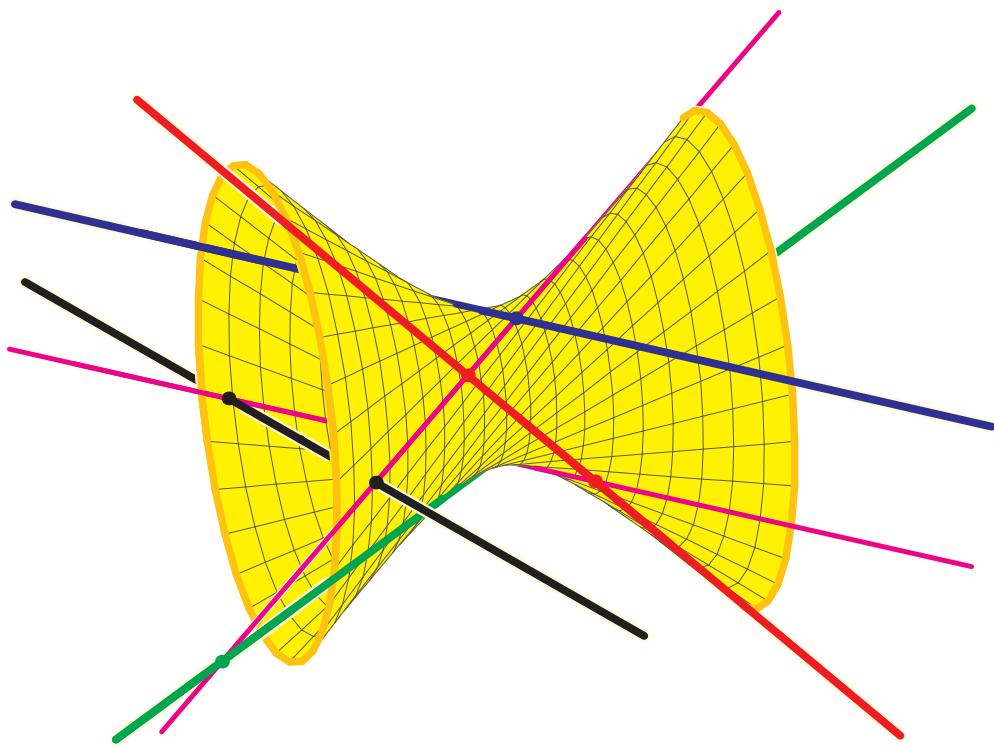


Applicable Algebraic Geometry

Frank Sottile
Thorsten Theobald



December 26, 2018

Contents

1 Varieties	11
1.1 Affine varieties	11
1.2 The algebra-geometry dictionary I	17
1.3 The algebra-geometry dictionary II	24
1.4 Projective varieties	30
1.5 Maps of projective varieties	39
1.6 Notes	46
2 Symbolic algorithms	47
2.1 Resultants and Bézout’s Theorem	47
2.2 Gröbner basics	58
2.3 Algorithmic aspects of Gröbner bases	67
2.4 Solving equations with Gröbner bases	74
2.5 Solving equations with linear algebra	84
2.6 Notes	89
3 Structure of varieties	91
3.1 Generic properties of varieties	91
3.2 Unique factorization for varieties	97
3.3 Rational functions and maps	104
3.4 Smooth and singular points	109
3.5 Hilbert functions and dimension	115
3.6 Bertini Theorems	125
3.7 Notes	125
4 Numerical Algebraic Geometry	127
4.1 Core Numerical Algorithms	127
4.2 Numerical Homotopy Continuation	137
4.3 Numerical Algebraic Geometry	148
4.4 Numerical Irreducible Decomposition	157
4.5 Smale’s α -theory	160
4.6 Notes	161

5 Real algebraic and semialgebraic geometry	163
5.1 Real roots of univariate polynomials	163
5.2 Univariate stable polynomials	169
5.3 Real roots and the trace form	176
5.4 Semialgebraic sets and polyhedra	179
5.5 Spectrahedra	185
5.6 Stable and hyperbolic polynomials	189
5.7 Notes	195
6 Positive polynomials	197
6.1 Nonnegative univariate polynomials	198
6.2 Positive polynomials and sums of squares	201
6.3 Hilbert's 17th problem	209
6.4 Systems of polynomial inequalities	212
6.5 The Positivstellensatz	217
6.6 Theorems of Pólya and Handelman	221
6.7 Representation theorems	225
6.8 Notes	231
7 Polynomial optimization	233
7.1 Linear programming relaxations	234
7.2 Unconstrained optimization and sums of squares	239
7.3 Semidefinite programming	242
7.4 Unconstrained optimization and semidefinite programming	247
7.5 Duality and the moment problem	250
7.6 Optimization over compact sets	255
7.7 Finite convergence and detecting optimality	261
7.8 Notes	267
8 Toric Varieties	269
8.1 Toric Ideals and Affine Toric Varieties	269
8.2 Projective Toric Varieties	275
8.3 Kushnirenko's Theorem	281
8.4 Toric Degenerations and regular subdivisions	289
8.5 Real Toric Varieties	290
8.6 Bernstein's Theorem and Polyhedral Homotopies	290
8.7 Notes	299
9 Tropical geometry	301
9.1 Tropical hypersurfaces	301
9.2 Tropical prevarieties and stable intersections	310
9.3 Amoebas	314
9.4 Valuations and Kapranov's Theorem	320

CONTENTS	5
----------	---

9.5 Tropical varieties	327
9.6 Tropical bases	331
9.7 Notes	337
10 Non-Linear Computational Geometry	339
10.1 Grassmann varieties and Plücker coordinates	339
10.2 Duality and the Plücker relations	342
10.3 Case study: Lines tangent to spheres	349
10.4 The Stewart platform and robotics	359
10.5 Notes	362
A Appendix	363
A.1 Algebra	363
A.1.1 Fields and rings	363
A.1.2 Fields and polynomials	366
A.1.3 Polynomials in one variable	367
A.1.4 Multilinear algebra	370
A.1.5 Real algebra	371
A.2 Topology	372
A.3 Convex geometry	373
A.3.1 Polytopes and polyhedra	373
A.3.2 Minkowski sum and mixed volumes	376
A.3.3 Positive semidefinite matrices	380
A.4 Complex analysis	382
A.5 Moments	382

Notation and conventions

Global:

$\mathbb{R}, \mathbb{C}, \mathbb{Q}$	real, complex, rational numbers
\mathbb{R}_+^n	nonnegative orthant
Use \mathbb{N}, \mathbb{Z}	Natural numbers ($0 \in \mathbb{N}!$), integers
\subset	(not necess. strict) subset, (\subseteq has to be avoided)
\mathbb{K}	ground field
n, m	natural numbers
x_1, \dots, x_n	usual ring variables (lowercase letters, n variables)
$(a_1, \dots, a_n), b, c$	constants from field, sometime points in \mathbb{K}^n
$\mathbb{K}[x_1, \dots, x_n]$	polynomial ring in x_1, \dots, x_n over \mathbb{K} We use the abbreviation $\mathbb{K}[x]$ in later chapters.
X, Y, Z	Varieties
I	ideal
$\mathcal{V}, \mathcal{V}_{\mathbb{R}}$	variety, real variety
$\mathcal{I}(S)$	ideal of set S of polynomials
$\mathbb{P}_{\mathbb{C}}^n, \mathbb{P}_{\mathbb{R}}^n, \mathbb{P}_{\mathbb{K}}^n$	n -dim. projective space over \mathbb{C} , over \mathbb{R} , over \mathbb{K}
\mathbb{P}^n	n -dim. projective space (over \mathbb{C}), short version
$\mathbb{P}(V)$	alternative notation, used in $\mathbb{P}(\bigwedge^k (\mathbb{C}^n)^*)$ (for Plücker coordinates)
x^α	for a monomial Need to be careful, as α collides with Smale's α-theory.
\mathcal{B}	Basis of standard monomials (Ch.2)
$\text{conv}(X)$	The convex hull of a set $X \subset \mathbb{R}^n$

Chapter Ch:One: Varieties b-deg($x^\alpha y^\beta$) bi-degree of monomial $x^\alpha y^\beta$ and of polynomials

Chapter 6: Positive polynomials

$\Sigma[x_1, \dots, x_n]$	sums of squares in $x = (x_1, \dots, x_n)$
$\Sigma_{n,d}$	homogeneous sums of squares in $x = (x_1, \dots, x_n)$ of degree d
$\mathcal{P}_{n,d}$	homogeneous non-negative polynomials in $x = (x_1, \dots, x_n)$ of degree d
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 8 Toric Varieties

$\text{Aff}(X)$	The affine span of a set $\subset \mathbb{R}^n$
Σ	Fan in a real vector space
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)
$\text{int } P, \text{relint } P$	interior, relative interior

Chapter 8 Toric Varieties

\mathbb{C}^\times non-zero complex numbers (Do not use \mathbb{C}^* as that clashes with dual of a cone).

Introduction

Recent years have seen many developments in applying methods and ideas from algebraic geometry within various other disciplines inside and outside mathematics. The purpose of this book is to provide a broad but friendly introduction to *applicable algebraic geometry*, where the potential of applicability is to be seen in a broad range of contexts and in our treatment often involves *combinatorial*, *algorithmic*, or *real* aspects. 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 change 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 mathematics 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.

In 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 introduces to symbolic techniques and algorithms for handling polynomial equations. We discuss the primary concepts of the theory of Gröbner bases, which have been introduced by Buchberger in 1965. Gröbner bases facilitate to establish a unique basis of a given polynomial ideal and have provided a landmark for the development of symbolic computation. They 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 eigenvalue technique to study the complex and real roots of a zero-dimensional ideals.

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 based on the Hilbert function and the Hilbert polynomial.

Chapter 4 discusses the viewpoint of numerical algebraic geometry. Starting from some core routines of numerical analysis, homotopy continuation methods provide a method to numerically compute the solutions of systems of polynomial equations. Building upon that framework, we then present some more advanced techniques, such as witness sets.

Chapter 5 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. We then deal with some recent developments. Firstly, we introduce to semialgebraic sets which arise as linear matrix inequalities, so-called spectrahedra, or as the projections of linear matrix inequalities and which are particularly useful with regard to computational handling. Secondly, we discuss the classes of stable and hyperbolic polynomials.

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's 17th problem from his famous list of 23 problems from 1900. Statements which certify the emptiness of set defined by real polynomial equations or inequalities can be seen as analogues to Hilbert's Nullstellensatz in the algebraically closed

case. Chapter 6 begins by discussing univariate positive polynomials and then treats the connection between positive polynomials and sums of squares. We then deal with some powerful theorems, such as the 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 7 on polynomial optimization. From the dual point of view, positive polynomials relate to the rich and classical world of moment problems.

Chapter 7 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 matrices. The chapter both provides theoretical as well as practical issues of sums of squares based relaxation schemes for polynomial optimization.

Chapter 8 deals with toric varieties which provide an important link between algebraic geometry and combinatorics. Building upon the notion of a toric variety, we give a detailed presentation of Bernstein’s Theorem which provides a tight upper bound on the number of isolated solutions of a sparse polynomial system in terms of the mixed volume of the underlying Newton polytopes.

Chapter 9 provides an access to the field of tropical geometry. In recent times, 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, Viro’s patchworking method, in the Maslov dequantization of real numbers as well as the use of idempotent semiring in optimization and control theory. We first concentrate on the combinatorial viewpoint and on tropical hypersurfaces, discussing the dual subdivision and polyhedral characterizations. Then we discuss in detail Kapranov’s theorem which – for the case of hypersurfaces – connects the two distinct viewpoints on tropical geometry stated above. We then introduce 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 10 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 geometry and by considering some aspects of the Stewart 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 practitioners in pure and applied mathematics as well inclined 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 [25].

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 5–7 (which depend sequentially on each other) for a focus on semialgebraic geometry and optimization, Chapters 8–10 for a focus on discrete aspects, and Chapter 10 for a glimpse into 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 include 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 concerns the interplay of algebra and geometry, with its two most basic objects the ring $\mathbb{K}[x_1, \dots, x_n]$ of polynomials in variables x_1, \dots, x_n with coefficients in \mathbb{K} , and the space \mathbb{K}^n of n -tuples $a = (a_1, \dots, a_n)$ of numbers from \mathbb{K} , called *affine space*. Evaluating a polynomial $f \in \mathbb{K}[x_1, \dots, x_n]$ at points of \mathbb{K}^n defines a function $f: \mathbb{K}^n \rightarrow \mathbb{K}$ on affine space. We use these polynomial functions to define our primary object of interest. We will often abbreviate $\mathbb{K}[x_1, \dots, x_n]$ as $\mathbb{K}[x]$, when it is clear from the context that we are working with multivariate polynomials (and not univariate polynomials).

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

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

This is a *subvariety* of \mathbb{K}^n or simply a *variety* or (*affine*) *algebraic variety*. When S consists of a single polynomial f , then $\mathcal{V}(S) = \mathcal{V}(f)$ is called a *hypersurface*.

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 $\mathcal{V}(S) \supset \mathcal{V}(T)$.

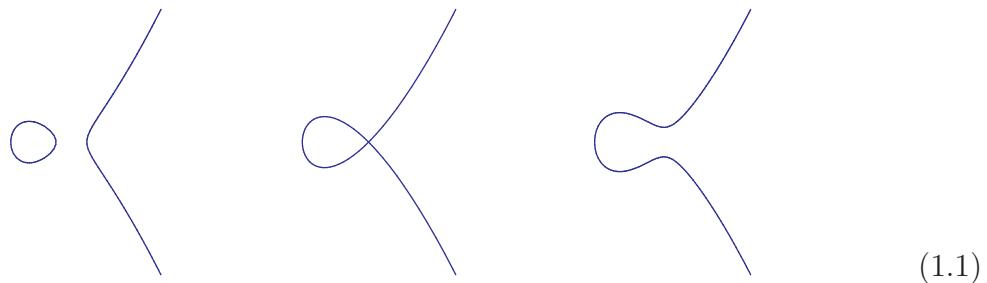
The empty set $\emptyset = \mathcal{V}(1)$ and affine space itself $\mathbb{K}^n = \mathcal{V}(0)$ are varieties. Any linear or affine subspace L of \mathbb{K}^n is a variety. Indeed, an affine subspace L has an equation $Ax = b$, where A is a matrix and b is a vector, and so $L = \mathcal{V}(Ax - b)$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

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 at points of 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 real plane cubic curves $\mathcal{V}(f + \frac{1}{20})$, $\mathcal{V}(f)$, and $\mathcal{V}(f - \frac{1}{20})$, where $f(x, y) := y^2 - x^2 - x^3$.



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 (a formal definition of smooth variety is given in Section 3.4). 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 intersection of an ellipse

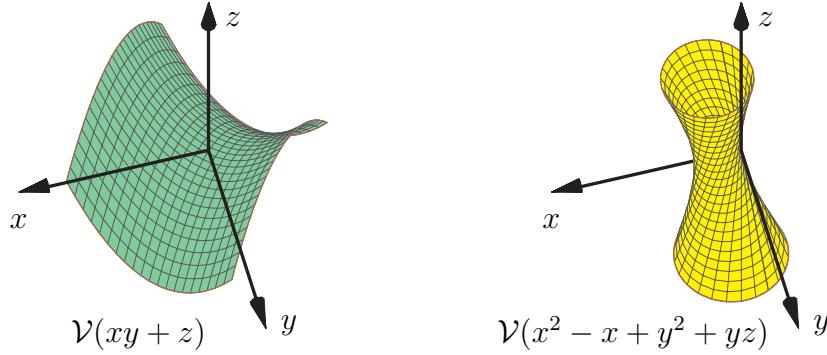
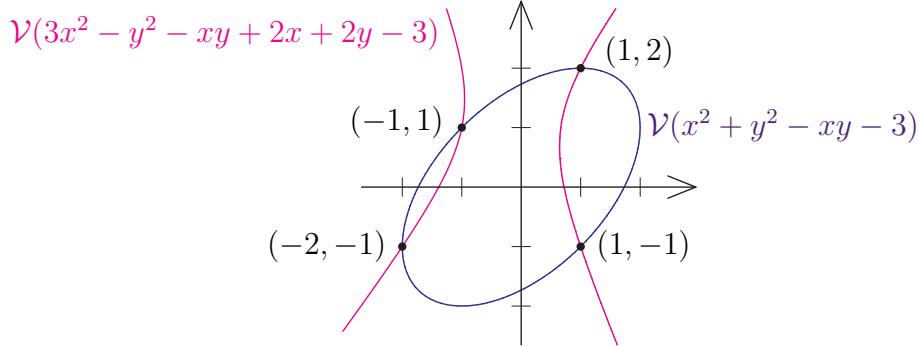


Figure 1.1: Two hyperboloids.

$\mathcal{V}(x^2 + y^2 - xy - 3)$ 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

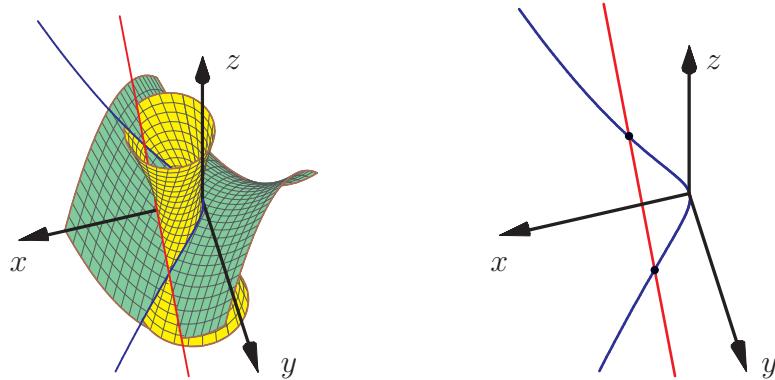


Figure 1.2: Intersection of two quadrics.

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. We will have more to say on this in Section 3.5 (or 3.6).

The intersection of the hyperboloid $x^2 + (y - \frac{3}{2})^2 - z^2 = \frac{1}{4}$ with the sphere $x^2 + y^2 + z^2 = 4$ centered at the origin with radius 2 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.

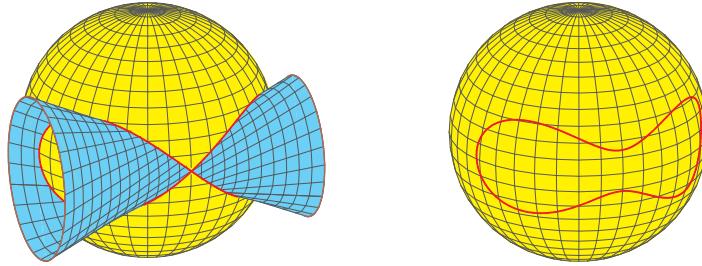


Figure 1.3: Quartics on spheres.

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 $f_1, \dots, f_s \in \mathbb{K}[x_1, \dots, x_n]$ and that $Y \subset \mathbb{K}^m$ is defined by the polynomials $g_1, \dots, g_t \in \mathbb{K}[y_1, \dots, y_m]$. Then $X \times Y \subset \mathbb{K}^n \times \mathbb{K}^m = \mathbb{K}^{n+m}$ is defined by the polynomials $f_1, \dots, f_s, g_1, \dots, g_t \in \mathbb{K}[x_1, \dots, x_n, y_1, \dots, y_m]$. 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 $\text{Mat}_{m \times n}$ or $\text{Mat}_{m \times n}(\mathbb{K})$ of $m \times n$ matrices with entries in \mathbb{K} is identified with the affine space \mathbb{K}^{mn} , which may be written $\mathbb{K}^{m \times n}$. An interesting class of varieties are linear algebraic groups, which are algebraic subvarieties of the space $\text{Mat}_{n \times n}$ square matrices that are closed under multiplication and taking inverses. The *special linear group* is the set of matrices with determinant 1,

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

which is a linear algebraic group. Since the determinant of a matrix in $\text{Mat}_{n \times n}$ 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.)

The general linear group $GL_n := \{M \in \text{Mat}_{n \times n} \mid \det M \neq 0\}$ at first does not appear to be a variety as it is defined by an inequality. You will show in Exercise 7 that it may be identified with the set $\{(t, M) \in \mathbb{K} \times \text{Mat}_{n \times n} \mid t \det M = 1\}$, which is a variety.

There is a general construction of other linear algebraic groups. Let g^T be the transpose of a matrix $g \in \text{Mat}_{n \times n}$. For a fixed matrix $M \in \text{Mat}_{n \times n}$, 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 $\text{Mat}_{2n \times 2n}$. 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 is known as the split form of the special orthogonal group which is not compact.

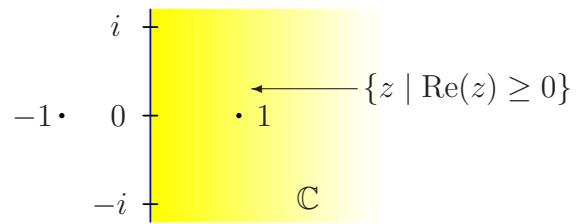
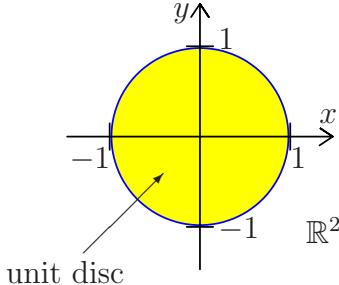
When $n = 2$, consider the two different real groups:

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

Note that in the Euclidean topology (see Appendix A.2) $SO_{2,0}(\mathbb{R})$ 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 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 5.

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. Let $S \subset T$ be sets of multivariate polynomials in $\mathbb{K}[x]$. Show that $\mathcal{V}(S) \supset \mathcal{V}(T)$.
3. Show that any finite subset Z of \mathbb{K}^n is a variety. (Hint: for a linear form $\Lambda : \mathbb{K}^n \rightarrow \mathbb{K}$, the polynomial

$$\Lambda_Z := \prod_{z \in Z} (\Lambda(x) - \Lambda(z))$$

vanishes on Z . Show that there is a set L of linear forms such that the polynomials Λ_Z for $\Lambda \in L$ define Z .)

4. Prove that in \mathbb{K}^2 we have $\mathcal{V}(y-x^2) = \mathcal{V}(y^3-y^2x^2, x^2y-x^4)$.
5. Show that the following sets are not algebraic varieties.
 - (i) $\{(x, y) \in \mathbb{R}^2 | y = \sin x\}$.
 - (ii) $\{(\cos t, \sin t, t) \in \mathbb{R}^3 | t \in \mathbb{R}\}$.
 - (iii) $\{(x, e^x) \in \mathbb{R}^2 | x \in \mathbb{R}\}$.
6. Express the cubic space curve C with parametrization (t, t^2, t^3) for $t \in \mathbb{K}$ as a variety 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.
7. 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 $GL_n(\mathbb{K}) \rightarrow \mathbb{K}^{n^2+1}$.

8. An $n \times n$ matrix with complex entries is *unitary* if its columns are orthonormal under the complex inner product $\langle z, w \rangle = z \cdot \bar{w}^t = \sum_{i=1}^n z_i \bar{w}_i$. 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.
9. 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 at most r is an algebraic variety.
 - (b) Show that the set of matrices of rank exactly r is not an algebraic variety when $r > 0$.

1.2 The algebra-geometry dictionary I

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—in fact an equivalence—between algebraic concepts and geometric concepts that we refer to as the algebra-geometry dictionary.

We defined varieties $\mathcal{V}(S)$ associated to sets $S \subset \mathbb{K}[x]$ of multivariate 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 ,

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

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

These two inclusion-reversing maps

$$\{\text{Subsets } S \text{ of } \mathbb{K}[x]\} \quad \begin{array}{c} \xrightarrow{\mathcal{V}} \\ \xleftarrow{\mathcal{I}} \end{array} \quad \{\text{Subsets } Z \text{ of } \mathbb{K}^n\} \tag{1.2}$$

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 $I \subset \mathbb{K}[x]$ which is closed under addition and under multiplication by polynomials in $\mathbb{K}[x]$. If $f, g \in I$ then $f + g \in I$ and if we also have $h \in \mathbb{K}[x]$, then $hf \in I$. The *ideal* $\langle S \rangle$ generated by a subset S of $\mathbb{K}[x]$ 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 $f_1, \dots, f_m \in S$ and $h_1, \dots, h_m \in \mathbb{K}[x]$. 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 $\mathbb{K}[x]$. In fact, we lose nothing if we restrict the left-hand-side of the correspondence (1.2) to the ideals of $\mathbb{K}[x]$.

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

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 $\mathbb{K}[x]$. This shows that $\mathcal{I}(S)$ is an ideal of $\mathbb{K}[x]$. \square

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 $\mathbb{K}[x]$ is finitely generated.*

We will prove a stronger form of this (Theorem 2.2.10) in Chapter 2, but use it here. Hilbert's Basis Theorem implies 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, \dots, f_s , and so $Z = \mathcal{V}(f_1, \dots, f_s) = \mathcal{V}(f_1) \cap \dots \cap \mathcal{V}(f_s)$. \square

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 .*

Foreshadowing the discussion of Zariski topology in Section 3.1, we write \overline{Z} for $\mathcal{V}(\mathcal{I}(Z))$, the smallest variety containing Z , and call it the closure of 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 4. \square

Thus we also lose nothing if we restrict the right-hand-side of the correspondence (1.2) to the subvarieties of \mathbb{K}^n . Our correspondence now becomes

$$\{\text{Ideals } I \text{ of } \mathbb{K}[x]\} \quad \xrightleftharpoons[\mathcal{I}]{\mathcal{V}} \quad \{\text{Subvarieties } X \text{ of } \mathbb{K}^n\}. \quad (1.3)$$

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 Chapters 6 and 7.

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

$$(y - x^2)^2 = y^2 - yx^2 - x(xy - x^3) \in \langle y^2 - yx^2, xy - x^3 \rangle. \quad (1.4)$$

These two cases reveal a source for lack of injectivity of the map \mathcal{V} —the polynomials f and f^N have the same set of zeroes, for any positive integer N . For example, if f_1, \dots, f_s are polynomials, then the two ideals

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

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 $I \subset \mathbb{K}[x]$ 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 $\mathbb{K}[x]$ is

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

You will show in Exercise 3 that \sqrt{I} is the smallest radical ideal containing I . For example (1.4) 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 $I \subset \mathbb{K}[x]$ is an ideal, then $\mathcal{V}(I) = \mathcal{V}(\sqrt{I})$.*

When \mathbb{K} is algebraically closed, the precise nature of the correspondence (1.3) 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 $\mathbb{K}[x]$ with $\mathcal{V}(I) = \emptyset$, then $I = \mathbb{K}[x]$.*

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

Theorem 1.2.7. *Suppose that \mathbb{K} is algebraically closed. Then every maximal ideal \mathfrak{m} of $\mathbb{K}[x_1, \dots, x_n]$ has the form \mathfrak{m}_b for some $b \in \mathbb{K}^n$.*

Proof. We prove this when \mathbb{K} is an uncountable field, e.g. $\mathbb{K} = \mathbb{C}$. As \mathfrak{m} is a maximal ideal, $\mathbb{K}[x]/\mathfrak{m}$ is a field, L , that contains \mathbb{K} whose dimension as a \mathbb{K} -vector space is at most countable (L 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 ξ that does not satisfy any algebraic equations with coefficients in \mathbb{K} (ξ is transcendental over \mathbb{K}). But then the subfield of L generated by \mathbb{K} and ξ is isomorphic to the field $\mathbb{K}(t)$ of rational functions (quotients of polynomials) in the indeterminate t , under the map $t \mapsto \xi$. 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 over \mathbb{K} . Indeed, suppose that there is a linear dependency among elements of this set,

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

For any $i = 1, \dots, m$, if we multiply this by $(t - a_i)$ and simplify, and then substitute $t = a_i$, we obtain the equation $\lambda_i = 0$. This shows that the elements $\frac{1}{t-a}$ for $a \in \mathbb{K}$ are linearly independent over \mathbb{K} . Thus $\mathbb{K}(t)$ has uncountable dimension over \mathbb{K} and so L cannot contain a subfield isomorphic to $\mathbb{K}(t)$.

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 both are maximal ideals, they are equal. \square

Proof of the weak Nullstellensatz. We prove the contrapositive, if $I \subsetneq \mathbb{K}[x]$ 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 $\mathbb{K}[x]$ 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 $I = \mathbb{K}[x]$, which proves the weak Nullstellensatz. \square

A consequence of this proof is that there is a 1-1 correspondence

$$\{\text{Points } b \in \mathcal{V}(I)\} \longleftrightarrow \{\text{Maximal ideals } \mathfrak{m}_b \supset I\}.$$

The Fundamental Theorem of Algebra states that any nonconstant univariate 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). *Let \mathbb{K} be an algebraically closed field. If the polynomials $f_1, \dots, f_m \in \mathbb{K}[x_1, \dots, x_n]$ generate a proper ideal, 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 (1.3). For this, we assume that \mathbb{K} is algebraically closed.

Theorem 1.2.9 (Nullstellensatz). *Let \mathbb{K} be an algebraically closed field. If $I \subset \mathbb{K}[x]$ 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 using the ‘trick of Rabinowitsch’. Suppose that we have a polynomial $f \in \mathcal{I}(\mathcal{V}(I))$. Let us 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, \dots, a_n, b) were a point of this variety, then (a_1, \dots, a_n) would be a point of $\mathcal{V}(I)$. But then $f(a_1, \dots, a_n) = 0$, and so the polynomial $tf - 1$ evaluates to 1 (and not 0) at the point (a_1, \dots, a_n, b) .

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

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

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, \dots, 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 $G_1, \dots, G_m \in \mathbb{K}[x]$. But this shows that $f \in \sqrt{I}$, and completes the proof of the Nullstellensatz. \square

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

$$\{\text{Radical ideals } I \text{ of } \mathbb{K}[x]\} \xleftrightarrow[\mathcal{I}]{\mathcal{V}} \{\text{Subvarieties } X \text{ of } \mathbb{K}^n\} \quad (1.5)$$

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. Thus \mathcal{I} and \mathcal{V} are inverse bijections. \square

Corollary 1.2.10 is only the beginning of the algebra-geometry 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

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

Note that $I + J$ is the ideal $\langle I, J \rangle$ generated by $I \cup J$, and that $I \cap J$ is also an ideal.

