous form* for a homogeneous polynomial.

Definition 1.4.5. Let $f_1, \ldots, f_m \in \mathbb{K}[x_0, \ldots, x_n]$ be homogeneous polynomials. These define a *projective variety*

$$\mathcal{V}(f_1, \ldots, f_m) \;:=\; \{a \in \mathbb{P}^n \mid f_i(a) = 0,\ i = 1, \ldots, m\}\,.$$

[1] Define degree!

An ideal $I \subset \mathbb{K}[x_0, \ldots, x_n]$ is *homogeneous* if whenever $f \in I$ then all homogeneous components of f lie in I. Thus projective varieties are defined by homogeneous ideals. Given a subset $Z \subset \mathbb{P}^n$ of projective space, its ideal is the collection of polynomials which vanish on Z,

$$\mathcal{I}(Z) \;:=\; \{f \in \mathbb{K}[x_0, x_1, \ldots, x_n] \mid f(z) = 0 \text{ for all } z \in Z\}\,.$$

In the exercises, you are asked to show that this ideal is homogeneous.

It is often convenient to work in an affine space when treating projective varieties. The *(affine) cone* $CZ \subset \mathbb{K}^{n+1}$ over a subset Z of projective space $\mathbb{P}^n$ is the union of the one-dimensional linear subspaces $\ell \subset \mathbb{K}^{n+1}$ corresponding to points of Z. Then the ideal $\mathcal{I}(X)$ of a projective variety X is equal to the ideal $\mathcal{I}(CX)$ of the affine cone over X.

Example 1.4.6. Let $\Lambda := a_0x_0 + a_1x_1 + \cdots + a_nx_n$ be a linear form. Then $\mathcal{V}(\Lambda)$ is a *hyperplane*. Let $V \subset \mathbb{K}^{n+1}$ be the kernel of Λ which is an n-dimensional linear subspace. It is also the affine variety defined by Λ. We have $\mathcal{V}(\Lambda) = \mathbb{P}(V)$.

The weak Nullstellensatz does not hold for projective space, as $\mathcal{V}(x_0, x_1, \ldots, x_n) = \emptyset$. We call this ideal, $\mathfrak{m}_0 := \langle x_0, x_1, \ldots, x_n\rangle$, the *irrelevant ideal*. It plays a special role in the projective algebraic-geometric dictionary.

Theorem 1.4.7 (Projective Algebraic-Geometric Dictionary)**.** *Over any field* $\mathbb{K}$*, the maps* $\mathcal{V}$ *and* $\mathcal{I}$ *give an inclusion reversing correspondence*

$$\left\{\begin{array}{c}\textit{Radical homogeneous ideals } I \textit{ of}\\ \mathbb{K}[x_0, \ldots, x_n] \textit{ properly contained in } \mathfrak{m}_0\end{array}\right\} \quad \underset{\mathcal{I}}{\overset{\mathcal{V}}{\rightleftarrows}} \quad \{\textit{Subvarieties } X \textit{ of } \mathbb{P}^n\}$$

with $\mathcal{V}(\mathcal{I}(X)) = X$*. When* $\mathbb{K}$ *is algebraically closed, the maps* $\mathcal{V}$ *and* $\mathcal{I}$ *are inverses, and this correspondence is a bijection.*

We can deduce this from the algebraic-geometric dictionary for affine space (Corollary 1.2.10), if we replace a subvariety X of projective space by its affine cone CX.

If we relax the condition that an ideal be radical, then the corresponding geometric objects are *projective schemes*. This comes at a price, for many homogeneous ideals will define the same projective scheme. This non-uniqueness comes from the irrelevant ideal, $\mathfrak{m}_0$. Recall the construction of colon ideals. Let I and J be ideals. Then the *colon ideal* (or *ideal quotient* of I by J) is

$$(I : J) \;:=\; \{f \mid fJ \subset I\}\,.$$

An ideal $I \subset \mathbb{K}[x_0, x_1, \ldots, x_n]$ is *saturated* if

$$I \;=\; (I : \mathfrak{m}_0) \;:=\; \{f \mid x_i f \in I \text{ for } i = 0, 1, \ldots, n\}\,.$$

The reason for this definition is that I and $(I : \mathfrak{m}_0)$ define the same projective scheme.

Given a projective variety $X \subset \mathbb{P}^n$, we may consider its intersection with any affine open subset $U_i = \{x \in \mathbb{P}^n \mid x_i \neq 0\}$. For simplicity of notation, we will work with $U_0 = \{[1, x_1, \ldots, x_n] \mid (x_1, \ldots, x_n) \in \mathbb{K}^n\} \simeq \mathbb{K}^n$. Then

$$X \cap U_0 \;=\; \{a \in U_0 \mid f(a) = 0 \text{ for all } f \in \mathcal{I}(X)\}\,.$$

and

$$\mathcal{I}(X \cap U_0) \;=\; \{f(1, x_1, \ldots, x_n) \mid f \in \mathcal{I}(X)\}\,.$$

We call the polynomial $f(1, x_1, \ldots, x_n)$ the *dehomogenization* of the homogeneous polynomial f. This shows that the ideal of $X \cap U_0$ is obtained by dehomogenizing the polynomials in the ideal of X. Note that f and $x_0^m f$ both dehomogenize to the same polynomial.

Conversely, given an affine subvariety $Y \subset U_0$, we have its Zariski closure[2] $\overline{Y} := \mathcal{V}(\mathcal{I}(Y)) \subset \mathbb{P}^n$. The relation between the ideal of the affine variety Y and homogeneous ideal of its closure $\overline{Y}$ is through homogenization.

$$\begin{aligned}
\mathcal{I}(\overline{Y}) &= \{f \in \mathbb{K}[x_0, \ldots, x_n] \mid f|_Y = 0\} \\
&= \{f \in \mathbb{K}[x_0, \ldots, x_n] \mid f(1, x_1, \ldots, x_n) \in \mathcal{I}(Y) \subset \mathbb{K}[x_1, \ldots, x_n]\} \\
&= \{x_0^{\deg(g)+m} g(\tfrac{x_1}{x_0}, \ldots, \tfrac{x_n}{x_0}) \mid g \in \mathcal{I}(Y),\ m \geq 0\}\,.
\end{aligned}$$

The point of this is that every projective variety X is naturally a union of affine varieties

$$X \;=\; \bigcup_{i=0}^{n} (X \cap U_i)\,.$$

This gives a relationship between varieties and manifolds: Affine varieties are to varieties what open subsets of $\mathbb{R}^n$ are to manifolds.

Could define quasi-projective varieties

[2]Use this notion earlier for closures of maps, but mention it is developed in Chapter 3.

1.5 Coordinate rings and maps of projective varieties

Given a projective variety $X \subset \mathbb{P}^n$, its *homogeneous coordinate ring* $\mathbb{K}[X]$ is the quotient

$$\mathbb{K}[X] \;:=\; \mathbb{K}[x_0, x_1, \ldots, x_n]/\mathcal{I}(X)\,.$$

If we set $\mathbb{K}[X]_d$ to be the images of all degree d homogeneous polynomials, $\mathbb{K}[x_0, \ldots, x_n]_d$, then this ring is graded,

$$\mathbb{K}[X] \;=\; \bigoplus_{d\geq 0} \mathbb{K}[X]_d\,,$$

where if $f \in \mathbb{K}[X]_d$ and $g \in \mathbb{K}[X]_e$, then $fg \in \mathbb{K}[X]_{d+e}$. More concretely, we have

$$\mathbb{K}[X]_d \;=\; \mathbb{K}[x_0, \ldots, x_n]_d/\mathcal{I}(X)_d\,,$$

where $\mathcal{I}(X)_d = \mathcal{I}(X) \cap \mathbb{K}[x_0, \ldots, x_n]_d$.

This differs from the coordinate ring of an affine variety as its elements are not functions on X, as we already observed that, apart from constant polynomials, elements of $\mathbb{K}[x_0, \ldots, x_n]$ do not give functions on $\mathbb{P}^n$.

Maps of projective varieities need to be treated much more carefully

However, given two homogeneous polynomials f and g which have the same degree, d, the quotient f/g does give a well-defined function, at least on $\mathbb{P}^n - \mathcal{V}(g)$. Indeed, if $[a_0, \ldots, a_n]$ and $[\lambda a_0, \ldots, \lambda a_n]$ are two representatives of the point $a \in \mathbb{P}^n$ and $g(a) \neq 0$, then

$$\frac{f(\lambda a_0, \ldots, \lambda a_n)}{g(\lambda a_0, \ldots, \lambda a_n)} \;=\; \frac{\lambda^d f(a_0, \ldots, a_n)}{\lambda^d g(a_0, \ldots, a_n)} \;=\; \frac{f(a_0, \ldots, a_n)}{g(a_0, \ldots, a_n)}\,.$$

It follows that if $f, g \in \mathbb{K}[X]$ with $g \neq 0$, then the quotient f/g gives a well-defined function on $X - \mathcal{V}(g)$.

More generally, let $f_0, f_1, \ldots, f_m \in \mathbb{K}[X]$ be elements of the same degree with at least one f_i non-zero on X. These define a *rational map*

$$\begin{array}{rccl} \varphi : & X & \dashrightarrow & \mathbb{P}^m \\ & x & \longmapsto & [f_0(x), f_1(x), \ldots, f_m(x)]\,. \end{array}$$

This is defined at least on the set $X - \mathcal{V}(f_0, \ldots, f_m)$. A second list $g_0, \ldots, g_m \in \mathbb{K}[X]$ of elements of the same degree (possible different from the degrees of the f_i) defines the same rational map if we have

$$\operatorname{rank} \begin{bmatrix} f_0 & f_1 & \cdots & f_m \\ g_0 & g_1 & \cdots & g_m \end{bmatrix} \;=\; 1 \qquad \text{i.e.} \quad f_i g_j - f_j g_i \;\in\; \mathcal{I}(X) \quad \text{for } i \neq j\,.$$

The map φ is regular at a point $x \in X$ if there is some system of representatives $f_0, \ldots, f_m$ for the map φ for which $x \notin \mathcal{V}(f_0, \ldots, f_m)$. The set of such points is an open subset of X called the *domain of regularity* of φ. The map φ is *regular* if it is regular at all points of X. The *base locus* of a rational map $\varphi\colon X \dashrightarrow Y$ is the set of points of X at which φ is not regular.

Example 1.5.1. An important example of a rational map is a linear projection. Let $\Lambda_0, \Lambda_1, \ldots, \Lambda_m$ be linear forms. These give a rational map φ which is defined at points of $\mathbb{P}^n - E$, where E is the common zero locus of the linear forms $\Lambda_0, \ldots, \Lambda_m$, that is $E = \mathbb{P}(\text{kernel}(L))$, where L is the matrix whose columns are the Λ_i.

The identification of $\mathbb{P}^1$ with the points on the circle $\mathcal{V}(x^2 + (y-1)^2 - 1) \subset \mathbb{K}^2$ from Example 1.4.2 is an example of a linear projection. Let $X := \mathcal{V}(x^2 + (y-z)^2 - z^2)$ be the plane conic which contains the point $[0, 0, 1]$. The identification of Example 1.4.2 was the map

$$\mathbb{P}^1 \ \ni \ [a, b] \ \longmapsto \ [2ab, 2a^2, a^2 + b^2] \ \in \ X \, .$$

Its inverse is the linear projection $[x, y, z] \mapsto [x, y]$.

Figure 1.5 shows another linear projection. Let C be the cubic space curve with parametrization $[1, t, t^2, 2t^3 - 2t]$ and $\pi \colon \mathbb{P}^3 \dashrightarrow L \simeq \mathbb{P}^1$ the linear projection defined by the last two coordinates, $\pi \colon [x_0, x_1, x_2, x_3] \mapsto [x_3, x_4]$. We have drawn the image $\mathbb{P}^1$ in the picture to illustrate that the inverse image of a linear projection is a linear section of the variety (after removing the base locus). The center of projection is a line, E, which meets

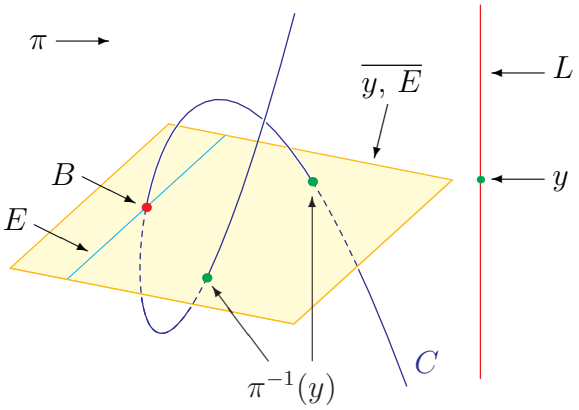


Figure 1.5: A linear projection π with center E.

the curve in a point, B.

Projective varieties $X \subset \mathbb{P}^n$ and $Y \subset \mathbb{P}^m$ are *isomorphic* if we have regular maps $\varphi \colon X \to Y$ and $\psi \colon Y \to X$ for which the compositions $\psi \circ \varphi$ and $\varphi \circ \psi$ are the identity maps on X and Y, respectively.

Exercises

1. A transition function $\varphi_{i,j}$ expresses how to change from the local coordinates from U_i of a point $p \in U_i \cap U_j$ to the local coordinates from U_j. Write down the transition functions for $\mathbb{P}^n$ provided by the affine charts $U_0, \ldots, U_n$.

2. Show that an ideal I is homogeneous if and only if it is generated by homogeneous polynomials.

3. Let $Z \subset \mathbb{P}^n$. Show that $\mathcal{I}(Z)$ is a homogeneous ideal.

4. Show that a radical homogeneous ideal is saturated.

5. Show that the homogeneous ideal $\mathcal{I}(Z)$ of a subset $Z \subset \mathbb{P}^n$ is equal to the ideal $\mathcal{I}(CZ)$ of the affine cone over Z.

6. Verify the claim in the text concerning the relation between the ideal of an affine subvariety $Y \subset U_0$ and of its Zariski closure $\overline{Y} \subset \mathbb{P}^n$:

$$\mathcal{I}(\overline{Y}) \;=\; \{x_0^{\deg(g)+m} g(\tfrac{x_1}{x_0}, \ldots, \tfrac{x_n}{x_0}) \mid g \in \mathcal{I}(Y) \subset \mathbb{K}[x_1, \ldots, x_n],\ m \geq 0\}\,.$$

7. Let $X \subset \mathbb{P}^n$ be a projective variety and suppose that $f, g \in \mathbb{K}[X]$ are homogeneous forms of the same degree with $g \neq 0$. Show that the quotient f/g gives a well-defined function on $X - \mathcal{V}(g)$.

8. Show that if I is a homogeneous ideal and J is its *saturation*,

$$J \;=\; \bigcup_{d \geq 0} (I : \mathfrak{m}_0^d)\,,$$

   then there is some integer N such that

$$J_d \;=\; I_d \qquad \text{for} \quad d \geq N\,.$$

9. Verify the claim in the text that if $X \subset \mathbb{P}^n$ is a projective variety, then its homogeneous coordinate ring is graded with

$$\mathbb{K}[X]_d \;=\; \mathbb{K}[x_0, \ldots, x_n]_d / \mathcal{I}(X)_d\,.$$

1.6 Notes

Most of the material in this chapter is standard material within courses of algebraic geometry or related courses. User-friendly, introductory texts to these topics include the books of Beltrametti, Carletti, Gallarati, and Monti Bragadin [5], Cox, Little, O'Shea [20], Holme [40], Hulek [42], Perrin [67], Smith, Kahanpää, Kekäläinen, and Traves [85]. Advanced, in-depth treatments from the viewpoint of modern, abstract algebraic geometry can be found in the books of Eisenbud [25], Harris [35], Hartshorne [36], and Shafarevich [84].

Chapter 2

Algorithms for Algebraic Geometry

Outline:

1. Gröbner basics.
2. Algorithmic applications of Gröbner bases.
3. Resultants and Bézout's Theorem.
4. Solving equations with Gröbner bases.
5. Numerical Homotopy continuation.
6. Numerical Algebraic Geometry

2.1 Gröbner basics

Gröbner bases are a foundation for many algorithms to represent and manipulate varieties on a computer. While these algorithms are important in applications, we shall see that Gröbner bases are also a useful theoretical tool.

A motivating problem is that of recognizing when a polynomial $f \in \mathbb{K}[x_1, \ldots, x_n]$ lies in an ideal I. When the ideal I is radical and $\mathbb{K}$ is algebraically closed, this is equivalent to asking whether or not f vanishes on $\mathcal{V}(I)$. For example, we may ask which of the polynomials $x^3z - xz^3$, $x^2yz - y^2z^2 - x^2y^2$, and/or $x^2y - x^2z + y^2z$ lies in the ideal

$$\langle x^2y - xz^2 + y^2z,\ y^2 - xz + yz \rangle \ ?$$

This *ideal membership problem* is easy for univariate polynomials. Suppose that $I = \langle f(x), g(x), \ldots, h(x) \rangle$ is an ideal and $F(x)$ is a polynomial in $\mathbb{K}[x]$, the ring of polynomials in a single variable x. We determine if $F(x) \in I$ via a two-step process.

1. Use the Euclidean Algorithm to compute $\varphi(x) = \gcd(f(x), g(x), \ldots, h(x))$.

2. Use the Division Algorithm to determine if $\varphi(x)$ divides $F(x)$.

This is valid, as $I = \langle \varphi(x) \rangle$. The first step is a simplifications, where we find a simpler (lower-degree) polynomial which generates I, while the second step is a reduction, where we compute F modulo I. Both steps proceed systematically, operating on the terms of the polynomials involving the highest power of x. A good description for I is a prerequisite for solving our ideal membership problem.

We shall see how Gröbner bases give algorithms which extend this procedure to multivariate polynomials. In particular, a Gröbner basis of an ideal I gives a sufficiently good description of I to solve the ideal membership problem. Gröbner bases are also the foundation of algorithms that solve many other problems.

2.1.1 Monomial ideals

Monomial ideals are central to what follows. A *monomial* is a product of powers of the variables $x_1, \ldots, x_n$ with nonnegative integer exponents. The *exponent* α of a monomial $x^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}$ is a vector $\alpha \in \mathbb{N}^n$. If we identify monomials with their exponent vectors, the multiplication of monomials corresponds to addition of vectors, and divisibility to the partial order on $\mathbb{N}^n$ of componentwise comparison.

Definition 2.1.1. A *monomial ideal* $I \subset \mathbb{K}[x_1, \ldots, x_n]$ is an ideal which satisfies the following two equivalent conditions.

(*i*) I is generated by monomials.

(*ii*) If $f \in I$, then every monomial of f lies in I.

One advantage of monomial ideals is that they are essentially combinatorial objects. By Condition (*ii*), a monomial ideal is determined by the set of monomials which it contains. Under the correspondence between monomials and their integer vector exponents, divisibility of monomials corresponds to componentwise comparison of vectors.

$$x^\alpha | x^\beta \iff \alpha_i \leq \beta_i \,,\ i = 1, \ldots, n \iff \alpha \leq \beta \,,$$

which defines a partial order on $\mathbb{N}^n$. Thus

$$(1,1,1) \leq (3,1,2) \qquad \text{but} \qquad (3,1,2) \not\leq (2,3,1) \,.$$

The set $O(I)$ of exponent vectors of monomials in a monomial ideal I has the property that if $\alpha \leq \beta$ with $\alpha \in O(I)$, then $\beta \in O(I)$. Thus $O(I)$ is an (upper) *order ideal* of the *poset* (*p*artially *o*rdered *set*) $\mathbb{N}^n$.

A set of monomials $G \subset I$ generates I if and only if every monomial in I is divisible by at least one monomial of G. A monomial ideal I has a unique minimal set of generators—these are the monomials x^α in I which are not divisible by any other monomial in I.

Let us look at some examples. When $n = 1$, monomials have the form x^d for some natural number $d \geq 0$. If d is the minimal exponent of a monomial in I, then $I = x^d$. Thus all monomial ideals have the form $\langle x^d \rangle$ for some $d \geq 0$.

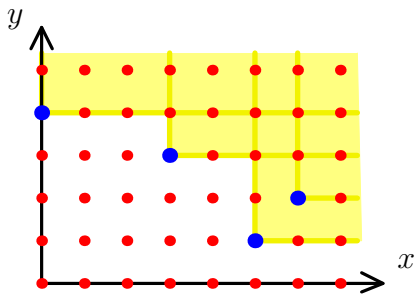
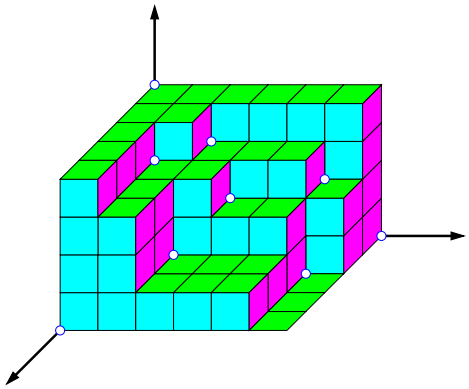


Figure 2.1: Exponents of monomials in the ideal $\langle y^4, x^3y^3, x^5y, x^6y^2\rangle$.

When $n = 2$, we may plot the exponents in the order ideal associated to a monomial ideal. For example, the lattice points in the shaded region of Figure 2.1 represent the monomials in the ideal $I := \langle y^4, x^3y^3, x^5y, x^6y^2\rangle$. From this picture we see that I is minimally generated by y^4, x^3y^3, and x^5y.

Since $x^ay^b \in I$ implies that $x^{a+\alpha}y^{b+\beta} \in I$ for any $(\alpha, \beta) \in \mathbb{N}^2$, a monomial ideal $I \subset \mathbb{K}[x, y]$ is the union of the shifted positive quadrants $(a, b) + \mathbb{N}^2$ for every monomial $x^ay^b \in I$. It follows that the monomials in I are those above the staircase shape that is the boundary of the shaded region. The monomials not in I lie under the staircase, and they form a vector space basis for the quotient ring $\mathbb{K}[x, y]/I$.

This notion of staircase for two variables makes sense when there are more variables. The *staircase* of an ideal consists of the monomials which are on the boundary of the ideal, in that they are visible from the origin of $\mathbb{N}^n$. For example, here is the staircase for the ideal $\langle x^5, x^2y^5, y^6, x^3y^2z, x^2y^3z^2, xy^5z^2, x^2yz^3, xy^2z^3, z^4\rangle$.



We offer a purely combinatorial proof that monomial ideals are finitely generated, which is independent of the Hilbert Basis Theorem.

Lemma 2.1.2 (Dickson's Lemma)**.** *Monomial ideals are finitely generated.*

Proof. We prove this by induction on n. The case $n = 1$ was covered in the preceding examples.

Let $I \subset \mathbb{K}[x_1, \ldots, x_n, y]$ be a monomial ideal. For each $d \in \mathbb{N}$, observe that the monomials

$$\{x^\alpha \mid x^\alpha y^d \in I\}$$

form a monomial ideal I_d of $\mathbb{K}[x_1, \ldots, x_n]$, and the union of all such monomials

$$\{x^\alpha \mid x^\alpha y^d \in I \text{ for some } d \geq 0\}\,.$$

form a monomial ideal I_∞ of $\mathbb{K}[x_1, \ldots, x_n]$. By our inductive hypothesis, I_d has a finite generating set G_d, for each $d = 0, 1, \ldots, \infty$.

Note that $I_0 \subset I_1 \subset \cdots \subset I_\infty$. We must have $I_\infty = I_d$ for some $d < \infty$. Indeed, each generator $x^\alpha \in G_\infty$ of I_∞ comes from a monomial $x^\alpha y^b$ in I, and we may let d be the maximum of the numbers b which occur. Since $I_\infty = I_d$, we have $I_b = I_d$ for any $b > d$. Note that if $b > d$, then we may assume that $G_b = G_d$ as $I_b = I_d$.

We claim that the finite set

$$G \;=\; \bigcup_{b=0}^{d} \{x^\alpha y^b \mid x^\alpha \in G_b\}$$

generates I. Indeed, suppose that $x^\alpha y^b$ is a monomial in I. We find a monomial in G which divides $x^\alpha y^b$. Since $x^\alpha \in I_b$, there is a generator $x^\gamma \in G_b$ which divides x^α. If $b \leq d$, then $x^\gamma y^b \in G$ is a monomial dividing $x^\alpha y^b$. If $b > d$, then $x^\gamma y^d \in G$ as $G_b = G_d$ and $x^\gamma y^d$ divides $x^\alpha y^b$. □

A simple consequence of Dickson's Lemma is that any strictly increasing chain of monomial ideals is finite. Suppose that

$$I_1 \;\subset\; I_2 \;\subset\; I_3 \;\subset\; \cdots$$

is an increasing chain of monomial ideals. Let I_∞ be their union, which is another monomial ideal. Since I_∞ is finitely generated, there must be some ideal I_d which contains all generators of I_∞, and so $I_d = I_{d+1} = \cdots = I_\infty$. We used this fact crucially in our proof of Dickson's lemma.

2.1.2 Monomial orders and Gröbner bases

The key idea behind Gröbner bases is to determine what is meant by 'term of highest power' in a polynomial having two or more variables. There is no canonical way to do this, so we must make a choice, which is encoded in the notion of a term or monomial order. An order $\succ$ on monomials in $\mathbb{K}[x_1, \ldots, x_n]$ is *total* if for monomials x^α and x^β exactly one of the following holds

$$x^\alpha \;\succ\; x^\beta \qquad \text{or} \qquad x^\alpha \;=\; x^\beta \qquad \text{or} \qquad x^\alpha \;\prec\; x^\beta\,.$$

Definition 2.1.3. A *monomial order* on $\mathbb{K}[x_1, \ldots, x_n]$ is a total order $\succ$ on the monomials in $\mathbb{K}[x_1, \ldots, x_n]$ such that

(*i*) 1 is the minimal element under $\succ$.

(*ii*) $\succ$ respects multiplication by monomials: If $x^\alpha \succ x^\beta$ then $x^\alpha \cdot x^\gamma \succ x^\beta \cdot x^\gamma$, for any monomial x^γ.

Conditions (*i*) and (*ii*) in Definition 2.1.3 imply that if x^α is divisible by x^β, then $x^\alpha \succ x^\beta$. A *well-ordering* is a total order with no infinite descending chain, equivalently, one in which every subset has a minimal element.

Lemma 2.1.4. *Monomial orders are exactly the well-orderings $\succ$ on monomials that satisfy Condition (ii) of Definition* 2.1.3.

Proof. Let $\succ$ be a well-ordering on monomials which satisfies Condition (ii) of Definition 2.1.3. Suppose that $\succ$ is not a monomial order, then there is some monomial x^a with $1 \succ x^a$. By Condition (ii), we have $1 \succ x^a \succ x^{2a} \succ x^{3a} \succ \cdots$, which contradicts $\succ$ being a well-order.

Let $\succ$ be a monomial order and M be any set of monomials. Set I to be the ideal generated by M. By Dickson's Lemma, M has a finite subset G which generates I. Since G is finite, let x^γ be the minimal monomial in G under $\succ$. We claim that x^γ is the minimal monomial in M.

Let $x^\alpha \in M$. Since G generates I and $M \subset I$, there is some $x^\beta \in G$ which divides x^α and thus $x^\alpha \succ x^\beta$. But x^γ is the minimal monomial in G, so $x^\alpha \succ x^\beta \succ x^\gamma$. □

The well-ordering property of monomials orders is key to what follows, as many proofs use induction on $\succ$, which is only possible as $\succ$ is a well-ordering.

Example 2.1.5. The (*total*) *degree* $\deg(x^\alpha)$ of a monomial $x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ is $\alpha_1 + \cdots + \alpha_n$. We describe four important monomial orders.

1. The *lexicographic order* $\succ_{\text{lex}}$ on $\mathbb{K}[x_1, \ldots, x_n]$ is defined by

$$x^\alpha \succ_{\text{lex}} x^\beta \iff \left\{ \begin{array}{l} \text{The first non-zero entry of the} \\ \text{vector } \alpha - \beta \text{ in } \mathbb{Z}^n \text{ is positive.} \end{array} \right\}$$

2. The *degree lexicographic order* $\succ_{\text{dlx}}$ on $\mathbb{K}[x_1, \ldots, x_n]$ is defined by

$$x^\alpha \succ_{\text{dlx}} x^\beta \iff \left\{ \begin{array}{ll} \deg x^\alpha > \deg x^\beta & \text{or}, \\ \deg x^\alpha = \deg x^\beta & \text{and } \ x^\alpha \succ_{\text{lex}} x^\beta . \end{array} \right.$$

3. The *degree reverse lexicographic order* $\succ_{\text{drl}}$ $\mathbb{K}[x_1, \ldots, x_n]$ is defined by

$$x^\alpha \succ_{\text{drl}} x^\beta \iff \left\{ \begin{array}{ll} \deg x^\alpha > x^\beta & \text{or}, \\ \deg x^\alpha = \deg x^\beta & \text{and the last non-zero entry of the} \\ & \text{vector } \alpha - \beta \text{ in } \mathbb{Z}^n \text{ is negative}. \end{array} \right.$$

4. More generally, we may have *weighted orders*. Let $c \in \mathbb{R}^n$ be a vector with non-negative components, called a weight. This defines a partial order $\succ_c$ on monomials

$$x^\alpha \succ_c x^\beta \iff c \cdot \alpha > c \cdot \beta \,.$$

If all components of c are positive, then $\succ_c$ satisfies the two conditions of Definition 2.1.3. Its only failure to be a monomial order is that it may not be a total order on monomials. (For example, consider $c = (1, 1, \ldots, 1)$.) This may be remedied by picking a monomial order to break ties. For example, if we use $\succ_{\text{lex}}$, then we get a monomial order

$$x^\alpha \succ_{c,\text{lex}} x^\beta \iff \begin{cases} \omega \cdot \alpha > \omega \cdot \beta & \text{or}\,, \\ \omega \cdot \alpha = \omega \cdot \beta & \text{and } x^\alpha \succ_{\text{lex}} x^\beta \end{cases}$$

Another way to do this is to break the ties with a different monomial order, or a different weight.

You are asked to prove these are monomial orders in Exercise 7.

Remark 2.1.6. We compare these three orders on monomials of degrees 1 and 2 in $\mathbb{K}[x, y, z]$ where the variables are ordered $x \succ y \succ z$.

$$\begin{aligned} &x^2 \succ_{\text{lex}} xy \succ_{\text{lex}} xz \succ_{\text{lex}} x \succ_{\text{lex}} y^2 \succ_{\text{lex}} yz \succ_{\text{lex}} y \succ_{\text{lex}} z^2 \succ_{\text{lex}} z \\ &x^2 \succ_{\text{dlx}} xy \succ_{\text{dlx}} xz \succ_{\text{dlx}} y^2 \succ_{\text{dlx}} yz \succ_{\text{dlx}} z^2 \succ_{\text{dlx}} x \succ_{\text{dlx}} y \succ_{\text{dlx}} z \\ &x^2 \succ_{\text{drl}} xy \succ_{\text{drl}} y^2 \succ_{\text{drl}} xz \succ_{\text{drl}} yz \succ_{\text{drl}} z^2 \succ_{\text{drl}} x \succ_{\text{drl}} y \succ_{\text{drl}} z \end{aligned}$$

For the remainder of this section, $\succ$ will denote a fixed, but arbitrary monomial order on $\mathbb{K}[x_1, \ldots, x_n]$. A *term* is a product ax^α of a scalar $a \in \mathbb{K}$ with a monomial x^α. We may extend any monomial order $\succ$ to an order on terms by setting $ax^\alpha \succ bx^\beta$ if $x^\alpha \succ x^\beta$ and $ab \neq 0$. Such a term order is no longer a partial order as different terms with the same monomial are incomparable. For example $3x^2y$ and $5x^2y$ are incomparable. Term orders are however well-founded in that they have no infinite strictly decreasing chains.

The *initial term* $\text{in}_\succ(f)$ of a polynomial $f \in \mathbb{K}[x_1, \ldots, x_n]$ is the term of f that is maximal with respect to $\succ$ among all terms of f. For example, if $\succ$ is lexicographic order with $x \succ y$, then

$$\text{in}_\succ(3x^3y - 7xy^{10} + 13y^{30}) \;=\; 3x^3y \,.$$

When $\succ$ is understood, we may write $\text{in}(f)$. The *initial ideal* $\text{in}_\succ(I)$ (or $\text{in}(I)$) of an ideal $I \subset \mathbb{K}[x_1, \ldots, x_n]$ is the ideal generated by the initial terms of polynomials in I,

$$\text{in}_\succ(I) \;=\; \langle \text{in}_\succ(f) \mid f \in I \rangle \,.$$

We make the most important definition of this section.

Definition 2.1.7. Let $I \subset \mathbb{K}[x_1, \ldots, x_n]$ be an ideal and $\succ$ a monomial order. A set $G \subset I$ is a *Gröbner basis* for I with respect to the monomial order $\succ$ if the initial ideal $\mathrm{in}_\succ(I)$ is generated by the initial terms of polynomials in G, that is, if

$$\mathrm{in}_\succ(I) \;=\; \langle \mathrm{in}_\succ(g) \mid g \in G \rangle \,.$$

Notice that if G is a Gröbner basis and $G \subset G'$, then G' is also a Gröbner basis.

We justify our use of the term 'basis' in 'Gröbner basis'.

Lemma 2.1.8. *If G is a Gröbner basis for I with respect to a monomial order $\succ$, then G generates I.*

Proof. We will prove this by induction on $\mathrm{in}(f)$ for $f \in I$. Let $f \in I$. Since $\{\mathrm{in}(g) \mid g \in G\}$ generates $\mathrm{in}(I)$, there is a polynomial $g \in G$ whose initial term $\mathrm{in}(g)$ divides the initial term $\mathrm{in}(f)$ of f. Thus there is some term ax^α so that

$$\mathrm{in}(f) \;=\; ax^\alpha \mathrm{in}(g) \;=\; \mathrm{in}(ax^\alpha g)\,,$$

as $\succ$ respects multiplication. If we set $f_1 := f - cx^\alpha g$, then $\mathrm{in}(f) \succ \mathrm{in}(f_1)$.

It follows that if $\mathrm{in}(f)$ is the minimal monomial in $\mathrm{in}I$, then $f_1 = 0$ and so $f \in \langle G \rangle$. In fact, $f \in G$ up to a scalar multiple. Suppose now that whenever $\mathrm{in}(f) \succ \mathrm{in}(h)$ and $h \in I$, then $h \in \langle G \rangle$. But then $f_1 = f - cx^\alpha g \in \langle G \rangle$, and so $f \in \langle G \rangle$. □

An immediate consequence of Dickson's Lemma and Lemma 2.1.8 is the following Gröbner basis version of the Hilbert Basis Theorem.

Theorem 2.1.9 (Hilbert Basis Theorem)**.** *Every ideal $I \subset \mathbb{K}[x_1, \ldots, x_n]$ has a finite Gröbner basis with respect to any given monomial order.*

Example 2.1.10. Different monomial orderings give different Gröbner bases, and the sizes of the Gröbner bases can vary. Consider the ideal generated by the three polynomials

$$xy^3 + xz^3 + x - 1, \quad yz^3 + yx^3 + y - 1, \quad zx^3 + zy^3 + z - 1$$

In the degree reverse lexicographic order, where $x \succ y \succ z$, this has a Gröbner basis

$x^3z + y^3z + z - 1$,
$xy^3 + xz^3 + x - 1$,
$x^3y + yz^3 + y - 1$,
$y^4z - yz^4 - y + z$,
$2xyz^4 + xyz + xy - xz - yz$,
$2y^3z^3 - x^3 + y^3 + z^3 + x^2 - y^2 - z^2$,
$y^6 - z^6 - y^5 + y^3z^2 - 2x^2z^3 - y^2z^3 + z^5 + y^3 - z^3 - x^2 - y^2 + z^2 + x$,
$x^6 - z^6 - x^5 - y^3z^2 - x^2z^3 - 2y^2z^3 + z^5 + x^3 - z^3 - x^2 - y^2 + y + z$,
$2z^7 + 4x^2z^4 + 4y^2z^4 - 2z^6 + 3z^4 - x^3 - y^3 + 3x^2z + 3y^2z - 2z^3 + x^2 + y^2 - 2xz - 2yz - z^2 + z - 1$,
$2yz^6 + y^4 + 2yz^3 + x^2y - y^3 + yz^2 - 2z^3 + y - 1$,

$2xz^6 + x^4 + 2xz^3 - x^3 + xy^2 + xz^2 - 2z^3 + x - 1,$

consisting of 11 polynomials with largest coefficient 4 and degree 7. If we consider instead the lexicographic monomial order, then this ideal has a Gröbner basis

$$64z^{34} - 64z^{33} + 384z^{31} - 192z^{30} - 192z^{29} + 1008z^{28} + 48z^{27} - 816z^{26} + 1408z^{25} + 976z^{24}$$
$$-1296z^{23} + 916z^{22} + 1964z^{21} - 792z^{20} - 36z^{19} + 1944z^{18} + 372z^{17} - 405z^{16} + 1003z^{15}$$
$$+879z^{14} - 183z^{13} + 192z^{12} + 498z^{11} + 7z^{10} - 94z^9 + 78z^8 + 27z^7 - 47z^6 - 31z^5 + 4z^3$$
$$-3z^2 - 4z - 1,$$

$$64yz^{21} + 288yz^{18} + 96yz^{17} + 528yz^{15} + 384yz^{14} + 48yz^{13} + 504yz^{12} + 600yz^{11} + 168yz^{10}$$
$$+200yz^9 + 456yz^8 + 216yz^7 + 120yz^5 + 120yz^4 - 8yz^2 + 16yz + 8y - 64z^{33} + 128z^{32}$$
$$-128z^{31} - 320z^{30} + 576z^{29} - 384z^{28} - 976z^{27} + 1120z^{26} - 144z^{25} - 2096z^{24} + 1152z^{23}$$
$$+784z^{22} - 2772z^{21} + 232z^{20} + 1520z^{19} - 2248z^{18} - 900z^{17} + 1128z^{16} - 1073z^{15} - 1274z^{14}$$
$$+229z^{13} - 294z^{12} - 966z^{11} - 88z^{10} - 81z^9 - 463z^8 - 69z^7 + 26z^6 - 141z^5 - 32z^4 + 24z^3$$
$$-12z^2 - 11z + 1$$

$$589311934509212912y^2 - 11786238690184258240yz^{20} - 9428990952147406592yz^{19}$$
$$-2357247738036851648yz^{18} - 4832357862975545878 4yz^{17} - 4832357862975545878 4yz^{16}$$
$$-2003660577331323900 8yz^{15} - 8191435889678059476 8yz^{14} - 9782578112852934339 2yz^{13}$$
$$-5303807410582916208 0yz^{12} - 7867314325697992375 2yz^{11} - 9988837289931158858 4yz^{10}$$
$$-6364568892699499449 6yz^9 - 3712665187408041345 6yz^8 - 4390373912093636194 4yz^7$$
$$-3447474816878895535 2yz^6 - 9134334984892800136yz^5 - 5893119345092129120yz^4$$
$$-4125183541564490384yz^3 - 1178623869018425824yz^2 - 2062591770782245192yz$$
$$-1178623869018425824y + 46665645155349846336z^{33} - 52561386330338650688z^{32}$$
$$+25195872352020329920z^{31} + 28156769162372952723 2z^{30} - 19392177430724378694 4z^{29}$$
$$-22383823960598695936z^{28} + 81706533724600969099 2z^{27} - 16308104685758723524 8z^{26}$$
$$-42770559036883403033 6z^{25} + 13905781683718208538 08z^{24} + 39000434368484674580 8z^{23}$$
$$-98032219788785598166 4z^{22} + 13454251172212979738 76z^{21} + 12879560659390367316 76z^{20}$$
$$-95338316228249822884 4z^{19} + 63120234731058122985 6z^{18} + 17043019678692273960 24z^{17}$$
$$-15520856778655514998 8z^{16} - 16764066862257396505z^{15} + 12574754032771507009 61z^{14}$$
$$+52668596890136716959 8z^{13} - 16475153000055626488 0z^{12} + 49124953163927565405 0z^{11}$$
$$+45712630887118688230 6z^{10} - 87008396189513562747z^9 + 15803768907185828750z^8$$
$$+13932068156394410127 3z^7 - 17355919586383317961z^6 - 50777365233910819054z^5$$
$$-4630862847055988750z^4 + 8085080238139562826z^3 + 1366850803924776890z^2$$
$$-3824545208919673161z - 2755936363893486164,$$

$$589311934509212912x + 589311934509212912y - 87966378396509318592z^{33}$$
$$+13338340253167146649 6z^{32} - 59115312141727767552z^{31} - 50692680764859328012 8z^{30}$$
$$+52214177181017233427 2z^{29} + 48286434009450032640z^{28} - 14347259883383736388752z^{27}$$
$$+62997181176686959171 2z^{26} + 91798600277439166526 4z^{25} - 23898711989748432051 36z^{24}$$
$$-24698231483106694188 8z^{23} + 20389689261052715195 36z^{22} - 21748963896433430866 20z^{21}$$
$$-17581387825462211569 76z^{20} + 20253901854065627985 52z^{19} - 77454264142036382836 4z^{18}$$
$$-23653906414512782784 84z^{17} + 62782483555936330499 92z^{16} + 39848463323285911590 7z^{15}$$
$$-15486831101309342203 22z^{14} - 50019266671009151041 9z^{13} + 55192142799847475851 0z^{12}$$
$$-49036879434510228641 0z^{11} - 48050400484189905738 4z^{10} + 22051400745440117561 5z^9$$

$+3851598490198004730 5z^8 - 1366443016356866846 09z^7 + 1741071269413252079 4z^6$
$+5872455235409422580 3z^5 + 1570234197189530735 6z^4 - 7440058907697789332z^3$
$-1398341089468668912z^2 + 3913205630531612397z + 2689145244006168857,$
consisting of 4 polynomials with largest degree 34 and significantly larger coefficients.

Exercises for Section 1

1. Prove the equivalence of conditions (i) and (ii) in Definition 2.1.1.

2. Show that a monomial ideal is radical if and only if it is square-free. (Square-free means that it has generators in which no variable occurs to a power greater than 1.)

3. Show that the elements of a monomial ideal I which are minimal with respect to division form a minimal set of generators of I in that they generate I and are a subset of any generating set of I.

4. Which of the polynomials $x^3z - xz^3$, $x^2yz - y^2z^2 - x^2y^2$, and/or $x^2y - x^2z + y^2z$ lies in the ideal
$$\langle x^2y - xz^2 + y^2z,\ y^2 - xz + yz \rangle \ ?$$

5. Using Definition 2.1.1, show that a monomial order is a linear extension of the divisibility partial order on monomials.

6. Show that if an ideal I has a square-free initial ideal, then I is radical. Give an example to show that the converse of this statement is false.

7. Show that each of the order relations $\succ_{\text{lex}}$, $\succ_{\text{dlx}}$, and $\succ_{\text{drl}}$, are monomial orders. Show that if the coordinates of $\omega \in \mathbb{R}^n_>$ are linearly independent over $\mathbb{Q}$, then $\succ_\omega$ is a monomial order. Show that each of $\succ_{\text{lex}}$, $\succ_{\text{dlx}}$, and $\succ_{\text{drl}}$ are weighted orders.

8. Show that a term order is well-founded.

9. Show that for a monomial order $\succ$, $\text{in}(I)\text{in}(J) \subseteq \text{in}(IJ)$ for any two ideals I and J. Find I and J such that the inclusion is proper.

10. Let $I := \langle x^2 + y^2, x^3 + y^3 \rangle \subset \mathbb{Q}[x, y]$. Let our monomial order be $\succ_{\text{lex}}$, the lexicographic order with $x \succ_{\text{lex}} y$.

    (a) Prove that $y^4 \in I$.

    (b) Show that the reduced Gröbner basis for I is $\{y^4, xy^2 - y^3, x^2 + y^2\}$.

    (c) Show that $\{x^2 + y^2, x^3 + y^3\}$ cannot be a Gröbner basis for I for any monomial ordering.

2.2 Algorithmic applications of Gröbner bases

Many practical algorithms to study and manipulate ideals and varieties are based on Gröbner bases. The foundation of algorithms involving Gröbner bases is the multivariate division algorithm. The subject began with Buchberger's thesis which contained his algorithm to compute Gröbner bases [15, 17].

2.2.1 Ideal membership and standard monomials

Both steps in the algorithm for ideal membership in one variable relied on the the same elementary procedure: using a polynomial of low degree to simplify a polynomial of higher degree. This same procedure was also used in the proof of Lemma 2.1.8. Those ideas lead to the *multivariate division algorithm*, which is a cornerstone of the theory of Gröbner bases.

Algorithm 2.2.1 (Multivariate division algorithm)**.** INPUT: Polynomials $g_1, \ldots, g_m$ and f in $\mathbb{K}[x_1, \ldots, x_n]$ and a monomial order $\succ$.
OUTPUT: Polynomials $q_1, \ldots, q_m$ and r such that

$$f \;=\; q_1 g_1 + q_2 g_2 + \cdots + q_m g_m + r\,, \tag{2.1}$$

where no term of r is divisible by an initial term of any polynomial g_i and we also have $\operatorname{in}(f) \succeq \operatorname{in}(r)$, and $\operatorname{in}(f) \succeq \operatorname{in}(q_i g_i)$, for each $i = 1, \ldots, m$.
INITIALIZE: Set $r := f$ and $q_1 := 0$, $\ldots$, $q_m := 0$.

(1) If no term of r is divisible by an initial term of some g_i, then exit.

(2) Otherwise, let ax^α be the largest (with respect to $\succ$) term of r divisible by some $\operatorname{in}(g_i)$. Choose j minimal such that $\operatorname{in}(g_j)$ divides x^α and suppose that $ax^\alpha = bx^\beta \cdot \operatorname{in}(g_j)$. Replace r by $r - bx^\beta g_j$ and q_j by $q_j + bx^\beta$, and return to step (1).

Proof of correctness. Each iteration of (2) is a *reduction* of r by the polynomials $g_1, \ldots, g_m$. With each reduction, the largest term in r divisible by some $\operatorname{in}(g_i)$ decreases with respect to $\succ$. Since the monomial order $\succ$ is well-founded, this algorithm must terminate after a finite number of steps. Every time the algorithm executes step (1), condition (2.1) holds. We also always have $\operatorname{in}(f) \succeq \operatorname{in}(r)$ because it holds initially, and with every reduction the new terms of r are less than the term which was canceled. Lastly, $\operatorname{in}(f) \succeq \operatorname{in}(q_i g_i)$ always holds, because it held initially, and the initial terms of the $q_i g_i$ are always terms of r. □

Given a list $G = (g_1, \ldots, g_m)$ of polynomials and a polynomial f, let r be the remainder obtained by the multivariate division algorithm applied to G and f. Since $f - r$ lies in the ideal generated by G, we write $f \bmod G$ for this remainder r. While it is clear (and expected) that $f \bmod G$ depends on the monomial order $\succ$, in general it will also depend

upon the order of the polynomials $(g_1, \ldots, g_m)$. For example, in the degree lexicographic order

$$\begin{aligned} x^2y \mod (x^2,\, xy + y^2) &= 0\,, \quad \text{but} \\ x^2y \mod (xy + y^2,\, x^2) &= y^3\,. \end{aligned}$$

This example shows that we cannot reliably use the multivariate division algorithm to test when f is in the ideal generated by G. However, this does not occur when G is a Gröbner basis.

Lemma 2.2.2 (Ideal membership test)**.** *Let G be a finite Gröbner basis for an ideal I with respect to a monomial order $\succ$. Then a polynomial $f \in I$ if and only if $f \mod G = 0$.*

Proof. Set $r := f \mod G$. If $r = 0$, then $f \in I$. Suppose $r \neq 0$. Since no term of r is divisible any initial term of a polynomial in G, its initial term $\mathrm{in}(r)$ is not in the initial ideal of I, as G is a Gröbner basis for I. But then $r \notin I$, and so $f \notin I$. □

When G is a Gröbner basis for an ideal I, the remainder $f \mod G$ is a linear combination of monomials that do not lie in the initial ideal of I. A monomial x^α is *standard* if $x^\alpha \notin \mathrm{in}(I)$. The images of standard monomials in the ring $\mathbb{K}[x_1, \ldots, x_n]/\mathrm{in}(I)$ form a vector space basis. Much more interesting is the following theorem of Macaulay [55].

Theorem 2.2.3. *Let $I \subset \mathbb{K}[x_1, \ldots, x_n]$ be an ideal and $\succ$ a monomial order. Then the images of standard monomials in $\mathbb{K}[x_1, \ldots, x_n]/I$ form a vector space basis.*

Proof. Let G be a finite Gröbner basis for I with respect to $\succ$. Given a polynomial f, both f and $f \mod G$ represent the same element in $\mathbb{K}[x_1, \ldots, x_n]/I$. Since $f \mod G$ is a linear combination of standard monomials, the standard monomials span $\mathbb{K}[x_1, \ldots, x_n]/I$.

A linear combination f of standard monomials is zero in $\mathbb{K}[x_1, \ldots, x_n]/I$ only if $f \in I$. But then $\mathrm{in}(f)$ is both standard and lies in $\mathrm{in}(I)$, and so we conclude that $f = 0$. Thus the standard monomials are linearly independent in $\mathbb{K}[x_1, \ldots, x_n]/I$. □

Because of this result, if we have a monomial order $\succ$ and an ideal I, then for every polynomial $f \in \mathbb{K}[x_1, \ldots, x_n]$, there is a unique polynomial $\overline{f}$ which involves only standard monomials such that f and $\overline{f}$ have the same image in the quotient ring $\mathbb{K}[x_1, \ldots, x_m]/I$. Moreover, this polynomial $\overline{f}$ may be computed from f with the division algorithm and any Gröbner basis G for I with respect to the monomial order $\succ$. This unique representative $\overline{f}$ of f is typically called the *normal form* of f modulo I and the division algorithm called with a Gröbner basis for I is often called normal form reduction.

Macaulay's Theorem shows that a Gröbner basis allows us to compute in the quotient ring $\mathbb{K}[x_1, \ldots, x_n]/I$ using the operations of the polynomial ring and ordinary linear algebra. Indeed, suppose that G is a finite Gröbner basis for an ideal I with respect to a given monomial order $\succ$ and that $f, g \in \mathbb{K}[x_1, \ldots, x_n]/I$ are in normal form, expressed as a linear combination of standard monomials. Then $f + g$ is a linear combination of standard monomials, and we can compute the product fg in the quotient ring as $fg \mod G$, where this product is taken in the polynomial ring.

2.2.2 Buchberger's algorithm

Theorem 2.1.9, which asserted the existence of a finite Gröbner basis, was purely existential. To use Gröbner bases, we need methods to detect and generate them. Such methods were given by Bruno Buchberger in his 1965 Ph.D. thesis [17]. Key ideas about Gröbner bases had appeared earlier in work of Gordan and of Macaulay, and in Hironaka's resolution of singularities [39]. Hironaka called Gröbner bases "standard bases", a term which persists. For example, in the computer algebra package `Singular` [31] the command `std(I);` computes the Gröbner basis of an ideal `I`. Despite these precedents, the theory of Gröbner bases rightly begins with these Buchberger's contributions.

A given set of generators for an ideal will fail to be a Gröbner basis if the initial terms of the generators fail to generate the initial ideal. That is, if there are polynomials in the ideal whose initial terms are not divisible by the initial terms of our generators. A necessary step towards generating a Gröbner basis is to generate polynomials in the ideal with 'new' initial terms. This is the *raison d'etre* for the following definition.

Definition 2.2.4. The *least common multiple*, $\mathrm{lcm}\{ax^\alpha, bx^\beta\}$ of two terms ax^α and bx^β is the minimal monomial x^γ divisible by both x^α and x^β. Here, γ is the exponent vector that is the componentwise maximum of α and β.

Let $0 \neq f, g \in \mathbb{K}[x_1, \ldots, x_n]$ and suppose $\succ$ is a monomial order. The *S-polynomial* of f and g, $\mathrm{Spol}(f, g)$, is the polynomial linear combination of f and g,

$$\mathrm{Spol}(f,g) \;:=\; \frac{\mathrm{lcm}\{\mathrm{in}(f),\mathrm{in}(g)\}}{\mathrm{in}(f)} f \;-\; \frac{\mathrm{lcm}\{\mathrm{in}(f),\mathrm{in}(g)\}}{\mathrm{in}(g)} g\,.$$

Note that both terms in this expression have initial term equal to $\mathrm{lcm}\{\mathrm{in}(f), \mathrm{in}(g)\}$.

Buchberger gave the following simple criterion to detect when a set G of polynomials is a Gröbner basis for the ideal it generates.

Theorem 2.2.5 (Buchberger's Criterion)**.** *A set G of polynomials is a Gröbner basis for the ideal it generates if and only if for for all pairs $f, g \in G$,*

$$\mathrm{Spol}(f,g) \bmod G \;=\; 0\,.$$

Proof. Suppose first that G is a Gröbner basis for I with respect to $\succ$. Then, for $f, g \in G$, their S-polynomial $\mathrm{Spol}(f, g)$ lies in I and the ideal membership test implies that $\mathrm{Spol}(f, g) \bmod G = 0$.

Now suppose that $G = \{g_1, \ldots, g_m\}$ satisfies Buchberger's criterion. Let $f \in \langle G \rangle$ be a polynomial in the ideal generated by G. We will show that $\mathrm{in}(f)$ is divisible by $\mathrm{in}(g)$, for some $g \in G$. This implies that G is a Gröbner basis for $\langle G \rangle$.

Given a list $h = (h_1, \ldots, h_m)$ of polynomials in $\mathbb{K}[x_1, \ldots, x_n]$ let $m(h)$ be the largest monomial appearing in one of $h_1 g_1, \ldots, h_m g_m$. This will necessarily be the leading monomial in at least one of $\mathrm{in}(h_1 g_1), \ldots, \mathrm{in}(h_m g_m)$. Let $j(h)$ be the minimum index i for which $m(h)$ is the monomial of $\mathrm{in}(h_i g_i)$.

Consider lists $h = (h_1, \ldots, h_m)$ of polynomials with

$$f \;=\; h_1 g_1 + \cdots + h_m g_m \tag{2.2}$$

for which $m(h)$ minimal among all lists satisfying (2.2). Of these, let h be a list with $j := j(h)$ maximal. We claim that $m(h)$ is the monomial of $\mathrm{in}(f)$, which implies that $\mathrm{in}(g_j)$ divides $\mathrm{in}(f)$.

Otherwise, $m(h) \succ \mathrm{in}(f)$, and so the initial term $\mathrm{in}(h_j g_j)$ must be canceled in the sum (2.2). Thus there is some index k such that $m(h)$ is the monomial of $\mathrm{in}(h_k g_k)$. By our assumption on j, we have $k > j$. Let $x^\beta := \mathrm{lcm}\{\mathrm{in}(g_j), \mathrm{in}(g_k)\}$, the monomial which is canceled in $\mathrm{Spol}(g_j, g_k)$. Since $\mathrm{in}(g_j)$ and $\mathrm{in}(g_k)$ both divide $m(h)$ and thus both must divide $\mathrm{in}(h_j g_j)$, there is some term ax^α such that $ax^\alpha x^\beta = \mathrm{in}(h_j g_j)$. Let $cx^\gamma := \mathrm{in}(h_j g_j)/\mathrm{in}(g_k)$. Then

$$ax^\alpha \mathrm{Spol}(g_j, g_k) \;=\; ax^\alpha \frac{x^\beta}{\mathrm{in}(g_j)} g_j \;-\; ax^\alpha \frac{x^\beta}{\mathrm{in}(g_k)} g_k \;=\; \mathrm{in}(h_j) g_j \;-\; cx^\gamma g_k \,.$$

By our assumption that Buchberger's criterion is satisfied, there are polynomials $q_1, \ldots, q_m$ such that

$$\mathrm{Spol}(g_j, g_k) \;=\; q_1 g_1 + \cdots + q_m g_m \,.$$

Define a new list h' of polynomials,

$$h' \;=\; (h_1 + ax^\alpha q_1, \ldots, h_j - \mathrm{in}(h_j) + ax^\alpha q_j, \ldots, h_k + cx^\gamma + ax^\alpha q_k, \ldots, h_m + ax^\alpha q_m)\,,$$

and consider the sum $\sum h'_i g_i$, which is

$$\sum_i h_i g_i + ax^\alpha \sum_i q_i g_i - \mathrm{in}(h_j) g_j + cx^\gamma g_k$$
$$= \; f + ax^\alpha \mathrm{Spol}(g_j, g_k) - ax^\alpha \mathrm{Spol}(g_j, g_k) \;=\; f \,.$$

By the division algorithm, $\mathrm{in}(q_i g_i) \preceq \mathrm{in}(\mathrm{Spol}(g_j, g_k))$, so $\mathrm{in}(ax^\alpha q_i g_i) \prec x^\alpha x^\beta = m(h)$. But then $m(h') \preceq m(h)$. By our assumption, $m(h') = m(h)$. Since $\mathrm{in}(h_j - \mathrm{in}(h_j)) \prec \mathrm{in}(h_j)$, we have $j(h') > j = j(h)$, which contradicts our choice of h. □

Buchberger's algorithm to compute a Gröbner basis begins with a list of polynomials and augments that list by adding reductions of S-polynomials. It halts when the list of polynomials satisfy Buchberger's Criterion.

Algorithm 2.2.6 (Buchberger's Algorithm)**.** Begin with a list of generators $G = (g_1, \ldots, g_m)$ for an ideal. For each $i < j$, let $h_{ij} := \mathrm{Spol}(g_i, g_j) \bmod G$. If each of these is zero, then we have a Gröbner basis, by Buchberger's Criterion. Otherwise append all the non-zero h_{ij} to the list G and repeat this process.

This algorithm terminates after finitely many steps, because the initial terms of polynomials in G after each step generate a strictly larger monomial ideal and Dickson's Lemma implies that any increasing chain of monomial ideals is finite. Since the manipulations in Buchberger's algorithm involve only algebraic operations using the coefficients of the input polynomials, we deduce the following corollary, which is important when studying real varieties. Let $\mathbb{K}$ be any field containing $\mathbb{K}$.

Corollary 2.2.7. *Let $f_1, \ldots, f_m \in \mathbb{K}[x_1, \ldots, x_n]$ be polynomials and $\succ$ a monomial order. Then there is a Gröbner basis $G \subset \mathbb{K}[x_1, \ldots, x_n]$ for the ideal $\langle f_1, \ldots, f_m \rangle$ in $\mathbb{K}[x_1, \ldots, x_n]$*

Example 2.2.8. Consider applying the Buchberger algorithm to $G = (x^2, xy + y^2)$ with any monomial order where $x \succ y$. First

$$\mathrm{Spol}(x^2, xy + y^2) \;=\; y \cdot x^2 - x(xy + y^2) \;=\; -xy^2 \,.$$

Then

$$-xy^2 \bmod (x^2, xy + y^2) \;=\; -xy^2 + y(xy + y^2) \;=\; y^3 \,.$$

Since all S-polynomials of $(x^2, xy + y^2, y^3)$ reduce to zero, this is a Gröbner basis.

Among the polynomials h_{ij} computed at each stage of the Buchberger algorithm are those where one of $\mathrm{in}(g_i)$ or $\mathrm{in}(g_j)$ divides the other. Suppose that $\mathrm{in}(g_i)$ divides $\mathrm{in}(g_j)$ with $i \neq j$. Then $\mathrm{Spol}(g_i, g_j) = g_j - mg_i$, where m is some term. This has strictly smaller initial term than does g_j and so we never use g_j to compute $h_{ij} := \mathrm{Spol}(g_i, g_j) \bmod G$. It follows that $g_j - h_{ij}$ lies in the ideal generated by $G \setminus \{g_j\}$, and so we may replace g_j by h_{ij} in G without changing the ideal generated by G, and only possibly increasing the ideal generated by the initial terms of polynomials in G.

This gives the following elementary improvement to the Buchberger algorithm:

$$\begin{array}{l}\text{In each step, initially compute } h_{ij} \text{ for those } i \neq j \\ \text{where } \mathrm{in}(g_i) \text{ divides } \mathrm{in}(g_j), \text{ and replace } g_j \text{ by } h_{ij}.\end{array} \tag{2.3}$$

In some important cases, this step computes the Gröbner basis. Another improvement, which identifies S-polynomials which reduce to zero and therefore do not need to be computed, is given in the exercises.

There are additional improvements in Buchberger's algorithm (see Ch. 2.9 in [20] for a discussion), and even a series of completely different algorithms due to Jean-Charles Faugère [27] based on linear algebra with vastly improved performance.

A Gröbner basis G is *reduced* if the initial terms of polynomials in G are monomials with coefficient 1 and if for each $g \in G$, no monomial of g is divisible by an initial term of another Gröbner basis element. A reduced Gröbner basis for an ideal is uniquely determined by the monomial order. Reduced Gröbner bases are the multivariate analog of unique monic polynomial generators of ideals of $\mathbb{K}[x]$. Elements f of a reduced Gröbner basis also have a special form,

$$x^\alpha - \sum_{\beta \in B} c_\beta x^\beta \,,$$

where $x^\alpha = \text{in}(f)$ is the initial term and B consists of exponent vectors of standard monomials. This rewrites the nonstandard initial monomial as a linear combination of standard monomials. The reduced Gröbner basis has one generator for every generator of the initial ideal.

Example 2.2.9. Let M be a $m \times n$ matrix, which we consider to be the matrix of coefficients of m linear forms $g_1, \ldots, g_m$ in $\mathbb{K}[x_1, \ldots, x_n]$, and suppose that $x_1 \succ x_2 \succ \cdots \succ x_n$. We can apply (2.3) to two forms g_i and g_j when their initial terms have the same variable. Then the S-polynomial and subsequent reductions are equivalent to the steps in the algorithm of Gaussian elimination applied to the matrix M. If we iterate our applications of (2.3) until the initial terms of the forms g_i have distinct variables, then the forms $g_1, \ldots, g_m$ are a Gröbner basis for the ideal they generate.

If the forms g_i are a reduced Gröbner basis and are sorted in decreasing order according to their initial terms, then the resulting matrix $\overline{M}$ of their coefficients is an *echelon matrix*: The initial non-zero entry in each row is 1 and is the only non-zero entry in its column and these columns increase with row number.

Gaussian elimination produces the same echelon matrix from M. In this way, we see that the Buchberger algorithm is a generalization of Gaussian elimination to non-linear polynomials.

Exercises for Section 2

1. Describe how Buchberger's algorithm behaves when it computes a Gröbner basis from a list of monomials. What if we use the elementary improvement (2.3)?

2. Use Buchberger's algorithm to compute the reduced Gröbner basis of $\langle y^2 - xz + yz,\ x^2y - xz^2 + y^2z \rangle$ in the degree reverse lexicographic order where $x \succ y \succ z$.

3. Let $f, g \in \mathbb{K}[x_1, \ldots, x_m]$ be such that $\text{in}(f)$ and $\text{in}(g)$ are relatively prime and the leading coefficients of f and g are 1. Show that

$$S(f,g) \ = \ -(g - \text{in}(g))f + (f - \text{in}(f))g\,.$$

   Deduce that the leading monomial of $S(f,g)$ is a multiple of either the leading monomial of f or of g in this case.

4. The following problem shows that every ideal has a finite generating set that is a Gröbner basis with respect to all term orderings. Such a generating set is called a *universal Gröbner basis*. Let $I \subset \mathbb{K}[x_1, \ldots, x_n]$ be an ideal.

   (a) Show that there are only finitely many initial ideals of I. More precisely show that

$$\{\text{in}_\succ(I) \mid \succ \text{ is a term order on } \mathbb{K}[x_1, \ldots, x_n]\}$$

   is a finite set.

(b) Show that every ideal $I \subset \mathbb{K}[x_1, \ldots, x_n]$ has a set of generators that is a Gröbner basis for every term order.

5. Let U be a universal Gröbner basis for an ideal I in $\mathbb{K}[x_1, \ldots, x_n]$. Show that for every subset $Y \subset \{x_1, \ldots, x_n\}$ the *elimination ideal* $I \cap \mathbb{K}[Y]$ is generated by $U \cap \mathbb{K}[Y]$.

6. Let I be a ideal generated by homogeneous linear polynomials. We call a nonzero linear form f in I a *circuit* of I if f has minimal support (with respect to inclusion) among all polynomials in I. Prove that the set of all circuits of I is a universal Gröbner basis of I.

7. (a) Prove that the ideal $\langle x, y \rangle \subset \mathbb{Q}[x, y]$ is not a principal ideal.
   (b) Is $\langle x^2 + y, x + y \rangle$ already a Gröbner basis with respect to some term ordering?
   (c) Use Buchberger's algorithm to compute a Gröbner basis of the ideal $I = \langle y - z^2, z - x^3 \rangle \in \mathbb{Q}[x, y, z]$ with lexicographic and the degree reverse lexicographic monomial orders.

8. This exercise illustrates an algorithm to compute the saturation of ideals. Let $I \subset \mathbb{K}[x_1, \ldots, x_n]$ be an ideal, and fix $f \in \mathbb{K}[x_1, \ldots, x_n]$. Then the *saturation* of I with respect to f is the set

$$(I : f^\infty) \;=\; \{g \in \mathbb{K}[x_1, \ldots, x_n] \mid f^m g \in I \text{ for some } m > 0\}\,.$$

   (a) Prove that $(I : f^\infty)$ is an ideal.
   (b) Prove that we have an ascending chain of ideals

$$(I : f) \;\subset\; (I : f^2) \;\subset\; (I : f^3) \;\subset \cdots$$

   (c) Prove that there exists a nonnegative integer N such that $(I : f^\infty) = (I : f^N)$.
   (d) Prove that $(I : f^\infty) = (I : f^m)$ if and only if $(I : f^m) = (I : f^{m+1})$.

   When the ideal I is homogeneous and $f = x_n$ then one can use the following strategy to compute the saturation. Fix the degree reverse lexicographic order $\succ_{\text{drl}}$ where $x_1 \succ_{\text{drl}} x_2 \succ_{\text{drl}} \cdots \succ_{\text{drl}} x_n$ and let G be a reduced Gröbner basis of a homogeneous ideal $I \subset \mathbb{K}[x_1, \ldots, x_n]$.

   (e) Show that the set

$$G' \;=\; \{f \in G \mid x_n \text{ does not divide } f\} \bigcup \{f/x_n \mid f \in G \text{ and } x_n \text{ divides } f\}$$

   is a Gröbner basis of $(I : x_n)$.
   (f) Show that a Gröbner basis of $(I : x_n^\infty)$ is obtained by dividing each element $f \in G$ by the highest power of x_n that divides f.

9. Suppose that $\prec$ is the lexicographic order with $x \prec y \prec z$.

   (a) Apply Buchberger's algorithm to the ideal $\langle x + y, xy \rangle$.

   (b) Apply Buchberger's algorithm to the ideal $\langle x + y + z, xy + xz + yz, xyz \rangle$.

   (c) Define the *elementary symmetric polynomials* $e_i(x_1, \ldots, x_n)$ by

$$\sum_{i=0}^{n} t^{n-i} e_i(x_1, \ldots, x_n) \;=\; \prod_{i=1}^{n} (t + x_i)\,,$$

   that is, $e_0 = 1$ and if $i > 0$, then

$$e_i(x_1, \ldots, x_n) \;:=\; e_i(x_1, \ldots, x_{n-1}) \;+\; x_n e_{i-1}(x_1, \ldots, x_{n-1})\,.$$

   Alternatively, $e_i(x_1, \ldots, x_n)$ is also the sum of all square-free monomials of total degree i in $x_1, \ldots, x_n$.
   The *symmetric ideal* is $\langle e_i(x_1, \ldots, x_n) \mid 1 \leq i \leq n \rangle$. Describe its Gröbner basis, and the set of standard monomials with respect to lexicographic order when $x_1 \prec x_2 \prec \cdots \prec x_n$.
   What about degree reverse lexicographic order?

2.3 Resultants and Bézout's Theorem

Algorithms based on Gröbner bases are universal in that their input may be any list of polynomials. This comes at a price as Gröbner basis algorithms may have poor performance and the output is quite sensitive to the input. An alternative foundation for some algorithms is provided by resultants. These are are special polynomials having determinantal formulas which were introduced in the 19th century. A drawback is that they are not universal—different inputs require different algorithms, and for many inputs, there are no formulas for resultants.

The key algorithmic step in the Euclidean algorithm for the greatest common divisor (gcd) of two univariate polynomials f and g in $\mathbb{K}[x]$ with $n = \deg(g) \geq \deg(f) = m$,

$$\begin{aligned} f &= f_0 x^m + f_1 x^{m-1} + \cdots + f_{m-1} x + f_m \\ g &= g_0 x^n + g_1 x^{n-1} + \cdots + g_{n-1} x + g_n \,, \end{aligned} \tag{2.4}$$

is to replace g by

$$g - \frac{g_0}{f_0} x^{n-m} \cdot f \,,$$

which has degree at most $n-1$. In some applications (for example, when $\mathbb{K}$ is a function field), we will want to avoid division. Resultants give a way to detect common factors without using division. We will use them for much more than this.

2.3.1 Sylvester Resultant

Let S_ℓ or $S_\ell(x)$ be the set of polynomials in $\mathbb{K}[x]$ of degree at most ℓ. This is a vector space over $\mathbb{K}$ of dimension $\ell + 1$ with a canonical ordered basis of monomials $x^\ell, \ldots, x, 1$. Given f and g as above, consider the linear map

$$\begin{aligned} \varphi_{f,g} \,:\, S_{n-1} \times S_{m-1} &\longrightarrow S_{m+n-1} \\ (h(x), k(x)) &\longmapsto f \cdot h + g \cdot k \,. \end{aligned}$$

The domain and range of $\varphi_{f,g}$ each have dimension $m + n$.

Lemma 2.3.1. *The polynomials f and g have a nonconstant common divisor if and only if* $\ker \varphi_{f,g} \neq \{(0,0)\}$.

Proof. Suppose first that f and g have a nonconstant common divisor, p. Then there are polynomials h and k with $f = pk$ and $g = ph$. As p is nonconstant, $\deg k < \deg f = n$ and $\deg h < \deg g = m$ so that $(h, -k) \in S_{n-1} \times S_{m-1}$. Since

$$fh - gk \;=\; pkh - phk \;=\; 0 \,,$$

we see that $(h, -k)$ is a nonzero element of the kernel of $\varphi_{f,g}$.

Suppose that f and g are relatively prime and let $(h,k) \in \ker \varphi_{f,g}$. Since $\langle f, g\rangle = \mathbb{K}[x]$, there exist polynomials p and q with $1 = gp + fq$. Using $0 = fh + gk$ we obtain

$$k \;=\; k\cdot 1 \;=\; k(gp+fq) \;=\; gkp + fkq \;=\; -fhp + fkq \;=\; f(kq - hp)\,.$$

This implies that $k = 0$ for otherwise $m-1 \geq \deg k > \deg f = m$, which is a contradiction. We similarly see that $h = 0$, and so $\ker \varphi_{f,g} = \{(0,0)\}$. □

The matrix of the linear map $\varphi_{f,g}$ in the ordered bases of monomials for $S_{m-1} \times S_{n-1}$ and S_{m+n-1} is called the *Sylvester matrix*. When f and g have the form (2.4), it is

$$\mathrm{Syl}(f,g;x) \;=\; \mathrm{Syl}(f,g) \;:=\; \left(\begin{array}{ccccc|ccc} f_0 & & & & & g_0 & & 0 \\ f_1 & f_0 & & 0 & & g_1 & \ddots & \\ \vdots & \vdots & \ddots & & & \vdots & & g_0 \\ f_m & \vdots & & \ddots & & \vdots & & \vdots \\ & f_m & & & f_0 & g_{n-1} & & \vdots \\ & & \ddots & & \vdots & g_n & & \vdots \\ & 0 & & \ddots & \vdots & & \ddots & \vdots \\ & & & & f_m & 0 & & g_n \end{array}\right) . \qquad (2.5)$$

Note that the sequence $f_0, \ldots, f_0, g_n, \ldots, g_n$ lies along the main diagonal and the left side of the matrix has n columns while the right side has m columns.

The *(Sylvester) resultant* $\mathrm{Res}(f,g)$ is the determinant of the Sylvester matrix. To emphasize that the Sylvester matrix represents the map $\varphi_{f,g}$ in the basis of monomials in x, we also write $\mathrm{Res}(f,g;x)$ for $\mathrm{Res}(f,g)$. We summarize some properties of resultants, which follow from its formula as the determinant of the Sylvester matrix (2.5) and from Lemma 2.3.1.

Theorem 2.3.2. *The resultant of two nonconstant polynomials $f, g \in \mathbb{K}[x]$ is an irreducible integer polynomial in the coefficients of f and g. The resultant vanishes if and only if f and g have a nonconstant common factor.*

We only need to prove that the resultant is irreducible. The path we choose to this will give another expression for the resultant as well as a geometric interpretation of the resultant.

Lemma 2.3.3. *Suppose that $\mathbb{K}$ contains all the roots of the polynomials f and g so that we have*

$$f(x) \;=\; f_0 \prod_{i=1}^{m} (x - \alpha_i) \qquad \text{and} \qquad g(x) \;=\; g_0 \prod_{i=1}^{n} (x - \beta_i)\,,$$

where $\alpha_1, \ldots, \alpha_m$ are the roots of f and $\beta_1, \ldots, \beta_n$ are the roots of g. Then

$$\mathrm{Res}(f, g; x) \;=\; (-1)^{mn} f_0^n g_0^m \prod_{i=1}^{m} \prod_{j=1}^{n} (\alpha_i - \beta_j)\,. \tag{2.6}$$

A corollary of this lemma is the Poisson formula,

$$\mathrm{Res}(f, g; x) \;=\; (-1)^{mn} f_0^m \prod_{i=1}^{m} g(\alpha_i) \;=\; g_0^n \prod_{i=1}^{n} f(\beta_i)\,.$$

Proof. Consider these formulas as expressions in $\mathbb{Z}[f_0, g_0, \alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_n]$. Recall that the coefficients of f and g are essentially the elementary symmetric polynomials in their roots,

$$f_i \;=\; (-1)^i f_0 e_i(\alpha_1, \ldots, \alpha_m) \qquad \text{and} \qquad g_i \;=\; (-1)^i g_0 e_i(\beta_1, \ldots, \beta_n)\,.$$

We claim that both sides of (2.6) are homogeneous polynomials of degree mn in the variables $\alpha_1, \ldots, \beta_n$. This is straightforward for the right hand side. To see this for the resultant, we extend our notation, setting $f_i := 0$ when $i < 0$ or $i > m$ and $g_i := 0$ when $i < 0$ or $i > n$. Then the entry in row i and column j of the Sylvester matrix is

$$\mathrm{Syl}(f, g; x)_{i,j} \;=\; \begin{cases} f_{i-j} & \text{if } j \le n\,, \\ g_{n+i-j} & \text{if } n < j \le m+n\,. \end{cases}$$

The determinant is a signed sum over permutations w of $\{1, \ldots, m+n\}$ of terms

$$\prod_{j=1}^{n} f_{w(j)-j} \;\cdot\; \prod_{j=n+1}^{m+n} g_{n+w(j)-j}\,.$$

Since f_i and g_i are each homogeneous of degree i in the variables $\alpha_1, \ldots, \beta_n$, this term is homogeneous of degree

$$\sum_{j=1}^{m} w(j)-j \quad + \quad \sum_{j=n+1}^{m+n} n+w(j)-j \;=\; mn + \sum_{j=1}^{m+n} w(j)-j \;=\; mn\,.$$

Both sides of (2.6) vanish exactly when some $\alpha_i = \beta_j$. Since they have the same degree, they are proportional. We compute this constant of proportionality. The term in $\mathrm{Res}(f, g)$ which is the product of diagonal entries of the Sylvester matrix is

$$f_0^n g_n^m \;=\; f_0^n g_0^m e_n(\beta_1, \ldots, \beta_n)^m \;=\; f_0^n g_0^m \beta_1^m \cdots \beta_n^m\,.$$

This is the only term of $\mathrm{Res}(f, g)$ involving the monomial $\beta_1^m \cdots \beta_n^m$. The corresponding term on the right hand side of (2.6) is

$$(-1)^{mn} f_0^n g_0^m (-\beta_1)^m \cdots (-\beta_n)^m \;=\; f_0^n g_0^m \beta_1^m \cdots \beta_n^m\,,$$

which completes the proof. □

Suppose that $\mathbb{K}$ is algebraically closed and consider the variety of all triples consisting of a pair of polynomials with a common root, together with a common root,

$$\Sigma \;:=\; \{(f,g,a) \in S_m \times S_n \times \mathbb{A}^1 \mid f(a) = g(a) = 0\}\,.$$

This has projections $p\colon \Sigma \to S_m \times S_n$ and $\pi\colon \Sigma \to \mathbb{A}^1$. The image $p(\Sigma)$ is the set of pairs of polynomials having a common root, which is the variety $\mathcal{V}(\mathrm{Res})$ of the resultant polynomial, $\mathrm{Res} \in \mathbb{Z}[f_0,\ldots,f_m,g_0,\ldots,g_n]$.

The fiber of π over a point $a \in \mathbb{A}^1$ consists all pairs of polynomials f, g with $f(a) = g(a) = 0$. Since each equation is linear in the coefficients of the polynomials f and g, this fiber is isomorphic to $\mathbb{A}^m \times \mathbb{A}^n$. Since $\pi\colon \Sigma \to \mathbb{A}^1$ has irreducible image ($\mathbb{A}^1$) and irreducible fibers, we see that Σ is irreducible, and has dimension $1+m+n$.[†]

This implies that $p(\Sigma)$ is irreducible. Furthermore, the fiber $p^{-1}(f,g)$ is the set of common roots of f and g. This is a finite set when $f, g \neq (0,0)$. Thus $p(\Sigma)$ has dimension $1+m+n$, and is thus an irreducible hypersurface in $S_m \times S_n$. Let F be a polynomial generating the ideal $\mathcal{I}(p(\Sigma))$, which is necessarily irreducible. As $\mathcal{V}(\mathrm{Res}) = p(\Sigma)$, we must have $\mathrm{Res} = F^N$ for some positive integer N. The formula (2.6) shows that $N = 1$ as the resultant polynomial is square-free.

Corollary 2.3.4. *The resultant polynomial is irreducible. It is the unique (up to sign) irreducible integer polynomial in the coefficients of f and g that vanishes on the set of pairs of polynomials (f,g) which have a common root.*

We only need to show that the greatest common divisor of the coefficients of the integer polynomial Res is 1. But this is clear as Res contains the term $f_0^n g_n^m$, as we showed in the proof of Lemma 2.3.3.

When both f and g have the same degree n, there is an alternative determinantal formula for their resultant. The *Bezoutian polynomial* of f and g is the bivariate polynomial

$$\Delta_{f,g}(y,z) \;:=\; \frac{f(y)g(z) - f(z)g(y)}{y-z} \;=\; \sum_{i,j=0}^{n-1} b_{i,j} y^i z^j \;.$$

The $n \times n$ matrix $\mathrm{Bez}(f,g)$ whose entries are the coefficients $(b_{i,j})$ of the Bezoutian polynomial is called the *Bezoutian matrix* of f and g. Note that each entry of the Bezoutian matrix $\mathrm{Bez}(f,g)$ is a linear combination of the brackets $[ij] := f_i g_j - f_j g_i$. For example, when $n = 2$ and $n = 3$, the Bezoutian matrices are

$$\begin{pmatrix} [02] & [12] \\ [01] & [02] \end{pmatrix} \qquad \begin{pmatrix} [03] & [13] & [23] \\ [02] & [03]+[12] & [13] \\ [01] & [02] & [03] \end{pmatrix} \;.$$

Theorem 2.3.5. *When f and g both have degree n,* $\mathrm{Res}(f,g) = (-1)^{\binom{n}{2}} \det(\mathrm{Bez}(f,g))$.

[†]This will need to cite results from Chapter 1 on dimension and irreducibility

Proof. Suppose that $\mathbb{K}$ is algebraically closed. Let B be the determinant of the Bezoutian matrix and Res the resultant of the polynomials f and g, both of which lie in the ring $\mathbb{K}[f_0, \ldots, f_n, g_0, \ldots, g_n]$. Then B is a homogeneous polynomial of degree $2n$, as is the resultant. Suppose that f and g are polynomials having a common root, $a \in \mathbb{K}$ with $f(a) = g(a) = 0$. Then the Bezoutian polynomial $\Delta_{f,g}(y, z)$ vanishes when $z = a$,

$$\Delta_{f,g}(y, a) \;=\; \frac{f(y)g(a) - f(a)g(y)}{y - a} \;=\; 0\,.$$

Thus

$$0 \;=\; \sum_{i,j=0}^{n-1} b_{i,j} y^i a^j \;=\; \sum_{i=0}^{n-1} \Big(\sum_{j=0}^{n-1} b_{i,j} a^j \Big)\, y^i\,.$$

Since every coefficient of this polynomial in y must vanish, the vector $(1, a, a^2, \ldots, a^{d-1})^T$ lies in the kernel of the Bezoutian matrix, and so the determinant $B(f, g)$ of the Bezoutian matrix vanishes.

Since the resultant generates the ideal of the pairs (f, g) of polynomial that are not relatively prime, Res divides B. As they have the same degree B is a constant multiple of Res. In Exercise 4 you are asked to show this constant is $(-1)^{\binom{n}{2}}$. □

Example 2.3.6. We give an application of resultants. A polynomial $f \in \mathbb{K}[x]$ of degree n has fewer than n distinct roots in the algebraic closure of $\mathbb{K}$ when it has a factor in $\mathbb{K}[x]$ of multiplicity greater than 1, and in that case f and its derivative f' have a factor in common. The *discriminant* of f is a polynomial in the coefficients of f which vanishes precisely when f has a repeated factor. It is defined to be

$$\mathrm{disc}(f) \;:=\; \frac{(-1)^{\binom{n}{2}}}{f_0} \mathrm{Res}(f, f')\,.$$

The discriminant is a polynomial of degree $2n - 2$ in the coefficients $f_1, \ldots, f_n$.

2.3.2 Resultants and Elimination

Resultants do much more than detect the existence of common factors in two polynomials. One of their most important uses is to eliminate variables from multivariate equations. The first step towards this is another interesting formula involving the Sylvester resultant. Not only is it a polynomial in the coefficients, but it has a canonical expression as a polynomial linear combination of f and g.

Lemma 2.3.7. *Given univariate polynomials $f, g \in \mathbb{K}[x]$, there are polynomials $h, k \in \mathbb{K}[x]$ whose coefficients are integer polynomials in the coefficients of f and g such that*

$$f(x)h(x) + g(x)k(x) \;=\; \mathrm{Res}(f, g)\,. \tag{2.7}$$

Proof. Set $\mathbb{K} := \mathbb{Q}(f_0, \ldots, f_m, g_0, \ldots, g_n)$, the field of rational functions (quotients of integer polynomials) in the indeterminates $f_0, \ldots, f_m, g_0, \ldots, g_n$ and let $f, g \in \mathbb{K}[x]$ be univariate polynomials as in (2.4). Then $\gcd(f, g) = 1$ and so the map $\varphi_{f,g}$ is invertible.

Set $(h, k) := \varphi_{f,g}^{-1}(\mathrm{Res}(f, g))$ so that

$$f(x)h(x) \;+\; g(x)k(x) \;=\; \mathrm{Res}(f, g)\,,$$

with $h, k \in \mathbb{K}[x]$ where $h \in S_{n-1}(x)$ and $k \in S_{m-1}(x)$.

Recall the adjoint formula for the inverse of a $n \times n$ matrix A,

$$\det(A) \cdot A^{-1} \;=\; \mathrm{ad}(A)\,. \tag{2.8}$$

Here $\mathrm{ad}(A)$ is the *adjoint* of the matrix A. Its (i, j)-entry is $(-1)^{i+j} \cdot \det A_{i,j}$, where $A_{i,j}$ is the $(n-1) \times (n-1)$ matrix obtained from A by deleting its ith column and jth row.

Since $\det \varphi_{f,g} = \mathrm{Res}(f, g) \in \mathbb{K}$, we have

$$\varphi_{f,g}^{-1}(\mathrm{Res}(f, g)) \;=\; \det \varphi_{f,g} \cdot \varphi_{f,g}^{-1}(1) \;=\; \mathrm{ad}(\mathrm{Syl}(f, g))(1)\,.$$

In the monomial basis for S_{m+n-1} the polynomial 1 corresponds to the vector $(0, \ldots, 0, 1)$. Thus, the coefficients of $\varphi_{f,g}^{-1}(\mathrm{Res}(f, g))$ are given by the entries of the last column of $\mathrm{ad}(\mathrm{Syl}(f, g))$, which are $\pm$ the minors of the Sylvester matrix $\mathrm{Syl}(f, g)$ with its last row removed. In particular, these are polynomials in the indeterminates $f_0, \ldots, g_n$ having integer coefficients. □

This proof shows that $h, k \in \mathbb{Z}[f_0, \ldots, f_m, g_0, \ldots, g_m][x]$ and that (2.7) holds as an expression in this polynomial ring with $m+n+3$ variables. It also shows that if $f, g \in \mathbb{K}[x_1, \ldots, x_n]$ are multivariate polynomials, and we hide all of the variables except x_1 in their coefficients, considering them as polynomials in x_n with coefficients in $\mathbb{K}(x_2, \ldots, x_n)$, then the resultant lies in both the ideal generated by f and g, and in the subring $\mathbb{K}[x_2, \ldots, x_n]$. We examine the geometry of this elimination of variables.

Suppose that $1 \leq i < n$ and let $\pi\colon \mathbb{A}^n \to \mathbb{A}^{n-i}$ be the coordinate projection

$$\pi \;:\; (a_1, \ldots, a_n) \;\longmapsto\; (a_{i+1}, \ldots, a_n)\,.$$

Lemma 2.3.8. *Let $I \subset \mathbb{K}[x_1, \ldots, x_n]$ be an ideal. Then $\pi(\mathcal{V}(I)) \subset \mathcal{V}(I \cap \mathbb{K}[x_{i+1}, \ldots, x_n])$.*

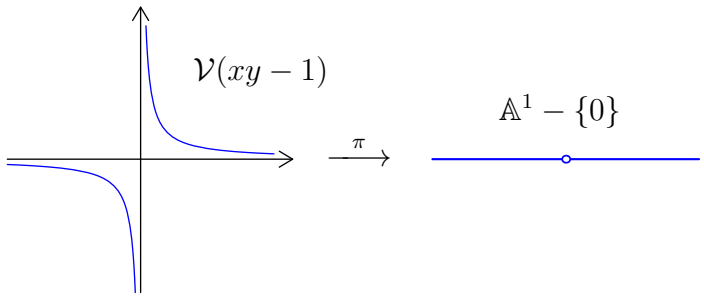
Proof. Suppose that $a = (a_1, \ldots, a_n) \in \mathcal{V}(I)$. If $f \in I \cap \mathbb{K}[x_{i+1}, \ldots, x_n]$, then

$$0 \;=\; f(a) \;=\; f(a_{i+1}, \ldots, a_n) \;=\; f(\pi(a))\,.$$

Note that we may view f as a polynomial in either $x_1, \ldots, x_n$ or in $x_{i+1}, \ldots, x_n$. The statement of the lemma follows. □

The ideal $I \cap \mathbb{K}[x_{i+1}, \ldots, x_n]$ is called an *elimination ideal* as the variables $x_1, \ldots, x_i$ have been eliminated from the ideal I. By Lemma 2.3.8, elimination is the algebraic counterpart to projection, but the correspondence is not exact. For example, the inclusion

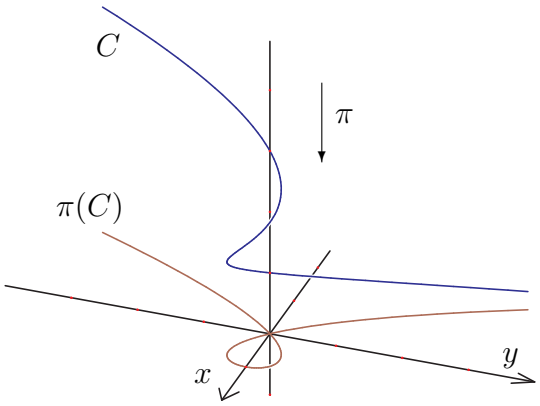
$\pi(\mathcal{V}(I)) \subset \mathcal{V}(I \cap \mathbb{K}[x_{i+1}, \ldots, x_n])$ may be strict. Let $\pi\colon \mathbb{A}^2 \to \mathbb{A}^1$ be the map which forgets the first coordinate. Then $\pi(\mathcal{V}(xy-1)) = \mathbb{A}^1 - \{0\} \subsetneq \mathbb{A}^1 = V(0)$ and $\{0\} = \langle xy-1 \rangle \cap F[y]$.



Elimination of variables enables us to solve the implicitization problem for plane curves. For example, consider the parametric plane curve

$$x \;=\; t^2 - 1, \qquad y \;=\; t^3 - t\,. \tag{2.9}$$

This is the image of the space curve $C := \mathcal{V}(t^2 - 1 - x, t^3 - t - y)$ under the projection $(t, x, y) \mapsto (x, y)$. We display this with the t-axis vertical and the xy-plane at $t = -2$.



By lemma 2.3.8, the equation for the plane curve is $\langle t^2 - x - x, t^3 - t - y \rangle \cap \mathbb{K}[x, y]$. If we set

$$f(t) \;:=\; t^2 - 1 - x \qquad \text{and} \qquad g(t) \;:=\; t^3 - t - y\,,$$

then the Sylvester resultant is

$$\det \left[\begin{array}{ccc|cc} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ -x-1 & 0 & 1 & -1 & 0 \\ 0 & -x-1 & 0 & -y & -1 \\ 0 & 0 & -x-1 & 0 & -y \end{array}\right] \;=\; y^2 + x^2 - x^3\,,$$

which is the implicit equation of the parameterized $\pi(C)$ cubic (2.9).

2.3.3 Resultants and Bézout's Theorem

We use resultants to study the variety $\mathcal{V}(f,g) \subset \mathbb{A}^2$ for $f, g \in \mathbb{K}[x,y]$. A by-product will be a form of Bézout's Theorem bounding the number of points in the variety $\mathcal{V}(f,g)$.

Suppose now that we have two variables, x and y. The ring $\mathbb{K}[x,y]$ of bivariate polynomials is a subring of the ring $\mathbb{K}(y)[x]$ of polynomials in x whose coefficients are rational functions in y. Suppose that $f, g \in \mathbb{K}[x,y]$. If we consider f and g as elements of $\mathbb{K}(y)[x]$, then the resultant $\mathrm{Res}(f,g;x)$ is the determinant of their Sylvester matrix expressed in the basis of monomials in x. Theorem 2.3.2 implies that $\mathrm{Res}(f,g;x)$ is a univariate polynomial in y which vanishes if and only if f and g have a common factor in $\mathbb{K}(y)[x]$. In fact it vanishes if and only if $f(x,y)$ and $g(x,y)$ have a common factor in $\mathbb{K}[x,y]$ with positive degree in y, by the following version of Gauss's lemma for $\mathbb{K}[x,y]$.

Lemma 2.3.9. *Polynomials f and g in $\mathbb{K}[x,y]$ have a common factor of positive degree in x if and only if they have a common factor in $\mathbb{K}(y)[x]$.*

Proof. The forward direction is clear. For the reverse, suppose that

$$f \;=\; h \cdot \overline{f} \qquad \text{and} \qquad g \;=\; h \cdot \overline{g} \tag{2.10}$$

is a factorization in $\mathbb{K}(y)[x]$ where h has positive degree in x.

There is a polynomial $d \in \mathbb{K}[y]$ which is divisible by every denominator of a coefficient of h, $\overline{f}$, and $\overline{g}$. Multiplying the expressions (2.10) by d^2 gives

$$d^2 f \;=\; (dh) \cdot (d\overline{f}) \qquad \text{and} \qquad d^2 g \;=\; (dh) \cdot (d\overline{g})\,,$$

where dh, $d\overline{f}$, and $d\overline{g}$ are polynomials in $\mathbb{K}[x,y]$. Let $k(x,y)$ be an irreducible factor of dh having positive degree in x. Then k divides both $d^2 f$ and $d^2 g$. However, k cannot divide d as $d \in \mathbb{K}[y]$ and k has positive degree in x. Therefore $k(x,y)$ is the desired common polynomial factor of f and g. □

Let $\pi\colon \mathbb{A}^2 \to \mathbb{A}^1$ be the projection which forgets the first coordinate, $\pi(x,y) = y$. Set $I := \langle f, g\rangle \cap \mathbb{K}[y]$. By Lemma 2.3.7, the resultant $\mathrm{Res}(f,g;x)$ lies in I. Combining this with Lemma 2.3.8 gives the chain of inclusions

$$\pi(\mathcal{V}(f,g)) \;\subset\; \mathcal{V}(I) \;\subset\; \mathcal{V}(\mathrm{Res}(f,g;x))\,.$$

We now suppose that $\mathbb{K}$ is algebraically closed. Let $f, g \in \mathbb{K}[x,y]$ and write

$$\begin{aligned} f &= f_0(y)x^m + f_1(y)x^{m-1} + \cdots + f_{m-1}(y)x + f_m(y) \\ g &= g_0(y)x^n + g_1(y)x^{n-1} + \cdots + g_{n-1}(y)x + g_n(y)\,, \end{aligned}$$

where neither $f_0(y)$ nor $g_0(y)$ is the zero polynomial.

Theorem 2.3.10 (Extension Theorem)**.** *If $b \in \mathcal{V}(I) - \mathcal{V}(f_0(y), g_0(y))$, then there is some $a \in \mathbb{K}$ with $(a,b) \in \mathcal{V}(f,g)$.*

This establishes the chain of inclusions of subvarieties of $\mathbb{A}^1$

$$\mathcal{V}(I) - \mathcal{V}(f_0, g_0) \ \subset \ \pi(\mathcal{V}(f,g)) \ \subset \ \mathcal{V}(I) \subset \ \mathcal{V}(\mathrm{Res}(f,g;x)) \, .$$

Proof. Let $b \in \mathcal{V}(I) - \mathcal{V}(f_0, g_0)$. Suppose first that $f_0(b) \cdot g_0(b) \neq 0$. Then $f(x,b)$ and $g(x,b)$ are polynomials in x of degrees m and n, respectively. It follows that the Sylvester matrix $\mathrm{Syl}(f(x,b), g(x,b))$ has the same format (2.5) as the Sylvester matrix $\mathrm{Syl}(f,g;x)$, and it is in fact obtained from $\mathrm{Syl}(f,g;x)$ by substituting $y = b$.

This implies that $\mathrm{Res}(f(x,b), g(x,b))$ is the evaluation of the resultant $\mathrm{Res}(f,g;x)$ at $y = b$. Since $\mathrm{Res}(f,g;x) \in I$ and $b \in \mathcal{V}(I)$, this evaluation is 0. By Theorem 2.3.2, $f(x,b)$ and $g(x,b)$ have a nonconstant common factor. As $\mathbb{K}$ is algebraically closed, they have a common root, say a. But then $(a,b) \in \mathcal{V}(f,g)$, and so $b \in \pi(\mathcal{V}(f,g))$.

Now suppose that $f_0(b) \neq 0$ but $g_0(b) = 0$. Since $\langle f, g \rangle = \langle f, g + x^l f \rangle$, if we replace g by $g + x^l f$ where $l + m > n$, then we are in the previous case. □

Observe that if f_0 and g_0 are constants, then we showed that $\mathcal{V}(I) = \mathcal{V}(\mathrm{Res}(f,g;x)$. We record this fact.

Corollary 2.3.11. *If the coefficients of the highest powers of x in f and g do not involve y, then $\mathcal{V}(I) = \mathcal{V}(\mathrm{Res}(f,g;x))$.*

Lemma 2.3.12. *Suppose that $\mathbb{K}$ is algebraically closed. The system of bivariate polynomials*

$$f(x,y) \ = \ g(x,y) \ = \ 0$$

has finitely many solutions in $\mathbb{A}^2$ if and only if f and g have no common nonconstant factor.

Proof. We instead show that $\mathcal{V}(f,g)$ is infinite if and only if f and g do have a common nonconstant factor. If f and g have a common nonconstant factor $h(x,y)$ then their common zeroes $\mathcal{V}(f,g)$ include $\mathcal{V}(h)$ which is infinite as h is nonconstant and $\mathbb{K}$ is algebraically closed.

Now suppose that $\mathcal{V}(f,g)$ is infinite. Then the projection of $\mathcal{V}(f,g)$ to at least one of the two axes is infinite. Suppose that the projection π onto the y-axis is infinite. Set $I := \langle f, g \rangle \cap \mathbb{K}[y]$, the elimination ideal. By the Extension Theorem 2.3.10, we have $\pi(\mathcal{V}(f,g)) \subset \mathcal{V}(I) \subset \mathcal{V}(\mathrm{Res}(f,g;x))$. Since $\pi(\mathcal{V}(f,g))$ is infinite, $\mathcal{V}(\mathrm{Res}(f,g;x)) = \mathbb{A}^1$, which implies that $\mathrm{Res}(f,g;x)$ it is the zero polynomial. By Theorem 2.3.2 and Lemma 2.3.9, f and g have a common nonconstant factor. □

Let $f, g \in \mathbb{K}[x,y]$ and suppose that neither $\mathrm{Res}(f,g;x)$ nor $\mathrm{Res}(f,g;y)$ vanishes so that f and g have no nonconstant common factor. Then $\mathcal{V}(f,g)$ consists of finitely many points. The Extension Theorem gives the following algorithm to compute $\mathcal{V}(f,g)$.

Algorithm 2.3.13 (Elimination Algorithm)**.** INPUT: Polynomials $f, g \in \mathbb{K}[x,y]$.
OUTPUT: $\mathcal{V}(f,g)$.

First, compute the resultant $\mathrm{Res}(f, g; x)$, which is not the zero polynomial. Then, for every root b of $\mathrm{Res}(f, g; x)$, find all common roots a of $f(x, b)$ and $g(x, b)$. The finitely many pairs (a, b) computed are the points of $\mathcal{V}(f, g)$.

The Elimination Algorithm reduces the problem of solving a bivariate system

$$f(x, y) \;=\; g(x, y) \;=\; 0\,, \tag{2.11}$$

to that of finding the roots of univariate polynomials.

Often we only want to count the number of solutions to a system (2.11), or give a realistic bound for this number which is attained when f and g are generic polynomials. The most basic of such bounds was given by Etienne Bézout in his 1779 treatise *Théorie Générale des Équations Algébriques* [10, 11]. Our first step toward establishing Bézout's Theorem is an exercise in algebra and some book-keeping. The monomials in a polynomial of degree n in the variables x, y are indexed by the set

$$n\Delta \;:=\; \{(i, j) \in \mathbb{N}^2 \mid i + j \leq n\}\,.$$

Let $F := \{f_{i,j} \mid (i, j) \in m\Delta\}$ and $G := \{g_{i,j} \mid (i, j) \in n\Delta\}$ be indeterminates and consider generic polynomials f and g of respective degrees m and n in $\mathbb{K}[F, G][x, y]$,

$$f(x, y) \;:=\; \sum_{(i,j)\in m\Delta} f_{i,j} x^i y^j \qquad \text{and} \qquad g(x, y) \;:=\; \sum_{(i,j)\in n\Delta} g_{i,j} x^i y^j\;.$$

Lemma 2.3.14. *The generic resultant* $\mathrm{Res}(f, g; x)$ *is a polynomial in* y *of degree* mn.

Proof. Write

$$f \;:=\; \sum_{j=0}^{m} f_j(y) x^{m-j} \qquad \text{and} \qquad g \;:=\; \sum_{j=0}^{n} g_j(y) x^{n-j}\,,$$

where the coefficients are univariate polynomials in x

$$f_j(y) \;:=\; \sum_{i=0}^{j} f_{i,j} y^i \qquad \text{and} \qquad g_j(y) \;:=\; \sum_{i=0}^{j} g_{i,j} y^i\,.$$

Then the Sylvester matrix $\mathrm{Syl}(f, g; x)$ has the form

$$\mathrm{Syl}(f, g; x) \;:=\; \left(\begin{array}{ccc|ccc}
f_0(y) & & 0 & g_0(y) & & 0 \\
\vdots & \ddots & & \vdots & \ddots & \\
\vdots & & \ddots & \vdots & & g_0(y) \\
f_m(y) & & f_0(y) & g_{n-1}(y) & & \vdots \\
 & \ddots & \vdots & g_n(y) & & \vdots \\
 & & \ddots \;\; \vdots & & \ddots & \vdots \\
0 & & f_m(y) & 0 & & g_n(y)
\end{array}\right)\,,$$

and so the resultant $\mathrm{Res}(f, g; x) = \det(\mathrm{Syl}(f, g; x))$ is a univariate polynomial in y.

As in the proof of Lemma 2.3.3, if we set $f_j := 0$ when $j < 0$ or $j > m$ and $g_j := 0$ when $j < 0$ or $j > n$, then the entry in row i and column j of the Sylvester matrix is

$$\mathrm{Syl}(f, g; x)_{i,j} \;=\; \begin{cases} f_{i-j}(y) & \text{if } j \le n \\ g_{n+i-j}(y) & \text{if } n < j \le m+n \end{cases}$$

The determinant is a signed sum over permutations w of $\{1, \ldots, m+n\}$ of terms

$$\prod_{j=1}^{n} f_{w(j)-j}(y) \;\cdot\; \prod_{j=n+1}^{m+n} g_{n+w(j)-j}(y)\,.$$

This is a polynomial of degree at most

$$\sum_{j=1}^{m} w(j)-j \quad + \quad \sum_{j=n+1}^{m+n} n+w(j)-j \;=\; mn + \sum_{j=1}^{m+n} w(j)-j \;=\; mn\,.$$

Thus $\mathrm{Res}(f, g; x)$ is a polynomial of degree at most mn.

We complete the proof by showing that the resultant does indeed have degree mn. The product $f_0(y)^n \cdot g_n(y)^m$ of the entries along the main diagonal of the Sylvester matrix has constant term $f_{0,0}^n \cdot g_{0,n}^m$. The coefficient of y^{mn} in this product is $f_{0,0}^n \cdot g_{n,n}^m$, and these are the only terms in the expansion of the determinant of the Sylvester matrix involving either of these monomials in the coefficients $f_{i,j}, g_{k,l}$. □

We now state and prove Bézout's Theorem, which bounds the number of points in the variety $\mathcal{V}(f, g)$ in $\mathbb{A}^2$.

Theorem 2.3.15 (Bézout's Theorem)**.** *Two polynomials $f, g \in \mathbb{K}[x, y]$ either have a common factor or else $|\mathcal{V}(f, g)| \le \deg f \cdot \deg g$.*

When $|\mathbb{K}|$ is at least $\max\{\deg f, \deg g\}$, this inequality is sharp in that the bound is attained. When $\mathbb{K}$ is algebraically closed, the bound is attained when f and g are general polynomials of the given degrees.

Proof. Suppose that $m := \deg f$ and $n = \deg g$. By Lemma 2.3.12, if f and g are relatively prime, then $\mathcal{V}(f, g)$ is finite. Let us extend our field to its algebraic closure $\overline{\mathbb{K}}$, which in infinite. If we change coordinates, replacing f by $f(A(x, y))$ and g by $g(A(x, y))$, where A is an invertible affine transformation,

$$A(x, y) \;=\; (ax + by + c,\; \alpha x + \beta y + \gamma)\,, \tag{2.12}$$

with $a, b, c, \alpha, \beta, \gamma \in \overline{\mathbb{K}}$ with $a\beta - \alpha b \neq 0$. We can choose these parameters so that both f and g have non-vanishing constant terms and nonzero coefficients of their highest powers (m and n, respectively) of y. By Lemma 2.3.14, this implies that the resultant $\mathrm{Res}(f, g; x)$

has degree at most mn and thus at most mn zeroes. If we set $I := \langle f, g\rangle \cap \overline{\mathbb{K}}[y]$, then this also implies that $\mathcal{V}(I) = \mathcal{V}(\mathrm{Res}(f, g; x))$, by Corollary 2.3.11.

We can furthermore choose the parameters in A so that the projection $\pi\colon (x, y) \mapsto y$ is 1-1 on $\mathcal{V}(f, g)$, as $\mathcal{V}(f, g)$ is a finite set. Thus

$$\pi(\mathcal{V}(f, g)) \;=\; \mathcal{V}(I) \;=\; \mathcal{V}(\mathrm{Res}(f, g; x))\,,$$

which implies the inequality of the theorem as $|\mathcal{V}(\mathrm{Res}(f, g; x))| \leq mn$.

To see that the bound is sharp when $|\mathbb{K}|$ is large enough, let $a_1, \ldots, a_m$ and $b_1, \ldots, b_n$ be distinct elements of $\mathbb{K}$. Note that the system

$$f \;:=\; \prod_{i=1}^{m}(x - a_i) \;=\; 0 \qquad \text{and} \qquad g \;:=\; \prod_{i=1}^{n}(y - b_i) \;=\; 0$$

has mn solutions $\{(a_i, b_j) \mid 1 \leq i \leq m, 1 \leq j \leq n\}$, so the inequality is sharp.

Suppose now that $\mathbb{K}$ is algebraically closed. If the resultant $\mathrm{Res}(f, g; x)$ has fewer than mn distinct roots, then either it has degree strictly less than mn or else it has a multiple root. In the first case, its leading coefficient vanishes and in the second case, its discriminant vanishes. But the leading coefficient and the discriminant of $\mathrm{Res}(f, g; x)$ are polynomials in the coefficients of f and g. Thus the set of pairs of polynomials (f, g) with $\mathcal{V}(f, g)$ consisting of mn points in $\mathbb{A}^2$ forms a nonempty open subset in the space $\mathbb{A}^{\binom{m+2}{2}+\binom{n+2}{2}}$ of pairs of polynomials (f, g) with $\deg f = m$ and $\deg g = n$. □

Exercises for Section 3

1. Using the formula (2.6) deduce the Poisson formula for the resultant of univariate polynomials f and g,

$$\mathrm{Res}(f, g; x) \;=\; (-1)^{mn} f_0^n \prod_{i=1}^{m} g(\alpha_i)\,,$$

   where $\alpha_1, \ldots, \alpha_m$ are the roots of f. where $\alpha_1, \ldots, \alpha_n$ are the roots of f.

2. Suppose that the polynomial $g = g_1 \cdot g_2$ factorizes. Show that the resultant also factorizes, $\mathrm{Res}(f, g; x) = \mathrm{Res}(f, g_1; x) \cdot \mathrm{Res}(f, g_2; x)$.

3. Compute the Bezoutian matrix when $n = 4$. Give a general formula for the entries of the Bezoutian matrix.

4. Compute the constant in the proof of Theorem 2.3.5, by computing the resultant and Bezoutian polynomials when $f(x) := x^m$ and $g(x) = x^n + 1$.

5. Write out the 5×5 matrix used to compute the discriminant of a general cubic $x^3 + ax^2 + bx + c$ and take its determinant to show that the discriminant is

$$27d^2 - 18b^2d - 18acd + 9bc^2 + 30a^2bd - 12ab^2c + 3b^4 - 8a^4d + 4a^3bc - a^2b^3 \,.$$

6. Show that the discriminant of a polynomial f of degree n may also be expressed as

$$\prod_{i \neq j} (\alpha_i - \alpha_j)^2 \,,$$

where $\alpha_1, \ldots, \alpha_n$ are the roots of f.

7. Prove the adjoint formula for the inverse of a matrix A, $\det(A) \cdot A^{-1} = \mathrm{ad}(A)$.

8. Let $f(x, y)$ be a polynomial of total degree n. Show that there is a non-empty Zariski open subset of parameters $(a, b, c, \alpha, \beta, \gamma) \in \mathbb{A}^6$ with $a\beta - \alpha b \neq 0$ such that if A is the affine transformation (2.12), then every monomial $x^i y^j$ with $0 \leq i, j$ and $i + j \leq n$ appears in the polynomial $f(A(x, y))$ with a non-zero coefficient.

9. Use Lemma 2.3.12 to show that $\mathbb{A}^2$ has dimension 2, in the sense of the combinatorial definition of dimension (3.2).

10. Use Lemma 2.3.12 and induction on the number of polynomials defining a proper subvariety X of $\mathbb{A}^2$ to show that X consists of finitely many irreducible curves and finitely many isolated points.

2.4 Solving equations with Gröbner bases

Algorithm 2.3.13 used resultants to reduce the problem of solving two equations in two variables to that of solving univariate polynomials. A modification of this algorithm can be used to find all roots of a zero-dimensional ideal $I \subset \mathbb{K}[x_1, \ldots, x_n]$, if we can compute elimination ideals $I \cap \mathbb{K}[x_i, x_{i+1}, \ldots, x_n]$. This may be accomplished using Gröbner bases and more generally, ideas from the theory of Gröbner bases can help us to understand solutions to systems of equations.

Suppose that we have N polynomial equations in n variables $(x_1, \ldots, x_n)$

$$f_1(x_1, \ldots, x_n) \;=\; \cdots \;=\; f_N(x_1, \ldots, x_n) \;=\; 0\,, \tag{2.13}$$

and we want to understand the solutions or *roots* to this system. By understand, we mean answering (any of) the following questions.

(i) Does (2.13) have finitely many solutions?

(ii) If not, can we understand the isolated solutions of (2.13)?

(iii) Can we count them, or give (good) upper bounds on their number?

(iv) Can we *solve* the system (2.13) and find all complex solutions?

(v) When the polynomials have real coefficients, can we count (or bound) the number of real solutions to (2.13)? Or simply find them?

We describe symbolic algorithms based upon Gröbner bases that begin to address these questions.

The solutions to (2.13) in $\mathbb{A}^n$ constitute the affine variety $\mathcal{V}(I)$, where I is the ideal generated by the polynomials $f_1, \ldots, f_N$. Symbolic algorithms to address Questions (i)-(v) involve studying the ideal I. An ideal I is *zero-dimensional* if, over the algebraic closure of $\mathbb{K}$, $\mathcal{V}(I)$ is finite, that is, if the dimension of $\mathcal{V}(I)$ is zero. Thus I is zero-dimensional if and only if the radical $\sqrt{I}$ of I is zero-dimensional.

Theorem 2.4.1. *Let $I \subset \mathbb{K}[x_1, \ldots, x_n]$ be an ideal. Then I is zero-dimensional if and only if $\mathbb{K}[x_1, \ldots, x_n]/I$ is a finite-dimensional $\mathbb{K}$-vector space.*

Proof. We may assume the $\mathbb{K}$ is algebraically closed, as this does not change the dimension of quotient rings.

Suppose first that I is radical. Then $\mathbb{K}[x_1, \ldots, x_n]/I$ is the coordinate ring $\mathbb{K}[X]$ of $X := \mathcal{V}(I)$, and therefore consists of all functions obtained by restricting polynomials to $\mathcal{V}(I)$. If X is finite, then $\mathbb{K}[X]$ is finite-dimensional as the space of functions on X has dimension equal to the number of points in X. Suppose that X is infinite. Then there is some coordinate, say x_1, such that the projection of X to the x_1-axis is infinite and therefore dense. Restriction of polynomials in x_1 to X is an injective map from $\mathbb{K}[x_1]$ to $\mathbb{K}[X]$ which shows that $\mathbb{K}[X]$ is infinite-dimensional.

Now suppose that I is any ideal. If $\mathbb{K}[x_1,\ldots,x_n]/I$ is finite-dimensional, then so is $\mathbb{K}[x_1,\ldots,x_n]/\sqrt{I}$ as $I \subset \sqrt{I}$. For the other direction, we suppose that $\mathbb{K}[x_1,\ldots,x_n]/\sqrt{I}$ is finite-dimensional. For each variable x_i, there is some linear combination of $1, x_i, x_i^2, \ldots$ which is zero in $\mathbb{K}[x_1,\ldots,x_n]/\sqrt{I}$ and hence lies in $\sqrt{I}$. But this is a univariate polynomial $g_i(x_i) \in \sqrt{I}$, so there is some power $g_i(x_i)^{M_i}$ of g which lies in I. But then we have $\langle g_1(x_1)^{M_1},\ldots,g_n(x_n)^{M_n}\rangle \subset I$, and so the map

$$\mathbb{K}[x_1,\ldots,x_n]/\langle g_1(x_1)^{M_1},\ldots,g_n(x_n)^{M_n}\rangle \quad \longrightarrow \quad \mathbb{K}[x_1,\ldots,x_n]/I$$

is a surjection. But $\mathbb{K}[x_1,\ldots,x_n]/\langle g_1(x_1)^{M_1},\ldots,g_n(x_n)^{M_n}\rangle$ has dimension $M_1M_2\cdots M_n$, which implies that $\mathbb{K}[x_1,\ldots,x_n]/I$ is finite-dimensional. □

A consequence of the proof is the following criterion for an ideal to be zero-dimensional.

Corollary 2.4.2. *An ideal $I \subset \mathbb{K}[x_1,\ldots,x_n]$ is zero-dimensional if and only if for every variable x_i, there is a univariate polynomial $g_i(x_i)$ which lies in I.*

Together with Macaulay's Theorem 2.2.3, Theorem 2.4.1 leads to a Gröbner basis criterion/algorithm to solve Question (i).

Corollary 2.4.3. *An ideal $I \subset \mathbb{K}[x_1,\ldots,x_n]$ is zero-dimensional if and only if for any monomial order $\succ$, the initial ideal $\mathrm{in}_\succ I$ of I contains some power of every variable.*

Thus we can determine if I is zero-dimensional and thereby answer Question (i) by computing a Gröbner basis for I and checking that the leading terms of elements of the Gröbner basis include pure powers of all variables.

When I is zero-dimensional, its *degree* is the dimension of $\mathbb{K}[x_1,\ldots,x_n]/I$ as a $\mathbb{K}$-vector space, which is the number of standard monomials, by Macaulay's Theorem 2.2.3. A Gröbner basis for I gives generators of the initial ideal which we can use to count the number of standard monomials to determine the degree of an ideal.

Suppose that the ideal I generated by the polynomials f_i of (2.13) is not zero-dimensional, and we still want to count the isolated solutions to (2.13). In this case, there are symbolic algorithms that compute a zero-dimensional ideal J with $J \supset I$ having the property that $\mathcal{V}(J)$ consists of all isolated points in $\mathcal{V}(I)$, that is all isolated solutions to (2.13). These algorithms successively compute the ideal of all components of $\mathcal{V}(I)$ of maximal dimension, and then strip them off. One such method would be to compute the primary decomposition of an ideal. Another method, when the non-isolated solutions are known to lie on a variety $\mathcal{V}(J)$, is to saturate I by J to remove the excess intersection.[†]

When I is a zero-dimensional radical ideal and $\mathbb{K}$ is algebraically closed, the degree of I equals the number of points in $\mathcal{V}(I) \subset \mathbb{A}^n$ (see Exercise 1) and thus we obtain an answer to Question (iii).

[†]Develop this further, either here or somewhere else, and then refer to that place.

Theorem 2.4.4. *Let I be the ideal generated by the polynomials f_i of (2.13). If I is zero-dimensional, then the number of solutions to the system (2.13) is bounded by the degree of I. When $\mathbb{K}$ is algebraically closed, the number of solutions is equal to this degree if and only if I is radical.*

In many important cases, there are sharp upper bounds for the number of isolated solutions to the system (2.13) which do not require a Gröbner basis. For example, Theorem 2.3.15 (Bézout's Theorem in the plane) gave such bounds when $N = n = 2$. Suppose that $N = n$ so that the number of equations equals the number of variables. This is called a *square system*. Bézout's Theorem in the plane has a natural extension in this case, which we will prove in Section 3.6. A common zero a to a square system of equations is *nondegenerate* if the differentials of the equations are linearly independent at a.

Theorem 2.4.5 (Bézout's Theorem)**.** *Given polynomials $f_1, \ldots, f_n \in \mathbb{K}[x_1, \ldots, x_n]$ with $d_i = \deg(f_i)$, the number of nondegenerate solutions to the system*

$$f_1(x_1, \ldots, x_n) \;=\; \cdots \;=\; f_n(x_1, \ldots, x_n) \;=\; 0$$

in $\mathbb{A}^n$ is at most $d_1 \cdots d_n$. When $\mathbb{K}$ is algebraically closed, this is a bound for the number of isolated solutions, and it is attained for generic choices of the polynomials f_i.

This product of degrees $d_1 \cdots d_n$ is called the *Bézout bound* for such a system. While this bound is sharp for generic square systems, few practical problems involve generic systems and other bounds are often needed (see Exercise 2).

We discuss a symbolic method to solve systems of polynomial equations (2.13) based upon elimination theory and the Shape Lemma, which describes the form of a Gröbner basis of a zero-dimensional ideal I with respect to a lexicographic monomial order.

Let $I \subset \mathbb{K}[x_1, \ldots, x_n]$ be an ideal. A univariate polynomial $g(x_i)$ is an *eliminant for I* if g generates the elimination ideal $I \cap \mathbb{K}[x_i]$.

Theorem 2.4.6. *Suppose that $g(x_i)$ is an eliminant for an ideal $I \subset \mathbb{K}[x_1, \ldots, x_n]$. Then $g(a_i) = 0$ for every $a = (a_1, \ldots, a_n) \in \mathcal{V}(I) \in \mathbb{A}^n$. When $\mathbb{K}$ is algebraically closed, every root of g occurs in this way.*

Proof. We have $g(a_i) = 0$ as this is the value of g at the point a. Suppose that $\mathbb{K}$ is algebraically closed and that ξ is a root of $g(x_i)$ but there is no point $a \in \mathcal{V}(I)$ whose ith coordinate is ξ. Let $h(x_i)$ be a polynomial whose roots are the other roots of g. Then h vanishes on $\mathcal{V}(I)$ and so $h \in \sqrt{I}$ and so some power, h^N, of h lies in I. Thus $h^N \in I \cap \mathbb{K}[x_i] = \langle g \rangle$. But this is a contradiction as $h(\xi) \neq 0$ while $g(\xi) = 0$. □

Theorem 2.4.7. *If $g(x_i)$ is a monic eliminant for an ideal $I \subset \mathbb{K}[x_1, \ldots, x_n]$, then g lies in any reduced Gröbner basis for I with respect to a lexicographic order in which x_i is the minimal variable.*

Proof. Suppose that $\succ$ is a lexicographic monomial order with x_i the minimal variable. Since g generates the elimination ideal $I \cap \mathbb{K}[x_i]$, it is the lowest degree monic polynomial in x_i lying in I. In particular $x_i^{\deg(g)}$ is a generator of the initial ideal of I. Therefore there is a polynomial f in the reduced Gröbner basis whose leading term is $x_i^{\deg(g)}$ and whose remaining terms involve smaller standard monomials. As x_i is the minimal variable and this is a lexicographic monomial order, the only smaller standard monomials are x^d with $d < \deg(g)$. Thus $f \in I$ is a monic polynomial in x_i with degree $\deg(g)$. By the uniqueness of g, $f = g$, which proves that g lies in the reduced Gröbner basis. □

Theorem 2.4.7 gives an algorithm to compute eliminants—simply compute a lexicographic Gröbner basis. This is not recommended, as lexicographic Gröbner bases appear to be the most expensive to compute in practice. We instead offer the following algorithm.

Algorithm 2.4.8. INPUT: Ideal $I \subset \mathbb{K}[x_1, \ldots, x_n]$ and a variable x_i.
OUTPUT: Either a univariate eliminant $g(x_i) \in I$ or else a certificate that one does not exist.

(1) Compute a Gröbner basis G for I with respect to any term order.

(2) If no initial term of any element of G is a pure power of x_i, then halt and declare that I does not contain a univariate eliminant for x_i.

(3) Otherwise, compute the sequence $1 \mod G$, $x_i \mod G$, $x_i^2 \mod G$, ..., until a linear dependence is found,

$$\sum_{j=0}^{m} a_j (x_i^j \mod G) = 0\,, \tag{2.14}$$

where m is minimal. Then

$$g(x_i) = \sum_{j=0}^{m} a_j x_i^j$$

is the univariate eliminant.

Proof of correctness. If I does not have an eliminant in x_i, then $I \cap \mathbb{K}[x_i] = \{0\}$, so $1, x_i, x_i^2, \ldots$ are all standard, and no Gröbner basis contains a polynomial with initial monomial a pure power of x_i. This shows that the algorithm correctly identifies when no eliminant exists.

Suppose now that I does have an eliminant $g(x_i)$. Since $g \mod G = 0$, the Gröbner basis G must contain a polynomial whose initial monomial divides that of g and is hence a pure power of x_i. If $g = \sum b_j x_i^j$ and has degree N, then

$$0 = g \mod G = \Big(\sum_{j=0}^{N} b_j x_i^j\Big) \mod G = \sum_{j=0}^{N} b_j (x_i^j \mod G)\,,$$

which is a linear dependence among the elements of the sequence $1 \mod G$, $x_i \mod G$, $x_i^2 \mod G$, ... Thus the algorithm halts. The minimality of the degree of g implies that $N = m$ and the uniqueness of such minimal linear combinations implies that the coefficients b_j and a_j are proportional, which shows that the algorithm computes a scalar multiple of g, which is also an eliminant. □

Elimination using Gröbner bases leads to an algorithm for Question (v). The first step is to understand the optimal form of a Gröbner basis of a zero-dimensional ideal.

Lemma 2.4.9 (Shape Lemma)**.** *Suppose g is an eliminant of a zero-dimensional ideal I with $\deg(g) = \deg(I)$. Then I is radical if and only if g has no multiple factors.*

If we further have $g = g(x_n)$, then in the lexicographic term order with $x_1 \succ x_2 \succ \cdots \succ x_n$, the ideal I has a Gröbner basis of the form:

$$x_1 - g_1(x_n), \quad x_2 - g_2(x_n), \quad \ldots, \quad x_{n-1} - g_{n-1}(x_n), \quad g(x_n)\,, \tag{2.15}$$

where $\deg(g) > \deg(g_i)$ for $i = 1, \ldots, n-1$.

If I is generated by real polynomials, then the number of real roots of I equals the number of real roots of g.

Proof. We have

$$\#\text{roots of } g \;\leq\; \#\text{roots of } I \;\leq\; \deg(I) \;=\; \deg(g)\,,$$

the first inequality by Theorem 2.4.6 and the second by Theorem 2.4.4. If the roots of g are distinct, then their number is $\deg(g)$ and so these inequalities are equalities. This implies that I is radical, by Theorem 2.4.4. Conversely, if $g = g(x_i)$ has multiple roots, then there is a polynomial h with the same roots as g but with smaller degree. Since $\langle g \rangle = I \cap \mathbb{K}[x_i]$, we have that $h \notin I$, but since $h^{\deg(g)}$ is divisible by g, $h^{\deg(g)} \in I$, so I is not radical.

To prove the second statement, let d be the degree of the eliminant $g(x_n)$. Then each of $1, x_n, \ldots, x_n^{d-1}$ is a standard monomial, and since $\deg(g) = \deg(I)$, there are no others. Thus the initial ideal in the lexicographic monomial order is $\langle x_1, \ldots, x_{n-1}, x_n^d \rangle$. Each element of the reduced Gröbner basis for I expresses a generator of the initial ideal as a $\mathbb{K}$-linear combination of standard monomials. It follows that the reduced Gröbner basis has the form claimed.

For the last statement, observe that the common zeroes of the polynomials (2.15) are

$$\{(a_1, \ldots, a_n) \mid g(a_n) = 0 \text{ and } a_i = g_i(a_n)\,,\ i = 1, \ldots, n-1\}\,.$$

By Corollary 2.2.7, the polynomials g_i are all real, and so a component a_i is real if the root a_n of $g(x_n)$ is real. □

Not all ideals can have such a Gröbner basis. For example, the ideal

$$\langle x, y \rangle^2 \;=\; \langle x^2, xy, y^2 \rangle$$

cannot. Nevertheless, the key condition on the eliminant g, that $\deg(g) = \deg(I)$, often holds after a generic change of coordinates, just as in the proof of Bézout's Theorem in the plane (Theorem 2.3.15). This gives the following symbolic random algorithm to count the number of real solutions to a system of equations.

Algorithm 2.4.10 (Counting real roots)**.** Given a system of polynomial equations (2.13), let I be the ideal generated by the polynomials. If I is zero-dimensional, then compute an eliminant $g(x_i)$ and check if $\deg(g) = \deg(I)$, if not, then perform a generic change of variables and compute a different eliminant, possibly after a linear change of variables.

Given such a eliminant g satisfying the hypotheses of the Shape Lemma, use Sturm sequences or any other method to determine its number of real solutions, which is the number of real solutions to the original system.

While this algorithm will not necessarily halt (for example, if $I = \langle x^2, xy, y^2\rangle$), it will halt with the correct answer when $\mathcal{V}(I)$ is reduced and zero-dimensional. This simple algorithm is the idea behind the efficient algorithm REALSOLVING of Faugère and Roullier.

While the Shape Lemma describes an optimal form of a Gröbner basis for a zero-dimensional ideal, it is typically not optimal to compute such a Gröbner basis directly. An alternative to direct computation of a lexicographic Gröbner basis is the *FGLM algorithm* of Faugère, Gianni, Lazard, and Mora [28], which is an algorithm for Gröbner basis conversion. That is, given a Gröbner basis for a zero dimensional ideal with respect to one monomial order and a different monomial order $\succ$, FGLM computes a Gröbner basis for the ideal with respect to $\succ$.

Algorithm 2.4.11 (FGLM)**.** Input: A Gröbner basis G for a zero-dimensional ideal $I \subset \mathbb{K}[x_1, \ldots, x_n]$ with respect to a monomial order $\geq$, and a different monomial order $\succ$.
Output: A Gröbner basis H for I with respect to $\succ$.
Initialize: Set $H := \{\}$, $x^\alpha := 1$, and $S := \{\}$.

(1) Compute $\overline{x^\alpha} := x^\alpha \mod G$.

(2) If $\overline{x^\alpha}$ does not lie in the linear span of S, then set $S := S \cup \{\overline{x^\alpha}\}$.

Otherwise, there is a (unique) linear combination of elements of S such that

$$\overline{x^\alpha} \;=\; \sum_{\overline{x^\beta} \in S} c_\beta \overline{x^\beta}\,.$$

Set $H := H \cup \{x^\alpha - \sum_\beta c_\beta x^\beta\}$.

(3) If

$$\{x^\gamma \mid x^\gamma \succ x^\alpha\} \;\subset\; \mathrm{in}_\succ(H) := \langle \mathrm{in}_\succ(h) \mid h \in H\rangle\,,$$

then halt and output H. Otherwise, set x^α to be the $\succ$-minimal monomial in the set $\{x^\gamma \notin \mathrm{in}_\succ(H) \mid x^\gamma \succ x^\alpha\}$ and return to (1).

Proof of correctness. By construction, H always consists of elements of I, and elements of S are linearly independent in the quotient ring $F[x_1, \ldots, x_n]/I$. Thus $\mathrm{in}_\succ(H)$ is a subset of the initial ideal $\mathrm{in}_\succ I$, and we always have the inequalities

$$|S| \;\leq\; \dim_{\mathbb{K}}(\mathbb{K}[x_1, \ldots, x_n]/I) \qquad \text{and} \qquad \mathrm{in}_\succ(H) \;\subset\; \mathrm{in}_\succ I\,.$$

Every time we return to (1) either the set S or the set H (and also $\mathrm{in}_\succ(H)$) increases. Since the cardinality of S is bounded and the monomial ideals generated by the different $\mathrm{in}_\succ(H)$ form a strictly increasing chain, the algorithm must halt.

When the algorithm halts, every monomial is either in the set SM $:= \{\mathrm{x}^\beta \mid \overline{\mathrm{x}^\beta} \in \mathrm{S}\}$ or else in the monomial ideal $\mathrm{in}_\succ(H)$. By our choice of x^α in (3), these two sets are disjoint, so that SM is the set of standard monomials for $\mathrm{in}_\succ(H)$. Since $\mathrm{in}_\succ(H) \subset \mathrm{in}_\succ\langle H\rangle \subset \mathrm{in}_\succ I$, and elements of S are linearly independent modulo $\mathrm{in}_\succ I$, we have

$$|S| \;\leq\; \dim_{\mathbb{K}}(\mathbb{K}[x]/\mathrm{in}_\succ I) \;\leq\; \dim_{\mathbb{K}}(\mathbb{K}[x]/\mathrm{in}_\succ\langle H\rangle) \;\leq\; \dim_{\mathbb{K}}(\mathbb{K}[x]/\mathrm{in}_\succ(H)) \;=\; |S|\,.$$

Thus $\mathrm{in}_\succ I = \mathrm{in}_\succ(H)$, which proves that H is a Gröbner basis for I with respect to the monomial order $\succ$. By the form of the elements of H, it is the reduced Gröbner basis. □

Exercises for Section 4

1. Suppose $I \subset \mathbb{K}[x_1, \ldots, x_n]$ is radical, $\mathbb{K}$ is algebraically closed, and $\mathcal{V}(I) \subset \mathbb{A}^n$ consists of finitely many points. Show that the coordinate ring $\mathbb{K}[x_1, \ldots, x_n]/I$ of restrictions of polynomial functions to $\mathcal{V}(I)$ has dimension as a $\mathbb{K}$-vector space equal to the number of points in $\mathcal{V}(I)$.

2. Compute the number of solutions to the system of polynomials

$$1 + 2x + 3y + 5xy \;=\; 7 + 11xy + 13xy^2 + 17x^2y \;=\; 0\,.$$

   Show that each is nondegenerate and compare this to the Bézout bound for this system. How many solutions are real?

3. In this and subsequent exercises, you are asked to use computer experimentation to study the number of solutions to certain structured polynomial systems. This is a good opportunity to become acquainted with symbolic software.

   For several small values of n and d, generate n random polynomials in n variables of degree d, and compute their numbers of isolated solutions. Does your answer agree with Bézout's Theorem?

4. A polynomial is *multilinear* if no variable occurs to a power greater than 1. For example,

$$3xyz - 7xy + 13xz - 19yz + 29x - 37y + 43z - 53$$

is a multilinear polynomial in the indeterminate x, y, z. For several small values of n generate n random multilinear polynomials and compute their numbers of common zeroes, Does your answer agree with Bézout's Theorem?

5. Let $\mathcal{A} \subset \mathbb{N}^n$ be a finite set of integer vectors, which we regard as exponents of monomials in $\mathbb{K}[x_1, \ldots, x_n]$. A polynomial with support $\mathcal{A}$ is a linear combination of monomials whose exponents are from $\mathcal{A}$. For example

$$1 + 3x + 9x^2 + 27y + 81xy + 243xy^2$$

is a polynomial with support $\mathcal{A} = \begin{pmatrix} 0 & 1 & 2 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 2 \end{pmatrix}$.

For $n = 2, 3$ and many $\mathcal{A}$ with $|\mathcal{A}| > n$ and $0 \in \mathcal{A}$, generate random systems of polynomials with support $\mathcal{A}$ and determine their numbers of isolated solutions. Formulate a precise conjecture about this number of solutions as a function of $\mathcal{A}$.

6. Fix $m, p \geq 2$. For $\alpha \colon 1 \leq \alpha_1 < \cdots < \alpha_p \leq m+p$, let E_α be a $p \times (m+p)$ matrix whose entries in the columns indexed by α form the identity matrix, and the entries in position i, j are either variables if $j < \alpha_i$ or 0 if $\alpha_i < j$. For example, when $m = p = 3$, here are E_{245} and E_{356},

$$E_{245} = \begin{pmatrix} a & 1 & 0 & 0 & 0 & 0 \\ b & 0 & c & 1 & 0 & 0 \\ d & 0 & e & 0 & 1 & 0 \end{pmatrix} \qquad E_{356} = \begin{pmatrix} a & b & 1 & 0 & 0 & 0 \\ c & d & 0 & e & 1 & 0 \\ f & g & 0 & h & 0 & 1 \end{pmatrix} .$$

Set $|\alpha| := \alpha_1 - 1 + \alpha_2 - 2 + \cdots + \alpha_p - p$ be the number of variables in E_α. For all small m, p, and α, generate $|\alpha|$ random $m \times (m+p)$ matrices $M_1, \ldots, M_{|\alpha|}$ and determine the number of isolated solutions to the system of equations

$$\det \begin{pmatrix} E_\alpha \\ M_1 \end{pmatrix} = \det \begin{pmatrix} E_\alpha \\ M_2 \end{pmatrix} = \cdots = \det \begin{pmatrix} E_\alpha \\ M_{|\alpha|} \end{pmatrix} = 0 .$$

Try to formulate a conjecture for the number of solutions as a function of m, p, and α.

2.5 Eigenvalue techniques

In this section we discuss a central bridge from the solutions of polynomial systems to eigenvalue methods of linear algebra and analytic geometry. This connection will provide further methods to compute and analyze the roots of a zero-dimensional ideal. The techniques are based on very classical results, but their computational aspects have only been developed systematically within the last 20 years.

Let $\mathbb{K}$ be algebraically closed, $f_1, \ldots, f_N \in \mathbb{K}[x_1, \ldots, x_n]$, and $I := \langle f_1, \ldots, f_N \rangle$ be zero-dimensional. Our goal is to interpret the points in $\mathcal{V}(I)$ as eigenvalues of suitable matrices. For numerically determining the eigenvalues of a complex matrix, there are well-investigated numerical methods.

It is instructive to start with the univariate case. Given a monic, univariate polynomial $p = \sum_{i=0}^{d} c_i x^i \in \mathbb{K}[x]$, the matrix

$$C_p \;=\; \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -c_0 & -c_1 & -c_2 & \ldots & -c_{n-1} \end{pmatrix} \;\in\; \mathbb{K}^{d \times d}\,.$$

is called the *companion matrix* of p.

For any given matrix $A \in \mathbb{K}^{n \times n}$, the eigenvalues of A are the roots of the characteristic polynomial $\chi_A(x) \;=\; \det(A - xI)$. Thus the following statement tells us that the roots of p coincide with the eigenvalues of the companion matrix C_p.

Theorem 2.5.1. *For any monic, univariate polynomial $p = \sum_{i=0}^{d} c_i x^i \in \mathbb{K}[x]$ of degree $d \geq 1$, the characteristic polynomial of its companion matrix C_p is*

$$\det(C_p - xI) \;=\; (-1)^d p(x)\,.$$

Proof. For $d = 1$, the statement is clear, and for $d > 1$ an expansion by the first column yields

$$\det(C_p - xI) \;=\; (-x)\det(C_q - xI) + (-1)^{d+1}(-c_0)\,,$$

where C_q is the companion matrix of the polynomial $q = x^{d-1} + \sum_{i=0}^{d-2} c_{i+1} x^i$. Inductively, we obtain $\det(C_p - xI) \;=\; (-1)^d x q(x) + (-1)^d c_0 = (-1)^d p(x)$. □

For the multivariate case, let $I := \langle f_1, \ldots, f_N \rangle$ be a zero-dimensional ideal in $R := \mathbb{K}[x_1, \ldots, x_n]$. By Theorems 2.4.1 and 2.4.4, the $\mathbb{K}$-vector space R/I is finite-dimensional, and the cardinality of the variety $\mathcal{V}(I)$ is bounded from above by the dimension R/I. Denote the residue class of a polynomial f in R/I by $[f]$.

For any $i \in \{1, \ldots, n\}$, multiplication of an element in R/I with the residue class $[x_i]$ of a variable x_i defines an endomorphism m_i,

$$\begin{aligned} m_i \;:\; R/I \;&\to\; R/I\,, \\ [f] \;&\mapsto\; [x_i]\cdot[f] \;=\; [x_i f]\,. \end{aligned}$$

Since the vector space R/I is finite-dimensional, for a fixed basis of R/I the linear mapping m_i can be expressed in terms of a representation matrix. For algorithmic purposes the basis of the standard monomials is particularly suited. Let $\mathcal{B}$ denote the set of standard monomials of an ideal I, and let $M_1, \ldots, M_n \in \mathbb{R}^{|\mathcal{B}|\times|\mathcal{B}|}$ be the representation matrices of the endomorphisms $m_1, \ldots, m_n$ with respect to the basis $\mathcal{B}$. M_i is called the *i-th companion matrix* of the ideal I with respect to $\mathcal{B}$. The rows and the columns of the representation matrix M_i are indexed with the monomials in $\mathcal{B}$. For $x^\alpha, x^\beta \in \mathcal{B}$, the entry of M_i in row x^α and column x^β is the coefficient of x^α in the normal form of the polynomial $x^\alpha \cdot x^\beta$ modulo I.

Lemma 2.5.2. *The companion matrices commute pairwise, i.e.,*

$$M_i \cdot M_j \;=\; M_j \cdot M_i \quad \textit{for } 1 \le i < j \le n\,.$$

Proof. The matrices M_iM_j and M_jM_i are the representation matrices of the compositions $m_i \circ m_j$ and $m_j \circ m_i$, respectively. Since multiplication in R/I is commutative, the claim follows. □

With regard to algorithms, our main goal is to deduce the subsequent characterization for the components of the zeroes of an ideal I.

Theorem 2.5.3 (Stickelberger's Theorem)**.** *Let $\mathbb{K}$ be algebraically closed and $I \subset R$ be a zero-dimensional ideal. For any $i \in \{1, \ldots, n\}$ and any $\lambda \in \mathbb{C}$, the value λ is an eigenvalue of the endomorphism m_i if and only if there exists a point $a \in \mathcal{V}(I)$ with $a_i = \lambda$.*

To prove the theorem, we recall some facts from linear algebra known in connection with the Theorem of Cayley-Hamilton.

Definition 2.5.4. Let V be a vector space over $\mathbb{K}$ and f be an endomorphism on V. For a polynomial $p = \sum_{i=0}^{d} c_i t^i \in \mathbb{K}[t]$, the polynomial $p(f)$ is defined by $p(f) \;=\; \sum_{i=0}^{d} c_i f^i$, where f^i denotes the i-times application of the endomorphism f. The ideal $I_f \;=\; \{p \in \mathbb{K}[t] \,:\, p(f) = 0\}$ is called the *ideal of f*. The uniquely determined monique polynomial h with $\langle h \rangle = I_f$ is called the *minimal polynomial of f* and is denoted by h_f.

The eigenvalues and the minimal polynomial of an endomorphism are related as follows.

Lemma 2.5.5. *Let V be a finite-dimensional vector space over an algebraically closed field $\mathbb{K}$ and f be an endomorphism on V. Then for each $\lambda \in \mathbb{K}$, the value λ is an eigenvalue of f if and only if λ is a zero of the minimal polynomial h_f.*

Proof. We show that the minimal polynomial h_f divides the characteristic polynomial χ_f and that χ_f divides h_f^k, where $k := \dim V$. This readily implies the lemma.

The first statement follows from the Theorem of Cayley-Hamilton, which says that each endomorphism is a zero of its characteristic polynomial. For the second statement

we first note that χ_f and h_f decompose over $\mathbb{K}$ in linear factors. Let A_f be a representation matrix of the endomorphism f, and

$$\chi_f \;=\; \det(A_f - tI_n) \;=\; \pm(t-\lambda_1)^{d_1}\cdots(t-\lambda_m)^{d_m}$$

with $\lambda_1,\ldots,\lambda_m \in \mathbb{K}$ and $d_1,\ldots,d_m \in \mathbb{N}$. Since we have already seen that h_f divides χ_f, the minimal polynomial must be of the form $h_f = \prod_{i=1}^m (t-\lambda_i)^{e_i}$. with $0 \le e_i \le d_i$. Now it suffices to show that $e_i \ge 1$ for all $i \in \{1,\ldots,m\}$. Assume that $e_i = 0$ for some i, and let v be an eigenvector to λ_i. Then for each eigenvalue $\lambda_j \neq \lambda_i$ we have

$$(A_f - \lambda_j I_n)v \;=\; (\lambda_i - \lambda_j)v \;\neq\; 0\,,$$

and hence for the application of the matrix $h_f(A_f)$ on the vector v

$$h_f(A_f)v \;=\; \prod_{j\neq i}\Big((A_f-\lambda_j I_n)^{e_j}\Big)v \;=\; \prod_{j\neq i}\Big((\lambda_i-\lambda_j)^{e_j}\Big)v \;\neq\; 0\,.$$

This contradicts the property that h_f is a minimal polynomial of f. □

With these tools we can prove the eigenvalue characterization in Theorem 2.5.3.

Proof of Theorem 2.5.3. Let λ be an eigenvalue of the endomorphism m_i on R/I and $[v]$ be an eigenvector to the eigenvalue λ. That is, $[x_i \cdot v] = [\lambda \cdot v]$ and hence $[(x_i - \lambda)\cdot v] = 0$ in the vector space R/I. We now assume that the second property of the theorem does not hold, i.e., that for $a \in \mathcal{V}(I)$ we have $a_i \neq \lambda$.

In order to lead this statement to a contradiction, it suffices to show that the element $[x_i - \lambda]$ has a multiplicative inverse in the ring R/I; namely, then multiplying this inverse with the eigenvalue equation $[(x_i - \lambda)\cdot v] = 0$ would give the contradiction $[v] = 0$

To any point $a \in \mathcal{V}(I)$, associate a polynomial $g_a \in R$ with the property that for any $b \in \mathcal{V}(I)$ the condition

$$g_a(b) \;=\; \begin{cases} 1 & \text{if } a = b\,, \\ 0 & \text{otherwise} \end{cases}$$

holds. If the first coordinates of all points in $\mathcal{V}(I)$ are distinct then we can — like in the well-known Lagrange interpolation formulas — specifically set

$$g_a \;=\; g_a(x_1) \;=\; \frac{\prod_{b\in\mathcal{V}(I)\setminus\{a\}}(x_1-b_1)}{\prod_{b\in\mathcal{V}(I)\setminus\{a\}}(a_1-b_1)}\,.$$

(Otherwise, using a linear transformation, we can reduce our situation to that one.)

Let g^* be the polynomial $g^* = \sum_{a\in\mathcal{V}(I)} \frac{1}{a_i-\lambda} g_a$. Then $(a_i-\lambda)g^*(a) = 1$ for all $a \in \mathcal{V}(I)$, in other words, the polynomial $1-(x_i-\lambda)g^*$ vanishes on all zeroes of the ideal I. By Hilbert's Nullstellensatz 1.2.9, there exists an $l \ge 1$ such that $(1-(x_i-\lambda)g^*)^l$ is contained in I. Expanding this polynomial and extracting the factors $(x_i - \lambda)$ we see that there exists a polynomial $f \in R$ such that $1-(x_i-\lambda)f$ is contained in I. Passing over to the

residue classes in R/I gives $[x_i - \lambda][f] = [1]$, so that f is the inverse element of $[x_i - \lambda]$ in R/I. This yields the desired contradiction.

Conversely, let $a \in \mathcal{V}(I)$ with $a_i = \lambda$. Let h_i be the minimal polynomial of m_i. By Lemma 2.5.5 it suffices to show that $h_i(\lambda) = 0$. Since by definition of the minimal polynomial the function $h_i(m_i)$ is the zero endomorphism on R/I, for the application of $h_i(m_i)$ on the element $[1]$ this means $h_i([x_i]) = h_i(m_i)([1]) = 0$ in R/I. For the polynomial $h_i(x_i)$ considered as polynomial in R we obtain $h_i(x_i) \in I$, so that the polynomial $h_i(x_i)$ vanishes on each element of $\mathcal{V}(I)$. Hence, $h_i(\lambda) = h_i(a_i) = 0$. □

Example 2.5.6. Let $I = \langle xy^2 + 1, x^2 - 1\rangle$. A Gröbner basis of I with respect to the lexicographic ordering is given by $\{y^4 - 1, y^2 + x\}$, hence a basis of R/I is $\{y^3, y^2, y, 1\}$. With respect to this basis, the representing matrices of the endomorphisms m_x and m_y are

$$M_x = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad M_y = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

In MAPLE, they can be computed using the command `MulMatrix`. The eigenvalues of M_x are -1 (twice) and 1 (twice), and the eigenvalues of M_y are $-1, 1, -i, i$. Indeed, we have $\mathcal{V}(I) = \{(1, i), (1, -i), (-1, 1), (-1, -1)\}$.

Note that from a computational point of view, Theorem 2.5.3 requires to know a basis of the coordinate ring R/I and the companion matrices. Then the resulting computational efforts depend on the dimension of R/I.

The situation is particularly nice if there exists a joint basis of eigenvectors, i.e., if there exists a matrix $S \in \mathbb{K}^{n\times n}$ and a diagonal matrix $D \in \mathbb{K}^{n\times n}$ with

$$M_i S = SD\,, \quad 1 \le i \le n\,.$$

In this case we say that the companion matrices are *simultaneously diagonalizable*.

Theorem 2.5.7. *The companion matrices $M_1, \ldots, M_n$ are simultaneously diagonalizable if I is radical.*

Proof. Let a be a point in $\mathcal{V}(I)$. As in the proof of Theorem 2.5.3, there exists a polynomial g be a polynomial with $g(a) = 1$ and $g(b) = 0$ for all $b \in \mathcal{V}(I) \setminus \{a\}$. Hence, the polynomial $(x_i - \lambda_i)g$ vanishes on all points of $\mathcal{V}(I)$. Hilbert's Nullstensatz then implies $(x_i - \lambda_i)[g] \in \sqrt{I} = I$, and thus $[g]$ is a joint eigenvector of $M_1, \ldots, M_n$. □

For numerical methods concerning the simultaneous diagonalization of matrices we refer the reader to Bunse-Gerstner, Byers, and Mehrmann [18]. In Section 4.2, a further refinement of the eigenvalue techniques will be used to study real roots.

Exercises

1. Let $G := \{yz - 1, x - z\}$ and $I := \langle G \rangle$ be an ideal in $\mathbb{C}[x, y]$. Show that G is a Gröbner basis of I for the lexicographic order $x \succ y \succ z$, determine the set of standard monomials of $\mathbb{C}[x, y]/I$ and compute the multiplication matrices M_x and M_y.

2. Let $p = \sum_{i=0}^{d} a_i x$ be monic, univariate polynomial and $I := \langle p \rangle$. Show that the representation matrix M_x of the endomorphism $m_x : R/I \to R/I$, $[f] \mapsto [xf]$ with respect to a natural basis coincides with the companion matrix C_p.

3. Let $R := \mathbb{K}[x_1, \ldots, x_n]$ and $f \in R$. Show that the mapping $m_f : R/I \to R/I$, $[g] \mapsto [f] \cdot [g]$ is an endomorphism.

4. Use a computer algebra system to determine the common roots of $f = x^2 + 3xy + y^2 - 1$ and $g = x^2 + 2xy + y + 3$ over $\mathbb{C}$, based on Stickelberger's Theorem.

5. If two endormorphisms f and g on a finite-dimensional vector space V are diagonalizable and $f \circ g = g \circ f$, then there exists a joint diagonalization. Conclude that in the situation of Stickelberger's Theorem for only two variables, there always exist a basis of joint eigenvectors.

2.6 Numerical Homotopy continuation

In Sections 2.2–Section 2.4, we discussed symbolic approaches to solving systems of polynomial equations. Resultants, for which we have formulas, do not always give precise information about the corresponding ideal, and are not universally applicable. On the other hand, Gröbner bases are universally applicable and may be readily computed. The drawback of Gröbner bases is that they contain too much information and therefore may be expensive or impossible to compute. Moreover, note that the eigenvalue techniques presented in Section 2.5 still require the set of standard monomials.

It is natural to ask for a universal method to find numerical solutions that does not require a Gröbner basis or the set of standard monomials. *Homotopy algorithms* furnish one such class of methods. These find all isolated solutions to a system of polynomial equations [86]. For large classes of systems, there are optimal homotopy algorithms, and they have the additional advantage of being inherently parallelizable.

Example 2.6.1. Suppose we want to compute the (four) solutions to the equations

$$x(y-1) \;=\; 1 \qquad \text{and} \qquad 2x^2 + y^2 \;=\; 9\,. \tag{2.16}$$

If we consider instead the system

$$x(y-1) \;=\; 0 \qquad \text{and} \qquad 2x^2 + y^2 \;=\; 9\,, \tag{2.17}$$

then we obtain the following solutions by inspection

$$(0, \pm 3) \qquad \text{and} \qquad (\pm 2, 1)\,.$$

Figure 2.2 shows the two systems, the first seeks the intersection of the hyperbola with the ellipse, while the second replaces the hyperbola by the two lines.

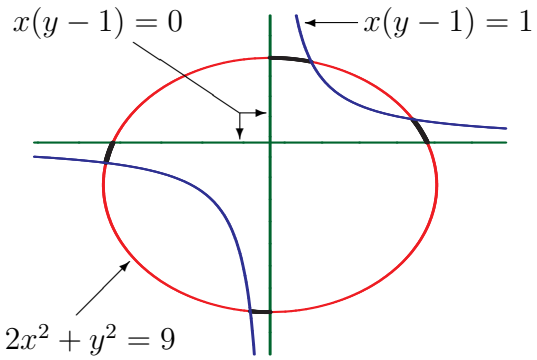


Figure 2.2: The intersection of a hyperbola with an ellipse.

The 1-parameter family of polynomial systems

$$x(y-1) \;=\; t \qquad \text{and} \qquad 2x^2 + y^2 \;=\; 9\,,$$

interpolates between these two systems as t varies from 0 to 1. This defines four solution curves $z_i(t)$ (the thickened arcs in Figure 2.2) for $t \in [0,1]$, which connect each solution $(0, \pm 3)$, $(\pm 2, 1)$ of (2.17) to a solution of (2.16). We merely trace each solution curve $z_i(t)$ from the known solution $z_i(0)$ at $t = 0$ to the desired solution $z_i(1)$ at $t = 1$, obtaining all solutions of (2.16).

More generally, suppose that we want to find all solutions to a 0-dimensional *target system* of polynomial equations

$$f_1(x_1, \ldots, x_n) \;=\; f_2(x_1, \ldots, x_n) \;=\; \cdots \;=\; f_N(x_1, \ldots, x_n) \;=\; 0\,, \qquad (2.18)$$

written compactly as $F(x) = 0$. Numerical homotopy continuation finds these solutions if we have a *homotopy*, which is a system $H(x,t)$ of polynomials in $n+1$ variables such that

1. The systems $H(x,1) = 0$ and $F(x) = 0$ both have the same solutions;

2. We know all solutions to the *start system* $H(x,0) = 0$;

3. The components of the variety defined by $H(x,t) = 0$ include curves whose projection to $\mathbb{C}$ (via the second coordinate t) is dominant; and

4. The solutions to the system $H(x,t) = 0$, where $t \in [0,1)$, occur at smooth points of curves from (3) in the variety $H(x,t) = 0$.

We summarize these properties: The homotopy $H(x,t)$ interpolates between our original system (1) and a trivial system (2) such that all isolated solutions are attained (3) and we can do this avoiding singularities (4).

Given such a homotopy, we restrict the variety $H(x,t) = 0$ to $t \in [0,1]$ and obtain finitely many real arcs in $\mathbb{C}^n \times [0,1]$ which connect (possibly singular) solutions of the target system $H(x,1) = 0$ to solutions of the start system $H(x,0) = 0$. We then numerically trace each arc from $t = 0$ to $t = 1$, obtaining all isolated solutions to the target system.

The homotopy is *optimal* if every solution at $t = 0$ is connected to a unique solution at $t = 1$ along an arc. This is illustrated in Figure 2.3.

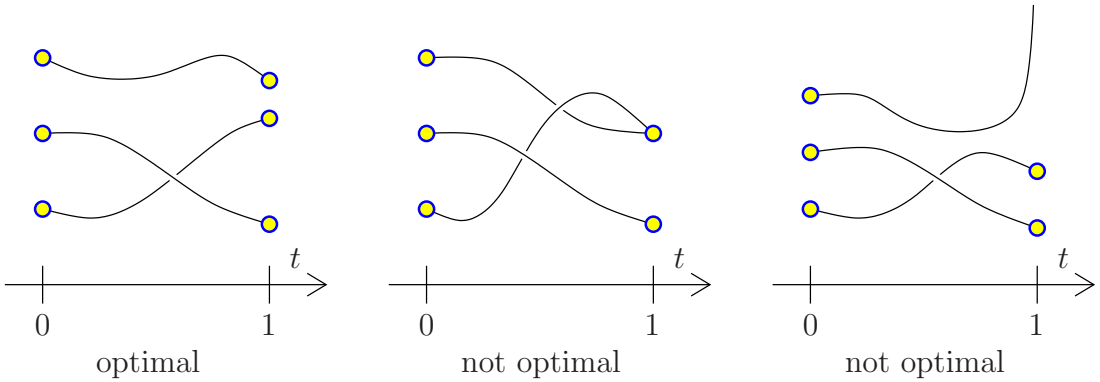


Figure 2.3: Optimal and non-optimal homotopies

Remark 2.6.2. Homotopy continuation software often constructs a homotopy as follows. Let $F(x)$ be the target system (2.18) and suppose we have solutions to a *start system* $G(x)$. Then for a number $\gamma \in \mathbb{C}$ with $|\gamma| = 1$ define

$$H(x,t) \;:=\; \gamma t F(x) + (1-t)G(x)\,.$$

Then $H(x,t)$ satisfies the definition of a homotopy for all but finitely many γ. The software detects the probability 0 event that $H(x,t)$ does not satisfy the definition when it encounters a singularity, and then it recreates the homotopy with a different γ.

Example 2.6.3. We illustrate these ideas. Suppose we have a square system

$$f_1(x_1,\ldots,x_n) \;=\; \cdots \;=\; f_n(x_1,\ldots,x_n) \;=\; 0\,, \tag{2.19}$$

where $\deg f_i = d_i$. If, for each $i = 1,\ldots,n$, we set

$$g_i \;:=\; \prod_{i=1}^{d_i}(x_i - j)\,, \tag{2.20}$$

then we immediately see that the start system

$$g_1(x_1,\ldots,x_n) \;=\; \cdots \;=\; g_n(x_1,\ldots,x_n) \;=\; 0 \tag{2.21}$$

has the $d_1 d_2 \cdots d_n$ solutions

$$\prod_{i=1}^{n}\{1,2,\ldots,d_i\} \;=\; \{(a_1,\ldots,a_n)\in\mathbb{Z}^n \mid 1\le a_i\le d_i,\ i=1,\ldots,n\}\,.$$

The *Bézout homotopy* $H(x,t)$ consists of the polynomials

$$h_i(x,t) \;:=\; \gamma t\cdot f_i(x) + (1-t)g_i(x) \text{ for } i=1,\ldots,n\,, \tag{2.22}$$

where γ is an arbitrary complex number. If the isolated solutions to the original system (2.19) $F(x)=0$ occur without multiplicities, then $H(x,t)$ is a homotopy, as Bézout's Theorem 2.4.5 implies that all intermediate systems $H(x,t_0)$ have at most $d_1\cdots d_n$ isolated solutions. Furthermore, if γ is chosen from a generic set, then there are no multiple solutions for $t\in[0,1]$. We may replace the polynomials g_i (2.20) by any other polynomials (of the same degree) so that the start system has $d_1\cdots d_n$ solutions. For example, we may take $g_i := x_i^{d_i}-1$.

Path following algorithms use predictor-corrector methods, which are conceptually simple for *square systems*, where the number of equations equals the number of variables.

Given a point $(x^{(0)},t^{(0)})$ on an arc such that $t^{(0)}\in[0,1)$, the $n\times n$ matrix

$$H_x \;:=\; \left(\frac{\partial H_i}{\partial x_j}\right)_{i,j=1}^{n}$$

is regular at $(x^{(0)},t^{(0)})$, which follows from the definition of the homotopy. Let $H_t = (\partial H_1/\partial t,\ldots,\partial H_n/\partial t)^T$, given Δt we set

$$\Delta x \;:=\; -\Delta t\, H_x(x^{(0)},t^{(0)})^{-1}\, H_t(x^{(0)},t^{(0)})\,.$$

Then the vector $(\Delta x,\Delta t)$ is tangent to the arc $H(x,t)=0$ at the point $(x^{(0)},t^{(0)})$. For $t^{(1)} = t^{(0)}+\delta t$, the point $(x',t^{(1)}) = (x^{(0)}+\Delta x,t^{(1)})$ is an approximation to the point $(x^{(1)},t^{(1)})$ on the same arc. This constitutes a first order predictor step. A corrector step uses the multivariate Newton method for the system $H(x,t^{(1)})=0$, refining the approximate solution x' to a solution $x^{(1)}$. In practice, the points $x^{(0)}$ and $x^{(1)}$ are numerical (approximate) solutions, and both the prediction and correction steps require that $\det H_x\neq 0$ at every point where the computation of the Jacobian matrix H_x is done.

When the system is not square, additional strategies must be employed to enable the path following.

These numerical methods for following a solution curve $x(t)$ will break down if $x(t)$ approaches a singularity of the variety $H(x,t)=0$, as the rank of the Jacobian matrix

$H_x(x(t), x)$ will drop at such points. This will occur when t lies in an algebraic subvariety of $\mathbb{C}$, that is, for finitely many $t \in \mathbb{C}$. In practice, the choice of random complex number γ in (2.22) precludes this situation.

Another possibility is that the target system has singular solutions at which the homotopy must necessarily end. Numerical algorithms to detect and handle these and other end-game situations have been devised and implemented. This is illustrated in the middle picture in Figure 2.3.

Another, more serious possibility is when the solution curve $x(t)$ does not converge to an isolated solution of the original system. By assumptions (III) and (IV), this can only occur of there are fewer isolated solutions to our original system than to the start system. In this case, some solution curves $x(t)$ will diverge to ∞ as t approaches 1. This illustrated in the rightmost picture in Figure 2.3. A homotopy $H(x,t)$ is optimal if this does not occur and if all intermediate systems $H(x,t) = 0$ (including $t = 0, 1$) have the same number of isolated solutions.

For example, suppose our original system (2.18) is a generic square system ($N = n$) with polynomials of degrees $d_1, \ldots, d_n$. (This is also called a generic dense system.) Then Bézout's Theorem implies that it has $d_1 d_2 \cdots d_n$ isolated solutions. Since the start system also has this number of solutions, almost all intermediate solutions will, too, and so we obtain the following fundamental result.

Theorem 2.6.4. *For generic dense systems, the Bézout homotopy is optimal.*

The only difficulty with Theorem 2.6.4 and the Bézout homotopy is that polynomial systems arising 'in nature' from applications are rarely generic dense systems. The Bézout bound for the number of isolated solutions is typically not achieved. A main challenge in this subject is to construct optimal homotopies.

For example, consider the system of cubic polynomials

$$\begin{array}{rcl} 1 + 2x + 3y + 4xy + 5x^2y + 6xy^2 & = & 0 \\ 5 + 7x + 11y + 13xy + 17x^2y + 19xy^2 & = & 0 \end{array} \tag{2.23}$$

we leave it an an exercise to check that this has 5 solutions and not 9, which is the Bézout bound.

Another source of optimal homotopies are *Cheater homotopies* [53], which are constructed from families of polynomial systems. For example, consider pairs of polynomials (f, g) whose monomials are as in (2.23),

$$\begin{array}{rcl} f & := & f_1 + f_2 x + f_3 y + f_4 xy + f_5 x^2 y + f_6 xy^2 \,, \\ g & := & g_1 + g_2 x + g_3 y + g_4 xy + g_5 x^2 y + g_6 xy^2 \,. \end{array}$$

The coefficients of these polynomials show that the space of such pairs is $\mathbb{C}^{12}$. Then the total space of the polynomial system $f(x,y) = g(x,y) = 0$

$$U \;:=\; \{(x, y, f, g) \in \mathbb{C}^{14} \mid f(x,y) = g(x,y) = 0\}$$

has dimension 12, and for a general $(f, g) \in \mathbb{C}^{12}$, there are 5 solutions (x, y) to the equations.

If $\varphi\colon \mathbb{C} \to \mathbb{C}^{12}$ is an embedding of $\mathbb{C}$ into $\mathbb{C}^{12}$ in which $\varphi(0)$ and $\varphi(1)$ are general pairs of polynomials and we write $\varphi(t) = (f_t, g_t)$, then

$$\varphi^* U \;=\; \{(x, y, f_t, g_t) \mid f_t(x, y) = g_t(x, y) = 0\} \; .$$

This is defined by a system $H(x, t) = 0$, which gives an optimal homotopy between the start stystem $f_0 = g_0 = 0$ and $f_1 = g_1 = 0$.[†]

The computational complexity of solving systems of polynomials using an optimal homotopy is roughly linear in the number of solutions, for a fixed number of variables. The basic idea is that the cost of following each solution curve is essentially constant. This happy situation is further enhanced as homotopy continuation algorithms are inherently massively parallelizable—once the initial precomputation of solving the start system and setting up the homotopies is completed, then each solution curve may be followed independently of all other solution curves.

Exercises for Section 5

1. Verify the claim in the text that the system (2.23) has five solutions. Show that only one is real.

2.7 Notes

Gröbner bases have been introduced by Hironaka (under the term"standard bases") in 1964 [39], and independently by Bruno Buchberger in his dissertation in 1965 [15, 16]. The term "Gröbner basis" honors Buchberger's doctoral advisor Wolfgang Gröbner.

For further information on technqiues for solving systems of polynomial equations see the books of Cox, Little, and O'Shea [21, 20], Sturmfels [90] as well as Emiris and Dickenstein [23].

[†]Fix this example!

Chapter 3

Structure of varieties

Outline:

1. Zariski topology.
2. Irreducible decomposition and dimension.
3. Rational functions.
4. Maps of projective varieties
5. Smooth and singular points.
6. Hilbert functions and degree.

Put this somewhere about projective varieties and Zariski Topology Many of the basic notions from affine varieties extend to projective varieties for many of the the same reasons. In particular, we have the Zariski topology with open and closed sets, and the notion of generic sets. Projective varieties are finite unions of irreducible varieties, in an essentially unique way. The definitions, statements, and proofs are the same as in Sections 3.1 and 3.2.

Furthermore, a subset $Z \subset \mathbb{P}^n$ of projective space is Zariski closed if and only if its intersection with each U_i is closed.

3.1 Generic properties of varieties

Many properties in algebraic geometry hold for almost all points of a variety or for almost all objects of a given type. For example, matrices are almost always invertible, univariate polynomials of degree d almost always have d distinct roots, and multivariate polynomials are almost always irreducible. We develop the terminology 'generic' and 'Zariski open' to describe this situation.

A starting point is that intersections and unions of affine varieties behave well.

Theorem 3.1.1. *The intersection of any collection of affine varieties is an affine variety. The union of any finite collection of affine varieties is an affine variety.*

Proof. For the first statement, let $\{I_t \mid t \in T\}$ be a collection of ideals in $\mathbb{F}[x_1, \ldots, x_n]$. Then we have

$$\bigcap_{t\in T} \mathcal{V}(I_t) \;=\; \mathcal{V}\Big(\bigcup_{t\in T} I_t\Big)\,.$$

Arguing by induction on the number of varieties, shows that it suffices to establish the second statement for the union of two varieties but that case is Lemma 1.2.11 (3). □

Theorem 3.1.1 shows that affine varieties have the same properties as the closed sets of a topology on $\mathbb{A}^n$. This was observed by Oscar Zariski.

Definition 3.1.2. We call an affine variety a *Zariski closed set*. The complement of a Zariski closed set is a *Zariski open set*. The *Zariski topology* on $\mathbb{A}^n$ is the topology whose closed sets are the affine varieties in $\mathbb{A}^n$. The *Zariski closure* of a subset $Z \subset \mathbb{A}^n$ is the smallest variety containing Z, which is $\overline{Z} := \mathcal{V}(\mathcal{I}(Z))$, by Lemma 1.2.4. Any subvariety X of $\mathbb{A}^n$ inherits its Zariski topology from $\mathbb{A}^n$, the closed subsets are simply the subvarieties of X. A subset $Z \subset X$ of a variety X is *Zariski dense* in X if its closure is X.

We emphasize that the purpose of this terminology is to aid our discussion of varieties, and not because we will use notions from topology in any essential way. This Zariski topology is behaves quite differently from the usual *Euclidean* topology on $\mathbb{R}^n$ or $\mathbb{C}^n$ with which we may be familiar. A topology on a space may be defined by giving a collection of basic open sets which generate the topology jn that any open set is a union or a finite intersection of basic open sets. In the Euclidean topology, the basic open sets are balls. Let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. The *ball* with radius $\epsilon > 0$ centered at $z \in \mathbb{A}^n$ is

$$B(z, \epsilon) \;:=\; \{a \in \mathbb{A}^n \mid \sum |a_i - z_i|^2 \;<\; \epsilon\}\,.$$

In the Zariski topology, the basic open sets are complements of hypersurfaces, called principal open sets. Let $f \in \mathbb{F}[x_1, \ldots, x_n]$ and set

$$U_f \;:=\; \{a \in \mathbb{A}^n \mid f(a) \neq 0\}\,.$$

In both these topologies the open sets are unions of basic open sets—we do not need intersections to generate the given topology.

We give two examples to illustrate the Zariski topology.

Example 3.1.3. The Zariski closed subsets of $\mathbb{A}^1$ are the empty set, finite collections of points, and $\mathbb{A}^1$ itself. Thus when $\mathbb{F}$ is infinite the usual separation property of Hausdorff spaces (any two points are covered by two disjoint open sets) fails spectacularly as any two nonempty open sets meet.

Example 3.1.4. The Zariski topology on a product $X \times Y$ of affine varieties X and Y is in general not the product topology. In the product topology on $\mathbb{A}^2$, the closed sets are finite unions of sets of the following form: the empty set, points, vertical and horizontal lines of the form $\{a\} \times \mathbb{A}^1$ and $\mathbb{A}^1 \times \{a\}$, and the whole space $\mathbb{A}^2$. On the other hand, $\mathbb{A}^2$ contains a rich collection of 1-dimensional subvarieties (called *plane curves*), such as the cubic plane curves of Section 1.1.

We compare the Zariski topology with the Euclidean topology.

Theorem 3.1.5. *Suppose that $\mathbb{F}$ is one of $\mathbb{R}$ or $\mathbb{C}$. Then*

1. *A Zariski closed set is closed in the Euclidean topology on $\mathbb{A}^n$.*
2. *A Zariski open set is open in the Euclidean topology on $\mathbb{A}^n$.*
3. *A nonempty Euclidean open set is Zariski dense.*
4. *$\mathbb{R}^n$ is Zariski dense in $\mathbb{C}^n$.*
5. *A Zariski closed set is nowhere dense in the Euclidean topology on $\mathbb{A}^n$.*
6. *A nonempty Zariski open set is dense in the the Euclidean topology on $\mathbb{A}^n$.*

Proof. For statements 1 and 2, observe that a Zariski closed set $\mathcal{V}(I)$ is the intersection of the hypersurfaces $\mathcal{V}(f)$ for $f \in I$, so it suffices to show this for a hypersurface $\mathcal{V}(f)$. But then Statement 1 (and hence also 2) follows as the polynomial function $f\colon \mathbb{A}^n \to \mathbb{F}$ is continuous in the Euclidean topology, and $\mathcal{V}(f) = f^{-1}(0)$.

We show that any ball $B(z, \epsilon)$ is Zariski dense. If a polynomial f vanishes identically on $B(z, \epsilon)$, then all of its partial derivatives do as well. This implies that its Taylor series expansion at z is identically zero. But then f is the zero polynomial. This shows that $\mathcal{I}(B) = \{0\}$, and so $\mathcal{V}(\mathcal{I}(B)) = \mathbb{A}^n$, that is, B is dense in the Zariski topology on $\mathbb{A}^n$.

For statement 4, we use the same argument. If a polynomial vanishes on $\mathbb{R}^n$, then all of its partial derivatives vanish and so f must be the zero polynomial. Thus $\mathcal{I}(\mathbb{R}^n) = \{0\}$ and $\mathcal{V}(\mathcal{I}(\mathbb{R}^n)) = \mathbb{C}^n$.

For statements 5 and 6, observe that if f is nonconstant, then the interior of the (Euclidean) closed set $\mathcal{V}(f)$ is empty and so $\mathcal{V}(f)$ is nowhere dense. A subvariety is an intersection of nowhere dense hypersurfaces, so varieties are nowhere dense. The complement of a nowhere dense set is dense, so Zariski open sets are dense in $\mathbb{A}^n$. □

The last statement of Theorem 3.1.5 leads to the useful notions of genericity and generic sets and properties.

Definition 3.1.6. Let X be a variety. A subset $Y \subset X$ is called *generic* if it contains a Zariski dense open subset U of X. That is, we have $U \subset Y \subset X$ with U Zariski open and $\overline{U} = X$. A property is generic if the set of points on which it holds is a generic set. Points of a generic set are called general points.

This notion of general depends on the context, and so care must be exercised in its use. For example, we may identify $\mathbb{A}^3$ with the set of quadratic polynomials in x via

$$(a, b, c) \longmapsto ax^2 + bx + c\,.$$

Then, the general quadratic polynomial does not vanish when $x = 0$. (We just need to avoid quadratics with $c = 0$.) On the other hand, the general quadratic polynomial has

two roots, as we need only avoid quadratics with $b^2 - 4ac = 0$. The quadratic $x^2 - 2x + 1$ is general in the first sense, but not in the second, while the quadratic $x^2 + x$ is general in the second sense, but not in the first. Despite this ambiguity, we will see that general is a very useful concept.

When $\mathbb{F}$ is $\mathbb{R}$ or $\mathbb{C}$, generic sets are dense in the Euclidean topology, by Theorem 3.1.5(6). Thus generic properties hold almost everywhere, in the standard sense.

Example 3.1.7. The generic $n \times n$ matrix is invertible, as it is a nonempty principal open subset of $\mathrm{Mat}_{n\times n} = \mathbb{A}^{n\times n}$. It is the complement of the variety $\mathcal{V}(\det)$ of singular matrices. Define the *general linear group* GL_n to be the set of all invertible matrices,

$$GL_n \;:=\; \{M \in \mathrm{Mat}_{n\times n} \mid \det(M) \neq 0\} \;=\; U_{\det}\,.$$

Example 3.1.8. The general univariate polynomial of degree n has n distinct complex roots. Identify $\mathbb{A}^n$ with the set of univariate polynomials of degree n via

$$(a_1, \ldots, a_n) \;\in\; \mathbb{A}^n \;\longmapsto\; x^n + a_1 x^{n-1} + \cdots + a_{n-1}x + a_n \;\in\; \mathbb{F}[x]\,. \tag{3.1}$$

The classical discriminant $\Delta \in \mathbb{F}[a_1, \ldots, a_n]$ (See Example 2.3.6) is a polynomial of degree $2n - 2$ which vanishes precisely when the polynomial (3.1) has a repeated factor. This identifies the set of polynomials with n distinct complex roots as the set U_Δ. The discriminant of a quadratic $x^2 + bx + c$ is $b^2 - 4c$.

Example 3.1.9. The generic complex $n \times n$ matrix is semisimple (diagonalizable). Let $M \in \mathrm{Mat}_{n\times n}$ and consider the (monic) characteristic polynomial of M

$$\chi(x) \;:=\; \det(xI_n - M)\,.$$

We do not show this by providing an algebraic characterization of semisimplicity. Instead we observe that if a matrix $M \in \mathrm{Mat}_{n\times n}$ has n distinct eigenvalues, then it is semisimple. The coefficients of the characteristic polynomial $\chi(x)$ are polynomials in the entries of M. Evaluating the discriminant at these coefficients gives a polynomial ψ which vanishes when the characteristic polynomial $\chi(x)$ of M has a repeated root.

We see that the set of matrices with distinct eigenvalues equals the basic open set U_ψ, which is nonempty. Thus the set of semisimple matrices contains an open dense subset of $\mathrm{Mat}_{n\times n}$ and is therefore generic.

When $n = 2$,

$$\det\left(xI_2 \;-\; \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\right) \;=\; x^2 - x(a_{11} + a_{22}) + a_{11}a_{22} - a_{12}a_{21}\,,$$

and so the polynomial ψ is $(a_{11} + a_{22})^2 - 4(a_{11}a_{22} - a_{12}a_{21})$.

In each of these examples, we used the following easy fact.

Proposition 3.1.10. *A set $X \subset \mathbb{A}^n$ is generic if and only if there is a nonconstant polynomial that vanishes on its complement, if and only if it contains a basic open set U_f.*

More generally, if $X \subset \mathbb{A}^n$ is a variety and $f \in \mathbb{F}[x_1, \ldots, x_n]$ is a polynomial which is not identically zero on X ($f \notin \mathcal{I}(X)$), then we have the *principal open subset* of X,

$$X_f \;:=\; X - \mathcal{V}(F) \;=\; \{x \in X \mid f(x) \neq 0\}\,.$$

Lemma 3.1.11. *Any Zariski open subset U of a variety X is a finite union of principal open subsets.*

Proof. The complement $Y := X - U$ of a Zariski open subset U of X is a Zariski closed subset. The ideal $\mathcal{I}(Y)$ of Y in $\mathbb{A}^n$ contains the ideal $\mathcal{I}(X)$ of X. By the Hilbert Basis Theorem, there are polynomials $f_1, \ldots, f_m \in \mathcal{I}(Y)$ such that

$$\mathcal{I}(Y) \;=\; \langle \mathcal{I}(X),\; f_1, \ldots, f_m \rangle\,.$$

Then $X_{f_1} \cup \cdots \cup X_{f_m}$ is equal to

$$(X - \mathcal{V}(f_1)) \cup \cdots \cup (X - \mathcal{V}(f_m)) \;=\; X - (\mathcal{V}(f_1) \cap \cdots \cap \mathcal{V}(f_m)) \;=\; X - Y \;=\; U\,. \quad \square$$

Exercises

1. Look up the definition of a topology in a text book and verify the claim that the collection of affine subvarieties of $\mathbb{A}^n$ form the closed sets in a topology on $\mathbb{A}^n$.

2. Prove that a closed set in the Zariski topology on $\mathbb{A}^1$ is either the empty set, a finite collection of points, or $\mathbb{A}^1$ itself.

3. Let $n \leq m$. Prove that a generic $n \times m$ matrix has rank n.

4. Prove that the generic triple of points in $\mathbb{A}^2$ are the vertices of a triangle.

5. (a) Describe all the algebraic varieties in $\mathbb{A}^1$.
   (b) Show that any open set in $\mathbb{A}^1 \times \mathbb{A}^1$ is open in $\mathbb{A}^2$.
   (c) Find a Zariski open set in $\mathbb{A}^2$ which is not open in $\mathbb{A}1 \times \mathbb{A}^1$.

6. (a) Show that the Zariski topology in $\mathbb{A}^n$ is not Hausdorff if $\mathbb{F}$ is infinite.
   (b) Prove that any nonempty open subset of $\mathbb{A}^n$ is dense.
   (c) Prove that $\mathbb{A}^n$ is compact.

3.2 Unique factorization for varieties

Every polynomial factors uniquely as a product of irreducible polynomials. A basic structural result about algebraic varieties is an analog of unique factorization. Any algebraic variety is the finite union of irreducible varieties, and this decomposition is unique.

A polynomial $f \in \mathbb{F}[x_1, \ldots, x_n]$ is *decomposable* if we may factor f notrivially, that is, if $f = gh$ with neither g nor h a constant polynomial. Otherwise f is *indecomposable*. Any polynomial $f \in \mathbb{F}[x_1, \ldots, x_n]$ may be factored

$$f \;=\; g_1^{\alpha_1} g_2^{\alpha_2} \cdots g_m^{\alpha_m} \tag{3.2}$$

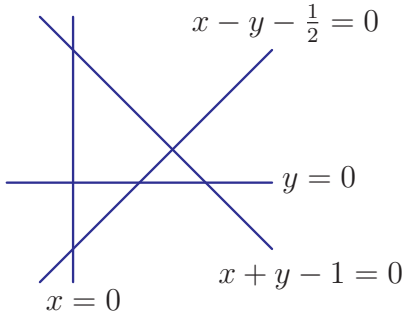
where the exponents α_i are positive integers, each polynomial g_i is irreducible and nonconstant, and when $i \neq j$ the polynomials g_i and g_j are not proportional. This factorization is essentially unique as any other such factorization is obtained from this by permuting the factors and possibly multiplying each polynomial g_i by a constant. The polynomials g_j are *irreducible factors* of f.