Lemma 1.2.11. *Let I, J be ideals in $\mathbb{K}[x]$ 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 by the Nullstellensatz we also have

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

You are asked to prove this in Exercise 8.

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 $J = \langle y^2 - x^2y \rangle$, 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.

A polynomial $f \in \mathbb{K}[x]$ has an essentially unique factorization $f = f_1 \cdots f_s$ into irreducible polynomials. It is unique in that any other factorization into irreducible polynomials will have the same length, and after permuting factors, the corresponding factors in each factorization are proportional. Collecting proportional factors and extracting a constant α if necessary, have $f = \alpha g_1^{n_1} \cdots g_r^{n_r}$ with each $n_i \geq 1$, where, if $i \neq j$, then g_i is not proportional to g_j . The *square-free* part of f is $\sqrt{f} = g_1 \cdots g_r$, and we have

$$\sqrt{\langle f \rangle} = \langle \sqrt{f} \rangle = \langle g_1 \rangle \cap \langle g_2 \rangle \cap \cdots \cap \langle g_r \rangle,$$

so that $\mathcal{V}(f) = \mathcal{V}(g_1) \cup \cdots \cup \mathcal{V}(g_r)$.

Exercises

1. Show that the map \mathcal{I} reverses inclusions so that $Z \subset Y$ implies $\mathcal{I}(Z) \supset \mathcal{I}(Y)$.
2. Verify the claim that the smallest ideal containing a set $S \subset \mathbb{K}[x]$ of polynomials consists of all expressions of the form

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

where $f_1, \dots, f_m \in S$ and $h_1, \dots, h_m \in \mathbb{K}[x]$.

3. Let I be an ideal of $\mathbb{K}[x]$. Show that

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

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

4. If $Y \subsetneq X$ are varieties, show that $\mathcal{I}(X) \subsetneq \mathcal{I}(Y)$.
5. Suppose that I and J are radical ideals. Show that $I \cap J$ is also a radical ideal.
6. Give radical ideals I and J for which $I + J$ is not radical.
7. Let I be an ideal in $\mathbb{K}[x]$. 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 $I = \mathbb{K}[x]$.
8. Give a proof of Lemma 1.2.11. Hint: Numbers 1. and 2. are set-theoretic.
9. Give two algebraic varieties Y and Z such that $\mathcal{I}(Y \cap Z) \neq \mathcal{I}(Y) + \mathcal{I}(Z)$.
10. (a) Let I be an ideal of $\mathbb{K}[x]$. Show that if $\mathbb{K}[x]/I$ 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 $\mathcal{V}(I)$ where $I = \langle xy, xz - yz \rangle$. Show that $\sqrt{I} = J$.
11. Let $f, g \in \mathbb{K}[x, y]$ be polynomials without a common factor. Use Exercise 10(a) to show that $\mathcal{V}(f) \cap \mathcal{V}(g)$ is a finite set.
12. 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.
13. Deduce the weak Nullstellensatz from the statement of the Strong Nullstellensatz, showing that they are equivalent.

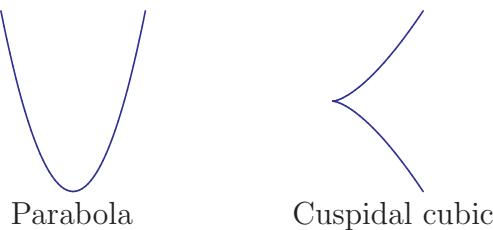
1.3 The algebra-geometry dictionary II

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 $\mathbb{K}[x_1, \dots, x_n]$ 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 varieties and homomorphisms of their coordinate rings.

We have used that a polynomial $f \in \mathbb{K}[x_1, \dots, x_n]$ gives a function $f: \mathbb{K}^n \rightarrow \mathbb{K}$, defined by evaluation at points of \mathbb{K}^n . When \mathbb{K} is infinite, the function is identically zero if and only if f is the zero polynomial, so this representation of polynomials by functions is faithful. Further suppose that $X \subset \mathbb{K}^n$ is an affine variety. Any polynomial function $f \in \mathbb{K}[x]$ restricts to give a *regular function* on X , $f: X \rightarrow \mathbb{K}$. 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 $\mathbb{K}[x] \twoheadrightarrow \mathbb{K}[X]$. 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 $\mathbb{K}[x]/\mathcal{I}(X)$.

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 any polynomial $f(x, y) \in \mathbb{K}[x, y]$ as a polynomial $g(x)$ in x alone. The resulting map $\mathbb{K}[x, y]/\langle y - x^2 \rangle \rightarrow \mathbb{K}[x]$ is well-defined and surjective. Since $y - x^2$ divides the difference $f(x, y) - g(x)$, the map is injective.



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 irreducible elements, while polynomials $f(x)$ in one variable have a unique factorization into irreducible polynomials. \diamond

Let $X \subset \mathbb{K}^n$ be a variety. Its coordinate ring $\mathbb{K}[X] = \mathbb{K}[x]/\mathcal{I}(X)$ has the structure of a vector space over \mathbb{K} , where addition is defined by the addition in the ring and scalar multiplication is defined by multiplication with an element in \mathbb{K} .

Definition 1.3.2. A \mathbb{K} -algebra is a ring that contains the field \mathbb{K} as a subring. \diamond

Hence, any \mathbb{K} -algebra has the structure of a vector space over \mathbb{K} . Using this terminology, the coordinate ring $\mathbb{K}[X]$ is a \mathbb{K} -algebra. Observe that $\mathbb{K}[X] = \mathbb{K}[x]/\mathcal{I}(X)$ is finitely generated by the images of the variables x_i . Since $\mathcal{I}(X)$ is radical, Exercise 4 implies that the coordinate ring $\mathbb{K}[X]$ has no nilpotent elements (elements f such that $f^N = 0$ for some N). Such a ring with no nilpotent elements is called *reduced*. When \mathbb{K} is algebraically closed, these two properties characterize coordinate rings of algebraic varieties.

Theorem 1.3.3. Suppose that \mathbb{K} is algebraically closed. 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, \dots, r_n for some $n \in \mathbb{N}$. Then there is a surjective ring homomorphism

$$\varphi : \mathbb{K}[x_1, \dots, x_n] \twoheadrightarrow R$$

given by $x_i \mapsto r_i$. Let $I \subset \mathbb{K}[x]$ be the kernel of φ . This identifies R with $\mathbb{K}[x]/I$. Since R is reduced, we have that I is radical. Indeed, a polynomial $f \notin I$ with $f^N \in I$ gives a nonzero element $\varphi(f) \in R$ with $(\varphi(f))^N = 0$.

As \mathbb{K} is algebraically closed, the algebra-geometry dictionary of Corollary 1.2.10 shows that $I = \mathcal{I}(\mathcal{V}(I))$ and so $R \simeq \mathbb{K}[x]/I \simeq \mathbb{K}[\mathcal{V}(I)]$. \square

A different choice s_1, \dots, s_m of generators for R in this proof will give a different affine variety with the same coordinate ring R . We seek to understand this apparent ambiguity.

Example 1.3.4. 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 in \mathbb{K}^2 . \diamond

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

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

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

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

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

and by Lemma 1.2.11 intersections 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 $\mathbb{K}[X] = \mathbb{K}[x]/\mathcal{I}(X)$ of X . (Recall that from abstract algebra, ideals of a quotient ring R/I have the form J/I , where J is an ideal of R which contains I .) 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 $\mathbb{K}[x] \rightarrow \mathbb{K}[X]$ and corresponding inclusion $X \hookrightarrow \mathbb{K}^n$.

Theorem 1.3.5 (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 algebra-geometry 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 you in Exercise 2.

Theorem 1.3.6. *Let X be an affine variety. Then the maps \mathcal{V} and \mathcal{I} give an inclusion reversing correspondence*

$$\{ \text{Radical ideals } I \text{ of } \mathbb{K}[X] \} \quad \begin{array}{c} \xleftarrow{\mathcal{V}} \\[-1ex] \xrightarrow{\mathcal{I}} \end{array} \quad \{ \text{Subvarieties } Y \text{ of } X \} \quad (1.6)$$

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

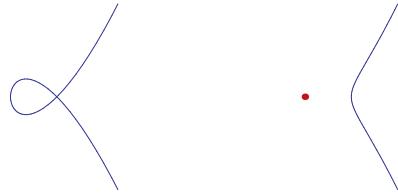
$$\begin{aligned} \varphi : X &\longrightarrow \mathbb{K}^m \\ x &\longmapsto (f_1(x), f_2(x), \dots, f_m(x)), \end{aligned}$$

which we call a *regular map*.

Example 1.3.8. The elements $t^2, t, -t^3 \in \mathbb{K}[t]$ define the map $\mathbb{K}^1 \rightarrow \mathbb{K}^3$ whose image is the cubic curve of Figure 1.2.

The elements t^2, t^3 of $\mathbb{K}[t]$ define a map $\mathbb{K}^1 \rightarrow \mathbb{K}^2$ whose image is the cuspidal cubic that we saw in Example 1.3.1.

Let $x = t^2 - 1$ and $y = t^3 - t$, which are elements of $\mathbb{K}[t]$. These define a map $\mathbb{K}^1 \rightarrow \mathbb{K}^2$ 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 $\mathbb{K}^1 \rightarrow \mathbb{K}^2$ whose image is the curve $\mathcal{V}(y^2 - (x^3 - x^2))$ on the right below. Both are singular at the origin.



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 $\pm\sqrt{-1}$.

Another regular map is matrix multiplication, $\mathbb{K}^{m \times n} \times \mathbb{K}^{n \times p} \rightarrow \mathbb{K}^{m \times p}$, because the product of two matrices $(a_{i,j}) \in \mathbb{K}^{m \times n}$ and $(b_{k,l}) \in \mathbb{K}^{n \times p}$ is the matrix in $\mathbb{K}^{m \times p}$ whose (i, l) -entry is $\sum_{j=1}^n a_{i,j} b_{j,l}$. Similarly, Cramer's rule shows that operation of taking inverse of a matrix is a regular map from $GL_n(\mathbb{K})$ to itself. \diamond

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

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

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

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

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

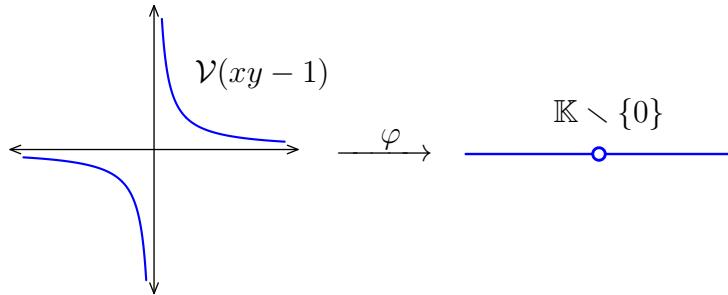
We have just shown the following basic fact.

Lemma 1.3.9. *The association $\varphi \mapsto \varphi^*$ defines a bijection*

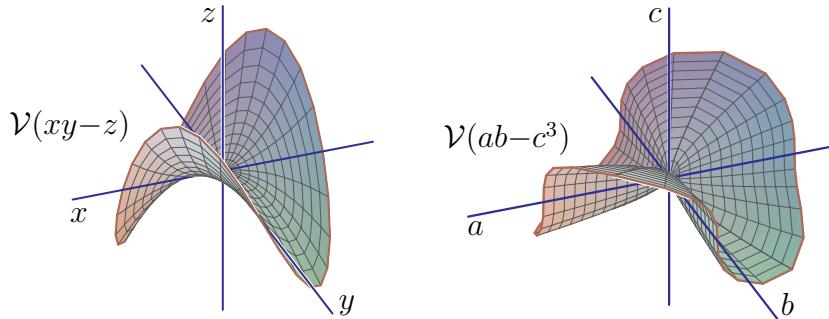
$$\left\{ \begin{array}{l} \text{Regular maps} \\ \varphi: X \rightarrow \mathbb{K}^m \end{array} \right\} \longleftrightarrow \left\{ \begin{array}{l} \mathbb{K}\text{-algebra homomorphisms} \\ \psi: \mathbb{K}[x_1, \dots, x_m] \rightarrow \mathbb{K}[X] \end{array} \right\}$$

In each of the regular maps of Example 1.3.8, the image $\varphi(X)$ of X under φ was equal to a subvariety. This is not always the case.

Example 1.3.10. Let $X = \mathcal{V}(xy - 1)$ be the hyperbola in \mathbb{K}^2 and $\varphi: \mathbb{K}^2 \rightarrow \mathbb{K}$ the map which forgets the second coordinate. Then $\varphi(X) = \mathbb{K} \setminus \{0\} \subsetneq \mathbb{K}$.



For a more interesting example, let $X = \mathcal{V}(xy - z) \subset \mathbb{K}^3$, the hyperbolic paraboloid. Consider the map $\varphi: X \rightarrow \mathbb{K}^3$ given by the three regular functions on X which are the images in $\mathbb{K}[X]$ of yx, xz, xy . Let (a, b, c) be coordinates for the image \mathbb{K}^3 . Then $\varphi^*(a) = yz$, $\varphi^*(b) = xz$, and $\varphi^*(c) = xy = z$, as $xy = z$ in $\mathbb{K}[X]$. But then $\varphi^*(ab - c^3) = xyz^2 - z^3 = 0$ as again $xy = z$ in $\mathbb{K}[X]$. Consequently, $\varphi(X) \subset \mathcal{V}(ab - c^3)$. We show these two varieties $\mathcal{V}(xy - z)$ and $\mathcal{V}(ab - c^3)$.



We do not have $\varphi(X) = \mathcal{V}(ab - c^3)$. Let $(a, b, c) \in \mathcal{V}(ab - c^3)$. If $c \neq 0$, then you may check that $(b/c, a/c, c) \in \mathcal{V}(xy - z)$, and $\varphi(b/c, a/c, c) = (a, b, c)$. However, if $c = 0$, then either $a = 0$ or $b = 0$. If $(a, b) \neq (0, 0)$, then the point (a, b, c) does not lie in the image of φ , but $(0, 0, 0) = \varphi(0, 0, 0)$. Thus the image of X under φ is the complement of the a - and b -axes in $\mathcal{V}(ab - c^3)$, together with the origin. This image is neither a subvariety, nor the complement of a subvariety. \diamond

Lemma 1.3.11. Let X be an affine variety, $\varphi: X \rightarrow \mathbb{K}^m$ 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 . We call this the subvariety of \mathbb{K}^m *parameterized* by φ . It is also the Zariski closure $\overline{\varphi(X)}$ of $\varphi(X)$. (This notion of Zariski closure is developed in Section 3.1.)

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 φ^*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 $\varphi(X) \subset Y$. \square

Definition 1.3.12. Affine varieties X and Y are *isomorphic* if there are regular maps $\varphi: X \rightarrow Y$ and $\psi: Y \rightarrow X$ such that both $\varphi \circ \psi$ and $\psi \circ \varphi$ are the identity maps on Y and X , respectively. In this case, we say that φ and ψ are isomorphisms.

Corollary 1.3.13. Let X be an affine variety, $\varphi: X \rightarrow \mathbb{K}^m$ 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 $\varphi: X \rightarrow Y$, then $\varphi^*: \mathbb{K}[x_1, \dots, x_m] \rightarrow \mathbb{K}[X]$ factors through $\mathbb{K}[Y]$ inducing a homomorphism $\varphi^*: \mathbb{K}[Y] \rightarrow \mathbb{K}[X]$.
- (4) φ is an isomorphism of varieties if and only if $\varphi^*: \mathbb{K}[Y] \rightarrow \mathbb{K}[X]$ is an isomorphism of \mathbb{K} -algebras.

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 1.3.11, $\mathcal{I}(Y) \subset \ker \varphi^*$ 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 $\varphi^*: \mathbb{K}[x_1, \dots, x_m] \rightarrow \mathbb{K}[X]$ factors through the quotient map $\mathbb{K}[x_1, \dots, x_m] \rightarrow \mathbb{K}[x_1, \dots, x_m]/\mathcal{I}(Y) = \mathbb{K}[Y]$.

Statement (4) is immediate from the definitions. \square

Thus we may refine the correspondence of Lemma 1.3.9. Let X and Y be affine varieties. Then the association $\varphi \mapsto \varphi^*$ gives a bijective correspondence

$$\left\{ \begin{array}{l} \text{Regular maps} \\ \varphi: X \rightarrow Y \end{array} \right\} \leftrightarrow \left\{ \begin{array}{l} \mathbb{K}\text{-algebra homomorphisms} \\ \psi: \mathbb{K}[Y] \rightarrow \mathbb{K}[X] \end{array} \right\}.$$

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 $L: V \rightarrow W$, its adjoint is a map $L^*: W^* \rightarrow V^*$. 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. Suppose that \mathbb{K} is an infinite field. Show that $f \in \mathbb{K}[x_1, \dots, x_n]$ defines the zero function $f: \mathbb{K}^n \rightarrow \mathbb{K}$ 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.)
2. Give a proof of Theorem 1.3.5.
3. 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.

4. Let $I \subset \mathbb{K}[x_1, \dots, x_n]$ be an ideal. Show that the quotient ring $\mathbb{K}[x_1, \dots, x_n]/I$ has nilpotent elements if and only if I is not a radical ideal.
5. Suppose that $I \subset \mathbb{K}[x_1, \dots, x_n]$ is a radical ideal and that $X := \mathcal{V}(I)$ is a finite set. Prove that the restriction of polynomial functions to X is a surjective map from the ring of polynomials $\mathbb{K}[x_1, \dots, x_n]$ to the finite vector space of functions from $X \rightarrow \mathbb{K}$.
6. Verify the claims about the parametrizations in Example 1.3.8, that the image of \mathbb{K} under the map $t \mapsto (t^2 - 1, t^3 - t)$ is $\mathcal{V}(y^2 - (x^3 + x^2))$ and its image under $t \mapsto (t^2 + 1, t^3 + t)$ is $\mathcal{V}(y^2 - (x^3 - x^2))$.
7. Show that $A \mapsto A^{-1}$ is a regular map on $GL_m(\mathbb{K})$.

1.4 Projective varieties

Projective space and projective varieties are of central importance in algebraic geometry. We motivate projective space with an example.

Example 1.4.1. 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

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

There are several cases to consider, illustrated below (1.7).

- (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 picture in \mathbb{R}^2 . In \mathbb{C}^2 , 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 2.1.17).

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

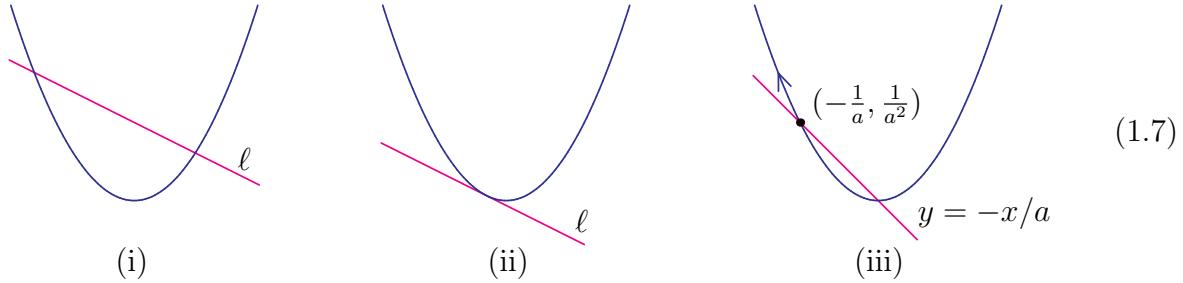
$$a(x - \frac{b}{2a})^2 = 0 \quad \text{and} \quad y = x^2.$$

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) $a = 0$, so that the line ℓ is vertical. There is a single, unique solution, $x = -c/b$ and $y = c^2/b^2$.

Let us examine this passage to a vertical line. 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 $a \rightarrow 0$, the second solution disappears off to infinity.

We illustrate these three possibilities.



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

Definition 1.4.2. 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 $\mathbb{P}_{\mathbb{K}}^n$. 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 $\mathbb{P}(V) \simeq \mathbb{P}^{\dim V - 1}$.

Example 1.4.3. The projective line \mathbb{P}^1 is the set of lines through the origin in \mathbb{K}^2 . When $\mathbb{K} = \mathbb{R}$, the line $x = ay$ through the origin intersects the circle $\mathcal{V}(x^2 + (y - 1)^2 - 1)$ in the origin and in the second point $(\frac{2a}{1+a^2}, \frac{2}{1+a^2})$, as shown in Figure 1.4. Identifying the nonhorizontal line $x = ay$ with this point $(\frac{2a}{1+a^2}, \frac{2}{1+a^2})$ and the horizontal x -axis with the origin, this identifies $\mathbb{P}_{\mathbb{R}}^1$ with the circle. \diamond

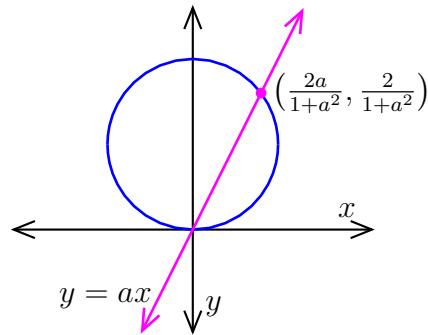


Figure 1.4: Lines through the origin meet the circle in a second point.

This definition of \mathbb{P}^n leads to a system of coordinates for \mathbb{P}^n . We may represent a point, ℓ , of \mathbb{P}^n by the coordinates $[a_0, a_1, \dots, a_n]$ 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 $[a_0, a_1, \dots, a_n]$ and $[\lambda a_0, \lambda a_1, \dots, \lambda a_n]$ both represent the same point. This non-uniqueness is the reason that we use rectangular brackets $[\dots]$ in our notation for these *homogeneous coordinates*. Some authors prefer the notation $[a_0 : a_1 : \dots : a_n]$.

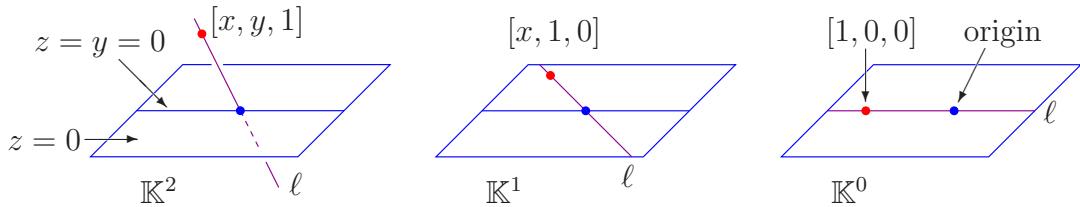
Example 1.4.4. When $\mathbb{K} = \mathbb{R}$, observe that a 1-dimensional subspace of \mathbb{R}^{n+1} meets the unit sphere S^n in two antipodal points, v and $-v$. The group $S^1 = \{-1, 1\}$ of real numbers of absolute value 1 acts on S^n by scalar multiplication interchanging antipodal points. This identifies real projective space $\mathbb{P}_{\mathbb{R}}^n$ with the quotient $S^n/\{\pm 1\}$, showing that $\mathbb{P}_{\mathbb{R}}^n$ is a compact manifold in the usual (Euclidean) topology.

Suppose that $\mathbb{K} = \mathbb{C}$. Given a point $a \in \mathbb{P}_{\mathbb{C}}^n$, after scaling, we may assume that $|a_0|^2 + |a_1|^2 + \dots + |a_n|^2 = 1$. 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 $[a_0, \dots, a_n] = [b_0, \dots, b_n]$ 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 $\mathbb{P}_{\mathbb{C}}^n$ with the quotient of S^{2n+1}/S^1 , showing that $\mathbb{P}_{\mathbb{C}}^n$ is a compact manifold. This is a version of the Hopf fibration. Since $\mathbb{P}_{\mathbb{R}}^n \subset \mathbb{P}_{\mathbb{C}}^n$, we again see that $\mathbb{P}_{\mathbb{R}}^n$ is compact. \diamond

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 $[a_0, a_1, \dots, a_n]$ in projective space \mathbb{P}^n with $a_i \neq 0$, but $a_{i+1} = \dots = a_n = 0$. Given a point $a \in A_i$, we may divide by its i th coordinate to get a representative of the form $[a_0, \dots, a_{i-1}, 1, 0, \dots, 0]$. These i numbers (a_0, \dots, 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

$$\mathbb{P}^n = \mathbb{K}^n \sqcup \dots \sqcup \mathbb{K}^1 \sqcup \mathbb{K}^0.$$

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, such as the Grassmannians of Section 10.1.

Figure 1.5: Affine paving of \mathbb{P}^2 .

It is instructive to look at this closely for \mathbb{P}^2 . Figure 1.5 shows 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 . Note that the last two charts give \mathbb{P}^1 , so we have $\mathbb{P}^2 = \mathbb{K}^2 \sqcup \mathbb{P}^1$, which is the familiar decomposition of the projective plane as the plane plus the line at infinity.

Projective space also admits systems of local coordinates. For $i = 0, \dots, n$, let U_i be the set of points $a \in \mathbb{P}^n$ in projective space whose i th coordinate is non-zero. Dividing by this i th coordinate, we obtain a representative of the point having the form

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

The n coordinates $(a_0, \dots, a_{i-1}, a_{i+1}, \dots, a_n)$ determine this point, identifying U_i with affine n -space, \mathbb{K}^n . Geometrically, U_i is the set of lines in \mathbb{K}^{n+1} that meet the affine plane defined by $x_i = 1$, with the point of intersection identifying U_i with this affine plane. Every point of \mathbb{P}^n lies in some U_i ,

$$\mathbb{P}^n = U_0 \cup U_1 \cup \dots \cup U_n.$$

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 .

These affine charts have a coordinate-free description. Let $\Lambda: \mathbb{K}^{n+1} \rightarrow \mathbb{K}$ be a linear map, and let $H \subset \mathbb{K}^{n+1}$ be the set $\{x \in \mathbb{K}^{n+1} \mid \Lambda(x) = 1\}$. Then $H \simeq \mathbb{K}^n$, and the map

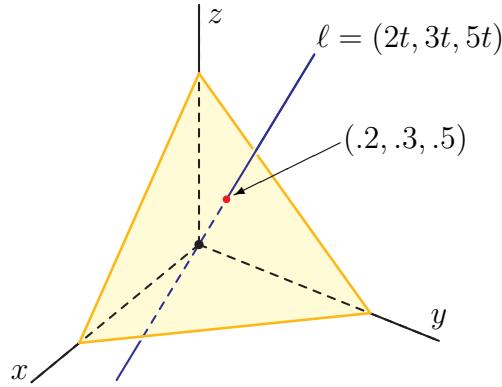
$$H \ni x \mapsto [x] \in \mathbb{P}^n$$

identifies H with the complement $U_\Lambda := \mathbb{P}^n - \mathbb{P}(\mathcal{V}(\Lambda))$ of the hyperplane defined by Λ .

Example 1.4.5 (Probability simplex). The second and more general description of affine charts leads to an application of algebraic geometry to statistics. Here $\mathbb{K} = \mathbb{R}$, the real numbers and we set $\Lambda(x) := x_0 + \dots + x_n$. If we consider those points x where $\Lambda(x) = 1$ which have nonnegative coordinates, we obtain the *probability simplex*

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

where \mathbb{R}_+^{n+1} is the *nonnegative orthant*, the points of \mathbb{R}^{n+1} with nonnegative coordinates. Here p_i represents the probability that event i occurs, and the condition $p_0 + \dots + p_n = 1$ reflects that every event does occur. Figure 1.6 shows this when $n = 2$. \diamond

Figure 1.6: Probability simplex when $n = 2$.

We wish to extend the definitions and structures of affine algebraic varieties to projective space. One problem arises immediately: given a polynomial $f \in \mathbb{K}[x_0, \dots, x_n]$ 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 natural number d , let f_d be the sum of the terms of f of degree d . We call f_d the d th *homogeneous component* of f . If $[a_0, \dots, a_n]$ and $[\lambda a_0, \dots, \lambda a_n]$ are two representatives of a point $a \in \mathbb{P}^n$, and f has degree m , then

$$f(\lambda a_0, \dots, \lambda a_n) = f_0(a_0, \dots, a_n) + \lambda f_1(a_0, \dots, a_n) + \dots + \lambda^m f_m(a_0, \dots, a_n), \quad (1.8)$$

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 (1.8) in λ is constant. That is, if and only if

$$f_i(a_0, \dots, a_n) = 0 \quad i = 1, \dots, \deg(f).$$

For a particular case, observe that 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* or simply *form* for a homogeneous polynomial.

Definition 1.4.6. Let $f_1, \dots, f_m \in \mathbb{K}[x_0, \dots, x_n]$ be homogeneous polynomials. These define a *projective variety*

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

An ideal $I \subset \mathbb{K}[x_0, \dots, x_n]$ 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 ,

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

In Exercise 2, 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 1.4.7. 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 $\mathcal{V}(\Lambda) = \mathbb{P}(V) \subset \mathbb{P}^n$. \diamond

Example 1.4.8. Let $[x, y, z]$ be homogeneous coordinates for the projective plane \mathbb{P}^2 , and consider the two subvarieties $\mathcal{V}(yz - x^2)$ and $\mathcal{V}(x + ay)$. In the affine patch U_z where $z \neq 0$, these subvarieties are the parabola and the line $x = -ay$ of Example 1.4.1. Their intersection, $\mathcal{V}(x + ay, yz - x^2)$, consists of the points $[0, 0, 1]$ and $[-a, 1, a^2]$. We see that as $a \rightarrow 0$, the second point approaches $[0, 1, 0]$, and does not “disappear off to infinity” as in Example 1.4.1(iii). \diamond

The weak Nullstellensatz does not hold for projective space, as $\mathcal{V}(x_0, x_1, \dots, x_n) = \emptyset$. We call this ideal, $\mathfrak{m}_0 := \langle x_0, x_1, \dots, x_n \rangle$, the *irrelevant ideal*.