When $\mathbb{F}$ is algebraically closed, this algebraic property has a consequence for the geometry of hypersurfaces. Suppose that a polynomial f has a factorization (3.2) into irreducible polynomials. Then the hypersurface $X = \mathcal{V}(f)$ is the union of hypersurfaces $X_i := \mathcal{V}(g_i)$, and this decomposition

$$X \;=\; X_1 \cup X_2 \cup \cdots \cup X_m$$

of X into hypersurfaces X_i defined by irreducible polynomials is unique.

For example, $\mathcal{V}(xy(x+y-1)(x-y-\frac{1}{2}))$ is the union of four lines in $\mathbb{A}^2$.



This decomposition property is shared by general varieties.

Definition 3.2.1. A variety X is *reducible* if it is the union $X = Y \cup Z$ of proper closed subvarieties $Y, Z \subsetneq X$. Otherwise X is *irreducible*. In particular, if an irreducible variety is written as a union of subvarieties $X = Y \cup Z$, then either $X = Y$ or $X = Z$.

Example 3.2.2. Figure 1.2 in Section 1.2 shows that $\mathcal{V}(xy+z,\, x^2-x+y^2+yz)$ consists of two space curves, each of which is a variety in its own right. Thus it is reducible. To see this, we solve the two equations $xy + z = x^2 - x + y^2 + yz = 0$. First note that

$$x^2 - x + y^2 + yz \;-\; y(xy+z) \;=\; x^2 - x + y^2 - xy^2 \;=\; (x-1)(x-y^2)\,.$$

Thus either $x = 1$ or else $x = y^2$. When $x = 1$, we see that $y + z = 0$ and these equations define the line in Figure 1.2. When $x = y^2$, we get $z = -y^3$, and these equations define the cubic curve parametrized by $(t^2, t, -t^3)$.

Figure 3.1 shows another reducible variety. It has six components, one is a surface,

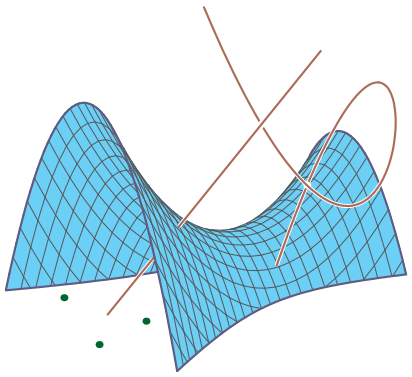


Figure 3.1: A reducible variety

two are space curves, and three are points.

Theorem 3.2.3. *A product $X \times Y$ of irreducible varieties is irreducible.*

Proof. Suppose that $Z_1, Z_2 \subset X \times Y$ are subvarieties with $Z_1 \cup Z_2 = X \times Y$. We assume that $Z_2 \neq X \times Y$ and use this to show that $Z_1 = X \times Y$. For each $x \in X$, identify the subvariety $\{x\} \times Y$ with Y. This irreducible variety is the union of two subvarieties,

$$\{x\} \times Y \;=\; \big((\{x\} \times Y) \cap Z_1\big) \;\cup\; \big((\{x\} \times Y) \cap Z_2\big)\,,$$

and so one of these must equal $\{x\} \times Y$. In particular, we must either have $\{x\} \times Y \subset Z_1$ or else $\{x\} \times Y \subset Z_2$. If we define

$$\begin{aligned} X_1 &= \{x \in X \mid \{x\} \times Y \subset Z_1\}\,, \quad \text{and} \\ X_2 &= \{x \in X \mid \{x\} \times Y \subset Z_2\}\,, \end{aligned}$$

then we have just shown that $X = X_1 \cup X_2$. Since $Z_2 \neq X \times Y$, we have $X_2 \neq X$. We claim that both X_1 and X_2 are subvarieties of X. Then the irreducibility of X implies that $X = X_1$ and thus $X \times Y = Z_1$.

We will show that X_1 is a subvariety of X. For $y \in Y$, set

$$X_y \;:=\; \{x \in X \mid (x, y) \in Z_1\}\,.$$

Since $X_y \times \{y\} = (X \times \{y\}) \cap Z_1$, we see that X_y is a subvariety of X. But we have

$$X_1 \;=\; \bigcap_{y \in Y} X_y\,,$$

which shows that X_1 is a subvariety of X. An identical argument for X_2 completes the proof. $\square$

The geometric notion of an irreducible variety corresponds to the algebraic notion of a prime ideal. An ideal $I \subset \mathbb{F}[x_1, \ldots, x_n]$ is *prime* if whenever $fg \in I$ with $f \notin I$, then we have $g \in I$. Equivalently, if whenever $f, g \notin I$ then $fg \notin I$.

Theorem 3.2.4. *A variety X is irreducible if and only if its ideal $\mathcal{I}(X)$ is prime.*

Proof. Let X be a variety. First suppose that X is irreducible. Let $f, g \notin \mathcal{I}(X)$. Then neither f nor g vanishes identically on X. Thus $Y := X \cap \mathcal{V}(f)$ and $Z := X \cap \mathcal{V}(z)$ are proper subvarieties of X. Since X is irreducible, $Y \cup Z = X \cap \mathcal{V}(fg)$ is also a proper subvariety of X, and thus $fg \notin \mathcal{I}(X)$.

Suppose now that X is reducible. Then $X = Y \cup Z$ is the union of proper subvarieties Y, Z of X. Since $Y \subsetneq X$ is a subvariety, we have $\mathcal{I}(X) \subsetneq \mathcal{I}(Y)$. Let $f \in \mathcal{I}(Y) - \mathcal{I}(X)$, a polynomial which vanishes on Y but not on X. Similarly, let $g \in \mathcal{I}(Z) - \mathcal{I}(X)$ be a polynomial which vanishes on Z but not on X. Since $X = Y \cup Z$, fg vanishes on X and therefore lies in $\mathcal{I}(X)$. This shows that I is not prime. □

We have seen examples of varieties with one, two, four, and six irreducible components. Taking products of distinct irreducible polynomials (or dually unions of distinct hypersurfaces), gives varieties having any finite number of irreducible components. This is all that can occur as Hilbert's Basis Theorem implies that a variety is a union of finitely many irreducible varieties.

Lemma 3.2.5. *Any affine variety is a finite union of irreducible subvarieties.*

Proof. An affine variety X either is irreducible or else we have $X = Y \cup Z$, with both Y and Z proper subvarieties of X. We may similarly decompose whichever of Y and Z are reducible, and continue this process, stopping only when all subvarieties obtained are irreducible. *A priori*, this process could continue indefinitely. We argue that it must stop after a finite number of steps.

If this process never stops, then X must contain an infinite chain of subvarieties, each properly contained in the previous,

$$X \supsetneq X_1 \supsetneq X_2 \supsetneq \cdots .$$

Their ideals form an infinite increasing chain of ideals in $\mathbb{F}[x_1, \ldots, x_n]$,

$$\mathcal{I}(X) \subsetneq \mathcal{I}(X_1) \subsetneq \mathcal{I}(X_2) \subsetneq \cdots .$$

The union I of these ideals is again an ideal. Note that no ideal $\mathcal{I}(X_m)$ is equal to I. By the Hilbert Basis Theorem, I is finitely generated, and thus there is some integer m for which $\mathcal{I}(X_m)$ contains these generators. But then $I = \mathcal{I}(X_m)$, a contradiction. □

A consequence of this proof is that any decreasing chain of subvarieties of a given variety must have finite length. When $\mathbb{F}$ is infinite, there are such decreasiung chains of arbitrary length. There is however a bound for the length of the longest decreasing chain of irreducible subvarieties.

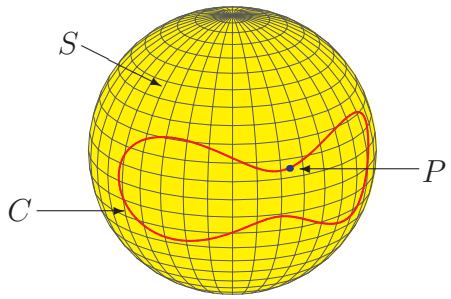
*[*Combinatorial Definition of Dimension] The *dimension* of a variety X is essentially the length of the longest decreasing chain of irreducible subvarieties of X. If

$$X \supset X_0 \supsetneq X_1 \supsetneq X_2 \supsetneq \cdots \supsetneq X_m \supsetneq \emptyset,$$

with each X_i irreducible is such a chain of maximal length, then X has dimension m.

Since maximal ideals of $\mathbb{C}[x_1, \ldots, x_n]$ necessarily have the form $\mathfrak{m}_a$, we see that X_m must be a point when $\mathbb{F} = \mathbb{C}$. The only problem with this definition is that we cannot yet show that it is well-founded, as we do not yet know that there is a bound on the length of such a chain. In Section 3.6 we shall prove that this definition is correct by relating it to other notions of dimension.

Example 3.2.6. The sphere S has dimension at least two, as we have the chain of subvarieties $S \supsetneq C \supsetneq P$ as shown below.



It is quite challenging to show that any maximal chain of irreducible subvarieties of the sphere has length 2 with what we know now.

By Lemma 3.2.5, an affine variety X may be written as a finite union

$$X = X_1 \cup X_2 \cup \cdots \cup X_m$$

of irreducible subvarieties. We may assume this is irredundant in that if $i \neq j$ then X_i is not a subvariety of X_j. If we did have $i \neq j$ with $X_i \subset X_j$, then we may remove X_i from the decomposition. We prove that this decomposition is unique, which is the main result of this section and a basic structural result about varieties.

Theorem 3.2.7 (Unique Decomposition of Varieites)**.** *A variety X has a unique irredundant decomposition as a finite union of irreducible subvarieties*

$$X = X_1 \cup X_2 \cup \cdots \cup X_m.$$

We call these distinguished subvarieties X_i the *irreducible components* of X.

Proof. Suppose that we have another irredundant decomposition into irreducible subvarieties,

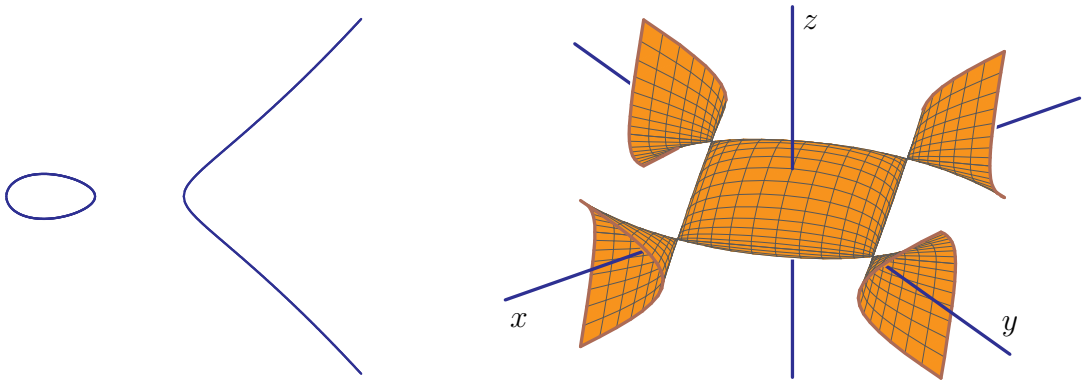
$$X = Y_1 \cup Y_2 \cup \cdots \cup Y_n,$$

where each Y_i is irreducible. Then

$$X_i \;=\; (X_i \cap Y_1) \;\cup\; (X_i \cap Y_2) \;\cup\; \cdots \;\cup\; (X_i \cap Y_n)\,.$$

Since X_i is irreducible, one of these must equal X_i, which means that there is some index j with $X_i \subset Y_j$. Similarly, there is some index k with $Y_j \subset X_k$. Since this implies that $X_i \subset X_k$, we have $i = k$, and so $X_i = Y_j$. This implies that $n = m$ and that the second decomposition differs from the first solely by permuting the terms. □

When $\mathbb{F} = \mathbb{C}$, we will show that an irreducible variety is connected in the usual Euclidean topology. We will even show that the smooth points of an irreducible variety are connected. Neither of these facts are true over $\mathbb{R}$. Below, we display the irreducible cubic plane curve $\mathcal{V}(y^2 - x^3 + x)$ in $\mathbb{A}^2_{\mathbb{R}}$ and the surface $\mathcal{V}((x^2-y^2)^2 - 2x^2 - 2y^2 - 16z^2 + 1)$ in $\mathbb{A}^3_{\mathbb{R}}$.



Both are irreducible hypersurfaces. The first has two connected components in the Euclidean topology, while in the second, the five components of smooth points meet at the four singular points.

Exercises

1. Show that the ideal of a hypersurface $\mathcal{V}(f)$ is generated by the *squarefree* part of f, which is the product of the irreducible factors of f, all with exponent 1.

2. For every positive integer n, give a decreasing chain of subvarieties of $\mathbb{A}^1$ of length $n+1$.

3. Prove that the dimension of a point is 0 and the dimension of $\mathbb{A}^1$ is 1.

4. Show that an irreducible affine variety is zero-dimensional if and only if it is a point.

5. Prove that the dimension of an irreducible plane curve is 1 and use this to show that the dimension of $\mathbb{A}^2$ is 2.

6. Write the ideal $\langle x^3 - x, x^2 - y\rangle$ as the intersection of two prime ideals. Describe the corresponding geometry.

7. Show that $f(x, y) = y^2 + x^2(x - 1)^2 \in \mathbb{R}[x, y]$ is an irreducible polynomial but that $V(f)$ is reducible.

8. Fix the hyperbola $H = V(xy - 5) \subset \mathbb{A}^2_{\mathbb{R}}$ and let C_t be the circle $x^2 + (y - t)^2 = 1$ for $t \in \mathbb{R}$.

   (a) Show that $H \cap C_t$ is zero-dimensional, for any choice of t.

   (b) Determine the number of points in $H \cap C_t$ (this number depends on t).

3.3 Rational functions

In algebraic geometry, we also use functions and maps between varieties which are not defined at all points of their domains. Workin with functions and maps not defined at all points is a special feature of algebraic geometry that sets it apart from other branches of geometry.

Suppose X is any irreducible affine variety. By Theorem 3.2.4, its ideal $\mathcal{I}(X)$ is prime, so its coordinate ring $\mathbb{F}[X]$ has no zero divisors ($0 \neq f, g \in \mathbb{F}[X]$ with $fg = 0$). A ring without zero divisors is called an *integral domain*. In exact analogy with the construction of the rational numbers $\mathbb{Q}$ as quotients of integers $\mathbb{Z}$, we may form the *function field* $\mathbb{F}(X)$ of X as the quotients of regular functions in $\mathbb{F}[X]$. Formally, $\mathbb{F}(X)$ is the collection of all quotients f/g with $f, g \in \mathbb{F}[X]$ and $g \neq 0$, where we identify

$$\frac{f_1}{g_1} = \frac{f_2}{g_2} \iff f_1 g_2 - f_2 g_1 = 0 \text{ in } \mathbb{F}[X]\,.$$

Example 3.3.1. The function field of affine space $\mathbb{A}^n$ is the collection of quotients of polynomials P/Q with $P, Q \in \mathbb{F}[x_1, \ldots, x_n]$. This field $\mathbb{F}(x_1, \ldots, x_n)$ is called the *field of rational functions* in the variables $x_1, \ldots, x_n$.

Given an irreducible affine variety $X \subset \mathbb{A}^n$, we may also express $\mathbb{F}(X)$ as the collection of quotients f/g of polynomials $f, g \in \mathbb{F}[\mathbb{A}^n]$ with $g \notin \mathcal{I}(X)$, where we identify

$$\frac{f_1}{g_1} = \frac{f_2}{g_2} \iff f_1 g_2 - f_2 g_1 \in \mathcal{I}(X)\,.$$

Rational functions on an affine variety X do not in general have unique representatives as quotients of polynomials or even quotients of regular functions.

Example 3.3.2. Let $X := \mathcal{V}(x^2 + y^2 + 2y) \subset \mathbb{A}^2$ be the circle of radius 1 and center at $(0, -1)$. In $\mathbb{F}(X)$ we have

$$-\frac{x}{y} = \frac{y^2 + 2y}{x}\,.$$

A point $x \in X$ is a *regular point* of a rational function $\varphi \in \mathbb{F}(X)$ if φ has a representative f/g with $f, g \in \mathbb{F}[X]$ and $g(x) \neq 0$. From this we see that all points of the neighborhood X_g of x in X are regular points of φ. Thus the set of regular points of φ is a nonempty Zariski open subset of X. Call this the *domain of regularity of* φ.

When $x \in X$ is a regular point of a rational function $\varphi \in \mathbb{F}(X)$, we set $\varphi(x) := f(x)/g(x) \in \mathbb{F}$, where φ has representative f/g with $g(x) \neq 0$. The value of $\varphi(x)$ does not depend upon the choice of representative f/g of φ. In this way, φ gives a function from a dense subset of X (its domain of regularity) to $\mathbb{F}$. We write this as

$$\varphi : X \dashrightarrow \mathbb{F}$$

with the dashed arrow indicating that φ is not necessarily defined at all points of X.

The rational function φ of Example 3.3.2 has domain of regularity $X - \{(0,0)\}$. Here $\varphi : X \dashrightarrow \mathbb{F}$ is stereographic projection of the circle onto the line $y = -1$ from the point $(0,0)$.

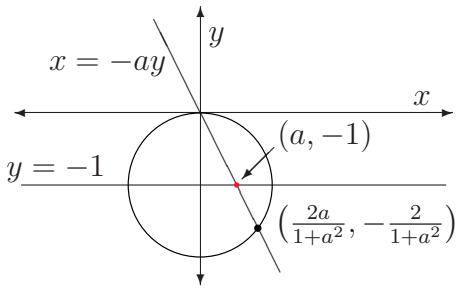


Figure 3.2: Projection of the circle $\mathcal{V}(x^2 + (y-1)^2 - 1)$ from the origin.

Example 3.3.3. Let $X = \mathbb{A}^1_{\mathbb{R}}$ and $\varphi = 1/(1+x^2) \in \mathbb{R}(X)$. Then every point of X is a regular point of φ. The existence of rational functions which are regular at every point, but are not elements of the coordinate ring, is a special feature of real algebraic geometry. Observe that φ is not regular at the points $\pm\sqrt{-1} \in \mathbb{A}^1_{\mathbb{C}}$.

Theorem 3.3.4. *When $\mathbb{F}$ is algebraically closed, a rational function that is regular at all points of an irreducible affine variety X is a regular function in $\mathbb{C}[X]$.*

Proof. For each point $x \in X$, there are regular functions $f_x, g_x \in \mathbb{F}[X]$ with $\varphi = f_x/g_x$ and $g_x(x) \neq 0$. Let $\mathcal{I}$ be the ideal generated by the regular functions g_x for $x \in X$. Then $\mathcal{V}(\mathcal{I}) = \emptyset$, as φ is regular at all points of X.

If we let $g_1, \ldots, g_s$ be generators of $\mathcal{I}$ and let $f_1, \ldots, f_s$ be regular functions such that $\varphi = f_i/g_i$ for each i. Then by the Weak Nullstellensatz for X (Theorem 1.3.4(3)), there are regular functions $h_1, \ldots, h_s \in \mathbb{C}[X]$ such that

$$1 \;=\; h_1 g_1 + \cdots + h_s g_s \,.$$

multiplying this equation by φ, we obtain

$$\varphi \;=\; h_1 f_1 + \cdots + h_s f_s \,,$$

which proves the theorem. □

A list $f_1, \ldots, f_m$ of rational functions gives a *rational map*

$$\begin{array}{rcl} \varphi : X & \dashrightarrow & \mathbb{A}^m \,, \\ x & \longmapsto & (f_1(x), \ldots, f_m(x)) \,. \end{array}$$

This rational map φ is only defined on the intersection U of the domains of regularity of each of the f_i. We call U the *domain of* φ and write $\varphi(X)$ for $\varphi(U)$.

Let X be an irreducible affine variety. Since $\mathbb{F}[X] \subset \mathbb{F}(X)$, any regular map is also a rational map. As with regular maps, a rational map $\varphi : X \dashrightarrow \mathbb{A}^m$ given by functions $f_1, \ldots, f_m \in \mathbb{F}(X)$ defines a homomorphism $\varphi^* : \mathbb{F}[\mathbb{A}^m] \to \mathbb{F}(X)$ by $\varphi^*(g) = g(f_1, \ldots, f_m)$. If Y is an affine subvariety of $\mathbb{A}^m$, then $\varphi(X) \subset Y$ if and only if $\varphi(\mathcal{I}(Y)) = 0$. In particular, the kernel J of the map $\varphi^* : \mathbb{F}[\mathbb{A}^m] \to \mathbb{F}(X)$ defines the smallest subvariety $Y = \mathcal{V}(J)$ containing $\varphi(X)$, that is, the Zariski closure of $\varphi(X)$. Since $\mathbb{F}(X)$ is a field, this kernel is a prime ideal, and so Y is irreducible.

When $\varphi : X \dashrightarrow Y$ is a rational map with $\varphi(X)$ dense in Y, then we say that φ is *dominant*. A dominant rational map $\varphi : X \dashrightarrow Y$ induces an embedding $\varphi^* : \mathbb{F}[Y] \hookrightarrow \mathbb{F}(X)$. Since Y is irreducible, this map extends to a map of function fields $\varphi^* : \mathbb{F}(Y) \to \mathbb{F}(X)$. Conversely, given a map $\psi : \mathbb{F}(Y) \to \mathbb{F}(X)$ of function fields, with $Y \subset \mathbb{A}^m$, we obtain a dominant rational map $\varphi : X \dashrightarrow Y$ given by the rational functions $\psi(x_1), \ldots, \psi(x_m) \in \mathbb{F}(X)$ where $x_1, \ldots, x_m$ are the coordinate functions on $Y \subset \mathbb{A}^m$.

Suppose we have two rational maps $\varphi\colon X \dashrightarrow Y$ and $\psi\colon Y \dashrightarrow Z$ with φ dominant. Then $\varphi(X)$ intersects the set of regular points of ψ, and so we may compose these maps $\psi \circ \varphi\colon X \dashrightarrow Z$. Two irreducible affine varieties X and Y are *birationally equivalent* if there is a rational map $\varphi : X \dashrightarrow Y$ with a rational inverse $\psi : Y \dashrightarrow X$. By this we mean that the compositions $\varphi \circ \psi$ and $\psi \circ \varphi$ are the identity maps on their respective domains. Equivalently, X and Y are birationally equivalent if and only if their function fields are isomorphic, if and only if they have isomorphic open subsets.

For example, the line $\mathbb{A}^1$ and the circle of Figure 3.2 are birationally equivalent. The inverse of stereographic projection from the circle to $\mathbb{A}^1$ is the map from $\mathbb{A}^1$ to the circle given by $a \mapsto (\frac{2a}{1+a^2}, -\frac{2}{1+a^2})$.

Exercises for Section 6

1. Show that irreducible affine varieties X and Y are birationally equivalent if and only if they have isomorphic open sets.

3.4 Maps of projective varieties

Some of the material in the section on projective varieties should be moved here.

3.5 Smooth and singular points

Given a polynomial $f \in \mathbb{F}[x_1, \ldots, x_n]$ and $a = (a_1, \ldots, a_n) \in \mathbb{A}^n$, we may write f as a polynomial in new variables $t = (x_1, \ldots, x_n)$, with $t_i = x_i - a_i$ and obtain

$$f \;=\; f(a) \;+\; \sum_{i=1}^{n} \tfrac{\partial f}{\partial x_i}(a) \cdot t_i \;+\; \cdots\,, \tag{3.3}$$

where the remaining terms have degrees greater than 1 in the variables t. When $\mathbb{F}$ has characteristic zero, this is the usual Taylor expansion of f at the point a. The coefficient of the monomial t^α is the mixed partial derivative of f evaluated at a,

$$\frac{1}{\alpha_1!\alpha_2!\cdots\alpha_n!}\left(\frac{\partial}{\partial x_1}\right)^{\alpha_1}\left(\frac{\partial}{\partial x_2}\right)^{\alpha_2}\cdots\left(\frac{\partial}{\partial x_n}\right)^{\alpha_n} f(a)\,.$$

In the coordinates t for $\mathbb{F}^m$, the linear term in the expansion (3.3) is a linear map

$$d_a f \;:\; \mathbb{F}^n \;\longrightarrow\; \mathbb{F}$$

called the *differential* of f at the point a.

Definition 3.5.1. Let $X \subset \mathbb{A}^n$ be a subvariety. The *(Zariski) tangent space* T_aX to X at the point $a \in X$ is the joint kernel[†] of the linear maps $\{d_a f \mid f \in \mathcal{I}(X)\}$. Since

$$\begin{aligned} d_a(f+g) &= d_a f + d_a g \\ d_a(fg) &= f(a)d_a g + g(a)d_a f \end{aligned}$$

we do not need all the polynomials in $\mathcal{I}(X)$ to define T_aX, but may instead take any finite generating set.

Theorem 3.5.2. *Let X be an affine variety. Then the set of points of X whose tangent space has minimal dimension is a Zariski open subset of X.*

Proof. Let $f_1, \ldots, f_m$ be generators of $\mathcal{I}(X)$. Let $M \in \mathrm{Mat}_{m\times n}(\mathbb{F}[\mathbb{A}^n])$ be the matrix whose entry in row i and column j is $\partial f_i/\partial x_j$. For $a \in \mathbb{A}^n$, the components of the vector-valued function

$$\begin{aligned} M \;:\; \mathbb{F}^n &\longrightarrow \mathbb{F}^m \\ t &\longmapsto M(a)t \end{aligned}$$

[†]Intersection of the nullspaces?

are the differentials $d_a f_1, \ldots, d_a f_m$.

For each $\ell = 1, 2, \ldots, \max\{n, m\}$, the *degeneracy locus* $\Delta_\ell \subset \mathbb{A}^m$ is the variety defined by all $\ell \times \ell$ subdeterminants (*minors*) of the matrix M, and set $\Delta_{1+\max\{n,m\}} := \mathbb{A}^n$. Since we may expand any $(i+1) \times (i+1)$ determinant along a row or column and express it in terms of $i \times i$ subdeterminants, these varieties are nested

$$\Delta_1 \ \subset \ \Delta_2 \ \subset \ \cdots \ \subset \ \Delta_{\max\{n,m\}} \ \subset \ \Delta_{1+\max\{n,m\}} \ = \ \mathbb{A}^n \, .$$

By definition, a point $a \in \mathbb{A}^n$ lies in $\Delta_{i+1} - \Delta_i$ if and only if the matrix $M(a)$ has rank exactly i. In particular, if $a \in \Delta_{i+1} - \Delta_i$, then the kernel of $M(a)$ has dimension $n - i$.

Let i be the minimal index with $X \subset \Delta_{i+1}$. Then

$$X - (X \cap \Delta_i) \ = \ \{a \in X \mid \dim T_a X = n - i\}$$

is a nonempty open subset of X and $n-i$ is the minimum dimension of a tangent space at a point of X. □

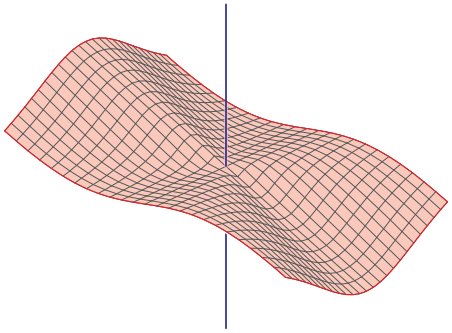
Suppose that X is irreducible and let m be the minimum dimension of a tangent space of X. Points $x \in X$ whose tangent space has this minimum dimension are call *smooth* and we write $X_{\rm sm}$ for the non-empty open subset of smooth points. The complement $X - X_{\rm sm}$ is the set $X_{\rm sing}$ of *singular* points of X. The set of smooth points is dense in X, for otherwise we may write the irreducible variety X as a union $\overline{X_{\rm sm}} \cup X_{\rm sing}$ of two proper closed subsets.

When X is irreducible, this minimum dimension of a tangent space is the dimension of X. This gives a second definition of dimension which is distinct from the combinatorial definition of Definition 3.2.

We have the following facts concerning the locus of smooth and singular points on a real or complex variety

Proposition 3.5.3. *The set of smooth points of an irreducible complex affine subvariety X of dimension d whose complex local dimension in the Euclidean topology is d is dense in the Euclidean topology.*

Example 3.5.4. Irreducible real algebraic varieties need not have this property. The Cartan umbrella $\mathcal{V}(z(x^2 + y^2) - x^3)$



is a connected irreducible surface in $\mathbb{A}^3_{\mathbb{R}}$ where the local dimension of its smooth points is either 1 (along the z axis) or 2 (along the 'canopy' of the umbrella).

3.6 Hilbert functions and degree

The homogeneous coordinate ring $\mathbb{F}[X]$ of a projective variety $X \subset \mathbb{P}^n$ is an invariant of the variety X which determines it up to a linear automorphism of $\mathbb{P}^n$. Basic numerical invariants, such as the dimension of X, are encoded in the combinatorics of $\mathbb{F}[X]$. Another invariant is the degree of X, which is the number of solutions on X to systems of linear equations.

The homogeneous coordinate ring $\mathbb{F}[X]$ of a projective variety $X \subset \mathbb{P}^n$ is a graded ring,

$$\mathbb{F}[X] \;=\; \bigoplus_{m=0}^{\infty} \mathbb{F}[X]_m \,,$$

with degree m piece $\mathbb{F}[X]_m$ equal to the quotient $\mathbb{F}[x_0, x_1, \ldots, x_n]_m / \mathcal{I}(X)_m$. The most basic numerical invariant of this ring is the *Hilbert function* of X, whose value at $m \in \mathbb{N}$ is the dimension of the m-th graded piece of $\mathbb{F}[X]$,

$$\mathrm{HF}_X(m) \;:=\; \dim_{\mathbb{F}}(\mathbb{F}[X]_m) \,.$$

This is also the number of linearly independent degree m homogeneous polynomials on X. We may also define the Hilbert function of a homogeneous ideal $I \subset \mathbb{F}[x_0, \ldots, x_n]$,

$$\mathrm{HF}_I(m) \;:=\; \dim_{\mathbb{F}}\big(\mathbb{F}[x_0, \ldots, x_n]_m / I_m\big) \,.$$

Note that $\mathrm{HF}_X = \mathrm{HF}_{\mathcal{I}(X)}$.

Example 3.6.1. Consider the space curve C of Figure 1.5. This is the image of $\mathbb{P}^1$ under the map

$$\varphi \;:\; \mathbb{P}^1 \ni [s,t] \;\longmapsto\; [s^3, s^2t, st^2, 2t^3 - 2s^2t] \in \mathbb{P}^3 \,.$$

If the coordinates of $\mathbb{P}^3$ are $[w,x,y,z]$, then $C = \mathcal{V}(2y^2{-}xz{-}2yw,\; 2xy{-}2xw{-}zw,\; x^2{-}yw)$. This map has the property that the pullback $\varphi^*(f)$ of a homogeneous form f of degree m is a homogenous polynomial of degree $3m$ in the variables s,t, and all homogeneous forms of degree $3m$ in s,t occur as pullbacks. Since there are $3m+1$ forms of degree $3m$ in s,t, we see that $\mathrm{HF}(m) = 3m+1$.

The Hilbert function of a homogeneous ideal I may be computed using Gröbner bases. First observe that any reduced Gröbner basis of I consists of homogeneous polynomials.

Theorem 3.6.2. *Any reduced Gröbner basis for a homogeneous ideal I consists of homogeneous polynomials.*

Proof. Buchberger's algorithm is friendly to homogeneous polynomials. If f and g are homogeneous, then so is $\mathrm{Spol}(f,g)$. Since Buchberger's algorithm consists of forming S-polynomials and of reductions (a form of S-polynomial), if given homogeneous generators of an ideal, it will compute a reduced Gröbner basis consisting of homogeneous polynomials.

A homogeneous ideal I has a finite generating set B consisting of homogeneous polynomials. Therefore, given a monomial order, Buchberger's algorithm will transform B into a reduced Gröbner basis G consisting of homogeneous polynomials. But reduced Gröbner bases are uniquely determined by the term order, so Buchberger's algorithm will transform any generating set into G. □

A consequence of Theorem 3.6.2 is that it is no loss of generality to use graded term orders when computing a Gröbner basis of a homogeneous ideal. Theorem 3.6.2 also implies that the linear isomorphism of Theorem 2.2.3 between $\mathbb{F}[x_0,\ldots,x_n]/I$ and $\mathbb{F}[x_0,\ldots,x_n]/\mathrm{in}(I)$ respects degree and so the Hilbert functions of I and of $\mathrm{in}(I)$ agree.

Corollary 3.6.3. *Let I be a homogeneous ideal. Then $\mathrm{HF}_I(m) = HF_{\mathrm{in}(I)}$, the number of standard monomials of degree m.*

Proof. The image in $\mathbb{F}[\mathbb{P}^n]/I$ of a standard monomial of degree m lies in the mth graded piece. Since the images of standard monomials are linearly independent, we only need to show that they span the degree m graded piece of this ring. Let $f \in \mathbb{F}[\mathbb{P}^n]$ be a homogeneous form of degree m and let G be a reduced Gröbner basis for I. Then the reduction f mod G is a linear combination of standard monomials. Each of these will have degree m as G consists of homogeneous polynomials and the division algorithm is homogeneous-friendly. □

Example 3.6.4. In the degree-reverse lexicographic monomial order where $x \succ y \succ z \succ w$, the polynomials

$$\underline{2y^2} - xz - 2yw\underline{2xy} - 2xw - zw,\ \underline{x^2} - yw,$$

form the reduced Gröbner basis for the ideal of the cubic space curve C of Example 3.6.1. As the underlined terms are the initial terms, the initial ideal is the monomial ideal $\langle y^2, xy, x^2\rangle$.

The standard monomials of degree m are exactly the set

$$\{z^a w^b,\ xz^c w^d,\ yz^c w^d \mid a+b=m,\ c+d=m-1\}$$

and so there are exactly $m+1+m+m = 3m$ standard monomials of degree m. This agrees with the Hilbert function of C, as computed in Example 3.6.1.

Thus we need only consider monomial ideals when studying Hilbert functions of arbitrary homogeneous ideals. Once again we see how some questions about arbitrary ideals may be reduced to the same questions about monomial ideals, where we may use combinatorics.

Because an ideal and its saturation both define the same projective scheme, and because Hilbert functions are difficult to compute, we introduce the Hilbert polynomial.

Definition 3.6.5. Two functions $f, g\colon \mathbb{N} \to \mathbb{N}$ are *stably equivalent*, $f \sim g$, if $f(d) = g(d)$ for d sufficiently large.

The following result may be proved purely combinatorially.†

Proposition-Definition 3.6.6. *The Hilbert function of a monomial ideal I is stably equivalent to a polynomial, HP_I, called the Hilbert polynomial of I.*

The degree of HP_I is the dimension of the largest linear subspace contained in $\mathcal{V}(I)$.

Corollary 3.6.7. *If I is a monomial ideal, then the Hilbert polynomials HP_I and $\mathrm{HP}_{\sqrt{I}}$ have the same degree.*

Together with Corollary 3.6.3, we may define the Hilbert polynomial of any homogeneous ideal I and the Hilbert polynomial HP_X of any projective variety $X \subset \mathbb{P}^n$. The Hilbert polynomial of a projective variety encodes virtually all of its numerical invariants. We explore two such invariants.

Definition 3.6.8. Let $X \subset \mathbb{P}^n$ be a projective variety and suppose that the initial term of its Hilbert polynomial is

$$\mathrm{in}(\mathrm{HP}_X(t)) \;=\; a\frac{t^d}{d!}\,.$$

Then the *dimension of X* is the degree, d, of the Hilbert polynomial and the number a is the *degree of X*.

We computed the Hilbert function of the curve C of Example 3.6.1 to be $3m+1$. This is also its Hilbert polynomial, as we see that C has dimension 1 and degree 3, which justifies our calling it a cubic space curve.

We may similarly define the dimension and degree of a homogeneous ideal I.

Example 3.6.9. In Exercise 4 you are asked to show that if X consists of a distinct points, then the Hilbert polynomial of X is the constant, a. Thus X has dimension 0 and degree a.

Suppose that X is a linear space, $\mathbb{P}(V)$, where $V \subset \mathbb{F}^{n+1}$ has dimension $d+1$. We may choose coordinates $x_0, \ldots, x_n$ on $\mathbb{P}^n$ so that V is defined by $x_{d+1} = \cdots = x_n = 0$, and so $\mathbb{F}[X] \simeq \mathbb{F}[x_0, \ldots, x_d]$. Then $\mathrm{HF}_X(m) = \binom{m+d}{d}$, which has initial term $\frac{m^d}{d!}$ and so X has dimension d and degree 1.

Suppose that $I = \langle f \rangle$, where f is homogeneous of degree a. Then

$$\big(\mathbb{F}[x_0, \ldots, x_n]/I\big)_m \;=\; \frac{\mathbb{F}[x_0, \ldots, x_n]_m}{f \cdot \mathbb{F}[x_0, \ldots, x_n]_{m-a}}\,,$$

so that if $m > 0$ we have $\mathrm{HF}_I(m) = \binom{m+n}{n} - \binom{m-a+n}{n}$. Thus the leading term of the Hilbert polynomial of I is $a\frac{m^{n-1}}{(n-1)!}$, and so I has dimension $n-1$ and degree a. When f is square-free, so that $I = \mathcal{I}(\mathcal{V}(f))$, we see that the hypersurface defined by f has dimension $n-1$ and degree equal to the degree of f.

†A proof is given in [20].

Theorem 3.6.10. *Let X be a subvariety of $\mathbb{P}^n$ and suppose that $f \in \mathbb{F}[X]_d$ has degree d and is not a zero divisor. Then the ideal $\langle \mathcal{I}(X), f\rangle$ has dimension $\dim(X)-1$ and degree $d \cdot \deg(X)$.*

Proof. For $m \geq d$, the degree m piece of the quotient ring $\mathbb{F}[x_0,\ldots,x_n]/\langle \mathcal{I}(X), f\rangle$ is the quotient

$$\mathbb{F}[X]_m / f \cdot \mathbb{F}[X]_{m-d}\,,$$

and so it has dimension $\dim_{\mathbb{F}}(\mathbb{F}[X]_m) - \dim_{\mathbb{F}}(\Lambda \cdot \mathbb{F}[X]_{m-d})$.

Suppose that m is large enough so that the Hilbert function of X is equal to its Hilbert polynomial at $m-d$ and all larger integers. Since f is not a zero divisor, multiplication by f is injective. Thus this dimension is

$$\mathrm{HP}_X(m) - \mathrm{HP}_X(m-d)\,.$$

which is a polynomial of degree $\dim(X)-1$ and leading coefficient $d\cdot\deg(X)/(\dim(X)-1)!$, as you are asked to show in Exercise 3. □

Lemma 3.6.11. *Suppose that X is a projective variety of dimension d and degree a. All subvarieties of X have dimension at most d and at least one irreducible component of X has dimension d, and a is the sum of the degrees of the irreducible components of dimension d.*

Furthermore, if X is irreducible, then every proper subvariety has dimension at most $d-1$ and X has a subvariety of dimension $d-1$.

Proof. Let Y be a subvariety of X. Then the coordinate ring of Y is a quotient of the coordinate ring of X, so $HF_Y(m) \leq HF_X(m)$ for all m, which shows that the degree of the Hilbert polynomial of Y is bounded above by the degree of the Hilbert polynomial of X.

Suppose that $X = X_1 \cup \cdots \cup X_r$ is the decomposition of X into irreducible components. Consider the map of graded vector spaces which is induced by restriction

$$\mathbb{F}[X] \;\longrightarrow\; \mathbb{F}[X_1] \oplus \mathbb{F}[X_2] \oplus \cdots \oplus \mathbb{F}[X_r]\,.$$

This is injective, which gives the inequality

$$\mathrm{HF}_X(m) \;\leq\; \sum_{i=1}^{r} \mathrm{HF}_{X_i}(m)\,.$$

Thus at least one irreducible component must have dimension d. Fill in the argument about the sum.

Suppose now that X is irreducible, let Y be a proper subvariety of X and let $0 \neq f \in \mathcal{I}(Y) \subset \mathbb{F}[X]$. Since $\mathbb{F}[X]/\langle f\rangle \twoheadrightarrow \mathbb{F}[X]/\mathcal{I}(Y) = \mathbb{F}[Y]$, we see that the Hilbert polynomial of $\mathbb{F}[Y]$ has degree at most that of $\mathbb{F}[X]/\langle f\rangle$, which is $d-1$.

Let $I = \langle \mathcal{I}(X), f \rangle$, where we write f both for the element $f \in \mathcal{I}(Y)$ and a homogeneous polynomial which restricts to it. If I is radical, then we have just shown that $\mathcal{V}(I) \subset X$ is a subvariety of dimension $d-1$. Otherwise, let $\succ$ be a monomial order, and we have the chain of inclusions

$$\mathrm{in}(I) \;\subset\; \mathrm{in}(\sqrt{I}) \;\subset\; \sqrt{\mathrm{in}(I)}\,, \tag{3.4}$$

and thus,

$$\deg(\mathrm{HP}_I) \;=\; \deg(\mathrm{HP}_{\mathrm{in}(I)}) \;\geq\; \deg(\mathrm{HP}_{\sqrt{I}}) \;\geq\; \deg(\mathrm{HP}_{\sqrt{\mathrm{in}(I)}})\,.$$

Since $\mathrm{HP}_{\mathrm{in}(I)}$ and $\mathrm{HP}_{\sqrt{\mathrm{in}(I)}}$ have the same degree and $\sqrt{I}$ is the ideal of $\mathcal{V}(I)$, we conclude that $\mathcal{V}(I)$ is a subvariety of X having dimension $d-1$. □

We may now show that the combinatorial definition (Definition 3.2) of dimension is correct.

Corollary 3.6.12 (Combinatorial definition of dimension)**.** *The dimension of a variety X is the length of the longest decreasing chain of irreducible subvarieties of X. If*

$$X \;\supset\; X_0 \;\supsetneq\; X_1 \;\supsetneq\; X_2 \;\supsetneq\; \cdots \;\supsetneq\; X_m \;\supsetneq\; \emptyset\,,$$

is such a chain of maximal length, then X has dimension m.

Proof. Suppose that

$$X \;\supset\; X_0 \;\supsetneq\; X_1 \;\supsetneq\; X_2 \;\supsetneq\; \cdots \;\supsetneq\; X_m \;\supsetneq\; \emptyset$$

is a chain of irreducible subvarieties of a variety X. By Lemma 3.6.11 $\dim(X_{i-1}) > \dim(X_i)$ for $i = 1, \ldots, m$, and so $\dim(X) \geq \dim(X_0) \geq m$.

For the other inequality, we may assume that X_0 is an irreducible component of X with $\dim(X) = \dim(X_0)$. Since X_0 has a subvariety X_1' with dimension $\dim(X_0) - 1$, we may let X_1 be an irreducible component of X' with the same dimension. In the same fashion, for each $i = 2, \ldots, \dim(X)$, we may construct an irreducible subvariety X_i of dimension $\dim(X)-i$. This gives a chain of irreducible subvarieties of X of length $\dim(X)+1$, which proves the combinatorial definition of dimension. □

A consequence of a Bertini's Theorem[†] is that if X is a projective variety, then for almost all homogeneous polynomials f of a fixed degree, $\langle \mathcal{I}(X), f \rangle$ is radical and f is not a zero-divisor in $\mathbb{F}[X]$.

Consequently, if Λ is a generic linear form and set $Y := \mathcal{V}(\Lambda) \cap X$, then $\mathcal{I}(Y) = \langle \mathcal{I}(X), \Lambda \rangle$, and so

$$\mathrm{HP}_Y \;=\; \mathrm{HP}_{\langle \mathcal{I}(X), \Lambda \rangle}\,,$$

and so by Theorem 3.6.10, $\deg(Y) = \deg(X)$. If $Y \subset \mathbb{P}^n$ has dimension d, then we say that Y has *codimension* $n - d$.

[†]Not formulated here, yet!

Corollary 3.6.13 (Geometric meaning of degree)**.** *The degree of a projective variety $X \subset \mathbb{P}^n$ of dimension d is the number of points in an intersection*

$$X \cap L\,,$$

where $L \subset \mathbb{P}^n$ is a generic linear subspaces of codimension d.

For example, the cubic curve of Figure 1.5 has degree 3, and we see in that figure that it meets the plane $z = 0$ in 3 points.

Exercises

1. Show that the dimension of the space $\mathbb{F}[x_0, \ldots, x_n]_m$ of homogeneous polynomials of degree m is $\binom{m+n}{n} = \frac{m^n}{n!} +$ lower order terms in m.

2. Let I be a homogeneous ideal. Show that $HF_I \sim HF_{(I\colon \mathfrak{m}_0)} \sim HF_{I_{\geq d}}$.

3. Suppose that $f(t)$ is a polynomial of degree d with initial term $a_0 t^d$. Show that $f(t) - f(t-1)$ has initial term $ma_0 t^{m-1}$. Show that $f(t) - f(t-b)$ has initial term $mba_0 t^{m-1}$.

4. Show that if $X \subset \mathbb{P}^n$ consists of a points, then, for m sufficiently large, we have $\mathbb{F}[X]_m \simeq \mathbb{F}^a$, and so $HP_X(t) = a$.

5. Compute the Hilbert functions and polynomials the following projective varieties. What are their dimensions and degrees?

   (a) The union of three skew lines in P^3, say $\mathcal{V}(x-w, y-z) \cup \mathcal{V}(x+w, y+z) \cup \mathcal{V}(y-w, x+z)$, whose ideal has reduced Gröbner basis

   $$\langle \underline{x^2} + y^2 - z^2 - w^2,\ \underline{y^2 z} - xz^2 - z^3 + xyw + yzw - zw^2,\ \underline{xyz} - y^2 w - xzw + yw^2,$$
   $$\underline{y^3} - yz^2 - y^2 w + z^2 w,\ \underline{xy^2} - xyw - yzw + zw^2 \rangle$$

   (b) The union of two coplanar lines and a third line not meeting the first two, say the x- and y-axes and the line $x = y = 1$.

   (c) The union of three lines where the first meets the second but not the third and the second meets the third. For example $\mathcal{V}(wy, wz, xz)$.

   (d) The union of three coincident lines, say the x-, y-, and z- axes.

Chapter 4

Basic concepts of real algebraic and semialgebraic geometry

Many applications of algebraic geometry deal – at least partially – with real solutions to polynomial equations. Depending on the type of question we ask, the problems become a quite different flavor. E.g., we might ask for (algorithmic) methods to analyze the real roots for the case of a given polynomial system (e.g., count them). A different type of question is to consider a whole class of problems with a finite number of complex solutions, and to ask how many solutions can be real.

In this chapter, we deal with some foundational material of real algebraic geometry. Our main focus is on the first of the two mentioned questions and on algorithmic aspects. At the end of the chapter, we discuss some aspects of the second question.

4.1 Real roots of univariate polynomials

We start by considering some classical results for univariate situations.

Let p be a univariate polynomial with real coefficients, i.e., $p \in \mathbb{R}[x]$. The *Sturm sequence* of p is the following sequence of polynomials of decreasing degree:

$$p_0(x) := p(x)\,, \quad p_1(x) := p'(x)\,, \quad p_i(x) := -\operatorname{rem}(p_{i-2}(x), p_{i-1}(x)) \text{ for } i \geq 2\,,$$

where rem denotes the remainder of a division with remainder. Let p_m be the last non-zero polynomial in the sequence.

Theorem 4.1.1. *(Sturm.) Let $p \in \mathbb{R}[x]$ and $a < b$ with $p(a), p(b) \neq 0$. Then the number of distinct real zeroes of p in the interval $[a, b]$ is the number of sign changes in the sequence $p_0(a), p_1(a), p_2(a), \ldots, p_m(a)$ minus the number of sign changes in the sequences $p_0(b), p_1(b), p_2(b), \ldots, p_m(b)$.*

Here, any zeroes are ignored when counting the number of sign changes in a sequence of real numbers. E.g., the sequence $+0+0-+0$ has two sign changes. Further note that

in the special case $m = 0$ the polynomial p is constant and thus due to $p(a)$, $p(b) \neq 0$ it has no roots.

In order to prove Sturm's Theorem, we concentrate on the case where all roots have multiplicity one. Let $N(x)$ be the number of sign changes at a point $x \in \mathbb{R}$.

Lemma 4.1.2. *For any $x \in \mathbb{R}$, the Sturm sequence cannot have two consecutive zeroes.*

Proof. By our assumption on the multiplicities, p_0 and p_1 cannot simultaneously vanish at x. Moreoever, inductively, if p_i and p_{i+1} both vanish at x then the division with remainder

$$p_{i-1} \;=\; s_i p_i - p_{i+1} \quad \text{with some polynomial } s_i$$

implies $p_{i-1}(x) = 0$ as well, contradicting the induction hypothesis. □

Proof of Sturm's Theorem. We imagine a left to right sweep on the real number line. By continuity of polynomial functions, it suffices to show that $N(x)$ decreases by 1 for a root of p and stays constant for a root of p_i, $i > 0$.

If $p(x) = 0$: If p switches from positive to negative then it is locally decreasing, so that the sequence of signs switches from $+-\ldots$ to $--\ldots$. If instead p switches from negative to positive then it is locally increasing, so that the sequence of signs switches from $-+\ldots$ to $++\ldots$.

If $p_i(x) = 0$ for some $i > 0$ (for $i \geq 2$ this might also happen at a zero of p): Assume that p_i switches from positive to negative (as before, the other case is analogous). Then by definition of p_{i+1}, the numbers $p_{i-1}(x)$ and $p_{i+1}(x)$ have opposite signs. So the sequence of sign switches either from $\ldots{+}{+}{-}\ldots$ to $\ldots{+}{-}{-}\ldots$ or from $\ldots{-}{+}{+}\ldots$ to $\ldots{-}{-}{+}\ldots$. In both cases, the number of sign changes remains invariant. Even *at* x, the pattern of signs is $\ldots{+}0{-}\ldots$ or $\ldots{-}0{+}\ldots$, so $N(x)$ is constant in the neighborhood of x. □

In order to count all real roots of a polynomial $p(x)$ we can apply Sturm's Theorem to $a = -\infty$ and $b = \infty$, which corresponds to looking at the signs of the leading coefficients of the polynomials p_i in the Sturm sequences. Using bisection, one can develop a procedure for isolating the real roots by rational intervals. This method is implemented, e.g., in Maple.

A second classical result for counting the number of real roots of a univariate polynomial is the Hermite form. Let $p \in \mathbb{R}[x]$ of degree n. Further, let $q \in \mathbb{R}[x]$ be a fixed polynomial, and let $H_q(p)$ be the symmetric $n \times n$-*Hankel matrix* defined by

$$(H_q(p))_{ij} \;=\; \sum_{k=1}^{n} q(x_k) x_k^{i+j-2}\,,$$

where $x_1, \ldots, x_n$ are the roots of p (over $\mathbb{C}$). Every symmetric matrix naturally defines a

quadratic form; here, we obtain

$$\begin{aligned} & z^T H_q(p) z \\ = & \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_{n-1} \end{pmatrix}^T \begin{pmatrix} \sum_{k=1}^n q(x_k) & \sum_{k=1}^n q(x_k)x_k & \cdots & \sum_{k=1}^n q(x_k)x_k^{n-1} \\ \sum_{k=1}^n q(x_k)x_k & \sum_{k=1}^n q(x_k)x_k^2 & \cdots & \sum_{k=1}^n q(x_k)x_k^n \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^n q(x_k)x_k^{n-1} & \sum_{k=1}^n q(x_k)x_k^n & \cdots & \sum_{k=1}^n q(x_k)x_k^{2n-2} \end{pmatrix} \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_{n-1} \end{pmatrix} \\ = & \sum_{k=1}^n q(x_k)(z_0 + z_1 x_k + \cdots + z_{n-1}x_k^{n-1})^2 \,. \end{aligned}$$

Denoting by V the Vandermonde matrix

$$V = \begin{pmatrix} 1 & x_1 & \cdots & x_1^{n-1} \\ 1 & x_2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^{n-1} \end{pmatrix},$$

we can write

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

Theorem 4.1.3. *The rank of $H_q(p)$ is equal to the number of distinct roots x_j of p for which $q(x_j) \neq 0$. The signature of $H_q(p)$ is equal to the number of distinct real roots x_j of p for which $q(x_j) > 0$ minus the number of real roots x_j of p for which $q(x_j) < 0$.*

Proof. Again, we first consider the case that all roots are distinct. Setting $z(x_k) := \sum_{i=0}^{n-1} z_i x_k^i$ we obtain

$$\begin{aligned} z^T H_q(p) z &= \sum_{k=1}^n q(x_k)(z_0 + z_1 x_k + \cdots + z_{n-1}x_k^{n-1})^2 \\ &= \sum_{k=1}^n q(x_k)(z(x_k))^2 \,. \end{aligned}$$

We write this quadratic form in x as

$$\begin{aligned} & z^T H_q(p) z \\ = & \sum_{x_k \in \mathbb{R}} q(x_k) z(x_k)^2 + \sum_{x_k, x_k^* \in \mathbb{C}\setminus\mathbb{R}} q(x_k) z(x_k)^2 + q(x_k^*) z(x_k^*)^2 \\ = & \sum_{x_k \in \mathbb{R}} q(x_k) z(x_k)^2 + 2 \sum_{x_k, x_k^* \in \mathbb{C}\setminus\mathbb{R}} \begin{pmatrix} \Re z(x_k) \\ \Im z(x_k) \end{pmatrix}^T \begin{pmatrix} \Re q(x_k) & -\Im q(x_k) \\ -\Im q(x_k) & -\Re q(x_k) \end{pmatrix} \begin{pmatrix} \Re z(x_k) \\ \Im z(x_k) \end{pmatrix}. \end{aligned}$$

Since the zeroes x_k are pairwise distinct, the polynomials $z(x_k)$ are linearly independent (by Vandermonde), and therefore also

$$\{z(x_k)\}_{x_k \in \mathbb{R}} \cup \{\Re z(x_k), \Im z(x_k)\}_{x_k, x_k^* \in \mathbb{C} \setminus \mathbb{R}} ,$$

which correspond to linear forms in $z_0, \ldots, z_{k-1}$. Hence, we have represented the quadratic form defined by $H_q(p)$ in a different basis. Due to the invariance of the signature under basis transformations we can compute the signature by adding the signatures of the scalar elements $q(x_k)$ and of the 2×2-blocks. The latter signatures are zero (because the trace is zero), which proves the claim.

For the general case, if $x_1, \ldots, x_s$ are the distinct roots with multiplicity $\mu(x_i)$, we have

$$z^T H_q(p) z \;=\; \sum_{k=1}^{s} \mu(x_k) q(x_k) (z(x_k))^2 ,$$

from which the statement follows analogously. □

In particular, for counting the number of roots choose $q(x) = 1$. The matrix corresponding to this quadratic form is

$$H_1(p) \;=\; \begin{pmatrix} n & s_1 & \cdots & s_{n-1} \\ s_1 & s_2 & \cdots & s_n \\ \vdots & \vdots & \ddots & \vdots \\ s_{n-1} & s_n & \cdots & s_{2n-2} \end{pmatrix} , \tag{4.1}$$

where $s_k = \sum_{i=1}^{n} x_i^k$ is the k-th *Newton sum* of p. The Newton sums can be expressed as polynomials in the coefficients a_i of $p = \sum_{i=0}^{n} a_i x^i$. Namely, the s_i and the a_j are related by *Newton's identities*

$$\begin{aligned} s_k + a_{n-1} s_{k-1} + \cdots + a_0 s_{k-n} &= 0 \qquad (k \geq n) , \\ s_k + a_{n-1} s_{k-1} + \cdots + a_{n-k+1} s_1 &= -k a_{n-k} \qquad (1 \leq k < n) . \end{aligned}$$

In particular, we obtain:

Corollary 4.1.4. *For a polynomial $p \in \mathbb{R}[x]$, all zeroes are real if and only if its associated matrix $H_1(p)$ is positive semidefinite.*

We consider another classical result:

Theorem 4.1.5. *(Déscarte's Rule of Signs.) The number of distinct positive real roots of a polynomial is at most the number of sign changes in its coefficient sequence.*

Proof. By induction on n. For $n = 1$, the statement is clear. Now assume that is already known for $n-1$, with $n > 1$. Let $p \in \mathbb{R}[x]$ be of degree n. We may assume that x does not divide p, so let p be of the form

$$p \;=\; \sum_{i=k}^{m} a_i x_i + a_0 \quad \text{with some } k \in \{1, \ldots, m\}$$

and $a_m, a_k, a_0 \neq 0$. Then $p' = \sum_{i=k}^{m} a_i i x^{i-1}$. Since the signs of the coefficients of p' coincide with the signs of the coefficients of p except a_0, the induction hypothesis implies that the number of sign changes in the coefficient sequence $a_n, \ldots, a_q$ bounds the number of positive roots of p'. Denote by x_0 the smallest positive root of p' (and set $x_0 = \infty$ if there is none). Then p' has the same sign in $(0, x_0)$ as a_k. Since $p(0) = a_0$, the polynomial p may have roots in $(0, x_0)$ only if $a_k a_0 < 0$, which is the case if the number of sign changes in $a_n, \ldots, a_0$ exceeds by 1 the number of sign changes in $a_n, \ldots, a_k$. Since between any two zeroes of p there must be a zero of p', this proves the statement. □

By replacing x by $-x$ in Déscarte's Rule, we obtain a bound on the number of negative real roots. In fact, both bounds are tight when all roots of p are real (see Theorem 4.2.3). In general, we have the following corollary to Déscarte's Rule.

Corollary 4.1.6. *A polynomial with m terms has at most $2m-1$ distinct real zeroes.*

This bound is optimal, as we see from the example

$$x \cdot \prod_{j=1}^{m-1} (x^2 - j)\,.$$

All $2m-1$ zeroes of this polynomial are real, and its expansion has m terms.

4.2 Real roots and the trace form

In the following let I be a zero-dimensional ideal in $\mathbb{C}[x_1, \ldots, x_n]$ generated by polynomials in $\mathbb{R}[x_1, \ldots, x_n]$. Further $R = \mathbb{C}[x_1, \ldots, x_n]$, and let $\mathcal{B}$ be a monomial basis of the coordinate ring R/I.

In generalization to the the previous section, for any polynomial $g \in R$, we can define the multiplication operation m_g by

$$\begin{aligned} R/I &\;\to\; R/I\,, \\ m_g([f]) &\;:=\; [g]\cdot[f] \;=\; [gf]\,. \end{aligned}$$

We fix a polynomial $q \in \mathbb{R}[x_1, \ldots, x_n]$ and construct the bilinear form T_q by

$$\begin{aligned} T_q : R/I \times R/I &\;\to\; \mathbb{R}\,, \\ (g, h) &\;\mapsto\; \mathrm{Tr}(m_{qgh})\,. \end{aligned}$$

T_q is called the *trace form* of q. Since I is generated by real polynomials, the representation matrix of the bilinear form is a symmetric real matrix, and hence its eigenvalues are real.

Recall that for a real quadratic form S, the *signature* $\sigma(S)$ is the number of positive eigenvalues minus the number of negative eigenvalues of its representing matrix. The *rank* $\rho(S)$ of S is the rank of the representing matrix.

Theorem 4.2.1. *For $q \in \mathbb{R}[x_1, \ldots, x_n]$, the signature and rank of the bilinear form T_q satisfy*

$$\begin{aligned} \sigma(T_q) &= \#\{a \in V(I) \cap \mathbb{R}^n \,:\, q(a) > 0\} - \#\{a \in V(I) \,:\, q(a) < 0\}\,, \\ \rho(T_q) &= \#\{a \in V(I) \,:\, q(a) \neq 0\}\,. \end{aligned}$$

Proof. Once more, for simplicity, we assume that all multiplicities are 1.

The entry (i, j) of the representing matrix M_q of T_q with respect to the monomial basis $\mathcal{B} = \{x^{\alpha(1)}, \ldots, x^{\alpha(d)}\}$ is

$$\mathrm{Tr}(m_{q \cdot x^{\alpha(i)} \cdot x^{\alpha(j)}})\,. \tag{4.2}$$

We will express (4.2) by the sum of the eigenvalues of T_q (or, equivalently, of M_q).

Let $f \in R$. By a slight generalization of Stickelberger's Theorem 2.5.3, the set of eigenvalues of m_f coincides with the set of values of f at the points in $\mathcal{V}(I)$. Let $p_1, \ldots, p_d$ be the points in I (which are distinct by our assumption). Hence, the sum of the eigenvalues of $m_{q \cdot x^{\alpha(i)} \cdot x^{\alpha(j)}}$ is

$$\sum_{p \in V(I)} q(p) p^{\alpha(i)} p^{(\alpha(j))}\,, \tag{4.3}$$

where in particular $p^{\alpha(i)}$ denotes the value of the monomial $x^{\alpha(i)}$ at the point p.

Similar to Theorem 4.1.3 we compute the signature in a different basis. Denoting by C the $d \times d$-matrix whose j-th column consists of the values $p_j^{\alpha(i)}$, $1 \leq i \leq d$, the expression (4.3) implies the decomposition

$$M_q \;=\; C D C^T\,,$$

where D is the diagonal matrix with entries $q(p_1), \ldots, q(p_d)$. In general C and D are complex matrices. However, the nonreal points occur in conjugate pairs, so the same arguments as in Theorem 4.1.3 can be applied to neglect these conjugate pairs. For the real points, the corresponding eigenvalues of T_q are

$$q(p) \qquad \text{for } p \in V(I) \cap \mathbb{R}^n\,,$$

which shows the claim. □

For the special case $q = 1$ we obtain:

Corollary 4.2.2. *The signature of T_1 yields the number of distinct real roots of I.*

For the special case $q = 1$ and $n = 1$, we can think of a principal ideal $I = \langle p \rangle$ with a univariate polynomial $p \in \mathbb{R}[x]$ of degree d. We set $\mathcal{B} = \{1, x, \ldots, x^{d-1}\}$. Then (4.3) implies that

$$(M_1)_{ij} \;=\; \sum_{p \in V(I)} p^{i-1} p^{j-1}$$

(in our univariate case this remains true for multiple roots). Thus we have recovered the Hankel matrix $H_1(p)$ from (4.1) containing the Newton sums of p.

In fact, the signature can be compute without actually determining the positive and negative eigenvalues.

Theorem 4.2.3. *Let A be a symmetric real matrix. Then the number of positive eigenvalues equals the number of sign changes in its characteristic polynomial $\chi_A(t)$.*

Proof. Let $p(t)$ be a real polynomial whose roots are all real. By Déscarte's rule, the number σ of positive eigenvalues is bounded by the number of sign changes in $p(t)$. Similarly, the number σ' of negative eigenvalues is bounded by the number of sign changes in $p(-t)$. Hence the total number of positive and negative eigenvalues is bounded by $\sigma + \sigma'$. Now $\sigma + \sigma' \leq n$ and the fact that all eigenvalues of a symmetric real matrix are real imply that the bound of Décarte's rule of signs holds with equality. □

4.3 Semialgebraic sets

A *semialgebraic set* in $\mathbb{R}^n$ is a subset of $\mathbb{R}^n$ satisfying a Boolean combination (i.e., using intersections, unions, and complements) of sets of the form

$$\{x \in \mathbb{R}^n \,:\, f(x) > 0\}$$

with $f \in \mathbb{R}[x_1, \ldots, x_n]$.

Example 4.3.1.

1. The semialgebraic subsets of $\mathbb{R}$ are the unions of finitely many points and open intervals.

2. An algebraic subset of $\mathbb{R}^n$ (defined by polynomial equations) is semialgebraic.

3. Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a polynomial mapping : $f = (f_1, \ldots, f_n)$, where $f_i \in \mathbb{R}[X_1, \ldots, X_n]$. Let A be a semialgebraic subset of $\mathbb{R}^n$. Then $f^{-1}(A)$ is a semialgebraic subset of $\mathbb{R}^m$.

4. If A is a semialgebraic subset of $\mathbb{R}^n$ and $L \subset \mathbb{R}^n$ a line, then $L \cap A$ is the union of finitely many points and open intervals.

5. If $A \subset \mathbb{R}^m$ and $B \subset \mathbb{R}^n$ are semialgebraic, then $A \times B$ is a semialgebraic subset of $\mathbb{R}^m \times \mathbb{R}^n$.

E.g., the condition that a matrix is positive semidefinite is a semialgebraic condition in the entries of the matrix.

Exercise 4.3.2. Let $Q = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_1 > 0 \ldots, x_n > 0\}$ be the interior of the positive orthant in $\mathbb{R}^n$. Show that Q cannot be written in the form

$$Q = \{x \in \mathbb{R}^n : p_1(x) > 0, \ldots, p_{n-1}(x) > 0\} \text{ with } p_1, \ldots, p_{n-1} \in \mathbb{R}[x_1, \ldots, x_n]\,.$$

We remark that by a deep Theorem of Bröcker [14] every semialgebraic set of the form $S = \{x \in \mathbb{R}^n : p_1(x) > 0, \ldots, p_k(x) > 0\}$ with $k \in \mathbb{N}$ can be written using at most n strict inequalities, i.e., in the form

$$S = \{x \in \mathbb{R}^n : p_1(x) > 0, \ldots, p_n(x) > 0\}\,.$$

There are various normal forms for the Boolean combinations in the definition of a semialgebraic sets. One of them is:

Exercise 4.3.3. Every semialgebraic set $S \subset \mathbb{R}^n$ can be written as a finite union of sets of the form

$$\{x \in \mathbb{R}^n : g(x) = 0 \text{ and } f_i(x) > 0,\ 1 \leq i \leq m\}\,.$$

Theorem 4.3.4 (Projection theorem.)**.** *Let $n \geq 2$, S be a semialgebraic set in $\mathbb{R}^n$ and $\pi : \mathbb{R}^n \to \mathbb{R}^{n-1}$ be the natural projection onto the first $n-1$ coordinates. Then $\pi(S)$ is semialgebraic.*

The statement can be deduced from the general principle of quantifier elimination (see Theorem A.3.4 in Appendix A.3).

Proof. We can write $\pi(S)$ in the form

$$\pi(S) = \{x \in \mathbb{R}^{n-1} : \exists x_n \text{ with } (x_1, \ldots, x_n) \in S\}\,.$$

By Exercise 4.3.3, we can assume that S can be represented as $S = S_1 \cup \cdots \cup S_m$ where each S_i is of the form

$$\{x \in \mathbb{R}^n : g = 0\} \cap \{x \in \mathbb{R}^n : f_i > 0,\ 1 \leq i \leq m\}$$

with $g, f_1, \ldots, f_m \in \mathbb{R}[x]$. First we consider the special case $i = 1$, i.e., $S = S_1$. Let $c = (c_1, \ldots, c_N)$ (for some N) be the sequence of all coefficients of the polynomials $g, f_1, \ldots, f_m$. Let $G, F_1, \ldots, F_m \in \mathbb{Z}[C_1, \ldots, C_N, x_1, \ldots, x_n]$ be the polynomials which result from $g, f_1, \ldots, f_m$ by replacing any c_i by some new indeterminate C_i. By applying quantifier elimination to the G and the F_i, we obtain polynomials $G_i, F_{ij} \in \mathbb{Z}[C_1, \ldots, C_N, x_1, \ldots, x_n]$ such that $\pi(S)$ is of the form $T_1, \ldots, T_l$ with

$$T_i = \{x \in \mathbb{R}^{n-1} : G_i(c, x) = 0 \text{ and } F_{ij}(c, x) > 0,\ 1 \leq j \leq r_i\}\,, \quad 1 \leq i \leq l\,.$$

Hence, $\pi(S)$ is defined semialgebraically by the real polynomials $G_i(c, x_1, \ldots, x_{n-1})$ and $F_{ij}(c, x_1, \ldots, x_{n-1})$.

In the case $S = S_1 \cup \cdots \cup S_m$ with $m > 1$, we can apply quantifier elimination to each of the S_i separately, then taking the union of the results. □

Let $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ be semialgebraic sets. A map $f : X \to Y$ is called *semialgebraic* if the graph of f,

$$\{(x, f(x)) \in X \times Y \, : x \in X\},$$

is a semialgebraic set in $\mathbb{R}^{n+m}$.

Theorem 4.3.5. *Let $f : X \to Y$ be a semialgebraic map. Then the image $f(X) \subset Y$ is a semialgebraic set.*

Proof. By definition, the graph of f is a semialgebraic set. Now the statement follows immediately from a (possibly repeated) application of the Projection Theorem 4.3.4. □

Exercises

1. Show that the composition of semialgebraic maps is semialgebraic.

2. Let $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ be semialgebraic sets.

   (a) Show that $g = (g_1, \ldots, g_m) : X \to Y$ is semialgebraic if and only if all the functions g_i are semialgebraic.

   (b) If $h : X \to Y$ is semialgebraic then $h^{-1}(Y)$ is a semialgebraic set.

3. Show that the set $\{(x, y) \in \mathbb{R}^2 \, : \, y = \lfloor x \rfloor$ or $(x \in \mathbb{Z}$ and $x \leq y \leq x + 1)\}$ ("infinite staircase") is not semialgebraic.

4.4 LMI-representable sets and spectrahedra

Let $\mathrm{Sym}_n(\mathbb{R})$ denote the set of real symmetric $n \times n$-matrices. $\succeq$ denotes positive semidefiniteness ...

Let $A_0, \ldots, A_n \in \mathrm{Sym}_k(\mathbb{R})$ be real symmetric matrices. We consider the *linear matrix polynomial*

$$A(x) \; := \; A_0 + \sum_{i=1}^{n} x_i A_i \, .$$

Then the set

$$S \; := \; \{x \in \mathbb{R}^n \, : \, A(x) \succeq 0\}$$

is called a *spectrahedron*. The inequality $A_0 + \sum_{i=1}^n x_i A_i \; \succeq 0$ is called a *linear matrix inequality (LMI)*.

E.g., if all matrices A_i are diagonal then for all $x \in \mathbb{R}^n$ the matrix $A(x)$ is a diagonal matrix and thus S is a polyhedron.

Example 4.4.1. The unit disc $\{x \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 1\}$ is a spectrahedron. This follows from setting

$$A_1 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A_2 := \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad A_3 := \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

and observing that

$$A(x) \;=\; \begin{pmatrix} 1 + x_1 & x_2 \\ x_2 & 1 - x_1 \end{pmatrix}$$

is positive semidefinite if and only if $1 - x_1^2 - x_2^2 \geq 0$.

Linearity of the operator $A(\cdot)$ immediately implies that any spectrahedron is convex. Moreover, any spectrahedron S is a basic closed semialgebraic sets. This can be seen by writing $S \;=\; \{x \in \mathbb{R}^n \,:\, p_i(x) \geq 0\,, i \in I\}$ where the $p_i(x)$ are the principal minors of $A(x)$. A slightly more concise representation is given by the following statement.

Theorem 4.4.2. *Any spectrahedron S is a basic closed semialgebraic set. In particular, given the modified characteristic polynomial*

$$t \mapsto \det(F(x) + tI_k) \;=: t^k + \sum_{i=0}^{k-1} p_i(x) t^i\,,$$

S has the representation $S \;=\; \{x \in \mathbb{R}^n \,:\, p_i(x) \geq 0,\, 1 \leq i \leq k\}$.

Proof. Denoting by $\lambda_1(x), \ldots, \lambda_k(x)$ the eigenvalues of the linear pencil $A(x)$, we observe

$$\det(F(x) + tI_k) \;=\; (t + \lambda_1(x)) \cdots (t + \lambda_k(x))\,.$$

Since $A(x)$ is symmetric all $\lambda_i(x)$ are real, for any $x \in \mathbb{R}^n$. Comparing the coefficients we have

$$p_{k-i} \;=\; \sum_{t_1 < \cdots < t_i} \lambda_{t_1}(x) \cdots \lambda_{t_i}(x)\,, \quad 1 \leq i \leq k$$

with $p_m(x) = 1$.

Now "$\subset$" of the desired representation follows from the fact that positive semidefiniteness of $F(x)$ implies non-negativity of all eigenvalues. Conversely, if $p_i(x) \geq 0$ for all i, the determinant polynomial $\det(F(x) + tI_k)$ has no sign changes. By Descartes' rule of signs this implies that $F(x)$ is positive semidefinite.[1] □

In the following we discuss some geometric properties of spectrahedra. Let C be a closed convex subset of $\mathbb{R}^n$ with non-empty interior. A face of C is a convex subset F such that whenever $a, b \in C$ and $ta + (1 - \lambda)b$ for some $\lambda \in (0, 1)$ we have $a, b \in F$. A *supporting hyperplane* of C is an affine hyperplane H in $\mathbb{R}^n$ such that $C \cap H \neq \emptyset$ and $C \setminus H$ is connected. A face F of C is *exposed* if either $F = S$ or there exists a supporting

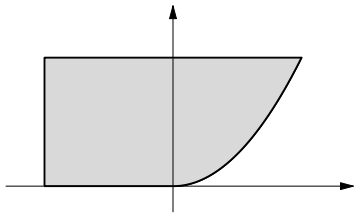


Figure 4.1: The origin is a non-exposed face of the shaded set S, since there does not exist a supporting hyperplane to S which intersects S only in the origin.

hyperplane H of S such that $H \cap S = F$. We also say that the hyperplane H *exposes* F. See Figure 4.1 for an example.

Every point $x \in C$ is contained in the relative interior of a unique face of C. We denote this face by $F_C(x)$ and observe

$$\begin{aligned} \operatorname{aff} F_C(x) &= \{x + y \,:\, \exists \lambda > 0 \text{ with } x - \lambda y, x + \lambda y \in C\}\,, && (4.4)\\ F_C(x) &= \operatorname{aff} F_C(x) \cap C\,. && (4.5) \end{aligned}$$

The following basic investigations on the faces of spectrahedra are mainly due to Ramana and Goldman [73].

Lemma 4.4.3. *Let $A(x)$ be a linear matrix polynomial of dimension k (i.e., defined by $k \times k$-matrices) and*

$$S \;=\; \{x \in \mathbb{R}^n \,:\, A(x) = A_0 + x_1 A_1 + \cdots + x_n A_n \succeq 0\}\,.$$

Then for any point $x \in S$

$$\operatorname{aff} F_S(x) \;=\; \{z \in \mathbb{R}^n \,:\, \ker A(z) \supset \ker A(x)\}\,.$$

We begin with an auxiliary result, where $A \pm B \succeq 0$ means that both $A+B$ and $A-B$ are positive semidefinite.

Lemma 4.4.4. *For $A, B \in \operatorname{Sym}_k$ and $A \succeq 0$ the following statements are equivalent.*

1. *$A \pm B \succeq 0$;*
2. *$\ker A \subset \ker B$, and for any $w \in (\ker A)^\perp$ we have $w^T(A \pm B)w \geq 0$.*

Proof. Let $A \pm B \succeq 0$, and for any $x \in \mathbb{R}^n$ consider the decomposition $x = v + w$ with $v \in (\ker A)$ and $w \in (\ker A)^\perp$. By considering vectors with $w = 0$ the positive semidefiniteness of $A \pm B$ implies $v^T B v = 0$. In order to show $\ker A \subset \ker B$, it suffices to show $w^T B v = 0$ for all $w \in \mathbb{R}^n$ and $v \in \ker A$. However, this follows from the observations that the

[1] connext appropriately ... pos vs. neg. zeros

directional derivative D_w of the function $f(x) = x^T Bx$ in v is $2w^T Bv = 2w^T(A+B)v \geq 0$ by the positive semidefiniteness of $A+B$ and $2w^T Bv = 2w^T(B-A)v \leq 0$ by the positive semidefinitess of $A-B$. Moreover, the second part of ii) is clear.

Conversely, with the same decomposition $x = v + w$, we have $x^T Ax = w^T Aw$ and due to $\ker A \subset \ker B$ also $x^T Bx = w^T Bw$. Hence, via the second part of ii) we obtain i). □

Proof. (of Lemma 4.4.3). Let $A^h(x) := A(x) - A_0$ be the homogeneous part of $A(x)$.

First assume $z \in \operatorname{aff} F_S(x)$, and define $y := z - x$. By (4.4) there exists some $\mu > 0$ with $x - \lambda y, x + \lambda y \in S$, i.e., $A(x) - A^h(\mu y) \succeq 0$ and $A(x) + A^h(\mu y) \succeq 0$. By Lemma 4.4.4 this implies $\ker A^h(\mu y) \supset \ker A(x)$, and due to $\ker A^h(y) = \ker A(z)$ and the homogeneity of A^h also $\ker A(z) \supset \ker A(\mu y)$.

Conversely, let $z \in \mathbb{R}^n$ with $\ker A(z) \supset \ker A(x)$. Using as before $y := z - x$, we can assume $y \neq 0$. The first goal is to construct some $\mu > 0$ such that

$$w^T A(x) w \;\succeq\; \mu w^T A^h(\mu y) w \;\succeq\; -w^T A(x) w \tag{4.6}$$

for all $w \in (\ker A(x))$. If $A^h(y) = 0$, choosing $\mu = 1$ does the job because $A(x) \succeq 0$. Otherwise, due to $\ker A(x) \subset \ker A(z) = \ker A^h(y)$ the matrix $A(x)$ cannot be the zero matrix. Let λ be the smallest positive eigenvalue of $A(x)$ and ρ be the spectral radius of $A^h(y)$. Then for any unit vector $v \in (\ker A(x))^\perp$ we have $v^T A(x) v \geq \lambda$ and $|v^T A^h(\mu y) v| \leq \rho$, which guarantees (4.6). Since by the homogeneity of A^h we also have $\ker A(z) \subset \ker A^h(\mu y)$, applying Lemma 4.4.4 implies

$$A(x) - \mu A^h(\mu y) \succeq 0 \text{ and } A(x) + \mu A^h(\mu y) \succeq 0\,.$$

Hence, $x + \mu y$, $x - \mu y \in S$, and by (4.4) further $z = x + y \in \operatorname{aff} F_S(x)$. □

Theorem 4.4.5. *Any face of a spectrahedron is exposed.*

Proof. (of Theorem 4.4.5). Let F be a face of a spectrahedron $S = \{x \in \mathbb{R}^n : A(x) \succeq 0\}$, and let y be a point in the relative interior of F. We can assume $y = 0$.

By considering the spectral decomposition $A_0 = U^T DU$ of A_0 with a diagonal matrix D and setting D' as the submatrix of D corresponding to the nonzero eigenvalues, we can further assume that $A(x)$ is of the form

$$A(x) \;=\; \begin{pmatrix} D + U(x) & B(x)^T \\ B(x) & C(x) \end{pmatrix} \tag{4.7}$$

with some linear matrix polynomials $C(x)$ of size j and $D \succ 0$. Note that in the case $j = 0$, the point $y = 0$ is an interior point of F, and hence F is exposed. So let $j > 0$.

Since $\ker A(0) = \{x \in \mathbb{R}^n : x_1 = \cdots = x_{n-j} = 0\}$, Lemma 4.4.3 implies $\operatorname{aff} F = \{x \in \mathbb{R}^n : B(x) = 0, C(x) = 0\}$. In order to construct a supporting hyperplane $H = \{x \in \mathbb{R}^n : a^T x = 0\}$ with $H \cap S = F$, we set $a_i = \operatorname{Tr}(C_i)$ for $1 \leq i \leq m$. To see that any $x \in F$ is contained in H and thus in $H \cap S$, note that $a^T x = \sum \operatorname{Tr}(C_i) x_i = \operatorname{Tr}(C(x)) = 0$. Conversely, to see that any $x \in H \subset S$ is also contained in F, note that from $C(x) \succeq 0$ and $0 = a^T x = 0 = \operatorname{Tr} C(x)$ we can deduce $C(x) = 0$ as well as $B(x) = 0$ and thus $x \in F$.

Since in the case $a = 0$, F would coincide with S, we can conclude $a \neq 0$ and thus F is an exposed face. □

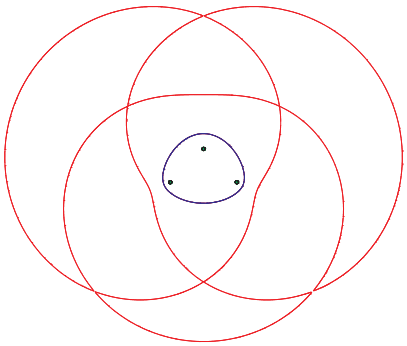


Figure 4.2: The closed curve in the center is a 3-ellipse of the three depicted points.

4.4.1 Rigid convexity

In the following we assume that there exists a point x_0 in the interior of the set $\mathcal{F}$, *i.e.*, $x_0 \in \operatorname{int} \mathcal{F}$. Any line through x_0 intersects the boundary of $\mathcal{F}$ if the determinant $\det F(x) = p_0(x)$ vanishes.

...

Wie oben bereits erwähnt sind die Zulässigkeitsbereiche semidefiniter Programme konvexe, semialgebraische Mengen. Viele der Fragen im Zusammenspiel zwischen semidefiniter Optimierung und semialgebraischer Geometrie sind daher eng verknüpft mit der Frage, welche semialgebraischen Mengen *semidefinit repräsentierbar sind*, das heisst, als Zulässigkeitsbereich eines semidefiniten Programms dargestellt werden können.

Definition 4.4.6. A closed subset $\mathcal{C} \subset \mathbb{R}^n$ is called an algebraic interior if there exists a polynomial $p \in \mathbb{R}[X_1, \ldots, X_n]$ such that $\mathcal{C}$ is the closure of a connected component of the positivity domanin $\{x \in \mathbb{R}^n : p(x) > 0\}$ of p.

For a given algebraic interior C the defining polynomial of $\mathcal{C}$ of minimal degree is unique up to a positive multiple.

Helton and Vinnikov have shown the following geometric characterization [37].

Theorem 4.4.7. *If a closed, convex set $C \subseteq \mathbb{R}^n$ can be semidefinitely represented, then C is* rigidly convex*, that is, for each point z in the interior of C and each generic line ℓ through z the line ℓ intersects the real algebraic hypersurface $p(x) = 0$ of degree d in exactly d points.*

For the case of dimension 2 the converse is true as well [37].

Theorem 4.4.8. *If an algebraic interior $C \subseteq \mathbb{R}^2$ is rigidly convex with a defining polynomial of degree d, then C is semidefinite representable.*

We illustrate these statements by two examples.

Example 4.4.9. For given k points $(a_1, b_1)^T, \ldots, (a_k, b_k)^T \in \mathbb{R}^2$ the *k-ellipse* with focal points $(a_i, b_i)^T$ and radius d is the plane curve $\mathcal{E}_k$ given by

$$\left\{(x, y)^T \in \mathbb{R}^2 \,:\, \sum_{i=1}^{k} \sqrt{(x - a_i)^2 + (y - b_i)^2} = d\right\} \tag{4.8}$$

(see Figure 4.2). For the special case $k = 2$ we obtain usual ellipses. The convex hull C of a k-ellipse $\mathcal{E}_k$ is a semidefinitely representable set in $\mathbb{R}^2$ (see [62]). In order to see this for the example in Figure 4.2, consider the Zariski closure $\mathcal{E}_3'$ of the set defined by (4.8); the real points hereof are depicted in the Figure. Actually the curve $\mathcal{E}_3'$ is of degree 8. Considering now an arbitrary point z in the interior of the 3-Ellipse, then each generic (not passing through a point of higher multiplicity of the curve) line through z contains exactly 8 points of $\mathcal{E}_3'$. By Theorem 4.4.8 this property implies semidefinite representability.

Example 4.4.10. Let p be the irreducible polynomial

$$p(x, y) \;=\; x^3 - 3xy^2 - (x^2 + y^2)^2$$

(see Figure 4.3). The positivity domain consists of three bounded connected components, as can be seen in the figure. We consider the bounded component C in the right half-plane, which is given by the topological closure

$$\text{cl}\,\left\{(x, y)^T \in \mathbb{R}^2 \,:\, p(x, y) > 0,\, x > 0\right\}.$$

Let a be a fixed point in the the interior of this component, for example $a = (1/2, 0)^T$. There exists an open set of lines through a which intersects the real zero set $V_{\mathbb{R}}(p)$ in only two points. Thus C is not semidefinite representable.

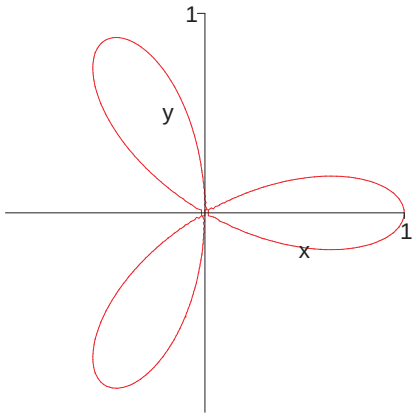


Figure 4.3: The real variety $V_{\mathbb{R}}(p)$ of the polynomial p.

For the case of general dimension ($n \geq 3$) no generalization of the exact geometric characterization of positive semidefinite representable sets is known so far.

Spectrahedra are rigidly convex.

real zero polynomials

Theorem 4.4.11 (Helton, Vinnikov)**.** *Every spectrahedron is rigidly convex. Conversely, every rigidly convex set in* $\mathbb{R}^2$ *is a spectrahedron.*

In fact, Theorem **??** is a special case of the following situation of rigidly convex sets.

Theorem 4.4.12 (Netzer, Plaumann, Schweighofer [61])**.** *The faces of a rigidly convex set are exposed. Since every spectrahedron is rigidly convex, every face of a spectrahedron set is exposed.*

Exercises

Exercise 4.4.13. Show that the intersection of two spectrahedra is again a spectrahedron.

Exercise 4.4.14. The property of spectrahedrality is unaltered under bijective affine transformations. It is also unaltered when a nonsingular congruence transformation is applied to the matrix map: i.e., if V is a nonsingular matrix, then $G = \{x : V^T Q(x) V \succeq 0\}$ is a spectrahedron.

Exercise 4.4.15. Convex quadratic inequalities give spectrahedra: Let $f(x) = x^T L^T L x + b^T x + c$. Then $f(x) \leq 0$ if and only if

$$\begin{pmatrix} -(b^T x + c) & x^T L^T \\ Lx & I \end{pmatrix} \succeq 0 .$$

4.5 Semidefinitely representable sets

A set $S \subseteq \mathbb{R}^n$ is called semidefinitely representable if it can be written as a projection of a spectrahedron.

The following question has been asked by Nemirovski and by Helton and Nie (Sources ...): Is every convex semi-algebraic set semidefinitely representable?

Let $\mathbb{R}[X]_d$ denote the set of polynomials of total degree at most d.

Theorem 4.5.1. *Let* $S \subseteq \mathbb{R}^n$ *be semidefinitely representable. Then*

$$S^\circ = \{\ell \in \mathbb{R}[X]_1 : \ell \geq 0 \text{ on } S\}$$

is also semidefinitely representable.

Proof. We concentrate on the case where S is strictly feasible and defined by a strictly feasible matrix polynomial

$$\mathcal{A}(X) = A + X_1 B_1 + \cdots + X_n B_n .$$

with symmetric $k \times k$-matrices $A, B_1, \ldots, B_n$; here, strictly feasible means that there exists a The general case can be reduced to that situation (see [30]).