Lemma 1.4.9. *Let $I \subset \mathbb{K}[x]$ be a homogeneous ideal. Then $\mathcal{V}(I) = \emptyset$ if and only if there is some $d \geq 0$ such that $I \supset \mathfrak{m}_0^d$.*

Proof. Note that $\mathcal{V}(I) = \emptyset$ in projective space if and only if, in the affine \mathbb{K}^{n+1} cone over projective space, we have either $\mathcal{V}(I) = \emptyset$ or $\mathcal{V}(I) = \{0\}$. This is equivalent to either $I = \mathbb{K}[x]$ or $\sqrt{I} = \mathfrak{m}_0$, which is in turn equivalent to $I \supset \mathfrak{m}_0^d$ for some $d \geq 0$. \square

The irrelevant ideal plays a special role in the projective algebra-geometry dictionary.

Theorem 1.4.10 (Projective Algebra-Geometry Dictionary). *Over any field \mathbb{K} , the maps \mathcal{V} and \mathcal{I} give an inclusion reversing correspondence*

$$\left\{ \begin{array}{l} \text{Radical homogeneous ideals } I \text{ of} \\ \mathbb{K}[x_0, \dots, x_n] \text{ properly contained in } \mathfrak{m}_0 \end{array} \right\} \quad \xleftrightarrow[\tau]{\mathcal{V}} \quad \{ \text{Subvarieties } X \text{ of } \mathbb{P}^n \}$$

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.

This follows from Lemma 1.4.9 and the algebra-geometry dictionary for affine varieties (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 (and even the same projective variety). This non-uniqueness comes from the irrelevant ideal, \mathfrak{m}_0 . Recall the construction of colon ideals

from commutative algebra. Let I be an ideal and g a polynomial. Then the colon ideal $(I : g)$ is $\{f \mid fg \in I\}$. If J is an ideal, then the *colon ideal* (or *ideal quotient*) of I by J is

$$(I : J) := \{f \mid fJ \subset I\} = \bigcap \{(I : g) \mid g \in J\}.$$

The saturation of I by J is

$$(I : J^\infty) = \bigcup_{m \geq 0} (I : J^m).$$

The reason for these definitions are the following results for affine varieties.

Lemma 1.4.11. *Let I be an ideal and $g \in \mathbb{K}[x]$ a polynomial. Then $\mathcal{V}(I : g^\infty)$ is the smallest affine variety containing $\mathcal{V}(I) \setminus \mathcal{V}(g)$.*

Proof. First note that as $I \subset (I : g^\infty)$, we have $\mathcal{V}(I : g^\infty) \subset \mathcal{V}(I)$. Let $x \in \mathcal{V}(I) \setminus \mathcal{V}(g)$. If $f \in (I : g^\infty)$, then there is some $m \in \mathbb{N}$ with $fg^m \in I$, so that $fg^m(x) = 0$. Since $x \notin \mathcal{V}(g)$, we conclude that $f(x) = 0$ as $g(x) \neq 0$. Thus $\mathcal{V}(I) \setminus \mathcal{V}(g) \subset \mathcal{V}(I : g^\infty)$.

For the other inclusion, let $x \in \mathcal{V}(I : g^\infty)$. If $g(x) \neq 0$, then $x \in \mathcal{V}(I) \setminus \mathcal{V}(g)$. Suppose now that $g(x) = 0$. Let $f \in \mathcal{I}(\mathcal{V}(I) \setminus \mathcal{V}(g))$. Note that fg vanishes on $\mathcal{V}(I)$. By the Nullstellensatz, there is some m such that $f^m g^m \in I$, and so $f^m \in (I : g^\infty)$. But then $f^m(x) = 0$ and so $f(x) = 0$. This shows that $x \in \mathcal{V}(\mathcal{I}(\mathcal{V}(I) \setminus \mathcal{V}(g)))$, and completes the proof. \square

Corollary 1.4.12. *Let I and J be ideals in $\mathbb{K}[x]$. Then $\mathcal{V}(I : J^\infty) = \overline{\mathcal{V}(I) \setminus \mathcal{V}(J)}$.*

A homogeneous ideal $I \subset \mathbb{K}[x_0, x_1, \dots, x_n]$ is *saturated* if

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

The reason for this definition is that I and $(I : \mathfrak{m}_0)$ define the same projective scheme, by Corollary 1.4.12 applied to the affine cones these varieties define in \mathbb{K}^{n+1} .

Given a projective variety $X = \mathcal{V}(I) \subset \mathbb{P}^n$, consider its intersection with an affine chart $U_i = \{x \in \mathbb{P}^n \mid x_i \neq 0\}$. For simplicity of notation, suppose that $i = 0$. Then

$$X \cap U_0 = \{x \in U_0 \mid f(x) = 0 \text{ for all } f \in I\}.$$

If we identify U_0 with \mathbb{K}^n by $U_0 = \{[1, x_1, \dots, x_n] \mid (x_1, \dots, x_n) \in \mathbb{K}^n\}$, this is

$$X \cap U_0 = \{x \in \mathbb{K}^n \mid f(1, x_1, \dots, x_n) = 0 \text{ for all } f \in I\}. \quad (1.9)$$

We call the polynomial $f(1, x_1, \dots, x_n)$ the *dehomogenization* of the homogeneous polynomial f with respect to x_0 . The calculation (1.9) shows that $X \cap U_0$ is the affine variety defined by the ideal generated by the dehomogenizations of forms in I .

This proves the forward implication of the following characterization of projective varieties in terms of their intersections with these affine charts.

Lemma 1.4.13. *A subset $X \subset \mathbb{P}^n$ is a projective variety if and only if $X \cap U_i$ is an affine variety, for each $i = 0, \dots, n$.*

Proof. For the reverse implication, suppose that $X \subset \mathbb{P}^n$ is a subset such that for each $i = 0, \dots, n$, $X \cap U_i$ is an affine variety. For each i , let $H_i = \mathcal{V}(x_i)$ be the hyperplane that is the complement of U_i . Then $X \subset (X \cap U_i) \cup H_i = X \cup H_i$. We claim that $X \cup H_i$ is a projective variety. This will imply the lemma, as

$$\bigcap_{i=0}^n (X \cup H_i) = X \cup \bigcap_{i=0}^n H_i = X,$$

as $H_0 \cap H_1 \cap \dots \cap H_n = \mathcal{V}(x_0, \dots, x_n) = \emptyset$ in \mathbb{P}^n .

To prove the claim, let $i = 0$ for simplicity and identify U_0 with \mathbb{K}^n whose coordinate ring is $\mathbb{K}[x_1, \dots, x_n]$. For a polynomial $g \in \mathbb{K}[x_1, \dots, x_n]$ of degree d , we have the homogeneous form g_+ of degree $d+1$ defined by

$$g_+(x_0, x_1, \dots, x_n) := x_0^{d+1} g\left(\frac{x_1}{x_0}, \dots, \frac{x_n}{x_0}\right).$$

Let I_+ be the homogeneous ideal generated by $\{g_+ \mid g \in \mathcal{I}(X \cap U_0)\}$. Since x_0 always divides g_+ , we have that $H_0 \subset \mathcal{V}(I_+)$. Since the dehomogenization of g_+ is g , the dehomogenization of I_+ is $\mathcal{I}(X \cap U_0)$. Then our previous calculations show that $X \cap U_0 = \mathcal{V}(I_+) \cap U_0$, which completes the proof. \square

The point of Lemma 1.4.13 is that every projective variety X is naturally a union of affine varieties

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

Consequently, we may often prove results for projective varieties by arguing locally on each of these affine sets that cover it. It also illustrates a relationship between varieties and manifolds: Affine varieties are to varieties as open subsets of \mathbb{R}^n are to manifolds.

Just as with affine varieties, projective varieties have coordinate rings. Let $X \subset \mathbb{P}^n$ be a projective variety. Its *homogeneous coordinate ring* $\mathbb{K}[X]$ is the quotient

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

If we set $\mathbb{K}[X]_d$ to be the image of all degree d homogeneous polynomials, $\mathbb{K}[x_0, \dots, x_n]_d$, 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

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

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

This differs from the coordinate ring of an affine variety in that its elements are not functions on X . Indeed, we already observed that, apart from constant polynomials, elements of $\mathbb{K}[x_0, \dots, x_n]$ do not give functions on any subset of \mathbb{P}^n . Despite this, they will be used to define maps of projective varieties, and the homogeneous coordinate ring plays another role which will be developed in Section 3.5.

Exercises

1. Verify the claim in Example 1.4.4 that if a, b lie on the unit sphere S^{2n+1} in \mathbb{C}^{n+1} and define the same point in \mathbb{P}^n , then $a = \zeta b$ for some unit complex number ζ .
2. Let $Z \subset \mathbb{P}^n$. Show that $\mathcal{I}(Z)$ is a homogeneous ideal.
3. 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, \dots, U_n .
4. Show that an ideal I is homogeneous if and only if it is generated by homogeneous polynomials.
5. Show that a radical homogeneous ideal is saturated.
6. 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 .
7. Verify the claim concerning the relation between the ideal of an affine subvariety $Y \subset U_0$ and of its Zariski closure $\overline{Y} \subset \mathbb{P}^n$:

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

8. Show that if $X \subset \mathbb{P}^n$ is a projective variety, then the smallest projective variety containing its intersection with the principal affine set U_{x_0} , $\overline{X \cap U_{x_0}}$ has ideal the saturation $(\mathcal{I}(X) : x_0^\infty)$.
9. Show that if I is a homogeneous ideal and $J = (I : \mathfrak{m}_0^\infty)$ is its *saturation* with respect to the irrelevant ideal \mathfrak{m}_0 , then there is some integer N such that

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

10. 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

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

1.5 Maps of projective varieties

Many properties of a projective variety X are inherited from the affine cone CX over X , but with some changes. The same is true for maps from X to a projective space. Elements of its homogeneous coordinate ring give maps, but care must be taken for the map to be well-defined. We explain this and describe some important maps of projective varieties. This leads to the product of projective varieties, and one of the most important properties of projective varieties; that the image of a projective variety under a map is a subvariety, which is not true for affine varieties, as we have seen in Example 1.3.10.

Suppose that $X \subset \mathbb{P}^n$ is a projective variety. Let $f_0, \dots, f_m \in \mathbb{K}[X]$ be elements of its homogeneous coordinate ring. Under what circumstances does

$$X \ni x \mapsto [f_0(x), f_1(x), \dots, f_m(x)] \in \mathbb{P}^m$$

define a map $X \rightarrow \mathbb{P}^m$? (It always defines a map $CX \rightarrow \mathbb{K}^{m+1}$.) Already the evaluation $f_i(x)$ of f_i at $x \in \mathbb{P}^n$ is a problem as the value of $f_i(x)$ is ambiguous. When f is homogeneous of degree d we saw that $f(\lambda x) = \lambda^d f(x)$ for $\lambda \in \mathbb{K}$. Thus when f_0, \dots, f_m are all homogeneous of the same degree d , their values at $x \in \mathbb{P}^n$ all have the same ambiguity. In fact, as long as $x \notin \mathcal{V}(f_0, \dots, f_m)$, then

$$\varphi(x) := [f_0(x), f_1(x), \dots, f_m(x)] \tag{1.10}$$

is a well-defined element of \mathbb{P}^m . Indeed, for $\lambda \in \mathbb{K}$, $\varphi(\lambda x) = \lambda^d \varphi(x)$ in \mathbb{K}^{m+1} , so that $\varphi(\lambda x) = \varphi(x)$ in \mathbb{P}^m for $\lambda \neq 0$. When $\mathcal{V}(f_0, \dots, f_m) = \emptyset$, so that the f_i have no common zeroes on X , then (1.10) defines a *regular map* $\varphi: X \rightarrow \mathbb{P}^m$.

Example 1.5.1. Suppose that \mathbb{P}^1 has homogeneous coordinates $[s, t]$. Then s^2, st, t^2 are homogeneous elements of its coordinate ring of the same degree, 2, with $\mathcal{V}(s^2, st, t^2) = \emptyset$. These define a regular map $\varphi: \mathbb{P}^1 \rightarrow \mathbb{P}^2$ where

$$\varphi: \mathbb{P}^1 \ni [s, t] \mapsto [s^2, st, t^2] \in \mathbb{P}^2.$$

If $[x, y, z]$ are the standard coordinates for \mathbb{P}^2 , then the image of φ is $\mathcal{V}(xz - y^2)$. Indeed, the image is a subset of $\mathcal{V}(xz - y^2)$ as $(s^2)(t^2) - (st)^2 = 0$. Let $[x, y, z] \in \mathcal{V}(xz - y^2)$. If $x \neq 0$, then $z = y^2/x$, and we have

$$[x, y, z] = [1, y/x, z/x] = [1, y/x, y^2/x^2] = \varphi([1, y/x]) = \varphi([x, y]). \tag{1.11}$$

If $x = 0$, then $y = 0$ and $[x, y, z] = [0, 0, z] = [0, 0, 1] = \varphi([0, 1])$. Thus $C := \mathcal{V}(xz - y^2)$ is the image of φ . This is the parabola of Examples 1.4.1 and 1.4.8. \diamond

The map φ of Example 1.5.1 is injective, and we would like to have that \mathbb{P}^1 is isomorphic to its image, C . For that, we need a map $C \rightarrow \mathbb{P}^1$ that is inverse to φ . For this, we extend and refine our notion of regular map of projective varieties. Let X be a projective variety and suppose that $f_0, \dots, f_m \in \mathbb{K}[X]$ are homogeneous elements of the same degree

with $\mathcal{V}(f_0, \dots, f_m) = \emptyset$ which define a regular map $\varphi: X \rightarrow \mathbb{P}^m$ (1.10). A second list $g_0, \dots, g_m \in \mathbb{K}[X]$ of elements of the same degree (possibly different from the degree of the f_i) with $\mathcal{V}(g_0, \dots, g_m) = \emptyset$ defines the same regular map if we have

$$\text{rank} \begin{pmatrix} f_0 & f_1 & \dots & f_m \\ g_0 & g_1 & \dots & g_m \end{pmatrix} = 1, \quad \text{i.e., if } f_i g_j - f_j g_i \in \mathcal{I}(X) \text{ for } i \neq j. \quad (1.12)$$

More interesting is when $f_0, \dots, f_m, g_0, \dots, g_m$ satisfy (1.12), but we do not have that $\mathcal{V}(f_0, \dots, f_m) = \mathcal{V}(g_0, \dots, g_m) = \emptyset$. In Example 1.5.1, (1.11) shows that if $[x, y, z] \in C = \mathcal{V}(xz - y^2)$, then $[x, y, z] = \varphi([x, y])$, but this requires that $(x, y) \neq (0, 0)$. Similarly, if $(y, z) \neq (0, 0)$, then $[x, y, z] = \varphi([y, z])$. We may understand this in terms of (1.12), at least when $xyz \neq 0$ as $\det \begin{pmatrix} x & y \\ y & z \end{pmatrix} = xz - y^2$, which vanishes on C . On C , $\mathcal{V}(x, y) = [0, 0, 1]$ and $\mathcal{V}(y, z) = [1, 0, 0]$, so that at every point of C , at least one of (x, y) or (y, z) can be used to define a map to \mathbb{P}^1 which is the inverse of φ .

Definition 1.5.2. A map $\varphi: X \rightarrow \mathbb{P}^m$ from a projective variety X is a *regular map* if for every $x \in X$, there are elements $f_0, \dots, f_m \in \mathbb{K}[X]$ of the same degree with $x \notin \mathcal{V}(f_0, \dots, f_m)$ such that for every $y \in X \setminus \mathcal{V}(f_0, \dots, f_m)$,

$$\varphi(y) = [f_0(y), f_1(y), \dots, f_m(y)].$$

That is, φ has the form (1.10), but the elements f_0, \dots, f_m may change for different parts of X (but any two choices satisfy (1.12), where they are both defined). \diamond

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

The map φ of Example 1.5.1 is an isomorphism between \mathbb{P}^1 and its image C . More generally, the set $V_{n,d}$ of all $\binom{n+d}{n}$ homogeneous monomials in x_0, \dots, x_n of degree d is a basis for the degree d component of the irrelevant ideal, and thus generates \mathfrak{m}_0^d . By Lemma 1.4.9, $\mathcal{V}(V_{n,d}) = \emptyset$, and thus this list of monomials gives a regular map,

$$\nu_{n,d}: \mathbb{P}^n \longrightarrow \mathbb{P}^{\binom{n+d}{n}-1},$$

called the *dth Veronese map*. The map φ of Example 1.5.1 is $\nu_{1,2}$. Let us study the image of ν_d . We adopt a useful convention from Section 8.1 and label the coordinates of $\mathbb{P}^{\binom{n+d}{n}-1}$ by the exponents of monomials in $V_{n,d}$,

$$\mathcal{A}_{n,d} := \{(a_0, \dots, a_n) \mid a_i \in \mathbb{N} \text{ and } a_0 + \dots + a_n = d\}.$$

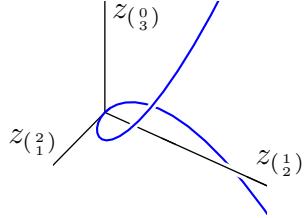
Then $V_{n,d} = \{x^\alpha \mid \alpha \in \mathcal{A}_{n,d}\}$ and $[z_\alpha \mid \alpha \in \mathcal{A}_{n,d}]$ are homogeneous coordinates of $\mathbb{P}^{\binom{n+d}{n}-1}$, with the z_α th coordinate of the Veronese map $\nu_{n,d}$ equal to x^α .

Observe that if $\alpha, \beta, \gamma, \delta \in \mathcal{A}_{n,d}$ satisfy $\alpha + \beta = \gamma + \delta$ (as integer vectors), then $z_\alpha z_\beta - z_\gamma z_\delta$ vanishes on the image $\nu_d(\mathbb{P}^n)$ as $\nu_d^*(z_\alpha z_\beta - z_\gamma z_\delta) = x^\alpha x^\beta - x^\gamma x^\delta = 0$. This

is the equation $xz - y^2 = 0$ that we found for $\nu_{1,2}$ in Example 1.5.1. When $n = 1$ and $d = 3$, we have $\mathcal{A}_{1,3} = \{(3), (2), (1), (0)\}$, $\nu_{1,3}([s, t]) = [s^3, s^2t, st^2, t^3]$, and the quadratic polynomials that vanish on the image include

$$z_{(0)} z_{(2)} - z_{(1)}^2, \quad z_{(0)} z_{(0)} - z_{(1)} z_{(1)}, \quad \text{and} \quad z_{(1)} - z_{(0)} z_{(2)}^2.$$

The image $\nu_{1,3}(\mathbb{P}^1)$ is the rational normal (or monomial) curve, depicted in $U_{(0)}$ below.



Theorem 1.5.3. *The image $\nu_{n,d}(\mathbb{P}^n) \subset \mathbb{P}^{\binom{n+d}{n}-1}$ is the subvariety defined by the vanishing of the quadratic polynomials*

$$z_\alpha z_\beta - z_\gamma z_\delta \quad \text{for } \alpha, \beta, \gamma, \delta \in \mathcal{A}_{n,d} \quad \text{with} \quad \alpha + \beta = \gamma + \delta, \quad (1.13)$$

and it is isomorphic to \mathbb{P}^n .

The image of $\nu_{n,d}$ is called the *Veronese variety*, and $\nu_{n,d}$ is the *Veronese embedding*.

Proof. We observed that these quadratics (1.13) vanish on $\nu_{n,d}(\mathbb{P}^n)$.

Let $X \subset \mathbb{P}^{\binom{n+d}{n}-1}$ be the variety defined by the vanishing of the quadratics (1.13) and let $y \in X$. In Exercise 3, you will show that there is at least one $i = 0, \dots, n$ such that $y_{de_i} \neq 0$. (Here, e_i is the i th standard basis vector so that de_i is the exponent of x_i^d .) Thus $y \in U_{de_i}$. Define φ_i on the affine patch $X \cap U_{de_i}$ by

$$\varphi_i(z) = [z_{e_j+(d-1)e_i} \mid j = 0, \dots, n].$$

(Here, the subscript $e_j + (d-1)e_i$ is the exponent of $x_j x_i^{d-1}$.) Then $\varphi_i: X \cap U_{de_i} \rightarrow U_i \subset \mathbb{P}^n$ is an inverse to $\nu_{n,d}$ on U_i , showing that $X \cap U_{de_i} = \nu_{n,d}(U_i) = \nu_{n,d}(\mathbb{P}^n) \cap U_{de_i}$. Thus these φ_i piece together to define a regular map $\varphi: \nu_{n,d}(\mathbb{P}^n) \rightarrow \mathbb{P}^n$ that is inverse to $\nu_{n,d}$. This completes the proof. \square

The value of the Veronese embedding is that if $f \in \mathbb{K}[x]$ is any form of degree d , then there is a linear form Λ_f on $\mathbb{P}^{\binom{n+d}{n}-1}$ such that $f = \nu_{n,d}^*(\Lambda_f)$. More precisely,

$$f = \sum_{\alpha \in \mathcal{A}_{n,d}} c_\alpha x^\alpha = \nu_{n,d}^* \left(\sum_{\alpha \in \mathcal{A}_{n,d}} c_\alpha z_\alpha \right). \quad (1.14)$$

Then $\nu_{n,d}(\mathcal{V}(f)) = \nu_{n,d}(\mathbb{P}^n) \cap \mathcal{V}(\Lambda_f)$. Extending this to a basis of $\mathcal{I}(X)_d$ for d large enough shows that any subvariety X of \mathbb{P}^n is isomorphic to a linear section of some

Veronese variety. Consequently, any projective variety is isomorphic to a variety defined by equations of degree at most two.

Furthermore, if f is a degree d form on \mathbb{P}^n with corresponding linear form Λ_f on $\mathbb{P}^{(n+d)-1}$, then $U_f = \mathbb{P}^n \setminus \mathcal{V}(f)$ is an affine variety as it is isomorphic to $\nu_{n,d}(\mathbb{P}^n) \cap U_{\Lambda_f}$. Consequently, for any projective variety $X \subset \mathbb{P}^n$ and any homogeneous element f of its coordinate ring, the set $X_f := X \setminus \mathcal{V}(f) = X \cap U_f$ is an affine variety. Lemma 1.4.13 extends to these more general affine charts U_f of \mathbb{P}^n .

This is related to maps of projective varieties. Suppose that $\varphi: X \rightarrow \mathbb{P}^m$ is a regular map defined on part of X by $\varphi(x) = [f_0(x), \dots, f_m(x)]$ for f_0, \dots, f_m homogeneous elements of $\mathbb{K}[X]$ of the same degree. Then φ is defined as a map of affine varieties on each affine patch X_{f_i} .

The product of affine varieties required no special treatment as the product $\mathbb{K}^m \times \mathbb{K}^n$ of two affine spaces is again an affine space, \mathbb{K}^{m+n} . This is not the case with projective spaces. To remedy this, we identify $\mathbb{P}^m \times \mathbb{P}^n$ with a subvariety of the projective space \mathbb{P}^{mn+m+n} , and use this identification to help understand subvarieties of $\mathbb{P}^m \times \mathbb{P}^n$.

Let x_0, \dots, x_m and y_0, \dots, y_n be homogeneous coordinates for \mathbb{P}^m and \mathbb{P}^n , respectively. Let $z_{i,j}$ for $i = 0, \dots, m$ and $j = 0, \dots, n$ be homogeneous coordinates for \mathbb{P}^{mn+m+n} . (Note that $(m+1)(n+1)-1 = mn+m+n$.) Define a map $\sigma_{m,n}: \mathbb{P}^m \times \mathbb{P}^n \rightarrow \mathbb{P}^{mn+m+n}$ by

$$\sigma_{m,n}(x, y) = z, \quad \text{where } z_{i,j} = x_i y_j.$$

This map becomes more clear when lifted to the affine cones over these projective spaces, where it is the map $\mathbb{K}^{m+1} \times \mathbb{K}^{n+1} \rightarrow \text{Mat}_{m+1, n+1}(\mathbb{K})$ that sends a pair of column vectors (x, y) to their outer product $xy^T \in \text{Mat}_{m+1, n+1}(\mathbb{K})$. The image is the set of rank 1 matrices, which is defined by the vanishing of the quadratic polynomials,

$$\det \begin{pmatrix} z_{i,j} & z_{i,l} \\ z_{k,j} & z_{k,l} \end{pmatrix} = z_{i,j}z_{k,l} - z_{i,l}z_{k,j} \quad \text{for } 0 \leq i < k \leq m \text{ and } 0 \leq j < l \leq n. \quad (1.15)$$

This is a special case of Exercise 9(a) in Section 1.1.

Theorem 1.5.4. *The image $\sigma_{m,n}(\mathbb{P}^m \times \mathbb{P}^n) \subset \mathbb{P}^{mn+m+n}$ is the subvariety defined by the vanishing of the quadratic polynomials (1.15). The map $\sigma_{m,n}$ admits an inverse.*

Call the map $\sigma_{m,n}$ the *Segre map* and its image the *Segre variety*. Exercise 5 explores the Segre variety $\mathbb{P}^1 \times \mathbb{P}^1 \subset \mathbb{P}^3$.

Proof. We sketch the proof, which is similar to that of Theorem 1.5.3, and leave the details as Exercise 6. For the inverse to $\sigma_{m,n}$, suppose that $X \subset \mathbb{P}^{mn+m+n}$ satisfies the equations (1.15). For each index k, l of a coordinate of \mathbb{P}^{mn+m+n} , we have an affine patch $U_{k,l} := \{z \in \mathbb{P}^{mn+m+n} \mid z_{k,l} \neq 0\}$. Define a map to $\mathbb{P}^m \times \mathbb{P}^n$ on the affine patch $X \cap U_{k,l}$ by

$$\varphi_{k,l}(z) = ([z_{i,l} \mid i = 0, \dots, m], [z_{k,j} \mid j = 0, \dots, n]).$$

Then $\varphi_{k,l}$ is an isomorphism between the affine varieties $X \cap U_{k,l}$ and $U_k \times U_l$. \square

This proof identifies affine patches $X \cap U_{k,l}$ with affine spaces $U_k \times U_l \simeq \mathbb{K}^m \times \mathbb{K}^n \subset \mathbb{P}^m \times \mathbb{P}^n$, and could be used to put the structure of an algebraic variety on the product $\mathbb{P}^m \times \mathbb{P}^n$, much as in differential geometry. Another approach is intrinsic: define subvarieties of $\mathbb{P}^m \times \mathbb{P}^n$ directly as we did with projective space. A third approach is extrinsic: use the Segre embedding to define subvarieties of $\mathbb{P}^m \times \mathbb{P}^n$. We develop the second and third approaches and show they coincide. That they give the same notion of subvariety as the first follows from the application of Lemma 1.4.13 to \mathbb{P}^{mn+m+n} and $\sigma_{m,n}(\mathbb{P}^m \times \mathbb{P}^n)$.

Both the intrinsic and extrinsic approaches begin with a definition. A monomial $x^\alpha y^\beta$ in $\mathbb{K}[x_0, \dots, x_n, y_0, \dots, y_m] = \mathbb{K}[x; y]$ has *bidegree* (a, b) where $a = \deg(x^\alpha)$ and $b = \deg(y^\beta)$. For example, the bidegree of $x_0 x_1^3 y_0 y_2 y_3$ is $(4, 3)$. A polynomial $g(x; y) \in \mathbb{K}[x; y]$ is *bihomogeneous* of *bidegree* (a, b) if each of its monomials has bidegree (a, b) . The same discussion that led us to understand the role of homogeneous ideals for projective varieties leads to bihomogeneous ideals defining subsets of $\mathbb{P}^n \times \mathbb{P}^m$.

We may also ask what are the subsets X of $\mathbb{P}^m \times \mathbb{P}^n$ whose image $\sigma_{m,n}(X)$ is a subvariety of \mathbb{P}^{mn+m+n} ? As $\sigma_{m,n}$ is given by bilinear monomials, the pullback $\sigma^*(f)$ of a form of degree d in z is a form of degree $2d$ that is bihomogeneous of bidegree (d, d) . Consequently, a subset X of $\mathbb{P}^m \times \mathbb{P}^n$ whose image $\sigma_{m,n}(X)$ is a subvariety is defined by bihomogeneous polynomials $g(x; y)$ with *diagonal* bidegree (a, a) .

To reconcile these two approaches, let $g(x; y)$ be a bihomogeneous polynomial with a non-diagonal bidegree (a, b) and suppose that $a = b + k$ with $k > 0$. Observe that

$$\mathcal{V}(g(x; y)) = \mathcal{V}(y_j^k g(x; y) \mid j = 0, \dots, n). \quad (1.16)$$

The same observation, but with x when $a < b$ shows that any subset of $\mathbb{P}^n \times \mathbb{P}^m$ defined by bihomogeneous polynomials may also be defined by bihomogeneous polynomials with a diagonal bidegree.

We follow the discussion leading up to Lemma 1.4.13 to define subvarieties of $\mathbb{P}^m \times \mathbb{K}^n$. If we restrict the second factor of $\mathbb{P}^m \times \mathbb{P}^n$ to $U_0 \simeq \mathbb{K}^n$ and dehomogenize bihomogeneous forms with respect to y_0 , we see that subvarieties of $\mathbb{P}^m \times \mathbb{K}^n$ are given by polynomials $f(x; y) \in \mathbb{K}[x_0, \dots, x_m, y_1, \dots, y_n]$ that are homogeneous in x , with no restriction on y . We have shown the following characterization of subvarieties of products.