In this strict situation, we can rename the variables and assume that $\mathcal{A}$ is of the form

$$\mathcal{A}(X,Y) \;=\; A + \sum_{i=1}^{n} X_i B_i + \sum_{i=1}^{n} Y_i C_i \,,$$

and $S = \{x \in \mathbb{R}^n \,:\, \exists y \in \mathbb{R}^m \, \mathcal{A}(x,y) \succeq 0\}$. Then, by applying a separation theorem, Nemirovski has shown that

$$\begin{aligned} S^\circ \;=\; \Big\{ l_0 + \sum_{i=1}^{n} l_i X_i \;\;:\;\; & \exists U \in \mathrm{Sym}_k(\mathbb{R}) \,:\, U \succeq 0,\, U \bullet A \le l_0, \\ & U \circ B_i = l_i \text{ for } 1 \le i \le n,\; U \circ C_j = 0 \text{ for } 1 \le j \le m \Big\} \,, \end{aligned}$$

which proves the theorem. □

Theorem 4.5.2. *The topological closure of a semidefinitely representable set is semidefinitely representable.*

Proof. Let S be semidefinitely representable. By Theorem 4.5.1 the set $(S^\circ)^\circ$ is semidefinitely representable, and by classical convex geometry we have $(S^\circ)^\circ = \overline{S}$. □

The conic hull of an sdr set is sdr.
The convex hull of a finite union of sdr is sdr.
Helton and Nie.

We close our discussion on methods for treating real roots by pointing out that this covered only a short glimpse of relevant aspects. In particular, throughout our discussion we always started from the situation of a given system and analyed the real roots of the system (in particular, counted them). A different viewpoint is to consider problem classes with a finite number of complex solutions (enumerative problems), and to ask how many solutions can be real.

An interesting class considered by Sottile is the *special Schubert calculus*. This special Schubert calculus asks for linear subspaces of a fixed dimension meeting some given (general) linear subspaces (whose dimensions and number ensure a finite number of solutions) in n-dimensional complex projective space $\mathbb{P}^n$. For any given dimensions of the subspaces, this problem is fully real, i.e., there exist *real* linear subspaces for which each of the a priori complex solutions is *real*. In particular, for $1 \le k \le n-2$ there are $d_{k,n} := (k+1)(n-k)$ real $(n-k-1)$-planes $U_1, \ldots, U_{d_{k,n}}$ in $\mathbb{P}^n$ with

$$\#_{k,n} := \frac{1!2! \cdots k!((k+1)(n-k))!}{(n-k)!(n-k+1)! \cdots n!}$$

real k-planes meeting $U_1, \ldots, U_{d_{k,n}}$. Here, $d_{k,n}$ and $\#_{k,n}$ are the dimension and the degree of the Grassmannian $\mathbb{G}_{k,n}$, respectively.

The simplest case of this type is the classical problem of common transversals to four lines in space. Let ℓ_1, ℓ_2, ℓ_3, and ℓ_4 be lines in general position in real 3-space. Then there

are two (in general complex) lines passing through $\ell_1, \ldots, \ell_4$, and there are configurations where both solution lines are real.

This can be seen as follows. The three mutually skew lines ℓ_1, ℓ_2, and ℓ_3 lie in one ruling of a doubly-ruled hyperboloid (see Figure 4.4). This is either (i) a hyperboloid of

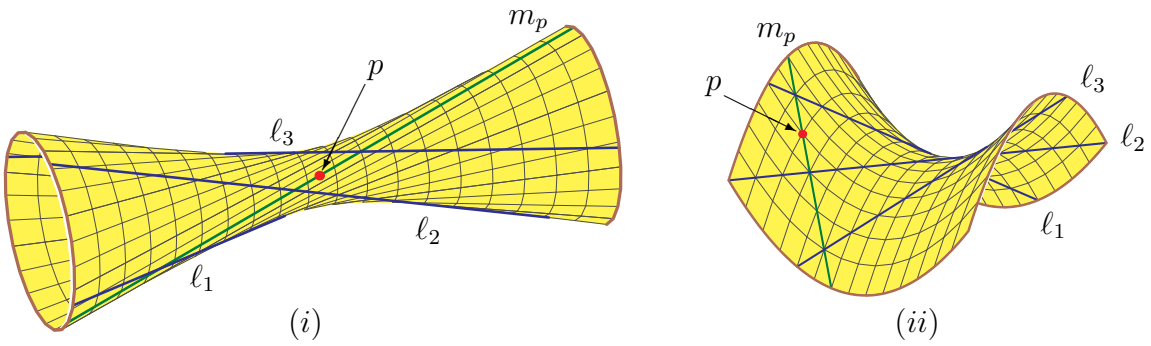


Figure 4.4: Hyperboloids through 3 lines.

one sheet, or (ii) a hyperbolic paraboloid. The line transversals to ℓ_1, ℓ_2, and ℓ_3 constitute the second ruling. Through every point p of the hyperboloid there is a unique line m_p in the second ruling which meets the lines ℓ_1, ℓ_2, and ℓ_3.

The hyperboloid is defined by a quadratic polynomial and so the fourth line ℓ_4 will either meet the hyperboloid in two points or it will miss the hyperboloid. In the first case there will be two real transversals to ℓ_1, ℓ_2, ℓ_3, and ℓ_4, and in the second case there will be no real transversal.

A related, recently well studied class of this type comes from nonlinear computational geometry. Sottile and Theobald showed that $2n-2$ general spheres in affine real space $\mathbb{R}^n$ have at most $3 \cdot 2^{n-1}$ common tangent lines in $\mathbb{C}^n$, and that there exist spheres for which all the a priori complex tangent lines are real.

The following construction (by Macdonald, Pach, and Theobald) illustrates this situation in dimension 3: Suppose that the spheres have equal radii, r, and have centers at the vertices of a regular tetrahedron with side length $2\sqrt{2}$,

$$(2,2,0)^T, \qquad (2,0,2)^T, \qquad (0,2,2)^T, \quad \text{and} \quad (0,0,0)^T.$$

There are real common tangents only if $\sqrt{2} \leq r \leq 3/2$, and exactly 12 when the inequality is strict. Note that in this case the spheres are non-disjoint. It is an open question whether it is possible for four disjoint *unit* spheres in $\mathbb{R}^3$ to have 12 common tangents.

If the spheres are unit spheres and the centers are coplanar, then Megyesi showed that the maximal number of solutions goes down to 8.

Macdonald, Pach, and Theobald also addressed the question of degenerate configurations of spheres.

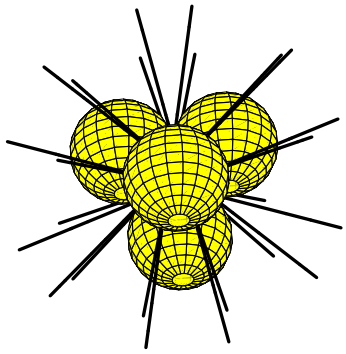


Figure 4.5: Four spheres with equal radii and 12 common tangents.

Theorem 4.5.3. *Four degenerate spheres in* $\mathbb{R}^3$ *of equal radii have colinear centers.*

This result was recently extended by by Borcea, Goaoc, Lazard, and Petitjean.

Theorem 4.5.4. *Four degenerate spheres in* $\mathbb{R}^3$ *have colinear centers.*

4.6 Notes

The term spectrahedron was introduced by Ramana and Goldman [73] who also showed exposedness of the faces. The facial structure of LMI-representable sets was studied by Netzer, Plaumann and Schweighofer [61]. For further information on eigenvalue techniques see [21, 63].

The lecture notes of Sturmfels [90] captures many of the start-of-the-art techniques and results.

For the preview on the nonlinear computational geometry in Chapter 11 see the surveys [87] and [88].

Chapter 5

Algebraic certificates for positive polynomials

We consider the question of when a given real polynomial $p \in \mathbb{R}[x] = \mathbb{R}[x_1, \ldots, x_n]$ is positive (in the sense of p being strictly positive or nonnegative on $\mathbb{R}^n$). This was a topic of discussion during Minkowski's public defense of his Ph.D. thesis on quadratic forms in Königsberg in 1885. Minkowski conjectured that there exist nonnegative real forms (real homogeneous polynomials) which cannot be written as a sum of squares of real forms. Hilbert, whose task as an 'opponent' was to attack the thesis, at the end of the defense attested that he was convinced by Minkowski's exposition that already in the ternary case $n = 3$ there may well be such remarkable forms "which are so stubborn as to remain positive without allowing themselves to submit a representation as sums of squares of forms"[1]. Hilbert proved Minkowski's conjecture in 1888 (see Theorem 5.2.3).

Among the 23 problems which Hilbert presented during his legendary lecture at the 1900 International Congress of Mathematicians in Paris, he asked whether every nonnegative polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]$ can be written as a finite sum of squares of real rational functions. This problem is widely known as *Hilbert's 17th problem*.

As we will see in this chapter and the next, this classical theory and its more recent developments are cornerstones of modern treatments of optimization of polynomial functions. Via the duality theory of optimization, algebraic certificates are of particular importance—they serve to "witness" that a certain polynomial is actually positive on a given set. The set of witnesses for global positivity is the set $\Sigma[x] = \Sigma[x_1, \ldots, x_n]$ of polynomials which can be written as a sum of squares of polynomials

$$\Sigma[x] \;:=\; \Big\{ \sum_{i=1}^{k} h_i^2 \text{ with } h_1, \ldots, h_k \in \mathbb{R}[x],\, k \in \mathbb{N} \Big\}\,. \tag{5.1}$$

In this chapter we provide the key elements of the fundamental classical results as well as of modern Positivstellensätze. This is a bit awkward.

[1] Thorsten, can you get Hilbert's exact words? (Auf Deutsch)

5.1 Nonnegative univariate polynomials

We start by discussing nonnegative univariate polynomials. Let $\mathbb{R}[x]$ denote the ring of polynomials in one variable. Recall that in Propoposition 2.5.1 we characterized the roots of a nonconstant monic polynomial $p \in \mathbb{R}[x]$ as the eigenvalues of the companion matrix C_p. This gives the following corollary:

Corollary 5.1.1. *A univariate polynomial $p \in \mathbb{R}[x]$ is strictly positive if and only if $p(0) > 0$ and, if it is nonconstant, its companion matrix C_p has no real eigenvalues.*

Polynomials of odd degree always have both positive and negative values, so nonnegative polynomials must have even degree. The set of nonnegative univariate polynomials coincides with set $\Sigma[x]$ of sum of squares of univariate polynomials.

Theorem 5.1.2. *A univariate polynomial of even degree is nonnegative if and only if it can be written as a sum of squares of univariate polynomials.*

Proof. A sum of squares is clearly nonnegative. For the converse, let $p \in \mathbb{R}[x]$ be nonnegative. As it is real, its non-real roots occur in conjugate pairs, and as it is nonnegative, its real roots have even multiplicity. Thus $p = r^2 \cdot q \cdot \overline{q}$, where $r \in \mathbb{R}[x]$ has the same real roots as does p, but each with half the multiplicity in p, and $q \in \mathbb{C}[x]$ has half of the non-real roots of p, one from each conjugate pair. Writing $q = q_1 + iq_2$ with $q_1, q_2 \in \mathbb{R}[x]$, this gives $p = r^2 q_1^2 + r^2 q_2^2$, a sum of squares. □

We now consider the problem of characterizing real polynomials that are nonnegative on an interval $I \subsetneq \mathbb{R}$. There are several cases to consider as I could be bounded or unbounded and it could be open, half-open, or closed. Observe that a polynomial p is nonnegative (respectively strictly positive) on a compact interval $[a, b]$ if and only if the polynomial q defined by

$$q(x) \;=\; p\left(\frac{(b-a)x + (b+a)}{2}\right) \tag{5.2}$$

is nonnegative (respectively strictly positive) on $[-1, 1]$, and the same is true for a half-open or open bounded interval. Similarly, a polynomial p is nonnegative on an interval $[a, \infty)$ with $a \in \mathbb{R}$ if and only if the polynomial $q(x) = p(x - a)$ is nonnegative on $[0, \infty)$.

These two principal cases of $I = [-1, 1]$ and $I = [0, \infty)$ are connected as follows. Given a polynomial p of degree at most d, the *(d-th degree) Goursat transform* $\widetilde{p}$ of p is

$$\widetilde{p}(x) \;:=\; (1+x)^d\, p\left(\frac{1-x}{1+x}\right) \;\in \mathbb{R}[x]\,.$$

Then $\widetilde{p}$ is a polynomial of degree at most d. Applying the Goursat transform to $\widetilde{p}$ yields the original polynomial p, up to a factor of 2^d,

$$(1+x)^d\, \widetilde{p}\left(\frac{1-x}{1+x}\right) \;=\; (1+x)^d \left(1 + \frac{1-x}{1+x}\right)^d p\left(\frac{1 - \frac{1-x}{1+x}}{1 + \frac{1-x}{1+x}}\right) \;=\; 2^d p(x)\,. \tag{5.3}$$

If $\deg \widetilde{p} < d$ then its $\deg \widetilde{p}$-th degree Goursat transform is a polynomial. The formula (5.3) implies that $(1+x)^{d-\deg \widetilde{p}}$ divides p. Hence, the Goursat transformation of a polynomial p of degree d is a polynomial of degree $d-k$ where k is the maximal power of $(1+x)$ dividing p. For example, the Goursat transform of $p = 1-x^2$ is $\widetilde{p} = 4x$ and $\widetilde{\widetilde{p}} = 4(1-x^2)$.

Lemma 5.1.3 (Goursat's Lemma)**.** *For a polynomial $p \in \mathbb{R}[x]$ of degree d we have:*

1. *p is nonnegative on $[-1,1]$ if and only if $\widetilde{p}$ is nonnegative on $[0,\infty)$.*
2. *p is strictly positive on $[-1,1]$ if and only if $\widetilde{p}$ is strictly positive on $[0,\infty)$ and $\deg \widetilde{p} = d$.*

Proof. Let $p \in \mathbb{R}[x]$ and consider the bijection $\varphi : (-1,1] \to [0,\infty)$, $x \mapsto \frac{1-x}{1+x}$. Since $(1+x)^d$ is strictly positive on $(-1,1]$, p is strictly positive on $(-1,1]$ if and only if $\widetilde{p}$ is strictly positive on $[0,\infty)$. The first assertion follows by continuity, $p(-1) \geq 0$ as p is positive on $(-1,1]$.

The second assertion follows from our observation that $\deg \widetilde{p} = d$ if and only $1+x$ does not divide p, so that p does not vanish at -1. □

Given polynomials $f_1, \ldots, f_r \in \mathbb{R}[x]$ and $I \subset \{1,\ldots,r\}$ set $f_I := \prod_{i \in I} f_i$, with the convention that $f_\emptyset = 1$. The *preorder generated by $f_1, \ldots, f_r$* is

$$P(f_1,\ldots,f_r) \;:=\; \left\{ \sum_{I \subset \{1,\ldots,r\}} s_I f_I \;:\; s_I \in \Sigma[x] \right\}.$$

Elements g of this preorder have the property that they are nonnegative on the set of $x \in \mathbb{R}$ where each of the polynomials f_i generating the preorder are nonnegative. Does this notion agree with the literature and with what comes later in the chapter?

Theorem 5.1.4 (Pólya and Szegő)**.** *If a univariate polynomial $p \in \mathbb{R}[x]$ is nonnegative on $[0,\infty)$ then we have*

$$p \;=\; f \;+\; xg \qquad \text{for some} \quad f, g \;\in\; \Sigma[x]\,, \tag{5.4}$$

with $\deg f, \deg xg \leq \deg p$. In particular, p lies in the preorder $P(x)$ generated by x.

In Section 5.6 we will see a multivariate statement of the same structure.

Proof. Dividing by the leading coefficient of p, we may assume that p is monic. We may also divide by any square factors in p (as in the proof of Theorem 5.1.2) and assume that p is square-free.

Consider first the case when p is irreducible. If p is linear, then, as its root is non-positive, $p = \alpha + x$ with $\alpha \geq 0$, an expression of the form (5.4). If p is quadratic, it has no real roots and so $p \in \Sigma[x]$ by Theorem 5.1.2, and we have $p = p + x \cdot 0$, again an expression of the form (5.4).

We complete the proof by induction on the number of irreducible factors of p. If $p = q_1 \cdot q_2$ with q_1, q_2 monic, real, and non-constant, then both q_1 and q_2 are nonnegative on $[0, \infty)$ and so we have expressions $q_i = f_i + x \cdot g_i$ for $f_i, g_i \in \Sigma[x]$ with $\deg f_i, \deg xg_i \leq \deg q_i$ for $i = 1, 2$. Setting $f := f_1 f_2 + x^2 g_1 g_2$ and $g := f_1 g_2 + g_1 f_2$ gives the desired expression (5.4) for p. □

Example 5.1.5. The polynomial $p = x^5 - 2x^3 + 2x^2 - 3x + 2 = (x+2)(x-1)^2(x^2+1)$, which is nonnegative on $[0, \infty)$, can be written as

$$p = 2\left((x(x-1))^2 + (x-1)^2\right) + x\left((x(x-1))^2 + (x-1)^2\right)$$

with the sum of squares polynomial $(x(x-1))^2 + (x-1)^2$.

Corollary 5.1.6. *A univariate polynomial $p \in \mathbb{R}[x]$ is nonnegative on an interval $[a, b]$ if and only if it lies in the preorder $P(x-a, b-x)$. Moerover, there is an expression*

$$p \;=\; f \;+\; (x-a)g \;+\; (b-x)h \;+\; (x-a)(b-x)k\,,$$

where $f, g, h, k \in \Sigma[x]$ and the degree of each term in this expression is at most $\deg(p)$.

An expression for a polynomial p such as that in the corollary that shows it lies in the preorder $P(x-a, b-x)$ is called a *certificate of nonegativity* for p on $[a, b]$.

Proof. It suffices to prove this when the interval $[a, b]$ is $[-1, 1]$, by the affine change of variables (5.2). Since elements of the preorder $P(x+1, 1-x)$ are automatically nonnegative on $[-1, 1]$, we only need to show that polynomials which are nonnegative on $[-1, 1]$ lie in this preorder.

Assume that p is nonnegative on $[-1, 1]$ and that it has degree d. By Goursat's Lemma, $\widetilde{p}$ is nonnegative on $[0, \infty)$ and by the Pólya-Szegő Theorem, there are polynomials $f, g \in \Sigma[x]$ with

$$\widetilde{p} \;=\; f + xg\,,$$

where $\deg f$ and $\deg xg$ are both at most $\deg \widetilde{p}$. Since $\widetilde{x} = 1 - x$, if we apply the dth order Goursat transform to this expression, we obtain

$$2^d p \;=\; (x+1)^{d - \deg f}\widetilde{f} \;+\; (1-x)(x+1)^{d-1-\deg g}\widetilde{g}\,. \tag{5.5}$$

The corollary follows as the Goursat transform is multiplicative, $\widetilde{hk} = \widetilde{h}\widetilde{k}$ and if $\deg(h) \geq \deg(k)$ with $\deg(h) = \deg(h+k)$, then

$$\widetilde{h+k} \;=\; \widetilde{h} \;+\; (1+x)^{\deg(h) - \deg(k)}\widetilde{k}\,,$$

and so $f, g \in \Sigma[x]$ implies that $\widetilde{f}, \widetilde{g} \in \Sigma[x]$. Absorbing even powers of $(x+1)^{d-\deg f}$ into $\widetilde{f}$, and the same for g, the expression (5.5) shows that $p \in P(x+1, 1-x)$. □

Example 5.1.7. The polynomial $p = 4 - 4x + 8x^2 - 4x^4$ is nonnegative on $[-1, 1]$. Indeed, we have

$$p = \left((2x-1)^2 + 1\right) + (x+1) + (1-x) + (x+1)(1-x)x^2\,.$$

Exercises

1. Use the Goursat transform and the algorithm in the proof of the Pólya and Szegő Theorem 5.1.4 to compute a certificate that

$$(x^2-9)(x^2-4) \;=\; x^4-13x^2+36$$

is positive on $[-1,1]$.

2. Find a certificate that the polynomial $(x^2-7)(x^2+x-5) = x^4+x^3-12x^2-7x+35$ is nonnegative on $[-2,1]$.

3. The Bernstein polynomials B_d^k, $0 \le k \le d$ are defined by

$$B_k^d(x) \;=\; 2^{-d}\binom{d}{k}(1+x)^k(1-x)^{d-k}\,.$$

Show that the Bernstein polynomials are nonnegative on $[-1,1]$ and constitute a partition of unity, that is, for $x \in [-1,1]$, $B_k^d(x) \ge 0$ and

$$\sum_{k=0}^{d} B_k^d(t) \;=\; 1\,.$$

4. Prove that the Bernstein polynomials satisfy the recursion

$$2B_k^d(x) \;=\; (1+x)B_{k-1}^{d-1}(x) + (1-x)B_k^{d-1}(x)$$

and form a basis of the vector space of polynomials of degree at most d.

5. Bernstein showed that every univariate polynomial p which is nonnonnegative on $[-1,1]$ can be written as a linear combination of Bernstein polynomials of some degree d with nonnegative coefficients. However, the smallest d, called the *Lorentz degree* of p, in a representation $p = \sum_{k=0}^{d} a_k B_k^d$ with nonnegative a_k is in general larger than the degree of p. Compute the Lorentz degree of the following polynomials.

   (a) $p = 4(1+x)^2 - 2(1+x)(1-x) + (1-x)^2$.

   (b) $p = 4(1+x)^2 - 3(1+x)(1-x) + (1-x)^2$.

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

5.2 Positive polynomials and sums of squares

As we saw in Section 5.1, every nonnegative univariate polynomial is a sum of squares. This property does not hold for multivariate polynomials and that phenomenon is at the root of many challenges in studying the set of nonnegative polynomials in n variables. For example, as we will see in more detail in Chapter 6, deciding the nonnegativity of a given polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]$ is a difficult problem.

There exist nonnegative multivariate polynomials which cannot be expressed as a sum of squares. While Hilbert gave a non-constructive proof of this in 1888, the first concrete example was given by Motzkin in 1967.

Theorem 5.2.1. *The polynomial* $p = 1 - 3x^2y^2 + x^2y^4 + x^4y^2 \in \mathbb{R}[x, y]$ *is nonnegative, but cannot be written as a sum of squares.*

See Figure 5.1 for an illustration of the graph $z = p(x, y)$ of the Motzkin polynomial.

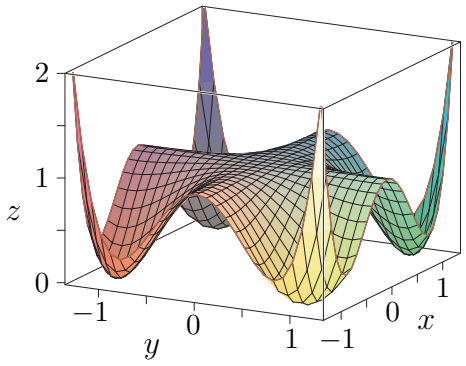


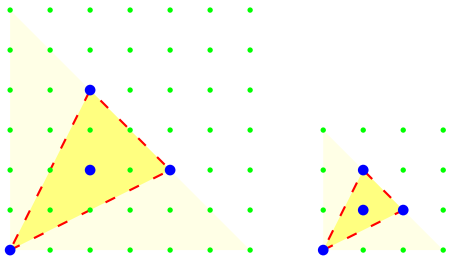
Figure 5.1: The Motzkin polynomial. Its zeroes are exactly at the four points $(\pm 1, \pm 1)$.

Proof. Recall the arithmetic-geometric mean inequality,

$$\frac{a+b+c}{3} \; - \; \sqrt[3]{abc} \; \geq \; 0 \quad \text{for } a, b, c \geq 0\,.$$

Setting $a = 1$, $b = x^2y^4$, and $c = x^4y^2$ shows the nonnegativity of p.

We now show that p cannot be written as a sum of squares by considering its Newton polytope and support, shown below on the left



(5.6)

Suppose that p is a sum of squares, so there exist polynomials $p_1, \ldots, p_k \in \mathbb{R}[x, y]$ with

$$p \;=\; p_1^2 \;+\; p_2^2 \;+\; \cdots \;+\; p_k^2 \,. \tag{5.7}$$

By Exercise 2, we must have that $2\,\mathrm{NP}(p_i) \subset \mathrm{NP}(p)$ for each $i = 1, \ldots, k$. Thus each summand p_i has Newton polytope contained in the polygon on the right of (5.6). That is, there exist $a_i, b_i, c_i, d_i \in \mathbb{R}$ with $p_i = a_i + b_i xy + c_i xy^2 + d_i x^2 y$. Then (5.7) implies that $-3 = \sum_i b_i^2$, which is impossible. □

For $d \geq 0$, let $\mathbb{R}_d[x_1, \ldots, x_n]$ be the space of polynomials in $x_1, \ldots, x_n$ that are homogeneous of degree d, called *d-forms*. For $n \geq 1$ and $d \geq 2$ let

$$\mathcal{P}_{n,d} \;:=\; \{p \in \mathbb{R}_d[x_1, \ldots, x_n] \,:\, p \geq 0\}$$

denote the set of nonnegative polynomials of degree d in $x_1, \ldots, x_n$ and

$$\Sigma_{n,d} \;:=\; \{p \in \mathcal{P}_{n,d} \,:\, p \text{ is a sum of squares}\}\,.$$

be its subset of polynomials that are sums of squares. Both of these are convex cones. For $\mathcal{P}_{n,d}$ this is because it is defined by the linear inequalities $p(x) \geq 0$ for $x \in \mathbb{R}^n$. This imlies that if $p, q \in \mathcal{P}_{n,d}$ and $\alpha, \beta \in \mathbb{R}_>$, then $\alpha f + \beta g \in \mathcal{P}_{n,d}$. For $\Sigma_{n,d}$, this is evident from its definition. Both are closed. For $\mathcal{P}_{n,d}$ this follows from its inequality description, and for $\Sigma_{n,d}$, this will be explained in Chapter ???.

For any polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]$ of degee at most d we have its homogenization to degree d,

$$\overline{p} \;:=\; x_0^d\, p\left(\frac{x_1}{x_0}, \ldots, \frac{x_1}{x_0}\right)\,,$$

which a form of degree d. Homogenization $p \mapsto \overline{p}$ is a vector space isomorphism from the space of all polynomials in $\mathbb{R}[x_1, \ldots, x_n]$ of total degree at most d to the space of homogeneous polynomials in $\mathbb{R}[x_0, x_1, \ldots, x_n]$ of degree d. The dehomogenization of a homogeneous polynomial $p \in \mathbb{R}[x_0, x_1, \ldots, x_n]$ is the polynomial $p(1, x_1, \ldots, x_n) \in \mathbb{R}[x_1, \ldots, x_n]$. For $p \in \mathbb{R}[x_1, \ldots, x_n]$ the dehomogenization of $\overline{p}$ is p.

Lemma 5.2.2. *Let d be even and $p \in \mathbb{R}[x]$ be a polynomial of degree d.*

1. *p is nonnegative on $\mathbb{R}^n$ if and only if $\overline{p}$ is nonnegative on $\mathbb{R}^{n+1}$.*
2. *p is a sum of squares of polynomials if and only if $\overline{p}$ is a sum of of squares of homogeneous polynomials of degree $\frac{d}{2}$.*

Proof. For the first statement, suppose that p is nonnegative on $\mathbb{R}^n$. For $a = (a_0, \ldots, a_n) \in \mathbb{R}^{n+1}$ with $a_0 \neq 0$ we have $\overline{p}(a) = a_0^d\, p\left(\frac{a_1}{a_0}, \ldots, \frac{a_1}{a_0}\right) > 0$, and the continuity of $\overline{p}$ implies that if $a_0 = 0$ then $\overline{p}(a) \geq 0$. The converse follows from dehomogenizing.

For the second statement if $p = \sum_{i=1}^k p_i^2$ is a sum of squares of polynomials p_i, then by Exercise 1, $\deg p_i \leq \frac{d}{2}$. Thus $p = \sum_{i=1}^k (x_0^{d/2} \cdot p_i(\frac{x_1}{x_0}, \ldots, \frac{x_1}{x_0}))^2$ is a representation as sum of squares of homogeneous polynomials of degree $\frac{d}{2}$. The converse follows from dehomogenizing. □

The following classical theorem due to Hilbert provides a complete classification of the cases (n, d) when the cone of nonnegative polynomials coincides with the cone of sums of squares of polynomials.

Theorem 5.2.3 (Hilbert)**.** *Let $n \geq 2$ and d even. For the inclusion $\Sigma_{n,d} \subset \mathcal{P}_{n,d}$ equality holds in exactly the following cases:*

1. *$n = 2$ (binary forms).*
2. *$d = 2$ (quadratic forms).*
3. *$n = 3$, $d = 4$ (ternary quartics).*

Proof. Under dehomogenization, a nonnegative binary form ($n = 2$) becomes a nonnegative univariate polynomial, so the first assertion is a consequence of Theorem 5.1.2.

When $d = 2$, note that a (homogeneous) quadratic form $x^T Ax$ with A symmetric is nonnegative if and only if A is positive semidefinite. Maybe this needs more explaination? By the Choleski decomposition, this holds if and only if the quadratic form is a sum of squares, and so $\Sigma_{n,2} = \mathcal{P}_{n,2}$.

The case $(n, d) = (3, 4)$ is Theorem 5.2.4 below, and thus it remains to show that $\mathcal{P}_{n,d} \setminus \Sigma_{n,d} \neq \emptyset$ for all pairs (n, d) not treated so far. By homogenizing the Motzkin polynomial $p \in \mathbb{R}[x, y]$ from Theorem 5.2.1 via

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

for $d \geq 6$ we see that $\mathcal{P}_{n,d} \setminus \Sigma_{n,d} \neq \emptyset$ for $n \geq 3$ and $d \geq 6$. For the cases (n, d) with $n \geq 4$ and $d = 4$ the difference follows from the nonnegative quartic form $w^4 + x^2y^2 + x^2z^2 + y^2z^2 - 4wxyz$ of Exercise 3. □

Theorem 5.2.4. *Every nonnegative ternary quartic can be written as a sum of squares of quadratic forms.*

The proof uses the following lemma:

Lemma 5.2.5. *For any positive polynomial $p \in \mathcal{P}_{3,4}$ there exists a nonzero quadratic form q with $p \geq q^2$ on $\mathbb{R}^3$.*

Proof. We consider three cases for the number of points in $\mathcal{V}(p) \subset \mathbb{P}^2(\mathbb{R})$, the real zeroes of the form p, either $\mathcal{V}(p)$ is empty, or it consists of one point, or more than one point.

Suppose that $\mathcal{V}(p)$ is empty. By compactness of the unit sphere $\mathbb{S}^2$, the positive continuous function $p/(x^2 + y^2 + z^2)^2$ on $\mathbb{S}^2$ is bounded from below by some $\varepsilon > 0$, and so $p \geq \varepsilon(x^2 + y^2 + z^2)^2$. By homogeneity, this implies that $p \geq \varepsilon(x^2 + y^2 + z^2)^2$, and so we set $q = \sqrt{\varepsilon}(x^2 + y^2 + z^2)$.

Suppose that $\mathcal{V}(p) \subset \mathbb{RP}^2$ consists of a single point. After a possible coordinate change, we may assume that this point is $[1 : 0 : 0]$. Expand p as a polynomial in x,

$$p \;=\; a_0x^4 + a_1x^3 + a_2x^2 + a_3x + a_4\,,$$

where $a_i \in \mathbb{R}[y,z]$ is homogeneous of degree i. Evaluating p at $(1,0,0)$, we see that $a_0 = 0$. Then a_3 must also be the zero polynomial for otherwise s takes negative values. Similarly, if a_2 is the zero polynomial, then a_3 is also zero, and so $p = a_4 \in \mathbb{R}[y,z]$ is a nonnegative binary form and thus a sum of squares.

Thus we may assume that p is a quadratic polynomial in x,

$$p \;=\; x^2 f + 2xg + h \tag{5.8}$$

for some $f, g, h \in \mathbb{R}[y,z]$ which are homogeneous forms of degree $2, 3, 4$, respectively. Since p does not vanish for any x with $(y,z) \neq (0,0)$, the discriminant $g^2 - fh$ of this quadratic in x is strictly negative on $\mathbb{R}^2 \smallsetminus \{(0,0)\}$. Since p is nonnegative on $\mathbb{R}^3$, the binary quadratic form f is also nonnegative, for if there is a (y,z) with $f(y,z) < 0$, then for x sufficiently large, $p(x,y,z) < 0$.

If f is irreducible over $\mathbb{R}$, then it is strictly positive on $\mathbb{R}^2 \smallsetminus \{(0,0)\}$. Since the negative of the discriminant $fh - g^2$ of the quadratic (5.8) is also strictly positive on $\mathbb{R}^2 \smallsetminus \{(0,0)\}$, the homogeneous function $(fh - g^2)/f^3$ of degree zero is bounded from below by some $\varepsilon > 0$, thus giving $fp \geq fh - g^2 \geq \varepsilon f^3$ and further $p \geq \varepsilon f^2$. So in this case we may set $q = \sqrt{\varepsilon} f$.

If the quadratic form f is reducible, then $f = f_1^2$ for some linear form f_1. The nonnegativity of $fh - g^2 = f_1^2 h - g^2$ implies that g vanishes on $\mathcal{V}(f) = \mathcal{V}(f_1)$ so that $g = g_1 f_1$ for some quadratic form $g_1 \in \mathbb{R}[y,z]$. Hence, $fs \geq (xf + g)^2 = f(xf_1 + g_1)^2$ and further $s \geq (xf_1 + g_1)^2$ on $\mathbb{R}^3$. Thus we can choose $q = xf_1 + g_1$.

Now suppose that $\mathcal{V}(p)$ has at least two points in $\mathbb{RP}^2$. We may assume that p vanishes at both $[1:0:0]$ and $[0:1:0]$ and therefore has at degree at most two in both x and y. Therefore, we may write

$$p \;=\; x^2 f + 2xzg + z^2 h\,,$$

where $f, g, h \in \mathbb{R}[y,z]$ are binary quadratic forms. As before, both f and h are nonnegative, and the discriminant $(zg)^2 - fh$ of p as a quadratic in x is non-positive for any $(y,z) \in \mathbb{R}^2$.

If either f or h vanishes and so is the square of a linear form, then the same arguments as before give the desired form q. Thus we may assume that both f and h are irreducible in $\mathbb{R}[y,z]$ and strictly positive on $\mathbb{R}^2 \smallsetminus \{(0,0)\}$. There remain two cases for the non-negative quartic $fh - g^2$.

If $fh - g^2$ has a zero at $(b,c) \neq (0,0)$, set $a := -g(b,c)/f(b,c)$, and define

$$p_1(x,y,z) \;:=\; p(x + az, y, z) \;=\; x^2 f + 2xz(g + af) + z^2(h + 2ag + a^2 f)\,.$$

Since $f(h + 2ag + a^2 f)(b,c) = (fh - g^2)(a,b) = 0$, the coefficient of z^2 in this expression for p_1 as a quadratic in x has a zero. By our previous arguments, there is a quadratic form q with $p \geq q^2$.

Finally, suppose that $fh - g^2$ is strictly positive. Then $(fh - g^2)/f/(y^2 + z^2)$ is strictly positive on the unit circle $\mathbb{S}^1 \subset \mathbb{R}^2$, and is therefore bounded below by some positive

constant $\varepsilon > 0$. We then conclude that

$$fp \;\geq\; z^2(fh - g^2) \;\geq\; \varepsilon z^2(y^2 + z^2)f\,,$$

and therefore $p \geq \varepsilon z^4$, and so we set $q := \sqrt{\varepsilon} z^2$. □

Before proving Theorem 5.2.4, note that $\mathcal{P}_{n,d}$ is a convex cone in a finite-dimensional space. A form $p \in \mathcal{P}_{n,d}$ is *extremal* if whenever we have $f = f_1 + f_2$ with $f_1, f_2 \in \mathcal{P}_{n,d}$, then $f_i = \lambda_i f$ for some λ_1, λ_2 with $\lambda_1 + \lambda_2 = 1$. Any form $s \in \mathcal{P}_{n,d}$ can be written as a finite sum of extremal forms.

Proof of Theorem 5.2.4. Given a form $s \in \mathcal{P}_{3,4}$, write $s = s_1 + \cdots + s_k$ as a sum of finitely many extremal forms $s_1, \ldots, s_k$. Lemma 5.2.5 implies that there exist quadratic forms $q_i \neq 0$ and nonnegative quartic forms t_i with $s_i = q_i^2 + t_i$, $1 \leq i \leq k$. Since s_i is extremal, t_i must be a nonnegative multiple of q_i^2, which implies that s_i is a sum of squares. □

Hilbert in fact showed that every ternary quartic is a sum of at most *three* squares, but all known results for this refinement are substantially more involved. [2]

The discriminant of a real symmetric matrix is a particularly beautiful example of a nonnegative polynomial that is a sum of squares. The *discriminant of a matrix* $A \in \mathbb{C}^{n \times n}$ is defined as the discriminant of its characteristic polynomial χ_A,

$$\operatorname{disc}(A) \;=\; \operatorname{disc}(\chi_A(t)) \;=\; \operatorname{Res}_t(\chi_A, \chi_A')\,, \tag{5.9}$$

hence

$$\operatorname{disc}(A) \;=\; \prod_{i<j} (\lambda_i - \lambda_j)^2\,,$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. If A is real and symmetric then its eigenvalues are all real which implies that $\operatorname{disc}(A)$ is nonnegative. Note that (5.9) expresses $\operatorname{disc}(A)$ as a homogeneous polynomial of degree $n(n-1)$ in the coefficients of A. In light of Theorem 5.2.3 it is remarkable that the nonnegative polynomial $\operatorname{disc}(A)$ can be written as a sum of squares.

Theorem 5.2.6 (Find a citation for the notes)**.** *Let $A = (a_{ij})$ be a symmetric $n \times n$-matrix with indeterminates a_{ij}. Then* $\operatorname{disc}(A)$ *is a sum of squares of polynomials in the a_{ij}.*

For a matrix $A \in \mathbb{R}^{n \times n}$ let $\operatorname{vec}(A)$ denote the vector in $\mathbb{R}^{n^2}$ obtained by writing the elements of A in a single column. Define $A^* \in \mathbb{R}^{n^2 \times n}$ to be the matrix whose i-th column contains $\operatorname{vec}(A^{i-1})$, for $1 \leq i \leq n$. When $n \geq 2$, the matrix A^* is not square. Theorem 5.2.6 follows from an explicit representation of $\operatorname{disc}(A)$ as a sum of squares.

[2] In the notes, write about the number of representations with 3 squares, and include the new work about varieties of minimal degree.

Lemma 5.2.7. *For a symmetric $n \times n$-matrix we have*

$$\operatorname{disc}(A) \;=\; \sum_{I \in \binom{[n^2]}{n}} (\det A_I^*)^2 \,,$$

where A_I^* is the submatrix of A^* formed by the rows with indices I. This notation also appears in the Plücker ideal

If $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix then the only non-zero rows I of A^* are those corresponding to the diagonal entries of A, and these form the Vandermonde matrix,

$$A_I^* \;=\; \begin{pmatrix} 1 & \lambda_1 & \cdots & \lambda_1^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_n & \cdots & \lambda_n^{n-1} \end{pmatrix} \,,$$

whose determinant is

$$\det(A_I^*) \;=\; \prod_{i<j} (\lambda_j - \lambda_i) \,,$$

and thus we have $(\det(A_I^*))^2 = \operatorname{disc}(A)$, which proves the lemma when A diagonal.

Proof. Since the discriminant of A is the square of the determinant of the Vandermonde matrix formed from the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A, we have

$$\begin{aligned} \operatorname{disc}(A) \;&=\; \det \begin{pmatrix} 1 & \cdots & 1 \\ \lambda_1 & \cdots & \lambda_n \\ \vdots & \ddots & \vdots \\ \lambda_1^{n-1} & \cdots & \lambda_n^{n-1} \end{pmatrix} \begin{pmatrix} 1 & \lambda_1 & \cdots & \lambda_1^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_n & \cdots & \lambda_n^{n-1} \end{pmatrix} \\ &=\; \det \begin{pmatrix} 1 & p_1(A) & p_2(A) & \cdots & p_{n-1}(A) \\ p_1(A) & p_2(A) & & \iddots & p_n(A) \\ p_2(A) & & \iddots & & \vdots \\ \vdots & \iddots & & & p_{n-3}(A) \\ p_{n-1}(A) & p_n(A) & \cdots & p_{n-3}(A) & p_{n-2}(A) \end{pmatrix} \,, \end{aligned} \tag{5.10}$$

where $p_k(A) = \lambda_1^k + \cdots + \lambda_n^k$ is the kth Newton power sum of the eigenvalues of A. This is the trace, $\operatorname{Tr}(A^k)$ of the matrix A^k, and so is a homogeneous polynomial of degree k in the entries a_{ij} of the matrix A.

If B and C are symmetric matrices, then we have that

$$\operatorname{Tr}(BC) \;=\; \sum_{k,l=1}^{n} B_{k,l} C_{l,k} \;=\; \sum_{k,l=1}^{n} B_{k,l} C_{k,l} \,.$$

Letting $B = C = A$, we see that the (i,j)-entry of $(A^*)^T A^*$ is

$$\sum_{k,l=1}^{n} A_{k,l}^{i-1} A_{k,l}^{j-1} \;=\; \mathrm{Tr}(A^{i-1}A^{j-1}) \;=\; p_{i+j-2}(A)\,,$$

which is the (i,j)-entry in the Hankel matrix (5.10). Thus $\mathrm{disc}(A) = \det((A^*)^T A^*)$, and the formula of the lemma is obtained from the Cauchy-Binet formula for the expansion of the determinant $\det((A^*)^T A^*)$. □

Example 5.2.8. For the symmetric matrix

$$A \;=\; \begin{pmatrix} a & c \\ c & b \end{pmatrix} \quad \text{we have} \quad A^* \;=\; \begin{pmatrix} 1 & a \\ 0 & c \\ 0 & c \\ 1 & b \end{pmatrix}$$

and thus from Lemma 5.2.7 we have the sum of squares representation

$$\mathrm{disc}(A) \;=\; c^2 + c^2 + (b-a)^2 + 0 + c^2 + c^2 = 4c^2 + (b-a)^2\,.$$

Exercises

1. Suppose that $p = \sum_{i=1}^{k} p_i^2$ is a sum of squares of polynomials $p_1, \ldots, p_k \in \mathbb{R}[x]$. If at least one of the polynomials p_i is non-zero then p is non-zero and $\deg p_i = 2\max\{\deg p_i \;:\; 1 \le i \le k\}$.

2. Suppose that $p = \sum_{i=1}^{k} p_i^2$ is a sum of squares of polynomials $p_1, \ldots, p_k \in \mathbb{R}[x_1, \ldots, x_n]$. For any weight vector $w \in \mathbb{R}^n$, let $c_\alpha x^\alpha$ be a term of p that is extreme in the direction of w, that is $w \cdot \alpha$ is maximal among all exponents in p. Show that, if $c_i x^{\beta_i}$ is the extreme monomial in p_i in the direction w, then $2w \cdot \beta_i \le w \cdot \alpha$, and there is some i for which this is an equality. Conclude that

$$\mathrm{NP}(p) \;=\; \mathrm{conv}\Big(\bigcup_{i=1}^{k} 2 \cdot \mathrm{NP}(p_i)\Big)\,,$$

   Where $\mathrm{NP}(p)$ is the Newton polytope of the polynomial f. Make sure that Newton polytopes and weighted term orders are done earlier.

3. Show that the polynomial $w^4 + x^2y^2 + x^2z^2 + y^2z^2 - 4wxyz$ is nonnegative but not a sum of squares.

4. In how many different ways can you write the ternary quartic $x^4 + y^4 + z^4$ as a sum of squares? Besides the obvious representation $(x^2)^2 + (y^2)^2 + (z^2)^2$ one other solution among the possible ones is, for example, $(x^2 - y^2)^2 + (2xy)^2 + (z^2)^2$. How about $x^2y^2 + x^2z^2 + y^2z^2$? Put the answer in the notes at the end

5.3 Preorders and quadratic modules

Let us begin with two examples.

Example 5.3.1. The set $\Sigma[x] = \Sigma[x_1, \ldots, x_n]$ of sums of squares of real polynomials satisfies the properties $\Sigma[x] + \Sigma[x] \subset \Sigma[x]$, $\Sigma[x] \cdot \Sigma[x] \subset \Sigma[x]$ and $a^2 \in \Sigma[x]$ for all $a \in \mathbb{R}[x]$.

More generally, given polynomials $f_1, \ldots, f_m$, let $P = P(g_1, \ldots, g_m)$ be the set of all polynomials of the form

$$\sum_{I \subset [m]} \sigma_I \cdot \prod_{i \in I} g_i \,,$$

where each coefficient σ_I is a sum of squares from $\Sigma[x]$. Then $P + P \subset P$, $PP \subset P$, and P contains all squares.

Example 5.3.2. Given a set of polynomials $g_1, \ldots, g_m$, let $M = M(g_1, \ldots, g_m)$ be the et of polynomials of the form

$$\sigma_0 + \sigma_1 g_1 + \cdots + \sigma_m g_m \,,$$

were each σ_i is a sum of squares from $\Sigma[x]$. Then this set of polynomials satisfies $M + M \subset M$, $1 \in M$, and $a^2 M \subset M$ for all $a \in \mathbb{R}[x]$.

Observe that, for $g_1, \ldots, g_m$, we have $M(g_1, \ldots, g_m) \subset P(g_1, \ldots, g_m)$, and if S is a set where each polynomial g is nonnegative, then any polynomial in $P(g_1, \ldots, g_m)$ (and hence also in $S(g_1, \ldots, g_m)$) is nonnegative on S.

We formalize the algebraic structures underlying these examples. Example 5.3.1 leads to the notion of a preorder and Example 5.3.2 leads to the notion of a quadratic module. While most of our preorders will be subsets of $\mathbb{R}[x]$, it is useful to define them over an aribitrary commutative ring R with 1. We always assume that $\frac{1}{2} \in R$.

Definition 5.3.3. A *preorder* of R is a subset P of R such that

$$P + P \subset P \,, \quad PP \subset P \,, \quad \text{and } a^2 \in P \text{ for all } a \in R \,.$$

A *quadratic module* of R is a subset M of R such that

$$M + M \subset M \,, \quad 1 \in M \,, \quad \text{and } a^2 M \subset M \text{ for all } a \in R \,.$$

Note that every preorder of R is a quadratic module of R. We call the set $P(g_1, \ldots, g_m)$ the preorder generated by $g_1, \ldots, g_m$. Note that it is the smallest preorder containing $g_1, \ldots, g_m$. Similarly, $M(g_1, \ldots, g_m)$, the quadratic module geerated by $g_1, \ldots, g_m$ is the smallest quadratic module containing $g_1, \ldots, g_m$.

Lemma 5.3.4. *Let M be a quadratic module of R. Then $I := M \cap -M$ is an ideal of R. Moreover, $-1 \in M$ if and only if $M = R$.*

Proof. The properties $I + I \subset I$ and $a^2 I \subset I$ for all $a \in R$ are clear. Writing a in the form $a = \frac{1}{4}((a+1)^2 - (a-1))^2$ shows $aI \subset I$. Hence, I is an ideal.

For the second statement, let $-1 \in M$. Since $1 \in M$ by definition, $1 \in M \cap -M = I$. Since I is an ideal, we have $M \cap -M = R$; hence $M = R$. □

A quadratic module M is *proper* if $-1 \notin M$ so that $M \neq R$. A proper quadratic module M is *maximal* if it is not contained in any strictly larger proper quadratic module.

Lemma 5.3.5. *If M is a maximal proper quadratic module of $\mathbb{R}[x]$ then $M \cup -M = \mathbb{R}[x]$.*

Proof. We argue by contradiction. Let $p \in \mathbb{R}[x]$ and assume that $p \notin M \cup -M$. By the maximality of M, neiter quadratic module $M' := M + p\Sigma[x]$ nor $M'' := M - p\Sigma[x]$ are proper. Hence there exist $m_1, m_2 \in M$ and $\sigma_1, \sigma_2 \in \Sigma[x]$ with $m_1 + \sigma_1 p = -1$ and $m_2 - \sigma_2 p = -1$. Thus

$$\sigma_1 m_2 + \sigma_2 m_1 \;=\; \sigma_1(m_2 - \sigma_2 p) + \sigma_2(m_1 + \sigma_1 p) \;=\; -(\sigma_1 + \sigma_2)\,,$$

which implies $\sigma_1 \in -\sigma_2 - M \subset -M$ and similarly $\sigma_2 \in -M$. We can conclude that σ_1, σ_2 are contained in the ideal $I := M \cap -M$ of Lemma 5.3.4. Thus $\sigma_1 p \in I$ and so $m_1 + \sigma_1 p \in M$ which contradicts $m_1 + \sigma_1 p = -1 \notin M$. □

Let M be a quadratic module of R. An ideal I of R is called *M-convex* if for any $m_1, m_2 \in M$ with $m_1 + m_2 \in I$, we have $m_1, m_2 \in I$.

Lemma 5.3.6. *Let M be a quadratic module of R and $I := M \cap -M$. Then any minimal prime ideal J containing I is M-convex.*

Proof. Let J be a minimal prime ideal containing I. Let $m_1, m_2 \in M$ with $m_1 + m_2 \in J$.

By Theorem A.1.1 in the Appendix, there exists a $u \in R \setminus J$ and some $N \geq 0$ with $u(m_1 + m_2)^N \in I$. Then also $u^2(m_1 + m_2)^N \in I$, and without loss of generality we can assume that N is odd (otherwise consider $u^2(m_1 + m_2)^{N+1}$). Thus one of i and $N - i$ is always even for i an integer, and so every term in the binomial expansion

$$u^2(m_1 + m_2)^N \;=\; \sum_{i=0}^{N} u^2 \binom{N}{i} m_1^i m_2^{N-i}$$

lies in M. By Exercise 3 the ideal I is M-convex, each term $u^2\binom{N}{i} m_1^i m_2^{N-i}$ lies in I, and thus in particular $u^2 m_1^N \in I$. As $I \subset J$, we have that $u^2 m_1^N \in J$. Since $u \in R \setminus J$ and J is prime, we have $m_1 \in J$, and thus also $m_2 \in J$. □

Given a proper M-convex prime ideal I of $\mathbb{R}[x]$ (or, more generally, of an integral domain), the quadratic module M can be extended to the quotient field $\mathbb{F}$ of $\mathbb{R}[x]/I$. Here, the extension $\widetilde{M}$ of M to $\mathbb{F}$ is given by all elements of the form $\sum_i (\frac{a_i}{b_i})^2 m_i$ with $a_i, b_i \in \mathbb{R}[x]/I$ and $m_i \in (M + I)/I$.

Lemma 5.3.7. *Let M be a quadratic module of $\mathbb{R}[x]$ and I be a proper M-convex prime ideal of $\mathbb{R}[x]$. Then the extension of M to the quotient field $\mathbb{F}$ of $\mathbb{R}[x]/I$ is proper.*

Proof. First note that any element $\sum_i (\frac{a_i}{b_i})^2 m_i$ of the extension $\widetilde{M}$ of M may be written as $\frac{m}{b^2}$ with $m \in M$ and b a non-zero element in $\mathbb{R}[x]/I$. For example, we may set $b := \prod_i b_i$ and $m := \sum_i (a_i \prod_{j \neq i} b_j)^2 m_i$.

Now assume that the extension $\widetilde{M}$ of M is not proper. Then there exists $m \in M$ and a non-zero element $b \in \mathbb{R}[x]/I$ such that $-1 = \frac{m}{b^2}$ in $\mathbb{R}[x]/I$. This implies $-b^2 = m \in (M+I)/I$, and therefore $b^2 \in (M+I)/I \cap -(M+I)/I$.

Since I is M-convex, Exercise 2 implies $(M+I) \cap -(M+I) = I$ and hence $(M+I)/I \cap -(M+I)/I = \{0\}$. Thus $b = 0$, a contradiction. □

Exercises

1. If P is a preorder of a field K of arbitrary characteristic, then the following statements are equivalent:

   (a) $P \cap -P = \{0\}$.

   (b) $P^\times + P^\times \subset P^\times$, where $P^\times := P \setminus \{0\}$.

   (c) $-1 \notin P$.

   For $\operatorname{char} K \neq 2$, the condition $P \neq K$ is equivalent to these as well.

2. An ideal I of R is M-convex if and only if $(M+I) \cap -(M+I) = I$.

3. Let M be a quadratic module of R. Show that the ideal $M \cap -M$ is M-convex and that $M \cap -M$ is the smallest M-convex ideal of R.

4. The intersection of M-convex ideals is M-convex.

5. Let $S := \{x \in \mathbb{R}^n : g_1(x) \geq 0, \ldots, g_m(x) \geq 0\}$, and let p be strictly positive on S. For each compact subset $C \subset \mathbb{R}^n$ there exists some $q \in M(g_1, \ldots, g_m)$ such that $p - q$ is strictly positive on C.

5.4 The Positivstellensatz

A representation of a polynomial p as a sum of squares is a *certificate* of nonnegativity of p, that is, a proof the p is nonnegative. Certificates play an important role in optimization and for algorithmic purposes, see the subsequent Chapter 6.

One of the most well-known forms of such certificates can be found in *Farkas' Lemma* in linear optimization, which has many different formulations. For affine-linear functions $g_1, \ldots, g_m \in \mathbb{R}[x_1, \ldots, x_n]$ let

$$S \;=\; \{x \in \mathbb{R}^n \,:\, g_i(x) \geq 0\,,\; 1 \leq i \leq m\}\,,$$

which is a polyhedron. We quote the following affine form of Farkas' Lemma.

Lemma 5.4.1 (Farkas' Lemma)**.** *Let p and $g_1, \ldots, g_m$ be affine-linear functions. If S is non-empty and p is nonnegative on S, then there exist scalars $\lambda_0, \ldots, \lambda_m \geq 0$ with*

$$p \;=\; \lambda_0 + \sum_{j=1}^{m} \lambda_j g_j \,.$$

Hence, providing nonnegative scalars $\lambda_0, \ldots, \lambda_m$ yields a certificate for the nonnegativity of the affine function p on K.

As we have seen in earlier chapters Where?, studying the solutions to a system of polynomial equations is a prominent problem in algebraic geometry. Hilbert's Nullstellensatz 1.2.9, which establishes a connection between the algebraic varieties in $\mathbb{C}^n$ and ideals in $\mathbb{C}[x_1, \ldots, x_n]$, yields a certificate for the nonexistence of a system of polynomial equations. Denoting the ideal generated by given polynomials $f_1, \ldots, f_r \in \mathbb{C}[x_1, \ldots, x_n]$ by $\mathcal{I}(f_1, \ldots, f_r)$, this can be particularly nicely seen from following formulation of the weak form of Hilbert's Nullstellensatz.

Theorem 5.4.2. *The following two statements are equivalent:* Make sure this is identical to the from in Chapter 1.

1. *The set $\{x \in \mathbb{C}^n \,:\, f_i(x) = 0$ for $1 \leq i \leq r\}$ is empty.*
2. *$1 \in \mathcal{I}(f_1, \ldots, f_r)$, i.e., there exist $g_1, \ldots, g_r \in \mathbb{C}[x_1, \ldots, x_n]$ with*

$$f_1 g_1 + \cdots + f_r g_r \;=\; 1 \,. \tag{5.11}$$

However, as discussed in earlier chapters Make this discussion!, the inherent difficulty is that the degrees of the polynomials in the representation (5.11) can grow doubly exponentially in the dimension n.

An analog of Hilbert's Nullstellensatz for semialgebraic sets, called the *Positivestellensatz*, was proven by Krivine [47] and Stengle [89]. This Positivstellensatz guarantees the existence of a certificate for the nonnegativity of a polynomial on a semialgebraic set. For this, let $P(g_1, \ldots, g_r)$ be the preorder generated by the polynomials $g_1, \ldots, g_m$,

$$P(g_1, \ldots, g_m) \;=\; \Big\{ p \in \mathbb{R}[x_1, \ldots, x_n] \,:\, p = \sum_{I \subseteq \{1,\ldots,n\}} \sigma_I \prod_{i \in I} g_i \Big\}$$

with sum of squares polynomials $\sigma_I \in \Sigma[x_1, \ldots, x_n]$. Further let $M(h_1, \ldots, h_t)$ be the monoid defined by the polynomials $h_1, \ldots, h_t$, i.e., the set of (finite) products of the polynomials including the empty product. This notation for $M(...)$ collides with earlier notation. Can the notation be excised?

Theorem 5.4.3 (Positivstellensatz)**.** *For polynomials $f_1, \ldots, f_r,\ g_1, \ldots, g_s,\ h_1, \ldots, h_t \in \mathbb{R}[x_1, \ldots, x_n]$ the following statements are equivalent:*

(1) *The set*
$$\{x \in \mathbb{R}^n \,:\, f_i(x) = 0,\, g_j(x) \geq 0,\, h_k(x) \neq 0 \quad \forall i,j,k\}$$
is empty.

(2) *There exist polynomials* $F \in \mathcal{I}(f_1,\ldots,f_r)$, $G \in P(g_1,\ldots,g_s)$ *and* $H \in M(h_1,\ldots,h_t)$ *with*
$$F + G + H^2 \;=\; 0\,.$$

Note that (2) $\Longrightarrow$ (1), for if x lies in the set of (1) and F, G, H are as in (2), then $F(x) = 0$, $G(x) \geq 0$, and $H^2(x) > 0$, which shows the emptyness of the set in (1). This has several important special cases:

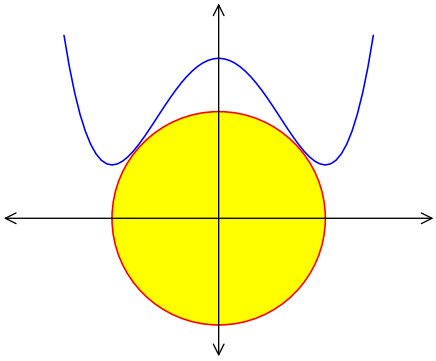
Corollary 5.4.4 (Real Nullstellensatz)**.** *Let* $f_1,\ldots,f_r \in \mathbb{R}[x] = \mathbb{R}[x_1,\ldots,x_n]$. *Then the real variety* $\mathcal{V}_{\mathbb{R}}(f_1,\ldots,f_r)$ *is emtpy if and only if there exists* $F \in \mathcal{I}(f_1,\ldots,f_r)$ *and a sum of squares* $G \in \Sigma[x]$ *with*
$$F + G + 1 = 0\,. \tag{5.12}$$

Proof. This is the Positivstellensatz when $s = t = 0$. □

Example 5.4.5. The polynomial $f(x) = x^2 + 1$ does not have a real zero. If $F = -(x^2+1) \in \mathcal{I}(f)$ and $G = x^2 \in \Sigma[x]$, then they satisfy (5.12) and thus provide a certificate for the emptyness of $\mathcal{V}_{\mathbb{R}}(f)$.

The general quadratic equation $x^2 + px + q = 0$ with coefficients $p, q \in \mathbb{R}$ has a real solution unless the disriminant $D := \frac{p^2}{4} - q$ is negative. In the latter case, setting
$$F \;:=\; \frac{1}{D}\left(x^2 + px + q\right) \qquad \text{and } G \;:=\; \left(\frac{1}{\sqrt{-D}}\left(x + \frac{p}{2}\right)\right)^2$$
yields a certicate of the form (5.12) for the emptyness of $\mathcal{V}(F)$.

The real algebraic curve in $\mathbb{R}^2$ given by $y = x^4 - 2x^2 + \frac{3}{2}$ does not intersect the unit disc, but it is hard to tell from a picture.



For a proof, set $f := y - (x^4 - 2x^2 + \frac{3}{2})$, $g := 1 - x^2 - y^2$, and $a := (\frac{2}{3})^{1/4}$. Then the identity
$$f \;+\; (ay - \tfrac{1}{2a})^2 \;+\; (x^2 + \tfrac{a^2}{2} - 1)^2 \;+\; a^2 g \;+\; 1 \;=\; 0\,,$$

gives a Positivstellensatz certificate for the emptyness of $\{(x,y) \in \mathbb{R}^2 : f(x,y) = 0,\, g(x,y) \geq 0\}$.

Before turning towards a proof of the Positivstellensatz, we record the following special cases.

Corollary 5.4.6. *Let $g_1, \ldots, g_m \in \mathbb{R}[x_1, \ldots, x_n]$. Set $K := \{x \in \mathbb{R}^n : g_j(x) \geq 0, 1 \leq j \leq m\}$ and let $P(g_1, \ldots, g_m)$ be the preorder generated by.* I think that T is supposed to be $P(g_1, \ldots, g_m)$, but this should be checked. *Then:*

1. *$f > 0$ on K if and only if there exist $G, H \in T$ with $fG = 1 + H$.*
2. *$f \geq 0$ on K if and only if there exist $G, H \in T$ and $k \geq 0$ with $fG = f^{2k} + H$.*
3. *$f = 0$ on K if and only if there exists a $G \in T$ and $k \geq 0$ with $f^{2k} + G = 0$.*

Proof. For the first statement, we consider the set

$$\{x \in \mathbb{R}^n : -f(x) \geq 0,\, g_j(x) \geq 0,\, 1 \leq j \leq m\}.$$

This set is empty if and only if there exist $G, H \in T$ with $H - fG + 1 = 0$.

Similarly, for the second statement, we apply the Positivstellensatz 5.4.3 to the set

$$\{x \in \mathbb{R}^n : -f(x) \geq 0,\, f(x) \neq 0,\, g_j(x) \geq 0,\, 1 \leq j \leq m\}.$$

This set is empty if and only if there exist $G, H \in T$ and $k \geq 0$ with $H - fG + f^{2k} = 0$.

And for the third statement, the set

$$\{x \in \mathbb{R}^n : f(x) \neq 0,\, g_j(x) \geq 0,\, 1 \leq j \leq m\}.$$

is empty if and only if there exist $G \in T$ and $k \geq 0$ with $G + f^{2k} = 0$. □

As exhibited in the following corollary, the Positivstellensatz can be seen as a generalization of the Artin and Schreier's solution to Hilbert's 17th Problem.

Corollary 5.4.7 (Solution to Hilbert's 17th problem)**.** *Any nonnegative polynomial $f \in \mathbb{R}[x]$ can be written as a sum of squares of rational functions.*

Proof. Choosing $K = \emptyset$, Corollary 5.4.6(2) implies for any nonnegative polynomial $f \in \mathbb{R}[x]$ there exists an identity $fg = f^{2k} + h$ with g, h sums of squares and k a nonnegative integer. We can assume that $f \neq 0$. Then $f^{2k} + h \neq 0$, and so $g \neq 0$. Hence, we have

$$f \;=\; \frac{1}{g}(f^{2k} + h) \;=\; \left(\frac{1}{g}\right)^2 g(f^{2k} + g),$$

a representation of f as a sum of squares of rational functions. □

This corollary explains why a treatment of the Positivstellensatz is closely connected to a treatement of Hilbert's 17th problem. Rather than providing a complete proof for the Positivstellensatz, we will sketch (building upon the appendix on real algebra) in Theorem 5.4.8 how the important special case of nonemptyness of the set can be deduced from some main tools of real algebra. Then we will show how to deduce some of the cases in Corollary 5.4.6 from Theorem 5.4.8.

Theorem 5.4.8. *Let $g_1, \ldots, g_m \in \mathbb{R}[x]$ and $S = \{x \in \mathbb{R}^n : g_i(x) \geq 0,\ 1 \leq i \leq m\}$. If $S = \emptyset$ then -1 is contained in the preorder $P(g_1, \ldots, g_m)$.*

Proof. We write for short $P = P(g_1, \ldots, g_m)$, and assume $-1 \notin P$. Since any preorder is a quadratic module, Lemma 5.3.4 implies that $P \cap -P$ is a proper ideal of $\mathbb{R}[x]$. Let I be a maximal ideal containing $P \cap -P$. Since I is also prime, we can apply Theorem A.1.1 in the Appendix to obtain a proper minimal prime ideal $Q \subset \mathbb{R}[x]$ which contains $P \cap -P$. Lemma 5.3.6 implies that Q is P-convex.

Since $-1 \notin P$, by Lemma 5.3.7 we may extend P to a proper preorder $\widetilde{P}$ of the quotient field $\mathbb{F}$ of $\mathbb{R}[x]/Q$. By Theorem A.3.2 in Appendix A.3 there exists an ordering P' on $\mathbb{F}$ with $\widetilde{P} \subset P'$. Let $\leq$ be the associated order relation, i.e., $a \leq b$ if and only if $b - a \in P'$. Embedding $\mathbb{R}$ canonically into $\mathbb{R}[x]$ and $\mathbb{R}[x]/Q$ naturally into $\mathbb{F}$, the natural residue class mapping $\mathbb{R}[x] \to \mathbb{R}[x]/Q$ shows that $\mathbb{F}$ is a field extension of $\mathbb{R}$. Restricting the ordering on $\mathbb{F}$ to $\mathbb{R}$, we obtain the unique (...) ordering on $\mathbb{R}$.

Our goal now is to show that there exists an element $x \in \mathbb{F}^n$ with $g_i(x) \geq 0$, $1 \leq i \leq m$. Define $[x_i]$ as the residue class of x_i in $\mathbb{R}[x]/Q$. For any $g = \sum_\alpha c_\alpha x^\alpha \in \mathbb{R}[x]$, the image $[g]$ of g in $\mathbb{F}$ is

$$[g] \;=\; \sum_\alpha [c_\alpha][x^\alpha] \;=\; g([x])\,.$$

Since $g \in P$ and P' extends P, we observe $[g] \geq 0$ in the ordering P', and hence $g([x]) \geq 0$.

Thus, by Tarski's Transfer Principle A.3.3, there exists some $z \in \mathbb{R}^n$ with $g_i(z) \geq 0$, $1 \leq i \leq m$. Therefore $S \neq \emptyset$. □

Now we show that the proofs of some special cases of the Positivstellensatz directly from the Emptyness Theorem 5.4.8.

Proof. (of the first statement of Corollary 5.4.6, from Theorem 5.4.8). Let $g_1, \ldots, g_m \in \mathbb{R}[x_1, \ldots, x_n]$, $K = \{x \in \mathbb{R}^n : g_j(x) \geq 0, 1 \leq j \leq m\}$ and $K' = K \cap \{x \in \mathbb{R}^n : -f \geq 0\}$. Let P be the preorder associated to the polynomials $g_1, \ldots, g_m$ and P' be the preorder associated to $g_1, \ldots, g_m, -f$. We observe $P' = P - fP$.

If $f \geq 0$ then $K' = \emptyset$ and thus, by Theorem 5.4.8, $-1 \in P$. Hence, there exist $g, H \in P$ with $-1 = H - fG$. □

Exercises

1. Write the Motzkin polynomal as a sum of squares of rational functions.

2. Provide a Nullstellensatz certificate showing that for $a, b > 1$ the polynomials $f := x^2 + y^2 - 1$ and $g := ax^2 + by^2 - 1$ have no common real zeroes.

5.5 Pólya's Theorem and Handelman's Theorem

In 1927 Pólya discovered a fundamental theorem on positive polynomials on a simplex, which will also be an ingredient for important representation theorems of positive polynomials. After discusssing Pólya's Theorem, we will also derive Handelman's Theorem from it, which provides a distinguished representation of a strictly positive polynomial on a polytope.

Theorem 5.5.1 (Pólya)**.** *Let $p \in \mathbb{R}[x]$ be a homogeneous polynomial which is strictly positive on $\mathbb{R}^n_+ \setminus \{0\}$. Then for all sufficiently large N, the polynomial*

$$(x_1 + \cdots + x_n)^N p$$

has only nonnegative coefficients.

By homogeneity, the condition of strict positivity on $\mathbb{R}^n_+ \setminus \{0\}$ can is equivalent to strict positivity on the unit simplex $\Delta_n := \{y \in \mathbb{R}^n_+ : \sum_{i=1}^n y_i = 1\}$. Further note that the when $n = 1$, Pólya's Theorem is trivial, since a homogenous polynomial in a single variable consists of a single term.

Proof. Let p be a homogenous polynomial of degree d, so that p is of the form $p = \sum_{|\alpha|=d} c_\alpha x^\alpha$. The key to the proof is to define the polynomial

$$g \;=\; \sum_{|\alpha|=d} c_\alpha \prod_{i=1}^n x_i(x_i - t)\cdots(x_i - (\alpha_i - 1)t)\,.$$

in the ring extension $\mathbb{R}[x_1, \ldots, x_n, t]$. Then $p = g(x_1, \ldots, x_n, 0)$ and the polynomial g provides a useful representation of $(x_1 + \cdots + x_n)^N p$ for any $N \in \mathbb{N}$, namely

$$(x_1 + \cdots + x_n)^N p \;=\; \sum_{|\beta|=d+N} \frac{N!(d+N)^d}{\beta_1! \cdots \beta_n!}\, g\left(\frac{\beta_1}{d+N}, \ldots, \frac{\beta_n}{d+N}, \frac{1}{d+N}\right) x^\beta\,. \qquad (5.13)$$

To see (5.13), first expand $(x_1 + \cdots + x_n)^N$ using the Multinomial Theorem and then collect terms with the same monomial x^β to obtain

$$\begin{aligned}(x_1 + \cdots + x_n)^N p \;&=\; \sum_{|\alpha|=d} \sum_{|\gamma|=N} c_\alpha \binom{N}{\gamma_1 \;\cdots\; \gamma_n} x_1^{\alpha_1+\gamma_1} \cdots x_n^{\alpha_n+\gamma_n} \\ &=\; \sum_{|\beta|=d+N} \sum_{|\alpha|=d,\, \alpha\le\beta} c_\alpha \binom{N}{\beta_1-\alpha_1 \;\cdots\; \beta_n-\alpha_n} x^\beta\,. \qquad (5.14)\end{aligned}$$

Equality of the coefficients in (5.13) and (5.14) follows from

$$\sum_{|\alpha|=d,\,\alpha\leq\beta} c_\alpha \binom{N}{\beta_1-\alpha_1\ \cdots\ \beta_n-\alpha_n} \;=\; \frac{N!}{\beta_1!\cdots\beta_n!}\sum_{|\alpha|=d} c_\alpha \prod_{i=1}^{d}\beta_i(\beta_i-1)\cdots(\beta_i-(\alpha_i-1))\,.$$

Since $\left(\frac{\beta_1}{d+N},\ldots,\frac{\beta_n}{d+N}\right)$ lies in the compact setsimplex Δ_n, and since $\lim_{N\to\infty}\frac{1}{d+N}=0$, it now suffices to show that there exists a neighborhood U of $0\in\mathbb{R}$ such that for all $x\in\Delta_n$ and for all $t\in U$ we have $g(x,t)>0$.

For any fixed $x\in\Delta_n$ we have $g(x,0)=p(x)>0$. By continuity, there exists a neighborhood of $(x,0)\in\mathbb{R}^{n+1}$ on which g remains strictly positive. Without loss of generality we can assume that this neighborhood is of the form $S_x\times U_x$ where S_x is a neighborhood of x and U_x is a neighborhood of 0.

The family of sets $\{S_x : x\in\Delta_n\}$ is a covering of the compact set Δ_n, and hence there exists finite covering $\{S_x : x\in X\}$ for a finite subset X of Δ_n. Choosing $U=\bigcap_{x\in X}U_x$ yields the desired neighborhood of 0. □

Pólya's Theorem implies that any homogeneous polynomial p which is strictly positive on Δ_n can be written as a quotient of two polynomials with positive coefficients, because $p=\frac{f}{(x_1+\cdots+x_n)^N}$ with some polynomial f that has nonnegative coefficients. Replacing each variable x_i with x_i^2 and transforming Δ_n into the unit sphere, we obtain a constructive solution to Hilbert's 17th problem, whenever p is an homogeneous, even, and strictly positive polynomial on $\mathbb{R}^n$. Explain this better!

Example 5.5.2. Let p be the strictly positive polynomial on $\mathbb{R}^3\setminus\{0\}$ defined by

$$p \;=\; (x-y)^2+(x+y)^2+(x-z)^2 \;=\; 3x^2-2xz+2y^2+z^2\,.$$

We have

$$(x_1+x_2+x_3)p \;=\; 3x^3+3x^2y+x^2z+2xy^2-2xyz-xz^2+2y^3+2y^2z+yz^2+z^3$$

and $(x_1+x_2+x_3)^2p$ also has some negative coefficients. The smallest N such that $(x_1+x_2+x_3)^3p$ has only nonnegative coefficients is 3. This smallest N is called the *Pólya exponent* of p.

We derive Handelman's Theorem from Pólya's Theorem. For a positive polynomial on a polytope, it characterizes (in contrast to Theorem 5.4.3) a representation of the polynomial p itself, rather than only a product of a sum of squares polynomial with p.

Theorem 5.5.3 (Handelman)**.** *Let $g_1,\ldots,g_m\in\mathbb{R}[x]$ be affine-linear polynomials such that $K=\{x\in\mathbb{R}^n : g_1(x)\geq 0,\ldots,g_m(x)\geq 0\}$ is non-empty and bounded, that is, a polytope. Any polynomial $p\in\mathbb{R}[x]$ which is strictly positive on K can be written in the form*

$$p \;=\; \sum_{e\in\mathbb{N}_0^m} c_e \prod_{j=1}^{m} g_j^{e_j}\,. \tag{5.15}$$

Why $e \in \mathbb{N}_0^m$, and not our usual exponent vector notation? What is c_e ?

Denoting by $\mathcal{S}(g_1, \ldots, g_m)$ the semiring

$$\mathcal{S}(g_1, \ldots, g_m) = \operatorname{pos}\left\{ \prod_{j=1}^{m} g_j^{e_j} : e \in \mathbb{N}_0^n \right\},$$

where pos abbreviates the positive hull, Handelman's Theorem states that $p \in \mathcal{S}(g_1, \ldots, g_m)$.

Since by Farkas' Lemma 5.4.1, any linear form which is strictly positive on K is contained in $\mathcal{S}(g_1, \ldots, g_m)$, it actually suffices to show that $p \in \mathcal{S}(g_1, \ldots, g_m, \ell_1, \ldots, \ell_r)$ with some some linear forms $\ell_1, \ldots, \ell_r$ which are strictly positive on S. The main idea then will be to encode linear forms by indeterminates and apply Pólya's Theorem to deduce that the coefficients of these extended polynomials are nonnegative. Handelman's Theorem 5.5.3 will follow from a suitible substitution.

Proof. For $1 \leq i \leq n$ let ℓ_i be the linear form $\ell_i = x_i + \tau_i$, with sufficently large $\tau_i \in \mathbb{R}$ such that ℓ_i is strictly positive on K. Further set

$$\ell_{n+1} \;=\; \tau - \sum_{j=1}^{m} g_j - \sum_{i=1}^{n} \ell_j$$

with $\tau \in \mathbb{R}$ sufficiently large such that ℓ_{n+1} is strictly positive on K. We show $p \in \mathcal{S}(g_1, \ldots, g_m, \ell_1, \ldots, \ell_{n+1})$, from which the desired statement then follows.

By applying a suitable variable transformation we can assume $\ell_i = x_i$, $1 \leq i \leq n$. Now extend the variable set $x_1, \ldots, x_n$ to $x_1, \ldots, x_{n+m+1}$ and consider the polynomial

$$h(x) \;=\; p(x) + c \sum_{j=1}^{m} (x_{n+j} - g_j(x))$$

with a positive constant c. Note that $h(x_1, \ldots, x_n, g_1(x), \ldots, g_m(x), x_{n+m+1}) = p(x)$. Let $\Delta = \{x \in \mathbb{R}_+^{n+m+1} : \sum_{i=1}^{m+n+1} x_i = \tau\}$ and

$$\Delta' \;=\; \{x \in \mathbb{R}_+^{n+m+1} : x_{n+1} = g_1(x), \ldots, x_{n+m} = g_m(x)\}.$$

For any $x \in \Delta'$ and $j \in \{1, \ldots, m\}$ we have $g_j(x) = x_{n+j} \geq 0$. Hence, for each $c > 0$ the polynomial h is strictly positive on Δ'. By compactness of Δ and Δ', we can fix some $c > 0$ such that h is strictly positive on Δ. Denoting by $\operatorname{tdeg} p$ the total degree of p, let

$$\overline{h} \;=\; \Big(\sum x_i\Big)^{\operatorname{tdeg} p} h\left(\frac{x_1}{\sum x_i}, \ldots, \frac{x_n}{\sum x_i}\right)$$

the homogenization of h with respect to $\sum x_i$. Since $\overline{h}$ is strictly positive on Δ, Pólya's Theorem 5.5.1 gives an $N \in \mathbb{N}_0$ such that all coefficients of $(\frac{1}{\tau}\sum x_i)^N h(x)$ are nonnegative. Substituting successively $z = \tau - \sum_{i=1}^{n+m} x_i$ and $x_{n+j} = g_j(x)$, $1 \leq j \leq m$, we see that

$$p(x_1, \ldots, x_n) = \left(\frac{1}{\tau}\sum x_i\right)^N h(x_1, \ldots, x_{n+m+1}),$$

which gives $p \in \mathcal{S}(g_1, \ldots, g_m, \ell_1, \ldots, \ell_{n+1})$ as needed. □

Example 5.5.4. It can happen that the degree of the right hand side of any Handelman representation (5.15) of p exceeds the degree of p. To see this, it suffices to consider the univariate situation with $g_1(x) = 1 + x$, $g_2(x) = 1 - x$. The parabola $x^2 + 1$ is strictly positive on $[-1, 1]$, but it cannot be represented as a nonnegative linear combination of 1, $1 + x$, $1 - x$, and $(1 + x)(1 - x)$. Indeed, this phenomenon has already been mentioned in connection with the Lorentz degree in the exercises to Section 5.1.

Exercises

1. Use a computer algebra system to determine the Pólya exponent of $p = (x - y)^2 + y^2 + (x - z)^2$.

2. Show that the precondition of strict positivity in Pólya's Theorem in general cannot be relaxed to nonnegativity. For this, inspect the counterexample $p = xz^3 + yz^3 + x^2y^2 - xyz^2$.

3. Give an example of Handleman's Theorem.

5.6 Representation Theorems

We derive the some fundamental theorems which (under certain conditions) provide beautiful and useful representations of polynomials p strictly positive on a semialgebraic set.

We begin with the Theorem of Jacobi-Prestel which can be derived from Handelman's Theorem and nicely exhibits some techniques which will also be applied in the subsequent Theorems of Schmüdgen and Putinar.

Let $g_1, \ldots, g_m \in \mathbb{R}[x] = \mathbb{R}[x_1, \ldots, x_n]$ and $S := \{x \in \mathbb{R}^n : g_i(x) \geq 0,\ 1 \leq i \leq m\}$.

Theorem 5.6.1 (Jacobi-Prestel [43]). *Let $S \neq \emptyset$ and bounded, let $M(g)$ contain some linear polynomials $\ell_1, \ldots, \ell_k$ with $k \geq 1$ such that $\{x \in \mathbb{R}^n : \ell_1(x) \geq 0, \ldots, \ell_k(x) \geq 0\}$ is bounded. If p is strictly positive on S, then $p \in M(g)$.*

What is S? Recall $M(g)$?

Proof. Without loss of generality we can assume that the polytope $L := \{x \in \mathbb{R}^n : \ell_1(x) \geq 0, \ldots, \ell_k(x) \geq 0\}$ is contained in the cube $[-1, 1]^n$.

We start by showing that $p \in M(g_1, \ldots, g_m, n - \|x\|^2 + 1)$. For this, first note that

$$\begin{aligned} n + 2 \pm x_i &= \frac{1}{2}\left((n+2) + (1 \pm x_i)^2 + \sum_{j \in \{1,\ldots,n\} \setminus \{i\}} x_j^2 + (n - \|x\|^2 + 1)\right) \\ &\subset M(n - \|x\|^2 + 1)\,. \end{aligned}$$

Hence, by linear combinations, $t+\ell_i \in M(n-\|x\|^2+1)$ for sufficiently large $t \in \mathbb{R}$. Since L is bounded, the set

$$P \;=\; \{x \in \mathbb{R}^n \,:\, t+\ell_1(x) \geq 0, \ldots, t+\ell_k(x) \geq 0\}$$

is bounded as well and thus a polytope. By Exercise 5 in Section 5.3 there exists a polynomial $q \in M(g)$ such that $p-q$ is strictly positive on P. Applying Handelman's Theorem 5.5.3, we can deduce

$$p-q \;\in\; \mathcal{S}(t+\ell_1, \ldots, t+\ell_k) \;\subset\; M(n-\|x\|^2+1)\,,$$

whence $p \in M(g_1, \ldots, g_m, n-\|x\|^2+1)$.

To deduce $p \in M(g_1, \ldots, g_m)$ from $p \in M(g_1, \ldots, g_m, n-\|x\|^2+1)$, observe

$$\begin{aligned} n-\|x\|^2 &= \frac{1}{2}\sum_{i=1}^{n}\big((1+x_i)^2(1-x_i)+(1-x_i)^2(1+x_i)\big) \\ &\in M(1-x_1, \ldots, 1-x_n, 1+x_1, \ldots, 1+x_n) \end{aligned}$$

Since by Farkas' Lemma we have $1 \pm x_i \in \mathbb{R}+\sum \mathbb{R}_+\ell_i$ and by assumption $\ell_1, \ldots, \ell_k \in M(g)$, we can conclude $n-\|x\|^2 \in M(g)$. Hence, the polynomial $n-\|x\|^2+1$ is contained in $M(g)$ as well, so that $p \in M(g_1, \ldots, g_m, n-\|x\|^2+1) \subset M(g)$. □

We now derive the fundamental Theorem of Schmüdgen which (under the condition of compactness of the feasible set) characterizes (in contrast to Theorem 5.4.3) a representation of the polynomial p itself in terms of the preorder of the polynomials definining the feasible set. Throughout this section let $g_1, \ldots, g_m \in \mathbb{R}[x] = \mathbb{R}[x_1, \ldots, x_n]$ and $K := \{x \in \mathbb{R}^n \,:\, g_i(x) \geq 0,\, 1 \leq i \leq m\}$.

Theorem 5.6.2 (Schmüdgen [79])**.** *If a polynomial $f \in \mathbb{R}[x]$ is strictly positive on a compact set K then $f \in P(g_1, \ldots, g_m)$. More precisely, f is strictly positive on K if and only if f can be written in the form*

$$\alpha + \sum_{e \in \{0,1\}^n} \left(\sum_i \alpha_{e_i} h_{e_i}^2\right) g_1^{e_1} \cdots g_n^{e_n} \tag{5.16}$$

with $\alpha > 0$, $\alpha_{ij} \geq 0$ and $h_{e_i} \in K[x]$.

Proof. Let $\gamma > 0$ such that $\gamma - \|x\|^2$ is strictly posivite von S. By Stengle's Positivstellensatz 5.4.3 (in the special form ...) there exist $G, H \in P(g_1, \ldots, g_m)$ with $\gamma - \|x\|^2 = (1+G)/(1+H)$. Hence, $(1+H)(\gamma-\|x\|^2) \in P(g_1, \ldots, g_m)$.

Similar to the proof of the Jacobi-Prestel Theorem one can show that there exists some $\gamma' > 0$ with $\gamma' - \|x\|^2 \in P(g_1, \ldots, g_m, (1+H)(\rho-\|x\|^2))$ and that $p \in M(g_1, \ldots, g_m, \rho' - \|x\|^2)$. The first of these statements implies $\gamma' - \|x\|^2 \in P(g_1, \ldots, g_m)$ and then the second one $p \in P(g_1, \ldots, g_m)$. □

Example 5.6.3. Let $g_1 = 1 - \sum_{i=1}^n x_i^2$, $g_2 = x_n$ and $S = \{x \in \mathbb{R}^n : g_1(x) \geq 0,\ g_2(x) \geq 0\}$. If we think of x_n in the vertical direction, S is the upper half of the unit ball. Schmüdgen's Theorem asserts that any strictly positive polynomial p on S can be written as

$$p = \sigma_0 + \sigma_1(1 - \sum_{i=1}^n x_i^2) + \sigma_2 x_n + \sigma_3(1 - \sum_{i=1}^n x_i^2)x_n$$

with sums of squares $\sigma_0, \ldots, \sigma_3 \in \Sigma[x]$.

The representation Theorem 5.6.2 of Schmüdgen is quite fundamental, but the number of terms is exponential in the number of variables n. In this section we discuss Putinar's Theorem which improves upon this situation, assuming that a certain precondition is satisfied.

Again let $g_1, \ldots, g_m \in \mathbb{R}[x] = \mathbb{R}[x_1, \ldots, x_n]$ and $K := \{x \in \mathbb{R}^n : g_i(x) \geq 0,\ 1 \leq i \leq m\}$. Before we state Putinar's Theorem we provide several equivalent formulations of the precondition which we need.

A quadratic module $M \subset \mathbb{R}[x]$ is called *Archimedean* if for every $h \in \mathbb{R}[x]$ there is some $N \in \mathbb{N}$ such that $N \pm h \in M$.

Theorem 5.6.4. *The following conditions are equivalent:*

1. *The module $M(g_1, \ldots, g_m)$ is Archimedean, i.e., For all $h \in \mathbb{R}[x]$ there is some $N \in \mathbb{N}$ such that $N \pm h \in M(g_1, \ldots, g_m)$.*
2. *There exists an $N \in \mathbb{N}$ such that $N - \sum_{i=1}^n x_i^2 \in M(g_1, \ldots, g_m)$.*
3. *There exists some $h \in M(g_1, \ldots, g_m)$ such that $\{x \in \mathbb{R}^n : h(x) \geq 0\}$ is compact.*
4. *There exist finitely many polynomials $h_1, \ldots, h_r \in M(g_1, \ldots, g_m)$ such that the set*

$$\{x \in \mathbb{R}^n : h_1(x) \geq 0, \ldots, h_s(x) \geq 0\} \tag{5.17}$$

(which contains K; explain why) is compact and $\prod_{i \in I} h_i \in M(g_1, \ldots, g_m)$ for all $I \subseteq \{1, \ldots, s\}$.

Proof. The implications $1 \Longrightarrow 2 \Longrightarrow 3 \Longrightarrow 4$ are obvious.

In order to show $4 \Longrightarrow 1$, let $h_1, \ldots, h_s \in M(g_1, \ldots, g_m)$ such that the set K' defined by (5.17) is compact and $\prod_{i \in I} h_i \in M(g_1, \ldots, g_m)$ for all $I \subseteq \{1, \ldots, s\}$. By the compactness of K', for any given $h \in \mathbb{R}[x]$ there exists an $N \in \mathbb{N}$ such that $N \pm h > 0$ on K'. Then Schmüdgen's Theorem 5.6.2 implies $N \pm h = \sum_{I \subseteq \{1, \ldots, h_r\}} s_I h_I$ for some sums of squares so that $h \in M(g_1, \ldots, g_m)$. □

The conditions in Theorem 5.6.4 are actually not conditions on K, but on the representation of K (in terms of the polynomials g_i). See Exercise 1 for an example which shows that the conditions are actually stronger than just requiring that K is compact. In many practical applications, the precondition in Theorem 5.6.4 can be imposed by adding an inequality $N - \sum_{i=1}^n x_i^2 \geq 0$ for a sufficiently large N.

Theorem 5.6.5 (Putinar)**.** *Let K be defined as above and $M(g_1, \ldots, g_m)$ be Archimedean. If a polynomial $f \in \mathbb{R}[x]$ is stricly positive on K then $f \in M(g_1, \ldots, g_m)$. That is, there exist sums of squares $\sigma_0, \ldots, \sigma_m \in \Sigma[x]$ with*

$$f \;=\; \sigma_0 + \sum_{i=1}^{m} \sigma_i g_i \,. \tag{5.18}$$

It is evident that each polynomial of the form (5.18) is nonnegative on K.

Proof. By Theorem 5.6.4, there exists some some $h \in M(g_1, \ldots, g_m)$ such that the set $C := \{x \in \mathbb{R}^n \,:\, h(x) \geq 0\}$ is compact. Hence, by Exercise 5 in Section 5.3, there exists some $q \in \mathbb{R}[x_1, \ldots, x_n]$ such that $p - q$ is strictly positive on C. Applying Schmüdgen's Theorem 5.6.2 yields $p - q \in P(h) = M(h) \subset M(g_1, \ldots, g_m)$. □

Example 5.6.6. The strict positivity in the precondition to Putinar's statement is essential, already in the univariate case. This can be seen in the example $p = 1 - x^2$, $g = g_1 = (1 - x^2)^3$.