Proposition 1.5.5. *A subvariety $X \subset \mathbb{P}^m \times \mathbb{P}^n$ is defined by a system of bihomogeneous polynomials $f_1(x; y), \dots, f_r(x; y)$. A subvariety $X \subset \mathbb{P}^m \times \mathbb{K}^n$ is defined by a system of polynomials $g_1(x; y), \dots, g_r(x; y)$ that are homogeneous in x .*

Let X, Y be varieties. Then $X \times Y$ is a variety, as it is defined by the set of polynomials $\{f(x)g(y) \mid f \in \mathcal{I}(X), g \in \mathcal{I}(Y)\}$. When both X and Y are affine, this was discussed in Section 1.1, when both are projective this is a set of bihomogeneous polynomials, and if X is projective and Y affine, then these are homogeneous in the first set of variables. This description of subvarieties of products of two projective spaces or of a projective space and an affine space extends in a natural way to arbitrary finite products of projective spaces with affine space; we leave the details to the reader.

In Example 1.4.4 we remarked that when $\mathbb{K} = \mathbb{C}$ or $\mathbb{K} = \mathbb{R}$, projective space is compact in the usual (Euclidean) topology, and consequently the projection maps $\mathbb{P}^n \times \mathbb{P}^m \rightarrow \mathbb{P}^m$ and $\mathbb{P}^n \times \mathbb{K}^m \rightarrow \mathbb{K}^m$ are proper in that the image of a closed set is also closed. This remains true, whatever the field, if we replace the property of being closed by that of being a subvariety. This will be a consequence of the following theorem.

Theorem 1.5.6. *The image of a subvariety $X \subset \mathbb{P}^m \times \mathbb{P}^n$ under the projection to \mathbb{P}^n is again a subvariety, and the same for subvarieties of $\mathbb{P}^m \times \mathbb{K}^n$ under projection to \mathbb{K}^n .*

Before proving Theorem 1.5.6, we will show that the image of a projective variety under a map is a variety. Let $\varphi: X \rightarrow Y$ be a regular map of algebraic varieties. This map, like all maps, factors as the composition of a projection with an injection. To see this, consider the *graph* of f ,

$$\Gamma_\varphi := \{(x, y) \in X \times Y \mid \varphi(x) = y\}.$$

This is a subvariety of $X \times Y$ defined by the vanishing of $f_j(x) - y_j$, where y_j runs over the coordinates of Y , at least on any affine patch X_{f_i} of X where φ is given by the $f_j \in \mathbb{K}[X]$. (This covers all cases of X, Y affine or projective.) The map $\iota: x \mapsto (x, f(x)) \in X \times Y$ is an isomorphism between X and the graph, and the map φ is its composition with the projection to Y .

Corollary 1.5.7. *If $\varphi: X \rightarrow Y$ is a regular map of projective varieties, then its image $\varphi(X)$ is a subvariety of Y .*

Proof. Let $\Gamma \subset X \times Y$ be the graph of φ . Suppose that X is a subvariety of \mathbb{P}^m and Y is a subvariety of \mathbb{P}^n . Then Γ is a subvariety of $\mathbb{P}^m \times \mathbb{P}^n$. By Theorem 1.5.6 the projection of Γ to \mathbb{P}^n , which is the image $\varphi(X)$, is a subvariety of \mathbb{P}^n , and hence of Y . \square

Proof of Theorem 1.5.6. By Lemma 1.4.13, it suffices to prove the statement about projection to \mathbb{K}^n , as we may argue locally on the affine patches U_0, \dots, U_n of \mathbb{P}^n . Let $X \subset \mathbb{P}^m \times \mathbb{K}^n$ be a subvariety. By Proposition 1.5.5, X is defined by the vanishing of finitely many polynomials

$$g_1(x; y), g_2(x; y), \dots, g_s(x; y) \in \mathbb{K}[x_0, \dots, x_m, y_1, \dots, y_n],$$

where each g_i is homogeneous of degree d_i in the x variables and with no condition on y .

Let $\pi: \mathbb{P}^m \times \mathbb{K}^n \rightarrow \mathbb{K}^n$ be the projection. A point $b \in \mathbb{K}^m$ lies in the image $\pi(X)$ if and only if the system of homogeneous polynomials

$$g_1(x; b) = g_2(x; b) = \dots = g_s(x; b) = 0,$$

has a solution in \mathbb{P}^m . By Lemma 1.4.9 this holds if and only if the ideal $I(b)$ these polynomials generate does not contain $\mathfrak{m}_0(x)^d$ for any d . Since $\mathfrak{m}_0(x)^d$ is generated by the vector space $\mathbb{K}[x]_d$ of all forms of degree d , this is equivalent to $I(b)_d \neq \mathbb{K}[x]_d$, for all d .

This degree d component $I(b)_d$ of $I(b)$ is the image of the linear map

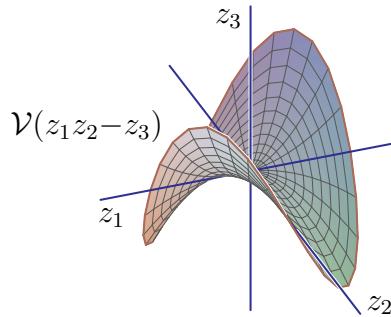
$$\Lambda_d(b) : \mathbb{K}[x]_{d-d_1} \oplus \cdots \oplus \mathbb{K}[x]_{d-d_s} \longrightarrow \mathbb{K}[x]_d,$$

given by $(f_1, \dots, f_s) \mapsto f_1g_1(x; b) + \cdots + f_sg_s(x; b)$. If we write the linear map $\Lambda_d(b)$ in terms of the bases of monomials of $\mathbb{K}[x]_d$ and $\mathbb{K}[x]_{d-d_i}$, we obtain a matrix $M_d(b)$ with entries the coefficients of the monomials in x in the $g_i(x; b)$, which are polynomials in b . Thus $I(b)_d \neq \mathbb{K}[x]_d$ if and only if $\Lambda_d(b)$ is not surjective if and only if the maximal minors of $M_d(b)$ vanish.

We conclude that b lies in $\pi(X)$ if and only if all the maximal minors of $M_d(b)$ vanish for all d . But this is a collection of polynomials in $\mathbb{K}[y_1, \dots, y_n]$, which shows that $\pi(X)$ is an affine subvariety of \mathbb{K}^n . \square

Exercises

1. Show that if $f_0, \dots, f_m \in \mathbb{K}[x]$ are homogeneous of the same degree that do not simultaneously vanish, $\mathcal{V}(f_0, \dots, f_m) = \emptyset$, then (1.10) defines a map $\varphi: \mathbb{P}^n \rightarrow \mathbb{P}^m$.
2. Show that the number of monomials in x_0, \dots, x_n of degree d is $\binom{n+d}{n} = \binom{n+d}{d}$.
3. Complete the proof of Theorem 1.5.3, verifying the claims made.
4. The quadratic Veronese map $\nu_{n,2}: \mathbb{P}^2 \rightarrow \mathbb{P}^{\binom{n+1}{2}}$ may be written as $z_{i,j} = x_i x_j$ for $0 \leq i \leq j \leq n$. Identify $\mathbb{P}^{\binom{n+1}{2}}$ with $(n+1) \times (n+1)$ symmetric matrices and show that the quadratic Veronese variety is the projectivization of the set of symmetric matrices of rank 1.
5. Show that the image of the Segre map $\sigma_{1,1}: \mathbb{P}^1 \times \mathbb{P}^1 \rightarrow \mathbb{P}^3$ is the projectivization of hyperbolic paraboloid $z_3 = z_1 z_2$.



6. Complete the proof of Theorem 1.5.4.
7. Explain why the set $\mathcal{V}(f) \subset \mathbb{P}^m \times \mathbb{P}^n$ is well-defined for a bihomogeneous polynomial $f(x, y) \in \mathbb{K}[x; y]$ and prove that for any subset $Z \subset \mathbb{P}^m \times \mathbb{P}^n$ its ideal $\mathcal{I}(Z) \subset \mathbb{K}[x; y]$ is bihomogeneous.

8. Prove the equality (1.16). Hint: saturate with respect to $\mathfrak{m}_0(y)$.
9. Prove that the graph Γ_φ of a regular map $\varphi: X \rightarrow Y$ of algebraic varieties is isomorphic to X .

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 [6], Cox, Little, O’Shea [25], Holme [62], Hulek [63], Perrin [102], Smith, Kahanpää, Kekäläinen, and Traves [129]. Advanced, in-depth treatments from the viewpoint of modern, abstract algebraic geometry can be found in the books of Eisenbud [35], Harris [48], Hartshorne [49], and Shafarevich [126]. Our treatment here and in Chapter 3 is most influenced by Shafarevich.

If the polynomials $f_1, \dots, f_m \in \mathbb{K}[x_1, \dots, x_n]$ over an algebraically closed K do not have a common zero, then Hilbert’s Nullstellensatz implies a polynomial identity of the form $\sum g_i f_i = 1$ with $g_1, \dots, g_m \in \mathbb{K}[x_1, \dots, x_n]$. However, the degrees of the polynomials in such a representation can grow doubly exponentially in the number n of variables, see Kollar [72].

Chapter 2

Symbolic algorithms

Symbolic algorithms, from resultants to Gröbner bases and beyond, have long been important in the use and application of algebraic geometry. The rise of computers has only increased their importance and they are now an indispensable part of the toolkit of modern algebraic geometry. We illustrate their utility for solving systems of equations.

2.1 Resultants and Bézout's Theorem

Resultants arose in the 19th century to provide symbolic algorithms for some operations such as elimination. They offer an approach to solving bivariate systems.

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

$$\begin{aligned} f &= f_0x^m + f_1x^{m-1} + \cdots + f_{m-1}x + f_m \\ g &= g_0x^n + g_1x^{n-1} + \cdots + g_{n-1}x + g_n, \end{aligned} \tag{2.1}$$

is to replace g by

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

which has degree at most $n-1$. (Note that $f_0 \cdot g_0 \neq 0$.) We often want to avoid division (e.g., when \mathbb{K} is a function field). Resultants detect common factors without division.

Let \mathbb{K} be any field. Let $\mathbb{K}[x]_\ell$ be the set of univariate polynomials of degree at most ℓ . (This differs from the use in Chapter 1, where $\mathbb{K}[X]_\ell$ consists of all homogeneous forms of degree ℓ .) This is a vector space over \mathbb{K} of dimension $\ell+1$ with a canonical ordered basis of monomials $1, x, \dots, x^\ell$. Given f and g as in (2.1), consider the linear map

$$\begin{aligned} L_{f,g} : \mathbb{K}[x]_{n-1} \times \mathbb{K}[x]_{m-1} &\longrightarrow \mathbb{K}[x]_{m+n-1} \\ (h(x), k(x)) &\longmapsto f \cdot h + g \cdot k. \end{aligned}$$

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

Lemma 2.1.1. *The polynomials f and g have a nonconstant common divisor if and only if $\ker L_{f,g} \neq \{(0,0)\}$.*

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, $\deg(k) < \deg(f) = m$ and $\deg(h) < \deg(g) = n$ so that $(h, -k) \in \mathbb{K}[x]_{n-1} \times \mathbb{K}[x]_{m-1}$. Since

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

we see that $(h, -k)$ is a non-zero element of the kernel of $L_{f,g}$.

Suppose that f and g are relatively prime and let $(h, k) \in \ker L_{f,g}$. 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 have $h = 0$, and so $\ker L_{f,g} = \{(0,0)\}$. \square

The *Sylvester matrix* is the matrix of the linear map $L_{f,g}$ in the ordered bases of monomials for $\mathbb{K}[x]_{m-1} \times \mathbb{K}[x]_{n-1}$ and $\mathbb{K}[x]_{m+n-1}$. When f and g have the form (2.1), it is

$$\text{Syl}(f, g; x) = \text{Syl}(f, g) := \left(\begin{array}{ccc|cc} f_m & & & g_n & 0 \\ f_{m-1} & f_m & 0 & g_{n-1} & \ddots \\ \vdots & \vdots & \ddots & \vdots & g_n \\ f_0 & \vdots & \ddots & \vdots & \vdots \\ & f_0 & f_m & g_1 & \vdots \\ & \ddots & \vdots & g_0 & \ddots & \vdots \\ 0 & \ddots & \vdots & f_0 & 0 & g_1 \\ & & & & & g_0 \end{array} \right). \quad (2.2)$$

Note that the sequence $f_m, \dots, f_0, g_0, \dots, g_0$ lies along the main diagonal and the left side of the matrix has n columns while the right side has m columns.

We often treat the coefficients $f_0, \dots, f_m, g_0, \dots, g_m$ of f and g as variables. That is, we will regard them as algebraically independent over \mathbb{Q} or \mathbb{Z} . Any formulas proven under this assumption remain valid when the coefficients of f and g lie in any field or ring.

The (*Sylvester*) *resultant* $\text{Res}(f, g)$ is the determinant of the Sylvester matrix. To emphasize that the Sylvester matrix represents the map $L_{f,g}$ in the basis of monomials in x , we also write $\text{Res}(f, g; x)$ for $\text{Res}(f, g)$. We summarize some properties of resultants, which follow from its definition and from Lemma 2.1.1.

Theorem 2.1.2. *The resultant of nonconstant polynomials $f, g \in \mathbb{K}[x]$ is an 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 give another expression for the resultant in terms of the roots of f and g .

Lemma 2.1.3. *Suppose that \mathbb{K} contains all the roots of the polynomials f and g so that*

$$f(x) = f_0 \prod_{i=1}^m (x - a_i) \quad \text{and} \quad g(x) = g_0 \prod_{i=1}^n (x - b_i),$$

where $a_1, \dots, a_m \in \mathbb{K}$ are the roots of f and $b_1, \dots, b_n \in \mathbb{K}$ are the roots of g . Then

$$\text{Res}(f, g; x) = (-1)^{mn} f_0^n g_0^m \prod_{i=1}^m \prod_{j=1}^n (a_i - b_j). \quad (2.3)$$

In Exercise 2 you are asked to show that this implies the Poisson formula,

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

Proof. We express these in $\mathbb{Z}[f_0, g_0, a_1, \dots, a_m, b_1, \dots, b_n]$. Recall that the coefficients of f and g are essentially the elementary symmetric polynomials in their roots,

$$f_i = (-1)^i f_0 e_i(a_1, \dots, a_m) \quad \text{and} \quad g_i = (-1)^i g_0 e_i(b_1, \dots, b_n).$$

We claim that both sides of (2.3) are homogeneous polynomials of degree mn in the variables a_1, \dots, b_n . This is immediate for the right hand side. 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

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

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

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

Since f_i and g_i are each homogeneous of degree i in the variables a_1, \dots, b_n and 0 is homogeneous of any degree, this term is homogeneous of degree

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

which proves the claim.

The resultant Res vanishes when $a_i = b_j$, which implies that Res lies in the ideal $\langle a_i - b_j \rangle$. Thus the resultant is a multiple of the double product in (2.3). As its degree is

mn , it is a scalar multiple. We determine this scalar. The term in $\text{Res}(f, g)$ which is the product of diagonal entries of the Sylvester matrix is

$$f_0^n g_n^m = f_0^n g_0^m e_n(b_1, \dots, b_n)^m = f_0^n g_0^m b_1^m \cdots b_n^m.$$

This is the only term of $\text{Res}(f, g)$ involving the monomial $b_1^m \cdots b_n^m$. The corresponding term on the right hand side of (2.3) is

$$(-1)^{mn} f_0^n g_0^m (-b_1)^m \cdots (-b_n)^m = f_0^n g_0^m b_1^m \cdots b_n^m,$$

which completes the proof. \square

Remark 3.2.12 uses geometric arguments to show that the resultant is irreducible and gives another characterization of resultants, which we give below.

Theorem 2.1.4. *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.*

Example 2.1.5. 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

$$\text{disc}(f) := (-1)^{\binom{n}{2}} f_0 \text{Res}(f, f') = f_0^{2n-2} \prod_{i < j} (a_i - a_j)^2,$$

where a_1, \dots, a_n are the roots of $f(x)$. \diamond

Resultants may also be used to eliminate variables from multivariate equations. The first step towards this is another interesting formula involving the Sylvester resultant, showing that it has a canonical expression as a polynomial linear combination of f and g .

Lemma 2.1.6. *Given polynomials $f, g \in \mathbb{K}[x]$, there are polynomials $h, k \in \mathbb{K}[x]$ whose coefficients are universal integer polynomials in the coefficients of f and g such that*

$$f(x)h(x) + g(x)k(x) = \text{Res}(f, g). \quad (2.4)$$

Proof. Set $\mathbb{K} := \mathbb{Q}(f_0, \dots, f_m, g_0, \dots, g_n)$, the field of rational functions (quotients of integer polynomials) in the variables $f_0, \dots, f_m, g_0, \dots, g_n$ and let $f, g \in \mathbb{K}[x]$ be univariate polynomials as in (2.1). Then $\gcd(f, g) = 1$ and so the map $L_{f,g}$ is invertible.

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

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

with $h \in \mathbb{K}[x]_{n-1}$ and $k \in \mathbb{K}[x]_{m-1}$.

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

$$\det(A) \cdot A^{-1} = \text{ad}(A). \quad (2.5)$$

Here $\text{ad}(A)$ 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 i th column and j th row.

Since $\det(L_{f,g}) = \text{Res}(f, g) \in \mathbb{K}$ and $L_{f,g}$ is \mathbb{K} -linear, we have

$$L_{f,g}^{-1}(\text{Res}(f, g)) = \text{Res}(f, g) \cdot L_{f,g}^{-1}(1) \det(L_{f,g}) \cdot L_{f,g}^{-1}(1) = \text{ad}(\text{Syl}(f, g))(1).$$

In the monomial basis of $\mathbb{K}[x]_{m+n-1}$ the polynomial 1 is the vector $(0, \dots, 0, 1)^T$. Thus, the coefficients of $L_{f,g}^{-1}(\text{Res}(f, g))$ are the entries of the last column of $\text{ad}(\text{Syl}(f, g))$, which are \pm the minors of the Sylvester matrix $\text{Syl}(f, g)$ with its last row removed. In particular, these are integer polynomials in the variables f_0, \dots, g_n . \square

This proof shows that $h, k \in \mathbb{Z}[f_0, \dots, f_m, g_0, \dots, g_n][x]$ and that (2.4) holds as an expression in this polynomial ring with $m+n+3$ variables. It leads to a method to eliminate variables. Suppose that $f, g \in \mathbb{K}[x_1, \dots, x_n]$ are multivariate polynomials. We may consider them as polynomials in the variable x_n whose coefficients are polynomials in the other variables, that is, as polynomials in $\mathbb{K}(x_1, \dots, x_{n-1})[x_n]$. Then the resultant $\text{Res}(f, g; x_n)$ both lies in the ideal generated by f and g and in the subring $\mathbb{K}[x_1, \dots, x_{n-1}]$. We examine the geometry of this elimination of variables.

Suppose that $1 \leq m < n$ and let $\pi: \mathbb{K}^n \rightarrow \mathbb{K}^m$ be the coordinate projection

$$\pi : (a_1, \dots, a_n) \longmapsto (a_1, \dots, a_m).$$

Also, for $I \subset \mathbb{K}[x_1, \dots, x_n]$ set $\text{I}_m := I \cap \mathbb{K}[x_1, \dots, x_m]$.

Lemma 2.1.7. *Let $I \subset \mathbb{K}[x_1, \dots, x_n]$ be an ideal. Then $\pi(\mathcal{V}(I)) \subset \mathcal{V}(I_m)$. When \mathbb{K} is algebraically closed $\mathcal{V}(I_m)$ is the smallest variety in \mathbb{K}^m containing $\pi(\mathcal{V}(I))$.*

Proof. Let us set $X := \mathcal{V}(I)$. For the first statement, suppose that $a = (a_1, \dots, a_n) \in X$. If $f \in I_m = I \cap \mathbb{K}[x_1, \dots, x_m]$, then

$$0 = f(a) = f(a_1, \dots, a_m) = f(\pi(a)),$$

which establishes the inclusion $\pi(X) \subset \mathcal{V}(I_m)$. (For this we view f as a polynomial in either x_1, \dots, x_n or in x_1, \dots, x_m .) This implies that $\mathcal{V}(\mathcal{I}(\pi(X))) \subset \mathcal{V}(I_m)$.

Now suppose that \mathbb{K} is algebraically closed. Let $f \in \mathcal{I}(\pi(X))$. Then $f \in \mathbb{K}[x_1, \dots, x_m]$ has the property that $f(a_1, \dots, a_m) = 0$ for all $(a_1, \dots, a_m) \in \pi(X)$. But then f is an element of $\mathbb{K}[x_1, \dots, x_n]$ that vanishes on $X = \mathcal{V}(I)$. By the Nullstellensatz, there is a positive integer N such that $f^N \in I$ (as elements of $\mathbb{K}[x_1, \dots, x_n]$). But then $f^N \in I \cap \mathbb{K}[x_1, \dots, x_m] = I_m$, which implies that $f \in \sqrt{I_m}$. Thus $\mathcal{I}(\pi(X)) \subset \sqrt{I_m}$, so that

$$\mathcal{V}(\mathcal{I}(\pi(X))) \supset \mathcal{V}(\sqrt{I_m}) = \mathcal{V}(I_m),$$

which completes the proof. \square

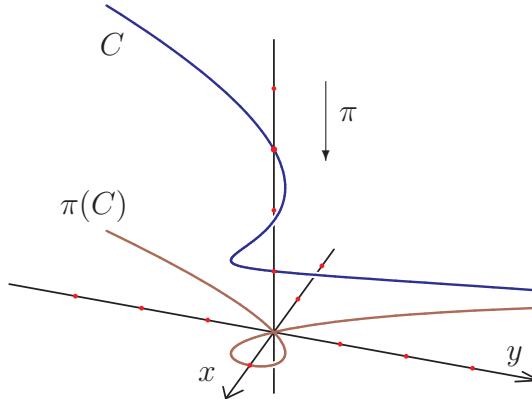
The ideal $I_m = I \cap \mathbb{K}[x_1, \dots, x_m]$ is called an *elimination ideal* as the variables x_{m+1}, \dots, x_n have been eliminated from the ideal I . By Lemma 2.1.7, elimination is the algebraic counterpart to projection, but the correspondence is not exact. For example, the inclusion $\pi(\mathcal{V}(I)) \subset \mathcal{V}(I \cap \mathbb{K}[x_1, \dots, x_m])$ may be strict. We saw this in Example 1.3.10 where the projection of the hyperbola $\mathcal{V}(xy - 1)$ the x -axis has image $\mathbb{K} - \{0\} \subsetneq \mathbb{K} = V(0)$, but $\langle 0 \rangle = \langle xy - 1 \rangle \cap \mathbb{K}[x]$. The missing point $\{0\}$ of \mathbb{K}^1 corresponds to the coefficient x of the highest power of y in $xy - 1$.

We may solve the implicitization problem for plane curves using elimination.

Example 2.1.8. Consider the parametric plane curve

$$x = 1 - t^2, \quad y = t^3 - t. \quad (2.6)$$

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



By Lemma 2.1.7, the plane curve is defined by $\langle t^2 - 1 + x, t^3 - t - y \rangle \cap \mathbb{K}[x, y]$. If we set

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

then the Sylvester resultant $\text{Res}(f, g; t)$ is

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

which is the implicit equation of the parameterized cubic $\pi(C)$ (2.6). \diamond

The ring $\mathbb{K}[x, y]$ of bivariate polynomials is a subring of the ring $\mathbb{K}(x)[y]$ of polynomials in y whose coefficients are rational functions in x . Suppose that $f, g \in \mathbb{K}[x, y]$. Considering f and g as elements of $\mathbb{K}(x)[y]$, the resultant $\text{Res}(f, g; y)$ is the determinant of their

Sylvester matrix expressed in the basis of monomials in y . By Theorem 2.1.2, $\text{Res}(f, g; y)$ is a univariate polynomial in x which vanishes if and only if f and g have a common factor in $\mathbb{K}(x)[y]$. 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 2.1.9. *Polynomials f and g in $\mathbb{K}[x, y]$ have a common factor of positive degree in y if and only if they have a common factor in $\mathbb{K}(x)[y]$.*

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

$$f = h \cdot \bar{f} \quad \text{and} \quad g = h \cdot \bar{g} \quad (2.7)$$

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

There is a polynomial $d \in \mathbb{K}[x]$ which is divisible by every denominator of a coefficient of h , \bar{f} , and \bar{g} . Multiplying the expressions (2.7) by d^2 gives

$$d^2 f = (dh) \cdot (d\bar{f}) \quad \text{and} \quad d^2 g = (dh) \cdot (d\bar{g}),$$

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

Let $\pi: \mathbb{K}^2 \rightarrow \mathbb{K}$ be the projection forgetting the last coordinate, $\pi(x, y) = x$. Set $I := \langle f, g \rangle \cap \mathbb{K}[x]$. By Lemma 2.1.6, the resultant $\text{Res}(f, g; y)$ lies in I . Combining this with Lemma 2.1.7 gives the chain of inclusions

$$\pi(\mathcal{V}(f, g)) \subset \mathcal{V}(I) \subset \mathcal{V}(\text{Res}(f, g; y)),$$

with the first inclusion an equality if \mathbb{K} is algebraically closed and $\pi(\mathcal{V}(f, g))$ is a variety. By Exercise 3 in Section 1.1 if $\mathcal{V}(f, g)$ is a finite set, then it is a variety.

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

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

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

Theorem 2.1.10 (Extension Theorem). *If $a \in \mathcal{V}(\langle f, g \rangle \cap \mathbb{K}[x]) \setminus \mathcal{V}(f_0(x), g_0(x))$, then there is some $b \in \mathbb{K}$ with $(a, b) \in \mathcal{V}(f, g)$.*

Writing $I := \langle f, g \rangle \cap \mathbb{K}[x]$, this establishes the chain of inclusions of subvarieties of \mathbb{K} ,

$$\mathcal{V}(I) \setminus \mathcal{V}(f_0, g_0) \subset \pi(\mathcal{V}(f, g)) \subset \mathcal{V}(I) \subset \mathcal{V}(\text{Res}(f, g; y)).$$

If either of f_0 or g_0 are constant, or if $\gcd(f, g) = 1$, then $\mathcal{V}(I) = \mathcal{V}(\text{Res}(f, g; y))$.

Proof. Let $a \in \mathcal{V}(I) \setminus \mathcal{V}(f_0, g_0)$. Suppose first that $f_0(a) \cdot g_0(a) \neq 0$. Then $f(a, y)$ and $g(a, y)$ are polynomials in y of degrees m and n , respectively. It follows that the Sylvester matrix $\text{Syl}(f(a, y), g(a, y))$ has the same format (2.2) as the Sylvester matrix $\text{Syl}(f, g; y)$, and it is in fact obtained from $\text{Syl}(f, g; y)$ by the substitution $x = a$.

This implies that $\text{Res}(f(a, y), g(a, y))$ is the evaluation of the resultant $\text{Res}(f, g; y)$ at $x = a$. Since $\text{Res}(f, g; y) \in I$ and $a \in \mathcal{V}(I)$, this evaluation is 0. By Theorem 2.1.2, $f(a, y)$ and $g(a, y)$ have a nonconstant common factor. As \mathbb{K} is algebraically closed, they have a common root, say b . But then $(a, b) \in \mathcal{V}(f, g)$, and so $a \in \pi(\mathcal{V}(f, g))$.

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

Example 2.1.11. Suppose that $f, g \in \mathbb{C}[x, y]$ are the polynomials,

$$\begin{aligned} f &= (5 - 10x + 5x^2)y^2 + (-14 + 42x - 24x^2)y + (5 - 28x + 19x^2) \\ g &= (5 - 10x + 5x^2)y^2 + (-16 + 46x - 26x^2)y + (19 - 36x + 21x^2) \end{aligned}$$

Figure 2.1 shows the curves $\mathcal{V}(f)$ and $\mathcal{V}(g)$, which meet in three points,

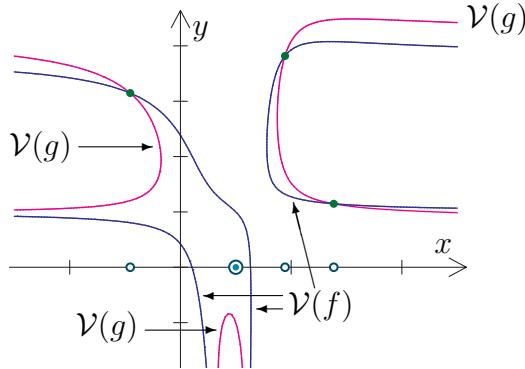


Figure 2.1: Comparing resultants to elimination.

$$\mathcal{V}(f, g) = \{(-0.9081601, 3.146707), (1.888332, 3.817437), (2.769828, 1.146967)\}.$$

Thus $\pi(\mathcal{V}(f, g))$ consists of three points which are roots of $h = 4x^3 - 15x^2 + 4x + 19$, where $\langle h \rangle = \langle f, g \rangle \cap \mathbb{K}[x]$. However, the resultant is

$$\text{Res}(f, g; y) = 160(4x^3 - 15x^2 + 4x + 19)(x - 1)^4,$$

whose roots are shown on the x -axis, including the point $x = 1$ with multiplicity four. \diamond

Corollary 2.1.12. *If the coefficients of the highest powers of y in f and g do not involve x and if $\gcd(f, g) = 1$, then $\mathcal{V}(\langle f, g \rangle \cap \mathbb{K}[x]) = \mathcal{V}(\text{Res}(f, g; x))$.*

Lemma 2.1.13. *When \mathbb{K} is algebraically closed, the system of bivariate polynomials*

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

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

Proof. We instead show that $\mathcal{V}(f, g)$ is infinite if and only if f and g do have a common factor. If f and g have a common 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. Its projection to at least one of the two coordinate axes is infinite. Suppose that the projection π onto the x -axis is infinite. Set $I := \langle f, g \rangle \cap \mathbb{K}[x]$, the elimination ideal. By the Theorem 2.1.10, we have $\pi(\mathcal{V}(f, g)) \subset \mathcal{V}(I) \subset \mathcal{V}(\text{Res}(f, g; y))$. Since $\pi(\mathcal{V}(f, g))$ is infinite, $\mathcal{V}(\text{Res}(f, g; y)) = \mathbb{K}$, which implies that $\text{Res}(f, g; y)$ is the zero polynomial. By Theorem 2.1.2 and Lemma 2.1.9, f and g have a common factor. \square

Let $f, g \in \mathbb{K}[x, y]$ and suppose that neither $\text{Res}(f, g; x)$ nor $\text{Res}(f, g; y)$ vanishes so that f and g have no 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 2.1.14 (Elimination Algorithm).

INPUT: Polynomials $f, g \in \mathbb{K}[x, y]$ with $\gcd(f, g) = 1$.

OUTPUT: $\mathcal{V}(f, g)$.

First, compute the resultant $\text{Res}(f, g; x)$, which is not the zero polynomial. Then, for every root a of $\text{Res}(f, g; y)$, find all common roots b of $f(a, y)$ and $g(a, y)$. 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

$$f(x, y) = g(x, y) = 0, \quad (2.8)$$

to that of finding the roots of univariate polynomials.

Remark 2.1.15. This method of finding a univariate polynomial $h(x)$ whose roots are the x -coordinates of points in $\mathcal{V}(f, g)$, then substituting the roots of h into f and g to compute $\mathcal{V}(f, g)$ is referred to as *back solving*.

Often we only want to count the number of solutions to a system (2.8), or give a realistic bound for this number which is attained when f and g are generic polynomials. The most basic such bound was given by Etienne Bézout in 1779. Our first step toward establishing Bézout's Theorem is an exercise in algebra and some bookkeeping. The monomials in a polynomial of degree n in the variables x, y are indexed by the set

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

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

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

Lemma 2.1.16. *The generic resultant $\text{Res}(f, g; y)$ is a polynomial in x of degree mn .*

Proof. Write

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

where the coefficients are univariate polynomials in x ,

$$f_j(x) := \sum_{i=0}^j f_{i,m-j} x^i \quad \text{and} \quad g_j(x) := \sum_{i=0}^j g_{i,n-j} x^i.$$

Then the Sylvester matrix $\text{Syl}(f, g; y)$ (2.2) has entries the polynomials $f_i(x)$ and $g_j(x)$, and so the resultant $\text{Res}(f, g; y) = \det(\text{Syl}(f, g; y))$ is a univariate polynomial in x .

As in the proof of Lemma 2.1.3, 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

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

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

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

This is a polynomial of degree at most

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

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

We complete the proof by showing that the resultant does indeed have degree mn . The product $f_m(x)^n \cdot g_0(x)^m$ of the entries along the main diagonal of the Sylvester matrix has leading term $f_{m,0}^n \cdot g_{0,n}^m x^{mn}$ and constant term $f_{0,0}^n \cdot g_{0,n}^m$, and these are the only terms in the expansion of the determinant of the Sylvester matrix involving either of these monomials in the coefficients $f_{i,j}, g_{k,l}$. \square

We now state and prove Bézout's Theorem. By general, we mean an element of the complement of a proper subvariety. This notion is covered in more detail on Section 3.1.

Theorem 2.1.17 (Bézout's Theorem). *Two polynomials $f, g \in \mathbb{K}[x, y]$ either have a common factor or else $|\mathcal{V}(f, g)| \leq \deg(f) \cdot \deg(g)$.*

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 2.1.13, if f and g are relatively prime, then $\mathcal{V}(f, g)$ is finite. Let us extend \mathbb{K} to its algebraic closure $\overline{\mathbb{K}}$, which is infinite. We may change coordinates, replacing f by $f(A(x, y))$ and g by $g(A(x, y))$, where A is an invertible affine transformation,

$$A(x, y) = (ax + by + c, \alpha x + \beta y + \gamma), \quad (2.9)$$

with $a, b, c, \alpha, \beta, \gamma \in \overline{\mathbb{K}}$ and $a\beta - ab \neq 0$. As $\overline{\mathbb{K}}$ is infinite, we can choose these parameters so that the constant terms and terms with highest power of x in each of f and g are non-zero. By Lemma 2.1.16, this implies that the resultant $\text{Res}(f, g; y)$ has degree at most mn and thus at most mn zeroes. If we set $I := \langle f, g \rangle \cap \overline{\mathbb{K}}[x]$, then this also implies that $\mathcal{V}(I) = \mathcal{V}(\text{Res}(f, g; x))$, by Corollary 2.1.12.

We can furthermore choose the parameters in A so that the projection $\pi: (x, y) \mapsto x$ is 1-1 on $\mathcal{V}(f, g)$, as $\mathcal{V}(f, g)$ is finite and $\overline{\mathbb{K}}$ infinite. Thus

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

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

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

$$f := \prod_{i=1}^m (x - a_i) = 0 \quad \text{and} \quad g := \prod_{i=1}^n (y - b_i) = 0 \quad (2.10)$$

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 $\text{Res}(f, g; y)$ 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 $\text{Res}(f, g; y)$ are polynomials in the $\binom{m+2}{2} + \binom{n+2}{2}$ coefficients of f and g . Neither is the zero polynomial, as they do not vanish when evaluated at the coefficients of the polynomials (2.10). Thus the set of pairs of polynomials (f, g) with $\mathcal{V}(f, g)$ consisting of mn points in \mathbb{K}^2 is the complement of a proper subvariety of $\mathbb{K}^{\binom{m+2}{2} + \binom{n+2}{2}}$. \square

Exercises

- Verify the claims in the proof of Lemma 2.1.3. This may involve unique factorization in polynomial rings and the Nullstellensatz.
- Using the formula (2.3) deduce the Poisson formula for the resultant of univariate polynomials f and g ,

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

where a_1, \dots, a_m are the roots of f .

3. Suppose that the polynomial $g = g_1 \cdot g_2$ factors. Show that the resultant also factors, $\text{Res}(f, g; x) = \text{Res}(f, g_1; x) \cdot \text{Res}(f, g_2; x)$.
4. Prove the equality of the two formulas for the discriminant in Example 2.1.5.
5. Compute the discriminant of a general cubic $x^3 + ax^2 + bx + c$ by taking the determinant of a 5×5 matrix. Show that the discriminant of the depressed quartic $x^4 + ax^2 + bx + c$ is

$$16a^4c - 4a^3b^2 - 128a^2c^2 + 144ab^2c - 27b^4 + 256c^3.$$

2.2 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, Gröbner bases are also a useful theoretical tool. They will reappear in later chapters in both guises.

A motivating problem is that of recognizing when a polynomial $f \in \mathbb{K}[x_1, \dots, x_n]$ lies in an ideal I . When 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 $I = \langle f(x), g(x), \dots, h(x) \rangle$ 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 $\varphi(x) := \text{gcd}(f(x), g(x), \dots, h(x))$.
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 simplification, 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.

A *monomial* is a product of powers of the variables x_1, \dots, x_n . The *exponent* 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, multiplication of monomials corresponds to vector addition.

Definition 2.2.1. A *monomial ideal* $I \subset \mathbb{K}[x_1, \dots, x_n]$ 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 . \diamond

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 exponents, divisibility of monomials corresponds to componentwise comparison of vectors.

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

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

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

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 (partially ordered 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 $I = \langle x^d \rangle$. Thus all univariate monomial ideals have the form $\langle x^d \rangle$ for some $d \geq 0$.

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 2.2 represent the

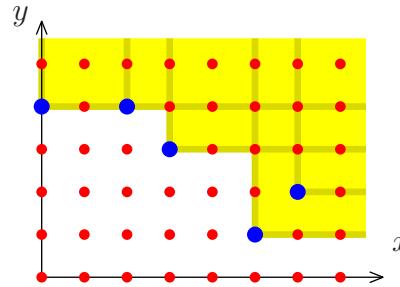


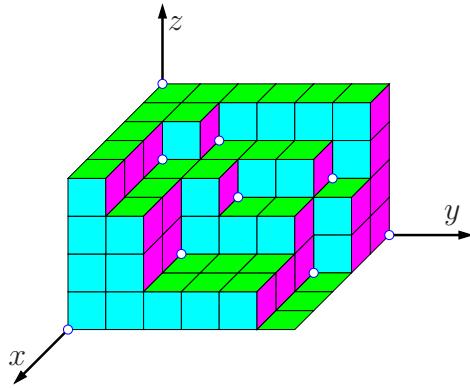
Figure 2.2: Exponents of monomials in the ideal $\langle y^4, x^2y^4, x^3y^3, x^5y, x^6y^2 \rangle$.

monomials in the ideal $I := \langle y^4, x^2y^4, x^3y^3, x^5y, x^6y^2 \rangle$, with the generators marked. From this picture we see that I is minimally generated by y^4 , x^3y^3 , and x^5y .

Since $x^a y^b \in I$ implies that $x^{a+c} y^{b+d} \in I$ for any $(c, d) \in \mathbb{N}^2$, 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

$x^a y^b \in I$. 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 I consists of the monomials which are on the boundary of $O(I)$. 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.

Lemma 2.2.2 (Dickson's Lemma). *Every monomial ideal is finitely generated.*

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

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

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

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

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

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

Note that $I_0 \subset I_1 \subset \dots \subset I_\infty$. 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, let $x^\alpha y^b$ be a monomial in I . 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 $x^\alpha y^b$. Thus G generates I . \square

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

$$I_1 \subset I_2 \subset I_3 \subset \dots$$

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

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 monomial order. An order \succ on monomials in $\mathbb{K}[x_1, \dots, x_n]$ is *total* if for monomials x^α and x^β exactly one of the following holds

$$x^\alpha \succ x^\beta \quad \text{or} \quad x^\alpha = x^\beta \quad \text{or} \quad x^\alpha \prec x^\beta.$$

(Note that we use both \succ and \prec , where $x^\alpha \prec x^\beta$ if and only if $x^\beta \succ x^\alpha$.)

Definition 2.2.3. A *monomial order* on $\mathbb{K}[x_1, \dots, x_n]$ is a total order \succ on the monomials in $\mathbb{K}[x_1, \dots, x_n]$ 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 2.2.3 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 2.2.4. *Monomial orders are exactly the well-orderings \succ on monomials that satisfy Condition (ii) of Definition 2.2.3.*

Proof. Let \succ be a well-ordering on monomials that satisfies Condition (ii) of Definition 2.2.3. Suppose that \succ is not a monomial order. Then there is some monomial x^α with $1 \succ x^\alpha$. By Condition (ii), we have $1 \succ x^\alpha \succ x^{2\alpha} \succ x^{3\alpha} \succ \dots$, which contradicts \succ being a well-order. Thus 1 is the \succ -minimal monomial.

Let \succ be a monomial order and M be any set of monomials. Let I be the ideal generated by M . By Dickson's Lemma, I is generated by a finite set G of monomials. We may assume that $G \subset M$, for if $x^\alpha \in G \setminus M$, then as M generates I , there is some $x^\beta \in M$ that divides x^α , and so we may replace x^α by x^β in G . After finitely many such replacements, we will have that $G \subset M$. 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 $x^\alpha \succ x^\beta \succ x^\gamma$. \square

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 2.2.5. Recall that 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_{lex} on $\mathbb{K}[x_1, \dots, x_n]$ 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_{dlx} on $\mathbb{K}[x_1, \dots, x_n]$ 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_{drl} on $\mathbb{K}[x_1, \dots, x_n]$ is defined by

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

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

$$x^\alpha \succ_\omega x^\beta \iff \omega \cdot \alpha > \omega \cdot \beta.$$

If all components of ω are positive, then \succ_ω satisfies the two conditions of Definition 2.2.3. Its only failure to be a monomial order is that it may not be a total order on monomials. (For example, consider $\omega = (1, 1, \dots, 1)$, then $\omega \cdot \alpha$ is the total degree of x^α .) This may be remedied by picking a monomial order to break ties. For example, if we use \succ_{lex} , then we get a monomial order

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

Another way to do this is to break the ties with a different monomial order, or a different weight, and this may be done recursively.

A monomial order is *graded* if it refines the total-degree partial order $\succ_{(1,1,\dots,1)}$. ◊

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

Remark 2.2.6. We compare the first three orders on monomials of degrees 1 and 2 in $\mathbb{K}[x, y, z]$ where the variables are ordered $x \succ y \succ z$.

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

A *term* is a product ax^α of a non-zero scalar $a \in \mathbb{K}^\times$ with a monomial x^α . Any monomial order \succ extends to terms by setting $ax^\alpha \succ bx^\beta$ if $x^\alpha \succ x^\beta$ and $ab \neq 0$. We also write $ax^\alpha \succeq bx^\beta$ when $ab \neq 0$ and $x^\alpha \succeq x^\beta$. This *term order* is not a partial order, but it is *well-founded* in that it does not admit an infinite strictly decreasing chain.

The *initial term* $\text{in}_\succ(f)$ of a polynomial $f \in \mathbb{K}[x_1, \dots, x_n]$ is the term of f that is maximal with respect to \succ . 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 $\text{in}(f)$. In Exercise 8, you will show that taking initial terms is multiplicative, which is a consequence that \succ respects the multiplication of monomials.

Example 2.2.7. The initial terms of a polynomial f with a weighted partial order \succ_ω have a geometric interpretation in terms of the Newton polytope (see Section A.1.1) of f . For example, suppose that f is

$$x^2 + 2x^3 + 3y + 5x^2y + 7y^2 + 11xy^2 + 13x^2y^2 + 17y^3 + 19xy^3 + 23y^4.$$

Figure 2.3 shows the exponent vectors of terms of f , along with the Newton polygon of f . Then $\text{in}_{(1,1)}f = 13x^2y^2 + 19xy^3 + 23y^4$, the terms of f of total degree 4. Also,

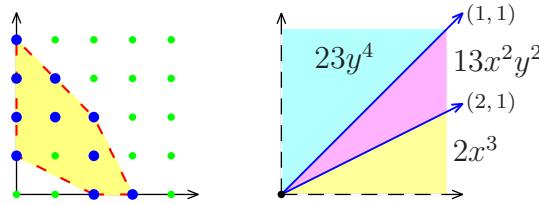


Figure 2.3: Newton polygon and weights.

$\text{in}_{(1,2)}f = 2x^3 + 13x^2y^2$. Other choices for $\omega \in \mathbb{R}_>^2$ give monomials, as shown on the right in Figure 2.3, where we label the cones with the corresponding monomials. \diamond

The *initial ideal* $\text{in}_\succ(I)$ (or $\text{in}(I)$) of an ideal $I \subset \mathbb{K}[x_1, \dots, x_n]$ is the ideal generated by the initial terms of polynomials in I ,

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

Note that every monomial in $\text{in}_\succ(I)$ arises as $\text{in}_\succ(f)$ for some $f \in I$.

We make the most important definition of this section.

Definition 2.2.8. Let $I \subset \mathbb{K}[x_1, \dots, x_n]$ 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 $\text{in}_\succ(I)$ is generated by the initial terms of polynomials in G , that is, if

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

Notice that if G is a Gröbner basis and $G \subset G'$, then G' is also a Gröbner basis. Note also that I is a Gröbner basis for I , and every Gröbner basis contains a finite subset that is also a Gröbner basis, by Dickson's Lemma.

We justify our use of the term ‘basis’ in ‘Gröbner basis’.

Lemma 2.2.9. *If G is a Gröbner basis for I with respect to a monomial order \succ , then G generates I .*

Proof. Let $f \in I$. Since $\{\text{in}(g) \mid g \in G\}$ generates $\text{in}(I)$, there is a polynomial $g \in G$ whose initial term $\text{in}(g)$ divides the initial term $\text{in}(f)$ of f . Thus there is some term ax^α so that

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

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

We will prove the lemma by induction on $\text{in}(f)$ for $f \in I$. Suppose first that $f \in I$ is a polynomial whose initial term $\text{in}(f)$ is the \succ -minimal monomial in $\text{in}(I)$. Then $f_1 = 0$ and so $f \in \langle G \rangle$. Suppose now that $I \neq \langle G \rangle$, and let $f \in I$ be a polynomial with $\text{in}(f)$ is \succ -minimal among all $f \in I \setminus \langle G \rangle$. But then $f_1 = f - cx^\alpha g \in I$ and as $\text{in}(f) \succ \text{in}(f_1)$, we must have that $f_1 \in \langle G \rangle$, which implies that $f \in \langle G \rangle$, a contradiction. \square

An immediate consequence of Dickson's Lemma and Lemma 2.2.9 is the following Gröbner basis version of the Hilbert Basis Theorem.

Theorem 2.2.10 (Hilbert Basis Theorem). *Every ideal $I \subset \mathbb{K}[x_1, \dots, x_n]$ has a finite Gröbner basis with respect to any given monomial order.*

Example 2.2.11. 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

$$\begin{aligned} & x^3z + y^3z + z - 1, \\ & xy^3 + xz^3 + x - 1, \\ & x^3y + yz^3 + y - 1, \\ & y^4z - yz^4 - y + z, \\ & 2xyz^4 + xyz + xy - xz - yz, \\ & 2y^3z^3 - x^3 + y^3 + z^3 + x^2 - y^2 - z^2, \\ & y^6 - z^6 - y^5 + y^3z^2 - 2x^2z^3 - y^2z^3 + z^5 + y^3 - z^3 - x^2 - y^2 + z^2 + x, \end{aligned}$$

$$\begin{aligned}
& x^6 - z^6 - x^5 - y^3z^2 - x^2z^3 - 2y^2z^3 + z^5 + x^3 - z^3 - x^2 - y^2 + y + z, \\
& 2z^7 + 4x^2z^4 + 4y^2z^4 - 2z^6 + 3z^4 - x^3 - y^3 + 3x^2z + 3y^2z - 2z^3 + x^2 + y^2 - 2xz - 2yz - z^2 + z - 1, \\
& 2yz^6 + y^4 + 2yz^3 + x^2y - y^3 + yz^2 - 2z^3 + y - 1, \\
& 2xz^6 + x^4 + 2xz^3 - x^3 + xy^2 + xz^2 - 2z^3 + x - 1,
\end{aligned}$$

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

$$\begin{aligned}
& 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} - 48323578629755458784yz^{17} - 48323578629755458784yz^{16} \\
& - 20036605773313239008yz^{15} - 81914358896780594768yz^{14} - 97825781128529343392yz^{13} \\
& - 53038074105829162080yz^{12} - 78673143256979923752yz^{11} - 99888372899311588584yz^{10} \\
& - 63645688926994994496yz^9 - 37126651874080413456yz^8 - 43903739120936361944yz^7 \\
& - 34474748168788955352yz^6 - 9134334984892800136yz^5 - 5893119345092129120yz^4 \\
& - 4125183541564490384yz^3 - 1178623869018425824yz^2 - 2062591770782245192yz \\
& - 1178623869018425824y + 46665645155349846336z^{33} - 52561386330338650688z^{32} \\
& + 25195872352020329920z^{31} + 281567691623729527232z^{30} - 193921774307243786944z^{29} \\
& - 22383823960598695936z^{28} + 817065337246009690992z^{27} - 163081046857587235248z^{26} \\
& - 427705590368834030336z^{25} + 1390578168371820853808z^{24} + 390004343684846745808z^{23} \\
& - 980322197887855981664z^{22} + 1345425117221297973876z^{21} + 1287956065939036731676z^{20} \\
& - 953383162282498228844z^{19} + 631202347310581229856z^{18} + 1704301967869227396024z^{17} \\
& - 155208567786555149988z^{16} - 16764066862257396505z^{15} + 1257475403277150700961z^{14} \\
& + 526685968901367169598z^{13} - 164751530000556264880z^{12} + 491249531639275654050z^{11} \\
& + 457126308871186882306z^{10} - 87008396189513562747z^9 + 15803768907185828750z^8 \\
& + 139320681563944101273z^7 - 17355919586383317961z^6 - 50777365233910819054z^5 \\
& - 4630862847055988750z^4 + 8085080238139562826z^3 + 1366850803924776890z^2 \\
& - 3824545208919673161z - 2755936363893486164, \\
& 589311934509212912x + 589311934509212912y - 87966378396509318592z^{33} \\
& + 133383402531671466496z^{32} - 59115312141727767552z^{31} - 506926807648593280128z^{30} \\
& + 522141771810172334272z^{29} + 48286434009450032640z^{28} - 1434725988338736388752z^{27} \\
& + 629971811766869591712z^{26} + 917986002774391665264z^{25} - 2389871198974843205136z^{24} \\
& - 246982314831066941888z^{23} + 2038968926105271519536z^{22} - 2174896389643343086620z^{21} \\
& - 1758138782546221156976z^{20} + 2025390185406562798552z^{19} - 774542641420363828364z^{18}
\end{aligned}$$

$$\begin{aligned}
& -2365390641451278278484z^{17} + 627824835559363304992z^{16} + 398484633232859115907z^{15} \\
& -1548683110130934220322z^{14} - 500192666710091510419z^{13} + 551921427998474758510z^{12} \\
& -490368794345102286410z^{11} - 480504004841899057384z^{10} + 220514007454401175615z^9 \\
& +38515984901980047305z^8 - 136644301635686684609z^7 + 17410712694132520794z^6 \\
& +58724552354094225803z^5 + 15702341971895307356z^4 - 7440058907697789332z^3 \\
& -1398341089468668912z^2 + 3913205630531612397z + 2689145244006168857,
\end{aligned}$$

consisting of 4 polynomials with largest degree 34 and significantly larger coefficients. \diamond

Exercises

1. Prove the equivalence of conditions (i) and (ii) in Definition 2.2.1.
2. Show that the radical of a monomial ideal is a monomial ideal, and that a monomial ideal is radical if and only if it is square-free. (Square-free means that in each of its minimal generators 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 2.2.1, 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_{lex} , \succ_{dlx} , and \succ_{drl} , are monomial orders. Show that if the coordinates of $\omega \in \mathbb{R}_>^n$ are linearly independent over \mathbb{Q} , then \succ_ω is a monomial order. Show that each of \succ_{lex} , \succ_{dlx} , and \succ_{drl} are weighted orders.
8. Suppose that \succ is a term order. Prove that for any two non-zero polynomials f, g , we have $\text{in}_\succ(fg) = \text{in}_\succ(f)\text{in}_\succ(g)$.
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.

2.3 Algorithmic aspects of Gröbner bases

Many practical algorithms to study and manipulate ideals and varieties are based on Gröbner bases. The foundations for algorithms involving Gröbner bases are the multivariate division algorithm and Buchberger's algorithm to compute Gröbner bases. As in Chapter 1, we will often write $\mathbb{K}[x]$ for the multivariate polynomial ring $\mathbb{K}[x_1, \dots, x_n]$.

Both steps in the algorithm for ideal membership in one variable relied on 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 2.2.9. This leads to the *multivariate division algorithm*, which is a cornerstone of the theory of Gröbner bases.

Algorithm 2.3.1 (Multivariate division algorithm).

INPUT: Polynomials g_1, \dots, g_m, f in $\mathbb{K}[x]$ and a monomial order \succ .

OUTPUT: Polynomials q_1, \dots, q_m and r such that

$$f = q_1g_1 + q_2g_2 + \dots + q_mg_m + r, \quad (2.11)$$

where no term of r is divisible by an initial term of any polynomial g_i and we also have $\text{in}(f) \succeq \text{in}(r)$, and $\text{in}(f) \succeq \text{in}(q_ig_i)$, for each $i = 1, \dots, m$.

INITIALIZE: Set $r := f$ and $q_1 := 0, \dots, q_m := 0$. Perform the following steps.

- (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 $\text{in}(g_i)$. Choose j minimal such that $\text{in}(g_j)$ divides x^α and set $bx^\beta := \text{in}(g_j)/ax^\alpha$. 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, \dots, g_m . With each reduction, the largest term in r divisible by some $\text{in}(g_i)$ decreases with respect to \succ . Since the term order \succ is well-founded, this algorithm must terminate after a finite number of steps. Every time the algorithm executes step (1), condition (2.11) holds. We also always have $\text{in}(f) \succeq \text{in}(r)$ because it holds initially, and with every reduction any new terms of r are less than the term that was canceled. Lastly, $\text{in}(f) \succeq \text{in}(q_ig_i)$ holds, because $\text{in}(q_ig_i)$ is a term of r in some previous step of the algorithm. \square

Given a list $G = (g_1, \dots, g_m)$ 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 $f \bmod G$ depends on the monomial order \succ , in general it will also depend upon the order of the polynomials (g_1, \dots, g_m) . For example, in the degree lexicographic order

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

Thus 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 2.3.2 (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 $f \bmod G = 0$.*

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

When G is a Gröbner basis for an ideal I and $f \in \mathbb{K}[x]$, no term of the remainder $f \bmod G$ lies in the initial ideal of I . A monomial x^α is *standard* if $x^\alpha \notin \text{in}(I)$. The images of standard monomials in the ring $\mathbb{K}[x]/\text{in}(I)$ form a vector space basis. Much more interesting is the following theorem.

Theorem 2.3.3. *Let $I \subset \mathbb{K}[x]$ be an ideal and \succ a monomial order. Then the images of standard monomials in $\mathbb{K}[x]/I$ 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 $f \bmod G$ represent the same element of $\mathbb{K}[x]/I$. Since $f \bmod G$ is a linear combination of standard monomials, the standard monomials span $\mathbb{K}[x]/I$.

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

By Theorem 2.3.3, if we have a monomial order \succ and an ideal I , then for every polynomial $f \in \mathbb{K}[x]$, there is a unique polynomial \bar{f} which involves only standard monomials such that f and \bar{f} have the same image in the quotient ring $\mathbb{K}[x]/I$. Moreover, $\bar{f} = f \bmod G$, where G is any finite Gröbner basis of I with respect to the monomial order \succ , and thus \bar{f} may be computed from f and G using the division algorithm. This unique representative \bar{f} of f is called the *normal form* of f modulo I and the division algorithm with a Gröbner basis for I is called *normal form reduction*.

A Gröbner basis enables computation in the quotient ring $\mathbb{K}[x]/I$ using the operations of the polynomial ring and linear algebra, by Theorem 2.3.3. Indeed, let G be a finite Gröbner basis for an ideal I with respect to a monomial order \succ and suppose that $f, g \in \mathbb{K}[x]/I$ are in normal form, 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 $fg \bmod G$, where this product is taken in the polynomial ring.

Theorem 2.2.10, 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.

A given set G 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 a Gröbner basis is to generate polynomials in the ideal with ‘new’ initial terms. This is the *raison d’être* for the following definition.

Definition 2.3.4. The *least common multiple*, $\text{lcm}\{ax^\alpha, bx^\beta\}$ of two terms ax^α and bx^β is the minimal monomial x^γ divisible by both x^α and x^β . Here, the exponent vector γ is the componentwise maximum of α and β .

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

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

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

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

Theorem 2.3.5 (Buchberger's Criterion). *A set G of polynomials is a Gröbner basis for the ideal $\langle G \rangle$ with respect to a monomial order \succ if and only if for all pairs $f, g \in G$,*

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

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

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

Given a list $h = (h_1, \dots, h_m)$ of polynomials in $\mathbb{K}[x_1, \dots, x_n]$ let $\text{mm}(h)$ be the largest monomial appearing in one of h_1g_1, \dots, h_mg_m . This will be the monomial in at least one of the initial terms $\text{in}(h_1g_1), \dots, \text{in}(h_mg_m)$. Let $j(h)$ be the minimum index i for which $\text{mm}(h)$ is the monomial of $\text{in}(h_ig_i)$.

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

$$f = h_1g_1 + \dots + h_mg_m \tag{2.12}$$

for which $\text{mm}(h)$ minimal among all lists satisfying (2.12). Of these, let h be a list with $j := j(h)$ maximal. We claim that $\text{mm}(h)$ is the monomial of $\text{in}(f)$, which implies that $\text{in}(g_j)$ divides $\text{in}(f)$.

Otherwise, $\text{mm}(h) \succ \text{in}(f)$, and the initial term $\text{in}(h_jg_j)$ is canceled in the sum (2.12). Thus there is some index k such that $\text{mm}(h)$ is the monomial of $\text{in}(h_kg_k)$. By our assumption on j , we have $k > j$. Let $x^\beta := \text{lcm}\{\text{in}(g_j), \text{in}(g_k)\}$, the monomial which is canceled in $\text{Spol}(g_j, g_k)$. Since $\text{in}(g_j)$ and $\text{in}(g_k)$ both divide $\text{mm}(h)$, both divide $\text{in}(h_jg_j)$, and there is some term ax^α such that $ax^\alpha x^\beta = \text{in}(h_jg_j)$. Set $cx^\gamma := \text{in}(h_jg_j)/\text{in}(g_k)$. Then

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

By Buchberger's criterion for G , there are polynomials q_1, \dots, q_m with

$$\text{Spol}(g_j, g_k) = q_1 g_1 + \dots + q_m g_m,$$

and we may assume that $\text{in}(q_i g_i) \preceq \text{in}(\text{Spol}(g_j, g_k)) \prec x^\beta$, by the division algorithm and the construction of $\text{Spol}(g_j, g_k)$.

Define a new list h' of polynomials,

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

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

$$\begin{aligned} \sum_i h_i g_i + ax^\alpha \sum_i q_i g_i - \text{in}(h_j) g_j + cx^\gamma g_k \\ = f + ax^\alpha \text{Spol}(g_j, g_k) - ax^\alpha \text{Spol}(g_j, g_k) = f, \end{aligned}$$

so h' is a list satisfying (2.12).

We have $\text{in}(q_i g_i) \preceq \text{in}(\text{Spol}(g_j, g_k))$, so $\text{in}(ax^\alpha q_i g_i) \prec x^\alpha x^\beta = \text{mm}(h)$. But then $\text{mm}(h') \preceq \text{mm}(h)$. By the minimality of $\text{mm}(h)$, we have $\text{mm}(h') = \text{mm}(h)$. Since $\text{in}(h_j - \text{in}(h_j)) \prec \text{in}(h_j)$, we have $j(h') > j = j(h)$, which contradicts our choice of h . \square

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 satisfies Buchberger's Criterion.

Algorithm 2.3.6 (Buchberger's Algorithm). Let $G = (g_1, \dots, g_m)$ be generators for an ideal I and \succ a monomial order. For each $1 \leq i < j \leq m$, let $h_{ij} := \text{Spol}(g_i, g_j) \bmod G$. If each reduction vanishes, then by Buchberger's Criterion, G is a Gröbner basis for I with respect to \succ . 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 subfield of \mathbb{K} .