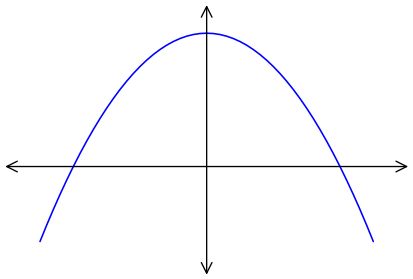


Figure 5.2: Graph of $p(x) = 1 - x^2$.

(see Figure 5.2). The feasible set K is the interval $K = [-1, 1]$, and hence the minima of the objective function p are at $x = -1$ and $x = 1$, both with function value 0. The precondition of Putinar's theorem satisfied since

$$\frac{2}{3} + \frac{4}{3}\big(x^3 - \frac{3}{2}x\big)^2 + \frac{4}{3}\big(1 - x^2\big)^3 \;=\; 2 - x^2 \,.$$

If a representation of the form (5.18) existed, i.e.,

$$1 - x^2 \;=\; \sigma_0(x) + \sigma_1(x)(1 - x^2)^3 \quad \text{ with } \sigma_0, \sigma_1 \in \Sigma[x] \,, \tag{5.19}$$

then the right hand side of (5.19) must vanish at $x = 1$ as well. The second term has at 1 a zero of at least third order, so that s_0 vanishes at 1 as well; by the SOS-condition this zero of s_0 is of order at least 2. Altogether, on the right hand side we have at 1 a zero of at least second order, in contradiction to the order 1 of the left side. Thus there exists no representation of the form (5.19).

When p is nonnegative on a compact set $K = \{x \in \mathbb{R}^n : g_1(x) \geq 0, \ldots, g_m(x) \geq 0\}$ then by Schmüdgen's Theorem the polynomial $p + \varepsilon$ is contained in $P(g_1, \ldots, g_m)$. And similarly, if the module $M(g_1, \ldots, g_m)$ is Archimedean, $p + \varepsilon \in M(g_1, \ldots, g_m)$. However, for $\varepsilon \to 0$, the smallest degrees of those representations might go to infinity.

Exercises

1. Let $g_1 = 2x_1 - 1$, $g_2 = 2x_2 - 1$, $g_3 = 1 - x_1x_2$. Show that $S := \{x \in \mathbb{R}^2 : g_i(x) \geq 0,\ 1 \leq i \leq 3\}$ is compact but that $M(g_1, g_2, g_3)$ is not archimedean.

2. Using $p = 1 - x^2$ and $g = g_1 = (1 - x^2)^3$ from Exercise 5.6.6, show that Schmüdgen's Theorem may fail if one replaces strict positivity in the precondition by nonnegativity.

3. Need a couple more exercises

5.7 Notes

A lot of theMuch material in this chapter is classical. Some standard references for these topics are the books of Basu, Pollack, and Roy [4], of Bochnak, Coste, and Roy [13] and of Prestel and Delzell [72].

The treatment on univariate polynomials is based on the exposition by Powers and Reznick [71]. Goursat's Lemma is due to Édouard Jean-Baptiste Goursat (1858–1936). It was shown by Bernstein [8] that every nonnegative polynomial on on $[0, 1]$ has a nonnegative representation in terms of the Bernstein polynomials (see also [70, vol. II, p. 83, Exercise 49]).

Have a discussion with reference to the Pólya-Szagő Theorem 5.1.4

The elementary proof of Hilbert's Classification Theorem 5.2.3 is due to Choi and Lam [19], and likewise the polynomial in Exercise 3.

The form we give of Farkas's Lemma 5.4.1 may be found in [81, Corollary 7.1h].

Our treatment of quadratic modules and the Positivstellensatz is based on Marshall's book [57]. The Positivstellensatz is due to Krivine and Stengle (for the historical development see [72]).

Pólya's Theorem was proven by Pólya ([69], see also [34]). The special cases $n = 2$ and $n = 3$ were shown before by Poincaré [68] and Meissner [59]. We remark that for the case of strictly positive polynomials, Habicht has shown how to derive a solution to Hilbert's 17th problem from Pólya's Theorem [32].

Handelman's Theorem was proven in [33], our proof follows Schweighofer's and Averkov's derivation from Pólya's Theorem [82, 83, 3]. The equivalences on the Archimedean property in Theorem 5.6.4 were shown by Schmüdgen [79].

Chapter 6

Optimization and real algebraic geometry

Do optim. .... use Positivstellensaetze ... different feature ... By the Positivstellensatz, a polynomial p is nonnegative on the set $K = \{x \in \mathbb{R}^n : g_i(x) \geq 0\,,\ 1 \leq i \leq m\}$ if and only if there exist a $k \in \mathbb{N}_0$ and an $F \in \mathcal{A}(-p, g_1, \ldots, g_m)$ with $F + p^{2k} = 0$. In order to minimize a polynomial on a set K, the task is therefore to determine the largest γ such that the polynomial $p - \gamma$ has such a certificate. In this way, we can also consider the algebraic certificates for the nonnegativity of polynomials on a semialgebraic set K from the viewpoint of optimization.

The main concern in this way of proceeding is that the existing proofs of the Positivstellensatz are nonconstructive, i.e., they do not yield an algorithmic method to determine a certificate. In particular, the degrees of the required polynomials F, G and H can be quite large. The best published bound is n-fold exponential. For the case that there are only equality constraints ("real Nullstellensatz"), an improvement – to 3-fold exponential – was announced by Lombardi and Roy.

Under certain restrictions to the semialgebraic set K "better suited" forms of Positivstellensätze can be provided. In view of the connection to optimization, the subsequently discussed version of Putinar has turned out to be particularly useful.

6.1 Global optimization of polynomials and sums of squares

6.1.1 Nonnegative polynomials versus sums of squares

Deciding the nonnegativity of a given polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]$ is a difficult problem. The fundamental idea of the approach is to replace such a problem by the decision problem "Is p a sum of squares of polynomials?" This problem turns out to be much easier.

Example 6.1.1. Let p be homogeneous of degree $2d$; then it suffices to investigate homogeneous polynomials of degree d for the decomposition.

Let

$$\begin{aligned} p(x,y) &= 2x^4 + 2x^3y - x^2y^2 + 5y^4 \\ &= (x^2, y^2, xy)\, Q \begin{pmatrix} x^2 \\ y^2 \\ xy \end{pmatrix} \end{aligned}$$

with a symmetric matrix $Q \in \mathbb{R}^{3\times 3}$. Since Q must be positive semidefinite, there exists a decomposition $Q = LL^T$. One specific solution is

$$L = \frac{1}{\sqrt{2}} \begin{pmatrix} 2 & 0 \\ -3 & 1 \\ 1 & 3 \end{pmatrix}, \qquad \text{hence} \quad Q = \begin{pmatrix} 2 & -3 & 1 \\ -3 & 5 & 0 \\ 1 & 0 & 5 \end{pmatrix}.$$

This implies the sum of squares (SOS) decomposition

$$p(x,y) = \frac{1}{2}(2x^2 - 3y^2 + xy)^2 + \frac{1}{2}(y^2 + 3xy)^2 .$$

This problem connects to a major theory of real algebraic geometry.

Let

$$\mathcal{P}_{n,d} = \{p \in \mathbb{R}[x_1, \ldots, x_n] \,:\, p \text{ of total degree} \le d \text{ and } p \ge 0\}$$

and

$$\Sigma_{n,d} = \{p \in \mathbb{R}[x_1, \ldots, x_n] \,:\, p \text{ is a sum of squares}\}.$$

Connect to Hilbert classification!

We consider the SOS relaxation for a global optimization problem.

For $p \in \mathbb{R}[x_1, \ldots, x_n]$:

$$\begin{aligned} p^\diamondsuit &:= \max \gamma \\ \text{s.t. } & p(x) - \gamma \text{ is SOS.} \end{aligned}$$

$p^\diamondsuit$ is a lower bound for the global minimum of p (where we usually assume that this minimum is finite). In many instances in practical applications, the exact value is found.

A nonzero-gap can be found, e.g., for the Motzkin polynomials. Consider

$$f(x,z) = M(x,1,z) = x^4 + x^2 + z^6 - 3x^2z^2 .$$

The global minimimum is 0 (which is attained for $(x,z) = (1,1)$). The best lower bound via SOS is

$$-\frac{729}{4096} \approx 0.17798 .$$

The corresponding SOS decomposition is

$$f(x,z) + \frac{729}{4096} = (-\frac{9}{8}z + z^3)^2 + \left(\frac{27}{64} + x^2 - \frac{3}{2}z^2\right)^2 + \frac{5}{32}x^2 .$$

An unbounded gap is possible, e.g., for

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

An improvement of the method (cf. the later sections for constrained opt.) would be to use representations of rational functions ($\rightsquigarrow$ Hilbert's 17th problem; connect!)

$$\begin{aligned} f(x,y) &= M(x,y,1) \\ &= \frac{(x^2y-y)^2 + (xy^2-x)^2 + (x^2y^2-1)^2 + \frac{1}{4}(xy^3-x^3y)^2 + \frac{3}{4}(xy^3+x^3y-2xy)^2}{x^2+y^2+1} \\ &\geq 0\,. \end{aligned}$$

No duality gap, duality result for extracting optimum

So far, we have not been much concerned with the amount of monomials needed for the SOS decomposition. In fact, the key to understanding this question quantitatively is the Newton polytope and the related sparsity issues which occur several times throughout this book.

Theorem 6.1.2. *If $p \in \mathbb{R}[X]$ can be written as $\sum q_i^2$ then $\mathrm{NP}(q_i) \subseteq \frac{1}{2}\,\mathrm{NP}(p)$ for $1 \leq i \leq m$.*

Proof. ... □

Example ...

In this part, our goal is to study polynomial optimization problems of the form

$$\begin{aligned} p_{\min} \;:=\; & \inf p(x) \\ & \text{s.t. } g_1(x) \geq 0, \ldots, g_m(x) \geq 0 \end{aligned}$$

with polynomials $p, g_1, \ldots, g_m \in \mathbb{R}[x_1, \ldots, x_n]$.

This class is a well-known "difficult" class of optimization problems. In general, these problems are non-convex optimization problems, and from the viewpoint of computational complexity these problems are in general **NP**-hard. Namely, e.g., the partition problem belongs to this class: Given $a_1, \ldots, a_m \in \mathbb{N}$, does there exist an $x \in \{-1,1\}^n$ with $\sum x_i a_i = 0$?

In the last years, an exciting development has taken place, showing how to approximate these problems in a hierarchical way using semidefinite programming and real algebraic geometry. The roots of this development go back to N.Z. Shor (1987), and the main developments of the SDP hierarchies have been initiated by A. Nemirovski, J. Lasserre and P. Parrilo. As we will see, these developments have been taken place in dual settings.

6.2 Basics of semidefinite programming

Wir first provide background on semidefinite programming, before we combine it with positive polynomials. Our point of departure is linear programming, one of the most fundamental tools in optimization. In the investigation of linear programs one usually starts from a normal form, such as

$$\begin{array}{lrcll} \inf & c^T x & & & \\ \text{s.t.} & Ax & = & b\,, & \\ & x & \geq & 0 & (x \in \mathbb{R}^n) \end{array} \tag{6.1}$$

with a matrix $A \in \mathbb{R}^{m\times n}$ and vectors $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$. If a linear optimization problem is given in a differing form, e.g. $\inf\{c^T x : Ax \leq b\}$, then by introducing additional variables it can be simply transferred in a normal form of type (6.1). With each linear program (6.1) one can associate a *dual program*

$$\begin{array}{lrcl} \sup & b^T y & & \\ \text{s.t.} & A^T y + s & = & c\,, \\ & s & \geq & 0\,. \end{array} \tag{6.2}$$

By a basic result of duality theory an admissible primal-dual pair $(x,(y,s))$ yields an optimal solution of the linear program if the Hadamard product $x \circ s = (x_i s_i)_{1\leq i\leq n}$ is the zero vector.

Linear programs can be solved both theoretically efficient (ellipsoid algorithm; Khachiyan, 1979) as well as practically efficient (simplex algorithm; Dantzig 1951). Moreover, in 1984 Karmarkar introduced the so-called *interior point methods* which are efficient as as well in the theoretical sense (polynomial time algorithm). Since this time, these methods were intensively developed further, such that meanwhile for large problems they can compete with the simplex algorithm also in practice.

The basic idea of primal-dual interior point methods can be explained as follows. Rather than aiminng directly on a primal-dual pair $(x,(y,s))$ with Hadamard produkt 0, we consider those primal-dual solution pairs, for which

$$x \circ s \;=\; \mu \mathbf{1}\,,$$

where $\mathbf{1}$ is the all-1-vector. For $\mu > 0$ this parameterizes a smooth, analytic curve of primal-dual solution pairs, which is known as *central path*. For $\mu \downarrow 0$ the central path converges to an optimal solution (x^*, y^*, s^*) of the linear program. Indeed, the limit point is contained in the relative interior of the set of all optimal solutions. The idea of interior point methods is, starting from an admissible primal-dual solution pair, to follow via numerical methods (with the Newton method as essential ingredient) the central path approximately. With this technique a linear optimization problem with rational input data can be solved in $O(\sqrt{n}\log(1/\varepsilon))$ arithmetic stepts up to a given precision $\varepsilon > 0$. For

the case of rational input data, these bounds imply the exact solvability in polynomial time.

From an abstract viewpoint the last inequality in (6.1) defines a cone. Thus linear programs can be seen as a special class of a *conic optimization problem*

$$\begin{array}{lrcl} \inf & c^T x & & \\ \text{s.t.} & Ax & = & b\,, \\ & x & \in & K \end{array} \tag{6.3}$$

with a cone K. Analogous to linear programs one can associated a dual program to (6.3):

$$\begin{array}{lrcl} \sup & b^T y & & \\ \text{s.t.} & A^T y + s & = & c\,, \\ & s & \in & K^*\,, \end{array} \tag{6.4}$$

where K^* denotes the dual cone to K. If K satisfies certain properties (closed, convex, pointed, nonempty interior), then for conic optimization problems as well a strong duality theorem holds (under the technical condition of the existence of a Slater point, compare condition (**??**) in Theorem **??**). We have seen above that the cone $\mathbb{R}^n_+$ is self-dual.

Semidefinite programming which has been studied intensively since the 90s, is a generalization of linear programming to matrix-based variables. First we choose a scalar product $\langle \cdot, \cdot \rangle$; usually the choice is $\langle A, B \rangle = \mathrm{Tr}(A \cdot B)$, on which the Frobenius norm $||A|| = (\sum_{i,j} a_{ij}^2)^{1/2}$ is based. Based on this, linear conditions can be specified. The condition that x is contained in the cone of nonnegative vectors, is replaced by the condition that a symmetric matrix $X \in \mathbb{R}^{n \times n}$ is *positive semidefinite*; this set defines as well a self-dual convex cone.

A normal form of a *semidefinite program (SDP)* is

$$\begin{array}{lrcl} \inf & \langle C, X \rangle & & \\ \text{s.t.} & \langle A_i, X \rangle & = & b_i\,, \quad 1 \le i \le m\,, \\ & X & \succeq & 0 \quad (X \in \mathbb{R}^{n \times n} \text{ symmetric})\,. \end{array} \tag{6.5}$$

A matrix $X \in \mathrm{Sym}_n(\mathbb{R})$ is called *(primal) feasible* if it satisfies all the constraints.

The optimal value is denoted by $\inf P^*$ (possibly $-\infty$), and we set $\inf P^* = \infty$ if there is no feasible solution.

From the viewpoint of optimization, these problems over the cone of positive semidefinite matrices are *convex* optimization problems.

6.2.1 Duality of semidefinite programs

To every SDP of the form (6.5) on can associate a dual SDP via

$$\text{(D)} \qquad \begin{array}{lrcl} \sup\limits_{y,S} & b^T y & & \\ & \sum\limits_{i=1}^m y_i A_i + S & = & C\,, \\ & S & \succeq & 0\,,\ y \in \mathbb{R}^m\,, \end{array}$$

whose optimal value is denoted by $\sup D^*$.
[1]

One often makes the assumptions that $A_1, \ldots, A_m$ are linearly independent. Then, in particular, y is uniquely determined by a dual feasible $S \in S_n^+$. Moreover one often assumes strict feasibility, i.e., that exists an $X \in \mathcal{P}$ and a $\mathcal{S} \in \mathcal{D}$ with $X \succ 0$ and $S \succ 0$. In particular, then Slater's condition from nonlinear programming is satisfied.

Let $X \in \mathcal{P}$ und $(y, S) \in \mathcal{D}$. Then

$$\operatorname{Tr}(CX) - b^T y$$

is called the duality gap of $(\mathcal{P})$ and $(\mathcal{D})$ in (X, y, S).

Theorem 6.2.1. (Weak duality theorem for SDP.) *Let $X \in \mathcal{P}$ und $(y, S) \in \mathcal{D}$. Then*

$$\operatorname{Tr}(CX) - b^T y \;=\; \operatorname{Tr}(SX) \;\geq 0\,.$$

Besides the weak duality statement, this theorem also gives an explicit description of the duality gap.

Proof. For any two feasible solutions X and (y, S) of the primal and the dual program we evaluate the difference

$$\operatorname{Tr}(CX) - b^T y \;=\; \operatorname{Tr}\left(\Big(\sum_{i=1}^m y_i A_i + S\Big)X\right) - \sum_{i=1}^m y_i \operatorname{Tr}(A_i X) \;=\; \operatorname{Tr}(SX)\,,$$

which is seen to be nonnegative by applying Féjer's Theorem A.4.3 on the positive semidefinite matrices S and X. □

Theorem 6.2.2. (Strong duality theorem for SDP). *Let $d^* < \infty$, and let the dual problem be strictly feasible. Then we have $\mathcal{P}^* \neq \emptyset$ and $p^* = d^*$.*

Analogously: Let $p^* > -\infty$, and let the primal problem be strictly feasible. Then $\mathcal{D}^* \neq \emptyset$ and $p^* = d^*$.

Note that the strict feasibility condition in the strong duality theorem is a specialization of Slater's condition in convex programming.
[2]

---

[1] Maybe the following notation becomes useful:

- Primal and dual feasibility region: $\mathcal{P}$, $\mathcal{D}$;
- sets of optimal solutions:

$$\begin{aligned} \mathcal{P}^* &:= \{X \in \mathcal{P} \,:\, \operatorname{Tr}(CX) = p^*\}\,, \\ \mathcal{D}^* &:= \{(S, y) \in \mathcal{D} \,:\, b^T y = d^*\}\,. \end{aligned}$$

[2] Notation S_n^+ ...

Proof. Let $d^* < \infty$ and let the dual problem (D) be strictly feasible. We can assume $b \neq 0$ since otherwise the dual objective function would be identically zero, thus implying the optimality of $X^* = 0$ for the primal problem (P). Define $M := \{S \in \mathrm{Sym}_n(\mathbb{R}) : S = C - \sum_{i=1}^m y_i A_i,\, b^T y \geq d^*\,, y \in \mathbb{R}^m\}$. The idea is to separate this convex set from the set of positive semidefinite matrices. The proof is carried out in three steps.

(1) We show that there exists a nonzero $Z \in \mathrm{Sym}_n(\mathbb{R})$ with $\sup_{S\in M} \mathrm{Tr}(SZ) \leq \inf_{U\in S_n^+} \mathrm{Tr}(UZ)$. We first observe that $\mathrm{relint}(M) \cap \mathrm{relint}(S_n^+) = \emptyset$, because the existence of some $S \in M \cap S_n^{++}$ would contradict the optimal value d^* of (D).

Identify $\mathrm{Sym}_n(\mathbb{R})$ with $\mathbb{R}^{\frac{1}{2}n(n+1)}$, where the scalar product is canonically induced from the scalar product on $\mathrm{Sym}_n(\mathbb{R})$ given by $\langle A, B\rangle = \sum_i a_{ii}b_{ii} + \sum_{i<j} 2a_{ij}b_{ij}$. By a standard separation theorem from convex analysis (see, e.g., [76, Cor. 11.4.1]) there exists a nonzero $Z \in \mathrm{Sym}_n(\mathbb{R})$ with $\sup_{S\in M} \mathrm{Tr}(SZ) \leq \inf_{U\in S_n^+} \mathrm{Tr}(UZ)$. Since $\mathrm{Sym}(\mathbb{R})^+$ is a cone, the right hand side must either be 0 or $-\infty$, where the latter possibility is ruled out by $M \neq \emptyset$.

Moreover, the statement $\inf_{U\in S_n^+} \mathrm{Tr}(UZ) = 0$ (which by Féjer implies $Z \succeq 0$) yields $\sup_{S\in M} \mathrm{Tr}(SZ) \leq 0$.

(2) We show that there exists some $\beta > 0$ with $\mathrm{Tr}(A_i Z) = \beta b_i$ for all $i \in \{1, \ldots, m\}$. First observe that on the halfspace $\{y \in \mathbb{R}^m : b^T y \geq d^*\}$, the linear function $f(y) := \sum_{i=1}^m y_i \mathrm{Tr}(A_i Z)$ is bounded from below (by $\mathrm{Tr}(CZ)$).

Let $y \in \mathbb{R}^m$ (where y uniquely determines an $S \in M$) with $b^T y \geq d^*$. Then

$$f(y) \;=\; \sum_{i=1}^m y_i \mathrm{Tr}(A_i Z) \;=\; -\mathrm{Tr}((S-C)Z) \;=\; -\mathrm{Tr}(SZ) + \mathrm{Tr}(CZ) \;\geq\; \mathrm{Tr}(CZ)\,.$$

Hence there exists a $\beta \geq 0$ such that $\mathrm{Tr}(A_i Z) = \beta b_i$ for all $i \in \{1, \ldots, m\}$ (since otherwise one can make f smaller on the halfspace.)

Assuming $\beta = 0$ would imply $\mathrm{Tr}(A_i Z) = 0$, $1 \leq i \leq m$, and therefore $\mathrm{Tr}(CZ) \leq 0$. By assumption there exist a $(y^\circ, S^\circ) \in \mathcal{D}$ with $S^\circ \succ 0$. Hence,

$$\mathrm{Tr}(S^\circ Z) \;=\; \mathrm{Tr}(CZ) - \sum_{i=1}^m y_i^\circ \mathrm{Tr}(A_i Z) \;=\; \mathrm{Tr}(CZ) \;\leq\; 0\,.$$

This is a contradiction, since $Z \succeq 0$ and $S^\circ \succ 0$ imply that $\mathrm{Tr}(S^\circ Z) > 0$ (due to Féjer, continuity, $Z \neq 0$). Hence, $\beta > 0$.

(3) Finally, the goal is to show that for $X^* := \frac{1}{\beta} Z$ we have $X^* \in \mathcal{P}$ and $\mathrm{Tr}(CX^*) = d^*$. We have $\mathrm{Tr}(A_i X^*) = b_i$ for $1 \leq i \leq m$ i.e., $X^* \in \mathcal{P}$). Hence, $\mathrm{Tr}(CX^*) \leq b^T y$ for all $y \in \mathbb{R}^m$ with $b^T y \geq d^*$, and further $\mathrm{Tr}(CX^*) \leq d^*$. The weak Duality Theorem implies $\mathrm{Tr}(CX^*) = d^*$, i.e., $X^* \in \mathcal{P}^*$.

The statement for which $p^* > -\infty$ and strict feasibility of the primal problem is assumed, can be proven analogously, or by exploiting symmetric (conic) formulations of the problems. □

Strong duality can fail. Indeed, consider the example

$$\sup -x_1 \text{ s.t. } \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix} x_1 + \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} x_2 \;\preceq\; \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Equivalently, a point (x_1, x_2) is feasible if and only if $\begin{pmatrix} y_1 & 1 \\ y & y_2 \end{pmatrix} \succeq 0$, i.e., if and only if $x_1 > 0$, $x_2 > 0$ and $x_1 x_2 > 1$.

The dual program is

$$\min \left\langle \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^T, X \right\rangle \text{ s.t. } \left\langle \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}^T, X \right\rangle = -1, \quad \left\langle \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}^T, X \right\rangle = 0, \quad X \succeq 0$$

which has only the feasible solution $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. Thus there is no duality gap, but the primal program does not attain the optimal value.

6.2.2 Algorithms

In 1991, Nesterov und Nemirovski and independently Alizadeh showed that interior point methods can be efficiently extended to semidefinite programs. Actually, Nesterov and Nemirovski investigated this connection in the extended context of conic optimization; for his resarch contributions Yurii Nesterov was awarded in 2000 the Dantzig Prize.[3] These developments on numerical optimization were strengthened by astonishing SDP-based approximation algorithms by Goemans and Williamson (0.878-approximation algorithm for MAX CUT[4]

Generalizing the interior point principle for linear programming, in interior point methods for semidefinite programs one considers not only primal-dual solution pairs, whose duality gap is 0, but the curve of all those pairs, which have the property $XS = \mu \cdot I_n$ with the unit matrix I_n and $\mu > 0$.f

—

Beyond these algorithmic properties, SDP has nice features: important special cases (linear programming, quadratic programming); important and partially surprisingly good applications in combinatorial optimization global optimization approximation theory control theory portfolio optimization distance geometry problems in mocelular biology, ...

Special classes of semidefinite optimization: linear programming, by restricting X onto diagonal matrices; convex-quadratic functions with convex-quadratic constraints, special case: "quadratic programmierung" (quadratic objective function; linear constraints).

Exercises

Exercise 6.2.3. Let $A \in \mathrm{Sym}(\mathbb{R})$ be positive definite. Show that if $B \in \mathrm{Sym}(R)$ is positive semidefinite with $\langle A, B \rangle = 0$ then $B = 0$.

[3] A. Nemirovski had obtained this prize already in 1991 for his earlier contributions to convex optimization.

[4] awarded with the Fulkerson Prize 2000.

Exercise 6.2.4. Let $A(x)$ be a symmetric real matrix, depending affinely on a vector x. Show that the problem of x in order to minimize the maximum eigenvalue can be phrased as semidefinite program.

Exercise 6.2.5. Show that the SDP in standard form with

$$C = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

and $b_1 = 0$, $b_2 = 2$ attains both its primal optimal value and its primal dual value, but has a duality gap of 1.

Exercise 6.2.6. Show that the SDP in standard form with

$$C = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

and $b_1 = 0$, $b_2 = 2$ is infeasible, but its dual program has a finite optimal value which is actually attained.

6.3 Sums of squares and semidefinite programming

Our two main actors *positive polynomials and semidefinite programming* now come together by the observation that the SOS-property of a polynomial can be formulated by a semidefinite program. For this, we consider a $p \in \mathbb{R}[X_1, \ldots, X_n]$ of even degree $2d$. Denote by Y the vector of all monomials in $X_1, \ldots, X_n$ of degree at most d; thus Y consists of $\binom{n+d}{d}$ components. In the following we identiy a polynomial $s = s(X)$ with the vector of its coefficients. A polynomial p is a sum of squares

$$p = \sum_j (s_j(X))^2 \quad \text{mit Polynomen } s_j \text{ vom Grad höchstens } d\,,$$

if and only if for its coefficient vectors s_j of the polynomials $s_j(X)$ we have

$$p \;=\; Y^T \big(\sum_j s_j s_j^T \big) Y\,.$$

By the Choleski decomposition of a matrix this is the case if and only if the matrix defined by $\sum_j s_j s_j^T$ is positive semidefinite. For deciding the SOS property via semidefinite programming we can therefor state:

Lemma 6.3.1. *A polynomial $p \in \mathbb{R}[X_1, \ldots, X_n]$ of degree $2d$ is a sum of squares if and only if there exists a positive semidefinite matrix Q with*

$$p \;=\; Y^T Q Y\,.$$

This is a system of matrix variables Y of order $\binom{n+d}{d}$, with $\binom{n+2d}{2d}$ equations; for fixed d or n this size is polynomial.

6.3.1 Semidefinite programming and sums of squares

For $t \in \mathbb{N}$, let $S_t = \{\alpha \in \mathbb{N}_0^n \,:\, \alpha \in \mathbb{N}_0^n \,:\, \sum_{i=1}^n \alpha_i \leq t\}$ be the set of monomials of total degree at most t.

Consider a polynomial $p \in \mathbb{R}[X_1, \ldots, X_n]$ of even degree $2d$. Let Y denote the vector of all monomials in $X_1, \ldots, X_n$ of degree at most d; Y consists of $\binom{n+d}{d}$ components. In the following, we identify a polynomial $s = s(X)$ with a vector of its coefficients. A polynomial p is a sum of squares,

$$p = \sum_j (s_j(X))^2 \quad \text{with polynomials } s_j \text{ of degree at most } d\,,$$

if and only if the coefficient vectors s_j of the polynomials $s_j(X)$ satisfy

$$p \;=\; Y^T \big(\sum_j s_j s_j^T \big) Y\,.$$

By the Choleski decomposition of a matrix this is the case if and only if the matrix $\sum_j s_j s_j^T$ is positive semidefinite. For deciding the SOS-property via semidefinite programming we record:

Lemma 6.3.2. *A polynomial $p \in \mathbb{R}[X_1, \ldots, X_n]$ of degree $2d$ is a sum of squares if and only if there exists a positive semidefinite matrix Q with*

$$p \;=\; Y^T Q Y\,.$$

The size of the SDP (i.e., #rows = # colums of X) : $\binom{n+d}{d}$. The number of equations is $\binom{n+2d}{d}$. Hence, this number is polynomial if n or d is fixed.

Hence, deciding the decomposition of an SOS decomposition is an SDP-feasibility problem.

Remark 6.3.3. The complexity of the ("exact") semidefinite feasibility problem SDFP in the Turing machine model (i.e., is SDFP $\in \mathbf{P}$? is still open and one of the most important open problems concerning the complexity of SDP. If the dimension n or the number of constraints m are constants, then SDFP is decidable in polynomial time. (Porkolab, Khachiyan '97). Hence, if n or d is fixed, then deciding an SOS decomposition can be done in polynomial time.

6.4 Semidefinite relaxations for constrained optimization

In this section we discuss a relaxation scheme due to Lasserre for the general constrained optimization problem

$$\begin{array}{lrcl} \inf & p(x) & & \\ \text{s.t. } & g_i(x) & \geq & 0\,, \quad 1 \leq i \leq m\,. \\ & x & \in & \mathbb{R}^n \end{array} \tag{6.6}$$

For convenience we set $g_0 := 1$. We assume that the module $\mathrm{QM}(g_1, \ldots, g_m)$ is Archimedean as introduced in Section **??**. Hence, there exists an $N \in \mathbb{N}$ with $N - X_1^2 \in \mathrm{QM}(g_1, \ldots, g_m)$. From an applied viewpoint this is no problem, since we can just add an inequality $\sum x_i^2 \leq N$ with a large N which causes that only solutions in a large ball around the origin are considered. Since in case of the mentioned precondition K is compact, Putinar's Positivstellensatz implies

$$\begin{aligned} p^* &= \sup \gamma \\ &\quad \text{s.t. } p(x) - \gamma \in \mathrm{QM}(g_1, \ldots, g_m)\,. \end{aligned}$$

via SDP

6.5 Duality and the moment problem

We have already seen in Section 6.2 that the concept of duality is fundamental for optimization. In the case of positive polynomials, duality will establish a connection to the classical *moment problem*. In the following assume that K is a closed set in $\mathbb{R}^n$, possibly $K = \mathbb{R}^n$. Given a sequence $y = (y_\alpha)_{\alpha \in \mathbb{N}_0^n}$, the moment problem asks for necessary and sufficient conditions on the sequence (y_α) concerning the the existence of a Borel measure μ with

$$y_\alpha \;=\; \int_K x^\alpha \, d\mu \,.$$

To begin with revealing this connection, recall that the dual cone C^* of a cone C in some finite-dimensional vector space is defined by

$$C^* \;=\; \{x \;:\; \langle x, y\rangle \geq 0 \text{ for all } y \in C\} \,,$$

where $\langle \cdot, \cdot \rangle$ denotes the usual dot product.

The set of nonnegative polynomials (on $\mathbb{R}^n$ or on a closed set K) defines a convex cone. For fixed number of variables n and fixed degree d, this cone lives in an ambient space of finite dimension. We would like to understand the dual cone of it. As a warm-up example, it is very instructive to consider the cone

$$C_d \;=\; \{p \in \mathbb{R}[x] \;:\; \deg p \leq d, p \geq 0\}$$

of non-negative univariate polynomials of even degree at most d. Let

$$\mathcal{M}_d = \left\{y = (y_0, \ldots, y_d)^T : y_i = \int_{\mathbb{R}} x^i \, d\mu \text{ for some Borel measure } \mu \text{ on } \mathbb{R}\right\}$$

be the *moment cone of order d*.

Exercise 6.5.1. For even d, show that that $\mathcal{M}_d$ is not closed. Hint: Consider the moment sequence $(1, \epsilon, 1/\epsilon)$ coming from a normal distribution with mean ϵ and variance $\epsilon - 1/\epsilon^2$.

For simplicity, we identify a univariate polynomial p of degree d with its coefficient vector $(p_0, \ldots, p_d)$. Denote by

$$\pi_d : \mathbb{R} \to \mathbb{R}^d \,, \quad t \mapsto (1, t, t^2, \ldots, t^d)^T \in \mathbb{R}^{d+1}$$

the *moment mapping of order d*. As shown in the following exercise, the image of π_d constitutes the set of extreme rays of $\mathcal{M}_d$.

Exercise 6.5.2. For even d, show that $\mathcal{M}_d = \operatorname{pos}\{\pi_d(t) \;:\; t \in \mathbb{R}\}$, where pos denotes the positive hull.

Lemma 6.5.3. *For even d, the dual cone C_d^* satisfies $C_d^* = \operatorname{cl} \mathcal{M}_d$ and $C_d = \mathcal{M}_d^*$, where* cl *denotes the topological closure.*

Proof. For the first equality, it suffices to prove

$$C_d^* \;=\; \mathrm{cl}\,\mathrm{pos}\{\pi_d(t) \,:\, t \in \mathbb{R}\} \tag{6.7}$$

due to the previous exercise. For any non-negative polynomial p of degree at most d and every $x \in \mathbb{R}$ we have $p(x) = p^T \pi_d(x) \geq 0$. Denoting the right hand side of (6.7) by C', this implies $C' \subset C_d^*$. Conversely, assume that there exists a $z \in C_d^* \setminus C'$. By the Separation Theorem, there exists a $b \in \mathbb{R}^{d+1}$ with $b^T z < 0 \leq b^T z'$ for all $z' \in C'$. By the right inequality, the polynomial defined by the coefficient vector b is non-negative, which then contradicts the left inequality.

To show $C_d = \mathcal{M}_d^*$, first observe that any $p \in \mathcal{M}_d^*$ satisfies $\sum_{i=0}^d p_i y_i \geq 0$ for all $\mathcal{M}_d$. In particular, if μ is a measure concentrated on the single point t (i.e., a multiple of the Dirac measure) then we have $\sum_{i=0}^d p_i t_i \geq 0$ for all $t \in \mathbb{R}$, i.e., $p \geq 0$. Conversely, let $p \in C_d$ and $y \in \mathcal{M}_d$. Then

$$p^T y \;=\; \sum_{i=0}^d p_i y_i \;=\; \int p(x) d\mu \;\geq\; 0\,,$$

i.e., $y \in \mathcal{M}_d^*$. □

To provide a an explicit characterization of the dual cone C_d^*, for a given $z = (z_0, z_1, \ldots, z_d)^d$ define the symmetric Hankel matrix

$$H_d(z) \;=\; \begin{pmatrix} z_0 & z_1 & z_2 & \cdots & z_{d/2} \\ z_1 & z_2 & z_3 & \cdots & z_{d/2+1} \\ z_2 & z_3 & z_4 & & z_{d/2+2} \\ \vdots & \vdots & & \ddots & \vdots \\ z_{d/2} & z_{d/2+1} & z_{d/2+2} & \cdots & z_d \end{pmatrix}.$$

Theorem 6.5.4. *For even d, we have*

$$C_d^* \;=\; \{z \in \mathbb{R}^{d+1} \,:\, H_d(z) \succeq 0\}\,.$$

Proof. Let $C' = \{z \in \mathbb{R}^{d+1} \,:\, H_d(z) \succeq 0\}$. In order to show $C_d^* \subset C'$, we start from $C_d^* \subset \mathrm{cl}\,\mathrm{pos}\{\pi_d(t) \,:\, t \in \mathbb{R}\}$ stated in Theorem 6.5.3. For every $t \in \mathbb{R}$ observe the decomposition $H_d(\pi_d(t)) = \pi_d(t)\pi_d(t)^T \succeq 0$, which yields $\pi_d(t) \in C'$. Linearity of H_d then implies $C_d^* \subset C'$.

Conversely, let $z \in C'$ and p be an arbitrary univariate polynomial of degree at most d. Writing p as a sum of squares $p = \sum_j (q^{(j)})^2$ with polynomials $q^{(j)}$, we obtain $p^T z = \sum_j ((q^{(j)})^2)^T z$. Since for an arbitrary polynomial q of degree at most $d/2$ we have

$$(q^2)^T z \;=\; \sum_{i=0}^d z_i \sum_{j+k=d} q_j q_k \;=\; q^T H_d(z) q\,,$$

we can conclude

$$p^T z \;=\; \sum_j (q^{(j)})^T H_d(z) q^{(j)} \;\geq\; 0$$

by the positive semidefiniteness of $H_d(z)$. □

In the multivariate situation, set

$$\begin{aligned} \mathcal{P}_n &= \{p \in \mathbb{R}[x_1, \ldots, x_n] \,:\, p(x) \geq 0 \text{ for all } x \in \mathbb{R}^n\}\,, \\ \Sigma_n &= \{p \in \mathbb{R}[x_1, \ldots, x_n] \,:\, p \text{ is SOS}\} \end{aligned}$$

denote the set of polynomials which are nonnegative on $\mathbb{R}^n$. These are convex cones in the infinite-dimensional vector space $\mathbb{R}[x_1, \ldots, x_n]$.

We can identify an element $\sum_\alpha c_\alpha x^\alpha$ in the vector space $\mathbb{R}[x_1, \ldots, x_n]$ with its coefficient vector (c_α); The dual space of $\mathbb{R}[x_1, \ldots, x_n]$ consists of the set of linear mappings on $\mathbb{R}[x_1, \ldots, x_n]$ and each such vector can be identified with a vector in the infinite dimensional space $\mathbb{R}^{\mathbb{N}_0^n}$. Topologically, $\mathbb{R}^{\mathbb{N}_0^n}$ is a locally convex space in the topology of pointwise convergence. We identify the dual space of a space $X \subset \mathbb{R}^{\mathbb{N}_0^n}$ with a subspace of $\mathbb{R}^{\mathbb{N}_0^n}$.

In order to characterize the dual cone $\mathcal{P}_n^*$, let $\mathcal{M}_n$ denote the set of (infinite) sequences $y = (y_\alpha)_{\alpha \in \mathbb{N}_0^n}$ admitting a representing Borel measure.

Theorem 6.5.5. *The cones $\mathcal{P}_n$ and $\mathcal{M}_n$ are dual to each other, i.e.,*

$$\mathcal{P}_n^* \;=\; \mathcal{M}_n\,, \quad \mathcal{M}_n^* \;=\; \mathcal{P}_n\,.$$

A main goal of this section is to enlighten Theorem 6.5.5, and to prove most inclusions of it.

First note that we can identify a sequence $y = (y_\alpha)_{\alpha \in \mathbb{N}_0^n}$ with a linear map $L : \mathbb{R}[x] \to \mathbb{R}$, since any such linear map L is uniquely determined by its images on the basis elements x^α. Hence, a vector $y = (y_\alpha)_{\alpha \in \mathbb{N}_0^n}$ admits a representing measure if and only a linear map $L : \mathbb{R}[x] \to \mathbb{R}$ comes from a measure μ via

$$L(f) \;=\; \int f d\mu \quad \text{for all } f \in \mathbb{R}[x]\,.$$

Lemma 6.5.6. *For a linear map $L : \mathbb{R}[x] \to \mathbb{R}$ and a polynomial $g \in \mathbb{R}[x]$, the following statements are equivalent:*

1. *$L(\sigma g) \geq 0$ for all $\sigma \in \Sigma[x]$;*
2. *$L(h^2 g) \geq 0$ for all $h \in \mathbb{R}[x]$;*
3. *The symmetric bilinear form $\langle \cdot, \cdot \rangle_g$ defined by $\langle p, q \rangle_g := L(pqg)$ is positive semidefinite.*

Proof. Clearly, the first condition implies the second one, and vice versa, the second one implies the first one via linearity. In order to see the equivalence of (2) and (3), simply observe that $\langle h, h\rangle_g = L(h^2 g)$. □

With regard to the canonical monomial basis, the symmetric bilinear form can be identified with an infinite matrix, whose rows and colums are indexed by vector $a \in \mathbb{N}_0^n$. For the choice $g := 1$, observe that the entry in row α and column β is

$$\langle x^\alpha, x^\beta\rangle_1 \;=\; L\left(x^{\alpha+\beta}\right) = y_{\alpha+\beta}\,.$$

If g is a general polynomial of the form $g = \sum_\gamma c_\gamma x^\gamma$, then the entry in row α and column β is

$$\langle x^\alpha, x^\beta\rangle_g \;=\; L\big(\sum_\gamma c_\gamma x^{\alpha+\beta+\gamma}\big) \;=\; \sum_\gamma c_\gamma y_{\alpha+\beta+\gamma}\,.$$

Proof. (of Theorem 6.5.5). We first consider the equality $\mathcal{M}_n^* = \mathcal{P}_n$. For each $p = \sum_\alpha c_\alpha x^\alpha \in \mathcal{M}_n^*$, by definition we have $\sum_\alpha c_\alpha y_\alpha \geq 0$ for all $y \in \mathcal{M}_n$. In particular this also holds true for the Dirac measure δ_x concentrated at a point x, which implies $\sum_\alpha c_\alpha x^\alpha \geq 0$ for all $x \in \mathbb{R}^n$. Hence, $p \in \mathcal{P}_n$.

Conversely, let $p = \sum_\alpha c_\alpha x^\alpha \in \mathcal{P}_n$. For each $y \in \mathcal{M}_n$ there exists a Borel measure μ with $y_\alpha = \int x^\alpha \, d\mu$, which implies

$$\sum_\alpha c_\alpha y_\alpha \;=\; \int p(x)\; d\mu \;\geq\; 0\,,$$

i.e., $p \in \mathcal{M}_n^*$.

Concerning the equality $\mathcal{M}_n = \mathcal{P}_n^*$, the inclusion $\mathcal{M}_n \subset \mathcal{P}_n^*$ is rather straightforward. Namely, for a moment sequence (y_α) and any $p = \sum_\alpha c_\alpha x^\alpha \in \mathcal{P}_n$ we clearly have $\sum_\alpha c_\alpha y_\alpha \geq 0$ by the non-negativity of p.

The converse direction $\mathcal{P}_n^* \subset \mathcal{M}_n$ is more involved and known as Haviland's Theorem. See the discussion afterwards, but we do not present the proof here. □

If for a Borel measure μ on a set K, a function $L : \mathbb{R}[x] \to \mathbb{R}$ is defined as $L(p) = \int_K p \, d\mu$ for all $p \in \mathbb{R}[x]$, then clearly $L(p) \geq 0$ for all polynomials p which are non-negative on K. Haviland's Theorem states that the converse of this statement holds as well, thus providing an exact characterization of linear maps coming from a measure.

Theorem 6.5.7 (Haviland)**.** *Let $K \subset \mathbb{R}^n$ be a measurable set. For a linear map $L : \mathbb{R}[x] \to \mathbb{R}$, the following statements are equivalent:*

1. *There exists a Borel measure μ with $L(p) = \int_K p \, d\mu$ for all $p \in \mathbb{R}[x]$.*

2. *$L(p) \geq 0$ for all $p \in \mathbb{R}[x]$ which are non-negative on K.*

For a proof of the statement we refer to the original paper [**?**] or to Marshall's book [57]. The univariate special cases $K = [0, \infty)$, $K = \mathbb{R}$, $K = [0, 1]$ (cf. Theorem 6.5.3) were proven earlier by Stieltjes, Hamburger, and Hausdorff.

Reminding the reader once more of the important connection between non-negative polynomials and sum of squares, we now turn towards the dual cone of the cone of sums of squares Σ_n.

In order to characterize the dual cone $(\Sigma_n)^*$, consider a real sequence $(y_\alpha)_{\alpha \in \mathbb{N}_0^n}$ indexed by non-negative integer vectors. let $\mathcal{M}_n^+ := \{y = (y_\alpha)_{\alpha \in \mathbb{N}_0^n} \,:\, M(y) \succeq 0\}$, where $M(y)$ is the (infinite) *moment matrix* $(M(y))_{\mathbb{N}_0^n \times \mathbb{N}_0^n}$ with

$$(M(y))_{\alpha,\beta} \;=\; y_{\alpha+\beta}\,.$$

Observe that in the univariate situation $M(y)$ with is an infinite Hankel matrix ,

$$M(y) \;=\; \begin{pmatrix} y_0 & y_1 & y_2 & \cdots \\ y_1 & y_2 & & \\ & & \ddots & \\ y_2 & & & \\ \vdots & & & \end{pmatrix}.$$

Lemma 6.5.8. *For $n \geq 1$, it holds $\mathcal{M}_n \subset \mathcal{M}_n^+$.*

For $t \in \mathbb{N}_0^n$, let $(y_\alpha)_{\alpha \in S_t}$ be the sequence of moments of μ up to order t.

Proof. Let μ be a representing measure for $y \in M$. Let $t \in \mathbb{N}_0$, $p(x) \in \mathbb{R}[x_1, \ldots, x_n]$ with degree $\leq t$, and $p \in \mathbb{R}^{S_t}$ be its vector of coefficients. The statement follows from

$$\begin{aligned} p^T M_t(y) p \;&=\; \sum_{\alpha,\beta \in S_t} p_\alpha p_\beta y_{\alpha+\beta} \;=\; \sum_{\alpha,\beta \in S_t} p_\alpha p_\beta \int x^{\alpha+\beta} d\mu \\ &=\; \int p(x)^2 d\mu \;\geq\; 0\,. \end{aligned}$$

□

We record the following classical result for the univariate and bivariate situation.

Corollary 6.5.9. *(Hamburger.) In the univariate situation, the cones $\mathcal{M}_n$ and $(\mathcal{M}_n^+)$ coincide. For $n \geq 2$, we have $\mathcal{M} \neq (\mathcal{M}_n^+)$.*

Proof. In the univariate situation, we have $\operatorname{cl} \mathcal{M}_1 = \mathcal{M}_1^+$. Thus the statement follows from Theorems 6.5.3 and 6.5.4. Indeed, it corresponds to the coincidence of the dual cones $\mathcal{P}_1$ and Σ_1.

The difference for $n \geq 2$ follows the fact that $\Sigma_n \neq \mathcal{P}_n$, provided by Hilbert's Theorem 5.2.3. □

For the multivariate situation, we record the following result, whose proof is partially covered in the exercises.

Theorem 6.5.10. *The cones Σ_n and $\mathcal{M}_n^+$ are dual to each other, i.e.*

$$\Sigma_n^* \;=\; \mathcal{M}_n\,, \quad (\mathcal{M}_n^+)^* \;=\; \Sigma_n\,.$$

Exercises

Exercise 6.5.11. For univariate polynomials of degree at most 4, show that the vector $z = (1, 0, 0, 0, 1)^T$ satisfies $H(z) \succeq 0$, but that it is not contained in $\mathcal{M}_d$.

6.6 Primal-dual semidefinite relaxations

The idea is to consider a primal-dual pair of convexifications of the problem. Since by the considerations in Section **??**, the cone of nonnegative polynomials is dual to the moment cone, it should not come as a surprise now that the primal-dual pair of optimization problems will involve the moment cone as well as the cone of nonnegative polynomials.

Let us convexify the problem by considering the equivalent formulation

$$p^* \;=\; \min_{x\in K} p(x) \;=\; \min_{\mu\in\mathcal{P}(K)} \int p(X)d\mu\,, \tag{6.8}$$

where $\mathcal{P}(K)$ denotes the set of all probability measures μ supported on the set K. The task now is to describe the measures in question.

Theorem 6.6.1 (Putinar)**.** *Suppose the module* $\mathrm{QM}(g_1,\ldots,g_m)$ *is Archimedean. A linear map* $L:\mathbb{R}[X]\to\mathbb{R}$ *is the integration with respect to a probability measure* μ *on* K *i.e.*

$$\exists\mu\;\forall p\in\mathbb{R}[X]:\; L(p)=\int_K p\,d\mu, \tag{6.9}$$

if and only if $L(1)=1$ *and all the bilinear forms*

$$\mathcal{L}_{g_i}:\mathbb{R}[X]\times\mathbb{R}[X] \;\to\; \mathbb{R}\,,\quad (p,q)\;\mapsto L(p\cdot q\cdot g_i) \tag{6.10}$$

$(0\le i\le m)$ *are positive semidefinite*

Before we start with the main part of the technical proof, note that if μ is a probability measure satisfying (6.9) then clearly the integral

$$\mathcal{L}_{g_i}(p,p) \;=\; L(p^2g_i) \;=\; \int_K p^2g_i\,d\mu$$

is nonnegative since μ is supported on K and g_i is nonnegative on K. Hence, the bilinear form $\mathcal{L}_{g_i}$ is positive semidefinite.

For the converse direction, let all the bilinear forms (6.10) by positive semidefinite. Our goal is show that the conditions of Riesz's Representation Theorem are satisfied, thus requiring us to work within the ring $\mathcal{C}(K,\mathbb{R})$ of of continous functions $K\to\mathbb{R}$. Formally, we will make use of the ring homomorphism

$$\varphi:\mathbb{R}[x]\to\mathcal{C}(K,\mathbb{R})\,,\quad p\mapsto p|_K\,. \tag{6.11}$$

Proof. For any non-negative polynomial p on K and any $\varepsilon>0$ Putinar's Positivstellensatz 5.6.5 implies that $p+\varepsilon\in\mathrm{QM}(g_1,\ldots,g_m)$ and thus by the precondition $L(p+\varepsilon)\ge 0$. By letting ε converge to zero, continuity implies $L(p)=L(p+\varepsilon)-\varepsilon L(1)=L(p+\varepsilon)-\varepsilon\ge 0$ as well.

Any element in the kernel of φ maps to zero under L. Thus, writing shortly $\hat{p} := \varphi(p)$, L induces a well-defined linear map $\hat{L} : \varphi(\mathbb{R}[x]) \to \mathbb{R}$, $\hat{L}(\hat{p}) \mapsto L(p)$ for $p \in \mathbb{R}[x]$.

Let $||\cdot||$ denote the supremum norm on $\mathcal{C}(K, \mathbb{R})$. For every $p \in \mathbb{R}[x]$, the functions $||\hat{p}|| \pm \hat{p}$ are nonnegative on K, so that our initial considerations imply $L(||\hat{p}|| \pm p) \geq 0$ and thus by linearity $\hat{L}(\hat{p}) \leq L(1)\,\|\hat{p}\| = \|\hat{p}\|$. As a consequence, the map L is continuous.

The map $\hat{L}$ can be extended to a linear map on $\mathcal{C}(K, \mathbb{R})$, since it can be shown that the image $\varphi(\mathbb{R}[x])$ of the map φ is dense in $\mathcal{C}(K, \mathbb{R})$. (For details on this density argument see [72, Theorem 5.2.6].) In order to see that $\hat{L}(f)$ is nonnegative for any given nonnegative $f \in \mathcal{C}(K, \mathbb{R})$, consider an $\varepsilon > 0$ and $p \in \mathbb{R}[x]$ such that the function $f_\varepsilon + \varepsilon$ satisfies $\|f_\varepsilon - \hat{p}\| < \varepsilon$. Then $\hat{p} > 0$ on K, so that $p \in \mathrm{QM}(g_1, \ldots, g_m)$ and further $\hat{L}(\hat{p}) = L(p) \geq 0$. Since $||f_\varepsilon - \hat{p}\| < \varepsilon$ we have $||f - \hat{p}\| < 2\varepsilon$, showing that f is in the closure of $\varphi(\mathrm{QM}(g_1, \ldots, g_m))$. Now $\hat{L}(f) \geq 0$ follows from the continuity of $\hat{L}$.

Ultimately, by the Riesz Representation Theorem (see, e.g., [77, Theorem 2.14]), there exists a probability measure μ supported on K with $\hat{L}(f) = \int_K f\, d\mu$ for any $f \in \mathcal{C}(K, \mathbb{R})$. In particular, this implies $L(p) = \hat{L}(\hat{p}) = \int_K \hat{p}\, d\mu = \int_K p\, d\mu$ for any $p \in \mathbb{R}[x]$. □

Applying Theorem 6.6.1, we can restate (6.8) as

$$p^* \;=\; \inf\{L(p) \;:\; L : \mathbb{R}[x] \to \mathbb{R} \text{ linear, } L(1) = 1 \text{ and each } \mathcal{L}_{g_i} \text{ is psd}\}\,. \tag{6.12}$$

Since every linear map $L : \mathbb{R}[x] \to \mathbb{R}$ is given by the values $L(x^\alpha)$ on the monomial basis $(x^\alpha)_{\alpha \in \mathbb{N}_0^n}$, Theorem 6.6.1 characterizes the families $(q_\alpha)_{\alpha \in \mathbb{N}_0^n}$ which arise as the sequences of moments of a probability measure on K, i.e. $(q_\alpha) = \int_K x^\alpha\, d\mu$ for every $\alpha \in \mathbb{N}_0^n$. Therefore Theorem 6.6.1 is also said to solve the *moment problem* on K.

Alternatively, and we will make use of this viewpoint below, the linear map $L : \mathbb{R}[x] \to \mathbb{R}$ can be interpreted (with regard to the natural monomial bases) as an infinite matrix $(L(X^\alpha)_{\alpha \in \mathbb{N}_0^n}$.

Now fix any (ordered) basis $\mathcal{B} = \{b_1, b_2, \ldots\}$ of the vector space $\mathbb{R}[x]$ (for example the monomial basis $(x^\alpha)_{\alpha \in \mathbb{N}_0^n}$ and consider the infinite-dimensional *moment matrix* $M(y)$ defined by

$$M(y)_{i,j} \;:=\; L(b_i \cdot b_j)\,.$$

As an example consider for $n = 2$ the following finite piece of $M(y)$ which only refers to the subset $\mathcal{B}' = \{x^{(0,0)}, x^{(1,0)}, x^{(0,1)}, x^{(2,0)}, x^{(1,1)}, x^{(0,2)}\} \subset \mathcal{B}$ of monomials with total degree at most 2:

$$\left(\begin{array}{c|cc|ccc} 1 & y_{10} & y_{01} & y_{20} & y_{11} & y_{02} \\ \hline y_{10} & y_{20} & y_{11} & y_{30} & y_{21} & y_{12} \\ y_{01} & y_{11} & y_{02} & y_{21} & y_{12} & y_{03} \\ \hline y_{20} & y_{30} & y_{21} & y_{40} & y_{31} & y_{22} \\ y_{11} & y_{21} & y_{12} & y_{31} & y_{22} & y_{13} \\ y_{02} & y_{12} & y_{03} & y_{22} & y_{13} & y_{04} \end{array}\right)\,.$$

Furthermore for each g_k define in an analogous manner the *localizing matrix* $M(g_k \cdot y)$ by

$$M(g_k \cdot y)_{i,j} \;:=\; L(g_k \cdot b_i \cdot b_j)\,.$$

In this language of moments, Theorem 6.6.1 can be stated as:

Theorem 6.6.2 (Putinar, moment matrix version)**.** *Suppose the module* $\mathrm{QM}(g_1,\dots,g_m)$ *is Archimedean. An infinite sequence* $y=(y_\alpha)_{\alpha\in\mathbb{N}_0^n}$ *of real numbers is the moment sequence of some positive measure* μ *supported on* K *if and only if the matrices* $M_k(g_j\cdot y)$ *are positive semidefinite for all* $j\in\{0,\dots,m\}$ *and all* $k\geq 0$*.*

Note the positive measure can be turned into a probability measure by additionally requiring entry $y_{(0,\dots,0)}$ of the moment matrix $M_k(y)$ to be 1.

Proof. Consider the canonical monomial basis $(x^\alpha)_{\alpha\in\mathbb{N}_0^n}$. By Theorem 6.6.1 the existence of the moment sequence is equivalent to $L(1)=1$ and the positive semidefiniteness of the bilinear forms (6.10). This translates to $\int_K p\,d\mu$ and $0\leq\int_K x^\alpha x^\beta g_j d\mu = M_k(g_j\cdot y)$ for $j\in\{0,\dots,m\}$. □

Example 6.6.3. With the notation $y_{ij}:=L(x_1^ix_2^j)$ introduced above and $g_0(x_1,x_2)=1$, the polynomial $g_1:=-4x_1^2+7x_1\geq 0$ has an (infinite) localizing matrix $M(g_1\cdot y)$ whose initial 3×3-submatrix is

$$\left(\begin{array}{c|cc} -4y_{20}+7y_{10} & -4y_{30}+7y_{20} & -4y_{21}+7y_{11}\\ \hline -4y_{30}+7y_{20} & -4y_{40}+7y_{30} & -4y_{31}+7y_{21}\\ -4y_{21}+7y_{11} & -4y_{31}+7y_{21} & -4y_{22}+7y_{12}\end{array}\right).$$

—

With these matrices a truncated version of (6.12) can be constructed. Let $k\geq k_0:=\max\{\lceil\deg p/2\rceil,\lceil\deg g_1/2\rceil,\dots,\lceil\deg g_m/2\rceil\}$, and consider the hierarchy of semidefinite relaxations:

$$Q_k:\quad \begin{array}{rcl}\inf_y \sum_\alpha p_\alpha y_\alpha & & \\ M_k(y) & \succeq & 0\,,\\ M_{k-\lceil\deg g_j/2\rceil}(g_j\,y) & \succeq & 0\,,\quad 1\leq j\leq m\end{array} \tag{6.13}$$

with optimal value denoted by $\inf Q_k$ (and $\min Q_k$ if the infimum is attained).

Although each of the relaxation values might not be optimal for the original problem, one has the following convergence result.

Theorem 6.6.4 (Lasserre [48])**.** *Let Assumption* **??** *hold and consider the hierarchy of SDP-relaxations* $(Q_k)_{k\geq k_0}$ *defined in (6.13). Then the sequence* $(\inf Q_k)_{k\geq k_0}$ *is monotone non-decreasing and converges to* p^**; that is,* $\inf Q_k\uparrow p^*$ *as* $k\to\infty$*.*

As mentioned before the infinite-dimensional cone $\mathrm{QM}(g_1,\dots,g_m)$ cannot be handled easily from a practical point of view. By restricting the degrees we replace it by a hierarchy of finite-dimensional cones.

Let $k_0=\max\left\{\lceil\frac{\deg p}{2}\rceil,\lceil\frac{\deg g_1}{2}\rceil,\dots,\lceil\frac{\deg g_m}{2}\rceil\right\}$, and for $k\geq k_0$ let

$$\begin{array}{rcll} a_k^* & := & \sup\gamma & \\ & & \text{s.t.} & p-\gamma=s_0+\sum_{j=1}^m s_jg_j\,,\\ & & & \text{where } s_0,\dots,s_m\in\Sigma \text{ with}\\ & & & \deg(s_0),\deg(s_1g_1),\dots,\deg(s_mg_m)\ \leq\ 2k\,.\end{array}$$

For each admissible k, by Lemma 6.3.1 this problem can be formulated as a semidefinite program. The dual semidefinite program results from the "truncated" finite version of the moment problem (**??**),

$$\begin{array}{rcll} b_k^* & := & \inf p^T y & \\ & & \text{s.t.} \quad y_0 & = & 1\,, \\ & & \qquad M_k(y) & \succeq & 0\,, \\ & & \qquad M_{k-\lceil \frac{\deg g_j}{2} \rceil}(g_j * y) & \succeq & 0\,, \quad 1 \leq j \leq m\,, \end{array} \tag{6.14}$$

where the M_k are the truncated versions of the localization matrices.

Theorem 6.6.5. *1. For each admissible k we have $a_k^* \leq b_k^*$.*

2. If Putinar's condition holds, we have

$$\lim_{k\to\infty} a_k^* \;=\; \lim_{k\to\infty} b_k^* \;=\; p^*\,.$$

Proof. The first statement immediately follows from weak duality.

For the second statements we first note that for each $\varepsilon > 0$ the polynomial $p - p^* + \varepsilon$ is strictly positive on K. By Putinar's Positivstellensatz $p - p^* + \varepsilon$ has a representation of the form (**??**). Hence, there exists a k with $a_k^* \geq p^* - \varepsilon$. Passing over to the limit $\varepsilon \downarrow 0$, this shows the claim. □

For $k \geq k_0$ this defines a hierarchy of semidefinite programs whose optimal values converges monotoneously to the optimum. It is possible that the optimum is reached already after finitely many steps ("finite convergence"). However, already to decide whether a value b_k^* obtained in the k-th relaxation is the optimal value is not easy. There only exist sufficient conditions.

Theorem 6.6.6. *Let $k \geq k_0$, y be an optimal value of the SDPs for b_k^*, and let $d = \max\left\{\frac{\lceil \deg g_1 \rceil}{2},\ \ldots,\ \frac{\lceil \deg g_m \rceil}{2}\right\}$. If $\operatorname{rank} M_k(y) = \operatorname{rank} M_{k-d}(y)$, then $b_k^* = p^*$.*

In the special case of 0-1-Problems we always have finite convergence.

Example 6.6.7. For $n \geq 2$ we consider the (parametric) optimization problem

$$\min \sum_{i=1}^{n+1} x_i^4 \quad \text{s.t.} \sum_{i=1}^{n+1} x_i^3 \;=\; 0\,, \quad \sum_{i=1}^{n+1} x_i^2 \;=\; 1\,, \quad \sum_{i=1}^{n+1} x_i \;=\; 0 \tag{6.15}$$

in the n variables $x_1, \ldots, x_n$. Systems of this type occur in the investigation of symmetric simplices. In order to show that a number α is a lower bound for the optimal value of (6.15), it suffices (due to the compactness of the feasible set) to show the existence of such a representation for $f(x) := \sum_{i=1}^{n+1} x_i^4 - \alpha + \varepsilon$ in view of $g_1(x) := \sum_{i=1}^{n+1} x_i^3$, $g_2(x) := -\sum_{i=1}^{n+1} x_i^3$, $g_3(x) := \sum_{i=1}^{n+1} x_i^2 - 1$, $g_4(x) := -\sum_{i=1}^{n+1} x_i^2 + 1$, $g_5(x) := \sum_{i=1}^{n+1} x_i$,

$g_6(x) := -\sum_{i=1}^{n+1} x_i$ for each $\varepsilon > 0$. For the case of odd n in (6.15) there exists a simple polynomial identity

$$\sum_{i=1}^{n+1} x_i^4 - \frac{1}{n+1} \;=\; \frac{2}{n+1}\left(\sum_{i=1}^{n+1} x_i^2 - 1\right) + \sum_{i=1}^{n+1}\left(x_i^2 - \frac{1}{n+1}\right)^2 , \tag{6.16}$$

which shows that the minimum is bounded from below by $1/(n+1)$; and since this value is attained at $x_1 = \ldots = x_{(n+1)/2} = -x_{(n+3)/2} = \ldots = -x_{n+1} = 1/\sqrt{n+1}$, the minimum is $1/(n+1)$. For each $\varepsilon > 0$ adding ε on both sides of (6.16) yields a representation of the positive polynomial in the quadratic module $\mathrm{QM}(g_1, \ldots, g_6)$. For each odd n this only uses polynomials $s_i g_i$ of (total) degree at most 4.

For the case n even (with minimum $1/n$) the situation looks different. A computer calculation with the software GloptiPoly shows that already for $n = 4$ it is necessary to go until degree 8 in order to obtain a Positivstellensatz-type certificate for optimality.

Exercises

Exercise 6.6.8.

Show that the semidefinite condition (6.10) in Theorem 6.6.1 can be equivalently replaced by the condition $L(M) \subset [0, \infty)$.

Exercise 6.6.9. Show the following moment matrix version of Schmüdgen's Theorem 5.6.2.

An infinite sequence $y = (y_\alpha)_{\alpha \in \mathbb{N}_0^n}$ of real numbers is the moment sequence of some measure μ supported on K if and only if the matrices $M_k(g_j y)$ are positive semidefinite for all $j \in \{0, \ldots, m\}$ and all $k \geq 0$.

Exercise 6.6.10. If the moment matrix and the localization matrices are indexed in terms of the monomials $\alpha \in \mathbb{N}_0^n$ of the monomial bases $(x^\alpha)_{\alpha \in \mathbb{N}_0^n}$ then the localizing matrix of a polynomial $g = \sum_\alpha c_\alpha x^\alpha$ is given by

$$M(g \cdot y)_{\alpha,\beta} \;=\; \sum_{\gamma \in \mathbb{N}_0^n} c_\gamma y_{\alpha+\beta+\gamma} \, .$$

6.6.1 The zero-dimensional case

In the general situation, convergence of the relaxation scheme is guaranteed. Moreover, as will be explained now, in some situations finite convergence can be guaranteed, that is, the optimal value will be attained at a finite relaxation order.

If there were a Putinar-type representation theorem for non-negative (rather than strictly positive) polynomials then finite convergence would just come out immediately from our discussion above (Theorem **??**).

When the definition of K involves a set of polynomial equations whose polynomials define a zero-dimensional radical, then such a representation is available rather explicitly.

This will be explained in the following. Later on, that result will be generalized towards the non-radical case.

Recall that over an arbitrary field K the radical ideal $\sqrt{I}$ is defined as

$$\sqrt{I} \;=\; \{p \in K[x] \,:\, p^k \in I \text{ for some } I\}$$

and that an ideal is called radical if $\sqrt{I} = I$.

Theorem 6.6.11 (Parrilo)**.** *Let $m \geq 0$, $l \geq 1$, $g_1, \ldots, g_m, h_1, \ldots, h_l \in \mathbb{R}[x]$ and*

$$K \;=\; \{x \in \mathbb{R}^n \,:\, g_1(x) \geq 0, \ldots, g_m(x) \geq 0, h_1(x) = 0, \ldots, h_l(x) = 0\}\,. \tag{6.17}$$

If the ideal $I := \langle h_1, \ldots, h_l$ is zero-dimensional and radical then any nonnegative polynomial p on K can be written in the form

$$p \;=\; s_0 + \sum_{i=1}^{m} s_j g_j + q$$

with $s_0, \ldots, s_m \in \Sigma[x]$ and $q \in I$.

Before we begin with the main proof, we collect some auxiliary results involving also the variety $\mathcal{V}_{\mathbb{C}}(I)$ of the ideal I over $\mathbb{C}$. Since the defining polynomials are real, the non-real zeroes in $\mathcal{V}_{\mathbb{C}}(I)$ come in conjugated pairs, thus giving a disjoint decomposition $V = \mathcal{V}_{\mathbb{R}}(I) \cup U \cup \overline{U}$ with $U \subset \mathbb{C}^n \setminus \mathbb{R}^n$.

Lemma 6.6.12. *Let $h_1, \ldots, h_l \in \mathbb{R}[x]$ and let $I := \langle h_1, \ldots, h_l \rangle$ be zero-dimensional and radical. Then any polynomial $f \in \mathbb{R}[x]$ which is nonnegative on $\mathcal{V}_{\mathbb{R}}(I)$ can be written as $f = s + q$ with $s \in \Sigma[x]$ and $q \in I$.*

Proof. Let $\mathcal{V}_{\mathbb{C}}(I) = \mathcal{V}_{\mathbb{R}}(I) \cup U \cup \overline{U}$ be the decomposition of the zeroes into real and complex ones. For any $a \in V_{\mathbb{R}}(I) \cup U$ let γ_a be a square root of a (so, for $a \in \mathcal{V}_{\mathbb{R}}(I)$ we have $\gamma_a \in \mathbb{R}$). Using Lagrange interpolation, we can find (complex) polynomials p_a satisfying $p_a(a) = 1$ and $p_a(b) = 0$ for $b \in \mathcal{V}_{\mathbb{C}}(I) \setminus \{a\}$. For $a \in \mathcal{V}_{\mathbb{R}}(I)$ we can clearly choose p_a with real coefficients (e.g., by choosing the real part of the complex polynomial), and for $a \in U$ we can choose p_a to satisfy $p_{\overline{a}} := \overline{p_a}$. Then the polynomials $q_a := \gamma_a p_a$ for $a \in \mathcal{V}_{\mathbb{R}}(I)$ and $q_a := \gamma_a p_a + \overline{\gamma_a p_a}$ for $a \in U$ are real polynomials. Since the polynomial $f_0 := f - \sum_{a \in \mathcal{V}_{\mathbb{R}}(I) \cup U} (q_a)^2$ satisfies $f_0(a) = 0$ for all $a \in V_{\mathbb{C}}(I)$, it is contained in the radical ideal I. This provides the desired representation. □

Proof of Theorem 6.6.11. Let $f \in \mathbb{R}[x]$ be nonnegative on K. Via Lagrange interpolation, we construct polynomials $r_0, \ldots, r_m$ with the following properties: Whenever a is a non-real point in I or whenever a is a real point in I with $f(a) \geq 0$, then we enforce $r_0(a) := f(a)$ and $r_j(a) := 0$, $1 \leq j \leq m$. Failing this, we have $a \notin K$ and there exists some $j_a \in \{1, \ldots, m\}$ with $g_{j_a}(a) < 0$. We enforce $r_{j_a}(a) := \frac{f(a)}{g_{j_a}(a)}$ as well as

$r_0(a) := r_j(a) = 0$ for all $j \neq j_a$. Note that each of the polynomials $r_0, \ldots, r_m$ is negative on K.

By Lemma 6.6.12, there exist $s_0, \ldots, s_m \in \Sigma[x]$ and $q_0, \ldots, q_m \in I$ with $r_j = s_j + q_j$, $1 \leq i \leq m$. The polynomial

$$f_0 \;:=\; f - r_0 - \sum_{j=1}^{m} r_j g_j \tag{6.18}$$

satisfies $f_0(a) = 0$ for all $a \in V_{\mathbb{C}}(I)$ and is therefore contained in the radical ideal I. Solving (6.18) for f and substituting r_j by $s_j + q_j$ provides the representation we were aiming for. □

Corollary 6.6.13. *If the set K is defined as in* (6.17) *and the ideal $I := \langle h_1, \ldots, h_l \rangle$ is zero-dimensional and radical then there is a $t \geq k_0$ with* $\inf Q_t = p^*$.

Theorem 6.6.14 (Laurent [51])**.** *If the ideal generated by $g_1, \ldots, g_m$ is zero-dimensional then there is an $t \geq k_0$ with* $\inf Q_t = p^*$.

Proof. Maybe give a proof (contains structural information on moment matrix and ideals) ... □

6.6.2 Detecting optimality and extracting optimal points

Setting ... to start from $M(y)$ ...

Via map L: $y_{\alpha\beta} = L(x^\alpha x^\beta) = \int_K x^\alpha x^\beta d\mu$. I.e., $M(y) = L(x^\alpha x^\beta)_{\alpha,\beta}$.

Theorem 6.6.15. *For a moment matrix $M(y)$ the following statements are equivalent:*

1. *$M(y) \succeq 0$ and $M(y)$ has finite rank r.*
2. *There exists a a unique probability measure μ representing y, and μ is r-atomic.*

For $M(y) \succeq 0$ we consider its kernel

$$I \;:=\; \ker M(y) \;=\; \{p \in \mathbb{R}[x] \,:\, M(y)p = 0\}\,. \tag{6.19}$$

We first note that I is an ideal, because $I + I \subset I$; in order to show that $p \in I$ and $q \in \mathbb{R}[x]$ we have $pq \in I$ recall that the kernel of a quadratic form coincides with the kernel of its representing matrix and observe that by definition $L(pq, pq) \;=\; L(pq^2, p) = 0$, since $p \in I$. Hence, $pq \in I$. Moreover, denoting bei $\mathcal{B}$ a set of monomials indexing a maximal nonsingular minor of $M(y)$, the set $\mathcal{B}$ is a vector space basis of the residue class ring $\mathbb{R}[x]/I$.

Theorem 6.6.16. *Let $M(y) \succeq 0$ and I be the kernel of $M(y)$ as defined in* (6.19)*.*

1. *I is a radical ideal in $\mathbb{R}[x]$.*

2. If $\operatorname{rank} M(y) < \infty$ *then* $|\mathcal{V}_{\mathbb{C}}(I)| = \operatorname{rank} M(y)$.

Proof. (a) Let $f^k \in I$. Setting l a the smallest power of k greater than or equal to l the ideal property of I implies $f^l \in I$. Hence $L(1, f^l) = L(f^{l/2}, f^{l/2})$ implies $f^{l/2} \in I$, which proves the claim after l iterations of this process.

(b) Since by ... (Connect!) for a radical ideal I the cardinality $|V_{\mathbb{C}}(I)|$ coincides with the dimension of the vector space $\mathbb{R}[x]/I$ the claim follows. □

Based on this ... Curto Fialkow ... leads to optimality criterion.

Theorem 6.6.17. *Let* $y \in \mathbb{R}^{S_{2t}}$ *be an optimal solution to the truncated moment relaxation* (6.14). *If*

$$\operatorname{rank} M_t(y) \;=\; \operatorname{rank} M_{t-d}(y)$$

then (moment values) $p_t^* = p^*$, *where* $d := max(\lceil \deg(g_j)/2 \rceil)$, $1 \le j \le m$.

Extraction of solutions ...
Sparsity and symmetry

Exercises

Exercise 6.6.18. Given $I = \langle -xy+z, -yz+x, xz-y \rangle$ and $f = -x+y^2-z^2+1$, determine $s \in \Sigma[x, y, z]$ and $q \in I$ such that $f = s + q$.

6.7 Notes

For the connection of global optimization and sums of squares see Parrilo [65, 66] or the survey of Laurent [52]. The sparsity result 6.1.2 is due to Reznick [74].

For an introduction to semidefinite programming see the book of De Klerk [22] or the survey article by Vandenberghe and Boyd [92].

The relaxation scheme for constrained global optimization has been introduced by Lasserre [48]. Theorem 6.6.1 is due to Putinar, where his proof of Theorem 6.6.1 was using methods from functional analysis and where he used this theorem to deduce then Theorem 5.6.5 from it. The link for the converse direction, from Theorem 5.6.5 to Theorem 6.6.1, as utilized in our treatment, goes back to Krivine [47], provided with a modern treatment in the book of Prestel and Delzell [72].

The classical moment problem arose during Stieltje's work on the analytical theory of continued fraction. Later, Hamburger established it as a question of its own right. Concerning the duality between non-negative polynomials and moment sequences, the univariate case is comprehensively treated by Karlin and Studden [44]. More information on the moment-based approach for optimization can be found in Lasserre's book [49] or in Laurent's extensive survey [52]. Haviland's proof of his Theorem 6.5.7 is contained

in [**?**]. ... Indeed, the statement is implicitly contained in Riesz' work in 1923 (cf. Helton, Putinar: Positive polynomials in scalar and matrix ...)

Theorem 6.6.11 was proven by Parrilo in [64], and the finite convergence result 6.6.14 of Laurent appears in [51]. The theory of the optimality criterion **??** is due to Curto and Fialkow who used functional-analytic methods. The transfer to a rather algebraic derivation has been achieved by Laurent [50, 51].

5

6

7

8

9

10

[5]integrate:

Proof. DRAFT on M_+ proof:
In order to show $\mathcal{M}_+ \subset \Sigma_n$ and $\Sigma \subset (\mathcal{M}_+)^*$, let $y \in \mathcal{M}_+$ and $p \in \mathbb{R}[x_1, \ldots, x_n]$. Then

$$y^T(p^2) \;=\; p^T M(y) p \;\geq 0\,,$$

i.e., $\mathcal{M}_+ \subset \Sigma_n^*$ and $\Sigma_n \subset (\mathcal{M}_+)^*$.
The inclusion $\Sigma^* \subset \mathcal{M}_+$ follows from reversing the argument (start from Choleski ...)
Berg, Christensen, and Jensen [6] and and independently Schmüdgen ([78], re-check!) have shown that the the infinite-dimensional cone (Σ_n) is closed (see also [7, Thm. 3.2] re-check!). Hence, by Theorem **??**, $(\mathcal{M}_+)^* = (\Sigma_n)^{**} = \Sigma_n$. □

[6]design question how to treat $\mathbb{R}^n$ vs. K
[7]A set $K \subset \mathbb{R}^n$ is called a *cone* if the following two conditions are satisified.

1. $x, y \in K \;\Rightarrow\; x + y \in K\,,$
2. $x \in K,\, \lambda \geq 0 \;\Longrightarrow\; \lambda x \in K\,.$

[8]Integrate:
Exercises duality:
Duality SOS vs. psd sequences
Variations of cones (e.g., bounded degree)
[9]Notation $\Sigma[x]$ vs. Σ_n
[10]Clarify use of bilinear form ...

Chapter 7

Algebra and Geometric Combinatorics

Outline:

1. Polytopes, polyhedra, cones, polyhedral complexes, fans, normal fans
2. Regular polyhedral subdivisions, regular triangulations, Minkowski sums, Mixed Volumes (See Ewald's book)
3. Universal Gröbner bases and Gröbner fan
4. Gröbner degeneration—calculation of Hilbert Function. Baby flatness, 1-parameter degenerations ??
5. Integer linear algebra
6. Toric ideals.

Geometric combinatorics, which we take to be the study of polytopes, polyhedra, and integer vectors, is fundamental to computation in algebraic geometry, to toric varieties which appear in many applications, and to tropical geometry. Many other algebraic and geometric structures are best understood in terms of objects from geometric combinatorics. We introduce some of the basic objects of geometric combinatorics and use them to derive some fundamental algebraic consequences.

7.1 Polytopes, complexes, and fans

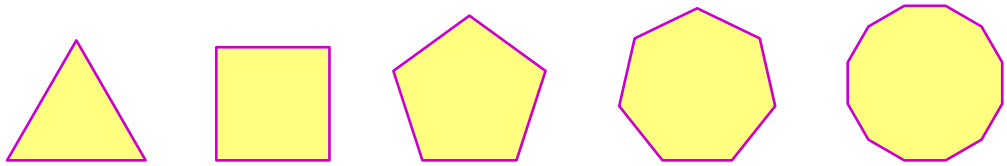
Perhaps the most fundamental objects in geometric combinatorics are polytopes. A *polytope* P is the convex hull of finitely many points $v_1, \ldots, v_m$ in the vector space $\mathbb{R}^d$,

$$P \;:=\; \operatorname{conv}\{v_1, \ldots, v_m\} \;:=\; \left\{ \sum_{i=1}^{m} \lambda_i v_i \mid \lambda_1, \ldots, \lambda_m \geq 0 \text{ and } \sum_i \lambda_i = 1 \right\}. \qquad (7.1)$$

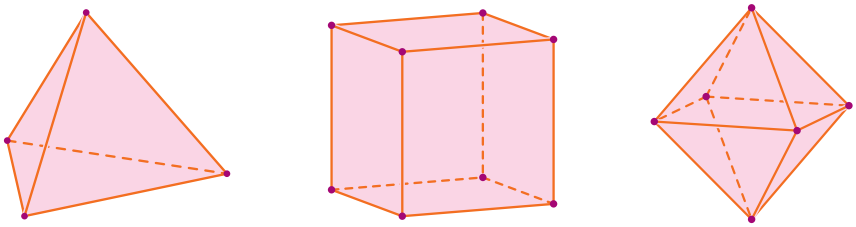
If this representation is irredundant, then the points $v_1, \ldots, v_m$ are the extreme points of P which are called its *vertices*. The dimension of a polytope (7.1) is the dimension of

its affine hull, which is also the dimension of the vector space spanned by the differences $\{v_2 - v_1, v_3 - v_1, \ldots, v_m - v_1\}$.

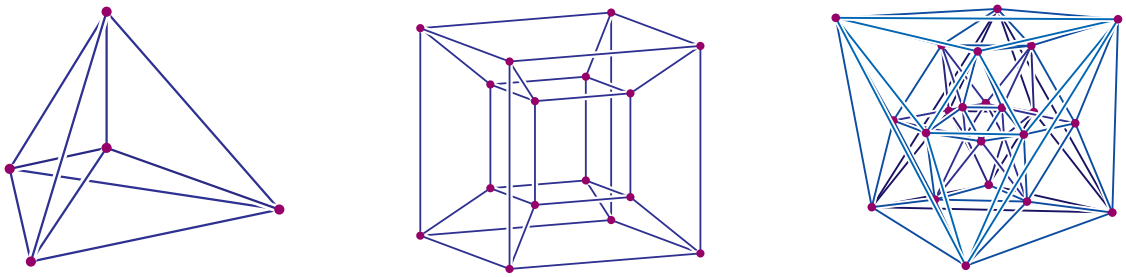
There is only one polytope (a point) when $d = 0$ and polytopes when $d = 1$ are line segments. When $d = 2$, we obtain polygons (or points and line segments when the polytope is degenerate).



When $d = 3$, we get the familiar three-dimensional polyhedra,



and it is increasingly hard to visualize polytopes when $d \geq 4$.



Perhaps the simplest polytope in each dimension is a *simplex*, which is the convex hull of $d+1$ affinely independent points in $\mathbb{R}^d$. For example, the standard, or probability, d-dimensional simplex is the convex hull of the $d+1$ linearly independent standard basis vectors $e_1, \ldots, e_{d+1}$ in $\mathbb{R}^{d+1}$. This lies on the affine hyperplane $\{x \in \mathbb{R}^{d+1} \mid x_1 + \cdots + x_{d+1} = 1\}$. The standard simplex is universal among all polytopes in that a polytope P (7.1) is the image of the standard $m-1$ simplex under the affine map that sends $e_i \in \mathbb{R}^m$ to $v_i \in \mathbb{R}^d$, for $i = 1, \ldots, m$. More generally, the image of any polytope under an affine map is anoter polytope—simply take the convex hull of the images of the vertices.

In addition to vertices, polytopes also have edges and faces that are themselves lower-dimensional polytopes. In general, given a vector $w \in \mathbb{R}^d$, the linear fuction $x \mapsto w^T \cdot x$ acheives its minimum on $P = \text{conv}\{v_1, \ldots, v_m\}$ along a *face* $F = F_w$ of P which is *supported* by this linear function. This face is itself a polytope

$$F_w \;:=\; \text{conv}\left\{v \in \{v_1, \ldots, v_m\} \mid w^T \cdot v = \min\{w^T \cdot v_1, \ldots, w^T \cdot v_m\}\right\} . \qquad (7.2)$$

If we let $\min(w)$ be this minimum, then we see that

$$P \;=\; \bigcap_{w\in\mathbb{R}^d} \{x \in \mathbb{R}^d \mid w^T \cdot x \geq \min(w)\}\,.$$

Each set in this intersection is a closed half-space. This description is highly redundant, we need only those w whose linear functions support maximal proper subspaces of P—called *facets*—and only one vector w for each facet. Letting $w_1, \ldots, w_n$ be a set of vectors supporting the facets of P, then we obtain a dual facet description of P as an intersection of finitely many half spaces,

$$P \;=\; \{x \in \mathbb{R}^d \mid w_i^T \cdot x \geq \min(w_i) \quad i = 1, \ldots, n\}\,. \tag{7.3}$$

These dual descriptions as the convex hull of finitely many points or the intersection of finitely many half-spaces are both important when studying or using polytopes. Unfortunately, it may be expensive to pass from one description to the other. For example, the cube in $\mathbb{R}^d$ has $2d$ facets but 2^d vertices and the octahedron (cross polytope) has $2d$ vertices and 2^d facets. Thus in the worst case there may be exponentially many more facets than vertices, or vice-versa.

Given the facet description (7.3) of a polytope P, let A be the $n \times d$ matrix whose rows are the facet normals w_i^T for $i = 1, \ldots, n$ and b the column vector with ith entry $-\min(w_i)$, then (7.3) becomes

$$P \;=\; \{x \in \mathbb{R}^d \mid Ax + b \geq 0\}\,, \tag{7.4}$$

where $\geq$ is coordinatewise comparison. Geometrically, this realizes P as the intersection of the affine space $A\mathbb{R}^d + b$ in $\mathbb{R}^n$ with the nonnegative orthant $\mathbb{R}^n_{\geq} := \{y \in \mathbb{R}^n \mid y_i \;\geq\; 0\}$, that is, an affine section of the orthant. Similarly, an affine section of a polytope is another polytope, as this affine section is equivalent to replacing the affine space $A\mathbb{R}^d + b$ in an affine section with a smaller affine space. Equivalently, one may add equality constraints to the facet description (7.3), such as $w_i^T \cdot x = \min(w_i)$, which will cut out a facet of P.

We single out two classes of polytopes, each of which is generic in a different sense. A polytope is *simplicial* if every face is a simplex. If the points $\{v_1, \ldots, v_m\}$ are in linear general position in that no $d+1$ of them lie on a hyperplane, then their convex hull is a simplicial polytope. A d-dimensional polytope is *simple* if every vertex lies on d facets. If the coordinates of the vector b in the facet description of a polytope P (7.4) are general, then P is simple. While both notions are general in different, dual senses, a polytope that is both simple and simplicial is just a simplex.

In addition to their geometric structure, polytopes also have combinatorial structures, which are often important. Among the faces of a polytope P we include the empty face $\emptyset$, as well as the polytope P itself. The inclusion relation among the faces of P is encoded by its *face lattice*, which is the partially ordered set of faces of P with minimal element $\emptyset$ and maximal element P. It is a lattice as any two faces intersect along their common subface (which may be $\emptyset$) and lie in a unique smallest face, and these operations are distributive.

This face lattice is a ranked poset, where the ranking function is the dimension of a face and we take $\emptyset$ to have dimension -1.

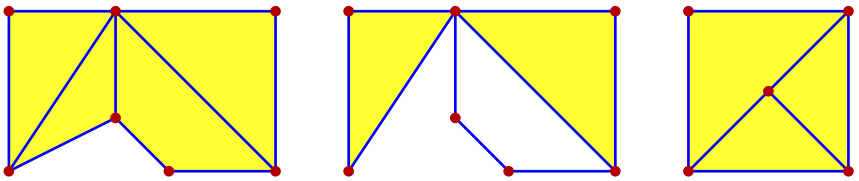
Polarity is an important construction which interchanges vertices and facets. Suppose that the origin 0 lies in the relative interior of a polytope P. Then the *polar* P° of P is

$$P^\circ := \{w \in \mathbb{R}^d \mid w^T \cdot x \geq -1 \text{ for all } x \in P\}.$$

The polar of P is another polytope whose vertices correspond to the facets of P and whose facets correspond to the vertices of P. Moreover, $(P^\circ)^\circ = P$. For example the polar of a cube is the octahedron and vice versa. Polarity is particularly pleasing on the face lattice—the face lattice of P° is simply the face lattice of P with all the order relations reversed. This illustrates that while the polar of a polytope changes if it is shifted with respect to the origin (keeping the origin in its interior), the combinatorial structure of the polar will not change. It also shows that the polar of a simplicial polytope is simple, and vice-versa.

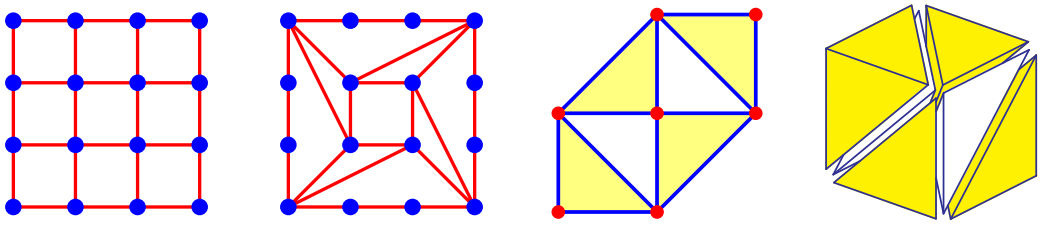
Show a polytope and its polar.

Another fundamental object is a *polyhedral complex*, which is a collection $\mathcal{P}$ of polytopes in $\mathbb{R}^d$ such that every face of a polytope P in $\mathcal{P}$ is another polytope in $\mathcal{P}$ and the intersection of any two polytopes P, P' in $\mathcal{P}$ is a common face of each. For example, of the three collections of vertices, line segments and polygons shown below, the first two are polyhedral complexes, while the last last is not; the large triangle does not meet either of the smaller triangles in one of its faces.



A common example of a polyhedral complex would be a polytope itself, togeter with the collection of its faces, or perhaps the boundary of a polytope.

The *support* of a polyhedral complex $\mathcal{P}$ is the union of all the poytopes in $\mathcal{P}$. A *polyhedral subdivision* of a polytope P is a polyhedral complex whose support is P and whose vertices are the vertices of P. More generally, given a finite set $\mathcal{A}$ of points in $\mathbb{R}^d$, a polyhedral subdivision of $\mathcal{A}$ is a polydehral complex whose vertices are a subset of $\mathcal{A}$ and whose support is the convex hull of $\mathcal{A}$. When every polytope in a polyhedral complex $\mathcal{P}$ is a simplex, we say that $\mathcal{P}$ is a *triangulation* of its support. Of the four polyhedral subdivisions below, only the last two are triangulations.



A *polyhedron* P is any intersection of finitey many half spaces

$$P \;=\; \bigcap_{i=1}^{m} \{x \in \mathbb{R}^d \mid w_i^T \cdot x \geq -b_i\}\,. \tag{7.5}$$

A polytope is a bounded polyhedron. The faces of a polyhedron P are subsets defined by replacing some of the inequalities by equalities in (7.5). Like (7.3), this description (7.5) may be made more concise. Let $\mathcal{A}$ be the $n \times d$ matrix whose rows are the vectors $w_1^T, \ldots, w_m^T$, and let b be the column vector $(b_1, \ldots, b_m)^T$. Then (7.5) becomes

$$P \;=\; \{x \in \mathbb{R}^d \mid Ax + b \geq 0\}\,,$$

where as in (7.4), $\geq$ means coordinatewise composition. Like polytopes, polyhedra have have dimension, faces, and a face lattice. Unlike polytopes, they need not have vertices. For example, a half-space isa polyhedron. Also, polyhedral complexes are more general than we defined them, they may not be solely composed of polytopes, but are more generally composed of polyhedra which meet along common faces.

A *(polyhedral) cone* is a polyhedron given by homogeneous iequalities

$$\{x \in \mathbb{R}^d \mid w_i^T \cdot x \geq 0 \text{ for } i = 1, \ldots, m\}\,. \tag{7.6}$$

This is a convex cone in the usual sense; it is a unital semigroup under vector addition that is closed under scalar multiplication by nonnegative numbers. Its *vertex* is the largest linear space contained in C,

$$\{x \in \mathbb{R}^d \mid w_i^T \cdot x = 0 \text{ for } i = 1, \ldots, m\}\,,$$

also called its *lineality space*, which is the largest linear subspace of $\mathbb{R}^d$ that acts on the cone by translation. The polar of a cone C is

$$C^\circ \;:=\; \{w \in \mathbb{R}^d \mid w^T \cdot x \geq 0 \text{ for all } x \in C\}\,.$$

The lineality space of C° is the orthogonal complement of the linear span of C, and the linear span of C° is the orthogonal complement of the lineality space of C.

A cone is *pointed* if it has trivial lineality space, $\{0\}$. A pointed cone is uniquely generated by a finite minimal set of 1-dimensional subcones, called its *rays*. If we choose one nonzero vector on each ray of a cone C, say $v_1, \ldots, v_m$, then the cone is generated by these vectors,

$$C \;=\; \Big\{\sum_{i=1}^{m} r_i v_i \mid r_i \in \mathbb{R}_{\geq 0}\Big\}\,. \tag{7.7}$$

A cone is *simplicial* if it is generated by minimally many rays, that is, if $n = \dim(C)$ in (7.7).

Pointed cones C in $\mathbb{R}^{d+1}$ correspond to polytopes in $\mathbb{R}^d$ in the following way. Let $w \in \mathbb{R}^{d+1}$ be any vector whose positive half space contains C in that $w^T \dot{v} > 0$ for any generator v of C. Then

$$C \cap \{x \in \mathbb{R}^{d+1} \mid w^T \cdot x = 1\},$$

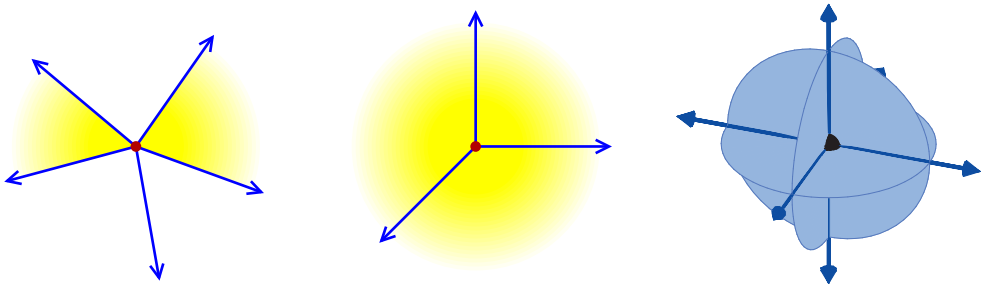
is a polytope in $\mathbb{R}^d$ (realized as an affine hyperplane in $\mathbb{R}^{d+1}$) that is the convex hull of

$$\left\{ \tfrac{1}{w^T \cdot v} v \mid v \text{ is a generator of } C \right\}.$$

Conversely, any polytope P in $\mathbb{R}^d$, may be realized as a polytope in $\mathbb{R}^{d+1}$ by identifying $\mathbb{R}^d$ with the affine hyperplane $\{x \in \mathbb{R}^{d+1} \mid x_{d+1} = 1\}$. Then the cone $C(P)$ generated by the vertices of P is the cone over P or the cone with base P.

This relationship is preserved under polarity. Suppose that P is a polytope in $\mathbb{R}^d$ of full dimension d with the origin in its relative interior, and construct $C(P)$ as above. Then $C(P^\circ) = C(P)^\circ$.

A *fan* is the analog of polyhedral complex for cones. Specifically, a fan is a polyhedral complex composed of cones with the same lineality space. Here are some examples.



Unlike these examples, the cones in a fan need not be pointed. A fan in $\mathbb{R}^d$ is *complete* if its support is $\mathbb{R}^n$.

An importat class of complete fans are normal fans of polytopes. Let $P \subset \mathbb{R}^d$ be a polytope. Recall that each vector $w \in \mathbb{R}^d$ supports a unique face F_w of P. For a nonempty face F of P, let $[F]$ be the set of those $w \in \mathbb{R}^d$ with $F = F_w$. By the description of F_w (7.2), we see that $[F]$ is defined by strict linear inequalities and linear equations, and so $\overline{[F]}$ is a cone with $[F]$ open in its linear span. We call $\overline{[F]}$ the *normal cone* of the face F. If F is a vertex, then $[F]$ is full-dimensional, and in general, $\dim F + \dim[F] = d$.

By the description of F_w (7.2), we see that if $F \subset G$ are faces of P, then $\overline{[G]} \subset \overline{[F]}$. Furthermore, if F, G are faces, then $\overline{[G]} \cap \overline{[F]} = \overline{[H]}$, where H isthe smallest face of P that contains both F and G. It follows that the set of cones $\overline{[F]}$ for F a nonempty face of P forms a complete fan, which is called the *normal fan of the polytope* P.

give an example of a polytope and its normal fan, or two examples.

Exercises for Section 7.1

1. Prove that the polar of a simple polytope is a simplicial polytope.

2. More

7.2 Regular subdivisions and mixed volumes

7.3 The Gröbner fan

7.4 Gröbner degenerations

7.5 Integer points in cones and polytopes

7.6 Toric ideals

Toric ideals are ideals of toric varieties, both of which play a special role in applications of algebraic geometry. We begin with some interesting geometry of polynomial equations. Here, we work with *Laurent polynomials*, which are polynomials whose monomials may have both positive and negative exponents.

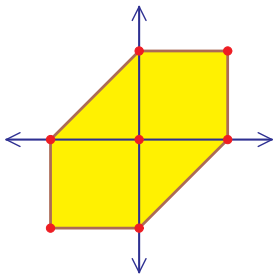
Let $\mathcal{A} \subset \mathbb{Z}^n$ be a finite collection of exponent vectors for Laurent monomials in $x_1, \ldots, x_n$. For example, when $n = 4$ the exponent vector $\alpha = (2, -3, -5, 1)$ corresponds to $x_1^2 x_2^{-3} x_3^{-5} x_4$. A *(Laurent) polynomial* f with *support* $\mathcal{A}$ is a linear combination of (Laurent) monomials,

$$f \;=\; \sum_{\alpha \in \mathcal{A}} c_\alpha x^\alpha \,. \tag{7.8}$$

For example, the polynomial

$$f \;:=\; 1 + 2x + 3xy + 5y + 7x^{-1} + 11x^{-1}y^{-1} + 13y^{-1}$$

has support $\mathcal{A}$ consisting of the integer points in the hexagon



$$\mathcal{A} \;=\; \begin{pmatrix} 0 & 1 & 1 & 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & 1 & 0 & -1 & -1 \end{pmatrix} .$$

The polynomial f (7.8) lies in the ring of Laurent polynomials, $\mathbb{F}[x_1, x_1^{-1}, \ldots, x_n, x_n^{-1}]$, which is the coordinate ring of the *algebraic torus*

$$\begin{aligned} (\mathbb{F}^\times)^n \;&:=\; \{x \in \mathbb{F}^n \mid x_1 \cdots x_n \neq 1\} \\ &\simeq\; \mathcal{V}(x_1 \cdots x_n x_{n+1} - 1) \;\subset\; \mathbb{A}^{n+1} . \end{aligned}$$

While it is often convenient to use the elements of $\mathcal{A}$ as indices, when the elements of $\mathcal{A}$ are in a list ($\mathcal{A} = \{\alpha_1, \ldots, \alpha_m\}$), then we will use the indices $1, \ldots, m$. For example, if we set $c_i := c_{\alpha_i}$, then the polynomial f in (7.8) becomes $\sum_{i=1}^m c_i x^{\alpha_i}$.

Given the set $\mathcal{A} \subset \mathbb{Z}^n$, we define a map,

$$\begin{array}{rccl} \varphi_{\mathcal{A}} : & (\mathbb{F}^\times)^n & \longrightarrow & \mathbb{P}^{\mathcal{A}} := [y_\alpha \mid \alpha \in \mathcal{A}] \\ & x & \longmapsto & [x^\alpha \mid \alpha \in \mathcal{A}]\,. \end{array}$$

One reason that we consider this map is that it turns non-linear polynomials on $(\mathbb{F}^\times)^n$ into linear equations on $\varphi((\mathbb{F}^\times)^n)$. Let

$$\Lambda \;=\; \Lambda(y) \;:=\; \sum_{\alpha\in\mathcal{A}} c_\alpha y_\alpha$$

be a linear form on $\mathbb{P}^{\mathcal{A}}$. Then the pullback

$$\varphi_{\mathcal{A}}^*(\Lambda) \;=\; \sum_{\alpha\in\mathcal{A}} c_\alpha x^\alpha$$

is a polynomial with support $\mathcal{A}$. This construction gives a correspondence

$$\begin{array}{ccc} \left\{ \begin{array}{c} \text{Polynomials } f \text{ on} \\ (\mathbb{F}^\times)^n \text{ with support } \mathcal{A} \end{array} \right\} & \Longleftrightarrow & \left\{ \begin{array}{c} \text{Linear forms } \Lambda \text{ in } \mathbb{P}^{\mathcal{A}} \\ \text{on } \varphi_{\mathcal{A}}((\mathbb{F}^\times)^n) \end{array} \right\} \\ f & \Longleftrightarrow & \varphi_{\mathcal{A}}^*(\Lambda)\,. \end{array}$$

In this way, a system of polynomials

$$f_1(x_1,\ldots,x_n) \;=\; f_2(x_1,\ldots,x_n) \;=\; \cdots \;=\; f_n(x_1,\ldots,x_n) \;=\; 0 \tag{7.9}$$

where each polynomial f_i has support $\mathcal{A}$ corresponds to a system of linear equations

$$\Lambda_1(y) \;=\; \Lambda_2(y) \;=\; \cdots \;=\; \Lambda_n(y) \;=\; 0$$

on $\varphi_{\mathcal{A}}((\mathbb{F}^\times)^n)$. Our approach to study these linear equations is to replace $\varphi_{\mathcal{A}}((\mathbb{F}^\times)^n)$ by its Zariski closure.

Definition 7.6.1. The *toric variety* $X_{\mathcal{A}} \subset \mathbb{P}^{\mathcal{A}}$ is the closure of the image $\varphi_{\mathcal{A}}((\mathbb{F}^\times)^n)$ of $\varphi_{\mathcal{A}}$. The *toric ideal* $I_{\mathcal{A}}$ is the ideal of the toric variety $X_{\mathcal{A}}$. It consists of all homogeneous polynomials which vanish on $\varphi_{\mathcal{A}}((\mathbb{F}^\times)^n)$.

Observe that the map $\varphi_{\mathcal{A}}\colon (\mathbb{F}^\times)^n \to X_{\mathcal{A}}$ parametrizes $X_{\mathcal{A}}$.
we are assuming here that $\mathcal{A}$ affinely spans $\mathbb{Z}^n$.

Corollary 7.6.2. *The number of solutions to a system of polynomials with support $\mathcal{A}$ (7.9) is at most the degree of the toric variety $X_{\mathcal{A}}$. When $\mathbb{F}$ is algebraically closed, it equals this degree when the polynomials are generic given their support $\mathcal{A}$.*

Proof. This follows from the geometric interpretation of degree of a projective variety (Corollary 3.6.13). The only additional argument that is needed is that the difference $\partial(X_{\mathcal{A}}) := X_{\mathcal{A}} - \varphi_{\mathcal{A}}((\mathbb{F}^\times)^n)$ has dimension less than n, and so a generic linear space of codimension n will not meet $\partial(X_{\mathcal{A}})$. □

It will greatly aid our discussion if we homogenize the map $\varphi_{\mathcal{A}}$.

$$\begin{array}{rcl} \mathbb{F}^\times \times (\mathbb{F}^\times)^n & \longrightarrow & \mathbb{P}^{\mathcal{A}} \\ (t,x) & \longmapsto & [tx^\alpha \mid \alpha \in \mathcal{A}] \end{array}$$

We can regard this homogenized version of $\varphi_{\mathcal{A}}$ as a map to $\mathbb{F}^{\mathcal{A}}$ whose image is equal to the cone over $\varphi_{\mathcal{A}}((\mathbb{F}^\times)^n)$ with the origin removed. Since $X_{\mathcal{A}}$ is projective, this does not change the image of $\varphi_{\mathcal{A}}$. The easiest way to ensure that $\varphi_{\mathcal{A}}$ is homogeneous is to assume that every vector in $\mathcal{A}$ has first coordinate 1, or more generally, to assume that the row space of $\mathcal{A}$ contains the vector all of whose components are 1.

For example, suppose that $\mathcal{A}$ consists of the 7 points in $\mathbb{Z}^2$ which are columns of the 2×7 matrix (also written $\mathcal{A}$),

$$\mathcal{A} = \begin{pmatrix} -1 & -1 & 0 & 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 1 & 0 & -1 & 0 \end{pmatrix}.$$

Then $\mathcal{A}$ is the integer points (in red) in the hexagon on the left of Figure 7.1. The

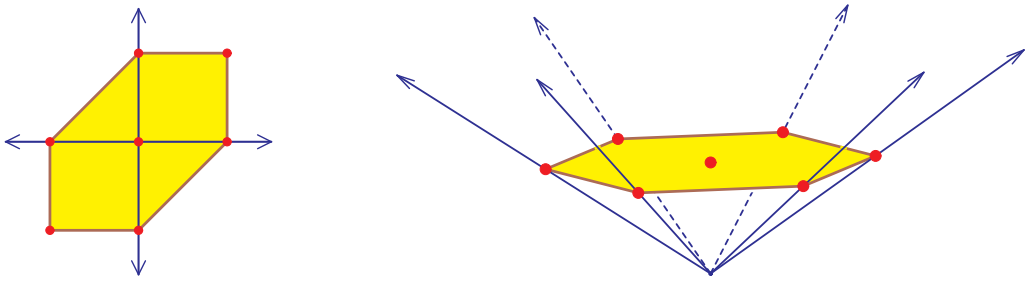


Figure 7.1: The hexagon and its lift

homogenized version of set $\mathcal{A}$ is the column vectors of the 3×7 matrix,

$$\mathcal{A}^+ = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & 0 & 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 1 & 0 & -1 & 0 \end{pmatrix}.$$

These are the red points in the picture on the right of Figure 7.1. (There, the first coordinate is vertical.)

Given such a homogeneous configuration of integer vectors $\mathcal{A}$ we have the associated map $\varphi_{\mathcal{A}}$ parametrizing the toric variety $X_{\mathcal{A}}$. The pull back of $\varphi_{\mathcal{A}}$ is the map

$$\begin{array}{rcl} \varphi_{\mathcal{A}}^* : \mathbb{F}[y_\alpha \mid \alpha \in \mathcal{A}] & \longrightarrow & \mathbb{F}[x_1, x_1^{-1}, \ldots, x_n, x_n^{-1}] \\ y_\alpha & \longmapsto & x^\alpha \end{array}$$

Its kernel is the toric ideal $I_{\mathcal{A}}$.

Theorem 7.6.3. *The toric ideal $I_{\mathcal{A}}$ is a prime ideal.*

Proof. The image of $\varphi_{\mathcal{A}}^*$ is a subring of $\mathbb{F}[x_1, x_1^{-1}, \ldots, x_n, x_n^{-1}]$, which is an integral domain. Since the image is isomorphic to the quotient $\mathbb{F}[y_\alpha \mid \alpha \in \mathcal{A}]/\ker \varphi^*$, this quotient is a domain and thus $I_{\mathcal{A}} = \ker \varphi^*$ is prime.

We may also see this as $(\mathbb{F}^\times)^n$ is irreducible, which implies that $X_{\mathcal{A}}$ is irreducible, and therefore its homogeneous ideal is prime. These two arguments are equivalent via the algebraic-geometric dictionary. Cite tis earlier result □

We describe some generating sets for the toric ideal $I_{\mathcal{A}}$. Let $u = (u_\alpha \mid \alpha \in \mathcal{A}) \in \mathbb{N}^{\mathcal{A}}$ be an integer vector. Then

$$\varphi_{\mathcal{A}}^*(y^u) \;=\; \prod_{\alpha \in \mathcal{A}} (x^\alpha)^{u_\alpha} \;=\; x^{\sum u_\alpha \cdot \alpha}\,.$$

If we regard these exponents $u \in \mathbb{Z}^{\mathcal{A}}$ as column vectors and the elements of $\mathcal{A}$ as the columns of a matrix $\mathcal{A}$ with $n+1$ rows, then

$$\varphi_{\mathcal{A}}^*(y^u) \;=\; x^{\mathcal{A}u}\,.$$

Notice that $\varphi_{\mathcal{A}}^*(y^u) = \varphi_{\mathcal{A}}^*(y^v)$ if and only if $Au = Av$, and thus $y^u - y^v \in I_{\mathcal{A}}$.

Theorem 7.6.4. *The toric ideal $I_{\mathcal{A}}$ is the linear span of binomials of the form*

$$\{y^u - y^v \mid Au = Av\}\,. \tag{7.10}$$

Proof. These binomials certainly lie in $I_{\mathcal{A}}$. Pick a monomial order $\prec$ on $\mathbb{F}[y_\alpha \mid \alpha \in \mathcal{A}]$. Let $f \in I_{\mathcal{A}}$,

$$f \;=\; c_u y^u \;+\; \sum_{v \prec u} c_v y^v \qquad c_u \neq 0\,,$$

so that $\mathrm{in}(f) = c_u y^u$. Then

$$0 \;=\; \varphi_{\mathcal{A}}^*(f) \;=\; c_u x^{\mathcal{A}u} \;+\; \sum_{v \prec u} c_v x^{\mathcal{A}v}\,.$$

There is some $v \prec u$ with $Av = Au$, for otherwise the initial term $c_u x^{\mathcal{A}u}$ is not canceled and $\varphi_{\mathcal{A}}^*(f) \neq 0$. Set $\overline{f} := f - c_u(y^u - y^v)$. Then $\varphi_{\mathcal{A}}^*(\overline{f}) = 0$ and $\mathrm{in}(\overline{f}) \prec \mathrm{in}(f)$.

If the leading term of f were $\prec$-minimal in $\mathrm{in}(I_{\mathcal{A}})$, then $\overline{f}$ must be zero, and so f is a linear combination of binomials of the form (7.10). Suppose by way of induction that every polynomial in $I_{\mathcal{A}}$ whose initial term is less that that of f is a linear combination of binomials of the form (7.10). Then $\overline{f}$ is a linear combination of binomials of the form (7.10), which implies that f is as well. □

Theorem 7.6.4 gives an infinite generating set for $I_{\mathcal{A}}$. We seek smaller generating sets. Suppose that $Au = Av$ with $u, v \in \mathbb{N}^{\mathcal{A}}$. Define vectors $t, w^{\pm}$ through their components, Set

$$\begin{aligned} t_\alpha &:= \min(u_\alpha, v_\alpha) \\ w^+_\alpha &:= \max(u_\alpha - v_\alpha, 0) \\ w^-_\alpha &:= \max(v_\alpha - u_\alpha, 0) \end{aligned}$$

Then $u - v = w^+ - w^-$ and $u = t + w^+$ and $v = t + w^-$, and

$$y^t(y^{w^+} - y^{w^-}) \;=\; y^u - y^v \;\in\; I_{\mathcal{A}}\,.$$

For $w \in \mathbb{Z}^{\mathcal{A}}$, let w^+ be the coordinatewise maximum of w and the 0-vector, and let w^- be the coordinatewise maximum of $-w$ and the 0-vector.

Corollary 7.6.5. $I_{\mathcal{A}} = \langle y^{w^+} - y^{w^-} \mid \mathcal{A}w = 0 \rangle$.

Thus $I_{\mathcal{A}}$ is generated by binomials coming from the integer kernel of the matrix $\mathcal{A}$.

Theorem 7.6.6. *Any reduced Gröbner basis of $I_{\mathcal{A}}$ consists of binomials.*

The point is that Buchberger's algorithm is binomial-friendly, in the same sense as it was homogeneous-friendly in the proof of Theorem 3.6.2. We omit the nearly verbatim proof.

In practice, Buchberger's algorithm is not the best way to compute a Gröbner basis for a toric ideal. Here is an alternative method due to Hoşten and Sturmfels [41]. We remark that there are other, often superior algorithms available, for example the project-and-lift algorithm of Hemmecke and Malkin [38] which is implemented in the software `4ti2`.

Let A be an integer matrix whose row space contains the vector all of whose components are 1. This is a map from $\mathbb{Z}^{\mathcal{A}}$ to $\mathbb{Z}^{n+1}$ whose kernel is a free abelian subgroup of $\mathbb{Z}^{\mathcal{A}}$. Let $\mathcal{W} := \{w_1, \ldots, w_\ell\}$ be a $\mathbb{Z}$-basis for the kernel of $\mathcal{A}$, and set

$$I_{\mathcal{W}} \;:=\; \{y^{w_i^+} - y^{w_i^-} \mid i = 1, \ldots, \ell\}\,.$$

Then $I_{\mathcal{W}}$ defines the image $\varphi((\mathbb{F}^\times)^n)$ as a subvariety of the dense torus in $\mathbb{P}^{\mathcal{A}}$,

$$(\mathbb{F}^\times)^{|\mathcal{A}-1|} \;=\; \{y \in \mathbb{P}^{\mathcal{A}} \mid \prod_\alpha y_\alpha \neq 0\}\,.$$

Thus the ideal $I_{\mathcal{W}}$ agrees with $I_{\mathcal{A}}$ off the coordinate axes of $\mathbb{P}^{\mathcal{A}}$.

The idea behind this method of Hoşten and Sturmfels is to use saturation to pass from the ideal $I_{\mathcal{W}}$ to the full toric ideal $I_{\mathcal{A}}$. Let $\alpha \in \mathcal{A}$, and consider the saturation of an ideal I with respect to the variable y_α,

$$(I \;:\; y_\alpha^\infty) \;:=\; \{f \mid y_\alpha^m f \in I \text{ for some } m\}\,.$$

Note that in the affine set $U_\alpha = \mathbb{P}^{\mathcal{A}} - \mathcal{V}(y_\alpha)$ the two ideals agree, $\mathcal{V}(I) = \mathcal{V}(I : y_\alpha^\infty)$.

Lemma 7.6.7. *Let I be a homogeneous ideal. Then $\mathcal{V}(I : y_\alpha^\infty) = \overline{\mathcal{V}(I) - \mathcal{V}(y_\alpha)}$.*

Proof. We have that $\overline{\mathcal{V}(I) - \mathcal{V}(y_\alpha)} \subset \mathcal{V}(I : y_\alpha^\infty)$ as $\mathcal{V}(I) = \mathcal{V}(I : y_\alpha^\infty)$ in $U_\alpha = \mathbb{P}^{\mathcal{A}} - \mathcal{V}(y_\alpha)$. For the other inclusion, let x be a point of $\mathcal{V}(y_\alpha)$ which lies in $\mathcal{V}(I : y_\alpha^\infty)$ and let f be a homogeneous polynomial which vanishes on $\mathcal{V}(I) - \mathcal{V}(y_\alpha)$. We show that $f(x) - 0$. Then $y_\alpha f$ vanishes on $\mathcal{V}(I)$. By the Nullstellensatz, there is some m such that $y_\alpha^m f^m \in I$, and so $f^m \text{in}(I : y_\alpha^\infty)$. But then $f^m(x) = 0$ and so $f(x) = 0$. □

Lemma 7.6.7 gives an algorithm to compute $I_{\mathcal{A}}$, namely, first compute a $\mathbb{Z}$-basis $\mathcal{W}$ for the kernel of $\mathcal{A}$ to obtain the ideal $I_{\mathcal{W}}$. Next, saturate this with respect to a variable y_α, and then saturate the result with respect to a second variable, and repeat.

Exercises for Section 7.6

1. Let $\mathcal{A} \subset \mathbb{Z}^n$ be a finite collection of exponent vectors for Laurent monomials. Show that the map $\varphi_{\mathcal{A}} \colon (\mathbb{F}^\times)^n \to \mathbb{P}^{\mathcal{A}}$ is injective if and only if $\mathcal{A}$ affinely spans $\mathbb{Z}^n$. (That is, if the differences $\alpha - \beta$ for $\alpha, \beta \in \mathcal{A}$ linearly span $\mathbb{Z}^n$.)

2. Show that $\mathbb{F}[x_1, x_1^{-1}, x_2, x_2^{-1}, \ldots, x_n, x_n^{-1}] \simeq \mathbb{F}[X]$, where $X = \mathcal{V}(x_1 \cdots x_n x_{n+1} - 1) \subset \mathbb{F}^{n+1}$.

3. Complete the implied proof of Theorem 7.6.6.

7.7 Notes

[96]

Chapter 8

Toric varieties

Vague Outline:

1. Toric varieties from polytopes, in particular the relation of the geometry to the polytope. One-parameter limits, torus orbits, and singularities.
2. Toric varieties from fans. Relations to all previous froms of toric varieties.
3. Real toric varieties, positive part, algebraic moment map, identification with polytope, linear precision, and structure of real toric varieties via glueing.
4. Regular polyhedral subdivision and toric degenerations.
5. Kushnirenko's Theorem via Hilbert series
6. Bernstein's Theorem (could use an earlier study of simplices or binimial systems)
7. Polyhedral homotopies

What follows are notes from a talk given by Frank, they have little relation to the eventual final form of the material of the chapter.

8.1 Toric varieties

Intro

I'm going be talking about projective toric varieties because I have a specific goal in mind. However, some of the things (with extra work) can translate to toric varieties as well.

8.2 Projective Toric Varieties

Before, I took $\mathcal{A} \subset \mathbb{Z}^n$ and I looked at the affine hull of $\mathcal{A}$. Now, I'm going to take

$$P := \text{conv}(\mathcal{A}) \;=\; \{\sum \lambda_\alpha \cdot \alpha : \sum \lambda_\alpha = 1\}.$$

This forms a *polytope*. Draw some polytopes

Polytopes have faces of various dimensions: the codimension 1 faces are caled *facets*. The zero-dimensional ones are called *vertices*. Facets have the nice feature that there's a normal vector to them.

Sometimes, I'll write $X_{\mathcal{A}}$ or X_P, and what I say really just depends upon these structures here. I assume the affine span of $\mathcal{A}$ is $\mathbb{Z}^n$, and I might as well assume that $0 \in \mathcal{A}$. Now I consider a map

$$\varphi_{\mathcal{A}}(\mathbb{F}^\times)^n \to \mathbb{P}^{\mathcal{A}}$$

defined as

$$x \mapsto [x^{\beta+\alpha} : \alpha \in \mathcal{A}]$$

With β, I just shift the vector, and so I'm not really doing anything at all. If I multiply a whole polynomial by a given monomial, I haven't changed the zero-set of my polynomial by one bit. So this map doesn't depend on $\mathcal{A}$, it depends on $\mathcal{A}$ upto translation. With zero in there, it has the nice fact that $x^0 = 1$ and so we have the map to projective coordinates of the form

$$[1, x_1^{\alpha_1}, \cdots].$$

This is just a sketch of a theory.

Then $(\mathbb{F}^\times)^n$ acts on $\mathbb{P}^{\mathcal{A}}$. Here, $x \in (\mathbb{F}^\times)^n$ hits $y \in \mathbb{P}^{\mathcal{A}}$ with the *diagonal action*

$$xy = [x^\alpha y_\alpha : \alpha \in \mathcal{A}.$$

Lemma 8.2.1. *$\varphi_{\mathcal{A}}((\mathbb{F}^\times)^n)$ is the orbit $(\mathbb{F}^\times)^n$ on $[1, 1, \cdots, 1]$.*

This implies that the algebraic torus $(\mathbb{F}^\times)^n$ acts on the toric variety $X_{\mathcal{A}}$. Moreover, $\varphi_{\mathcal{A}}$ maps $(\mathbb{F}^\times)^n$ isomorphically to its image $\varphi_{\mathcal{A}}((\mathbb{F}^\times)^n)$, and this is a dense subset of X_A. There is an open subset of the image that is the torus. Here, we apply a compactification.

But rather than talk about this in complete generality, let's just look at some simple examples.

Example 8.2.2. Let's suppose we have $\mathcal{A} = \{(0,0), (1,0), (0,1)\}$. Then P is the unit simplex. Then, I have the map

$$\begin{aligned}(\mathbb{F}^\times)^2 &\to \mathbb{P}^2\\ (t,u) &\mapsto [1,t,u] \subset U_0 \sim \mathbb{A}^2\end{aligned}$$

It turns out that $X_{\mathcal{A}}$ here is just $\mathbb{P}^2$.

I'd like to decompose $\mathbb{P}^2$ according to its orbits. What are the orbits? Well, there's $[1, x, y]$, sets of the form $xy \neq 0$. There's also $[1, x, 0]$ and $[1, 0, y]$. Another orbit of this action is $[0, 1, y]$. Lastly, there's three zero-dimensional orbits, which correspond to the three basis points: $[1, 0, 0]$, $[0, 1, 0]$, and $[0, 0, 1]$.

Let's decompose my polytope P. The whole polytope is a face, but it also has a horizontal edge, a vertical edge, and a diagonal edge. It also has a lower-left corner point, the right vertex, and the top vertex. When I draw these faces, I really mean the (relative)

interior of the face, because the boundary is composed of the lower-dimensional stuff. There's a map here from simplectic space that I won't discuss. There's a different map for non-projective varieties that are algebraic. In the projective case, the map is a *moment map*.

This always happens: you can always decompose a toric variety into pieces.

There's a very illustrative example where I take the unit square, but I won't do this. I might exemplify that by looking here. I'm going to show you how the facets glue together.

8.3 The Facets

(I'm going to talk about limits, which correspond with our usual notion in $\mathbb{R}$ or $\mathbb{C}$. In one-dimension, algebraic geometry gives us a notion. In complex analysis, this might be called the *Riemann extension theorem*.)

Now, let F be a facet of P. Let η be the inward-pointing normal (dual) vector. Now η a priori some vector, but I can choose $\eta \in (\mathbb{Z}^n)^*$, but I want it to be *primitive*: If I look at $\mathbb{Q}\eta \cap \mathbb{Z}^n$, then this should just be $\mathbb{Z} \cdot \eta$.

I need something to help me take the limit. Notice that

$$\mathcal{F}^\times \ni t \mapsto (t^{\eta_1}, t^{\eta_2}, \ldots, t^{\eta_n}) \in (\mathbb{F}^\times)^n$$

acts on $\mathbb{P}^{\mathcal{A}}$ by

$$t \cdot y = [t^{\eta \cdot \alpha} y_\alpha : \alpha \in \mathcal{A}],$$

which is exactly how η is evaluating.

Suppose that $a = \eta(F)$. Then η applied to anything on F should be constant. Notice that if $\alpha \in \mathcal{A}$, $\eta \cdot \alpha \geq a$ and equality holds iff $\alpha \in F$.

What happens if I apply t to $\varphi_{\mathcal{A}}(x)$, with $x \in (\mathbb{F}^\times)^n$? This looks like

$$\begin{aligned} t \cdot \varphi_{\mathcal{A}}(x) &= [t^{\eta_\alpha} \cdot y_\alpha : \alpha \in \mathcal{A}] \\ &= [t^{-a+\eta\cdot\alpha} \cdot x^\alpha : \alpha \in \mathcal{A}] \end{aligned}$$

So as we take the limit as $t \to 0$, $\alpha \notin F$, we get $t^{pos} x^\alpha$, $\alpha \in F$, x^α????????

If F is a facet of P, then I have that $\mathbb{P}^{\mathcal{A}\cap F}$ naturally embeds as a coordinate subspace of $\mathbb{P}^{\mathcal{A}}$. And if I consider the limit

$$\lim_{t\to 0} t\varphi_{\mathcal{A}}(x) = \varphi_{F\cap\mathcal{A}}(x).$$

In summary:

1. X_P is a disjoint union:

$$X_P = \bigcup_{\text{faces } F \subset P} \varphi_{F \cap \mathcal{A}}((\mathbb{F}^\times)^n).$$

These factors of the coproduct are the orbits. If you take the closures, every face F of P gives rise to a sub toric variety $X_{F\cap\mathcal{A}}$ of $X_\mathcal{A}$.

Now, the question is, how do they fit together? I won't get into that, because toric varieties can be singular. It turns out that what really matters (in terms of the singularities) is the angles of how things meet.

8.4 Over $\mathbb{R}$

What about over the real numbers? Let's get even more extreme. Let's say our "field" is $\mathbb{R}_>$. Let

$$X^+ := X_\mathcal{A} \cap \Delta_\mathcal{A}$$

and this is just the positive part of $\mathbb{RP}^\mathcal{A}_2$. This turns out to be the closure of $\varphi_\mathcal{A}(\mathbb{R}^n_>)$.

Here's a great fact: This is homeomorphic to the polytope. A polytope is a manifold with boundary. The homeomorphism really respects the angles. You can either go to your symplectic geometry friends and ask them for a moment map, or you can go to friends in algebraic statistics and ask them for what ever they call it.

Luis called it a *tautolagical map*. I call it the *algebraic moment map*. This is the map

$$\tau : \Delta_\mathcal{A} \to P$$

Recall that you can identify $\Delta_\mathcal{A}$ with the standard unit vectors in $\mathcal{RP}^\mathcal{A}$. What do I do? The map is

$$e_\alpha \mapsto \alpha.$$

A typical point in the probability simplex is $[y_\alpha : \alpha \in \mathcal{A}]$ with $y_\alpha \geq 0$ and $\sum y_\alpha = 1$. What does the map do to this? The image is $\sum y_\alpha \cdot \alpha$, and I call this a tautological map for this reason. It's essentially a linear map, but it's really an algebraic moment map. The fact is that under τ, X_A is homeomorphic to P.

This is the positive part of toric variety.

Notice that inside the torus $(\mathbb{R}^\times)^n = \{\pm 1\}^n \times \mathbb{R}^n_>$, if I take $\varepsilon \in \{\pm 1\}^n$, note that ε acts on $\mathbb{RP}^\mathcal{A}$. In turn, I can act with ε on $X_A \cap \Delta_\mathcal{A}$, and this is homeomorphic to $X_A \cap \Delta_\mathcal{A}$. So for all 2^n possible ε, I get a copy of P. What happens to my edges? Some of them have zero coordinates. Each facets have 2^{n-1} elements under this sign group. The way they're glued together come from the signs of the primitive normal vector and looking at its coefficients modulo 2.

8.5 The Punchline

I've talked about the geometry and structure of toric varieties. I was also giving a dictionary between these and polytopes. Now, I'm going to go on to our main question:

How many solutions are there to a system of polynomial equations

$$f_1(x_1, \ldots, x_n) = \cdots = f_n(x_1, \ldots, x_n) = 0$$

where the suppose of $f_i = \mathcal{A}$?

This is called a *sparse system*.

Sometimes these people require that all of these vectors in $\mathcal{A}$ lie in the convex hull, but we don't do that here.

These f_i are polynomials on the algebraic torus $(\mathbb{F}^\times)^n$. Recall that these correspond to linear polynomials on the toric variety $X_\mathcal{A} = \overline{\varphi_\mathcal{A}((\mathbb{F}^\times)^n)}$. Recall that we have dimension-many linear polynomials, namely n of them.

Last time, we said that we can in this situation count the number of solutions. But I want to replace $\varphi_\mathcal{A}((\mathbb{F}^\times)^n)$ by $X_\mathcal{A}$. So, I'm just going to answer Question when they're generic.

We need to determine the Hilbert polynomial. Here's it's relatively easy to determine the degree of HP of $X_\mathcal{A}$. I'll sketch the argument of my bounds.

What I'm going to do is, as before, I'll homogenize my map, but I'll write it a little bit differently:

$$\begin{aligned} \psi : \mathbb{F}^\times \times (\mathbb{F}^\times)^n &\to \mathbb{P}^\mathcal{A} \\ (t, x) &\mapsto [tx^\alpha : \alpha \in \mathcal{A}] \end{aligned}$$

Then

$$\mathbb{F}[X_\mathcal{A}] = \mathbb{F}[y_\alpha]/\ker\psi = \text{image}\psi = \mathbb{F}[tx^\alpha : \alpha \in \mathcal{A}]$$

A typical element in here is $t^{\text{mess}}x^{\text{bigmess}}$. What is the degree of this? It's just "mess".

Let me give you a "geometry of numbers"-way of looking at this. My exponent vectors $\mathcal{A}$ lie inside some $\mathbb{Z}^n$. When I homogenize, they are in $1 \times \mathbb{Z}^n$. So essentially, I have

$$\mathbb{N}\mathcal{A}^+ \leftrightsquigarrow \text{ monomials}$$

Points in $\mathbb{N}A^+$ that are sums of d monomials in $\mathcal{A}^+$ $\mathbb{R}_{>0}\mathcal{A}^+ \cap (\text{first coordinate} = d) = d \cdot P$.

Thus, we have $HF_\mathcal{A}(d) \leq$ the Ehrhart polynomial (we have a lattice polytope). This has the form $\text{Vol}(P) \cdot d^n +$ lower order terms in d.

Thus

$$\text{EP}_P(d - m) \leq HP_\mathcal{A} \leq \text{EP}_P(d)$$

so my HP looks like

$$n! \cdot \text{Vol}(P) \cdot \frac{d^n}{n!} + \text{ lower order terms}$$

We have Kouchnerenko's Theorem (1976):

Theorem 8.5.1. *The number of non-degenerate solutions to a system of polynomials $f_1 = \cdots = f_n = 0$ with support of each f_i being $\mathcal{A}$ is at most*

$$n! \operatorname{Vol}(\operatorname{conv}(\mathcal{A})).$$

When $\mathbb{F}$ is closed and the f_i are generic, then this is equal to $n! \operatorname{Vol}(\operatorname{conv}(\mathcal{A}))$.

Then, the question over the real numbers is interesting.

8.6 Bernstein's theorem

You want to solve equations

$$f_1(x_1, \dots, x_n) = \cdots = f_n(x_1, \dots, x_n) = 0 \tag{8.1}$$

Note that numerics almost always place you in the generic situation. The polynomial f_i has support $A_i := P_i \cap \mathbb{Z}^n$, where $P_1, \dots, P_n$ are integer polytopes.

Bernstein's Theorem says:

Theorem 8.6.1. *The number of non-degenerate solutions to (8.1) is less than or equal to the mixed volume* $\operatorname{MixedVol}(P_1, \dots, P_n)$.

When $\mathbb{F}$ is closed and the f_i are generic (given their support), the number is exactly the mixed volume.

Suppose each P_i is a segment (or rather just two points u_i and v_i). This is the same thing as the segment (0 and $v_i - u_i$). So, I'll assume that P_0 is of the form

$$P_0 = \operatorname{conv}(0, v_i : v_i \in \mathbb{Z}^n)$$

So, we can say $x^{v_i} = c_i$ for $i = 1, \dots, n$. This set of equations has the form

$$\varphi_V^{-1}(c_1, \dots, c_n)$$

where

$$\begin{aligned} \varphi_V : (\mathbb{F}^\times)^n &\to (\mathbb{F}^\times)^n \\ x &\mapsto (x^{v_1}, \dots, (x^{v_i}) \end{aligned}$$

Notice that φ_V is a group homomorphism. Thus, we compute the size:

$$|\varphi_V^{-1}(c)| = |\ker \varphi_V| = \{x : \varphi_V(x) = 1\}$$

It turns out that

$$\ker \varphi_V = \operatorname{Hom}\left(\frac{\mathbb{Z}^n}{\langle v_1, \dots, v_n \rangle}, \mathbb{F}^\times\right)$$

So, I just need to count this. There are a number of interpretations for this:

$$\left|\frac{\mathbb{Z}^n}{\langle v_1, \dots, v_n \rangle}\right|$$

This is equal to the volume $\det(v_1 | \cdots | v_n)$ of the fundamental parallelepiped Π determined by the vectors v_i:

$$\Pi = \{\sum \lambda_i v_i : 0 \le \lambda_i \le 1\} = P_1 + \cdots + P_n$$

the Minkowski sum of the polytopes P_i.

Then, the *mixed volume* is defined as the coefficient of $t_1 \cdots t_n$ in

$$\begin{aligned}
&\mathrm{Vol}(t_1 P_1 + \cdots t_n P_n) \\
&= \det(t_1 v_1 | \cdots | t_n v_n) \\
&= t_1 \cdots t_n \det(v_1 | \cdots | v_n) \\
&= \mathrm{MixedVol}(P_1, \ldots, P_n) \\
&= \text{Number of solutions}
\end{aligned}$$

The permutations $\sigma \in S_n$ index the simplices in a triangulation of the n-cube C_n. Given σ, take the convex hull of $0, 0 + e_{\sigma(1)}, 0 + e_{\sigma(1)} + e_{\sigma(2)}, \ldots, 0 + e_{\sigma(1)} + \cdots + e_{\sigma(n)}$.

8.7 Proof of Bernstein's Theorem

The polynomial f_i has the form

$$f_i = \sum_{\alpha \in A_i} c_{i,\alpha} x^\alpha$$

I will choose $v_{i,\alpha} \in \mathbb{Z}, \alpha \in A_i, i = 1, \ldots, n$. Given this choice, I will define:

$$f_i(x; t) := \sum_{\alpha \in A_i} c_{i,\alpha} t^{\nu_{i,\alpha}} x^\alpha \qquad t \in \mathbb{F}^\times \tag{8.2}$$

When $t = 1$, we get back the original system. This defines some algebraic curve C in $(\mathbb{F}^\times)^n \times \mathbb{F}^\times$. I'd like to look at C near $t = 0$. It will be a little strange because of how the exponents might behave.

8.7.1 Puiseaux series

To work on Puiseaux series, we really need our field to be algebraically closed. So, what I'll do here is prove the equality statement of Theorem 8.6.1.

Puiseaux series are the algebraic closure of the field of formal power series in t. Our formal power series are of the form

$$\sum_{n \geq N} c_n t^n$$

and thus our Puiseaux elements look like

$$\sum_{n \geq N} c_n t^{n/M}.$$

So, what are the Puiseaux solutions to (8.2)? We look for solutions of the form

$$X_i(t) = X_i(0) t^{u_i} + \text{ higher order terms in } t \qquad u_i \in \mathbb{Q}. \tag{8.3}$$

What is $X(t)$? Let's take (coordinate-wise) powers:

$$X(t)^\alpha = X(0)^\alpha t^{u\cdot\alpha} + \text{ higher order terms in } t$$

Now, when I plug this into (8.2), I obtain

$$f_i(X(t);t) = \sum_{\alpha\in A_i} c_{i,\alpha}\big(t^{\nu_{i,\alpha}+u\cdot\alpha}X(0)^\alpha + \text{ higher order terms in } t\big) = 0.$$

I have a system of equations like this. If a power series is zero, that means that all of the coefficients must cancel.

It's necessary[1] that

$$\min_{\alpha\in A_i}\{\nu_{i,\alpha} + U\cdot\alpha\} \tag{8.4}$$

occurs at **least** twice each $i = 1,\dots,n$.

Given (8.4), suppose the occurances are all exactly twice. Then the lowest order terms are binomials of the form (cancelling $t^{\min}$), you get the system

$$c_{i,\alpha}X(0)^\alpha + c_{i,\beta}X(0)^\beta = 0 \tag{8.5}$$

with $\alpha,\beta\in A_i$.

The number of solutions is exactly the volume of the parallelepiped Π, which is the sum of the volumes of line segments.

Now, we go back to the form for X in of our anzatz(sp) in (8.3) and we add a second term.

I need to thus count the number of solutions to the tropical problem (8.4), and for each of those, I need to solve the volume problem (8.5). So, that's what I'm going to do. From here on out, now it's going to be geometric combinatorics.

8.8 Counting the number of solutions in (8.5)

8.8.1 Some Geometric Combinatorics

Each $A_i\subset\mathbb{R}^n$, and I'm going to lift to $\mathbb{R}^{n+1}$ by considering $A_i\ni\alpha\mapsto(\alpha,\nu_{i,\alpha})$. I'll call this lifted set A_i^+.

Now, I'm going to take $P_i^+ := \operatorname{conv}(A_i^+)$, and I'll set $V = P_1^+ + \cdots + P_n^+$. The faces of V are places where a linear functional takes its minimum. These are sums of the corresponding faces of the P_i^+.

The facets whose inward-pointing normal vector has the form $(u,1)$ where $u\in\mathbb{Q}^n$ are called *lower facets*. These project to $\mathbb{R}^n$ to give some polytopes. The form a *polyhedral subdivision*, a union of polyhedra that meet certain intersection criteria. It is a subdivision of the projection of V, and this is of course our original Minkowski sum $P_1+\cdots+P_n$.

[1] No, that can't be right. Is that right?

We are going to try to identify parts of the subdivision with the parallelepipeds, and then compute the mixed volume. The proper term for this is called a *regular mixed subdivision*.

I have a lower facets F with normal vector of the form $(u, 1)$. Then $(u, 1)$ has a minimum face f_i on each P_i^+. Then

$$F = f_1 + \cdots + f_n. \tag{8.6}$$

If you take a point of coordinates $(\alpha, \nu_{i,\alpha})$ and dot this with $(u, 1)$, then you obtain

$$u \cdot \alpha + \nu_{i,\alpha}.$$

Notice that this is exactly the expression in (8.4). That means that on each of these, I need to have at least two of these f_i lifted.

The genericity assumption on these ν values is that if each is an edge, none of them contain two or points on them. It means that if you had a face, then one of them had to have been lifted higher. The point is $(u, 1)$ solves (8.4) exactly when each f_i is an edge with only two $(\alpha, \nu_{i,\alpha})$ on it.

Note (8.6) is a Minkowski sum of edges. So, the projection of the sum of the line segments $(\alpha, \nu_{i,\alpha}) - -(\beta, \nu_{i,\beta})$ are the Minkowski sums of the form $\alpha - -\beta$. Thus, the solutions to (8.4) are the parallelepipeds among the lower faces of $P_1^+, \cdots, P_n^+$ generated by edges $f_1, \ldots, f_n$, an edge in each P_i^+. Projecting one of these lower faces to $\mathbb{R}^n$ gives a parallelepiped generated by the segment $\alpha - -\beta$, where $\alpha, \beta \in A_i$.

I have an exact bijection between solutions to (8.4) and the Minkowski sums of edges. The term for this is *mixed cells*.

In this mixed subdivision, there are two classes of faces:

1. Mixed cells (These are special parallelepipeds. Each edge comes from a different P_i^+).
2. The rest (Each one excludes some P_i^+).

Recall the mixed volume is the coefficient $t_1 \cdots t_n$ in

$$t_1 P_1 + \cdots + t_n P_n.$$

What happens in each case?

1. Mixed cells: $\text{vol} = t_1 \cdots t_n \cdot \text{vol}(t = 1)$.
2. Here, each cell will lack some t_i.

The sum of the volumes of these mixed cells **IS** the mixed volume. (The second case doesn't contribute because some t_i is always missing).

Chapter 9

Tropical geometry

Tropical geometry denotes a young mathematical discipline in which the basic operations are performed over the semiring $(\mathbb{R}, \min, +)$ (or $(\mathbb{R}, \max, +)$. The name "tropical" was coined by French mathematicians, including Jean-Eric Pin, to honor the pioneering work of their Brazilian colleague Imre Simon on the max-plus algebra.

Tropical geometry can be seen as the geometry resulting from a degeneration process of toric geometry. As a consequence of this process, complex toric varieties can be replaced by the real space $\mathbb{R}^n$ and complex algebraic varieties by polyhedral cell complexes.

The origins of the tropical degeneration ideas go back to Viro's patchworking method (in the 70's), to the Bergman complex (in the 70's), and to Maslov's dequantization of positive real numbers (in the 80's). As a consequence of these developments, in different areas of mathematics different names were used for tropical varieties: logarithmic limit sets, Bergman fans, Bieri-Groves sets, and non-archimedean amoebas. In the last years, the various research directions have been fruitfully merged, generalized and advanced under what is now called tropical geometry. These developments were based on substantial progress in understanding the concept of an amoeba that was introduced by I. Gelfand, M. Kapranov and A. Zelevinsky (in the early 90's) as the logarithmic image of a complex variety.

While the roots of tropical geometry come from algebraic geometry and valuation theory, tropical varieties are profitably approached via polyhedral combinatorics. In fact, tropical hypersurfaces can be defined in a combinatorial and in an algebraic way. For the combinatorial definition, let $(\mathbb{R}, \oplus, \odot)$ denote the *tropical semiring*, where

$$x \oplus y \;=\; \min\{x, y\} \quad \text{and} \quad x \odot y \;=\; x + y\,.$$

Sometimes the underlying set $\mathbb{R}$ of real numbers is augmented by ∞.

In this chapter, we will discuss some elementary, rather combinatorial tropical concepts. Based on this, Chapter 10 will investigate some more advanced topics.

9.1 Tropical hypersurfaces and the dual subdivision

A *tropical monomial* is an expression of the form $c \odot x^\alpha = c \odot x_1^{\alpha_1} \odot \cdots \odot x_n^{\alpha_n}$ where the powers of the variables are computed tropically as well (e.g., $x_1^3 = x_1 \odot x_1 \odot x_1$). This tropical monomial represents the classical linear function

$$\tilde{f} : \mathbb{R}^n \to \mathbb{R}, \quad (x_1, \ldots, x_n) \mapsto \alpha_1 x_1 + \cdots + \alpha_n x_n + c.$$

A *tropical polynomial* $f = \oplus_{\alpha \in A} c_\alpha \odot x^\alpha$ is a finite tropical sum of tropical monomials and thus represents the (pointwise) minimum function of linear functions,

$$\tilde{f} : \mathbb{R}^n \to \mathbb{R}, \quad (x_1, \ldots, x_n) \mapsto \min_{\alpha \in A} \{c_\alpha + \sum_{i=1}^n \alpha_i x_i\}.$$

Let $\mathbb{R}_{\text{trop}}[x_1, \ldots, x_n]$ denote the semiring of tropical polynomials.

Theorem 9.1.1. *Every tropical polynomial $f \in \mathbb{R}_{\text{trop}}[x_1, \ldots, x_n]$ defines a continuous, concave and piecewise-linear function $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$.*

Proof. Continuity and piecewise-linearity are clear. In order to show concavity, let $f = \oplus_\alpha f_\alpha$ be a tropical polynomial with terms f_α, and let $b, c \in \mathbb{R}^n$ and $\lambda \in [0, 1]$.

Let $\beta, \gamma, \delta \in \operatorname{supp} f$ with $f(b) = f_\beta(b)$ and $f(c) = f_\gamma(c)$, and $f(b + (1-\lambda)c) = f_\delta(b + (1-\lambda)c)$. Then

$$\begin{aligned} \lambda f(b) + (1-\lambda) f(c) &= \lambda f_\beta(b) + (1-\lambda) f_\gamma(c) \\ &\leq \lambda f_\delta(b) + (1-\lambda) f_\delta(c) \;=\; f_\delta(\lambda b + (1-\lambda)c) \\ &= f(\lambda b + (1-\lambda)c). \end{aligned}$$

□

Considering f as a concave, piecewise linear function, Figure 9.1 shows the graph of f and the resulting curve $\mathcal{T}(f) \subset \mathbb{R}^2$ for a quadratic tropical polynomial.

At each given point $x \in \mathbb{R}^n$ the minimum of the piecewise linear function $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$ is either attained at a single linear function or at more than one of the linear functions ("at least twice"). The *tropical hypersurface* $\mathcal{T}(f)$ of a tropical polynomial f is defined as the set of points $x \in \mathbb{R}^n$ where that minimum is attained at least twice. Equivalently, $\mathcal{T}(f)$ is given by the *corner locus* of the piecewise linear function $\tilde{f}$, i.e., with the the set of points where $\tilde{f}$ is not differentiable.

In this section we will see that tropical hypersurfaces have the structure of a polyhedral complex. The combinatorial structure is given by a dual subdivision. To explain this in detail, we will apply some of the notation from Appendix A.5 on polyhedral geometry.

Let $\mathcal{A} \subset \mathbb{N}_0^n$ be finite and $f(x_1, \ldots, x_n) = \bigoplus_{\alpha \in \mathcal{A}} c_\alpha \cdot x^\alpha$ be a tropical polynomial with $c_\alpha \in \mathbb{R}$ for all $\alpha \in \mathcal{A}$. The *extended Newton polytope of a tropical polynomial* $f = \bigoplus_{\alpha \in \mathcal{A}} c_\alpha \odot x^\alpha$ is the convex hull

$$\text{NP}^{\text{e}}(f) \;=\; \text{conv}\{(\alpha, c_\alpha) : \alpha \in \mathcal{A}\} \;\subset\; \mathbb{R}^{n+1}.$$

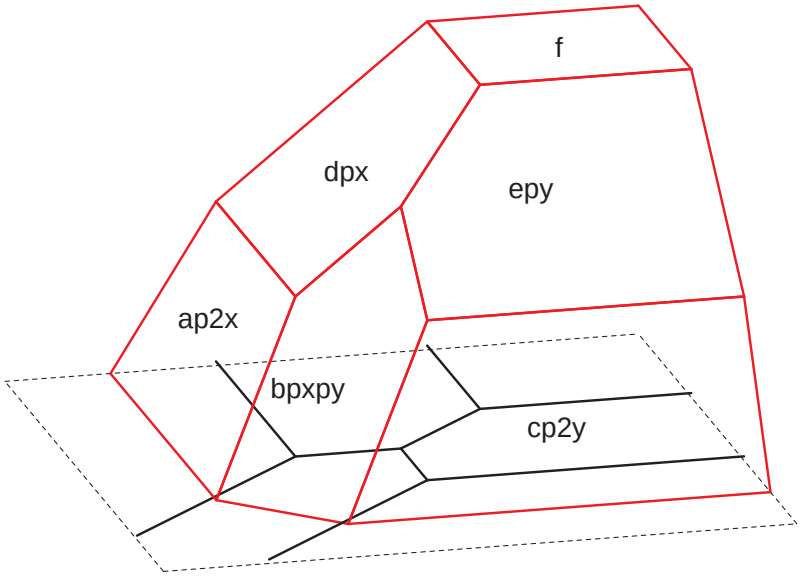


Figure 9.1: The graph of $\tilde{f} : \mathbb{R}^2 \to \mathbb{R}$ for a tropical polynomial $f = a \odot x^2 \oplus b \odot x \odot y \oplus c \odot y^2 \oplus d \odot x \oplus e \odot y \oplus f$.

Definition 9.1.2. Let f be a tropical polynomial in $x_1, \ldots, x_n$ and $\mathcal{N}(\mathrm{NP}^{\mathrm{e}}(f))$ be the normal fan of its extended Newton polytope. The set

$$\mathcal{L}_f \;=\; \{F \text{ face of } \mathrm{NP}^{\mathrm{e}}(f) \;:\; \text{there exists some } w \in C(F) \text{ with } w_{n+1} < 0\}$$

is called the *lower complex* of f, where $C(F)$ denotes the normal cone of F. $\mathcal{L}_f$ constitutes a polyhedral complex.

While we often identify a polyhedral complex with its underlying support set, the following statement reveals the structure of $\mathcal{T}(f)$ as a polyhedral complex. Here, for $k \in \mathbb{N}$, let $\mathcal{N}_k(\mathrm{NP}^{\mathrm{e}}(f))$ denote the set of faces of the normal fan $\mathcal{N}(\mathrm{NP}^{\mathrm{e}}(f))$ of dimension at most k.

Theorem 9.1.3. *Let f be a tropical polynomial in $x_1, \ldots, x_n$. Then*

$$\mathcal{T}(f) \;=\; -\mathcal{N}_n(\mathrm{NP}^{\mathrm{e}}(f)) \cap \{x \in \mathbb{R}^{n+1} \;:\; x_{n+1} = 1\}\,. \tag{9.1}$$

Moreover, $\mathcal{T}(f)$ is a pure polyhedral complex in $\mathbb{R}^n$ of dimension $n-1$.

Proof. By definition, $\mathcal{T}(f)$ is the set of all points $w \in \mathbb{R}^n$ such that the linear form with coefficients $(w_1, \ldots, w_n, 1)$ attains its minimum at more than one of the points $(\alpha_1, \ldots, \alpha_n, c)$ representing a term of f. Hence, a point w is in $\mathcal{T}(f)$ if and only if the face of $\mathrm{NP}^{\mathrm{e}}(f)$ maximizing the linear function $x \mapsto (-w_1, \ldots, -w_n, -1)^T \cdot x$ has dimension at least 1. This set coincides with the set (9.1).

The fact that $\mathcal{T}(f)$ is a polyhedral complex then follows from the first one. The dimension statement follows from the observation that none of the maximal cells of $\mathcal{N}(\mathrm{NP}^{\mathrm{e}}(f))$ is contained in the hyperplane $\{x \in \mathbb{R}^{n+1} \;:\; x_{n+1} = 1\}$. □

The Newton polytope $\mathrm{NP}(f)$ of a tropical polynomial f comes with a *privileged subdivision* $\mathrm{subdiv}(f)$. If we project down the faces of the lower complex $\mathcal{L}_f$ by forgetting the last coordinate we get a subdivision of $\mathrm{NP}(f)$. Each cell of the subdivision is convex.

There is a one-to-one correspondence between the faces of $\mathcal{T}(f)$ and the faces of $\mathrm{subdiv}(f)$. Explicitly, the dual face $F^\vee$ of a face F of $\mathcal{T}(f)$ is the maximal face G of $\mathrm{subdiv}(f)$ such that the set $-F \times \{-1\}$ is contained in the normal cone of the lifted face $\hat{G} \subset \mathrm{NP}^{\mathrm{e}}(f)$.

Theorem 9.1.4. *The tropical hypersurface $\mathcal{T}(f)$ and the subdivision* $\mathrm{subdiv}(f)$ *are* dual *to each other, i.e., we have*

1. *F and $F^\vee$ span orthogonal real affine space.*

2. $\dim F^\vee = n - \dim F$.

3. *If E is a face of F then $F^\vee$ is a face of $E^\vee$.*

Proof. The first statement is clear from the definition of the outer normal fan $\mathcal{N}_n(\mathrm{NP}^{\mathrm{e}}(f))$. Let $k := \dim F$. Then the lifted face $\hat{F}$ is of dimension k as well, and $\hat{F}$ is a lower face of the lifted polytope $\mathrm{NP}^{\mathrm{e}}(f)$. By Theorem (9.1.3), the dimension of the dual cell $F^\vee$ is $n - k$.

If E is a face of F then there are two corresponding faces $\tilde{G}$ and $\tilde{H}$ in the extended Newton polytope such that $\tilde{H}$ is a face of $\tilde{G}$. Hence, $F^\vee$ is a face of $E^\vee$. □

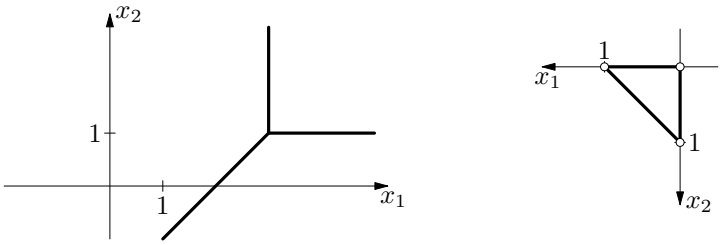


Figure 9.2: The tropical curve of a linear polynomial f in two variables and the Newton polygon of f.

We say that a tropical polynomial is *of degree at most d* if every term has (total) degree at most d. See Figure 9.2 for an example of a tropical line (i.e., the tropical variety of a linear polynomial in two variables) and Figure 9.3 for an example of a tropical cubic curve, as well as their dual subdivisions (whose coordinate axes are directed to the left and to the bottom to visualize the duality).

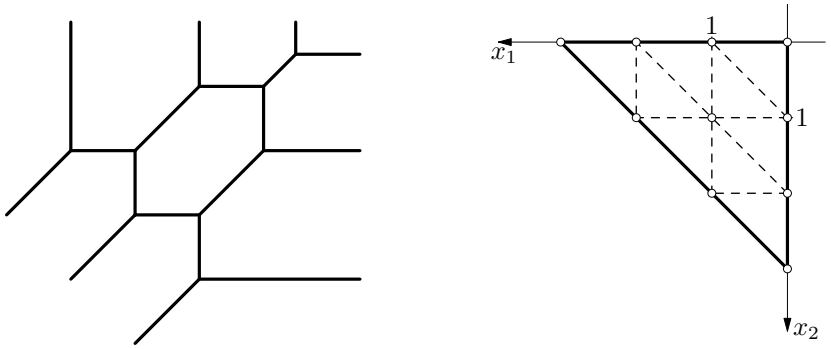


Figure 9.3: An example of a tropical cubic curve $\mathcal{T}(f)$ and the dual subdivision of the Newton polygon of f.

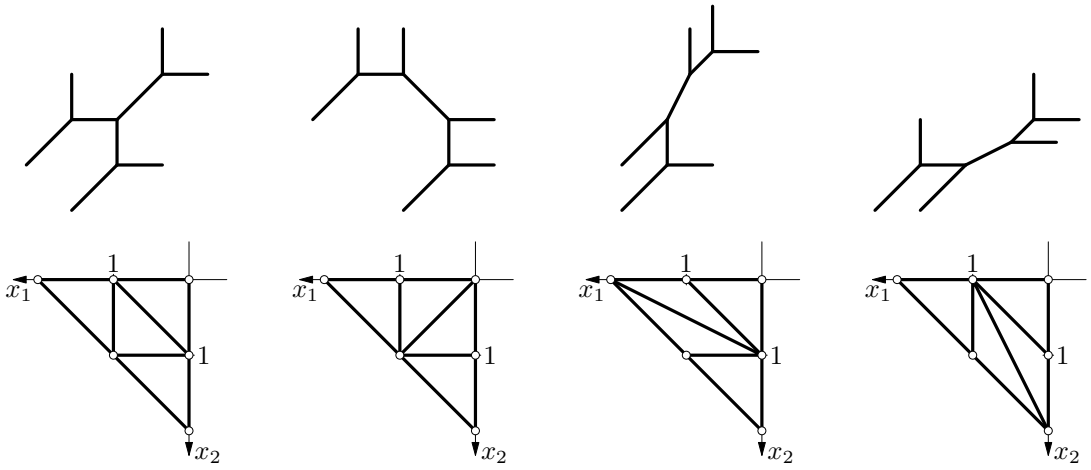


Figure 9.4: Types of tropical conics with a unimodular dual triangulation.

Example 9.1.5. Consider *quadratic curves in the plane* which are defined by a tropical polynomial of degree 2,

$$f \;=\; a_1 \odot x \odot x \oplus\; a_2 \odot x \odot y \;\oplus\; a_3 \odot y \odot y \;\oplus\; a_4 \odot x \;\oplus\; a_5 \odot y \;\oplus\; a_6\,.$$

The curve $\mathcal{T}(f)$ is a graph which has six unbounded edges and at most three bounded edges. The unbounded edges are pairs of parallel half rays in the three coordinate directions. Figure 9.4 shows the four possible combinatorial types of tropical curves if the dual subdivision is a unimodular triangulation. In Exercise ..., the reader will investigate the dual subdivision in dependence of the coefficients $a_1, \ldots, a_6$.

Any face F in a tropical hypersurface has a natural *multiplicity* (or *weight*). If F is j-dimensional then the dual cell $F^\vee$ is $(n-j)$-dimensional. Define

$$m_F \;=\; (n-j)!\,\mathrm{vol}'_{n-j}(F)\,,$$

where vol′ denotes the volume in the lattice $\mathbb{Z}(F)$ spanned by the integer vectors of F. In particular, if $\mathcal{T}(f)$ is a curve, then the multiplicity is the lattice length of the corresponding edge in the dual subdivision subdiv(f) (see Figure 9.5)

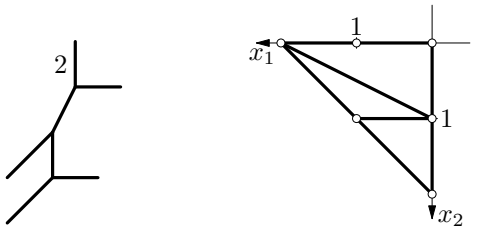


Figure 9.5: A tropical curve with a segment of weight 2 and its dual subdivision.

If f is a tropical polynomial whose Newton polytope Δ spans $\mathbb{Z}^n$, then the number of vertices N (counting multiplicities) results to

$$\sum_{C \text{ } n\text{-cell in } \Delta} n!\,\mathrm{vol}'_n(C) \;=\; \sum_{C \text{ } n\text{-cell in } \Delta} n!\,\mathrm{vol}_n(C) \;=\; \mathrm{vol}_n(\Delta)\,.$$

Similarly, the number of j-faces (counting multiplicity) is

$$\sum_{C \text{ } (n-j)\text{-cell in } \Delta} \mathrm{vol}'_{n-j}(C)\,.$$

Exercises

Exercise 9.1.6. Show that a face F of $\mathcal{T}(f)$ is unbounded if and only if $F^\vee$ is contained in the boundary of NP f.

9.2 Polyhedral characterization of tropical hypersurfaces

In this section we show that the balancing condition actually characterizes tropical hypersurfaces in n-space.

The polyhedral complexes $\mathcal{P}$ in $\mathbb{R}^n$ which we consider here consist of (possibly infinite) polyhedra. We call $\mathcal{P}$ *rational* if the slope of the affine span of each cell is rational.

For the case of a tropical curve it is easy to see that in each vertex a *balancing condition* (or *equilibrium condition*) is satisfied.

Lemma 9.2.1. *Let p be a vertex of a tropical curve $\mathcal{T}(f)$ in the plane, let $v^{(1)}, v^{(2)}, \ldots, v^{(r)}$ be the primitive lattice vectors in the directions of the edges emanating from p, and let $m_1, m_2, \ldots, m_r$ be the multiplicities of these edges. Then $\sum_{i=1}^{r} m_i v^{(i)} = 0$.*

Proof. Let Q be the convex r-gon dual to p in the subdivision subdiv(f). Up to a 90 degree rotation, the edges of Q coincide with the vectors $m_i \cdot v^{(i)}$. Since the edges of a convex polygon sum up the zero vector, the claim follows. □

This balancing condition generalizes to tropical hypersurfaces in $\mathbb{R}^n$ as follows. A weighted polyhedral $(n-1)$-complex $\mathcal{P} \subset \mathbb{R}^n$ is called *balanced* if for any $(n-2)$-face F of $\mathcal{P}$ the following condition holds: Let $F_1, \ldots, F_k$ be the neighboring $(n-1)$-faces of $\mathcal{P}$. A choice of a rotation direction w.r.t. F defines an orientation of these $(n-1)$-faces. The condition is that $\sum_{i=1}^{k} v_{F_i} = 0$, where v_{F_i} is the (uniformly oriented) normal vector in $\mathbb{Z}^n$ (including multiplicity).

Lemma 9.2.2. *Any tropical hypersurface in $\mathbb{R}^n$ is a pure rational balanced polyhedral complex.*

Proof. The only property which remains to be shown is the balancing condition. Let F be an $(n-2)$-face of the tropical hypersurface $\mathcal{T}(f)$ and $F_1, \ldots, F_k$ be the neighboring $(n-1)$-faces in the order consistent with the chosen rotation direction.

The dual face $F^\vee$ has dimension 2. Restricting the polynomial f to this dual face provides us with a lattice n-gon which lives in the lattice $\mathbb{Z}(F^\vee)$. Hence, $\sum_{i=1}^{k} v_{F_i} = 0$. □

In the following we discuss the converse and see that the balancing condition can be used to characterize tropical hypersurfaces.

Theorem 9.2.3. *The pure rational weighted balanced polyhedral complexes in $\mathbb{R}^n$ are exactly the tropical hypersurfaces in $\mathbb{R}^n$.*

Proof. By Lemma 9.2.2 any tropical hypersurfaces yields a pure rational weighted balanced polyhedral complex.

Conversely, let $\mathcal{P}$ be rational weighted balanced polyhedral complex. We show that $\mathcal{P}$ is the corner locus of a piecewiese linear function $f : \mathbb{R}^n \to \mathbb{R}$.

We define f inductively. First choose a connected component C_0 of $\mathbb{R}^n \backslash \mathcal{P}$ as a reference component, and define $f_{C_0} = 0$. Now let C' be a connected component of $\mathbb{R}^n \setminus \mathcal{P}$ such that a neighboring component C exist on which is already defined along an $(n-1)$-face.

Let $F := \overline{C} \cap \overline{C'}$ be the $(n-1)$-face separating C and C' and let ℓ_C be the affine-linear function which extends f_C to $\mathbb{R}^n$. Now define $f_{C'}$ as follows. Due to F the affine-linear extension of $f|_{C'}$ is fixed up to the rotation around the lifted $(n-1)$-face $\hat{F}$.

Let $u^* :=$ multiplicity $\cdot$ primitive normal vector of F.

Let

$$\ell'_C \;:=\; \ell_C + (u^*)^T x + \gamma$$

with γ a suitable constant such that g_C and g'_C coincide on F. Note that the term $(u^*)^T x$ is constant along F. Let $g_{C'}$ be the restriction of ℓ'_C to C'. Since the graph of g is simply connected, the construction is well-defined. □

An embedded graph G (allowing infinite half-rays as edges) in the plane $\mathbb{R}^2$ is a *rational graph* is a *rational graph* if all endpoints and directions have rational coordinates $\mathbb{Q}$, and each ray or segment has a positive integral multiplicity. A rational graph Γ is said to be *balanced* if at each vertex p the condition

$$\sum_{i=1}^{r} m_i v_i \ = 0$$

holds (...)

Corollary 9.2.4. *The rational, weighted, embedded graphs without isolated vertices are exactly the tropical curves in the plane.*

Proof. 90 degree rotation ... □

Exercises

Exercise 9.2.5. Let

$$f \ = \ a_1 \odot x \odot x \ \oplus \ a_2 \odot x \odot y \ \oplus \ a_3 \odot y \odot y \ \oplus \ a_4 \odot y \odot z \ \oplus \ a_5 \odot z \odot z \ \oplus \ a_6 \odot x \odot z.$$

be a homogeneous tropical quadratic form of degree 2. Determine conditions on $a_1, \ldots, a_6$ such that $\mathcal{T}(f)$ has six distinct half-rays.

Exercise 9.2.6. Show that a rational, weighted, embedded, balanced graph Γ in the plane without isolated vertices is a tropical curve of degree d if and only if Γ has d ends (counting multiplicity) in directions $(-1,-1)$, $(1,0)$ and $(0,1)$.

9.3 Tropical prevarieties

An intersection of finitely many tropical hypersurfaces is called a *tropical prevariety*. In this section, we will see that such an intersection obey certain intersection principles which we know from the intersection of projective hypersurfaces over algebraically closed fields. In particular, we will see the validity of a tropical Bézout and a tropical Bernstein Theorem.

Generalizing the dual subdivision of a single tropical hypersurface, a tropical prevariety can be regarded from a dual point of view. Let $f_1, \ldots, f_k$ be k tropical polynomials in $x_1, \ldots, x_n$. The coefficients of $f_1, \ldots, f_k$ induce a *privileged subdivision* $\Gamma(f_1, \ldots, f_k)$ of $\mathrm{NP}(f_1) + \cdots + \mathrm{NP}(f_k)$ by projecting down the lower hull of the Minkowski sum $\mathrm{NP}^{\mathrm{e}}(f_1) + \cdots + \mathrm{NP}^{\mathrm{e}}(f_k)$. For a generic choice of coefficients in the system $f_1, \ldots, f_k$ this subdivision is mixed (as defined in Appendix A.6).

The subdivision $\Gamma(f_1, \ldots, f_k)$ and the union $\mathcal{T}(f_1) \cup \cdots \cup \mathcal{T}(f_k)$ of tropical hypersurfaces are polyhedral complexes which are dual in the sense that there is a one-to-one correspondence between their cells which reverses the inclusion relations. Each cell

C in $\Gamma(f_1,\dots,f_k)$ corresponds to a cell A in the union $\mathcal{T}(f_1)\cup\cdots\cup\mathcal{T}(f_k)$ such that $\dim(C)+\dim(A)=n$, C and A span orthogonal real affine spaces and A is unbounded if and only if C lies on the boundary of $P(f_1)+\cdots+P(f_k)$. Furthermore we have that a cell A of $\mathcal{T}(f_1)\cup\cdots\cup\mathcal{T}(f_k)$ is in the intersection $\mathcal{I}:=\mathcal{T}(f_1)\cap\cdots\cap\mathcal{T}(f_k)$ if and only if the corresponding dual cell C in $\Gamma(f_1,\dots,f_k)$ is mixed.

A cell A in $\mathcal{I}$ can be written as $A=\bigcap_{i=1}^k A_i$ where $A_i\in X_i$. If we require that A lies in the relative interior of each A_i then this representation is unique. The dual cell C of A has then a unique decomposition into a Minkowski sum $C=F_1+\cdots+F_k$ where each F_i is dual to A_i. We will always refer to this decomposition if not stated otherwise.

An intersection $\mathcal{I}=X_1\cap\cdots\cap X_k$ is called *proper* if $\dim(\mathcal{I})=n-k$. $\mathcal{I}$ is *transversal along a cell* A of this complex if the dual cell $C=F_1+\cdots+F_k$ in the privileged subdivision of $P_1+\cdots+P_k$ satisfies

$$\dim(C)=\dim(F_1)+\cdots+\dim(F_k)\ .$$

We call the intersection *transversal* if for each subset $J\subset\{1,\dots,k\}$ the intersection is proper and transversal along each cell of the complex. In the dual picture a transversal intersection implies that the privileged subdivision of $P_1+\cdots+P_k$ is mixed. Note that in a transversal intersection each cell A of $\mathcal{I}$ lies in the relative interior of each cell A_i from X_i that is involved in the intersection.

In the case of a non-transversal intersection $\mathcal{I}$ we can perturb the hypersurfaces by a small parameter ε to obtain again a transversal intersection $\mathcal{I}_\varepsilon$. The *stable intersection* $\mathcal{I}_{\text{st}}$ is defined as the limit of these transversal intersections when ε goes to 0,

$$\mathcal{I}_{\text{st}}=X_1\cap_{\text{st}}\cdots\cap_{\text{st}}X_k=\lim_{\varepsilon\to 0}X_1^{(\varepsilon_1)}\cap\cdots\cap X_k^{(\varepsilon_k)}$$

Stable intersections are always proper and they have some more comfortable features. As mentioned above a tropical hypersurface $\mathcal{T}(g)\subset\mathbb{R}^n$ is a polyhedral complex of dimension $n-1$. The stable intersection of $\mathcal{T}(g)$ with itself gives the $(n-2)$-skeleton of $\mathcal{T}(g)$. In particular we can isolate the vertices of $\mathcal{T}(g)$ by stably intersecting $\mathcal{T}(g)$ $(n-1)$-times with itself.

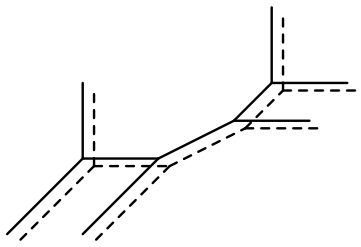


Figure 9.6: The stable intersection of a tropical quadratic curve with itself yields the vertices of the curve.

Every face of a tropical intersection $\mathcal{I}$ naturally comes with a multiplicity.

Definition 9.3.1. Each cell A in an intersection $\mathcal{I}$ can be assigned a *multiplicity* (or *weight*) as follows. Let $C = F_1 + \cdots + F_k$ be its dual cell in $P_1 + \cdots + P_k$. If A is of dimension j then C is of dimension $n - j$ and we denote its type by $(d_1, \ldots, d_k)$. For a transversal intersection define

$$\begin{aligned} m_A &:= \left(\prod_{i=1}^{k} d_i! \cdot \mathrm{vol}'_{d_i}(F_i) \right) \cdot \mathrm{vol}'_{n-j}(\mathcal{P}) \\ &= \mathrm{MV}'_{n-j}(F_1, d_1; \ldots; F_k, d_k) \end{aligned} \tag{9.2}$$

where $\mathcal{P}$ is a fundamental lattice polytope in the $(n-j)$-dimensional sublattice $\mathbb{Z}(F_1) + \cdots + \mathbb{Z}(F_k)$ and where vol'_{d_i} denotes the volume in the lattice $\mathbb{Z}(F_i)$ spanned by the integer vectors of F_i. (For more background on these relative volume forms and the proof that equality holds in (9.2) see [9].)

In the non-transversal case we have that $n - j \leq d_1 + \cdots + d_k$ and we define,

$$m_A := \sum_{\substack{(e_1,\ldots,e_k) \text{ s.t.} \\ \sum e_i = n-j;\ e_i \leq d_i}} \mathrm{MV}'_{n-j}(F_1, e_1; \ldots; F_k, e_k) \ .$$

Theorem 9.3.2 (Tropical Bernstein)**.** *Suppose the tropical hypersurfaces $X_1, \ldots, X_n \subset \mathbb{R}^n$ with Newton polytopes $P_1, \ldots, P_n$ intersect in finitely many points. Then the number of intersection points counted with multiplicity is $\mathrm{MV}_n(P_1, \ldots, P_n)$.*

Furthermore the stable intersection of n tropical hypersurfaces $X_1, \ldots, X_n$ always consists of $\mathrm{MV}_n(P_1, \ldots, P_n)$ points counted with multiplicities.

Corollary: Bezout

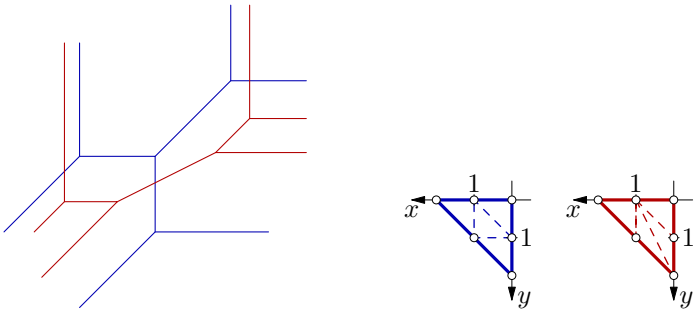


Figure 9.7: Bernstein.

Let $\mathcal{I} = X_1 \cap \cdots \cap X_k$ be a transversal intersection. Hence the intersection is proper which implies that the number of j-dimensional faces in $\mathcal{I}$ is 0 if $j \geq n - k$. By using the duality approach described in Section 9.1 the number of j-faces can be expressed in terms of mixed volumes.

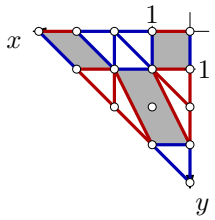


Figure 9.8: Bernstein ctd.

Figure 9.9: Non-transversal intersection of a line and a conic.

Theorem 9.3.3. *The number of j-faces in $\mathcal{I}$ counting multiplicities is*

$$\sum_{A\in\mathcal{I}^{(j)}} m_A = \sum_{\substack{(d_1,\ldots,d_k) \text{ s.t.} \\ d_i\geq 1 \text{ and } \sum_i d_i = n-j}} \mathrm{MV}'_{n-j}(P_1,d_1;\ldots;P_k,d_k)\ , \tag{9.3}$$

*where $\mathrm{MV}'_{n-j}(P_1,d_1;\ldots;P_k,d_k)$ is interpreted as in (**??**) as the sum over the lattice volume times $d_1!\cdots d_k!$ of all $(n-j)$-dimensional cells of type $(d_1,\ldots,d_k)$ in a mixed subdivision of $P_1+\cdots+P_k$.*

Note that this implies the tropical version of Bernstein's theorem (see Theorem 9.3.2) for $k=n$ and $j=0$.

Proof. Each j-dimensional cell C in the mixed subdivision of $P_1+\cdots+P_k$ is dual to an $(n-j)$-dimensional cell A of $X_1\cup\cdots\cup X_k$. If C is a mixed cell, i.e. $d_i\geq 1$ for all i, its dual A is contained in every X_i. Hence, by Definition 9.3.1

$$\sum_{A\in\mathcal{I}^{(j)}} m_A = \sum_{\substack{(d_1,\ldots,d_k) \text{ s.t.} \\ d_i\geq 1 \text{ and } \sum_i d_i = n-j}} \sum_{C=F_1+\cdots+F_k} \mathrm{MV}'_{n-j}(F_1,d_1;\ldots;F_k,d_k)$$

where the second sum runs over all cells C of type $(d_1,\ldots,d_k)$. If we denote by $\mathrm{vol}'_{d_i}(F_i)$ the volume of F_i in the lattice spanned by the integer points of F_i and furthermore denote by $\mathcal{P}$ the fundamental lattice parallelotope in $\mathbb{Z}^{n-j}$ defined by $F_1,\ldots,F_k$ then (9.2) implies

$$\begin{aligned}\mathrm{MV}'_{n-j}(F_1,d_1;\ldots;F_k,d_k) &= d_1!\cdots d_k!\ \mathrm{vol}'_{d_1}(F_1)\cdots\mathrm{vol}'_{d_k}(F_k)\ \mathrm{vol}'_{n-j}(\mathcal{P})\\ &= d_1!\cdots d_k!\ \mathrm{vol}'_{n-j}(C)\ .\end{aligned}$$

Hence we have

$$\begin{aligned}\sum_{A\in\mathcal{I}^{(j)}} m_A &= \sum_{\substack{(d_1,\ldots,d_k) \text{ s.t.} \\ d_i\geq 1 \text{ and } \sum_i d_i = n-j}} \sum_{\substack{C \text{ of type} \\ (d_1,\ldots,d_k)}} d_1!\cdots d_k!\ \mathrm{vol}'_{n-j}(C)\\ &= \sum_{\substack{(d_1,\ldots,d_k) \text{ s.t.} \\ d_i\geq 1 \text{ and } \sum_i d_i = n-j}} \mathrm{MV}'_{n-j}(P_1,d_1;\ldots;P_k,d_k)\end{aligned}$$

where we used (**??**) for the last identity. □

For $k = n-1$ and $j = 0$ we obtain a specific version for the number of vertices in tropical intersection curves

Corollary 9.3.4. *Let $\mathcal{I} = X_1 \cap \cdots \cap X_{n-1}$ be a transversal intersection curve in $\mathbb{R}^n$ of $n-1$ tropical hypersurfaces with corresponding Newton polytopes $P_1, \ldots, P_{n-1}$. Then the number of vertices in $\mathcal{I}$ counting multiplicities is*

$$\sum_{A \in \mathcal{I}^{(0)}} m_A = \mathrm{MV}_n(P_1, \ldots, P_{n-1}, P_1 + \cdots + P_{n-1}) \ . \tag{9.4}$$

Remark 9.3.5. Corollary 9.3.4 generalizes [93, Theorem 3.3] where each P_i is a standard simplex of the form $\mathrm{conv}\{s_i \cdot e^{(i)} \cup \{0\} \ : \ 1 \leq i \leq n\}$ where $e^{(i)}$ denotes the i-th unit vector and $s_i \in \mathbb{Z}_{>0}$. In this case (9.4) gives $s_1 \cdots s_{n-1} \cdot (s_1 + \cdots + s_{n-1})$ as the number of vertices counting multiplicities.

Proof. For $k = n-1$ the sum in (9.3) runs over all cells of type $(2, 1, \ldots, 1)$, $(1, 2, 1, \ldots, 1)$, $\ldots$, $(1, \ldots, 1, 2)$. Hence,

$$\sum_{A \in \mathcal{I}^{(0)}} m_A = \mathrm{MV}'_n(P_1, 2; P_2, 1; \ldots, P_{n-1}, 1) + \cdots + \mathrm{MV}'_n(P_1, 1; P_2, 1; \ldots, P_{n-1}, 2).$$

If the lattice $\sum_{i=1}^{n-1} \mathbb{Z}(P_i)$ spanned by the integer points of $P_1, \ldots, P_{n-1}$ coincides with $\mathbb{Z}^n$ then the volume forms MV_n and MV'_n coincide, and by the symmetry and linearity of the mixed volume (A.6) we get (9.4).

In the general situation, by [46] we can multiply all polytopes by a suitable factor N to obtain a lattice $\sum_{i=1}^{n-1} \mathbb{Z}(NP_i)$ which coincides with $\mathbb{Z}^n$. Then for all $l \in \mathbb{N}$, the lattice $\sum_{i=1}^{n-1} \mathbb{Z}(lNP_i)$ as well coincides with $\mathbb{Z}^n$, and hence (9.4) holds. Since then both sides of (9.4) are polynomials in l and N, the equation thus follows with regard to the lattice $\sum_{i=1}^{n-1} \mathbb{Z}(P_i)$ as well. □

We can also prove Corollary 9.3.4 independently of the dual approach by using stable intersections.