Corollary 2.3.7. *Let $f_1, \dots, f_m \in \mathbb{k}[x_1, \dots, x_n]$ be polynomials and \succ a monomial order. Then there is a Gröbner basis $G \subset \mathbb{k}[x_1, \dots, x_n]$ for the ideal $\langle f_1, \dots, f_m \rangle$ in $\mathbb{K}[x_1, \dots, x_n]$ with respect to the monomial order \succ .*

Example 2.3.8. Consider applying the Buchberger algorithm to $G = (x^2, xy + y^2)$ with any monomial order where $x \succ y$. First

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

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. \diamond

Among the polynomials h_{ij} computed at each stage of Buchberger's algorithm are those where one of $\text{in}(g_i)$ or $\text{in}(g_j)$ divides the other. Suppose that $\text{in}(g_i)$ divides $\text{in}(g_j)$ with $i \neq j$. Then $\text{Spol}(g_i, g_j) = g_j - ax^\alpha g_i$, where ax^α is some term. This has strictly smaller initial term than does g_j and so we never use g_j to compute $h_{ij} := \text{Spol}(g_i, g_j) \bmod G$. It follows that $g_j - h_{ij}$ lies in the ideal generated by $G \setminus \{g_j\}$ (and *vice-versa*), 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 Buchberger's algorithm:

$$\begin{aligned} &\text{In each step, initially compute } h_{ij} \text{ for those } i \neq j \\ &\text{where } \text{in}(g_i) \text{ divides } \text{in}(g_j), \text{ and replace } g_j \text{ by } h_{ij}. \end{aligned} \tag{2.13}$$

In some cases this computes the Gröbner basis. Another improvement, identifying S-polynomials that reduce to zero and therefore need not be computed, is given in Exercise 3.

A Gröbner basis G is *reduced* if the initial terms of polynomials in G have coefficient 1 and if for each $g \in G$, no monomial of g is divisible by an initial term of another element of G . 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 g of a reduced Gröbner basis have the form,

$$x^\alpha - \sum_{\beta \in \mathcal{B}} a_\beta x^\beta, \tag{2.14}$$

where $x^\alpha = \text{in}(g)$ is the initial term and \mathcal{B} consists of exponent vectors of standard monomials. This rewrites the nonstandard initial monomial in terms of standard monomials. In this way, a Gröbner basis is a system of rewriting rules for polynomials. A reduced Gröbner basis has one generator for every generator of the initial ideal.

Example 2.3.9. Let M be a $m \times n$ matrix which is the matrix of coefficients of m linear forms g_1, \dots, g_m in $\mathbb{K}[x_1, \dots, x_n]$, and suppose that $x_1 \succ x_2 \succ \dots \succ x_n$. We can apply (2.13) 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 (2.13) until the initial terms of the forms g_i have distinct variables, then the forms g_1, \dots, 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 \bar{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 . Thus the Buchberger algorithm is a generalization of Gaussian elimination to non-linear polynomials. \diamond

The form (2.14) of elements in a reduced Gröbner basis G for an ideal I with respect to a given monomial order \succ implies that G depends on the monomial ideal $\text{in}_\succ(I)$, and

thus only indirectly on \succ . That is, if \succ' is a second monomial order with $\text{in}_{\succ'}(I) = \text{in}_{\succ}(I)$, then G is also a Gröbner basis for I with respect to \succ' . While there are uncountably many monomial orders, any given ideal has only finitely many initial ideals.

Theorem 2.3.10. *The set $\text{In}(I)$ of initial ideals of an ideal $I \subset \mathbb{K}[x]$ is finite.*

Proof. For each initial ideal M of I , choose a monomial order \succ_M with $M = \text{in}_{\succ_M}(I)$. Let

$$\mathcal{T} := \{\succ_M \mid M \in \text{In}(I)\}$$

be this set of monomial orders, one for each initial ideal of I .

Suppose that $\text{In}(I)$ and hence T is infinite and let $g_1, \dots, g_m \in \mathbb{K}[x]$ be generators for I . Since each polynomial g_i has only finitely many terms, there is an infinite subset \mathcal{T}_1 of T with the property that any two monomial orders \succ, \succ' in \mathcal{T}_1 will select the same initial terms from each of the g_i ,

$$\text{in}_{\succ}(g_i) = \text{in}_{\succ'}(g_i) \quad \text{for } i = 1, \dots, m.$$

Set $M_1 := \langle \text{in}_{\succ}(g_1), \dots, \text{in}_{\succ}(g_m) \rangle$, where \succ is any monomial order in \mathcal{T}_1 . Either (g_1, \dots, g_m) is a Gröbner basis for I with respect to \succ or else there is a some polynomial g_{m+1} in I whose initial term does not lie in M_1 . Replacing g_{m+1} by $g_{m+1} \bmod (g_1, \dots, g_m)$, we may assume that g_{m+1} has no term in M_1 .

Then there is an infinite subset \mathcal{T}_2 of \mathcal{T}_1 such that any two monomial orders \succ, \succ' in \mathcal{T}_2 will select the same initial term of g_{m+1} , $\text{in}_{\succ}(g_{m+1}) = \text{in}_{\succ'}(g_{m+1})$. Let M_2 be the monomial ideal generated by M_1 and $\text{in}_{\succ}(g_{m+1})$ for some monomial order \succ in \mathcal{T}_2 . As before, either $(g_1, \dots, g_m, g_{m+1})$ is a Gröbner basis for I with respect to \succ , or else there is an element g_{m+2} of I having no term in M_2 .

Continuing in this fashion constructs an increasing chain $M_1 \subsetneq M_2 \subsetneq \dots$ of monomial ideals in $\mathbb{K}[x]$. By Dickson's Lemma, this process must terminate, at which point we will have an infinite subset \mathcal{T}_r of T and polynomials g_1, \dots, g_{m+r} that form a Gröbner basis for I with respect to a monomial order \succ in \mathcal{T}_r , and these have the property that for any other monomial order \succ' in \mathcal{T}_r , we have

$$\text{in}_{\succ}(g_i) = \text{in}_{\succ'}(g_i) \quad \text{for } i = 1, \dots, m+r.$$

But this implies that $\text{in}_{\succ}(I) = \text{in}_{\succ'}(I)$ is an initial ideal for two distinct monomial orders in $\mathcal{T}_r \subset T$, which contradicts the construction of the set T . \square

Definition 2.3.11. A consequence of Theorem 2.3.10 that an ideal I has only finitely many initial ideals is that it has only finitely many reduced Gröbner bases. The union of this finite set of reduced Gröbner bases is a finite generating set for I that is a Gröbner basis for I with respect to any monomial order. Such a generating set is called a *universal Gröbner basis* for the ideal I .

Exercises

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 (2.13)?
2. Use Buchberger's algorithm to compute by hand 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 $f, g \in \mathbb{K}[x]$ be polynomials with relatively prime initial terms, and suppose that their initial coefficients are 1.

- (a) Show that

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

Deduce that the initial monomial of $\text{Spol}(f, g)$ is a multiple of either the initial monomial of f or the initial monomial of g .

- (b) Analyze the steps of the reduction computing $\text{Spol}(f, g) \bmod (f, g)$ using the division algorithm to show that this is zero.

This gives another improvement to Buchberger's algorithm: avoid computing and reducing those S-polynomials of polynomials with relatively prime initial terms.

4. Let U be a universal Gröbner basis for an ideal I in $\mathbb{K}[x_1, \dots, x_n]$. Show that for every subset $Y \subset \{x_1, \dots, x_n\}$ the *elimination ideal* $I \cap \mathbb{K}[Y]$ is generated by $U \cap \mathbb{K}[Y]$.
5. Let \succ be any monomial order and G be a list of homogeneous polynomials. Then for any homogeneous polynomial f , its reduction modulo G is also homogeneous. Show that the reduced Gröbner basis computed by Buchberger's algorithm from G consists of homogeneous polynomials. Deduce that the reduced Gröbner basis of a homogeneous ideal consists of homogeneous polynomials.
6. Let I be a ideal generated by homogeneous linear polynomials. A non-zero linear form f in I is 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. Let $I := \langle x^2 + y^2, x^3 + y^3 \rangle \subset \mathbb{Q}[x, y]$ and suppose that the monomial order \succ is the lexicographic order with $x \succ y$.

- (a) Show 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.

8. (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 by hand Gröbner bases of the ideal $I = \langle y - z^2, z - x^3 \rangle \in \mathbb{Q}[x, y, z]$ with respect to the lexicographic and to the degree reverse lexicographic monomial orders.
9. Let $I \subset \mathbb{K}[x_1, \dots, x_n]$ be an ideal, and fix $f \in \mathbb{K}[x_1, \dots, x_n]$. Then the *saturation* of I with respect to f is the set

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

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

$$(I : f) \subset (I : f^2) \subset (I : f^3) \subset \dots$$

- (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 I is homogeneous and $f = x_n$ the following strategy computes the saturation. Fix the degree reverse lexicographic order \succ where $x_1 \succ x_2 \succ \dots \succ x_n$ and let G be a reduced Gröbner basis of a homogeneous ideal $I \subset \mathbb{K}[x_1, \dots, x_n]$.

- (e) Show that the set

$$\{f \in G \mid x_n \text{ does not divide } f\} \cup \{f/x_n \mid f \in G \text{ and } x_n \text{ divides } f\}$$

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 .

2.4 Solving equations with Gröbner bases

Algorithm 2.1.14 reduced the problem of solving two equations in two variables to that of solving univariate polynomials, using resultants to eliminate a variable. For an ideal $I \subset \mathbb{K}[x]$ whose variety $\mathcal{V}(I)$ consists of finitely many points, this same idea of back solving leads to an algorithm to compute $\mathcal{V}(I)$, provided we can compute the elimination ideals $I \cap \mathbb{K}[x_i]$. Gröbner bases provide a universal algorithm for computing elimination ideals. More generally, ideas from the theory of Gröbner bases can help to understand solutions to systems of equations.

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

$$f_1(x_1, \dots, x_n) = \dots = f_N(x_1, \dots, x_n) = 0, \quad (2.15)$$

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

- (i) Does (2.15) have finitely many solutions?
- (ii) Can we count them, or give (good) upper bounds on their number?
- (iii) Can we *solve* the system (2.15) and find all solutions?
- (iv) When the polynomials have real coefficients, can we count (or bound) the number of real solutions to (2.15)? Or simply find them?

The solutions to (2.15) in \mathbb{K}^n constitute the affine variety $\mathcal{V}(I)$, where I is the ideal generated by the polynomials f_1, \dots, f_N . Algorithms based on Gröbner bases to address Questions (i)-(iv) involve studying I . An ideal I is *zero-dimensional* if, over the algebraic closure $\overline{\mathbb{K}}$ of \mathbb{K} , $\mathcal{V}(I)$ is finite. Thus I is zero-dimensional if and only if its radical \sqrt{I} is zero-dimensional.

Theorem 2.4.1. *An ideal $I \subset \mathbb{K}[x]$ is zero-dimensional if and only if $\mathbb{K}[x]/I$ is a finite-dimensional \mathbb{K} -vector space, if and only if $\mathcal{V}(I) \subset \overline{\mathbb{K}}^n$ is finite.*

When an ideal I is zero-dimensional, we will call the points of $\mathcal{V}(I)$ the *roots of I* .

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, so that $I = \mathcal{I}(\mathcal{V}(I))$, by the Nullstellensatz. Then $\mathbb{K}[x]/I$ is the coordinate ring $\mathbb{K}[X]$ of $X := \mathcal{V}(I)$, consisting of all functions obtained by restricting polynomials to $\mathcal{V}(I)$, and is therefore a subring of the ring of functions on X . 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. In particular, no polynomial in x_1 , except the zero polynomial, vanishes on X . Restriction of polynomials in x_1 to X is therefore an injective map from $\mathbb{K}[x_1]$ to $\mathbb{K}[X]$ which shows that $\mathbb{K}[X]$ is infinite-dimensional.

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

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

is a surjection. But $\mathbb{K}[x]/\langle g_1(x_1)^{M_1}, \dots, g_n(x_n)^{M_n} \rangle$ has dimension $\prod_i M_i \deg(g_i)$, which implies that $\mathbb{K}[x]/I$ is finite-dimensional. \square

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

Corollary 2.4.2. *An ideal $I \subset \mathbb{K}[x]$ 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 Theorem 2.3.3, Theorem 2.4.1 leads to a Gröbner basis criterion/algorithm to solve Question (i).

Corollary 2.4.3. *An ideal $I \subset \mathbb{K}[x]$ is zero-dimensional if and only if for any monomial order \succ , the initial ideal $\text{in}_\succ I$ 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 initial terms of elements of the Gröbner basis include pure powers of all variables.

When I is zero-dimensional, its *degree* is the dimension of $\mathbb{K}[x]/I$ as a \mathbb{K} -vector space, which is the number of standard monomials, by Theorem 2.3.3. A Gröbner basis for I gives generators of the initial ideal which we can use to count the number of standard monomials to determine its degree.

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

Theorem 2.4.4. *Let I be the ideal generated by the polynomials f_i of (2.15). If I is zero-dimensional, then the number of solutions to the system (2.15) 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 (2.15) which do not require a Gröbner basis. For example, Theorem 2.1.17 (Bézout's Theorem in the plane) gives 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 3.5. A common solution x to a square system of equations is *nondegenerate* if the differentials of the equations are linearly independent at x .

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

$$f_1(x_1, \dots, x_n) = \dots = f_n(x_1, \dots, x_n) = 0$$

in \mathbb{K}^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 the *Bézout bound* for such a system. While sharp for generic square systems, few practical problems involve generic systems and other bounds are often needed (see Exercise 5). We discuss such bounds in Chapter 8, where we establish the polyhedral bounds of Kushnirenko's and Bernsteins's Theorems.

We discuss a symbolic method to solve systems of polynomial equations (2.15) 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 $I \subset \mathbb{K}[x]$ 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 $I \subset \mathbb{K}[x]$. Then $g(a_i) = 0$ for every $a = (a_1, \dots, a_n) \in \mathcal{V}(I) \in \mathbb{K}^n$. When \mathbb{K} is algebraically closed, every root of g is the i th coordinate of a point of $\mathcal{V}(I)$.

Proof. First, $g(a_i) = 0$ 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 i th 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 $h \in \sqrt{I}$. But then 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 $g(\xi) = 0$. \square

Theorem 2.4.7. If $g(x_i)$ is a monic eliminant for an ideal $I \subset \mathbb{K}[x]$, then g lies in the reduced Gröbner basis for I with respect to any monomial order in which the pure powers x_i^m of x_i precede variables x_j with $j \neq i$.

Proof. Suppose that \succ is such a monomial order. Then its minimal monomials are $1, x_i, x_i^2, \dots$. 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 . As $g \in I$, we have that $x_i^{\deg(g)} \in \text{in}_\prec(I)$. Let x_i^m be the generator of $\text{in}_\prec(I) \cap \mathbb{K}[x_i]$. Then $m \leq \deg(g)$. Let f be the polynomial in the reduced Gröbner basis of I with respect to \prec whose initial term is x_i^m . Then its remaining terms involve smaller standard monomials and are thus pure powers of x_i . We conclude that $f \in I \cap \mathbb{K}[x_i] = \langle g \rangle$, and so g divides f , so $m = \deg(g)$. As $f-g$ is a polynomial in x_i which lies in I but has degree less than $\deg(g)$, the minimality of f and g implies that $f-g=0$. This proves that g lies in the reduced Gröbner basis. \square

The following theorem relating Gröbner bases and elimination ideals is proven in the exercises.

Theorem 2.4.8. Let $I \subset \mathbb{K}[x]$ be an ideal and let \prec be the lexicographic monomial order with $x_1 \prec x_2 \prec \dots \prec x_n$ and G a Gröbner basis for I with respect to \prec . Then, for each $m = 1, \dots, n$, the polynomials in G that lie in $\mathbb{K}[x_1, \dots, x_m]$ form a Gröbner basis for the elimination ideal $I_m = I \cap \mathbb{K}[x_1, \dots, x_m]$.

These theorems give 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. As we saw in Example 2.2.11, their size can be significantly larger than other Gröbner bases. It is even expensive to compute a univariate eliminant $g(x_i)$ using an *elimination order*, (a monomial order \prec where any pure power x_i^d of x_i precedes any monomial involving any other variable x_j for $j \neq i$). We instead offer the following algorithm.

Algorithm 2.4.9.

INPUT: A zero-dimensional ideal $I \subset \mathbb{K}[x_1, \dots, x_n]$ and a variable x_i .

OUTPUT: A univariate eliminant $g(x_i) \in I$.

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

- (2) Compute the sequence $1 \bmod G, x_i \bmod G, x_i^2 \bmod G, \dots$, until a linear dependence is found,

$$\sum_{j=0}^m a_j(x_i^j \bmod G) = 0, \quad (2.16)$$

where m is minimal. Then

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

is a univariate eliminant.

Proof of correctness. Since I is zero-dimensional, by Corollary 2.4.2 it has an eliminant $g(x_i) \in I$. If $g = \sum_{j=0}^N b_j x_i^j$ then by the ideal membership test (Lemma 2.3.2),

$$0 = g \bmod G = \left(\sum_{j=0}^N b_j x_i^j \right) \bmod G = \sum_{j=0}^N b_j (x_i^j \bmod G),$$

which is a linear dependence among the elements of the sequence $1 \bmod G, x_i \bmod G, x_i^2 \bmod G, \dots$. Thus the algorithm halts during Step (2). 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. \square

Elimination using Gröbner bases gives algorithms for Questions (iii) and (iv). The first step is to understand the optimal form of a Gröbner basis of a zero-dimensional ideal.

Lemma 2.4.10 (Shape Lemma). *Suppose $g = g(x_i)$ 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.*

Suppose that $i = 1$ so that $g = g(x_1)$. Then in the lexicographic monomial order with $x_1 \prec x_2 \prec \dots \prec x_n$, the ideal I has a Gröbner basis of the form:

$$g(x_1), \quad x_2 - g_2(x_1), \quad \dots, \quad x_n - g_n(x_1), \quad (2.17)$$

where $\deg(g) > \deg(g_i)$ for $i = 2, \dots, n$.

If I is generated by polynomials with coefficients in a subfield \mathbb{k} , then the number of points of $\mathcal{V}(I)$ in \mathbb{k}^n equals the number of roots of g in \mathbb{k} .

This is a simplified version of the Shape Lemma, which describes the form of a reduced Gröbner basis for any zero-dimensional ideal in the lexicographic order. Example 2.2.11 gives a zero-dimensional ideal which does not satisfy the hypotheses of Lemma 2.4.10.

Proof. Replacing \mathbb{K} by its algebraic closure does not affect these algebraic statements, as the polynomials g and g_i have coefficients in \mathbb{k} , by Corollary 2.3.7. Suppose that $g = g(x_i)$ is an eliminant. We have

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

the first inequality is 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 has multiple roots, then there is a polynomial h with the same roots as g but with smaller degree. (We may select h to be the square-free part of g .) 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 $g(x_1)$. Then $1, x_1, \dots, x_1^{d-1}$ are standard monomials, and since $\deg(g) = \deg(I)$, there are no others. Thus the lexicographic initial ideal is $\langle x_1^d, x_2, \dots, x_n \rangle$. 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 (2.17) are

$$\{(a_1, \dots, a_n) \mid g(a_1) = 0 \text{ and } a_i = g_i(a_1), i = 2, \dots, n\}.$$

By Corollary 2.3.7, the polynomials g, g_2, \dots, g_n all have coefficients from \mathbb{k} , and so a component a_i lies in \mathbb{k} if the root a_1 of $g(x_1)$ lies in \mathbb{k} . \square

Not all ideals I can have an eliminant g with $\deg(g) = \deg(I)$. For example, let $\mathfrak{m}_0 := \langle x, y \rangle$ be the maximal ideal corresponding to the origin $\{(0, 0)\} \in \mathbb{K}^2$. Then its square $\mathfrak{m}_0^2 = \langle x^2, xy, y^2 \rangle$ has degree three, but any eliminant has degree two.

Failure of the condition $\deg(g) = \deg(I)$ in the Shape Lemma may occur when I is radical. Indeed, when I is radical, $\deg(g(x_i)) = \deg(I)$ if and only if the projection map π_i to the coordinate x_i -axis is one-to-one.

Example 2.4.11. Suppose that the ideal I is generated by the three polynomials,

$$\begin{aligned} f &:= 1574y^2 - 625yx - 1234y + 334x^4 - 4317x^3 + 19471x^2 \\ &\quad - 34708x + 19764 + 45x^2y - 244y^3, \\ g &:= 45x^2y - 305yx - 2034y - 244y^3 - 95x^2 + 655x + 264 + 1414y^2, \text{ and} \\ h &:= -33x^2y + 197yx + 2274y + 38x^4 - 497x^3 + 2361x^2 - 4754x \\ &\quad + 1956 + 244y^3 - 1414y^2. \end{aligned}$$

Then $\mathcal{V}(I)$ is the seven nondegenerate points of Figure 2.4. There are only five points in the projection to the x -axis and four in the projection to the y -axis. The corresponding eliminants have degrees five and four,

$$2x^5 - 29x^4 + 157x^3 - 391x^2 + 441x - 180 \quad 2y^4 - 13y^3 + 28y^2 - 23y + 6 \quad \diamond$$

Nevertheless, when I is radical, $\deg(g) = \deg(I)$ will hold after a generic change of coordinates, as we saw in Example 2.4.11 and as was used in the proof of Bézout's Theorem in the plane (Theorem 2.1.17). In this case, back solving, may be used to find all roots of I over an algebraically closed field, solving Question (iii). It also gives a symbolic algorithm to count the number of real solutions to a system of equations whose ideal satisfies the hypotheses of the Shape Lemma and solves Question (iv).

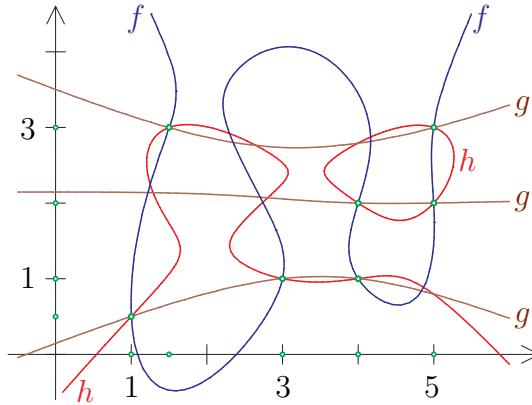


Figure 2.4: The seven points of $\mathcal{V}(f, g, h)$ and their projections.

Algorithm 2.4.12 (Counting real roots).

INPUT: An ideal $I \subset \mathbb{R}[x_1, \dots, x_n]$.

OUTPUT: The number of real points in $\mathcal{V}(I)$, if I satisfies the hypotheses of the Shape Lemma, or else “ I does not satisfy the hypotheses of the Shape Lemma”.

Compute $\dim(I)$ and $\deg(I)$. If I does not have dimension 0, then exit with “ I is not zero-dimensional”, else set $i := 1$.

1. Compute an eliminant $g(x_i)$ for I . If $\deg(g) = \deg(I)$ and $\gcd(g, g') = 1$, then output the number of real roots of g . Else if $i < n$, set $i := i + 1$ and return to (1).
2. If no eliminant has been computed and $i = n$, then output “ I does not satisfy the hypotheses of the Shape Lemma”.

While this algorithm will not successfully compute the number of real points in $\mathcal{V}(I)$ (it would fail for the ideal of Figure 2.4), it may be combined with more sophisticated methods to accomplish that task.

The Shape Lemma describes an optimal form of a Gröbner basis for a zero-dimensional ideal, we remarked that it is typically not optimal to compute a lexicographic Gröbner basis directly, and offered Algorithm 2.4.9 to compute eliminants. This idea extends to the *FGLM algorithm* for Gröbner basis conversion. This takes a Gröbner basis for a zero-dimensional ideal with respect to one monomial order \triangleright and computes a Gröbner basis with respect to a different monomial order \succ .

Algorithm 2.4.13 (FGLM).

INPUT: A Gröbner basis G for a zero-dimensional ideal $I \subset \mathbb{K}[x_1, \dots, x_n]$ with respect to a monomial order \triangleright , 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 $\overline{x^\alpha} := x^\alpha \bmod G$.

- (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 $\{x^\gamma \mid x^\gamma \succ x^\alpha\} \subset \text{in}_\succ H := \langle \text{in}_\succ h \mid h \in H \rangle$, then halt and output H . Otherwise, set x^α to be the \succ -minimal monomial in $\{x^\gamma \notin \text{in}_\succ H \mid x^\gamma \succ x^\alpha\}$ 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 $\mathbb{K}[x]/I$. Thus $\text{in}_\succ H$ is a subset of the initial ideal $\text{in}_\succ I$, and we always have the inequalities

$$|S| \leq \dim_{\mathbb{K}}(\mathbb{K}[x]/I) \quad \text{and} \quad \text{in}_\succ H \subset \text{in}_\succ I.$$

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

When the algorithm halts, every monomial is either in the set $\text{SM} := \{x^\beta \mid \overline{x^\beta} \in S\}$ or else in the monomial ideal $\text{in}_\succ H$. By our choice of x^α in (3), these two sets are disjoint, so that SM is the set of standard monomials for $\text{in}_\succ H$. Since

$$\text{in}_\succ H \subset \text{in}_\succ \langle H \rangle \subset \text{in}_\succ I,$$

and elements of S are linearly independent in $\mathbb{K}[x]/I$, we have

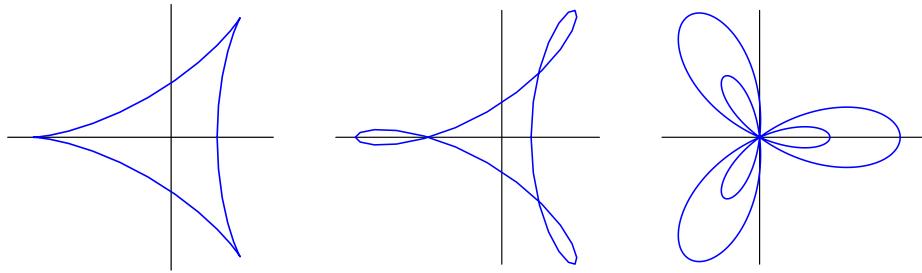
$$|S| \leq \dim_{\mathbb{K}}(\mathbb{K}[x]/I) = \dim_{\mathbb{K}}(\mathbb{K}[x]/\text{in}_\succ I) \leq \dim_{\mathbb{K}}(\mathbb{K}[x]/\text{in}_\succ H) = |S|.$$

Thus $\text{in}_\succ I = \text{in}_\succ H$, 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. \square

Exercises

- Suppose $I \subset \mathbb{K}[x]$ is radical, \mathbb{K} is algebraically closed, and $\mathcal{V}(I) \subset \mathbb{K}^n$ consists of finitely many points. Show that the coordinate ring $\mathbb{K}[x]/I$ of restrictions of polynomial functions to $\mathcal{V}(I)$ has dimension as a \mathbb{K} -vector space equal to the number of points in $\mathcal{V}(I)$.
- The trigonometric curves parameterized by $(\cos(\theta) - \frac{1}{2} \cos(2\theta), \sin(\theta) + \frac{1}{2} \sin(2\theta)/2)$, $(\cos(\theta) - \frac{2}{3} \cos(2\theta), \sin(\theta) + \frac{2}{3} \sin(2\theta))$, and the polar curve $r = 1 + 3 \cos(3\theta)$ are the

cuspidal and trinodal plane quartics, and the flower with three petals, respectively.

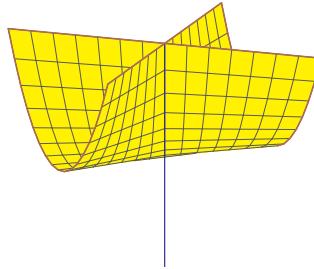


Use elimination to find their implicit equations: Write each as the projection to the (x, y) -plane of an algebraic variety in \mathbb{K}^4 . Hint: These are images of the circle $c^2 + s^2 = 1$ under maps to the (x, y) plane, where the variables (c, s) correspond to $(\cos(\theta), \sin(\theta))$. The graph of the first is given by the three polynomials

$$c^2 + s^2 - 1, \quad x - (c - \frac{1}{2}(c^2 - s^2)), \quad y - (s + sc),$$

using the identities $\cos(2\theta) = \cos^2(\theta) - \sin^2(\theta)$ and $\sin(2\theta) = 2\sin(\theta)\cos(\theta)$.

3. The Whitney umbrella is the image in \mathbb{K}^3 of the map $(u, v) \mapsto (uv, u, v^2)$. Use elimination to find an implicit equation for the Whitney umbrella.



Which points in \mathbb{K}^2 give the handle of the Whitney umbrella?

4. Show that every eliminant of $\mathfrak{m}_0^2 = \langle x^2, xy, y^2 \rangle$ has degree two, even after a change of coordinates.
5. 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?

6. 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?

7. A polynomial is *multilinear* if all exponents are 0 or 1. For example,

$$3xyz - 17xy + 29xz - 37yz + 43x - 53y + 61z - 71$$

is a multilinear polynomial in the variables 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?

8. Let $\mathcal{A} \subset \mathbb{N}^n$ be a finite set of integer vectors, which we regard as exponents of monomials in $\mathbb{K}[x_1, \dots, x_n]$. 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 whose support is the column vectors of $\mathcal{A} = \begin{pmatrix} 0 & 1 & 2 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 2 \end{pmatrix}$.

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. Try to formulate a conjecture about this number of solutions as a function of \mathcal{A} .