Proof. Define $\mathcal{J} := \mathcal{T}(f_1 \odot \cdots \odot f_{n-1}) = \mathcal{T}(f_1) \cup \cdots \cup \mathcal{T}(f_{n-1})$. We know that $\underbrace{\mathcal{J} \cap_{\mathrm{st}} \cdots \cap_{\mathrm{st}} \mathcal{J}}_{n\text{-times}}$ $= \mathcal{J}^{(0)}$. Since $\mathcal{I} \subset \mathcal{J}^{(1)}$ holds, this implies that $\mathcal{I} \cap_{\mathrm{st}} \mathcal{J} \subset \mathcal{J}^{(0)}$. Furthermore we have $\mathcal{I} \cap_{\mathrm{st}} \mathcal{J} \subset \mathcal{I} \cap \mathcal{J} = \mathcal{I}$ and $\mathcal{J}^{(0)} \cap \mathcal{I} = \mathcal{I}^{(0)}$ such that

$$\mathcal{I}^{(0)} = \mathcal{I} \cap_{\mathrm{st}} \mathcal{J} \ .$$

The Newton polytope of $f_1 \odot \cdots \odot f_{n-1}$ is $P_1 + \cdots + P_{n-1}$. Now using the tropical Bernstein theorem for stable intersections (Theorem 9.3.2) we have that the number of points in $\mathcal{I}^{(0)}$ counted with multiplicities is $\mathrm{MV}_n(P_1, \ldots, P_{n-1}, P_1 + \cdots + P_{n-1})$. □

Example 9.3.6. In the case of two tropical curves in the plane, the intersection multiplicity of two intersectin line segments specializes to

$$m_1 \cdot m_2 \cdot \left| \det \begin{pmatrix} v_1^{(1)} & v_1^{(2)} \\ v_2^{(1)} & v_2^{(2)} \end{pmatrix} \right| ,$$

where $v^{(1)}$ and $v^{(2)}$ are the primitive outgoing direction vectors of the segments and m_1 and m_2 are the weights of the segments.

In this planar case, the validity of Bézout's Theorem can also be seen from a nice homotopy argument. The statement clearly holds for curve where all intersection points occur among the half rays of the first curve in x-direction and the half rays of the second curve in y-direction (see Figure 9.10).

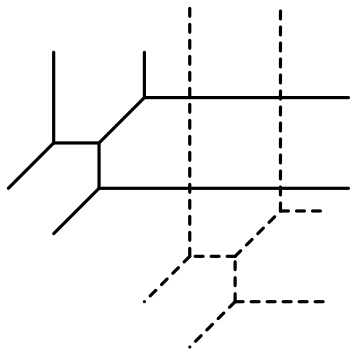


Figure 9.10: Two quadratic tropical curves in special position with points of intersection.

Using a homotopy, one can transform these curves into the given curves $C := \mathcal{T}(f_1)$ and $D := \mathcal{T}(f_2)$. Due to the equilibrium condition, whenever a vertex of one of the curves moves across a segment of the other, the sums of the intersection multiplicities remains locally constant (see Figure ...). Namely, if a vertex p of C passes over the interior of a segment S of D, the line ℓ underlying S decomposes the plane into two open half-planes. Let u be the weighted outgoing direction vector of p along ℓ, and let $v^{(1)}, \ldots, v^{(k)}$ and $w^{(1)}, \ldots, w^{(l)}$ be the weighted direction vectors of the outgoing edges of p into the two open half planes.

Just immediately before and after crossing the segment, the total intersection multiplicities at the neighborhoods of p are

$$m' \;=\; \sum_{i=1}^{k} \left| \det \begin{pmatrix} u_1 & u_2 \\ v_1^{(i)} & v_2^{(i)} \end{pmatrix} \right| \quad \text{and} \quad m'' \;=\; \sum_{j=1}^{l} \left| \det \begin{pmatrix} u_1 & u_2 \\ w_1^{(j)} & w_2^{(j)} \end{pmatrix} \right| .$$

Since within each of the two sums the determinants have the same sign, equality of m' and m'' follows immediately from the equilibrium condition at p.

In case of a non-transversal intersection, the intersection multiplicity is the (well-defined) multiplicity of any perturbation in which all intersections are transversal. The validity of Bézout's theorem then follows from the validity for the transversal case.

With similar techniques we will count now the number of unbounded faces in $\mathcal{I} = X_1 \cap \cdots \cap X_k$. Again, we formulate the result in a general manner though our main interest will later be the case $k = n-1$ and $j = 1$, i.e. the number of unbounded edges in a tropical intersection curve.

Theorem 9.3.7. *The number of unbounded j-faces in $\mathcal{I}$ is*

$$\sum_{F=(P_1)^v+\cdots+(P_k)^v} \mathrm{MV}'_{n-j}((P_1)^v,\ldots,(P_k)^v) \ . \tag{9.5}$$

Here the sum is taken over all $(n-j)$-faces F of $P := P_1 + \cdots + P_k$, $v \in \mathbb{S}^n$ is the outer unit normal vector of F and MV'_{n-j} denotes the $(n-j)$-dimensional mixed volume taken with respect to the lattice defined by the face F.

Proof. As seen in Section 9.1 the unbounded j-faces of the union $X_1 \cup \cdots \cup X_k$ correspond to $(n-j)$-dimensional cells in the boundary of $P = P_1 + \cdots + P_k$. So to count the unbounded j-faces in the intersection $\mathcal{I}$ we count mixed cells in all $(n-j)$-faces of P. Each face F of P has an outer unit normal vector v and $F = (P_1)^v + \cdots + (P_k)^v$ where $(P_i)^v$ denotes the face of P_i which is maximal with respect to v. So the number of unbounded j-faces counted with multiplicity (see Definition 9.3.1) which are dual to cells in F is $\mathrm{MV}'_{n-j}((P_1)^v,\ldots,(P_k)^v)$ and the result follows. □

Using the notion of stable intersection, we can provide the following variant of Bézout's Theorem as follows:

Corollary 9.3.8. *Given n tropical hypersurfaces of degrees $d_1,\ldots,d_n$, in n-space, the stable intersection consists of exactly $\prod_{i=1}^n$ points, counting multiplicities.*

Exercises

Exercise 9.3.9. Let $f \in \mathbb{R}_{\mathrm{trop}}[x_1,\ldots,x_n]$. Show that the unbounded cells of $\mathcal{T}(f)$ are associated to those cells of subdiv(f) which are contained in the boundary of the Newton polytope of f.

Exercise 9.3.10. Tropical conics ...

9.4 Valuations

From an algebraic point of view, tropical structures can be profitably approached via concepts from valuation theory. The goal of this section is to lay the foundations for this approach.

9.4.1 Definitions and the valuation ring

For a field K, a *real valuation* is a map $\mathrm{val} : K \to \mathbb{R}_\infty = \mathbb{R} \cup \{\infty\}$ with

1. $\mathrm{val}(x) = \infty \iff x = 0$,
2. $\mathrm{val}(xy) = \mathrm{val}(x) + \mathrm{val}(y)$ and
3. $\mathrm{val}(x + y) \geq \min\{\mathrm{val}(x), \mathrm{val}(y)\}$.

Example 9.4.1. 1.) The *trivial* valuation is given by $\mathrm{val}(0) = \infty$ and $\mathrm{val}(x) = 0$ for $x \neq 0$.
2) $K = \mathbb{Q}$ equipped with the *p-adic valuation* $v_p(\cdot)$. If q has the form

$$q \;=\; p^s \frac{m}{n},\; s \in \mathbb{Z},\, m \in \mathbb{Z}, n \in \mathbb{N}$$

where p neither divides m nor n, then the p-adic valuation of q is defined by $v_p(q) \;=\; s$.
3) If K is a field and x a single indeterminate, then

$$v_\infty\left(\frac{f}{g}\right) \;=\; \deg g - \deg f$$

defines a valuation $v_\infty : K(x) \to \mathbb{Z} \cup \{\infty\}$ on the quotient field $K(x)$.

Remark 9.4.2. Let val be a real valuation on a field K. Then for $c \in (0, 1)$ the function

$$|x| := c^{-\,\mathrm{val}(x)}$$

defines a non-archimedean norm on K, i.e. $|\cdot|$ is a map $K \to \mathbb{R}_+$ satisfying the three properties $|x| = 0 \iff x = 0$, $|xy| = |x| \cdot |y|$, and $|x+y| \leq \max\{|x|, |y|\}$ for all $x, y \in K$.

Let val be a real valuation on a field K. Then the *valuation ring* of val is defined by

$$R_{\mathrm{val}} \;:=\; \{x \in K \,:\, \mathrm{val}(x) \geq 0\}\,.$$

Theorem 9.4.3. *R_{val} is a local ring, i.e., it contains exactly one maximal ideal. The units of R_{val} are exactly the elements of K whose valuation is 0.*

Proof. Verifying that R_{val} is a ring can be done in a straightforward way. The most interesting item is the existence of an additive inverse which follows from $\mathrm{val}(a) = \mathrm{val}(-a)$, as will be shown in Exercise 9.4.9.

The set

$$M_{\mathrm{val}} \;=\; \{x \in K \,:\, \mathrm{val}(x) > 0\} \;\subset\; R_{\mathrm{val}}\,. \tag{9.6}$$

is a proper ideal in R_{val}; hence it cannot contain a unit. As a consequence, any unit of R_{val} must have valuation 0. Vice versa, for any $x \in R_{\mathrm{val}}$ with $\mathrm{val}(x) = 0$, the inverse x^{-1} in K satisfies $\mathrm{val}(x^{-1}) = \mathrm{val}(x) = 0$ by Exercise ..., which tells us that x is invertible in R_{val}. This proves the characterization of the units and also the maximality of M_{val}.

In order to show that M_{val} is a unique maximal ideal, we assume that I is another maximal ideal and the existence of an element $x \in I \setminus M_{\mathrm{val}}$. Since $\mathrm{val}(x) = 0$, I contains a unit and thus $I = R_{\mathrm{val}}$. Hence, M_{val} is a maximal ideal. □

The unique maximal ideal M_{val} from (9.6) is called the *valuation ideal* in R_{val}. The field $R_{\text{val}}/M_{\text{val}}$ is called the *residue class field of the valuation*, denoted K_{val}.

Theorem 9.4.4. *Let* val *be a real valuation on a field K. If K is algebraically closed then the residue class field K_{val} is algebraically closed as well.*

Proof. Denote by $\varphi : R_{\text{val}} \to K_{\text{val}}$, $x \mapsto x + M_{\text{val}}$ be the canonical residue class mapping.

Let $f \in K_{\text{val}}[x]$ be a non-constant polynomial and let $q \in R_{\text{val}}[x]$ be a pre-image of f under natural extension of φ to polynomials. q is non-constant, and we can choose any coefficient of q as a unit in R_{val}, since any non-zero element in K_{val} has a φ-preimage which is not contained in M_{val}.

Since $R_{\text{val}} \subset K$ and K is algebraically closed, there exists a $z \in K$ with $q(z) = 0$. We show that z is a unit in R_{val} which implies the desired statement, since then $\varphi(z) \in K_{\text{val}}$ is a zero of f.

Let q be of the form $q(x) = \sum c_i x^i$ with $c_i \in R^*_{\text{val}}$. Since $q(z) = \sum c_i z^i = 0$, by Exercise 9.4.10 there exist two indices $j \neq k$ with $\text{val}(c_j z^j) = \text{val}(c_k z^k)$. This is equivalent to $(j-k)\,\text{val}(z) = \text{val}(c_j) - \text{val}(c_k) = 0 - 0 = 0$. Since $j \neq k$, we obtain $\text{val}(z) = 0$, i.e., $z \in R^*_{\text{val}}$. □

We can extend val to a mapping $K^n \to \mathbb{R}^n$ by applying the valuation componentwise.

9.4.2 Puiseux series

A particular important example of a field with a real valuation is given by the field $\mathbb{C}\{\{t\}\}$ of *Puiseux series* with complex coefficients. These series are series of the form

$$p(t) \;=\; \sum_{k=m}^{\infty} a_k \cdot t^{k/N} \quad \text{with } m \in \mathbb{Z}, N \geq 1 \text{ and } a_k \in \mathbb{C}\,,$$

i.e., power series with complex coefficients and rational exponents whose exponents have a common denominator. We record the following classical statement which will be discussed in detail below.

Theorem 9.4.5. *$\mathbb{C}\{\{t\}\}$ is an algebraically closed field which is isomorphic to the algebraic closure of the field $\mathbb{C}((t)) = \{\sum_{k=m}^{\infty} a_k t^k : m \in \mathbb{Z}, a_k \in \mathbb{C}\}$ of formal Laurent series in t.*

There is a natural valuation on $K := \mathbb{C}\{\{t\}\}$, which generalizes the last example in Example 9.4.1. For $p(t) \neq 0$, the valuation $\text{val}\, p(t)$ is defined as the exponent of the lowest-order term of $p(t)$. For $p = 0$ set $\text{val}\, p(t) = \infty$. The image of K under this valuation map gives the set of rational numbers.

Alternatively (as made precise by Markwig [56]), it is often slightly more convenient in tropical geometry to work with a variant of Puiseux with real exponents, thus resulting in the real numbers as the image.

It is nice to observe that the standard proof of the Newton-Puiseux-Theorem 9.4.5 is constructive. Moreover, given a polynomial

$$p(x) \;=\; c_n x^n + \cdots + c_1 x + c_0$$

with coefficients $c_i = c_i(t) \in K$, it does not only compute one of the zeros of p, but all of them.

Example 9.4.6. Let $p = x^2 + x - t^2 \in K$. The zeroes of p in K are

$$\begin{aligned} x_1 &= -1 - t^2 + t^4 - 2t^6 + 5t^8 + \cdots , \\ x_2 &= t^2 - t^4 + 2t^6 - 5t^8 + \cdots . \end{aligned}$$

In order to compute these series (actually leading also to a proof of Theorem 9.4.5), we will first focus on the leading exponents and the leading coefficients of p. Let the coefficients $c_i(t)$ be of the form

$$c_i(t) \;=\; c_{i0} t^{\gamma_{i0}} + \text{ high order terms}\,.$$

We will write the desired solution for x in the form

$$x^* \;=\; b_0 t^{\mu_0} + b_1 t^{\mu_0+\mu_1} + b_2 t^{\mu_0+\mu_1+\mu_2} + \cdots$$

with $b_0 \neq 0$ and $\mu_i > 0$ for $i \geq 1$. If there exists a solution of this form then a substitution (focussing on the terms of lowest order) yields

$$\begin{aligned} 0 &= \sum_{i=0}^{n} c_{i0} t^{\gamma_{i0}} (b_0 t^{\mu_0})^i + \cdots \\ &= \sum_{i=0}^{n} c_{i0} b_0^i (t^{\gamma_{i0} + i\mu_0}) + \cdots \end{aligned}$$

A necessary condition for the right hand side to coincide with the zero Puiseux series, the (nominally) leading terms on the right hand side have to cancel. Thus, the minimum value of the sequence

$$\gamma_{00}\,,\ \gamma_{10} + \mu_0\,,\ \gamma_{20} + 2\mu_0\,,\ \ldots,\ \gamma_{n0} + n\mu_0$$

must be attained at least twice.

Example 9.4.7. For the polynomial $p = x^2 + x - t^2$ from Example 9.4.6 we obtain the condition that the minimum of

$$2\,,\ 0 + 1 \cdot \mu_0\,,\ 0 + 2 \cdot \mu_0$$

is attained at least twice. This implies

$$2 = \mu_0 \leq 2\mu_0 \ \text{ or } \ 2 = 2\mu_0 \leq \mu_0 \ \text{ or } \ \mu_0 = 2\mu_0 \leq 2$$

which yields $\mu_0 = 0$ or $\mu_0 = 2$, since the second case gives a contradiction. Thus we have determined the candidates for the leading exponent. The leading coefficient can then be determined by comparing coefficients.

It is very instructive to study the geometry of the zeroes in terms of an extended Newton polygon. For a univariate polynomial $p \in K[x]$, let the *support* and the *extended support* of p be defined by

$$\begin{aligned} \operatorname{supp}(p) &= \{i : c_i \neq 0,\, 0 \leq i \leq n\} \subset \mathbb{N}_0\,, \\ \operatorname{supp}^{\mathrm{e}}(p) &= \{(i, \operatorname{val} c_i) : c_i \neq 0,\, 0 \leq i \leq n\} \subset \mathbb{N}_0 \times \mathbb{R}\,, \end{aligned}$$

and let

$$\mathrm{NP}^{\mathrm{e}}(p) = \operatorname{conv}\{(\alpha, u) \in \operatorname{supp}(p) \times \mathbb{R} : u \geq \operatorname{val}(c_\alpha)\} \subset \mathbb{R}^2\,,$$

be the *extended Newton polygon of* p. Figure 9.11 shows the extended Newton polygon for our running example.

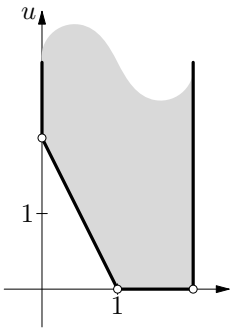


Figure 9.11: Extended Newton polygon.

Using this notation, our observation from above can be stated as follows:

Lemma 9.4.8. *The leading exponents of the zeroes of* p *(in* K*) are exactly the negative slopes of the lower edges of* $\mathrm{NP}^{\mathrm{e}}(p)$*. The lattice lengths of a lower edge gives the multiplicity, that is, how many zeroes with that leading exponent exist.*

We can assume that $c_1 \neq 0$.

By the considerations above, the first exponent μ_0 coincides with the negative slope of a lower segment of the extended Newton polygon. By considering the lowest terms in x of $p(t^{\mu_0}(x + b_0))$, we can determine the lowest coefficient b_0.

Using the values of γ_0, μ_0 and b_0, we consider

$$p_1 = t^{-\beta_0} p(t^{\mu_0}(x + b_0)) \in K[x]$$

and want to construct a zero of p_1. If we find a root x' of p_1 then $t^{\mu_0}(x' + c_0)$ will be a zero for p. Now repeat the process for p_1 to find μ_1 and c_1; with one exception: we consider only those segments of the extended Newton polygon which have negative slopes. By continuing this process, we obtain conditions for all the μ_i and c_i.

A crucial but tedious point for the proof is that in each iteration there exists an edge with negative slope and that the fractional exponents that occur have a common denominator. See, e.g., [95] for these details.

Exercises

Exercise 9.4.9. Let val be a real valuation on a field K. Show for $a \in K$, $n \in \mathbb{N}$:

1. $\mathrm{val}(1) = 0$;
2. $\mathrm{val}(a) = \mathrm{val}(-a)$;
3. $\mathrm{val}(a^{-1}) = -\mathrm{val}(a)$;
4. $\mathrm{val}(a^n) = n\,\mathrm{val}(a)$.

Exercise 9.4.10. Let val be a real valuation on a field K. Show

1. $\mathrm{val}(x_1 + \cdots + x_n) \geq \min(\mathrm{val}(x_1), \ldots, \mathrm{val}(x_n))$;
2. If there exists exactly one $k \in \{1, \ldots, n\}$ with $\mathrm{val}(x_k) = \min\{\mathrm{val}(x_1), \ldots, \mathrm{val}(x_n)\}$, then $\mathrm{val}(x_1 + \cdots + x_n) = \mathrm{val}(x_k)$.
3. If $x_1 + \cdots + x_n = 0$ for $n \geq 2$, then there exist $j \neq k \in \{1, \ldots, n\}$ with $\mathrm{val}(x_j) = \mathrm{val}(x_k) = \min\{\mathrm{val}(x_1), \ldots, \mathrm{val}(x_n)\}$.

Exercise 9.4.11. Let $K = \mathbb{Q}$ and val be the p-adic valuation. Then:

1. the valuation ring is $\{\frac{a}{b} \in \mathbb{Q}, : p \text{ does not divide } b\}$.
2. the valuation ideal is $\{\frac{a}{b} \in \mathbb{Q} : p \text{ does not divide } b \text{ and } p|a\}$.
3. the residue class field is $\mathbb{F}_p$ (the field with p elements).

Exercise 9.4.12. Let val be a real valuation on a field K and A be a nonempty subset of K. If $\mathrm{val}(\sum_{a\in A} a) > \min_{a\in A} \mathrm{val}(a)$ then the set

$$A_{\min} \;=\; \{x \in A \,:\, \mathrm{val}(x) = \min_{a\in A} \mathrm{val}(a)\}$$

consists of at least two elements.

Exercise 9.4.13. Let val be a real valuation on a field K and L an algebraic field extension of K. Show that val extends uniquely to L.

9.5 Kapranov's Theorem

For a polynomial $f = \sum_{\alpha\in\mathcal{A}} c_\alpha(t)x^\alpha \in K[x_1,\dots,x_n]$ with a finite support set $\mathcal{A} \subset \mathbb{N}_0^n$ and $c_\alpha(t) \neq 0$ for all $\alpha \in \mathcal{A}$, the *tropicalization* of f is defined by

$$\operatorname{trop} f \;=\; \bigoplus_{\alpha\in\mathcal{A}} (c_\alpha(t)) \odot x^\alpha \,.$$

Whenever there is no possibility of confusion we also write $\cdot$ instead of $\odot$.

In the following let val be a real valuation on a field K.

Theorem 9.5.1. (Kapranov.) *Let K be algebraically closed. Then for $f \in K[x]$ we have*

$$\mathcal{T}(\operatorname{trop} f) \cap \operatorname{val}(K^*)^n \;=\; \operatorname{val}\mathcal{V}(f)\,, \tag{9.7}$$

where $\mathcal{V}(f)$ denotes the zero set in $(K^*)^n$. In particular, the topological closure of the set $\operatorname{val} V(f)$ is contained in $\mathcal{T}(\operatorname{trop} f)$. If the valuation val *is surjective, then the equality*

$$\mathcal{T}(\operatorname{trop} f) \;=\; \operatorname{val}\mathcal{V}(f) \tag{9.8}$$

holds.

Before proving the statement, we provide some auxiliary lemmas:

Lemma 9.5.2. *Let $f \in K[x]$ be a univariate polynomial and $w \in \mathcal{T}(\operatorname{trop} f)$. Then there exists some $z \neq 0$ in the algebraic closure of K such that $f(z) = 0$ and $\operatorname{val}(z) = w$.*

Proof. Let $f \in K[x]$ be a univariate polynomial. By the Fundamental Theorem of Algebra, f factors into $f(x) = \prod_{i=1}^n (x - a_i)$ with $a_1,\dots,a_n \in \bar{K}$. Then, by Exercise ...,

$$\begin{aligned} \operatorname{trop} f &= \operatorname{trop}((x-a_1)\cdots(x-a_n)) \\ &= (x \oplus \operatorname{val}(a_1)) \odot \cdots \odot (x \oplus \operatorname{val}(a_n))\,. \end{aligned}$$

Hence, by Exercise ..., there exists some $i \in \{1,\dots,n\}$ with $w \in \mathcal{T}(x \oplus \operatorname{val}(a_i))$. Since $w = \operatorname{val}(a_i)$, choosing $z := a_i$ completes the proof. □

Lemma 9.5.3. *Let $f_1,\dots,f_k \in R_{\text{val}}[x_1,\dots,x_n]$, where R_{val} denotes the valuation ring. Assume that for $1 \le i \le n$ there exists a coefficient in f_i which is a unit in the valuation ring R_{val}. Then there exists a $y \in (R_{\text{val}}^*)^n$ with $f_i(y) \in R_{\text{val}}^*$.*

Proof. Denoting by $\varphi : R_{\text{val}} \to K_{\text{val}} = R_{\text{val}}/M_{\text{val}}$ the canonical residue class homomorphism, φ extends to a ring homomorphism $\varphi : R_{\text{val}}[x_1,\dots,x_n] \to K_{\text{val}}[x_1,\dots,x_n]$. By our assumptions on $f_1,\dots,f_k$, none of the polynomials $\overline{f_i} := \varphi(f_i)$ is the zero polynomial. Since K is algebraically closed, K_{val} is algebraically closed by Theorem 9.4.4. And since any algebraically closed field has infinitely many elements, we can choose an $r \in (K_{\text{val}}^*)^n$ with $\overline{f}_i(r) \neq 0$. Let $y \in R_{\text{val}}^n$ be a (componentwise) preimage of r. Since $r \in (K_{\text{val}}^*)^n$, we have $y \in (R_{\text{val}}^*)^n$. Hence

$$\varphi(f_i(y)) \;=\; \overline{f_i}(r) \;\neq\; 0\,,$$

and consequently $f_i(y) \in R_{\text{val}}^*$, $1 \le i \le k$. □

Lemma 9.5.4. *Let $f_1, \ldots, f_k \in K[x_1, \ldots, x_n]$ and $w \in \mathrm{val}(K^*)^n$. Then there exists an $a \in K^n$ with $\mathrm{val}(a) = w$ and $\mathrm{val}(f_i(a)) = (\mathrm{trop}\, f_i)(w)$ for $1 \leq i \leq k$.*

Proof. Let $y \in K^n$ with $\mathrm{val}(y) = w$. Further, for $i \in \{1, \ldots, n\}$ there exists $z_i \in K^*$ with $\mathrm{val}(z_i) = (\mathrm{trop}\, f_i)(w)$. For each i define the polynomial

$$g_i(x_1, \ldots, x_n) \;:=\; \frac{1}{z_i} f_i(y_1 x_1, \ldots, y_n x_n)\,.$$

Writing g_i in the form $g_i = \sum_\alpha c_{i\alpha} x^\alpha$, we see from the choice of z_i that $(\mathrm{trop}\, g_i)(0, \ldots, 0) = \oplus_\alpha \mathrm{val}(c_{i\alpha}) = 0$. Hence, for all α we have $\mathrm{val}(c_{i\alpha}) \geq 0$, and there exists an α with $\mathrm{val}(c_{i\alpha}) = 0$. In other words, $g_i \in R_{\mathrm{val}}[x_1, \ldots, x_n f]$, and there exists a coefficient of g_i which is unit in R_{val}. By Lemma 9.5.3, there exists an $u \in (R^*_{\mathrm{val}})^n$ with $\mathrm{val}(g_i(u)) = 0$. We show that the point $a = (y_1 u_1, \ldots, y_n u_n)$ satisfies the desired properties. The condition $\mathrm{val}(f_i(a)) = (\mathrm{trop}\, f_i)(w)$ can be verified straightforwardly. Moreover, since each u_i is a unit in the valuation ring, we observe

$$\mathrm{val}(a) \;=\; (\mathrm{val}(y_1 u_1), \ldots, \mathrm{val}(y_n u_n)) \;=\; (\mathrm{val}(y_1), \ldots, \mathrm{val}(y_n)) \;=\; w\,.$$

□

We can now prove Kapranov's Theorem:

Proof. Let $f = \sum_\alpha c_\alpha x^\alpha$ and $a \in \mathcal{V}(f)$. In order to show that $\mathrm{val}(a) \in \mathcal{T}(f_{\mathrm{trop}})$, we observe

$$\mathrm{val}(\sum_\alpha c_\alpha a^\alpha) \;=\; \mathrm{val}(0) \;=\; \infty \;>\; \min_\alpha \mathrm{val}(c_\alpha a^\alpha)\,.$$

By Exercise ..., the minimum in the right-hand expression is attained at at least two terms. Hence, $\mathrm{val}(a) \in \mathcal{T}(\mathrm{trop}\, f)$.

Conversely, let $w \in \mathcal{T}(\mathrm{trop}\, f) \cap (\mathrm{val}\, K^*)^n$. Choose $y \in (K^*)^n$ with $\mathrm{val}(y) = w$. By the definition of tropical surfaces, the minimum is attained at least twice, say at the terms with multiindices β and γ. Assume without loss of generality that $\beta_1 \neq \gamma_1$ and write f in the form

$$f(x_1, \ldots, x_n) \;=\; \sum_i h_i(x_2, \ldots, x_n) x_1^i\,.$$

By Lemma 9.5.4 we can pick a $y^* \in K^{n-1}$ with $\mathrm{val}(y^*) = (w_2, \ldots, w_n)$ and $\mathrm{val}(h_i(y^*)) = (\mathrm{trop}\, h_i)((w_2, \ldots, w_n))$ for all i. For the univariate polynomial

$$g(x_1) \;:=\; \sum_i h_i(y) x_1^i\,,$$

we have $w_1 \in \mathcal{T}(\mathrm{trop}\, g)$. By Lemma 9.5.2, there exists a $z \in K^*$ with $z = \mathrm{val}(w_1)$ and $g(z) = 0$. Hence, $g(z) = f(y^*, z) = 0$.

Since $\mathcal{T}(f)$ is topologically closed, the closure of $\mathrm{val}\, V(f)$ is contained in $\mathcal{T}(f)$. Moreover, if val is surjective then clearly (9.8) holds. □

connect to Puiseux, examples, concretely, construct polynomial

9.6 Tropicalization via dequantization

The tropical semiring results as a limit of certain *dequantizations* of the classical semiring of real positive numbers with the natural operations. For this, consider the operations

$$\begin{aligned} x \oplus_t y &= \log_t(t^x + t^y)\,, \\ x \odot y &= x + y \end{aligned}$$

for $0 < t < 1$. $(\mathbb{R}, \oplus_t, \odot)$ constitutes a semiring. Indeed, note that for $x, y, z \in \mathbb{R}$ we have the distributive law $(x \oplus_t y) \odot z = x \odot y \oplus_t x \odot z$.

In the limit case for $t \downarrow 0$, we obtain

$$x \oplus_0 y = \min\{x, y\}\,.$$

The following inequality holds for $k \in \mathbb{N}$ and $x_1, \ldots, x_k \in \mathbb{R}$:

$$\min\{x_1, \ldots, x_k\} + \underbrace{\log_t k}_{<0} \leq x_1 \oplus_t \cdots \oplus_t x_k \leq \min\{x_1, \ldots, x_k\}\,.$$

Given a polynomial $f \in \mathbb{R}[x_1, \ldots, x_n]$, let f_t be the polynomial obtained from using the operations $\oplus_t$, $\odot$. Then we consider the function $g_t : (\mathbb{R}_{>0})^n \to \mathbb{R}_{>0}$,

$$g_t(z) = t^{f_t(\mathrm{Log}_t z)}$$

Theorem 9.6.1. (Maslov [58], Viro [94].) *For any given $t \in (0,1)$, the function g_t is a polynomial function with regard to the usual arithmetic operations in $\mathbb{R}_{>0}$. Precisely, we obtain*

$$g_t(z) = \sum_j t^{c_\alpha} z^\alpha\,.$$

Proof. For $f = \sum_\alpha c_\alpha x^\alpha$, we have $f_t = \bigoplus_t c_\alpha \odot x^\alpha = \log_t(\sum_\alpha t^{c_\alpha + \sum_i \alpha_i x_i})$, and hence

$$g_t(z) = \sum_\alpha t^{c_\alpha + \sum_i \alpha_i \log_t z_i} = \sum_\alpha t^{c_\alpha} z^\alpha\,.$$

□

This statement can be seen as a special case of the *patchworking technique* for constructing real algebraic hypersurfaces introduced by Viro in 1979 (see [94]). In that technique, one considers a lattice polyhedron $\Delta \subset \mathbb{R}^n$ and a real function v on the lattice points of Δ. Projecting down the lower faces of the the corresponding upper convex hull defines a polyhedral subdivision of Δ into cells Δ_k. Denoting by $F = \sum_{j \in \Delta} a_j z^j$ a generic real polynomial with regard to the Newton polytope Δ, one considers the *truncations* $F_{\Delta_k} = \sum_{j \in \Delta_k} a_j z^j$ and the pathworking polynomials $f_t^v : \mathbb{R}^n \to \mathbb{R}$,

$$f_t^v = \sum a_j t^{v(j)} z^j$$

with a parameter $t > 1$ (which generalizes the polynomial in Theorem 9.6.1).

Let V_{Δ_k} and V_t be the varieties of F_{Δ_k} and f_t^v in $(\mathbb{C}^*)^n$, and let $\mathbb{R}V_{\Delta_k} = V_{\Delta_k} \cap (\mathbb{R}^*)^n$ and $\mathbb{R}V_t = V_t \cap (\mathbb{R}^*)^n$. Viro's patchworking theorem states that for large values of t the real hypersurface $\mathbb{R}V_t$ results from $\mathbb{R}V_{\Delta_k}$ by a certain patchworking procedure.

Exercises

9.7 Notes

The tropical Bernstein theorem can be found in [9, 75, 90].

The stable intersection was introduced in [75].

For the notion of intersection multiplicity there are various equivalent approaches, see [9, 45, 60, 91].

Tropical hypersurfaces of homogeneous polynomials naturally live in tropical projective space $\mathbb{TP}^{n-1} = \mathbb{R}^n/\mathbb{R}(1,1,\ldots,1)$.

Chapter 10

Advanced tropical geometry

In this chapter we discuss general tropical varieties. Throughout the chapter, let $K = \mathbb{C}\{\{t\}\}$ denote the field of Puiseux series with the natural valuation. We remark that many statements generalize to arbitrary nontrivial valuations.

10.1 Tropical varieties

For an ideal $I \lhd K[x_1, \ldots, x_n]$, the *tropical variety of* I is defined by the topological closure $\mathcal{T}(I) = \overline{\operatorname{val} \mathcal{V}(I)}$ where $\mathcal{V}(I)$ is the subvariety of I in $(K^*)^n$.

It is the main purpose of this section to provide several alternative characterizations of a tropical variety. The equivalences of these characterizations generalize Kapranov's Theorem 9.5.1.

We introduce the following notation for an ideal I in $K[x_1, \ldots, x_n]$. For $w \in \mathbb{R}^n$ the t-w-degree of a (Laurent) term $ct^b x^\alpha$ with $c \in \mathbb{C}^*$, $b \in \mathbb{Q}$ and $\alpha \in \mathbb{Z}^n$ is defined as $\operatorname{val}(ct^b) + w \cdot \alpha = b + w \cdot \alpha$. The *$t$-initial form* $\text{t-in}_w(f) \in \mathbb{C}[x_1, \ldots, x_n]$ of a polynomial $f \in K[x_1, \ldots, x_n]$ is the sum of all terms in f of minimal t-w-degree, evaluated at $t = 1$.

The *t-initial ideal* of I with respect to w is defined as:

$$\text{t-in}_w(I) \;:=\; \langle \text{t-in}_w(f) \,:\, f \in I \rangle \;\subset\; \mathbb{C}[x_1, \ldots, x_n]\,.$$

Example 10.1.1. Let $f = (t + t^2) * x_1^2 + (t^2 + t^3) * x_2 + t^3$. Then for $w = (1, 0)$ we have

$$\text{t-in}_w(f) \;=\; 2x_1^2 + 1\,,$$

and for the principal ideal $I = \langle f \rangle$ generated by f

$$\text{t-in}_w(I) \;=\; \langle \text{t-in}_w(f) \,:\, f \in I \rangle \;=\; \langle 2x_1^2 + 1 \rangle\,.$$

The main goal of this section to discuss the following theorem.

Theorem 10.1.2. *For an ideal $I \subset K[x_1, \ldots, x_n]$ the following subsets of $\mathbb{R}^n$ coincide:*

1. $\mathcal{T}(I)$.

2. $\bigcap_{f\in I}\mathcal{T}(f)$.

3. *The set of all vectors* $w\in\mathbb{R}^n$ *such that* $\text{t-in}_w(I)$ *does not contain a monomial.*

We first record the following auxiliary statement. For notational convenience, let $\mathcal{T}'(I)=\{w\in\mathbb{R}^n : \text{t-in}_w(I) \text{ does not contain a monomial }\}$.

Lemma 10.1.3. *Let* I *be an ideal in* $K[x_1,\dots,x_n]$ *and* $w\in\mathbb{R}^n$. *Then* $w\in\mathcal{T}'(I)$ *if and only if for all* $f\in I$, *the* t*-initial form* $\text{t-in}_w(f)$ *is not a monomial.*

Proof. The direction "$\Longrightarrow$" is clear. For the converse direction, we show that all elements in $\text{t-in}_w(I)$ have the form $\text{t-in}_w(f)$ for some $f\in I$. For this, it suffices to show that for $g,h\in I$ and $p=\sum_{\alpha\in\mathcal{A}}c_\alpha x^\alpha\in\mathbb{C}[x_1,\dots,x_n]$, the polynomials $p\cdot\text{t-in}_w(g)$ and $\text{t-in}_w(g)+\text{t-in}_w(h)$ are of the desired form. We have

$$\begin{aligned} p\cdot\text{t-in}_w(g) &= \sum_\alpha c_\alpha x^\alpha\,\text{t-in}_w(g) \;=\; \sum_\alpha \text{t-in}_w(c_\alpha x^\alpha g)\\ &= \sum_\alpha \text{t-in}_w(f_\alpha) \text{ with polynomials } f_\alpha\in I\,, \end{aligned}$$

thus reducing the case to the case of the sum. For $g,h\in I$ we have

$$\text{t-in}_w(g)+\text{t-in}_w(h) \;=\; \text{t-in}_w(g+t^{\tilde w}h) \;=\; \text{t-in}_w(f)$$

with some suitably chosen $\tilde w$ and some some polynomial $f\in I$. □

Based on this lemma, we can now prove the equivalence of the last two statements in Theorem 10.1.2.

Proof. (Equivalence of last two statements in Theorem 10.1.2). Choose a fixed polynomial $f\in I$. By Lemma 10.1.3 it suffices to show that for any $w\in\mathbb{R}^n$ the two conditions

1. $w\in\mathcal{T}(f)$,

2. the t-initial form $\text{t-in}_w(f)$ is not a monomial

are equivalent.

This equivalence follows from the observation that for a some $w\in\mathbb{R}^n$, the minimum of the tropicalization $\operatorname{trop} f$ is attained at least twice in w if and only $\text{t-in}_w(f)$ is not a monomial. □

The following lemma captures the equivalence of the first and the third statement in Theorem 10.1.2.

Lemma 10.1.4. *Let $I \subset K[x_1, \ldots, x_n]$ be an ideal. Then*

$$\mathcal{T}(I) \;=\; \mathcal{T}'(I)\,.$$

We will only prove the direction "$\subset$" as well as the zero-dimensional case. We start with some preparations.

Lemma 10.1.5. *Let $I, J \subset K[x_1, \ldots, x_n]$ be ideals. Then*

1. $\mathcal{T}'(I \cap J) \;=\; \mathcal{T}'(I) \cup \mathcal{T}'(J)$;
2. $\mathcal{T}'(I) \;=\; \mathcal{T}'(\sqrt{I})$.

Proof. In both statements, the inclusion $\supset$ is clear from $I \cap J \subset I$, $I \cap J \subset J$ and $I \subset \sqrt{I}$.

For the direction $\subset$ of the first statement, let $w \notin \mathcal{T}'(I) \cup \mathcal{T}'(J)$. Since both initial ideals $\text{t-in}_w(I)$ and $\text{t-in}_w(J)$ contain a monomial, by Lemma 10.1.3 there exist $f \in I$ and $g \in J$ such that $\text{t-in}_w(f)$ and $\text{t-in}_w(g)$ are monomials. The t-initial form of the product fg is a monomial as well, namely $\text{t-in}_\omega(f) \cdot \text{t-in}_\omega(g)$. Hence, $w \notin \mathcal{T}'(I \cap J)$.

Similarly, if the initial ideal $\text{t-in}_w(\sqrt{I})$ contains a monomial then there exists some $f \in I$ such that $\text{t-in}_w(f)$ is a monomial. Thus $\text{t-in}_w(f)^m = \text{t-in}_w(f^m) \in \text{t-in}_w(I)$, giving $w \notin \mathcal{T}'(I)$. □

Proof. (of Lemma 10.1.4). Let $p \in \mathcal{V}(I) \cap K^*$ and define $w := \text{val}(p)$. Since $\mathbb{Q}$ is dense in $\mathbb{R}$, we can assume that $\text{val}(p)$ is rational. Then for any polynomial $f \in I$, the t-initial form $\text{t-in}_w(f)$ cannot be a monomial. Hence, by Lemma 10.1.3 we have $w \in \mathcal{T}'(I)$.

In the zero-dimensional case, the converse direction can be proven as follows. Let $w \in \mathcal{T}'(I)$. Consider a minimal primary decomposition $I = \bigcap_i Q_i$ of I. By Lemma 10.1.5, we have

$$\mathcal{T}'(I) \;=\; \mathcal{T}'(\bigcap_i Q_i) \;=\; \bigcup_i \mathcal{T}'(Q_i) \;=\; \bigcup_i \mathcal{T}'(\sqrt{Q_i})\,.$$

Since $w \in \mathcal{T}'(I)$, we can assume without loss of generality that $w \in \mathcal{T}'(\sqrt{Q_1})$. The prime ideal $\sqrt{Q_1}$ of dimension 0 is maximal, and hence $\sqrt{Q_1} = \langle x_1 - p_1, \ldots, x_n - p_n \rangle$ for some $p \in K^n$. Since $w \in \mathcal{T}'(I)$, we can conclude $p \in (K^*)^n$ and $w = \text{val}\, p$ for $1 \le i \le n$. This means $w \in \mathcal{T}(I)$.

For the case of general dimension see [24]. □

Exercises: Constant coeff., lin. ideals

10.2 Projections of polyhedral complexes

We discuss here some techniques of projecting and reconstructing polyhedral complexes. These techniques will be very helpful in the next section, where we will discuss the polyhedrality of tropical varieties and tropical bases.

Throughout this section, let $\mathcal{C}$ be a pure polyhedral complex in $\mathbb{R}^n$ of dimension d (i.e., all maximal faces are of dimension d). We identify the polyhedral complex with its underlying support.

Let $\mathcal{X}$ be a finite set of affine linear subspaces in $\mathbb{R}^n$. We call $\mathcal{X}$ *complete* if for all $X, Y \in \mathcal{X}$ the intersection $X \cap Y$ is in $\mathcal{X}$. We always assume that $\mathcal{X}$ is complete (otherwise consider its completion by adding all the intersections of any subsets). The *dimension* of $\mathcal{X}$ is the maximal dimension among the subspaces in $\mathcal{X}$. The *support* of $|\,\mathcal{X}\,|$ is defined by $|\,\mathcal{X}\,| = \bigcup_{X \in \mathcal{X}} X \subset \mathbb{R}^n$.

With each polyhedral complex $\mathcal{C}$ we can associate a finite set $\mathcal{X}$ of affine linear subspaces by taking all underlying affine subspaces of the maximal cells of $\mathcal{C}$. Clearly, $\dim \mathcal{X} = \dim \mathcal{C}$.

In the following we consider *rational projections*, i.e., linear maps $\pi : \mathbb{R}^n \to \mathbb{R}^{d+1}$ described by a matrix $A = (a_{ij})$ with rational entries. π is called *geometrically regular with respect to* $\mathcal{X}$ if

1. for all $X \in \mathcal{X} : \dim(X) = \dim(\pi(X))$, and
2. for all $X, Y \in \mathcal{X} : X \subset Y \Leftrightarrow \pi(X) \subset \pi(Y)$.

Geometrically regular projections $\mathbb{R}^n \to \mathbb{R}^{d+1}$ with respect ot $\mathcal{X}$ are generic rational projections in the sense that all other projections form a non-full-dimensional subset in the space of all projections $\mathbb{R}^n \to \mathbb{R}^{d+1}$. More precisely, the two conditions above show that the set of projections which are not geometrically regular is contained in a finite union of hyperplanes within the space of all projections $\pi : \mathbb{R}^n \to \mathbb{R}^{d+1}$.

The following statement was proven by Bieri and Groves [12].

Lemma 10.2.1 (Regular projection lemma.)**.** *Let $\mathcal{X}$ be a finite set of affine subspaces of $\mathbb{R}^n$ with $d := \dim(\mathcal{X}) < n$, $\mathcal{Y}$ a subset of $\mathcal{X}$ with $\dim(\mathcal{Y}) = r \leq d$. Then there are $r+1$ projections $\pi_0, \ldots, \pi_r : \mathbb{R}^n \to \mathbb{R}^{d+1}$ with the property that for every point $x \in |\mathcal{Y}|$ there is an index $0 \leq i \leq r$ such that*

$$\pi_i^{-1}(\pi_i(x)) \cap \mathcal{X} \;=\; \{x\}\,.$$

Proof. The proof is by induction on r. If $r = -1$ then $\mathcal{Y} = \emptyset$ and there is nothing to prove. For $r \geq 0$, pick a projection $\pi : \mathbb{R}^n \to \mathbb{R}^{d+1}$ which is geometrically regular with respect to $\mathcal{Y}$. Let $T \;:=\; \{y \in |\mathcal{Y}| \,:\, \exists x \in |\mathcal{X}|, x \neq y : \pi(y) = \pi(x)\}$. Then for all subspaces $Y \in \mathcal{Y}$ we have

$$Y \cap T \;=\; \bigcup_{X \in \mathcal{X}, Y \not\subseteq X} Y \cap \pi^{-1}(\pi(X))\,.$$

Since π is regular, $\pi(Y) \not\subseteq \pi(X)$ for all $X \not\supseteq Y$. Therefore $\dim(Y \cap \pi^{-1}(\pi(X))) \;<\; \dim Y$. Let $\mathcal{Z}$ be the set of all such intersections,

$$\mathcal{Z} \;:=\; \{Y \cap \pi^{-1}(\pi(X)) \mid Y \in \mathcal{Y}, X \in \mathcal{X}, Y \not\subseteq X\}\,.$$

$\mathcal{Z}$ is a finite set of affine subspaces of $\mathbb{R}^n$ with $\dim(\mathcal{Z}) < r$. Let $\mathcal{Z}'$ be the smallest complete finite set of subspaces containing $\mathcal{X}$ and $\mathcal{Z}$. Then the induction hypothesis applied for $\mathcal{Z}$ and $\mathcal{Z}'$ yields projections $\pi_0, \dots \pi_{r-1}$ such that for every point $z \in \mathcal{Z}$ there is an index $0 \leq i \leq r-1$ with $\pi_i^{-1}(\pi_i(z)) \cap \mathcal{X} = \{z\}$. This holds in particular for all points $z \in T \subset \mathcal{Z}'$. But for all other points of $y \in \mathcal{Y}$ the projection $\pi := \pi_r$ satisfies $\pi_r^{-1}(\pi_r(y)) \cap \mathcal{X} = \{y\}$. This proves the assertion. □

So if to each polyhedral complex $\mathcal{C}$ a finite set of affine subspaces of the same dimension is assigned, then all points x of $\mathcal{C}$ can be reconstructed by an appropriate projection. Moreover, the following lemma states that also the affine subspaces can be obtained as an intersection of preimages of several projections.

Lemma 10.2.2. *Let $\mathcal{X}$ be a complete and finite set of affine subspaces pure of dimension d. Then there are $n-d$ projections $\pi_1, \dots, \pi_{n-d}$ such that $\bigcap_{i=1}^{n-d} \pi^{-1}(\pi(\mathcal{X}))$ is pure d-dimensional.*

Proof. We proceed by induction. Let $\mathcal{X}_0 = \{\mathbb{R}^n\}$, and let π_{n-d} be an arbitrary projection with d-dimensional image $\pi_{n-d}(\mathcal{X})$. Then define

$$\mathcal{X}_1 \;=\; \{\text{affine subspaces in the preimage } \pi_{n-d}^{-1}\pi_{n-d}(\mathcal{X})\}\,.$$

We observe that all maximal affine subspaces of $\mathcal{X}_1$ have dimension $n-1$.

Assume now $\mathcal{X}_t$ is constructed and is $(n-t)$-dimensional with $\mathcal{X}_t = \bigcap_{i=n-d-t+1}^{n-d} \pi_i^{-1}\pi_i(\mathcal{X})$. Let $\mathcal{B}$ be the finite set of all $(n-t)$-dimensional subspaces of $\mathbb{R}^n$ parallel to at least one of the affine subspaces of $\mathcal{X}_t$. Then choose $\pi_{n-d-t} : \mathbb{R}^n \to \mathbb{R}^{d+1}$ as a projection with $\mathbb{R}^n = \ker \pi_{n-d-t} + V$ for all $V \in \mathcal{B}$. Let $\mathcal{Y}$ be the set of all $(n-1)$-dimensional affine subspaces of $\pi_{n-d-t}^{-1}\pi_{n-d-t}(\mathcal{X})$ and let

$$\mathcal{X}_{t+1} \;=\; \{Y \cap X \mid Y \in \mathcal{Y}, X \in \mathcal{X}_t\}\,.$$

Now $\mathcal{X}_{t+1}$ is pure of dimension $n-t-1$. Altogether, it follows that $\pi_1, \dots, \pi_{n-d}$ are the projections we are searching for. □

Corollary 10.2.3. *Let $\mathcal{C}$ be a pure d-dimensional polyhedral complex in $\mathbb{R}^n$. Then there are $n+1$ projections $\pi_0, \dots, \pi_n : \mathbb{R}^n \to \mathbb{R}^{d+1}$ such that*

$$\mathcal{C} \;=\; \bigcap_{i=0}^{n} \pi_i^{-1}(\pi_i(\mathcal{C}))\,.$$

Proof. Clearly, $\bigcap_{i=1}^{n-d} \pi^{-1}(\pi(\mathcal{X}))$ contains $\mathcal{X}$. Hence, the statement follows from combining Lemmas 10.2.1 and 10.2.2. □

Picture ...

10.3 Tropical bases

Bieri-Groves ...

zero tension / total concavity ...

We show that every ideal $I \lhd K[x_1, \ldots, x_n]$ has a tropical basis ...

In particular, this shows that every tropical variety is a polyhedral complex.[1]

Let $I \lhd K[x_1, \ldots, x_n]$ be an m-dimensional prime ideal. The main geometric idea is to consider $n - m + 1$ different (rational) projections $\pi_0, \ldots, \pi_{n-m} : \mathbb{R}^n \to \mathbb{R}^{m+1}$. If these projections are sufficiently generic (as specified below) then we obtain

$$\bigcap_{i=0}^{n-m} \pi_i^{-1}(\pi_i(\mathcal{T}(I))) \;=\; \mathcal{T}(I)\,,$$

and each of the sets $\pi_i^{-1}(\pi_i(\mathcal{T}(I)))$ is a tropical hypersurface.

First we consider the image of the tropical variety $\mathcal{T}(I)$ under a single (rational) projection

$$\begin{aligned} \pi \,:\, \mathbb{R}^n &\to \mathbb{R}^{m+1}\,, \\ x &\mapsto Ax \end{aligned}$$

with a non-singular rational matrix A whose rows are denoted by $a^{(1)}, \ldots, a^{(m+1)}$. Let $u^{(1)}, \ldots, u^{(l)} \in \mathbb{Z}^n$ with $l := n - (m+1)$ be a basis of the orthogonal complement of $\operatorname{span}\{a^{(1)}, \ldots, a^{(m+1)}\}$.

Set $R = K[x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_l]$, and for any polynomial $f \in K[x_1, \ldots, x_n]$ let $\hat{f}$ be the composition of f with the monomial map $x_i \mapsto x_i \prod_{j=1}^{l} \lambda_j{}^{u_i^{(j)}}$, i.e.,

$$\hat{f}(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_l) \;=\; f(x_1 \prod_{j=1}^{l} \lambda_j{}^{u_1^{(j)}}, \ldots, x_n \prod_{j=1}^{l} \lambda_j{}^{u_n^{(j)}}) \;\in\; R\,.$$

Define the ideal $J \lhd R$ by

$$J \;=\; \left\langle \hat{f} \in R \,:\, f \in I \right\rangle.$$

We show the following characterization of $\pi^{-1}(\pi(\mathcal{T}(I)))$ in terms of elimination.

Theorem 10.3.1. *Let $I \lhd K[x_1, \ldots, x_n]$ be an m-dimensional prime ideal and $\pi : \mathbb{R}^n \to \mathbb{R}^{m+1}$ be a rational projection. Then $\pi^{-1}(\pi(\mathcal{T}(I)))$ is a tropical variety with*

$$\pi^{-1}(\pi(\mathcal{T}(I))) \;=\; \mathcal{T}(J \cap K[x_1, \ldots, x_n])\,. \tag{10.1}$$

In order to prove Theorem 10.3.1, we first consider *algebraically regular* projections (as defined below). At the end of this section we also cover the remaining special cases.

We start with an auxiliary statement which holds for an arbitrary rational projection π.

[1] intersection condition ...

Lemma 10.3.2. *For any $w \in \mathcal{T}(J \cap K[x_1, \ldots, x_n])$ and $u \in \operatorname{span}\{u^{(1)}, \ldots, u^{(l)}\}$ we have $w + u \in \mathcal{T}(J \cap K[x_1, \ldots, x_n])$.*

Proof. Let $u = \sum_{i=1}^{l} \mu_j u^{(j)}$ with $\mu_1, \ldots, \mu_l \in \mathbb{Q}$. The case of real μ_i then follows as well.

Let $w \in \mathcal{T}(J \cap K[x_1, \ldots, x_n])$. Since $\mathcal{T}(J \cap K[x_1, \ldots, x_n])$ is closed, we can assume without loss of generality that there exists $z \in \mathcal{V}(J \cap K[x_1, \ldots, x_n])$ with $\operatorname{ord} z = w$. Define $y = (y', y'') \in (\bar{K}^*)^{n+l}$ by

$$y \;=\; (y', y'') \;=\; \left(z_1 t^{\sum_{j=1}^{l} \mu_j u_1^{(j)}}, \ldots, z_n t^{\sum_{j=1}^{l} \mu_j u_n^{(j)}}, t^{-\mu_1}, \ldots, t^{-\mu_l} \right).$$

For any $f \in I$, the point y is a zero of the polynomial $\hat{f}$ in the ring R, and thus $y \in \mathcal{V}(J)$. Hence, $y' \in \mathcal{V}(J \cap K[x_1, \ldots, x_n])$. Moreover,

$$\operatorname{ord} y' \;=\; (w_1 + \sum_{j=1}^{l} \mu_j u_1^{(j)}, \ldots, w_n + \sum_{j=1}^{l} \mu_j u_n^{(j)}) \;=\; w + \sum_{j=1}^{l} \mu_j u^{(j)} \;=\; w + u\,,$$

which proves our claim. □

Lemma 10.3.3. *Let $I \lhd K[x_1, \ldots, x_n]$ be an ideal. Then $J \cap K[x_1, \ldots, x_n] \subset I$.*

Proof. Let $p = \sum_i h_i \hat{f}_i$ be a polynomial in $J \cap K[x_1, \ldots, x_n]$ with $f_i \in I$. Since p is independent of $\lambda_1, \ldots, \lambda_l$ we have

$$p \;=\; p|_{\lambda_1 = 1, \ldots, \lambda_l = 1} \;=\; \sum_i h_i|_{\lambda_1 = 1, \ldots, \lambda_l = 1} f_i \;\in\; I.$$

□

We call a rational projection *algebraically regular* for I if for each $i \in \{1, \ldots, l\}$ the elimination ideal $J \cap K[x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_i]$ has a finite basis $\mathcal{F}_i$ such that in every polynomial $f \in \mathcal{F}_i$ the coefficients of the powers of λ_i (when considering f as a polynomial in λ_i) are monomials in $x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_{i-1}$.

The following statement shows that the set of algebraically regular projections is dense in the set of all real projections $\pi : \mathbb{R}^n \to \mathbb{R}^{m+1}$.

Lemma 10.3.4. *The set of projections which are not algebraically regular is contained in a finite union of hyperplanes within the space of all projections $\pi : \mathbb{R}^n \to \mathbb{R}^{m+1}$*

Proof. It suffices to show that for the choice of $u^{(l)}$, we just have to avoid a lower-dimensional subset of $\mathbb{R}^n \setminus \{0\}$. For $u^{(1)}, \ldots, u^{(l-1)}$ we can then argue inductively (however, an explicit description then becomes more involved). Assume that I is generated by $f_1, \ldots, f_s$. Then $J = \langle \hat{f}_j \,:\, 1 \le j \le s \rangle$. For any fixed j, the polynomial $\hat{f}_j$ is of the form

$$\hat{f}_j \;=\; \sum_{\alpha \in \mathcal{A}_j} c_\alpha x^\alpha \lambda_1^{\sum \alpha_i u_i^{(1)}} \cdots \lambda_l^{\sum \alpha_i u_i^{(l)}}$$

with $\mathcal{A}_j \subset \mathbb{Z}^n$ finite. Thus all λ_l^k have monomial coefficients if

$$\sum \alpha_i u_i^{(l)} \;\neq\; \sum \beta_i u_i^{(l)}$$

for all $\alpha, \beta \in \mathcal{A}_j$ with $\alpha \neq \beta$. So we have to choose $u^{(l)}$ from the subset

$$\bigcap_j \{u \in \mathbb{R}^n \,:\, \sum \alpha_i u_i^{(l)} \neq \sum \beta_i u_i^{(l)} \text{ for all } \alpha, \beta \in \mathcal{A}_j \text{ with } \alpha \neq \beta\}\,.$$

Hence, the algebraically non-regular projections are contained in a finite number of hyperplanes. □

Theorem 10.3.5. *Let $I \lhd K[x_1, \ldots, x_n]$ be a prime ideal and $\pi : \mathbb{R}^n \to \mathbb{R}^{m+1}$ be an algebraically regular projection. Then $\pi^{-1}\pi(\mathcal{T}(I))$ is a tropical variety with*

$$\pi^{-1}\pi(\mathcal{T}(I)) \;=\; \mathcal{T}(J \cap K[x_1, \ldots, x_n])\,. \tag{10.2}$$

Proof. Let $w \in \pi^{-1}\pi(\mathcal{T}(I))$. Since the right hand set of (10.2) is closed, we can assume without loss of generality that there exists $z' \in \mathcal{V}(I)$ and $u \in \operatorname{span}\{u^{(1)}, \ldots, u^{(l)}\}$ with $\operatorname{ord} z' = w + u$. For any $f \in I$, the point

$$z \;:=\; (z', 1)$$

is a zero of the polynomial $\hat{f} \in R$, and thus $z \in \mathcal{V}(J)$. Hence, $z' \in \mathcal{V}(J \cap K[x_1, \ldots, x_n])$. By Lemma 10.3.2, $w \in \mathcal{T}(J \cap K[x_1, \ldots, x_n])$ as well.

Let now $w \in \mathcal{T}(J \cap K[x_1, \ldots, x_n])$. Again we can assume that there is a $z \in \mathcal{V}(J \cap K[x_1, \ldots, x_n]) \subset (\bar{K}^*)^n$ with $w = \operatorname{ord}(z)$. The projection is algebraically regular which means that the generators of the elimination ideals $J \cap K[x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_i]$ have only monomials as coefficients with respect to λ_i. By the Extension Theorem (see, e.g., [20]), we can extend the root z inductively to a root $\tilde{z} \in \mathcal{V}(J)$ with the same first n entries. The definition of J says that

$$z' \;:=\; (z_1 \tilde{z}_{n+1}^{u_1^{(1)}} \cdots \tilde{z}_{n+l}^{u_1^{(l)}}, \ldots, z_n \tilde{z}_{n+1}^{u_n^{(1)}} \cdots \tilde{z}_{n+l}^{u_n^{(l)}})$$

is a root of I. Then

$$\operatorname{ord}(z') \;=\; \operatorname{ord}(z) + \sum_{i=1}^{l} \operatorname{ord}(\tilde{z}_{n+i}) u^{(i)}$$

which means that $\operatorname{ord}(z) = w \in \pi^{-1}\pi(\mathcal{T}(I))$. □

This completes the proof of Theorem 10.3.1 for the case of algebraically regular projections.

In the following, we consider the notion of *geometric regularity*.

Definition 10.3.6. Let $\mathcal{C}$ be a polyhedral complex in $\mathbb{R}^n$. A projection $\pi : \mathbb{R}^n \to \mathbb{R}^{m+1}$ is called *geometrically regular* if the following two conditions hold.

1. For any k-face σ of $\mathcal{C}$ we have $\dim(\pi(\sigma)) = k$, $0 \leq k \leq \dim \mathcal{C}$.

2. If $\pi(\sigma) \subset \pi(\tau)$ then $\sigma \subset \tau$ for all $\sigma, \tau \in \mathcal{C}$.

These conditions ensure that we can recover the whole complex $\mathcal{C}$ from the projections.

Corollary 10.3.7. *In the situation of Theorem 10.3.5, if* $\dim \pi(\mathcal{T}(I)) = m$ *then* $\pi^{-1}\pi(\mathcal{T}(I))$ *is a tropical hypersurface.*

In particular, this holds when the projection is geometrically regular.

Proof. $\dim \pi^{-1}\pi(\mathcal{T}(I)) = \dim \pi(\mathcal{T}(I)) + \dim \ker \pi = m + (n - (m+1)) = n - 1$. □

Let $I \lhd K[x_1, \ldots, x_n]$ be a prime ideal and $m = \dim I$. Then $\mathcal{T}(I)$ is a pure m-dimensional polyhedral complex. Bieri and Groves [12] used the following geometric technique (which actually was also used to prove that $\mathcal{T}(I)$ has this polyhedral property).

There exists a finite family $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_s\}$ of m-dimensional affine subspaces with $\mathcal{T}(I) \subset \bigcup_{i=1}^{s} \mathcal{X}_s$. By the finiteness of $\mathcal{X}$, for a sufficiently generic choice of $n - m + 1$ geometrically regular projections $\pi_0, \ldots, \pi_{n-m}$ the set-theoretic intersection of the inverse projections exactly yields the original polyhedral complex. This follows from [12, Thm. 4.4] (and its proof) in connection with the pure-dimensionality of $\mathcal{T}(I)$.

Proposition 10.3.8 (Bieri, Groves [12])**.** *Let* $I \lhd K[x_1, \ldots, x_n]$ *be a prime ideal. For any dense set* $\mathcal{D}$ *of projections there exist* $\operatorname{codim} I + 1$ *projections* $\pi_0, \ldots, \pi_{\operatorname{codim} I} \in \mathcal{D}$ *such that*

$$\mathcal{T}(I) \;=\; \bigcap_{i=0}^{\operatorname{codim} I} \pi_i^{-1}\pi_i(\mathcal{T}(I))\,.$$

By Lemma 10.3.4, the set of algebraically regular projections is dense in the space of projections. Hence, combining Proposition 10.3.8 with Theorem 10.3.5 yields Theorem **??**. Note that by Lemma 10.3.3 the generators g_i are actually contained in I.

Using this knowledge about the existence of some tropical basis, we can also provide the proof of Theorem 10.3.1 for arbitrary rational projections.

Theorem 10.3.9 (Tropical Extension Theorem)**.** *Let* $I \lhd K[x_0, \ldots, x_n]$ *be an ideal and* $I_1 = I \cap K[x_1, \ldots, x_n]$ *be its first elimination ideal. For any* $w \in \mathcal{T}(I_1)$ *there exists a point* $\tilde{w} = (w_0, \ldots, w_n) \in \mathbb{R}^{n+1}$ *with* $w_i = \tilde{w}_i$ *for* $1 \leq i \leq n$ *and* $\tilde{w} \in \mathcal{T}(I)$*.*

Proof. First let $w \in \operatorname{ord}(\mathcal{V}(I_1))$, so that there exists $z \in \mathcal{V}(I_1)$ with $\operatorname{ord}(z) = w$. Let $\mathcal{G} = \{g_1, \ldots, g_s\}$ be a reduced Gröbner basis of I with respect to a lexicographical term order with $x_0 > x_i$, $1 \leq i \leq n$. I.e.,

$$g_i \;=\; h_i(x_1, \ldots, x_n) x_0^{\deg_{x_0} g_i} \;+\; \text{terms of lower degree in } x_0\,.$$

There are two cases to consider:

Case 1: $z \notin \mathcal{V}(h_1, \ldots, h_s)$. Then by the classical Extension Theorem there is a root $\tilde{z}$ of I which extends z, so $\operatorname{ord}(\tilde{z}) =: \tilde{w}$ extends w.

Case 2: $z \in \mathcal{V}(h_1, \ldots, h_s)$. Then $w = \mathrm{ord}(z) \in \mathcal{T}(h_1, \ldots, h_s)$. Let $\mathcal{P} = \{p_1, \ldots, p_t\}$ be a tropical basis of I.

Let p_j be any of these polynomials. p_j has the form

$$p_j \;=\; q_j(x_1, \ldots, x_n) x_0^{\deg_{x_0} p_j} \;+\; \text{terms of lower degree in } x_0\,.$$

Since $\mathcal{G}$ is a lexicographic Gröbner basis, we have $q_j(x_1, \ldots, x_n) =: \sum k_\alpha x^\alpha \in \langle h_1, \ldots, h_s \rangle$. Hence, the minimum

$$\min_\alpha \{\mathrm{ord}(k_\alpha) + \alpha_1 x_1 + \cdots + \alpha_n x_n\}$$

is attained twice at w. We can pick a sufficiently small value $w_0^{(j)} \in \mathbb{R}$ so that all terms $x_1^{m_1} \cdots x_n^{m_n} x_0^{m_0}$ of p_j with $m_0 < \deg_{x_0} p_j$ have a larger value $m_1 w_1 + \cdots + m_n w_n + m_0 w_0^{(j)}$. But then the minimum of all values of all terms of p_j is attained at least twice; it is

$$\min_\alpha \{\mathrm{ord}(k_\alpha) + \alpha_1 w_1 + \cdots + \alpha_n w_n\} + \deg_{x_0} p_j \cdot w_0^{(j)}\,.$$

So $(w_0^{(j)}, w_1, \ldots, w_n) \in \mathcal{T}(h_j)$.

By setting $w_0 = \min_j \{w_0^{(j)}\}$ and $\tilde{w} := (w_0, \ldots, w_n) \in \mathcal{T}(I)$, we obtain the desired extension of w. This completes case 2.

Let now $w = \lim_{i \to \infty} w^{(i)}$ be in the closure of $\mathrm{ord}(\mathcal{V}(I_1))$. Then there exist $\tilde{w}^{(i)} \in \mathcal{T}(I)$ with $\tilde{w}_j^{(i)} = w_j^{(i)}$ for $1 \le j \le n$. Let $\mathcal{P} = \{p_1, \ldots, p_t\}$ be again a tropical basis of I. Then we can assume w.l.og. that the minimum in $\mathrm{trop}(p_k)$, $1 \le k \le t$ for $\tilde{w}^{(i)}$ is attained at the same terms. This gives us conditions for the $\tilde{w}_0^{(i)}$:

$$k^{(i)} \le \tilde{w}_0^{(i)} \le l^{(i)} \quad (\text{one of them can be } \pm\infty)\,.$$

These bounds vary continuously with $w^{(i)}$. So we can choose $\tilde{w}_0$ arbitrarily in $[\lim k^{(i)}, \lim l^{(i)}]$ (only one of the limits can be $\pm\infty$). □

10.4 Tropical linear spaces

We remark that there are linear tropical spaces of dimension $n-2$ which are not complete intersections, i.e., which are not the intersection of two tropical hypersurfaces (see Proposition 6.3 in Speyer's and Sturmfels' paper on the tropical Grassmannian).

Definition 10.4.1. A *tropical linear space* is a subset of tropical projective space $\mathbb{TP}^{n-1}$ of the form $\mathcal{T}(I)$ where the ideal I is generated by linear forms

$$p_1(t) \cdot x_1 \;+\; p_2(t) \cdot x_2 \;+\; \cdots \;+\; p_n(t) \cdot x_n$$

with coefficients $p_i(t) \in K$.

Example 10.4.2. A *line in three-space* is the tropical variety $\mathcal{T}(I)$ of an ideal I which is generated by a two-dimensional space of linear forms in $K[x_1, x_2, x_3, x_4]$. A tropical basis of such an ideal I consists of four linear forms,

$$\begin{aligned} U \quad = \quad \{ \quad & p_{12}(t) \cdot x_2 \,+\, p_{13}(t) \cdot x_3 \,+\, p_{14}(t) \cdot x_4, \\ & -p_{12}(t) \cdot x_1 \,+\, p_{23}(t) \cdot x_3 \,+\, p_{24}(t) \cdot x_4, \\ & -p_{13}(t) \cdot x_1 \,-\, p_{23}(t) \cdot x_2 \,+\, p_{34}(t) \cdot x_4, \\ & -p_{14}(t) \cdot x_1 \,-\, p_{24}(t) \cdot x_2 \,-\, p_{34}(t) \cdot x_3 \,\}, \end{aligned}$$

where the coefficients of the linear forms satisfy the *Grassmann-Plücker relation*

$$p_{12}(t) \cdot p_{34}(t) \,-\, p_{13}(t) \cdot p_{24}(t) \,+\, p_{14}(t) \cdot p_{23}(t) \;=\; 0. \tag{10.3}$$

We abbreviate $a_{ij} = \text{order}\big(p_{ij}(t)\big)$. The line $\mathcal{T}(I)$ is the set of all points $w \in \mathbb{TP}^3$ which satisfy a Boolean combination of linear inequalities:

$$\begin{aligned} & \big(\, a_{12} + x_2 = a_{13} + x_3 \leq a_{14} + x_4 \quad \text{or} \\ & \quad a_{12} + x_2 = a_{14} + x_4 \leq a_{13} + x_3 \;\text{ or }\; a_{13} + x_3 = a_{14} + x_4 \leq a_{12} + x_2 \, \big) \\ \text{and} \quad & \big(\, a_{12} + x_1 = a_{23} + x_3 \leq a_{24} + x_4 \quad \text{or} \\ & \quad a_{12} + x_1 = a_{24} + x_4 \leq a_{23} + x_3 \;\text{ or }\; a_{23} + x_3 = a_{24} + x_4 \leq a_{12} + x_1 \, \big) \\ \text{and} \quad & \big(\, a_{13} + x_1 = a_{23} + x_2 \leq a_{34} + x_4 \quad \text{or} \\ & \quad a_{13} + x_1 = a_{34} + x_4 \leq a_{23} + x_2 \;\text{ or }\; a_{23} + x_2 = a_{34} + x_4 \leq a_{13} + x_1 \, \big) \\ \text{and} \quad & \big(\, a_{14} + x_1 = a_{24} + x_2 \leq a_{34} + x_3 \quad \text{or} \\ & \quad a_{14} + x_1 = a_{34} + x_3 \leq a_{24} + x_2 \;\text{ or }\; a_{24} + x_2 = a_{34} + x_3 \leq a_{14} + x_1 \, \big). \end{aligned}$$

To resolve this Boolean combination, one distinguishes three cases arising from (10.3):

$$\begin{aligned} \text{Case } [12,34]: \quad & a_{14} + a_{23} \;=\; a_{13} + a_{24} \;\leq\; a_{12} + a_{34}, \\ \text{Case } [13,24]: \quad & a_{14} + a_{23} \;=\; a_{12} + a_{34} \;\leq\; a_{13} + a_{24}, \\ \text{Case } [14,23]: \quad & a_{13} + a_{24} \;=\; a_{12} + a_{34} \;\leq\; a_{14} + a_{23}. \end{aligned}$$

In each case, the line $\mathcal{T}(I)$ consists of a line segment, with two of the four coordinate rays emanating from each end point. The two end points of the line segment are

$$\begin{aligned} \text{Case } [12,34]: \quad & (a_{23} + a_{34},\, a_{13} + a_{34},\, a_{14} + a_{23},\, a_{13} + a_{23}) \quad \text{and} \\ & (a_{13} + a_{24},\, a_{13} + a_{14},\, a_{12} + a_{14},\, a_{12} + a_{13}), \\ \text{Case } [13,24]: \quad & (a_{24} + a_{34},\, a_{14} + a_{34},\, a_{14} + a_{24},\, a_{12} + a_{34}) \quad \text{and} \\ & (a_{23} + a_{34},\, a_{13} + a_{34},\, a_{12} + a_{34},\, a_{13} + a_{23}), \\ \text{Case } [14,23]: \quad & (a_{23} + a_{34},\, a_{13} + a_{34},\, a_{12} + a_{34},\, a_{13} + a_{23}) \quad \text{and} \\ & (a_{24} + a_{34},\, a_{14} + a_{34},\, a_{14} + a_{24},\, a_{12} + a_{34}). \end{aligned}$$

The three types of lines in $\mathbb{TP}^3$ are depicted in Figure 10.1.

....

Figure 10.1: The three types of tropical lines in $\mathbb{TP}^3$.

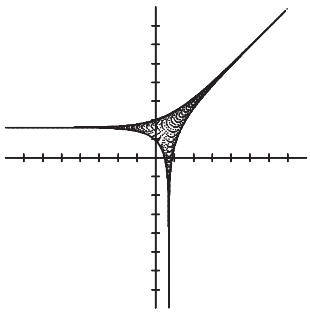


Figure 10.2: Amoeba Log $\mathcal{V}(f)$ for $f(z_1, z_2) = \frac{1}{2}z_1 + \frac{1}{5}z_2 - 1$

10.5 Counting curves

[2]

Discuss history of Kapranov for ideals ...

10.6 Amoebas, tropical geometry and deformations

10.6.1 Introduction

We consider algebraic varieties from the following viewpoint of amoebas.

Definition 10.6.1. For a polynomial $f \in \mathbb{C}[X_1, \ldots, X_n]$ the image set of its variety $\mathcal{V}(f) \subset (\mathbb{C}^*)^n$ under the map

$$\begin{array}{rcl} \text{Log} \ : (\mathbb{C}^*)^n & \to & \mathbb{R}^n \,, \\ z = (z_1, \ldots, z_n) & \mapsto & (\log|z_1|, \ldots, \log|z_n|) \end{array}$$

is called the *amoeba* of f, denoted $\mathcal{A}_f$.

In order to keep the setup simple, we often concentrate on the case of plane curves, i.e., $f \in \mathbb{C}[X_1, X_2]$.

Example 10.6.2. (a) The shaded area in Figure 10.2 shows the amoeba Log $\mathcal{V}(f)$ for the linear function

$$f(z_1, z_2) = \frac{1}{2}z_1 + \frac{1}{5}z_2 - 1 \,.$$

Note that this amoeba is a two-dimensional set. When denoting the coordinates in the amoeba plane by w_1 and w_2, the three tentacles have the asympotics $w_1 = \log 2$, $w_2 =$

[2] Throughout the tropical chapter, include examples, connections and applications

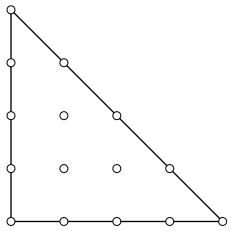


Figure 10.3: Newton polygon of a dense quartic in two variables

$\log 5$, and $w_2 = w_1 + \log(5/2)$. We remark that the amoeba of a two-dimensional variety $\mathcal{V}(f) \in (\mathbb{C}^*)^2$ is not always a two-dimensional set. Namely, e.g., for $f(z_1, z_2) := z_1 + z_2$, we obtain $\text{Log } \mathcal{V}(f) = \{(w_1, w_2) \in \mathbb{R}^2 \,:\, w_1 = w_2\}$.

(b) If $f \in \mathbb{C}[X_1^{\pm 1}, \ldots, X_n^{\pm 1}]$ is a binomial in n variables,

$$f(z) = z^\alpha - z^\beta$$

with $\alpha \neq \beta \in \mathbb{Z}^n$, then the amoeba $\text{Log } \mathcal{V}(f)$ is a hyperplane in $\mathbb{R}^n$ which passes through the origin. To see this, first note that for any complex solution z of $z^\alpha = z^\beta$, the real vector $|z| = (|z_1|, \ldots, |z_n|)$ is a solution as well. So it suffices to consider vectors $z \in (0, \infty)^n$. We can rewrite $|z|^\alpha = |z|^\beta$ as $|z|^{\alpha-\beta} = 1$, and by using the dot product of vectors we obtain

$$(\alpha - \beta) \cdot \text{Log } z = 0\,.$$

Since $\alpha \neq \beta$, this equation defines a hyperplane in the coordinates $\log|z_1|, \ldots, \log|z_n|$ which passes through the origin.

The following properties are the reason why it is often convenient to look at $\log|z_i|$ rather than $|z_i|$ itself.

Theorem 10.6.3. *The complement of a hypersurface amoeba* $\text{Log } \mathcal{V}(f)$ *consists of finitely many convex regions, and these regions are in bijective correspondence with the different Laurent expansions of the rational function* $1/f$.

The shape of the amoeba is also related to the support

$$\text{supp}(f) = \{\alpha \in \mathbb{Z}^n \,:\, c_\alpha \neq 0\}$$

of the function f and to the Newton polytope

$$\text{New}(f) = \text{conv}(\text{supp}(f))\,.$$

Example 10.6.4. Figure 10.3 shows the Newton polygon of a dense quartic polynomial f in two variables. Figure 10.4 depicts a series of amoebas Log $\mathcal{V}(f)$ for dense quartic polynomials $f \in \mathbb{R}[X_1, X_2]$. In the first picture in this series, f is the product of four linear functions f_1, f_2, f_3, f_4. The amoeba of $\mathcal{V}(f)$ is the union of the amoebas of $\mathcal{V}(f_1)$, $\mathcal{V}(f_2)$, $\mathcal{V}(f_3)$, and $\mathcal{V}(f_4)$. This polynomial f is perturbed by adding or subtracting to every coefficient c_α of f (with the exception of the coefficient corresponding to the constant term) independently a random value in the interval $[0, \frac{1}{5}|c_\alpha|)$; see the right picture in the top row. This perturbation process is then iterated another four times.

10.6.2 Background from complex analysis

A central theme here is that we are looking for convexity and linearity within algebraic varieties.

Suppose $f \in \mathbb{C}[X]$ is a univariate polynomial with zeroes $a_1, \ldots, a_k$ satisfying $|a_1| \leq \cdots \leq |a_k|$, and assume $f(0) \neq 0$. Then Jensen's formula (for entire functions, i.e., holomorphic functions with a countable number of solutions) implies

$$\frac{1}{2\pi i}\int_{|z|=R} \frac{\log|f(z)|}{z} dz \;=\; \frac{1}{2\pi}\int_0^{2\pi} \log|f(Re^{it})| dt \;=\; \log|f(0)| + \sum_{i=1}^{m_R} \log\frac{R}{|a_i|}\,,$$

where m_R is the largest index with $|a_{m_R}| < R$. Considering this expression as a function N_f of $\log R$, then obviously N_f is a piecewise linear convex function whose gradient is $\sum_{i=1}^{m_R} 1 = m_R$, i.e., the number of zeroes of f inside the disc $\{|z| \in \mathbb{C}^n \,:\, |z| < R\}$.

A main analytic tools in the study of amoebas is the Ronkin function which can be seen as a certain generalization of N_f to functions in several variables.

Definition 10.6.5. For a polynomial $f \in \mathbb{R}[X_1, \ldots, X_n]$ the *Ronkin function* $N_f : \mathbb{R}^n \to \mathbb{R}$ is defined by

$$N_f(w_1, \ldots, w_n) \;=\; \frac{1}{(2\pi i)^n}\int_{\mathrm{Log}^{-1}(w)} \frac{\log|f(z_1, \ldots, z_n)|}{z_1 \cdots z_n}\, dz_1 \cdots dz_n\,.$$

Example 10.6.6. Let $n = 2$ and f be the monomial

$$f(z_1, z_2) \;=\; c z_1^s z_2^t$$

with $c \in \mathbb{R}$. Then

$$\begin{aligned} N_f(w_1, w_2) \;&=\; \frac{1}{(2\pi i)^2}\left(\int_{\mathrm{Log}^{-1}(w_1,w_2)} \frac{\log|c|}{z_1 z_2} + \frac{s\log|z_1|}{z_1 z_2} + \frac{s\log|z_2|}{z_1 z_2}\right) dz_1 dz_2 \\ &=\; \log|c| + s w_1 + t w_2\,, \end{aligned}$$

since $\frac{1}{(2\pi i)^2}\int\limits_{\mathrm{Log}^{-1}(w_1,w_2)} \frac{s\log|z_1|}{z_1 z_2} dz_1 dz_2 \;=\; \frac{1}{2\pi}\int\limits_{t=0}^{2\pi} s(\log e^{w_1} + \log \underbrace{|e^{it}|}_{1}) dt \;=\; s w_1\,.$

N_f retains some properties from the one-dimensional case, while others are lost or attain a new form. For example, N_f is a convex function, but is not longer piecewise linear. However, on each component of $\mathbb{R}^n \setminus \mathcal{A}_f$, N_f behaves like the Ronkin function of a monomial: it is linear, and its gradient is the corresponding integer point of the Newton polytope NP_f.

Theorem 10.6.7. *i) The Ronkin function N_f is convex.*

ii) N_f is affine on each component of $\mathbb{R}^n \setminus \mathcal{A}_f$ and stricly convex on $\mathcal{A}_f$.

iii) The derivative of N_f with respect to z_j is the real part of

$$\nu_j = \frac{1}{(2\pi i)^n} \int\limits_{\mathrm{Log}^{-1}(w)} \frac{z_j \partial_j f(z)}{f(z)} \frac{dz_1 \cdots dz_n}{z_1 \cdots z_n}\,, \qquad 1 \le j \le n\,.$$

For x in a connected component C of $\mathbb{R}^n \setminus \mathcal{A}_f$, the vector $\nu = (\nu_1, \ldots, \nu_n)$ is defined to be the *order* of the component C (the invariance of ν in the same complement component can also be seen from complex analysis arguments). Moreover, two different points $x, x' \in {}^c\mathrm{Log}\ \mathcal{V}(f)$ have the same order if and only if they are contained in the same connected component E of ${}^c\mathrm{Log}\ \mathcal{V}(f)$. Moreover, it can be shown that the order ν of any component of ${}^c\mathrm{Log}\ \mathcal{V}(f)$ is contained in the Newton polytope $\mathrm{New}(f)$.

The maximum of the affine functions underlying the Ronkin function on the complement components is a piecewise linear convex function. The set where it is not differentiable is called the *spine*.

In order to compute an order, the following description is useful.

Theorem 10.6.8. *If x is in the complement of an amoeba $\mathcal{A}_f$, then* $\mathrm{grad}\, N_f(x)$ *is equal to the order of the complement component containing x*

The importance of the spine comes from the following statement.

Theorem 10.6.9. *Let $f \in \mathbb{R}[X_1, \ldots, X_n]$. Then the spine $\mathcal{S}_f$ is a polyhedral complex which is dual to a subdivision the Newton polytope of f. $\mathcal{S}_f$ is a deformation retract of the amoeba, i.e., the complement $\mathbb{R}^n \setminus \mathcal{S}_f$ consists of a finite number of polyhedra, and each of these polyhedra contains exactly one connected component of the amoeba complement $\mathbb{R}^n \setminus \mathcal{A}_f$.*

10.6.3 Maslov dequantization of amoebas

We now consider a deformation of an amoeba of a polynomial $f \in \mathbb{R}[x_1, \ldots, x_n]$ to the "natural" tropical hypersurface associated with f. For simplicity, let $n = 2$.

Lemma 10.6.10. *If a point $x \in \mathbb{R}^n$ belongs to the amoeba*

$$\mathrm{Log}_t(\{z \in (\mathbb{C}^*)^n \,:\, g_t(z) = 0\})$$

then for each multiindex α we have

$$c_\alpha \odot x^\alpha \;\geq\; \bigoplus_{\beta\neq\alpha} c_\beta \odot x^\beta \,.$$

(Here, the index t in $\bigoplus$ is omitted for notational convenience.)

Proof. If $x = \mathrm{Log}_t z$ with $g_t(z) = 0$ then for each α

$$t^{c_\alpha} z^\alpha \;=\; \sum_{\beta\neq\alpha} t^{c_\beta} z^\beta \,.$$

Passing over to the absolute value and applying the triangle inequality yields

$$t^{c_\alpha} |z|^\alpha \;\leq\; \sum_{\beta\neq\alpha} t^{c_\beta} |z|^\beta \,.$$

Now applying $\log_t$ (for $0 < t < 1$) on both sides gives

$$c_\alpha \odot x^\alpha \;\geq\; \bigoplus_{\beta\neq\alpha} c_\beta \odot x^\beta \,.$$

□

The *Hausdorff distance* between two closed subsets $A, B \subset \mathbb{R}^n$ is defined by

$$\max\{\sup_{a\in A} d(a,B), \sup_{b\in B} d(b,A)\}\,,$$

where $d(a,B)$ is the Euclidean from a to B. Let $\mathcal{A}_t = \mathrm{Log}_t(V_t)$ and $\mathcal{A}_{\mathrm{trop}}$ be the tropical hypersurface of the tropical polynomial with the coefficients of f.

Theorem 10.6.11. *For $t \downarrow 0$, the sequence of $\mathcal{A}_t$ converges in the Hausdorff metric to the tropical hypersurface $\mathcal{A}_{\mathrm{trop}}$.*

Exercises

Exercise 10.6.12. Give an example of a univariate polynomial f with pairwise distinct zeroes such that the amoeba $\mathcal{A}(f)$ consists of a single point.

Notes

All these results refer to the case where X is an algebraic hypersurface. A main difficulty in the treatment of amoebas of arbitrary varieties comes from the following simple observation. If X, Y, and Z are subvarieties of $(\mathbb{C}^*)^n$ with $X \cap Y = Z$, then $\mathrm{Log}\, Z \subset \mathrm{Log}\, X \cap \mathrm{Log}\, Y$, but in general the inclusion is proper.

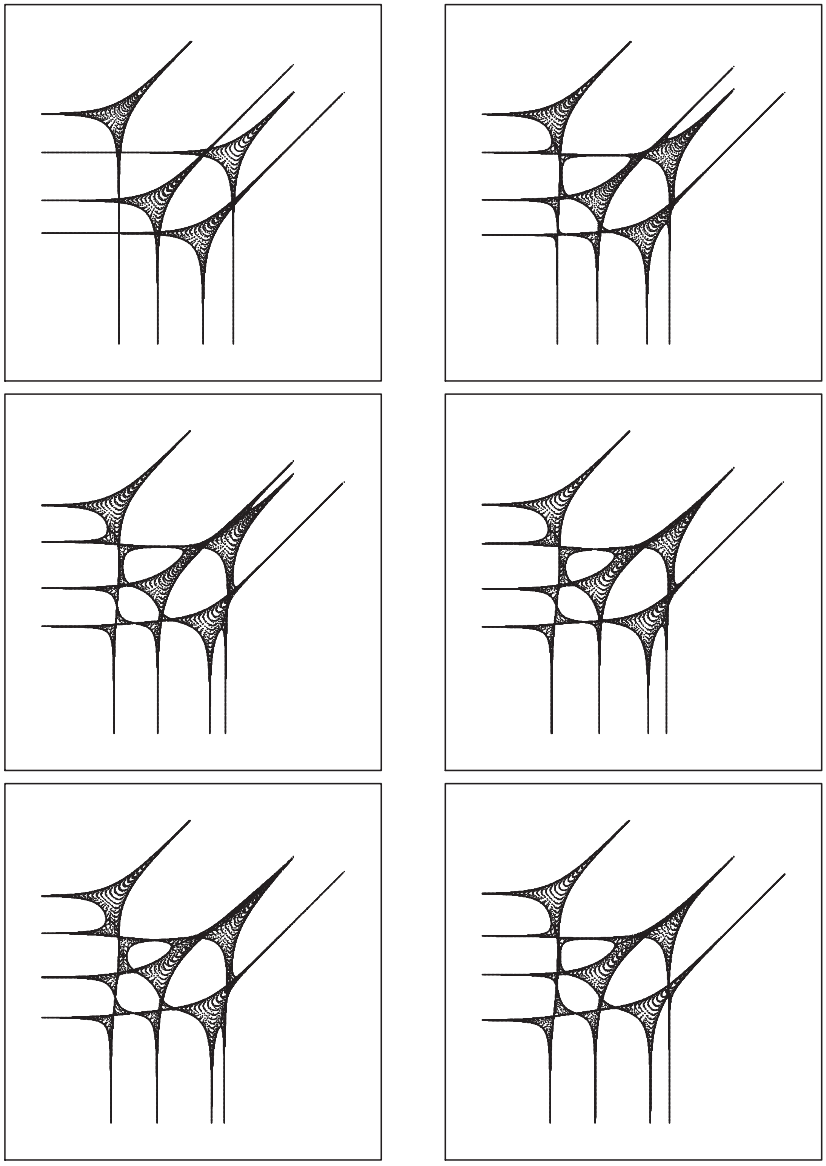


Figure 10.4: A series of quartic amoebas in two variables. The first picture shows the amoeba of $\mathcal{V}(f_1 \cdot f_2 \cdot f_3 \cdot f_4)$, where $f_1(z_1, z_2) = \left(\frac{1}{30}z_1 + \frac{1}{30}z_2 - 1\right)$, $f_2(z_1, z_2) = \left(\frac{1}{5}z_1 + 4z_2 - 1\right)$, $f_3(z_1, z_2) = \left(3z_1 + \frac{4}{7}z_2 - 1\right)$, $f_4(z_1, z_2) = \left(30z_1 + \frac{1}{300}z_2 - 1\right)$.

Chapter 11

Non Linear Computational Geometry

Outline:

1. Case Study: Lines tangent to four spheres.
2. Stewart platform and robotics
3. Geometric Modelling

11.1 Lines tangent to four spheres

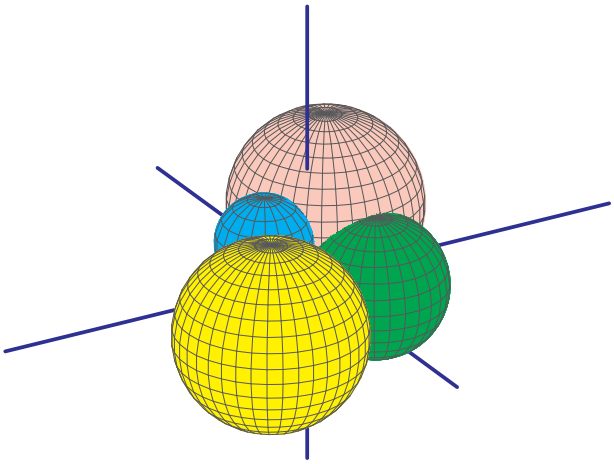
Methods and ideas from algebraic geometry can be fruitfully applied to solve problems in the ordinary geometry of three-dimensional space. To illustrate this, we consider the following geometrical problem:

How many lines are tangent to four general spheres?

We shall see that this simple problem contains quite a lot of interesting geometry.

We approach this problem computationally, formulating it as a system of equations, and then we shall solve one instance of the question, determining the lines that are tanget to the spheres with centers $(0, 1, 1)$, $(-1, -1, 0)$, $(1, -1, 1)$, $(-2, 2, 0)$, and respective radii

1, 3/2, 2, and 2, drawn below.



For this, we first give a system of coordinates for lines, then determine equations for a line to be tangent to a sphere, and finally, solve this instance.

11.1.1 Coordinates for lines

We will consider a line in $\mathbb{R}^3$ as a line in complex projective space, $\mathbb{P}^3$. Thus we study an *algebraic relaxation* of the original problem in $\mathbb{R}^3$, as our coordinates will include complex lines as well as lines at infinity. Recall that projective space $\mathbb{P}^3$ is the set of all 1-dimensional linear subspaces of $\mathbb{C}^4$. Under this correspondence, a line in $\mathbb{P}^3$ is the projective space of a 2-dimensional linear subspace of $\mathbb{C}^4$.

We will need some multilinear algebra[†]. Consider the inclusion $V \hookrightarrow \mathbb{C}^4$, where V is a 2-dimensional linear subspace of $\mathbb{C}^4$. Applying the second exterior power gives

$$\wedge^2 V \;\hookrightarrow\; \wedge^2 \mathbb{C}^4 \;\simeq\; \mathbb{C}^{\binom{4}{2}} \;\simeq\; \mathbb{C}^6 .$$

Since V is 2-dimensional, $\wedge^2 V \simeq \mathbb{C}$, and so $\wedge^2 V$ is a 1-dimensional linear subspace of $\wedge^2 \mathbb{C}^4$, which a point in the corresponding projective space. In this way, we associate each 2-plane in $\mathbb{C}^4$ to a point in the projective 5-space $\mathbb{P}(\wedge^2 \mathbb{C}^4) = \mathbb{P}^5$.

This map

$$\{\text{2-planes in } \mathbb{C}^4\} \;\longrightarrow\; \mathbb{P}^5 \;=\; \mathbb{P}(\wedge^2 \mathbb{C}^4)$$

is called the *Plücker embedding* and $\mathbb{P}(\wedge^2 \mathbb{C}^4)$ is called *Plücker space*. We identify its image. A tensor in $\wedge^2 \mathbb{C}^4$ is *decomposable* if it has the form $u \wedge v$. If $V = \langle u, v \rangle$ is the linear span of u and v, then $\wedge^2 V = \langle u \wedge v \rangle$ is spanned by decomposable tensors. Conversely, any non-zero decomposable tensor $u \wedge v \in \mathbb{P}(\wedge^2 \mathbb{C}^4)$ is the image of a unique 2-dimensional linear subspace $\langle u, v \rangle$ of $\mathbb{C}^4$ (Exercise 2).

[†]Do this in the algebra appendix, Appendix A

Let us investigate decomposable tensors. A basis e_0, e_1, e_2, e_3 for $\mathbb{C}^4$, gives the basis

$$e_0 \wedge e_1\,,\ e_0 \wedge e_2\,,\ e_0 \wedge e_3\,,\ e_1 \wedge e_2\,,\ e_1 \wedge e_3\,,\ e_2 \wedge e_3$$

for $\wedge^2\mathbb{C}^4$, showing that it is six-dimensional. If

$$u \ = \ u_0e_0 + u_1e_1 + u_2e_2 + u_3e_3 \qquad \text{and} \qquad v \ = \ v_0e_0 + v_1e_1 + v_2e_2 + v_3e_3$$

are vectors in $\mathbb{C}^4$, their exterior product $u \wedge v$ is

$$\begin{aligned} &(u_0v_1 - u_1v_0)e_0 \wedge e_1 \ + \ (u_0v_2 - u_2v_0)e_0 \wedge e_2 \ + \ (u_0v_3 - u_3v_0)e_0 \wedge e_3 \\ + \ &(u_1v_2 - u_2v_1)e_1 \wedge e_2 \ + \ (u_1v_3 - u_3v_1)e_1 \wedge e_3 \ + \ (u_2v_3 - u_3v_2)e_2 \wedge e_3\,. \end{aligned}$$

The components

$$p_{ij} \ := \ u_iv_j - u_jv_i \ = \ \det\begin{pmatrix} u_i & u_j \\ v_i & v_j \end{pmatrix} \qquad 0 \leq i < j \leq 3$$

of this decomposable tensor are the *Plücker coordinates* of the 2-plane $\langle u, v\rangle$.

Observe that $p_{03}p_{12}$ equals

$$(u_0v_3 - u_3v_0)(u_1v_2 - u_2v_1) \ = \ u_0u_1v_2v_3 - u_0u_2v_1v_3 - u_1u_3v_0v_2 + u_2u_3v_0v_1\,.$$

Another product which has the same leading term, $u_0u_1v_2v_3$, in the lexicographic monomial order where $u_0 > u_1 > \cdots > v_2 > v_3$ is $p_{02}p_{13}$, which is

$$(u_0v_2 - u_2v_0)(u_1v_3 - u_3v_1) \ = \ u_0u_1v_2v_3 - u_0u_3v_1v_2 - u_1u_2v_0v_3 + u_2u_3v_0v_1\,.$$

If we subtract these, $p_{03}p_{12} - p_{02}p_{13}$, we obtain

$$-(u_0u_2v_1v_3 - u_0u_3v_1v_2 - u_1u_2v_0v_3 + u_1u_3v_0v_2) \ = \ -(u_0v_1 - u_1v_0)(u_2v_3 - u_3v_2)\,,$$

which is $p_{01}p_{23}$. We have just applied the subduction algorithm to the polynomials p_{ij} to obtain the quadratic *Plücker relation*

$$p_{03}p_{12} - p_{02}p_{13} + p_{01}p_{23} \ = \ 0\,, \tag{11.1}$$

which holds on the Plücker coordinates of decomposable tensors. In the exercises, you are asked to show that any $p_{01}, \ldots, p_{23}$ satisfying this relation is a Plücker coordinate of a 2-plane in $\mathbb{C}^4$.

Definition 11.1.1. The *Grassmannian of 2-planes in* $\mathbb{C}^4$, $G(2,4)$, is the algebraic subvariety of Plücker space defined by the Plücker relation (11.1). We may also write $\mathbb{G}(1,3)$ for this Grassmannian, when we consider it as the space of lines in $\mathbb{P}^3$.

The Plücker coordinates of the Grassmannian as a subvariety of Plücker space give us a set of coordinates for lines.

11.1.2 Equation for a line to be tangent to a sphere

The points of a sphere S with center (a, b, c) and radius r are the homogeneous coordinates $X = [x_0, x_1, x_2, x_3]^T$ that satisfy the quadratic equation $X^T Q X = 0$, where

$$Q \;=\; \begin{bmatrix} a^2+b^2+c^2-r^2 & -a & -b & -c \\ -a & 1 & 0 & 0 \\ -b & 0 & 1 & 0 \\ -c & 0 & 0 & 1 \end{bmatrix} . \tag{11.2}$$

Indeed, $X^T Q X$ is $(x_1 - a x_0)^2 + (x_2 - b x_0)^2 + (x_3 - c x_0)^2 - r^2 x_0^2$. This matrix Q defines an isomorphism $\mathbb{C}^4 \xrightarrow{Q} (\mathbb{C}^4)^\vee$, between $\mathbb{C}^4$ and its linear dual, $(\mathbb{C}^4)^\vee$. The quadratic form $X^T Q X$ is simply the pairing between $X \in \mathbb{C}^4$ and $QX \in (\mathbb{C}^4)^\vee$.

If V is a 2-plane in $\mathbb{C}^4$, then its intersection with the sphere S is the zero set of this quadratic form restricted to V. There are three possibilities for such a homogeneous quadratic form on $V \simeq \mathbb{C}^2$. If it is non-zero, then it factors. Either it has two distinct factors, and thus the line corresponding to V meets the sphere in 2 distinct points, or else it is the square of a linear form, and thus the line is tangent to the sphere. The third possibility is that it is zero, in which case, the line lies on the sphere (such a line is necessarily imaginary) and is tangent to the sphere at every point of the line. The three cases are distinguished by the rank of the matrix representing the quadratic form (Exercise 3). In particular, the line is tangent to the sphere if and only if the determinant of this matrix vanishes.

We investigate its determinant. This restriction is defined by the composition of maps

$$V \;\hookrightarrow\; \mathbb{C}^4 \;\xrightarrow{Q}\; (\mathbb{C}^4)^\vee \;\twoheadrightarrow\; V^\vee , \tag{11.3}$$

where the last map is the restriction of a linear form on $\mathbb{C}^4$ to V. The line represented by V is tangent to S if and only if this quadratic form is degenerate, which means that the map does not have full rank. To take the determinant of this map between two-dimensional vector spaces, we apply the second exterior power $\wedge^2$ to the composition (11.3) and obtain

$$\wedge^2 V \;\hookrightarrow\; \wedge^2 \mathbb{C}^4 \;\xrightarrow{\wedge^2 Q}\; \wedge^2 (\mathbb{C}^4)^\vee \;\twoheadrightarrow\; \wedge^2 V^\vee .$$

Since the image of $\wedge^2 V$ in $\wedge^2 \mathbb{C}^4$ is spanned by the Plücker vector p of $\wedge^2 V$ and we restrict a linear form on $\wedge^2 \mathbb{C}^4$ to $\wedge^2 V^\vee$ by evaluating it at p, we obtain the equation

$$p^T \wedge^2 Q \; p \;=\; 0 ,$$

for the line with Plücker coordinate p to be tangent to the sphere defined by the quadratic form Q.

If we express $\wedge^2 Q$ as a matrix with respect to the basis $e_i \wedge e_j$ of $\wedge^2 \mathbb{C}^4$, it will have rows and columns indexed by pairs ij with $0 \le i < j \le 3$, where

$$(\wedge^2 Q)_{ij,kl} \;:=\; Q_{ik} Q_{jl} - Q_{il} Q_{jk} \;=\; \det \begin{pmatrix} Q_{ik} & Q_{il} \\ Q_{jk} & Q_{jl} \end{pmatrix} .$$

For our sphere (11.2), this is

$$\wedge^2 Q \;=\; \begin{pmatrix} b^2+c^2-r^2 & -ab & -ac & b & c & 0 \\ -ab & a^2+c^2-r^2 & -bc & -a & 0 & c \\ -ac & -bc & a^2+b^2-r^2 & 0 & -a & -b \\ b & -a & 0 & 1 & 0 & 0 \\ c & 0 & -a & 0 & 1 & 0 \\ 0 & c & -b & 0 & 0 & 1 \end{pmatrix} \begin{matrix} 01 \\ 02 \\ 03 \\ 12 \\ 13 \\ 23 \end{matrix} \tag{11.4}$$

We remark that there is nothing special about spheres in this discussion.

Theorem 11.1.2. *If q is any smooth quadric in $\mathbb{P}^3$ defined by a quadratic form Q, then a line with Plücker coordinate p is tangent to q if and only if $p^T \wedge^2 Q\; p = 0$, if and only if p lies on the quadric in Plücker space defined by $\wedge^2 Q$.*

11.1.3 Solving the equations?

We may now formulate our problem of lines tangent to four spheres as the solutions to a system of equations. Namely, the set of lines tangent to four spheres have Plücker coordinates in $\mathbb{P}^5$ which satisfy

1. The Plücker equation (11.1), and
2. Four quadratic equations of the form $p^T \wedge^2 Q\; p = 0$, one for each sphere.

By Bézout's theorem, we expect that there will be 2^5 solutions to these five quadratic equations on $\mathbb{P}^5$.

Let us investigate these equations. We will use the symbolic computation package Singular [31] and display both annotated code and output in `typewriter font`. Output lines begin with `//`, which are comment-line characters in Singular. First, we define our ground ring R to be $\mathbb{Q}[u, v, w, x, y, z]$ with the degree reverse lexicographic monomial order where $u > v > \cdots > z$. This is the coordinate ring of Plücker space, where we identify p_{01} with u, p_{02} with v, p_{03} with w, and so on. We also declare the types of some variables.

```
ring R = 0, (u,v,w,x,y,z), dp;
matrix wQ[6][6];
matrix Pc[6][1] = u,v,w,x,y,z;
```

We give a procedure to compute $\wedge^2 Q$ (11.4),

```
proc Wedge_2_Sphere (poly r, a, b, c)
 {
  wQ = b^2+c^2-r^2 , -a*b , -a*c ,  b , c , 0,
       -a*b , a^2+c^2-r^2 , -b*c , -a , 0 , c,
       -a*c , -b*c , a^2+b^2-r^2 ,  0 ,-a ,-b,
```

```
                    b  , -a  ,  0  ,  1 , 0 , 0,
                    c  ,  0  , -a  ,  0 , 1 , 0,
                    0  ,  c  , -b  ,  0 , 0 , 1;
  return(wQ);
 }
```

and a procedure to compute the quadratic form $p^T \wedge^2 Q\ p$.

```
proc makeEquation (poly r, a, b, c)
 {
  return((transpose(Pc)*WedgeTwoSphere(r,a,b,c)*Pc)[1][1]);
 }
```

Now we create the ideal defining the lines tangent to four spheres with radii 1, 3/2, 2, and 2, and respective centers $(0, 1, 1)$, $(-1, -1, 0)$, $(1, -1, 1)$, and $(-2, 2, 0)$.

```
ideal I =
   w*x-v*y+u*z,
   makeEquation(1 , 0, 1, 1),
   makeEquation(3/2 ,-1,-1, 0),
   makeEquation(2 , 1,-1, 1),
   makeEquation(2 ,-2, 2,0);
```

Lastly, we compute a Gröbner basis for I and determine its dimension and degree.

```
 I=std(I);
 degree(I);
// dimension (proj.)  = 1
// degree (proj.)   = 4
```

This computation shows that the set of lines tangent to these four spheres has dimension 1, so that there are infinitely many common tangents! This is not what we expected from Bézout's Theorem and we must conclude that our equations are *not* sufficiently general. We will try to understand the special structure in our equations.

The key to this, as it turns out, is to use some classical facts about spheres. It is well-knkown that circles are exactly the conics in the plane $\mathbb{P}^2$ which contain the imaginary circular points at infinity $[0, 1, \pm i]$. This is clear if we set $x_0 = 0$ in the equation for a circle with center (a, b) and radius r,

$$(x_1 - ax_0)^2 + (x_2 - ax_0)^2 \ = \ r^2 x_0^2 \, .$$

For the same reason, spheres are the quadrics in $\mathbb{P}^3$ which contain the imaginary *circular conic at infinity*, which is defined by

$$x_0 \ = \ 0 \qquad \text{and} \qquad x_1^2 + x_2^2 + x_3^2 \ = \ 0 \, .$$

The lines at infinity have Plücker coordinates satisfying $p_{01} = p_{02} = p_{03} = 0$. For such a point p, the equation $p^T \wedge^2 Q\, p$, where Q is (11.4), becomes

$$p_{12}^2 + p_{13}^2 + p_{23}^2 \;=\; 0\,. \tag{11.5}$$

This is the condition that the line at infinity is tangent to the spherical conic at infinity. Since the parameters r, a, b, c for the sphere do not appear in the equations $p_{01} = p_{02} = p_{03} = 0$ and (11.5), every line at infinity tangent to the spherical conic at infinity is tangent to every sphere.

In the language of enumerative geometry, this problem of lines tangent to four spheres has *excess intersection*. That is, our equations for a line to be tangent to four spheres not only define the lines we want (the tangent lines not at infinity), but also lines we did not intend, namely these lines tangent to the spherical conic at infinity.

If I is the ideal generated by our equations and J is the ideal of this excess component, then by Lemma 7.6.7, the saturation $(I\colon J^\infty)$ is the ideal of $\overline{\mathcal{V}(I) \setminus \mathcal{V}(J)}$, which should be the tangents that we seek. (We could also saturate by the ideal $K = \langle p_{01}, p_{02}, p_{03}\rangle$ of lines at infinity.) We return to our Singular computation, defining the ideal J and computing the quotient ideal $(I\colon J)$. We do this instead of saturation, as saturation is typically computationally expensive.

```
   ideal J = std(ideal(u,v,w,x^2+y^2+z^2));
   I = std(quotient(I,J));
   degree(I);
// dimension (proj.)  = 1
// degree (proj.)   = 2
```

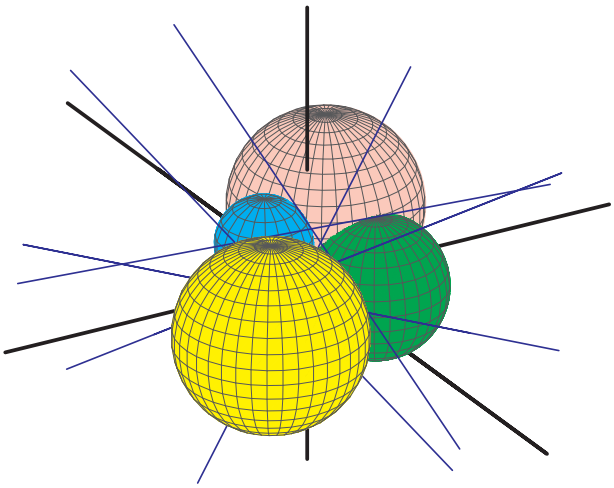
While the degree of $(I\colon J)$ is less than that of I, it is still 1-dimensional, so we take the quotient ideal again.

```
   I = std(quotient(I,J));
   degree(I);
// dimension (proj.)  = 0
// degree (proj.)   = 12
```

The dimension is now zero and we have removed the excess component from $\mathcal{V}(I)$.

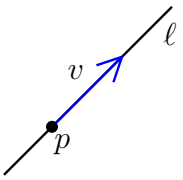
Since the dimension is 0 and the degree is 12, we expect that 12 is the answer to our original question. That is, we expect that there will be 12 *complex* lines tangent to amy four spheres in general position. In Exercise 6 we ask you to verify that there are indeed 12 complex common tangent lines to the four spheres. Of these 12, six are real, and we

display them with the spheres below.



11.1.4 Twelve lines tanget to four general spheres

We remark that the computation of Section 11.1.3, while convincing, does not constitute a proof. We will give a rigorous proof here. Our basic idea to handle the excess component is to simply define it away. Represent a line ℓ in $\mathbb{R}^3$ by a point $p \in \ell$ and a direction vector $v \in \mathbb{RP}^2$.



No such line can lie at infinity, so we are avoiding the excess component of lines at infinity tangent to the spherical conic at infinity.

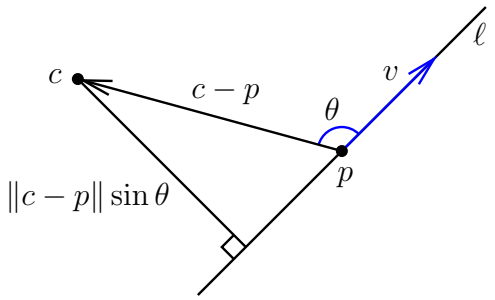
Lemma 11.1.3. *The set of direction vectors $v \in \mathbb{RP}^2$ of lines tangent to four spheres with affinely independent centers consists of the common solutions to a cubic and a quartic equation on $\mathbb{RP}^2$. Each direction vector gives one common tangent.*

Proof. For vectors $x, y \in \mathbb{R}^3$, let $x \cdot y$ be their ordinary Euclidean dot product and write x^2 for $x \cdot x$, which is $\|x\|^2$. Fix p to be the point of ℓ closest to the origin, so that

$$p \cdot v \;=\; 0\,. \tag{11.6}$$

The distance from the line ℓ to a point c is $\|c - p\| \sin\theta$, where θ is the angle between

v and the displacement vector from p to c.



This is the length of the cross product $(c-p)\times v$, multiplied by the length $\|v\|$ of the direction vector v. If ℓ is tangent to the sphere with radius r centered at $c\in\mathbb{R}^3$ if its distance to c is r, and so we have $\|(c-p)\times v\| = r\|v\|$. Squaring, we get

$$[(c-p)\times v]^2 \;=\; r^2v^2\,. \tag{11.7}$$

This formulation requires that $v^2\neq 0$. A line with $v^2=0$ has a complex direction vector and it meets the spherical conic at infinity.

We assume that one sphere is centered at the origin and has radius r, while the other three have centers and radii (c_i,r_i) for $i=1,2,3$. The condition for the line to be tangent to the sphere centered at the origin is

$$p^2 \;=\; r^2\,. \tag{11.8}$$

For the other spheres, we expand (11.7), use vector product identities, and the equations (11.6) and (11.8) to obtain the vector equation

$$2v^2\begin{pmatrix}c_1^T\\ c_2^T\\ c_3^T\end{pmatrix}\cdot p \;=\; -\begin{pmatrix}(c_1\cdot v)^2\\ (c_2\cdot v)^2\\ (c_3\cdot v)^2\end{pmatrix} \;+\; v^2\begin{pmatrix}c_1^2+r^2-r_1^2\\ c_2^2+r^2-r_2^2\\ c_3^2+r^2-r_3^2\end{pmatrix}\;. \tag{11.9}$$

Now suppose that the spheres have affinely independent centers. Then the matrix $(c_1,c_2,c_3)^T$ appearing in (11.9) is invertible. Assuming $v^2\neq 0$, we may use (11.9) to write p as a quadratic function of v. Substituting this expression into equations (11.6) and (11.8), we obtain a cubic and a quartic equation for $v\in\mathbb{RP}^2$. The lemma now follows from Bézout's Theorem. □

Bézout's Theorem implies that there are at most $3\cdot 4=12$ isolated solutions to these equations, and over $\mathbb{C}$ exactly 12 if they are generic. The equations are however far from generic as they involve only 13 parameters while the space of quartics has 14 parameters and the space of cubics has 9 parameters.

Example 11.1.4. Suppose that the spheres have equal radii, r, and have centers at the vertices of a regular tetrahedron with side length $2\sqrt{2}$,

$$(2,2,0)^T\,,\qquad (2,0,2)^T\,,\qquad (0,2,2)^T\,,\quad\text{and}\quad (0,0,0)^T\,.$$

In this symmetric case, the cubic factors into three linear factors. There are real common tangents only if $\sqrt{2} \leq r \leq 3/2$, and exactly 12 when the inequality is strict. If $r = \sqrt{2}$, then the spheres are pairwise tangent and there are three common tangents, one for each pair of non-intersecting edges of the tetrahedron. Each tangent has algebraic multiplicity 4. If $r = 3/2$, then there are six common tangents, each of multiplicity 2. The spheres meet pairwise in circles of radius $1/2$ lying in the plane equidistant from their centers. This plane also contains the centers of the other two spheres, as well as one common tangent which is parallel to the edge between those centers.

Figure 11.1 shows the cubic (which consists of three lines supporting the edges of an equilateral triangle) and the quartic, in an affine piece of the set $\mathbb{RP}^2$ of direction vectors. The vertices of the triangle are the standard coordinate directions $(1,0,0)^T$, $(0,1,0)^T$, and $(0,0,1)^T$. The singular cases, (i) when $r = \sqrt{2}$ and (ii) when $r = 3/2$, are shown first, and then (iii) when $r = 1.425$. The 12 points of intersection in this third case are visible in the expanded view in (iii'). Each point of intersection gives a real tangent, so there are

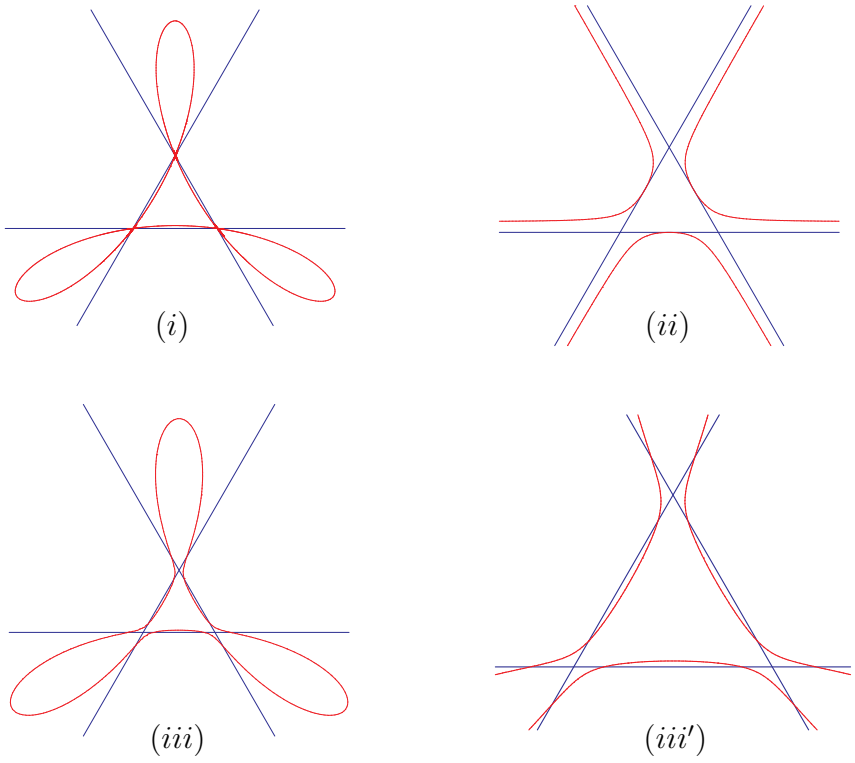


Figure 11.1: The cubic and quartic for symmetric configurations.

12 tangents to four spheres of equal radii 1.425 with centers at the vertices of the regular tetrahedron with edge length $2\sqrt{2}$.

One may also see this number 12 using group theory. The symmetry group of the tetrahedron, which is the group of permutations of the spheres, acts transitively on their common tangents and the isotropy group of any tangent has order 2. To see this, orient

a common tangent and suppose that it meets the spheres a, b, c, d in order. Then the permutation $(a, d)(b, c)$ fixes that tangent but reverses its orientation, and the identity is the only other permutation fixing that tangent.

This example shows that the bound of 12 common tangents from Lemma 11.1.3 is in fact attained.

Theorem 11.1.5. *There are at most* 12 *common real tangent lines to four spheres whose centers are not coplanar, and there exist spheres with* 12 *common real tangents.*

Example 11.1.6. We give an example when the radii are distinct, namely 1.4, 1.42, 1.45, and 1.474. Figure 4.5 shows the quartic and cubic and the configuration of 4 spheres and their 12 common tangents.

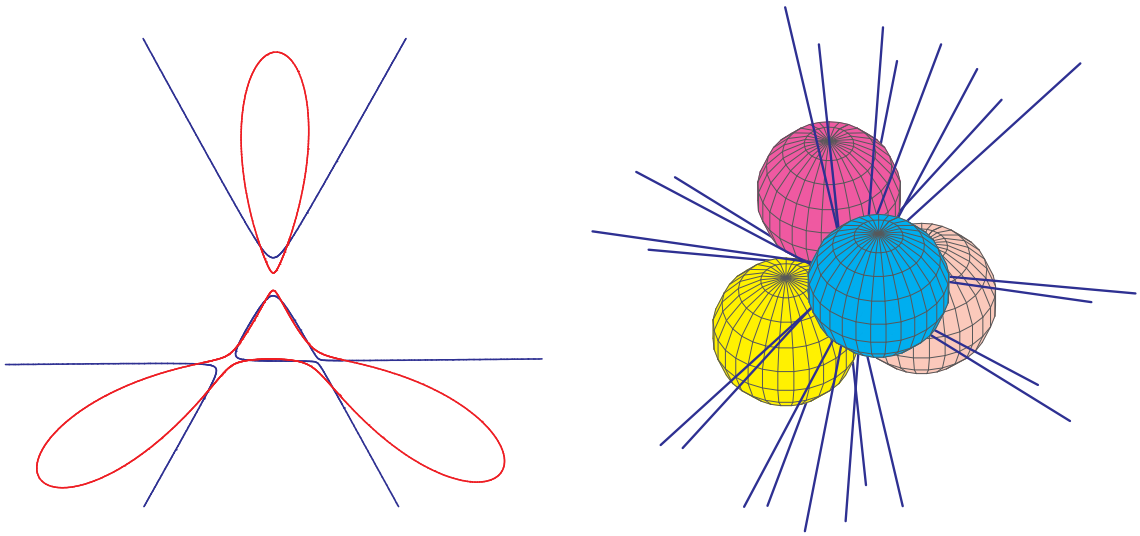


Figure 11.2: Spheres with 12 common tangents.

Now suppose that the centers are coplanar. A continuity argument shows that four general such spheres will have 12 complex common tangents (or infinitely many, but this possibility is precluded by the following example). Three spheres of radius 4/5 centered at the vertices of an equilateral triangle with side length $\sqrt{3}$ and one of radius 1/3 at the triangle's center have 12 common real tangents. We display this configuration in Figure 11.3. This configuration of spheres has symmetry group $\mathbb{Z}_2 \times D_3$, which has order 12 and acts faithfully and transitively on the common tangents.

In the symmetric configuration of Example 11.1.4 having 12 common tangents, every pair of spheres meet. It is however not necessary for the spheres to meet pairwise when there are 12 common tangents. In fact, in both Figures 4.5 and 11.3 not all pairs of spheres meet. However, the union of the spheres is connected. This turns out to be unnecessary.

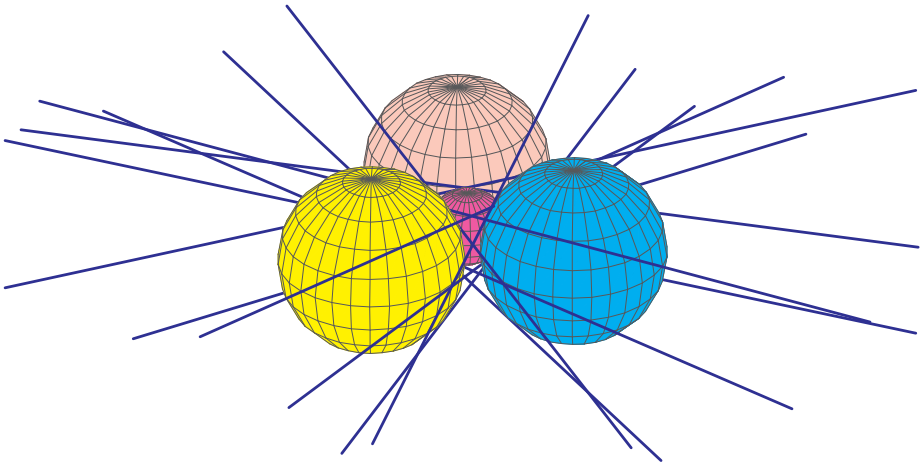


Figure 11.3: Spheres with coplanar centers and 12 common tangents.

In the tetrahedral configuration, if one sphere has radius 1.38 and the other three have equal radii of 1.44, then the first sphere does not meet the others, but there are still 12 tangents.

More interestingly, it is possible to have 12 common real tangents to four *disjoint* spheres. Figure 11.4 displays such a configuration. The three large spheres have radius

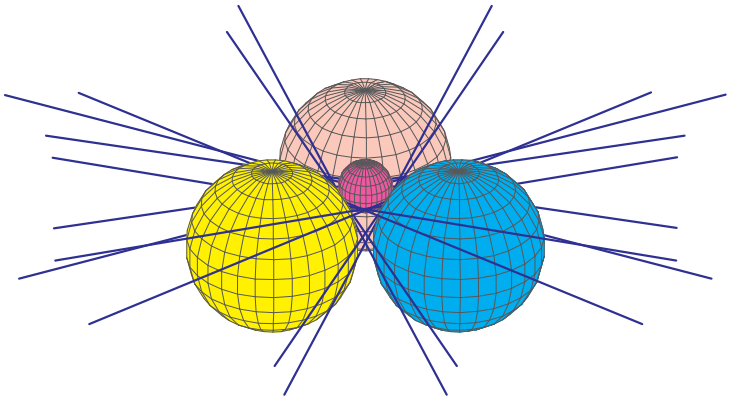


Figure 11.4: Four disjoint spheres with 12 common tangents.

4/5 and are centered at the vertices of an equilateral triangle of side length $\sqrt{3}$, while the smaller sphere has radius 1/4 and is centered on the axis of symmetry of the triangle, but at a distance of 35/100 from the plane of the triangle.

Exercises

1. Show that if $u \wedge v$ and $x \wedge y$ are two decomposable tensors representing the same point in $\mathbb{P}(\wedge^2\mathbb{C}^4)$, then they come from the same 2-plane in $\mathbb{C}^4$. That is, show that $\langle u, v\rangle = \langle x, y\rangle$.

2. Show that if $[p_{01}, \ldots, p_{23}]$ are the homogeneous coordinates of a point of $\mathbb{P}(\wedge^2\mathbb{C}^4)$ that satisfies the Plücker relation, then this is the Plücker coordinate of a 2-plane in $\mathbb{C}^4$.

3. Let Q be a symmetric 2×2 matrix. Show that the quadratic form X^TQX has distinct factors if and only if Q is invertible. If Q has rank 1, then show that X^TQX is the square of a linear form, and that X^TQX is the zero polynomial only when Q has rank zero.

4. Verify that there are indeed 12 lines tangent (four real and 8 complex) to the four spheres of the example in Section 11.1.3.

5. Show that four general quadrics in $\mathbb{P}^3$ will have 32 common tangents. By Bézout's theorem, it suffices to find a single instance of four quadrics with this number of tangents. You may do this by replacing the procedure `WedgeTwoSphere` with a procedure to compute $\wedge^2 Q$ for an arbitrary 4×4 symmetric matrix Q.

6. Verify the claim that there are 12 complex tangents to the four spheres discussed in Section 11.1.3. Show that six of these are real.

Appendix A

Appendix

A.1 Algebra

Algebra is the foundation of algebraic geometry here we collect some of the basic algebra on which we rely. We develop some algebraic background that is needed in the text. This may not be an adequate substitute for a course in abstract algebra. Proofs can be found in give some useful texts.

A.1.1 Fields and Rings

We are all familiar with the real numbers, $\mathbb{R}$, with the rational numbers $\mathbb{Q}$, and with the complex numbers $\mathbb{C}$. These are the most common examples of *fields*, which are the basic building blocks of both the algebra and the geometry that we study. Formally and briefly, a field is a set $\mathbb{F}$ equipped with operations of addition and multiplication and distinguished elements 0 and 1 (the additive and multiplicative identities). Every number $a \in \mathbb{F}$ has an additive inverse $-a$ and every non-zero number $a \in \mathbb{F}^\times := \mathbb{F} - \{0\}$ has a multiplicative inverse $a^{-1} =: \frac{1}{a}$. Addition and multiplication are commutative and associative and multiplication distributes over addition, $a(b + c) = ab + ac$. To avoid triviality, we require that $0 \neq 1$.

The set of integers $\mathbb{Z}$ is not a field as $\frac{1}{2}$ is not an integer. While we will mostly be working over $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$, at times we will need to discuss other fields. Most of what we do in algebraic geometry makes sense over any field, including the finite fields. In particular, linear algebra (except numerical linear algebra) works over any field.

Linear algebra concerns itself with *vector spaces*. A vector space V over a field $\mathbb{F}$ comes equipped with an operation of addition—we may add vectors and an operation of multiplication—we may multiply a vector by an element of the field. A linear combination of vectors $v_1, \ldots, v_n \in V$ is any vector of the form

$$a_1 v_1 + a_2 v_2 + \cdots + a_n v_n \,,$$

where $a_1, \ldots, a_n \in \mathbb{F}$. A collection S of vectors *spans* V if every vector in V is a linear

combination of vectors from S. A collection S of vectors is *linearly independent* if zero is not nontrivial linear combination of vectors from S. A *basis* S of V is a linearly independent spanning set. When a vector space V has a finite basis, every other basis has the same number of elements, and this common number is called the *dimension* of V.

A *ring* is the next most complicated object we encounter. A ring R comes equipped with an addition and a multiplication which satisfy almost all the properties of a field, except that we do not necessarily have multiplicative inverses. While the integers $\mathbb{Z}$ do not form a field, they do form a ring. An *ideal* I of a ring R is a subset which is closed under addition and under multiplication by elements of R. Every ring has two trivial ideals, the zero ideal $\{0\}$ and the unit ideal consisting of R itself. Given a set $S \subset R$ of elements, the smallest ideal containing S, also called the ideal *generated by* S, is

$$\langle S \rangle \;:=\; \{r_1 s_1 + r_2 s_2 + \cdots + r_m s_m \mid r_1, \ldots, r_m \in R \text{ and } s_1, \ldots, s_m \in S\}\,.$$

A primary use of ideals in algebra is through the construction of quotient rings. Let $I \subset R$ be an ideal. Formally, the *quotient ring* R/I is the collection of all sets of the form

$$[r] \;:=\; r + I \;=\; \{r + s \mid s \in I\}\,,$$

as r ranges over R. Addition and multiplication of these sets are defined in the usual way

$$\begin{aligned} [r] + [s] &= \{r' + s' \mid r' \in [r] \text{ and } s' \in [s]\} \overset{!}{=} [r+s], \quad \text{and} \\ [r] \cdot [s] &= \{r' \cdot s' \mid r' \in [r] \text{ and } s' \in [s]\} \overset{!}{=} [rs]\,. \end{aligned}$$

The last equality ($\overset{!}{=}$) in each line is meant to be surprising, it is a theorem and due to I being an ideal. Thus addition and multiplication on R/I are inherited from R. With these definitions (and also $-[r] = [-r]$, $0 := [0]$, and $1 := [1]$), the set R/I becomes a ring.

We say 'R-mod-I' for R/I because the arithmetic in R/I is just the arithmetic in R, but considered modulo the ideal I, as $[r] = [s]$ in R/I if and only if $r - s \in I$.

Ideals also arise naturally as kernels of homomorphisms. A *homomorphism* $\varphi\colon R \to S$ from the ring R to the ring S is a function that preserves the ring structure. Thus for $r, s \in R$, $\varphi(r+s) = \varphi(r) + \varphi(s)$ and $\varphi(rs) = \varphi(r)\varphi(s)$. We also require that $\varphi(1) = 1$. The *kernel* of a homomorphism $\varphi\colon R \to S$,

$$\ker \varphi \;:=\; \{r \in R \mid \varphi(r) = 0\}$$

is an ideal: If $r, s \in \ker\varphi$ and $t \in R$, then

$$\varphi(r+s) \;=\; \varphi(r) + \varphi(s) \;=\; 0 \;=\; t\varphi(r) \;=\; \varphi(tr)\,.$$

Homomorphisms are deeply intertwined with ideals. If I is an ideal of a ring R, then the association $r \mapsto [r]$ defines a homomorphism $\varphi\colon R \to R/I$ whose kernel is I. Dually, given a homomorphism $\varphi\colon R \to S$, the image of R in S is identified with $R/\ker\varphi$. More

generally, if $\varphi\colon R \to S$ is a homomorphism and $I \subset R$ is an ideal with $I \subset \ker \varphi$ (that is, $\varphi(I) = 0$), then φ induces a homomorphism $\varphi\colon R/I \to S$.

Properties of ideals induce natural properties in the associated quotient rings. An element r of a ring R is *nilpotent* if $r \neq 0$, but some power of r vanishes. A ring R is *reduced* if it has no nilpotent elements, that is, whenever $r \in R$ and n is a natural number with $r^n = 0$, then we must have $r = 0$. An ideal *radical* if whenever $r \in R$ and n is a natural number with $r^n \in I$, then we must have $r \in I$. It follows that a quotient ring R/I is reduced if and only if I is radical.

A ring R is a *domain* if whenever we have $r \cdot s = 0$ with $r \neq 0$, then we must have $s = 0$. An ideal is *prime* if whenever $r \cdot s \in I$ with $r \notin I$, then we must have $s \in I$. It follows that a quotient ring R/I is a domain if and only if I is prime.

A ring R with no nontrivial ideals must be a field. Indeed, if $0 \neq r \in R$, then the ideal rR of R generated by r is not the zero ideal, and so it must equal R. But then $1 = rs$ for some $s \in R$, and so r is invertible. Conversely, if R is a field and $0 \neq r \in R$, then $1 = r \cdot r^{-1} \in rR$, so the only ideals of R are $\{0\}$ and R. An ideal $\mathfrak{m}$ of R is *maximal* if $\mathfrak{m} \subsetneq R$, but there is no ideal I strictly contained between $\mathfrak{m}$ and R; if $\mathfrak{m} \subset I \subset R$ and $I \neq R$, then $I = \mathfrak{m}$. It follows that a quotient ring R/I is a field if and only if I is maximal.

Lastly, we remark that any ideal I of R with $I \neq R$ is contained in some maximal ideal. Suppose not. Then we may find an infinite chain of ideals

$$I =: I_0 \subsetneq I_1 \subsetneq I_2 \subsetneq \cdots$$

where each is proper so that 1 lies in none of them. Set $J := \bigcup_n I_n$. Then we claim that the union $I := \bigcup_n I_n$ of these ideals is an ideal. Indeed, if $r, s \in I$ then there are indices i, j with $r \in I_i$ and $s \in I_j$. Since $I_i, I_j \subset I_{\max(i,j)}$, we have $r + s \in I_{\max(i,j)} \subset J$. If $t \in R$, then $tr \in I_i \subset J$. We remark that maximal ideals are prime ideals.

Needed (T.T):

Theorem A.1.1. *Let I be an ideal of a ring R. Then:*

1. *$\sqrt{I}$ is the intersection of all prime ideals containing I.*
2. *For every prime ideal P containing I there exists a minimal prime ideal $P' \subset P$ containing I.*
3. *If P is a minimal prime ideal containing I and $a \in P$ then there exists $b \in R \setminus P$ and $n \geq 0$ with $a^n b \in I$.*

A.1.2 Fields and polynomials

Our basic algebraic objects are polynomials. A *univariate polynomial* p is an expression of the form

$$p = p(x) := a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m , \tag{A.1}$$

where m is a nonnegative integer and the coefficients $a_0, a_1, \ldots, a_m$ lie in $\mathbb{F}$. Write $\mathbb{F}[x]$ for the set of all polynomials in the variable x with coefficients in $\mathbb{F}$. We may add, subtract, and multiply polynomials and $\mathbb{F}[x]$ is a ring.

While a polynomial p may be regarded as a formal expression (A.1), evaluation of a polynomial defines a function $p\colon \mathbb{F} \to \mathbb{F}$: The value of the function p at a point $a \in \mathbb{F}$ is simply $p(a)$. When $\mathbb{F}$ is infinite, the polynomial and the function determine each other, but this is not the case when $\mathbb{F}$ is finite.

Our study requires polynomials with more than one variable. We first define a monomial.

Definition A.1.2. A *monomial* in the variables $x_1, \ldots, x_n$ is a product of the form

$$x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n} ,$$

where the exponents $\alpha_1, \ldots, \alpha_n$ are nonnegative integers. For notational convenience, set $\alpha := (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^{n\dagger}$ and write x^α for the expression $x_1^{\alpha_1} \cdots x_n^{\alpha_n}$. The *(total) degree* of the monomial x^α is $|\alpha| := \alpha_1 + \cdots + \alpha_n$.

A *polynomial* $f = f(x_1, \ldots, x_n)$ in the variables $x_1, \ldots, x_n$ is a linear combination of monomials, that is, a finite sum of the form

$$f \;=\; \sum_{\alpha \in \mathbb{N}^n} a_\alpha x^\alpha ,$$

where each *coefficient* a_α lies in $\mathbb{F}$ and all but finitely many of the coefficients vanish. The product $a_\alpha x^\alpha$ of an element a_α of $\mathbb{F}$ and a monomial x^α is called a *term*. The *support* $\mathcal{A} \subset \mathbb{N}^n$ of a polynomial f is the set of all exponent vectors that appear in f with a nonzero coefficient. We will say that f has support $\mathcal{A}$ when we mean that the support of f is a subset of $\mathcal{A}$.

After 0 and 1 (the additive and multiplicative identities), the most distinguished integers are the prime numbers, those $p > 1$ whose only divisors are 1 and themselves. These are the numbers $2, 3, 5, 7, 11, 13, 17, 19, 23, \ldots$ Every integer $n > 1$ has a unique factorization into prime numbers

$$n \;=\; p_1^{\alpha_1} p_1^{\alpha_2} \cdots p_n^{\alpha_n} ,$$

where $p_1 < \cdots < p_n$ are distinct primes, and $\alpha_1, \ldots, \alpha_n$ are (strictly) positive integers. For example, $999 = 3^3 \cdot 37$. Polynomials also have unique factorization.

Definition A.1.3. A nonconstant polynomial $f \in \mathbb{F}[x_1, \ldots, x_n]$ is *irreducible* if whenever we have $f = gh$ with g, h polynomials, then either g or h is a constant. That is, f has no nontrivial factors.

†Where have we defined $\mathbb{N}$?

Theorem A.1.4. *Every polynomial $f \in \mathbb{F}[x_1, \ldots, x_n]$ is a product of irreducible polynomials*

$$f \;=\; p_1 \cdot p_2 \cdots p_m \,,$$

where the polynomials $p_1, \ldots, p_m$ are irreducible and nonconstant. Moreover, this factorization is essentially unique. That is, if

$$f \;=\; q_1 \cdot q_2 \cdots q_s \,,$$

is another such factorization, then $m = s$, and after permuting the order of the factors, each polynomial q_i is a scalar multiple of the corresponding polynomial p_i.

A.1.3 Polynomials in one variable

While rings of polynomials have many properties in common with the integers, the relation is the closest for univariate polynomials. The *degree*, $\deg(f)$ of a univariate polynomial f is the largest degree of a monomial appearing in f. If this monomial has coefficient 1, then the polynomial is *monic*. This allows us to remove the ambiguity in the uniqueness of factorizations in Theorem A.1.4. A polynomial $f(x) \in \mathbb{F}[x]$ has a unique factorization of the form

$$f \;=\; f_m \cdot p_1^{\alpha_1} \cdot p_2^{\alpha_2} \cdots p_s^{\alpha_s} \,,$$

where $f_m \in \mathbb{F}^\times$ is the leading coefficient of f, the polynomials $p_1, \ldots, p_s$ are monic and irreducible, and the exponents α_i are positive integers.

Definition A.1.5. A *greatest common divisor* of two polynomials $f, g \in \mathbb{F}[x]$ (or $\gcd(f, g)$) is a polynomial h such that h divides each of f and g, and if there is another polynomial k which divides both f and g, then k divides h.

Any two polynomials f and g have a monic greatest common divisor which is the product of the common monic irreducible factors of f and g, each raised to the highest power that divides both f and g. Finding greatest common divisor would seem challenging as factoring polynomials is not an easy task. There is, however, a very fast and efficient algorithm for computing the greatest common divisor of two polynomials.

Suppose that we have polynomials f and g in $\mathbb{F}[x]$ with $\deg(g) \geq \deg(f)$,

$$\begin{aligned} f &= f_0 + f_1 x + f_2 x^2 + \cdots + f_m x^m \\ g &= g_0 + g_1 x + g_2 x^2 + \cdots + g_m x^m + \cdots + g_n x^n \,, \end{aligned}$$

where f_m and g_n are nonzero. Then the polynomial

$$S(f, g) \;:=\; g - \frac{g_n}{f_m} x^{n-m} \cdot f$$

has degree strictly less than $n = \deg(g)$. This simple operation of *reducing* f by the polynomial g forms the basis of the Division Algorithm and the Euclidean Algorithm for computing the greatest common divisor of two polynomials.

We describe the *Division Algorithm* in *pseudocode*, which is a common way to explain algorithms without reference to a specific programming language.

Algorithm A.1.6 (Division Algorithm)**.**
INPUT: Polynomials $f, g \in \mathbb{F}[x]$.
OUTPUT: Polynomials $q, r \in \mathbb{F}[x]$ with $g = qf + r$ and $\deg(r) < \deg(f)$.
Set $r := g$ and $q := 0$.
(1) If $\deg(r) < \deg(f)$, then exit.
(2) Otherwise, reduce r by f to get the expression

$$r \;=\; \frac{r_n}{f_m} x^{n-m} \cdot f \;+\; S(f, r)\,,$$

where $n = \deg(r)$ and $m = \deg(f)$. Set $q := q + \frac{r_n}{f_m} x^{n-m}$ and $r := S(f, r)$, and return to step (1).

To see that this algorithm does produce the desired expression $g = qf + r$ with the degree of r less than the degree of f, note first that whenever we are at step (1), we will always have $g = qf + r$. Also, every time step (2) is executed, the degree of r must drop, and so after at most $\deg(g) - \deg(f) + 1$ steps, the algorithm will halt with the correct answer.

The *Euclidean Algorithm* computes the greatest common divisor of two polynomials f and g.

Algorithm A.1.7 (Euclidean Algorithm)**.**
INPUT: Polynomials $f, g \in \mathbb{F}[x]$.
OUTPUT: The greatest common divisor h of f and g.
(1) Call the Division Algorithm to write $g = qf + r$ where $\deg(r) < \deg(f)$.
(2) If $r = 0$ then set $h := f$ and exit.
Otherwise, set $g := f$ and $f := r$ and return to step (1).

To see that the Euclidean algorithm performs as claimed, first note that if $g = qf + r$ with $r = 0$, then $f = \gcd(f, g)$. If $r \neq 0$, then $\gcd(f, g) = \gcd(f, r)$. Thus the greatest common divisor h of f and g is always the same whenever step (1) is executed. Since the degree of r must drop upon each iteration, r will eventually become 0, which shows that the algorithm will halt and return h.[†]

An ideal is *principal* if it has the form

$$\langle f \rangle \;=\; \{h \cdot f \mid h \in \mathbb{F}[x]\}\,,$$

for some $f \in \mathbb{F}[x]$. We say that f *generates* $\langle f \rangle$. Since $\langle f \rangle = \langle \alpha f \rangle$ for any $\alpha \in \mathbb{F}$, the principal ideal has a unique monic generator.

Theorem A.1.8. *Every ideal I of $\mathbb{F}[x]$ is principal.*

[†]This is poorly written!

Proof. Suppose that I is a nonzero ideal of $\mathbb{F}[x]$, and let f be a nonzero polynomial of minimal degree in I. If $g \in I$, then we may apply the Division Algorithm and obtain polynomials $q, r \in \mathbb{F}[x]$ with

$$g \;=\; qf + r \qquad \text{with} \qquad \deg(r) < \deg(f)\,.$$

Since $r = g - qf$, we have $r \in I$, and since $\deg(r) < \deg(f)$, but f had minimal degree in I, we conclude that f divides g, and thus $I = \langle f \rangle$. □

The ideal generated by univariate polynomials $f_1, \ldots, f_s$ is the principal ideal $\langle p \rangle$, where p is the greatest common divisor of $f_1, \ldots, f_s$.

For univariate polynomials p the quotient ring $\mathbb{F}[x]/\langle p \rangle$ has a concrete interpretation. Given $f \in \mathbb{F}[x]$, we may call the Division Algorithm to obtain polynomials q, r with

$$f \;=\; q \cdot p + r, \text{ where } \deg(r) < \deg(q)\,.$$

Then $[f] = f + \langle p \rangle = r + \langle p \rangle = [r]$ and in fact r is the unique polynomial of minimal degree in the coset $f + \langle p \rangle$. We call this the *normal form* of f in $\mathbb{F}[x]/\langle p \rangle$.

Since, if $\deg(r), \deg(s) < \deg(p)$, we cannot have $r - s \in \langle p \rangle$ unless $r = s$, we see that the monomials $1, x, x^2, \ldots, x^{\deg(p)-1}$ form a basis for the $\mathbb{F}$-vector space $\mathbb{F}[x]/\langle p \rangle$. This describes the additive structure on $\mathbb{F}[x]/\langle p \rangle$.

To describe its multiplicative structure, we only need to show how to write a product of monomials $x^a \cdot x^b$ with $a, b < \deg(p)$ in this basis. Suppose that p is monic with $\deg(p) = n$ and write $p(x) = x^n - q(x)$, where q has degree strictly less than p. Since $x^a \cdot x^b = (x^a \cdot x) \cdot x^{b-1}$, we may assume that $b = 1$. When $a < n$, we have $x^a \cdot x^1 = x^{a+1}$. When $a = n - 1$, then $x^{n-1} \cdot x^1 = x^n = q(x)$,

- Relate algebraic properties of $p(x)$ to properties of R, for example, zero divisors and domain.

- Prove that a field is a ring with only trivial ideals.

  Prove $I \subset J \subset R$ are ideals, then J/I is an ideal of R/I, and deduce that $R = \mathbb{F}[x]/p(x)$ is a field only if $p(x)$ is irreducible.

  Example $\mathbb{Q}[x]/(x^2 - 2)$ and explore $\mathbb{Q}(\sqrt{2})$.

  Example $\mathbb{R}[x]/(x^2 + 1)$ and show how it is isomorphic to $\mathbb{C}$.

  Work up to algebraically closed fields, the fundamental theorem of algebra (both over $\mathbb{C}$ and over $\mathbb{R}$).

  Explain that an algebraically closed field has no algebraic extensions (hence the name).

- Define the maximal ideal $\mathfrak{m}_a$ for $a \in \mathbb{A}^n$.

  Theorem A.1.9. *The maximal ideals of* $\mathbb{C}[x_1, \ldots, x_n]$ *all have the form* $\mathfrak{m}_a$ *for some* $a \in \mathbb{A}^n$.

A.2 Topology

Collect some topological statements here. Definition of topology, Closed/open duality, dense, nowhere dense... Describe the usual topology.

Recall that a function $f\colon X \to Y$ is continuous if and only if whenever $Z \subset Y$ is a closed set $f^{-1}(Z) \subset X$ is also closed.

A.3 Real algebra

In real algebra, the concept of an ordered field is central.

Definition A.3.1. A field K together with a relation $\leq$ on K is called an *ordered field* if

1. $\leq$ is a total order;
2. $a \leq b$ implies $a + c \leq b + c$;
3. $a \leq b$ and $0 \leq c$ implies $ac \leq bc$.

The order $\leq$ is completely determined by the *set of non-negative elements* $P = \{a \in K,\, a \geq 0\}$. Note that $P + P \subset P$, $PP \subset P$, $P \cap -P = \{0\}$ as well as $P \cup -P = K$. Conversely, any subset P of K satisfying these four conditions defines an ordered field via

$$a \leq b \iff b - a \in P\,.$$

This allows to identify P with an order.

The following result connecting preorderings (as defined in Section **??**) and orders was shown by Artin and Schreier [2]. See also, for example, [57, Theorem 1.4.4], [54, § 20, Theorem 1].

Theorem A.3.2. *Let T be a preordering on a field K which does not contain -1. Then there exists an order P of K such that $T \subseteq P$.*

...

Theorem A.3.3 (Tarski's transfer principle)**.** *Let $(\mathbb{F}, \leq)$ be an ordered field extension of $(\mathbb{R}, \leq)$. If there exists an $a \in \mathbb{F}^n$ satisfying some finite system of polynomial equations and inequalities with coefficients in $\mathbb{R}$, then there exists an $a \in \mathbb{R}^n$ satisfying the same equations and inequalities.*

Theorem A.3.4 (Quantifier elimination)**.** *Let $g, f_1, \ldots, f_m \in \mathbb{Z}[x_1, \ldots, x_n, y]$. Then there are $g_i, f_{ij} \in \mathbb{Z}[x_1, \ldots, x_n]$, $1 \leq i \leq l$, $1 \leq j \leq r_i$ (with $l, r_i \in \mathbb{N}$) such that for every real closed field R and for all $a \in R^n$*

$$\exists b \in R \left(g(a,b) = 0 \wedge \bigwedge_{j=1}^{m} f_j(a,b) > 0 \right) \iff \bigvee_{i=1}^{l} \left(g_i(a) = 0 \wedge \bigwedge_{j=1}^{t_i} f_{ij}(a) > 0 \right).$$

A.4 Positive semidefinite matrices

In the text we denote by $\mathrm{Sym}_n(\mathbb{R})$ the set of all symmetric $n \times n$-matrices. A matrix $A \in \mathrm{Sym}_n(\mathbb{R})$ is *positive semidefinite* if $x^T Ax \geq 0$ for all $x \in \mathbb{R}^n$ and it is *positive definite* if $x^T Ax > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$. By S_n^+ and S_n^{++} we denote the set of all positive semidefinite and positive definite matrices, respectively.

The following two statements characterize positive (semi)-definiteness from multiple viewpoints.

Theorem A.4.1. *For $A \in \mathbb{R}^{n\times n}$ the following statements are equivalent characterizations for the property $A \succeq 0$:*

1. *The smallest eigenvalue $\lambda_{\min}(A)$ of A is nonnegative.*
2. *All principal minors of A are nonnegative.*
3. *There exists an $L \in \mathbb{R}^{n\times n}$ with $A = LL^T$ (Choleski decomposition).*

Theorem A.4.2. *For $A \in \mathbb{R}^{n\times n}$ the following statements are equivalent characterizations for the property $A \succ 0$:*

1. *The smallest eigenvalue $\lambda_{\min}(A)$ of A is positive.*
2. *All principal minors of A are positive.*
3. *All leading principal minors of A (i.e., the determinants of the the submatrices $A_{\{1,\ldots,k\},\{1,\ldots,k\}}$) of A are positive.*
4. *There exists a non-singular matrix $L \in \mathbb{R}^{n\times n}$ with $A = LL^T$.*

Concerning the Choleski decomposition, let $A \in S_n^+$, and let $v^{(1)}, \ldots, v^{(n)}$ be an orthonormal system of eigenvectors with respect to the eigenvalues $\lambda_1, \ldots, \lambda_n$. Then

$$A \;=\; SDS^T \text{ with } S := (v^{(1)}, \ldots, v^{(n)}), \quad D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)\,.$$

For $A^{1/2} := \sum_{i=1}^n \sqrt{\lambda_i} v^{(i)} v^{(i)^T}$ we have $A^{1/2} \cdot A^{1/2} = A$, and $A^{1/2}$ is the only positive semidefinite matrix with this property.

For $A, B \in \mathbb{R}^{n\times n}$ we consider the inner product

$$\begin{aligned}\langle A, B\rangle \;&:=\; \mathrm{Tr}(A^T B) \;=\; \mathrm{Tr}(B^T A) \;=\; \mathrm{Tr}(AB^T) \;=\; \mathrm{Tr}(BA^T)\\ &=\; \mathrm{vec}(A)^T \mathrm{vec}(B)\,,\end{aligned}$$

where $\mathrm{vec}(A) := (a_{11}, a_{21}, \ldots, a_{n1}, a_{12}, a_{22}, \ldots, a_{nn})^T$ and Tr denotes the trace.

For $A \in \mathbb{R}^{n\times n}$ the definition $||A||_F^2 := \langle A, A\rangle = \mathrm{Tr}(A^T A) = \sum_{i,j=1}^n a_{ij}^2$ defines the *Frobenius norm* on $\mathbb{R}^{n\times n}$. If $A \in \mathrm{Sym}_n$ with eigenvalues $\lambda_1, \ldots, \lambda_n$, then $||A||_F^2 = \sum_{i=1}^n \lambda_i^2$.

Theorem A.4.3. (Féjer.) *A matrix $A \in S_n$ is positive semidefinite if and only if* $\mathrm{Tr}(AB) \geq 0$ *for all $B \in S_n^+$ (that is, S_n^+ is* self-dual*.*

We provide the illustrative proof of this statement.

Proof. Let $A \in S_n^+$ and $B \in S_n^+$. Then $\mathrm{Tr}(AB) = \mathrm{Tr}(A^{1/2}A^{1/2}B^{1/2}B^{1/2}) = \mathrm{Tr}(A^{1/2}B^{1/2}B^{1/2}A^{1/2})$. Since A and B are symmetric, this implies $\mathrm{Tr}(AB) = ||A^{1/2}B^{1/2}||_F^2 \geq 0$.

Conversely, let $A \in S_n$ and $\mathrm{Tr}(AB) \geq 0$ for all $B \in S_n^+$. Moreover, let $x \in \mathbb{R}^n$. For $B := xx^T \in S_n^+$ this implies $0 \leq \mathrm{Tr}(AB) = \mathrm{Tr}(Axx^T) = \sum_{i,j=1}^n a_{ij}x_ix_j = x^TAx$. □

Theorem A.4.4. (Schur complement.) *For $M = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$ with $A \succeq 0$ and C we have: M is positive (semi-)definite if and only if $C - B^TA^{-1}B$ is positive (semi-)definite. The matrix $C - B^TA^{-1}B$ is called the* Schur complement *of A in M.*

Proof. For $D := -A^{-1}B$ we have

$$\begin{pmatrix} I & 0 \\ D^T & I \end{pmatrix} \underbrace{\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}}_{=M^T=M} \begin{pmatrix} I & D \\ 0 & I \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & C - B^TA^{-1}B \end{pmatrix}.$$

The theorem now follows from the fact that a block diagonal matrix is positive (semi-)definite if and only if the diagonal blocks are positive (semi-)definite and from

$$X \succeq 0 \iff C^TXC \succeq 0 \text{ for all } C \in \mathbb{R}^{n\times n}.$$

□

A.5 Polyhedral geometry

polyhedron P

If $H = \{x \in \mathbb{R}^n : \sum_{i=1}^n a_ix_i = b\}$ is a supporting hyperplane to a face F of P with $P \subset \{x \in \mathbb{R}^n : \sum_{i=1}^n a_ix_i \leq b\}$ then $a = (a_1, \ldots, a_n)^T$ is an *outer normal vector to the face F*. The set of all outer normal vectors to F defines the *outer normal cone $C(F)$*.[1]

A *polyhedral complex* $\mathcal{C}$ is a finite family of polyhedra satisfying

1. $\emptyset \in \mathcal{C}$.
2. If $P \in \mathcal{C}$ then all faces of P are in $\mathcal{C}$.
3. The intersection $P \cap Q$ of two polyhedra $P, Q \in \mathcal{C}$ is a (possibly empty) face of P and of Q.

[1] zero vector ...

The elements of $\mathcal{C}$ are called *cells*. The *dimension* of $\mathcal{C}$ is the dimension of its maximal cells. $\mathcal{C}$ is called *pure* if all maximal faces have the same dimension.

A *fan* is a polyhedral complex whose cells are all cones. For a polyhedron P, the *normal fan* of P is defined by

$$\mathcal{N}(P) \;=\; \{C(F) \,:\, F \text{ is a face of } P\}\,.$$

polyhedral subdivision ...

A.6 Mixed volumes and mixed subdivisions

Mixed volumes are a classical, yet intriguing concept in convex geometry. As standard references, we refer to [80] as well as to [26, 29] for connections of mixed volumes and algebraic geometry.

The *Minkowski sum* of two sets $A_1, A_2 \subset \mathbb{R}^n$ is defined as

$$A_1 + A_2 = \{a_1 + a_2 \,:\, a_1 \in A_1, a_2 \in A_2\} \;.$$

Let $K_1, \ldots, K_n$ be convex bodies in $\mathbb{R}^n$ and $\lambda_1, \ldots, \lambda_n$ be non-negative real parameters. Then the function $\mathrm{vol}_n(\lambda_1 K_1 + \ldots + \lambda_n K_n)$ is a homogeneous polynomial of degree n in $\lambda_1, \ldots, \lambda_n$, where $+$ denotes the Minkowski sum and vol_n denotes the n-dimensional volume. The function $\mathrm{vol}_n(\lambda_1 K_1 + \cdots + \lambda_n K_n)$ is called the *mixed volume* of $K_1, \ldots, K_n$ and is denoted by $\mathrm{MV}_n(P_1, \ldots, P_n)$. We remark that some authors prefer to factor out $n!$ in the definition of the mixed volume. We prefer to keep that factor since this scaling yields integer values for the mixed volumes of polytopes with integral vertices.

The mixed volume can be expressed as an alternating sum in conventional volumes,

$$\mathrm{MV}_n(K_1, \ldots, K_n) \;=\; \sum_{j=1}^{n} (-1)^j \sum_{I \subset \{1,\ldots,n\}, |I|=j} \mathrm{vol}_n\Big(\sum_{i \in I} K_i\Big)\,.$$

Mixed volumes are always non-negative, and they are monotone with respect to inclusion, i.e.

$$\mathrm{MV}(K_1, \ldots, K_n) \;\le\; \mathrm{MV}(K_1', \ldots, K_n') \quad \text{if } K_1 \subset K_1' \text{ for all } i \in \{1, \ldots, n\}\,.$$

Moreover, $\mathrm{MV}(K_1, \ldots, K_n)$ is strictly positive if and only if there exist segments $S_i \subset K_i$ (for all i) whose directions are linearly independent. The mixed volume is invariant under permutation of its arguments and is linear in each argument, i.e.

$$\mathrm{MV}_n(\ldots, \alpha P_i + \beta P_i', \ldots) \;=\; \alpha\, \mathrm{MV}_n(\ldots, P_i, \ldots) + \beta\, \mathrm{MV}_n(\ldots, P_i', \ldots)\,.$$

It generalizes the usual volume in the sense that $\mathrm{MV}_n(P, \ldots, P) \;=\; n!\,\mathrm{vol}_n(P)$ holds.

We denote by $\mathrm{MV}_n(P_1, d_1; \ldots; P_k, d_k)$ the mixed volume where P_i is taken d_i times and $\sum_{i=1}^{k} d_i = n$.

Mixed subdivisions. Let $\mathcal{S} = (S^{(1)}, \ldots, S^{(m)})$ be a sequence of finite point sets in $\mathbb{R}^n$ which affinely span $\mathbb{R}^n$. A sequence $C^{(1)}, \ldots, C^{(m)}$ of subset $C^{(i)} \subset S^{(i)}$ is called a *cell* of $\mathcal{S}$. A *subdivision* of $\mathcal{S}$ is a collection $\Gamma = (C_1, \ldots, C_k)$ of cells such that

1. $\dim \operatorname{conv} C_i = n$ for all cells C_i,
2. $\operatorname{conv} C_i \cap \operatorname{conv} C_j$ is a face of both convex hulls and
3. $\bigcup_{i=1}^{k} \operatorname{conv} C_i = \operatorname{conv} \mathcal{S}$,

where $\operatorname{conv} A := \operatorname{conv}(A^{(1)} + \cdots + A^{(m)})$ for a sequence A of point sets. A subdivision is called *mixed* if additionally $\sum_{i=1}^{m} \dim \operatorname{conv} C_j^{(i)} = n$ for all cells C_j in Γ, and it is called fine mixed if moreover $\sum_{i=1}^{m} (|C_j^{(i)}| - 1) = n$ for all cells C_j in Γ. The *type* of a cell is defined as

$$\operatorname{type}(C) \;=\; (\dim \operatorname{conv} C^{(1)}, \ldots, \dim \operatorname{conv} C^{(m)})\,,$$

and cells of type $(d_1, \ldots, d_m)$ with $d_i \geq 1$ for all i are called *mixed cells*.

We remark that all concepts can be transferred from point sets to polytopes P_i by considering their vertex sets $\operatorname{vert}(P_i)$ as the point sets above. By abuse of notation we then speak of a subdivision of $P := P_1 + \cdots + P_k$ meaning a mixed subdivision of $(\operatorname{vert}(P_1), \ldots, \operatorname{vert}(P_m))$. As cells of such a subdivision we always consider Minkowski sums of faces $F_1 + \cdots + F_m$ where F_i is a face of P_i. If all cells of a subdivision Γ of $P_1 + \cdots + P_m$ are simplices then Γ is called a triangulation.

With this terminology the mixed volume can be calculated by

$$\operatorname{MV}_n(P_1, d_1; \ldots; P_k, d_k) \;=\; \sum_{C} d_1! \cdots d_k! \operatorname{vol}_n(C)$$

where the sum is over all cells C of type $(d_1, \ldots, d_k)$ in an arbitrary mixed subdivision of $P_1 + \cdots + P_k$.

Bibliography

[1] William W. Adams and Philippe Loustaunau, *An introduction to Gröbner bases*, Graduate Studies in Mathematics, vol. 3, American Mathematical Society, Providence, RI, 1994. MR 1287608 (95g:13025)

[2] E. Artin and O. Schreier, *Algebraische Konstruktion reeller Körper.*, Abhandlungen Hamburg **5** (1926), 85–99 (German).

[3] G. Averkov, *Constructive proofs of some positivstellens*, arXiv preprint arXiv:1201.4066, 2012.

[4] Saugata Basu, Richard Pollack, and Marie-Françoise Roy, *Algorithms in real algebraic geometry*, second ed., Algorithms and Computation in Mathematics, vol. 10, Springer-Verlag, Berlin, 2006. MR 2248869 (2007b:14125)

[5] Mauro C. Beltrametti, Ettore Carletti, Dionisio Gallarati, and Giacomo Monti Bragadin, *Lectures on curves, surfaces and projective varieties*, EMS Textbooks in Mathematics, European Mathematical Society (EMS), Zürich, 2009, A classical view of algebraic geometry, Translated from the 2003 Italian original by Francis Sullivan. MR 2549804 (2010k:14001)

[6] C. Berg, J. P. R. Christensen, and C. U. Jensen, *A remark on the multidimensional moment problem*, Math. Ann. **243** (1979), no. 2, 163–169. MR 543726 (81e:44008)

[7] Christian Berg, Jens Peter Reus Christensen, and Paul Ressel, *Harmonic analysis on semigroups*, Graduate Texts in Mathematics, vol. 100, Springer-Verlag, New York, 1984, Theory of positive definite and related functions. MR 747302 (86b:43001)

[8] S. Bernštein, *Sur la représentation des polynomes positifs.*, Charĭkov, Comm. Soc. Math. (2) **14** (1915), 227–228 (French).

[9] B. Bertrand and F. Bihan, *Euler characteristic of real nondegenerate tropical complete intersections*, Preprint, `arXiv:math/0710.1222`, 2007.

[10] Etienne Bézout, *Théorie générale des équations algébriques*, Ph.-D. Pierres, 1779.

[11] ———, *General theory of algebraic equations*, Princeton University Press, 2006, Translated from French original by Eric Feron.

[12] R. Bieri and J.R.J. Groves, *The geometry of the set of characters induced by valuations*, J. Reine Angew. Math. **347** (1984), 168–195.

[13] Jacek Bochnak, Michel Coste, and Marie-Françoise Roy, *Real algebraic geometry*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], vol. 36, Springer-Verlag, Berlin, 1998, Translated from the 1987 French original, Revised by the authors. MR 1659509 (2000a:14067)

[14] Ludwig Bröcker, *Spaces of orderings and semialgebraic sets*, Quadratic and Hermitian forms (Hamilton, Ont., 1983), CMS Conf. Proc., vol. 4, Amer. Math. Soc., Providence, RI, 1984, pp. 231–248. MR MR776457 (86m:12002)

[15] B. Buchberger, *Ein algorithmisches Kriterium für die Lösbarkeit eines algebraischen Gleichungssystems*, Aequationes Math. **4** (1970), 374–383.

[16] Bruno Buchberger, *Ein Algorithmus zum Auffinden der Basiselemente des Restklassenrings nach einem nulldimensionalen Polynomideal*, Ph.D. thesis, Universität Innsbruck, 1965.

[17] Bruno Buchberger, *An algorithm for finding the basis elements of the residue class ring of a zero dimensional polynomial ideal*, J. Symbolic Comput. **41** (2006), no. 3-4, 475–511, Translated from the 1965 German original by Michael P. Abramson.

[18] Angelika Bunse-Gerstner, Ralph Byers, and Volker Mehrmann, *Numerical methods for simultaneous diagonalization*, SIAM J. Matrix Anal. Appl. **14** (1993), no. 4, 927–949. MR 1238912 (94h:65036)

[19] Man Duen Choi and Tsit Yuen Lam, *Extremal positive semidefinite forms*, Math. Ann. **231** (1977/78), no. 1, 1–18. MR 0498384 (58 #16512)

[20] David Cox, John Little, and Donal O'Shea, *Ideals, varieties, and algorithms*, third ed., Undergraduate Texts in Mathematics, Springer, New York, 2007, An introduction to computational algebraic geometry and commutative algebra.

[21] David A. Cox, John Little, and Donal O'Shea, *Using algebraic geometry*, second ed., Graduate Texts in Mathematics, vol. 185, Springer, New York, 2005.

[22] Etienne de Klerk, *Aspects of semidefinite programming*, Applied Optimization, vol. 65, Kluwer Academic Publishers, Dordrecht, 2002, Interior point algorithms and selected applications. MR 2064921 (2005a:90001)

[23] Alicia Dickenstein and Ioannis Z. Emiris (eds.), *Solving polynomial equations*, Algorithms and Computation in Mathematics, vol. 14, Springer-Verlag, Berlin, 2005, Foundations, algorithms, and applications. MR 2161984 (2008d:14095)

[24] Jan Draisma, *A tropical approach to secant dimensions*, J. Pure Appl. Algebra **212** (2008), no. 2, 349–363. MR MR2357337 (2008j:14102)

[25] David Eisenbud, *Commutative algebra*, Graduate Texts in Mathematics, vol. 150, Springer-Verlag, New York, 1995, With a view toward algebraic geometry. MR 1322960 (97a:13001)

[26] G. Ewald, *Combinatorial Convexity and Algebraic Geometry*, Graduate Texts in Mathematics, vol. 168, Springer-Verlag, New York, 1996.

[27] J.-C. Faugère, *FGb*, See `http://fgbrs.lip6.fr/jcf/Software/FGb/index.html`.

[28] J. C. Faugère, P. Gianni, D. Lazard, and T. Mora, *Efficient computation of zero-dimensional Gröbner bases by change of ordering*, J. Symbolic Comput. **16** (1993), no. 4, 329–344.

[29] William Fulton, *Introduction to toric varieties*, Annals of Mathematics Studies, vol. 131, Princeton University Press, Princeton, NJ, 1993, The William H. Roever Lectures in Geometry. MR MR1234037 (94g:14028)

[30] J. Gouveia and T. Netzer, *Positive polynomials and projections of spectrahedra*, Preprint, arXiv:math.0911.2750, 2010.

[31] 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`.

[32] W. Habicht, *Über die Zerlegung strikter definiter Formen in Quadrate.*, Comment. Math. Helv. **12** (1940), 317–322 (German).

[33] David Handelman, *Representing polynomials by positive linear functions on compact convex polyhedra*, Pacific J. Math. **132** (1988), no. 1, 35–62. MR 929582 (90e:52005)

[34] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge Mathematical Library, Cambridge University Press, Cambridge, 1988, Reprint of the 1952 edition. MR 944909 (89d:26016)

[35] Joe Harris, *Algebraic geometry*, Graduate Texts in Mathematics, vol. 133, Springer-Verlag, New York, 1992, A first course. MR 1182558 (93j:14001)

[36] Robin Hartshorne, *Algebraic geometry*, Springer-Verlag, New York, 1977, Graduate Texts in Mathematics, No. 52. MR 0463157 (57 #3116)

[37] B. Helton and V. Vinnikov, *Linear matrix inequality representation of sets*, Communications on Pure and Applied Mathematics **60** (2007), 654–674.

[38] Raymond Hemmecke and Peter Malkin, *Computing generating sets of lattice ideals*, `math.CO/0508359`.

[39] H. Hironaka, *Resolution of singularities of an algebraic variety over a field of characteristic zero*, Ann. Math. **79** (1964), 109–326.

[40] Audun Holme, *A royal road to algebraic geometry*, Springer, Heidelberg, 2012. MR 2858123

[41] Serkan Hoşten and Bernd Sturmfels, *GRIN: an implementation of Gröbner bases for integer programming*, Integer programming and combinatorial optimization (Copenhagen, 1995), Lecture Notes in Comput. Sci., vol. 920, Springer, Berlin, 1995, pp. 267–276.

[42] Klaus Hulek, *Elementary algebraic geometry*, Student Mathematical Library, vol. 20, American Mathematical Society, Providence, RI, 2003, Translated from the 2000 German original by Helena Verrill. MR 1955795 (2003m:14002)

[43] Thomas Jacobi and Alexander Prestel, *Distinguished representations of strictly positive polynomials*, J. Reine Angew. Math. **532** (2001), 223–235. MR 1817508 (2001m:14080)

[44] Samuel Karlin and William J. Studden, *Tchebycheff systems: With applications in analysis and statistics*, Pure and Applied Mathematics, Vol. XV, Interscience Publishers John Wiley & Sons, New York-London-Sydney, 1966. MR 0204922 (34 #4757)

[45] Eric Katz, *A tropical toolkit*, Expo. Math. **27** (2009), no. 1, 1–36. MR MR2503041 (2010f:14069)

[46] G. Kempf, Finn Faye Knudsen, D. Mumford, and B. Saint-Donat, *Toroidal embeddings. I*, Lecture Notes in Mathematics, Vol. 339, Springer-Verlag, Berlin, 1973. MR MR0335518 (49 #299)

[47] J.-L. Krivine, *Anneaux préordonnés*, J. Analyse Math. **12** (1964), 307–326. MR MR0175937 (31 #213)

[48] Jean B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim. **11** (2000/01), no. 3, 796–817 (electronic). MR 1814045 (2002b:90054)

[49] Jean Bernard Lasserre, *Moments, positive polynomials and their applications*, Imperial College Press Optimization Series, vol. 1, Imperial College Press, London, 2010. MR 2589247

[50] Monique Laurent, *Revisiting two theorems of Curto and Fialkow on moment matrices*, Proc. Amer. Math. Soc. **133** (2005), no. 10, 2965–2976 (electronic). MR 2159775 (2006d:47027)

[51] ———, *Semidefinite representations for finite varieties*, Math. Program. **109** (2007), no. 1, Ser. A, 1–26. MR 2291590 (2008g:90094)

[52] ———, *Sums of squares, moment matrices and optimization over polynomials*, Emerging applications of algebraic geometry, IMA Vol. Math. Appl., vol. 149, Springer, New York, 2009, pp. 157–270. MR 2500468 (2010j:13054)

[53] T. Y. Li, Tim Sauer, and J. A. Yorke, *The cheater's homotopy: an efficient procedure for solving systems of polynomial equations*, SIAM J. Numer. Anal. **26** (1989), no. 5, 1241–1251.

[54] Falko Lorenz, *Algebra. Vol. II*, Universitext, Springer, New York, 2008, Fields with structure, algebras and advanced topics, Translated from the German by Silvio Levy, With the collaboration of Levy. MR 2371763 (2008k:12001)

[55] F.S. Macaulay, *Some properties of enumeration in the theory of modular systems*, Proc. London Math. Soc. **26** (1927), 531–555.

[56] T. Markwig, *A field of generalised puiseux series for tropical geometry*, Rend. Semin. Mat. Torino **xx** (xx), xx.

[57] Murray Marshall, *Positive polynomials and sums of squares*, Mathematical Surveys and Monographs, vol. 146, American Mathematical Society, Providence, RI, 2008. MR 2383959 (2009a:13044)

[58] V. P. Maslov, *On a new superposition principle for optimization problem*, Séminaire sur les équations aux dérivées partielles, 1985–1986, École Polytech., Palaiseau, 1986, pp. Exp. No. XXIV, 14. MR MR874583

[59] E. Meissner, *Über positive Darstellungen von Polynomen*, Math. Ann. **70** (1911), no. 2, 223–235. MR 1511619

[60] Grigory Mikhalkin, *Enumerative tropical algebraic geometry in* $\mathbb{R}^2$, J. Amer. Math. Soc. **18** (2005), no. 2, 313–377. MR MR2137980 (2006b:14097)

[61] T. Netzer, D. Plaumann, and M. Schweighofer, *Exposed faces of semidefinitely representable sets*, SIAM J. Optim. **20** (2010), 1944–1955.

[62] J. Nie, P. A. Parrilo, and B. Sturmfels, *Semidefinite representation of the k-ellipse*, Algorithms in algebraic geometry (A. Sommesse A. Dickenstein, F.-O.S̃chreyer, ed.), The IMA Volumes in Mathematics and its Applications, vol. 146, Springer, New York, 2008, pp. 117–132.

[63] M.-F. Roy P. Pedersen and A. Szpirglas, *Counting real zeros in the multivariate case*, Computational algebraic geometry (Nice, 1992), Progr. Math., vol. 109, Birkhäuser Boston, Boston, MA, 1993, pp. 203–224. MR 1230868 (94m:14075)

[64] P.A. Parrilo, *An explicit construction of distinguished representations of polynomials nonnegative over finite sets*, ETH Zürich, IfA Technical Report AUT02-02, 2002.

[65] Pablo A. Parrilo, *Semidefinite programming relaxations for semialgebraic problems*, Math. Program. **96** (2003), no. 2, Ser. B, 293–320, Algebraic and geometric methods in discrete optimization. MR 1993050 (2004g:90075)

[66] Pablo A. Parrilo and Bernd Sturmfels, *Minimizing polynomial functions*, Algorithmic and quantitative real algebraic geometry (Piscataway, NJ, 2001), DIMACS Ser. Discrete Math. Theoret. Comput. Sci., vol. 60, Amer. Math. Soc., Providence, RI, 2003, pp. 83–99. MR 1995016 (2004e:13038)

[67] Daniel Perrin, *Algebraic geometry*, Universitext, Springer-Verlag London Ltd., London, 2008, An introduction, Translated from the 1995 French original by Catriona Maclean. MR 2372337 (2008k:14001)

[68] H. Poincaré, *Sur les équations algébriques.*, C. R. XCVII, (1884), 1418–1419 (French).

[69] G. Pólya, *Über positive darstellung von polynomen*, Vierteljahresschrift der Naturforschenden Gesellschaft in Zürich **73** (1928), 141–145, Reprinted in: Collected Papers, Volume 2, 309–313, MIT Press, Cambridge.

[70] G. Pólya and G. Szegő, *Problems and theorems in analysis. Vol. II*, german ed., Springer-Verlag, New York, 1976, Theory of functions, zeros, polynomials, determinants, number theory, geometry, Die Grundlehren der Mathematischen Wissenschaften, Band 216. MR 0396134 (53 #2)

[71] Victoria Powers and Bruce Reznick, *Polynomials that are positive on an interval*, Trans. Amer. Math. Soc. **352** (2000), no. 10, 4677–4692. MR 1707203 (2001b:12002)

[72] Alexander Prestel and Charles N. Delzell, *Positive polynomials*, Springer Monographs in Mathematics, Springer-Verlag, Berlin, 2001, From Hilbert's 17th problem to real algebra. MR 1829790 (2002k:13044)

[73] M. Ramana and A.J. Goldman, *Some geometric results in semidefinite programming*, Journal of Global Optimization **7** (1995), 33–50.

[74] Bruce Reznick, *Extremal PSD forms with few terms*, Duke Math. J. **45** (1978), no. 2, 363–374. MR 0480338 (58 #511)

[75] J. Richter-Gebert, B. Sturmfels, and T. Theobald, *First steps in tropical geometry*, Idempotent Mathematics and Mathematical Physics, Contemp. Math., vol. 377, Amer. Math. Soc., Providence, RI, 2005, pp. 289–317.

[76] R. Tyrrell Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, N.J., 1970. MR 0274683 (43 #445)

[77] Walter Rudin, *Real and complex analysis*, 3 ed., McGraw-Hill Book Co., 1987.

[78] Konrad Schmüdgen, *An example of a positive polynomial which is not a sum of squares of polynomials. A positive, but not strongly positive functional*, Math. Nachr. **88** (1979), 385–390. MR 543417 (81b:12024)

[79] ———, *The K-moment problem for compact semi-algebraic sets*, Math. Ann. **289** (1991), no. 2, 203–206. MR 1092173 (92b:44011)

[80] Rolf Schneider, *Convex bodies: the Brunn-Minkowski theory*, Encyclopedia of Mathematics and its Applications, vol. 44, Cambridge University Press, Cambridge, 1993. MR MR1216521 (94d:52007)

[81] A. Schrijver, *Theory of Linear and Integer Programming*, Wiley-Interscience Series in Discrete Mathematics, John Wiley & Sons Ltd., Chichester, 1986. MR 874114 (88m:90090)

[82] Markus Schweighofer, *An algorithmic approach to Schmüdgen's Positivstellensatz*, J. Pure Appl. Algebra **166** (2002), no. 3, 307–319. MR 1870623 (2002j:14063)

[83] ———, *Optimization of polynomials on compact semialgebraic sets*, SIAM J. Optim. **15** (2005), no. 3, 805–825 (electronic). MR 2142861 (2006d:90136)

[84] Igor R. Shafarevich, *Basic algebraic geometry. 1*, second ed., Springer-Verlag, Berlin, 1994, Varieties in projective space, Translated from the 1988 Russian edition and with notes by Miles Reid. MR 1328833 (95m:14001)

[85] Karen E. Smith, Lauri Kahanpää, Pekka Kekäläinen, and William Traves, *An invitation to algebraic geometry*, Universitext, Springer-Verlag, New York, 2000. MR 1788561 (2001k:14002)

[86] Andrew J. Sommese and Charles W. Wampler, II, *The numerical solution of systems of polynomials*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2005, Arising in engineering and science.

[87] F. Sottile, *Enumerative real algebraic geometry*, Algorithmic and quantitative real algebraic geometry (Piscataway, NJ, 2001), DIMACS Ser. Discrete Math. Theoret. Comput. Sci., vol. 60, Amer. Math. Soc., Providence, RI, 2003, pp. 139–179. MR 1995019 (2004j:14065)

[88] F. Sottile and T. Theobald, *Line problems in nonlinear computational geometry*, Surveys on discrete and computational geometry, Contemp. Math., vol. 453, Amer. Math. Soc., Providence, RI, 2008, pp. 411–432. MR 2405690 (2010a:14096)

[89] Gilbert Stengle, *A nullstellensatz and a positivstellensatz in semialgebraic geometry*, Math. Ann. **207** (1974), 87–97. MR MR0332747 (48 #11073)

[90] B. Sturmfels, *Solving systems of polynomial equations*, CBMS Regional Conference Series in Mathematics, vol. 97, Published for the Conference Board of the Mathematical Sciences, Washington, DC, 2002.

[91] B. Sturmfels and J. Tevelev, *Elimination theory for tropical varieties*, Math. Res. Lett. **15** (2008), no. 3, 543–562. MR MR2407231 (2009f:14124)

[92] Lieven Vandenberghe and Stephen Boyd, *Semidefinite programming*, SIAM Rev. **38** (1996), no. 1, 49–95. MR 1379041 (96m:90005)

[93] M. D. Vigeland, *Tropical complete intersection curves.*, Preprint, `arXiv:math/0711.1962`, 2007.

[94] Oleg Viro, *Dequantization of real algebraic geometry on logarithmic paper*, European Congress of Mathematics, Vol. I (Barcelona, 2000), Progr. Math., vol. 201, Birkhäuser, Basel, 2001, pp. 135–146. MR MR1905317 (2003f:14067)

[95] Robert J. Walker, *Algebraic curves*, Dover Publications Inc., New York, 1962. MR MR0144897 (26 #2438)

[96] G.M. Ziegler, *Lectures on polytopes*, Springer-Verlag, New York, 1995.