9. Fix $m, p \geq 2$. For $\alpha: 1 \leq \alpha_1 < \dots < \alpha_p \leq m+p$, 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} ,

$$E_{245} = \begin{pmatrix} x_{1,1} & 1 & 0 & 0 & 0 & 0 \\ x_{2,1} & 0 & x_{2,3} & 1 & 0 & 0 \\ x_{3,1} & 0 & x_{3,3} & 0 & 1 & 0 \end{pmatrix} \quad E_{356} = \begin{pmatrix} x_{1,1} & x_{1,2} & 1 & 0 & 0 & 0 \\ x_{2,1} & x_{2,2} & 0 & x_{2,4} & 1 & 0 \\ x_{3,1} & x_{3,2} & 0 & x_{3,4} & 0 & 1 \end{pmatrix}.$$

Set $|\alpha| := \alpha_1 - 1 + \alpha_2 - 2 + \dots + \alpha_p - p$ be the number of variables in E_α . For all small m, p , and α , generate $|\alpha|$ random $m \times (m+p)$ matrices $M_1, \dots, M_{|\alpha|}$ and determine the number of isolated solutions to the system of equations

$$\det \begin{pmatrix} E_\alpha \\ M_1 \end{pmatrix} = \det \begin{pmatrix} E_\alpha \\ M_2 \end{pmatrix} = \dots = \det \begin{pmatrix} E_\alpha \\ M_{|\alpha|} \end{pmatrix} = 0.$$

Formulate a conjecture for the number of solutions as a function of m, p , and α .

2.5 Solving equations with linear algebra

We discuss a connection between the solutions to systems of polynomial systems and eigenvalues of linear algebra. This leads to further methods to compute and analyze the roots of a zero-dimensional ideal. The techniques are based on classical results, but their computational aspects have only recently been developed systematically.

Suppose that \mathbb{K} is algebraically closed and $I \subset \mathbb{K}[x_1, \dots, x_n]$ is a zero-dimensional ideal. Our goal is to interpret the coordinates of points in $\mathcal{V}(I)$ in terms of eigenvalues of suitable matrices. This is efficient as numerical linear algebra provides efficient methods to numerically determine the eigenvalues of a complex matrix, and the matrices we use are readily computed using Gröbner basis algorithms.

It is instructive to start with univariate polynomials. Given a monic univariate polynomial $p = c_0 + c_1x + \dots + c_{d-1}x^{d-1} + x^d \in \mathbb{K}[x]$, the *companion matrix* of p is

$$C_p = \begin{pmatrix} 0 & 0 & \cdots & 0 & -c_0 \\ 1 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & \cdots & 0 & -c_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -c_{d-1} \end{pmatrix} \in \mathbb{K}^{d \times d}. \quad (2.18)$$

The eigenvalues of a square matrix A are the roots of its characteristic polynomial $\chi_A(x) = \det(x\text{Id} - A)$, where Id is the appropriately-sized identity matrix. The roots of a polynomial p are the eigenvalues of its companion matrix C_p .

Theorem 2.5.1. *Let $p = c_0 + \dots + c_{d-1}x^{d-1} + x^d \in \mathbb{K}[x]$ be a monic univariate polynomial of degree $d \geq 1$. Then $p(x) = \chi_{C_p}(x)$, characteristic polynomial of its companion matrix C_p . Its companion matrix expresses multiplication by x in the ring $\mathbb{K}[x]/\langle p \rangle$ in the basis $1, x, \dots, x^{d-1}$ of standard monomials.*

Proof. For $d = 1$, the statement is clear, and for $d > 1$ expanding the determinant along the first row of $x\text{Id} - C_p$ yields

$$\det(x\text{Id} - C_p) = x \det(x\text{Id} - C_q) + (-1)^{d+1}(-1)^{d-1}c_0,$$

where C_q is the companion matrix of the polynomial

$$q := c_1 + c_2x + \dots + c_{d-1}x^{d-2} + x^{d-1} = (p - c_0)/x.$$

Applying the induction hypothesis gives the result.

The claim that the matrix C_p expresses multiplication by x in $\mathbb{K}[x]/\langle p \rangle$ in the basis $1, x, \dots, x^{d-1}$ of standard monomials is Exercise 1 below. \square

Let $I \subset \mathbb{K}[x_1, \dots, x_n]$ be a zero-dimensional ideal. By Theorems 2.4.1 and 2.4.4, the \mathbb{K} -vector space $\mathbb{K}[x_1, \dots, x_n]/I$ is finite-dimensional, and the cardinality of the variety

$\mathcal{V}(I)$ is bounded from above by the dimension of $\mathbb{K}[x_1, \dots, x_n]/I$. Given a polynomial $f \in \mathbb{K}[x_1, \dots, x_n]$, write \bar{f} for its residue class in the quotient ring $\mathbb{K}[x_1, \dots, x_n]/I$.

For any $i = 1, \dots, n$, multiplication of an element in $\mathbb{K}[x_1, \dots, x_n]/I$ with the residue class \bar{x}_i of a variable x_i defines an endomorphism m_i ,

$$\begin{aligned} m_i : \mathbb{K}[x_1, \dots, x_n]/I &\longrightarrow \mathbb{K}[x_1, \dots, x_n]/I, \\ \bar{f} &\longmapsto \bar{x}_i \cdot \bar{f} = \bar{x_i f}. \end{aligned}$$

Lemma 2.5.2. *The map $x_i \mapsto m_i$ induces an injection*

$$K[x_1, \dots, x_n]/I \hookrightarrow \text{End}(K[x_1, \dots, x_n]/I).$$

Proof. The map $x_i \mapsto m_i$ induces a map φ from $\mathbb{K}[x_1, \dots, x_n]$ to the endomorphism ring. For polynomials $p, f \in \mathbb{K}[x_1, \dots, x_n]$, the value of $p(m_1, \dots, m_n)(\bar{f})$ is $\bar{p(x_1, \dots, x_n)f}$. This implies that $I \subset \ker(\varphi)$. Setting $f = 1$ shows that $\ker(\varphi) \subset I$. \square

This map $K[x_1, \dots, x_n]/I \hookrightarrow \text{End}(K[x_1, \dots, x_n]/I)$ is the regular representation of $K[x_1, \dots, x_n]/I$. We will use it to study the variety $\mathcal{V}(I)$. Since the vector space $\mathbb{K}[x_1, \dots, x_n]/I$ is finite-dimensional, we may represent each linear multiplication map m_i as a matrix with respect to a fixed basis of $\mathbb{K}[x_1, \dots, x_n]/I$. For this, a basis of standard monomials is both convenient and readily computed.

Let \mathcal{B} be the set of standard monomials for I with respect a monomial order \prec . Let G be a Gröbner basis for I with respect to \prec . For each $i = 1, \dots, n$, let $M_i \in \text{Mat}_{\mathcal{B} \times \mathcal{B}}(K)$ be the matrix representing the endomorphism m_i of multiplication by the variable x_i with respect to the basis \mathcal{B} , which we call the *i-th companion matrix* of the ideal I with respect to \mathcal{B} . The rows and the columns of M_i are indexed by the monomials in \mathcal{B} . For a pair of monomials $x^\alpha, x^\beta \in \mathcal{B}$, the entry of M_i in the row corresponding to x^α and column corresponding to x^β is the coefficient of x^α in $x_i \cdot x^\beta \bmod G$, the normal form of $x_i \cdot x^\beta$.

Lemma 2.5.3. *The companion matrices commute,*

$$M_i \cdot M_j = M_j \cdot M_i \quad \text{for } 1 \leq i < j \leq n.$$

Proof. The matrices $M_i M_j$ and $M_j M_i$ represent the compositions $m_i \circ m_j$ and $m_j \circ m_i$, respectively. This follows as multiplication in $\mathbb{K}[x_1, \dots, x_n]/I$ is commutative. \square

The companion matrices M_1, \dots, M_n generate a subalgebra of $\text{Mat}_{\mathcal{B} \times \mathcal{B}}(\mathbb{K})$ isomorphic to $\mathbb{K}[x_1, \dots, x_n]/I$, by Lemma 2.5.2. As $\mathbb{K}[x_1, \dots, x_n]/I$ is commutative, when \mathbb{K} is algebraically closed this subalgebra has a collection of common eigenvectors whose eigenvalues are characters (homomorphisms to \mathbb{K}) of $\mathbb{K}[x_1, \dots, x_n]/I$. The following fundamental result allows us to identify the eigenvectors with the points of $a \in \mathcal{V}(I)$ with corresponding eigenvalue the evaluation of an element of $\mathbb{K}[x_1, \dots, x_n]/I$ at the point a .

Theorem 2.5.4 (Stickelberger's Theorem). *Suppose that \mathbb{K} is algebraically closed and $I \subset \mathbb{K}[x_1, \dots, x_n]$ is a zero-dimensional ideal. For each $i = 1, \dots, n$ and any $\lambda \in \mathbb{K}$, 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 $a_i = \lambda$.*

Corollary 2.5.5. Let $R \subset \text{End}(\mathbb{K}[x_1, \dots, x_n]/I)$ be the commutative subalgebra generated by the endomorphisms m_1, \dots, m_n . The joint eigenvectors of R correspond to points of $\mathcal{V}(I)$. For $p \in \mathbb{K}[x_1, \dots, x_n]$ and $a \in \mathcal{V}(I)$, the eigenvalue of $p(m_1, \dots, m_n)$ on the eigenvector corresponding to a is $p(a)$.

For the proof of this Stickelberger's Theorem, we recall some facts from linear algebra related to the Cayley-Hamilton Theorem.

Definition 2.5.6. Let V be a vector space over \mathbb{K} and ϕ an endomorphism on V . For any polynomial $p = \sum_{i=0}^d c_i t^i \in \mathbb{K}[t]$, set $p(\phi) := \sum_{i=0}^d c_i \phi^i \in \text{End}(V)$, where ϕ^i is the i -fold composition of the endomorphism ϕ with itself. The ideal $I_\phi := \{p \in \mathbb{K}[t] \mid p(\phi) = 0\}$ is the kernel of the homomorphism $\mathbb{K}[t] \rightarrow \text{End}(V)$ defined by $t \mapsto \phi$. Its unique monic generator h_ϕ is the *minimal polynomial of ϕ* .

The eigenvalues and the minimal polynomial of an endomorphism are related.

Lemma 2.5.7. Let V be a finite-dimensional vector space over an algebraically closed field \mathbb{K} and ϕ be an endomorphism of V . Then an element $\lambda \in \mathbb{K}$ is an eigenvalue of ϕ if and only if λ is a zero of the minimal polynomial h_ϕ .

Proof. The eigenvalues of ϕ are the roots of its characteristic polynomial χ_ϕ . By the Cayley-Hamilton Theorem, the characteristic polynomial vanishes on ϕ , $\chi_\phi(\phi) = 0$. Thus $\chi_\phi \in I_\phi$ and h_ϕ divides χ_ϕ .

Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of ϕ , which are the roots of χ_ϕ . Suppose there is some eigenvalue, say λ_1 , for which $h_\phi(\lambda_1) \neq 0$. That is, the roots of h_ϕ are a proper subset of the eigenvalues, and we may write

$$h_\phi(t) = (t - \lambda_2)^{d_2}(t - \lambda_3)^{d_3} \cdots (t - \lambda_m)^{d_m}.$$

Let $v \in V$ be an eigenvector of ϕ with eigenvalue λ_1 . For any other eigenvalue $\lambda_i \neq \lambda_1$, we have $(\phi - \lambda_i I).v = (\lambda_1 - \lambda_i).v \neq 0$, and so

$$h_\phi(\phi).v = (\phi - \lambda_2)^{d_2} \cdots (\phi - \lambda_m)^{d_m}.v = (\lambda_1 - \lambda_2)^{d_2} \cdots (\lambda_1 - \lambda_m)^{d_m}v \neq 0,$$

which contradicts h_ϕ being the minimal polynomial of ϕ , do that $h_\phi(\phi) = 0$. \square

We can now prove Stickelberger's Theorem 2.5.4.

Proof of Theorem 2.5.4. Let λ be an eigenvalue of the multiplication endomorphism m_i on $\mathbb{K}[x_1, \dots, x_n]/I$ with corresponding eigenvector \bar{v} . That is, $\bar{x}_i \bar{v} = \lambda \bar{v}$ and thus $(\bar{x}_i - \lambda) \cdot \bar{v} = 0$ in the vector space $\mathbb{K}[x_1, \dots, x_n]/I$ so that $(x_i - \lambda)v \in I$. Let us assume by way of contradiction that there is no point $a \in \mathcal{V}(I)$ with i th coordinate λ .

This implies that $x_i - \lambda$ vanishes at no point of $\mathcal{V}(I)$. We will use this to show that $\bar{x}_i - \lambda$ is invertible in $\mathbb{K}[x_1, \dots, x_n]/I$. Multiplying the equation $(\bar{x}_i - \lambda) \cdot \bar{v} = 0$ by this inverse implies that $\bar{v} = 0$, which is a contradiction as eigenvectors are non-zero.

By Exercise 5 of Section 1.3, the map $\mathbb{K}[x_1, \dots, x_n] \rightarrow \mathbb{K}^{\mathcal{V}(I)}$ is surjective, where $\mathbb{K}^{\mathcal{V}(I)}$ is the ring of functions on the finite set $\mathcal{V}(I)$. Its kernel is \sqrt{I} by Hilbert's Nullstellensatz. Thus there exists a polynomial $f \in \mathbb{K}[x_1, \dots, x_n]$ with image

$$\bar{f} = \sum_{a \in \mathcal{V}(I)} \frac{1}{a_i - \lambda} \delta_a$$

in $\mathbb{K}^{\mathcal{V}(I)} \simeq \mathbb{K}[x_1, \dots, x_n]/\sqrt{I}$, where δ_a is the Kronecker delta function, whose value at a point b is zero unless $b = a$, and then its value is 1. Then $f(a) = 1/(a_i - \lambda)$ for $a \in \mathcal{V}(I)$, from which we obtain

$$(1 - (x_i - \lambda)f(x)) \in \mathcal{I}(\mathcal{V}(I)) = \sqrt{I}.$$

By Hilbert's Nullstellensatz, there is a positive integer N such that $(1 - (x_i - \lambda)f(x))^N \in I$. Expanding this, we obtain

$$1 - N(x_i - \lambda)f + \binom{N}{2}(x_i - \lambda)^2 f^2 - \dots \in I,$$

and so there exists a polynomial g such that $1 - (x_i - \lambda)g \in I$. Then \bar{g} is the desired inverse to $x_i - \lambda$ in $\mathbb{K}[x_1, \dots, x_n]/I$.

Conversely, let $a \in \mathcal{V}(I)$ with $a_i = \lambda$. Let h_i be the minimal polynomial of m_i . By Lemma 2.5.7 we need only show that $h_i(\lambda) = 0$. By the definition of minimal polynomial, the function $h_i(m_i)$ is the zero endomorphism on $\mathbb{K}[x_1, \dots, x_n]/I$. In particular, $h_i(\bar{x}_i) = h_i(m_i)(\bar{1}) = 0$ in $\mathbb{K}[x_1, \dots, x_n]/I$, which implies that the polynomial $h_i(x_i) \in \mathbb{K}[x_1, \dots, x_n]$ lies in I . Evaluating this at a point $a \in \mathcal{V}(I)$ gives $0 = h(a) = h(a_i) = h(\lambda)$. \square

Example 2.5.8. Let $I = \langle x^2y + 1, y^2 - 1 \rangle$. Then $\{x^4 - 1, y + x^2\}$ is a lexicographic Gröbner basis of I . Hence $\{1, x, x^2, x^3\}$ is a basis of $\mathbb{K}[x, y]/I$. With respect to this basis, the representing matrices of the endomorphisms m_x and m_y are

$$M_x = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad M_y = \begin{pmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix}.$$

The eigenvalues of M_x are $-1, 1, -i, i$ and the eigenvalues of M_y are -1 (twice) and 1 (twice). Indeed, we have $\mathcal{V}(I) = \{(i, 1), (-i, 1), (1, -1), (-1, -1)\}$. \diamond

Computationally, Theorem 2.5.4 requires that we know a basis of the coordinate ring $\mathbb{K}[x_1, \dots, x_n]/I$ and the companion matrices in this basis. Given these data, the computational complexity depends on the dimension, d , of $\mathbb{K}[x_1, \dots, x_n]/I$.

These methods simplify when there exists a joint basis of eigenvectors. That is, if there exists a matrix $S \in \mathbb{K}^{d \times d}$ and diagonal matrices $D_i \in \mathbb{K}^{d \times d}$ for $i = 1, \dots, n$ with

$$M_i S = S D_i, \quad \text{for } i = 1, \dots, n \quad (1 \leq i \leq n).$$

That is if the companion matrices M_i are *simultaneously diagonalizable*.

Theorem 2.5.9. *The companion matrices M_1, \dots, M_n are simultaneously diagonalizable if and only if I is radical.*

Proof. Let $a = (a_1, \dots, a_n)$ be a point in $\mathcal{V}(I)$. As in the proof of Theorem 2.5.4, there exists a polynomial g with $g(a) = 1$ and $g(b) = 0$ for all $b \in \mathcal{V}(I) \setminus \{a\}$. Hence, the polynomial $(x_i - a_i)g$ vanishes on $\mathcal{V}(I)$. Hilbert's Nullstellensatz then implies $(x_i - a_i)[g] \in \sqrt{I} = I$, and thus $[g]$ is a joint eigenvector of M_1, \dots, M_n .

Conversely, if the companion matrices M_1, \dots, M_n are simultaneously diagonalizable, then for every polynomial $f \in \mathbb{K}[x_1, \dots, x_n]$, the matrix $f(M_1, \dots, M_n)$ is simultaneously diagonalizable, as $f(M_1, \dots, M_n)S = Sf(D_1, \dots, D_n)$. Thus $f(M_1, \dots, M_n)$ is nilpotent only if it is the zero matrix. By Lemma 2.5.2, this implies that I is radical. \square

Stickelberger's Theorem 2.5.4 not only connects classical linear algebra to the problem of finding the common zeroes of a zero-dimensional ideal, but it leads to another method to compute eliminants.

Corollary 2.5.10. *Suppose that $I \subset \mathbb{K}[x_1, \dots, x_n]$ is a zero-dimensional ideal. The eliminant $g(x_i)$ is the minimal polynomial of the operator m_i of multiplication by x_i on $\mathbb{K}[x_1, \dots, x_n]/I$. It is a factor of the characteristic polynomial χ_{m_i} of m_i which contains all the roots of χ_{m_i} .*

This leads to an algorithm to compute the eliminant $g(x_i)$ of the radical of I .

Algorithm 2.5.11.

INPUT: A zero-dimensional ideal $I \subset \mathbb{K}[x_1, \dots, x_n]$ and an index i .

OUTPUT: The eliminant $g(x_i)$ of the radical of I .

Compute a Gröbner basis G for I with respect to any monomial order \prec . If $\dim I \neq 0$, then exit, else let \mathcal{B} be the corresponding finite set of standard monomials.

Construct M_i , the matrix in $\text{Mat}_{\mathcal{B} \times \mathcal{B}}(\mathbb{K})$ representing multiplication by x_i in the quotient ring $\mathbb{K}[x_1, \dots, x_n]/I$ in the basis of standard monomials. Let χ_{m_i} be the characteristic polynomial of M_i , and set $g(x_i)$ to be the square-free part of χ_{m_i} , $\chi_{m_i}/\gcd(\chi_{m_i}, \chi'_{m_i})$.

The proof of correctness of this algorithm is Exercise 7.

Exercises

- Let $p = c_0 + \dots + c_{d-1}x^{d-1} + x^d$ be a monic, univariate polynomial and set $I := \langle p \rangle$. Show that the matrix M_x representing the endomorphism $m_x : R/I \rightarrow R/I$, $\bar{f} \mapsto xf$ with respect to a natural basis coincides with the companion matrix C_p .
- Let $G := \{x^4 - 3x^2 - 2x + 1, y + x^3 - 3x - 1\}$ 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 $x \prec y$, determine the set of standard monomials of $\mathbb{C}[x, y]/I$ and compute the multiplication matrices M_x and M_y .

3. Let $f \in \mathbb{K}[x_1, \dots, x_n]$. Show that $m_f: \mathbb{K}[x_1, \dots, x_n]/I \rightarrow \mathbb{K}[x_1, \dots, x_n]/I$, where $\bar{g} \mapsto \bar{f} \cdot \bar{g}$ is an endomorphism.
4. In a computer algebra system, use the method of Stickelberger's Theorem to determine the common complex zeroes of $x^2 + 3xy + y^2 - 1$ and $x^2 + 2xy + y + 3$.
5. If two endomorphisms f and g on a finite-dimensional vector space V are diagonalizable and $f \circ g = g \circ f$, then they are jointly diagonalizable. Conclude that for Stickelberger's Theorem for the ring $\mathbb{K}[x, y]$ with only two variables, there always exist a basis of joint eigenvectors.
6. Perform the following computational experiment.
Generate two bivariate polynomials $f, g \in \mathbb{K}[x, y]$.
 - (a) Compute their resultant $\text{Res}(f, g; x) \in \mathbb{K}[y]$.
 - (b) Compute their eliminant $\langle f, g \rangle \cap \mathbb{K}[y]$, using a lexicographic Gröbner basis.
 - (c) Compute the characteristic polynomial of the companion matrix M_y .
 Compare the timings for these three operations for a number of polynomial pairs of moderate order. Which is more efficient ?
7. Prove the correctness of Algorithm 2.5.11.

2.6 Notes

Resultants were developed in the nineteenth century by Sylvester, were part of the computational toolkit of algebra from that century, and have remained a fundamental symbolic tool in algebra and its applications. Even more classical is Bézout's Theorem, stated by Etienne Bézout in his 1779 treatise *Théorie Générale des Équations Algébriques* [12, 13].

Perhaps mention that Chinese mathematicians could eliminate up to 4 variables?

The subject of Gröbner bases began with Buchberger's 1965 Ph.D. thesis which contained his algorithm to compute Gröbner bases [20, 21]. The term “Gröbner basis” honors Buchberger's doctoral advisor Wolfgang Gröbner. Key ideas about Gröbner bases had appeared earlier in work of Gordan and of Macaulay, and in Hironaka's resolution of singularities [60]. Hironaka called Gröbner bases “standard bases”, a term which persists. For example, in the computer algebra package **Singular** [44] the command `std(I)`; computes the Gröbner basis of an ideal I . Despite these precedents, the theory of Gröbner bases rightly begins with Buchberger's contributions.

Theorem 2.3.3 was proven by Macaulay [84], who the Gröbner basis package **Macaulay 2** [43] is named after.

There are additional improvements in Buchberger's algorithm (see Ch. 2.9 in [25] for a discussion), and even a series of completely different algorithms due to Jean-Charles Faugère [37] based on linear algebra with vastly improved performance.

The FGLM Gröbner basis conversion algorithm for zero-dimensional ideals is due to Faugère, Gianni, Lazard, and Mora [38].

For further information on techniques for solving systems of polynomial equations see the books of Cox, Little, and O’Shea [26, 25], Sturmfels [138] as well as Emiris and Dickenstein [31].

For numerical methods concerning the simultaneous diagonalization of matrices we refer the reader to Bunse-Gerstner, Byers, and Mehrmann [22]. In Section 5.3, a further refinement of the eigenvalue techniques will be used to study real roots.

Where is a reference to Stickelberger’s Theorem? David Cox is chasing this down.

Chapter 3

Structure of varieties

Outline:

1. Zariski topology.
2. Irreducible decomposition and dimension.
3. Rational functions.
4. Smooth and singular points.
5. Hilbert functions and dimension
6. Bertini and Bézout Theorems

In Chapter 1 we introduced varieties and ideals and established the algebra-geometry dictionary, and developed basic symbolic algorithms in Chapter 2. We now turn to structural properties of varieties which we will need in subsequent chapters. This begins with the Zariski topology and the notion of genericity, and then the analog of unique factorization for varieties. After discussing smooth and singular points and tangent spaces, we introduce the notion of dimension. This sets the stage for the fundamental theorems of Bertini-type which deal with the dimension and smoothness of intersections of varieties and their images under maps. This chapter finishes with the Hilbert function and degree of a projective variety and Bézout’s Theorem.

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. This notion is much stronger than elsewhere in geometry, where almost always may mean the complement of a set of measure zero or the complement of a nowhere dense set. We develop the terminology ‘generic’ and ‘Zariski open’ to formalize this notion of almost always in algebraic geometry.

A starting point is the behaviour of intersections and unions of affine varieties.

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. The first statement generalizes Lemma 1.2.11(1). Let $\{I_t \mid t \in T\}$ be a collection of ideals in $\mathbb{K}[x_1, \dots, x_n]$. Then we have

$$\bigcap_{t \in T} \mathcal{V}(I_t) = \mathcal{V}\left(\bigcup_{t \in T} I_t\right),$$

as both containments are straightforward. 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 (2). \square

Theorem 3.1.1 shows that affine varieties have the same properties as the closed sets of a topology on \mathbb{K}^n . (See Section A.2 of the Appendix.)

Definition 3.1.2. An affine variety is a *Zariski closed set*. The complement of a Zariski closed set is a *Zariski open set*. The *Zariski topology* on \mathbb{K}^n is the topology whose closed sets are the affine varieties in \mathbb{K}^n . The *Zariski closure* of a subset $Z \subset \mathbb{K}^n$ is the smallest variety containing Z , which is $\overline{Z} := \mathcal{V}(\mathcal{I}(Z))$, by Lemma 1.2.4. A subvariety X of \mathbb{K}^n inherits its Zariski topology from \mathbb{K}^n , the closed subsets of X are its subvarieties. A subset $Z \subset X$ of a variety X is *Zariski dense* in X if its Zariski closure is X .

We emphasize that the purpose of this terminology is to aid our discussion of varieties, and not because we will use these notions from topology in an essential way. In a more advanced treatment of algebraic geometry, including sheaves, these topological notions are essential. This Zariski topology behaves quite differently from the usual, or *Euclidean* topology on \mathbb{R}^n or \mathbb{C}^n with which we are familiar. A topology on a space may be defined by giving a collection of basic open sets which generate the topology. In the Euclidean topology, the basic open sets are (Euclidean) balls. Let $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. The *ball* with radius $\epsilon > 0$ centered at $z \in \mathbb{K}^n$ is

$$B(z, \epsilon) := \{a \in \mathbb{K}^n \mid \sum |a_i - z_i|^2 < \epsilon\}.$$

In the Zariski topology, the basic open sets are complements of hypersurfaces, called *principal open sets*. Let $f \in \mathbb{K}[x_1, \dots, x_n]$ and set

$$U_f := \{a \in \mathbb{K}^n \mid f(a) \neq 0\}. \tag{3.1}$$

In both the Zariski topology and the Euclidean topology the open sets are unions of basic open sets. We give two examples to illustrate the Zariski topology.

Example 3.1.3. The Zariski closed subsets of \mathbb{K}^1 consist of the empty set, finite collections of points, and \mathbb{K}^1 itself. Thus when \mathbb{K} is infinite the familiar separation property of Hausdorff spaces (any two points are covered by two disjoint open sets) fails spectacularly as any two nonempty open sets meet. \diamond

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 Zariski topology on \mathbb{K}^2 , the closed sets are finite unions of sets of the following form: the empty set, points, vertical and horizontal lines $\{a\} \times \mathbb{K}^1$ and $\mathbb{K}^1 \times \{a\}$ for $a \in \mathbb{K}$, and the whole space \mathbb{K}^2 . On the other hand, \mathbb{K}^2 contains a rich collection of other subvarieties (called *plane curves*), such as the cubic plane curves of Section 1.1. \diamond

Example 3.1.4 illustrates a general fact: the product Zariski topology on $X \times Y$ is weaker than the Zariski topology on $X \times Y$, when both X and Y are infinite. If $Z \subset X$ and $W \subset Y$ are Zariski closed subsets, then $Z \times W$ is Zariski closed in the product Zariski topology on $X \times Y$, and the same is true for products of Zariski open sets. However, the diagonal $\{(x, x) \mid x \in X\}$ is not closed in the product Zariski topology on $X \times X$, even though it is a subvariety.

We compare the Zariski topology with the Euclidean topology. Recall that a set is nowhere dense in the Euclidean topology if its closure does not contain a ball.

Theorem 3.1.5. Suppose that \mathbb{K} is one of \mathbb{R} or \mathbb{C} . Then

1. A Zariski closed set is closed in the Euclidean topology on \mathbb{K}^n .
2. A Zariski open set is open in the Euclidean topology on \mathbb{K}^n .
3. A nonempty Euclidean open set is dense in the Zariski topology on \mathbb{K}^n .
4. \mathbb{R}^n is dense in the Zariski topology on \mathbb{C}^n .
5. A proper Zariski closed set is nowhere dense in the Euclidean topology on \mathbb{K}^n .
6. A nonempty Zariski open set is dense in the Euclidean topology on \mathbb{K}^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 $f: \mathbb{K}^n \rightarrow \mathbb{K}$ is continuous in the Euclidean topology, and $\mathcal{V}(f) = f^{-1}(0)$.

Any ball $B(z, \epsilon)$ is dense in the Zariski topology. If a polynomial f vanishes identically on $B(z, \epsilon)$, then all of its partial derivatives do as well. Thus its Taylor series expansion at z is identically zero. But then f is the zero polynomial. This shows that $\mathcal{I}(B(z, \epsilon)) = \{0\}$, and so $\mathcal{V}(\mathcal{I}(B(z, \epsilon))) = \mathbb{K}^n$, that is, $B(z, \epsilon)$ is dense in the Zariski topology on \mathbb{K}^n .

Statement 4 uses 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 $\mathcal{V}(\mathcal{I}(\mathbb{R}^n)) = \mathbb{C}^n$. In fact, we may replace \mathbb{R}^n by any set containing a Euclidean ball.

For statements 5 and 6, observe that if f is nonconstant, then by 4, the Euclidean closed set $\mathcal{V}(f)$ does not contain a Euclidean ball so $\mathcal{V}(f)$ is nowhere dense. A variety is an intersection of nowhere dense hypersurfaces, so varieties (Zariski closed sets) are nowhere dense. The complement of a nowhere dense set is dense, so nonempty Zariski open sets are dense in \mathbb{K}^n . \square

Theorem 3.1.5(6) leads to the useful notions of genericity and generic sets and properties. We will use the shorthand term “Zariski dense” for dense in the Zariski topology.

Definition 3.1.6. Let X be a variety. A subset $Y \subset X$ is *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.

Our notion of which points are general depends on the context, and so care must be exercised in the use of these terms. For example, we may identify \mathbb{K}^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 due to its reliance on context, general is a very useful concept.

When \mathbb{K} 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. A general $n \times n$ matrix is invertible, as the set of invertible matrices forms a nonempty principal open subset of $\text{Mat}_{n \times n}(\mathbb{K})$. It is the complement of the variety $\mathcal{V}(\det)$ of singular matrices. The *general linear group* GL_n is the set of all invertible matrices,

$$GL_n := \{M \in \text{Mat}_{n \times n} \mid \det(M) \neq 0\} = U_{\det}. \quad \diamond$$

Example 3.1.8. A general univariate polynomial of degree n has n distinct complex roots. Identify \mathbb{K}^n with the set of univariate polynomials of degree n via

$$(a_1, \dots, a_n) \in \mathbb{K}^n \longmapsto x^n + a_1x^{n-1} + \dots + a_{n-1}x + a_n \in \mathbb{K}[x]. \quad (3.2)$$

The classical discriminant $\text{disc} \in \mathbb{K}[a_1, \dots, a_n]$ (see Example 2.1.5) is a polynomial of degree $2n-2$ which vanishes precisely when the polynomial (3.2) has a repeated factor. This identifies the set of polynomials with n distinct complex roots as the set U_{disc} . The discriminant of a quadratic $x^2 + bx + c$ is $b^2 - 4c$. \diamond

Example 3.1.9. A general complex $n \times n$ matrix is semisimple (diagonalizable). We do not show this by providing an algebraic characterization of semisimplicity. Instead we observe that if a matrix $M \in \text{Mat}_{n \times n}$ has n distinct eigenvalues, then it is semisimple. Let $M \in \text{Mat}_{n \times n}$ and consider the (monic) characteristic polynomial of M

$$\chi(x) := \det(xI_n - M),$$

whose roots are the eigenvalues of M . 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.

It follows that the set of matrices with distinct eigenvalues equals the principal open set U_ψ , which is nonempty. Thus the set of semisimple matrices contains an open dense subset of $\text{Mat}_{n \times n}$ and is therefore generic.

When $n = 2$, the characteristic polynomial of a generic matrix is

$$\det \left(xI_2 - \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \right) = x^2 - x(a_{11} + a_{22}) + a_{11}a_{22} - a_{12}a_{21},$$

and so the polynomial ψ is $(a_{11} + a_{22})^2 - 4(a_{11}a_{22} - a_{12}a_{21})$. \diamond

In each of these examples, we used the following easy fact.

Proposition 3.1.10. *A set $X \subset \mathbb{K}^n$ is generic if and only if there is a nonconstant polynomial that vanishes on its complement, if and only if it contains a principal open set U_f .*

More generally, if $X \subset \mathbb{K}^n$ is a variety and $f \in \mathbb{K}[x_1, \dots, x_n]$ is a polynomial which is not identically zero on X ($f \notin \mathcal{I}(X)$), then we have the *principal open subset* of X ,

$$X_f := X - \mathcal{V}(F) = \{x \in X \mid f(x) \neq 0\}. \quad (3.3)$$

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{K}^n contains the ideal $\mathcal{I}(X)$ of X . By the Hilbert Basis Theorem, there are polynomials $f_1, \dots, f_m \in \mathcal{I}(Y)$ such that

$$\mathcal{I}(Y) = \langle \mathcal{I}(X), f_1, \dots, f_m \rangle.$$

Then $X_{f_1} \cup \dots \cup X_{f_m}$ is equal to

$$(X - \mathcal{V}(f_1)) \cup \dots \cup (X - \mathcal{V}(f_m)) = X - (\mathcal{V}(f_1) \cap \dots \cap \mathcal{V}(f_m)) = X - Y = U. \quad \square$$

In Section 1.4, we introduced the affine cover of projective space $\mathbb{P}^n = U_0 \cup U_1 \cup \dots \cup U_n$, where for each $i = 0, \dots, n$,

$$U_i = U_{x_i} := \{a = [a_0, a_1, \dots, a_n] \in \mathbb{P}^n \mid a_i \neq 0\} \simeq \mathbb{K}^n.$$

Using homogenization and dehomogenization, we then showed that every projective variety X is covered by affine varieties, $X = X_0 \cup X_1 \cup \dots \cup X_n$, where $X_i := X \cap U_{x_i}$. We may define the Zariski topology on projective varieties in two equivalent ways: Either

extend the definition of Zariski topology to projective space (projective varieties are the closed sets) or use such affine covers of projective varieties to define basic open subsets to generate the topology. A consequence of the Zariski topology on a projective variety X (or on \mathbb{P}^n) is that a subset $Z \subset X$ is Zariski closed if and only if it is closed in each principal affine set $X_i = X \cap U_i$ of X .

We expand our notion of a variety slightly. A subset $X \subset \mathbb{P}^n$ is a *quasi-projective variety* if it is an open subset of its closure in \mathbb{P}^n . That is, if there are projective subvarieties Y, Z of \mathbb{P}^n with $X = Y \setminus Z$. It inherits its Zariski topology from that of projective space. A closed subset of a quasi-projective variety X is its intersection with a projective subvariety Y , and the same for an open subset of X . A subvariety of a quasi-projective variety X is a Zariski closed subset $Y \subset X$; this will also be a quasi-projective variety. The notion of generic introduced for affine varieties also makes sense for quasi-projective varieties, and it has the same properties. We will henceforth often drop the adjective quasi-projective and simply refer to these as varieties.

A path in a topological space X is a continuous map $\gamma: [0, 1] \rightarrow X$ from the unit interval in \mathbb{R} to X . Such a path connects $\gamma(0)$ to $\gamma(1)$. A space X is path-connected if any two points are connected by a path in X .

Theorem 3.1.12. *A nonempty Zariski open subset $U \subset \mathbb{C}^n$ is path connected in the Euclidean topology.*

Proof. Let $x, y \in U$. The parametrization $\ell: tx + (1 - t)y$ identifies \mathbb{C} with the affine line spanned by x and y . The inverse image $V := \ell^{-1}(U)$ of U is a nonempty Zariski open subset of \mathbb{C} (containing 0 and 1); its complement consists of finitely many points by Exercise 2. This is path connected, which implies there is a path in $\ell(V) \subset U$ connecting x and y . \square

After the exercises for this section, the Zariski topology is the default topology; “open” means Zariski open and “closed” means Zariski closed.

Exercises

1. Verify the claim that the collection of affine subvarieties of \mathbb{K}^n form the closed sets in a topology on \mathbb{K}^n . (See Section A.2 of the Appendix for definitions.)
2. Prove that a closed set in the Zariski topology on \mathbb{K}^1 is either the empty set, a finite collection of points, or \mathbb{K}^1 itself.
3. Prove that if $Z \subset X$ and $W \subset Y$ are subvarieties of the varieties X and Y , respectively, then $Z \times W$ is closed in the product Zariski topology on $X \times Y$, and that $Z \times W$ is a subvariety of $X \times Y$. Prove that if X is an affine variety, then the diagonal $\{(x, x) \mid x \in X\}$ is a subvariety of $X \times X$.
4. Suppose that $n \leq m$. Prove that a general $n \times m$ matrix has rank n .

5. Prove that the general triple of points in \mathbb{R}^2 are the vertices of a triangle.
6. (a) Verify the claim in Example 3.1.4 about the closed sets in the product Zariski topology on $\mathbb{K}^1 \times \mathbb{K}^1$.
(b) Show that any open set in the product Zariski topology on $\mathbb{K}^1 \times \mathbb{K}^1$ is Zariski open in \mathbb{K}^2 .
(c) Find a Zariski open set in \mathbb{K}^2 which is not open in the product topology on $\mathbb{K}^1 \times \mathbb{K}^1$.
7. (a) Show that the Zariski topology in \mathbb{K}^n is not Hausdorff if \mathbb{K} is infinite.
(b) Prove that any nonempty Zariski open subset of \mathbb{K}^n is dense.
(c) Prove that \mathbb{K}^n is compact in the Zariski topology.
8. Show that the principal open subset U_f (3.1) is an affine variety by identifying it with $\mathcal{V}(yf - 1) \subset \mathbb{K}^{n+1}$. Show that its coordinate ring is $\mathbb{K}[x]_{\frac{1}{f}}$, the localization of the polynomial ring at f . Deduce that a principal open subset X_f (3.3) of an affine variety is an affine variety.

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 this unique factorization. Any algebraic variety is the finite union of irreducible varieties, and this decomposition is unique.

A polynomial $f \in \mathbb{K}[x_1, \dots, x_n]$ is *reducible* if we may factor f nontrivially, that is, if $f = gh$ with neither g nor h a constant polynomial. Otherwise f is *irreducible*. Any polynomial $f \in \mathbb{K}[x_1, \dots, x_n]$ may be factored

$$f = cg_1^{\alpha_1}g_2^{\alpha_2} \cdots g_m^{\alpha_m} \quad (3.4)$$

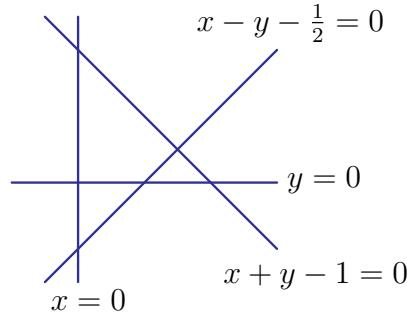
where c is a constant, the exponents α_i are positive integers, each polynomial g_i is irreducible and non-constant, and when $i \neq j$ the polynomials g_i and g_j are not proportional. The factorization (3.4) is essentially unique as any other such factorization is obtained from it by permuting the factors and possibly multiplying each polynomial g_i by a constant. The polynomials g_j are the *irreducible factors* of f .

When \mathbb{K} is algebraically closed, this algebraic property has a consequence for the geometry of hypersurfaces in \mathbb{K}^n . Suppose that a polynomial f has a factorization (3.4) 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, $\mathcal{V}(xy^2(x+y-1)^3(x-y-\frac{1}{2}))$ is the union of four lines in \mathbb{K}^2 .



We will show that 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 have $y+z = 0$ 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 parameterized by $(t^2, t, -t^3)$.

Figure 3.1 shows another reducible variety. It has six components, one is a surface,

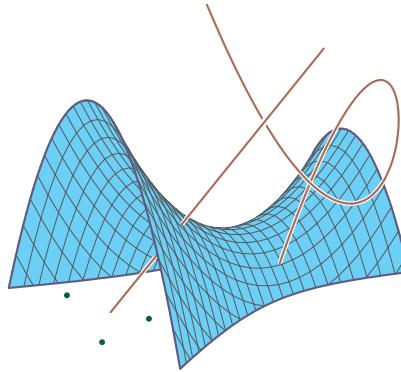


Figure 3.1: A reducible variety

two are space curves, and three are points. \diamond

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,

$$\{x\} \times Y = ((\{x\} \times Y) \cap Z_1) \cup ((\{x\} \times Y) \cap Z_2),$$

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 $X \times Y = Z_1$.

It suffices to 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 and completes the proof. \square

The geometric notion of an irreducible variety corresponds to the algebraic notion of a prime ideal. An ideal $I \subset \mathbb{K}[x_1, \dots, x_n]$ 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. *An affine 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 $Z := X \cap \mathcal{V}(g)$ 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 $\mathcal{I}(X)$ is not prime. \square

As a principal ideal $\langle f \rangle$ for $f \in \mathbb{K}[x_1, \dots, x_n]$ is prime if and only if f is irreducible, Theorem 3.2.4 implies that the unique decomposition of hypersurfaces into unions of hypersurfaces defined by irreducible polynomials is a decomposition of a hypersurface into irreducible hypersurfaces.

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), yields 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 closed 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 is reducible, and continue this process, stopping only when all subvarieties obtained are irreducible. *A priori*, this process could continue indefinitely. We show 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,

$$X \supsetneq X_1 \supsetneq X_2 \supsetneq \dots .$$

Their ideals form an infinite increasing chain of ideals in $\mathbb{K}[x_1, \dots, x_n]$,

$$\mathcal{I}(X) \subsetneq \mathcal{I}(X_1) \subsetneq \mathcal{I}(X_2) \subsetneq \dots .$$

The union I of these ideals is again an ideal. No ideal $\mathcal{I}(X_m)$ is equal to I as the chain of ideals is strict. 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. \square

Lemma 3.2.6. *Let X be a variety with $U \subset X$ a quasiprojective variety that is dense in X . Then X is irreducible if and only if U is irreducible.*

Proof. Assume that U is irreducible and suppose that $X = Y \cup Z$ is the union of two closed subvarieties. Then $U = (U \cap Y) \cup (U \cap Z)$ is the union of two closed subvarieties. As U is irreducible, we may assume that $U = U \cap Y$, but then $X = \overline{U} \subset Y$.

Now assume that X is irreducible and suppose that $U = V \cup W$ is union of two closed subvarieties of U . Then $X = \overline{U} = \overline{V} \cup \overline{W}$ is the union of two closed subvarieties. As X is irreducible, we may assume that $X = \overline{V}$, but then $U = U \cap \overline{V} = V$. \square

Corollary 3.2.7. *A variety X is a finite union of irreducible subvarieties.*

Proof. Suppose that $X \subset \mathbb{P}^n$ is a projective variety. Then $X = X_0 \cup X_1 \cup \dots \cup X_n$ where $X_i = X \cap U_i$. Each affine variety X_i is a finite union of irreducible closed subvarieties $U_{i,1}, \dots, U_{i,m_i}$. By Lemma 3.2.6, the closure in \mathbb{P}^n of each $U_{i,j}$ is irreducible. Noting that X equals the union of the closures of the $U_{i,j}$ shows that X is a finite union of irreducible closed subvarieties.

If $X \subset \mathbb{P}^n$ is a quasi-projective variety, then its Zariski closure is a projective variety. Thus \overline{X} may be written as a finite union of irreducible closed subvarieties. The intersection of each of these with X is an irreducible closed subvariety of X , by Lemma 3.2.6. Noting that X is the union of these intersections completes the proof. \square

A consequence of the proof of Lemma 3.2.5 and of Corollary 3.2.7 is that any decreasing chain of subvarieties of a given variety must have finite length. When \mathbb{K} is infinite, there are such decreasing chains of arbitrary length. There is however a bound for the length of the longest decreasing chain of irreducible subvarieties.

Definition 3.2.8 (Combinatorial Definition of Dimension). The *dimension* of a variety X is the length of the longest decreasing chain of irreducible subvarieties of X . If

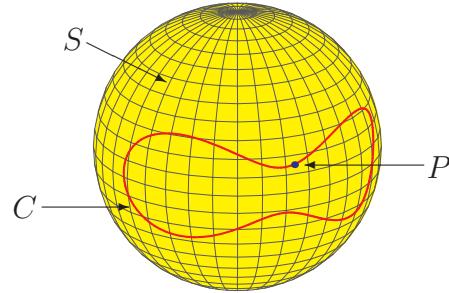
$$X \supset X_0 \supseteq X_1 \supseteq X_2 \supseteq \cdots \supseteq X_m \supseteq \emptyset, \quad (3.5)$$

with each X_i irreducible is such a chain of maximal length, then X has dimension m and we write $\dim X = m$. If (3.5) has length $m = \dim X$, the X_i has dimension $m-i$.

The discussion on elimination of bivariate polynomials in Section 2.1, takes care of making this well-founded, at least on \mathbb{K}^2 .

Since maximal ideals of $\mathbb{C}[x_1, \dots, x_n]$ have the form $\mathfrak{m}_a = \langle x_1 - a_1, \dots, x_n - a_n \rangle$ for some $a \in \mathbb{C}^n$, we see that X_m is a point when $\mathbb{K} = \mathbb{C}$. 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 3.5 we shall prove that this definition is correct by relating it to other notions of dimension.

Example 3.2.9. The sphere S has dimension at least two, as we have the chain of irreducible subvarieties $S \supsetneq C \supsetneq P$ as shown below.



It is challenging to show that any maximal chain of irreducible subvarieties of the sphere has length 2 with what we now know. \diamond

By Corollary 3.2.7, a variety X may be written as a finite union

$$X = X_1 \cup X_2 \cup \cdots \cup X_m$$

of irreducible closed 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 3.2.10 (Unique Decomposition of Varieties). *A variety X has a unique irredundant decomposition as a finite union of irreducible closed 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 closed subvarieties,

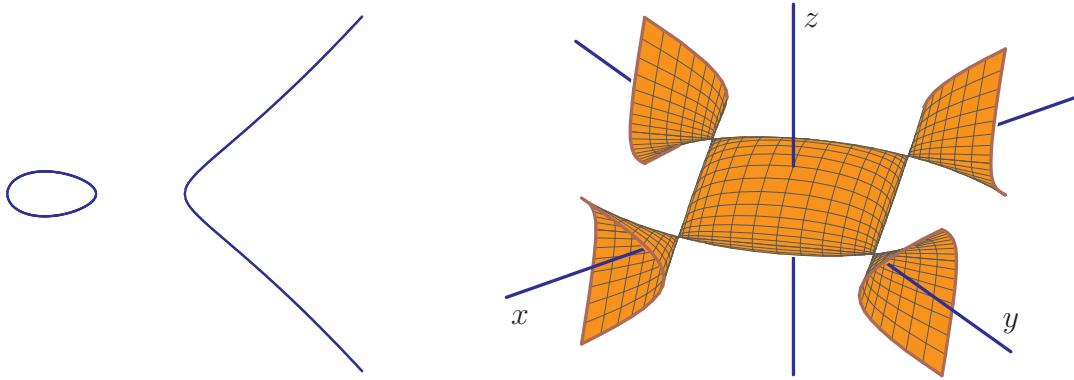
$$X = Y_1 \cup Y_2 \cup \cdots \cup Y_n,$$

where Y_j is irreducible and closed. Then for each $i = 1, \dots, m$,

$$X_i = (X_i \cap Y_1) \cup (X_i \cap Y_2) \cup \cdots \cup (X_i \cap Y_n).$$

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 $Y_j \subset X_k$, so that $X_i \subset X_k$. But then $i = k$, 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. \square

When $\mathbb{K} = \mathbb{C}$, 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{R}^2 and the surface $\mathcal{V}((x^2 - y^2)^2 - 2x^2 - 2y^2 - 16z^2 + 1)$ in \mathbb{R}^3 .



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. (While intuitive, smooth and singular points will be defined in Section 3.4.)

Theorem 3.2.11. *Suppose that X is irreducible and $f: X \rightarrow Y$ is a map. Then $\overline{f(X)}$ is an irreducible subvariety of Y .*

[†]When and where will we show this?

Proof. Suppose that the closure Z of $f(X)$ is the union of two subvarieties, $Z = Z_1 \cup Z_2$, with $Z_2 \neq Z$. For each $i = 1, 2$, set $X_i := f^{-1}(Z_i)$, which is closed in X and thus a subvariety. Then $X = X_1 \cup X_2$. Since $Z_2 \neq Z$, we have $X_2 \neq X$. As X is irreducible, this implies that $X = X_1$, and therefore that $Z = Z_1$. \square

We close this section with a proof that the resultant polynomial of Section 2.1 is irreducible. This uses facts about irreducibility and dimension that we have not yet established. We should also be able to do the discriminant.

Remark 3.2.12. Suppose that \mathbb{K} is algebraically closed and consider the variety of all triples consisting of a pair of univariate polynomials with a common root, together with a common root,

$$\Sigma := \{(f, g, a) \in \mathbb{K}_m[x] \times \mathbb{K}_n[x] \times \mathbb{K} \mid f(a) = g(a) = 0\},$$

where $\mathbb{K}_r[x]$ is the $(r+1)$ -dimensional vector space of polynomials of degree r . This has projections $p: \Sigma \rightarrow \mathbb{K}_m[x] \times \mathbb{K}_n[x]$ and $\pi: \Sigma \rightarrow \mathbb{K}$. The image $p(\Sigma)$ is the set of pairs of polynomials having a common root, which is the variety $\mathcal{V}(\text{Res})$ of the resultant polynomial, $\text{Res} \in \mathbb{Z}[f_0, \dots, f_m, g_0, \dots, g_n]$, where f_0, \dots, g_n are the coefficients of f and g ,

$$f = f_0x^m + f_1x^{m-1} + \dots + f_m \quad \text{and} \quad g = g_0x^n + g_1x^{n-1} + \dots + g_n.$$

The fiber of π over a point $a \in \mathbb{K}$ 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 $\mathbb{K}^m \times \mathbb{K}^n$. Since $\pi: \Sigma \rightarrow \mathbb{K}$ has irreducible image (\mathbb{K}) 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 $\mathbb{K}_m[x] \times \mathbb{K}_n[x]$. Let F be a polynomial generating the ideal $\mathcal{I}(p(\Sigma))$, which is necessarily irreducible. As $\mathcal{V}(\text{Res}) = p(\Sigma)$, we must have $\text{Res} = F^N$ for some positive integer N . The formula (2.3) shows that $N = 1$ as the resultant polynomial is square-free.

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 $f_0^n g_n^m$ with coefficient 1, as we showed in the proof of Lemma 2.1.3. \diamond

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 , each with exponent 1.
2. Suppose that \mathbb{K} is infinite. For every positive integer n , give a decreasing chain of subvarieties of \mathbb{K}^1 of length $n+1$.

3. Suppose that $I_1 \subset I_2 \subset \dots$ is an increasing chain of ideals in $\mathbb{K}[x_1, \dots, x_n]$. Show that its union is an ideal of $\mathbb{K}[x_1, \dots, x_n]$.
4. Prove that the dimension of a point is 0 and the dimension of \mathbb{K}^1 is 1.
5. Show that an irreducible affine variety is zero-dimensional if and only if it is a point.
6. Prove that the dimension of an irreducible plane curve is 1 and use this to show that the dimension of \mathbb{K}^2 is 2.
7. Write the ideal $\langle x^3 - x, x^2 - y \rangle$ as the intersection of two prime ideals. Describe the corresponding geometry.
8. 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.
9. Suppose that f and g are two polynomials on \mathbb{C}^n that are relatively prime. Show that every component of $\mathcal{V}(f, g)$ has dimension $n - 2$.
10. Let $f(x, y)$ be a polynomial of total degree n . Show that there is a non-empty Zariski open subset of parameters $(a, b, c, \alpha, \beta, \gamma) \in \mathbb{K}^6$ with $a\beta - \alpha b \neq 0$ such that if A is the affine transformation (2.9), 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.
11. Use Lemma 2.1.13 to show that \mathbb{K}^2 has dimension 2, in the sense of the combinatorial definition of dimension (3.2.8).
12. Use Lemma 2.1.13 and induction on the number of polynomials defining a proper subvariety X of \mathbb{K}^2 to show that X consists of finitely many irreducible curves and finitely many isolated points.

3.3 Rational functions and maps

This should be folded into the previous section.

In algebraic geometry, we also use functions and maps between varieties which are not defined at all points of their domains. Working 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 $\mathbb{K}[X]$ has no zero divisors ($0 \neq f, g \in \mathbb{K}[X]$ 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* $\mathbb{K}(X)$

of X as the quotients of regular functions in $\mathbb{K}[X]$. Formally, $\mathbb{K}(X)$ is the collection of all quotients f/g with $f, g \in \mathbb{K}[X]$ and $g \neq 0$, where we identify

$$\frac{f_1}{g_1} = \frac{f_2}{g_2} \iff f_1g_2 - f_2g_1 = 0 \text{ in } \mathbb{K}[X].$$

The map $f \mapsto \frac{f}{1}$ embeds $K[X]$ into the function field $\mathbb{K}(X)$.

Example 3.3.1. The function field of affine space \mathbb{K}^n is the collection of quotients of polynomials P/Q with $P, Q \in \mathbb{K}[x_1, \dots, x_n]$. This field $\mathbb{K}(x_1, \dots, x_n)$ is called the *field of rational functions* in the variables x_1, \dots, x_n . \diamond

Given an irreducible affine variety $X \subset \mathbb{K}^n$, we may also express $\mathbb{K}(X)$ as the collection of quotients f/g of polynomials $f, g \in \mathbb{K}[x_1, \dots, x_n]$ with $g \notin \mathcal{I}(X)$, where we identify

$$\frac{f_1}{g_1} = \frac{f_2}{g_2} \iff f_1g_2 - f_2g_1 \in \mathcal{I}(X).$$

Rational functions on an affine variety X do not in general have unique representatives as quotients of polynomials or even as quotients of regular functions.

Example 3.3.2. Let $X := \mathcal{V}(x^2 + y^2 + 2y) \subset \mathbb{K}^2$ be the circle of radius 1 and center at $(0, -1)$. In $\mathbb{K}(X)$ we have

$$-\frac{x}{y} = \frac{y+2}{x}. \quad \diamond$$

In Chapter 1, we showed that an affine variety is determined up to embedding in affine space by its coordinate ring, and that there is an equivalence of categories between affine varieties and finitely generated reduced \mathbb{K} -algebras. There is not as tight of a correspondence between irreducible varieties and their fields of rational functions. This however enables us to define fields of rational functions for arbitrary irreducible varieties.

Proposition-Definition 3.3.3. *Let X be an irreducible variety and $U, V \subset X$ non-empty affine open subvarieties of X . Then their function fields are equal, $\mathbb{K}(U) = \mathbb{K}(V)$, and we define the function field $\mathbb{K}(X)$ to be this common field.*

Thus the function field of an irreducible variety X depends rather weakly on X as any affine open subset has the same function field.

Proof. Suppose first that $X \subset \mathbb{K}^n$ is affine. As U is open, there is some $f \in \mathbb{K}[X]$ with $X \setminus U \subset \mathcal{V}(f)$, so that $X_f = X \setminus \mathcal{V}(f) \subset U$. By Exercise 1, $\mathbb{K}[X_f] = \mathbb{K}[X][\frac{1}{f}]$, and $\mathbb{K}[X] \subset \mathbb{K}[X_f]$. But then $\mathbb{K}(X) = \mathbb{K}(X_f)$. As the same holds for U in place of X , we have $\mathbb{K}(X) = \mathbb{K}(U)$.

This does not quite prove the general case, as $U \cap V$ need not be affine. Let $f \in \mathbb{K}[U]$ be such that $U_f \subset U \cap V$, which is an affine subset of both U and V . Then $\mathbb{K}(U_f) = \mathbb{K}(U)$, and the same for V completes the proof. \square

A point $x \in X$ is a *regular point* of a rational function $\varphi \in \mathbb{K}(X)$ if φ has a representative f/g with $f, g \in \mathbb{K}[X]$ and $g(x) \neq 0$. From this we see that all points of the principal affine set X_g , which is a neighborhood of x in X , are regular points of φ . Thus the set of regular points of φ is a nonempty open subset of X . This is the *domain of regularity of φ* .

When $x \in X$ is a regular point of a rational function $\varphi \in \mathbb{K}(X)$, we set $\varphi(x) := f(x)/g(x) \in \mathbb{K}$, 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{K} . We write this as

$$\varphi : X \dashrightarrow \mathbb{K}$$

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 $\varphi : X \dashrightarrow \mathbb{K}$ is stereographic projection of the circle onto the line $y = -1$ from the point $(0, 0)$.

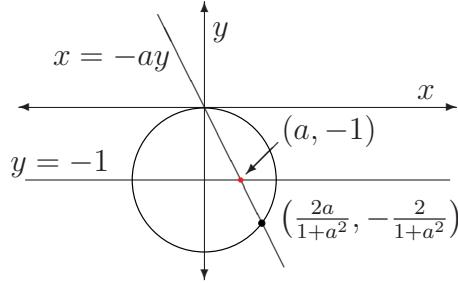


Figure 3.2: Projection of the circle $\mathcal{V}(x^2 + (y - 1)^2 - 1)$ from the origin.

Example 3.3.4. Let $X = \mathbb{R}$ 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 $\pm\sqrt{-1} \in \mathbb{C}$. \diamond

Theorem 3.3.5. *When \mathbb{K} is algebraically closed, a rational function that is regular at all points of an irreducible affine variety X is a regular function in $\mathbb{K}[X]$.*

Proof. For each point $x \in X$, there are regular functions $f_x, g_x \in \mathbb{K}[X]$ with $\varphi = f_x/g_x$ and $g_x(x) \neq 0$. Let \mathcal{I} be the ideal generated by the denominators g_x of φ for $x \in X$. Then $\mathcal{V}(\mathcal{I}) = \emptyset$, as φ is regular at all points of X .

If we let g_1, \dots, g_s be generators of \mathcal{I} that are denominators of φ and let f_1, \dots, f_s be regular functions such that $\varphi = f_i/g_i$ for each i . Then by the Weak Nullstellensatz for X (Theorem 1.3.5(3)), there are regular functions $h_1, \dots, h_s \in \mathbb{K}[X]$ such that in $\mathbb{K}[X]$,

$$1 = h_1 g_1 + \dots + h_s g_s.$$

Multiplying this equation by φ , we obtain

$$\varphi = h_1 f_1 + \cdots + h_s f_s,$$

which proves the theorem. \square

A list f_1, \dots, f_m of rational functions gives a *rational map*

$$\begin{aligned} \varphi : X &\dashrightarrow \mathbb{K}^m, \\ x &\longmapsto (f_1(x), \dots, f_m(x)). \end{aligned}$$

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 $\mathbb{K}[X] \subset \mathbb{K}(X)$, any regular map is also a rational map. As with regular maps, a rational map $\varphi : X \dashrightarrow \mathbb{K}^m$ given by functions $f_1, \dots, f_m \in \mathbb{K}(X)$ defines a homomorphism $\varphi^* : \mathbb{K}[y_1, \dots, y_m] \rightarrow \mathbb{K}(X)$ by $\varphi^*(g) = g(f_1, \dots, f_m)$. If Y is an affine subvariety of \mathbb{K}^m , then $\varphi(X) \subset Y$ if and only if $\varphi(\mathcal{I}(Y)) = 0$. In particular, the kernel J of the map $\varphi^* : \mathbb{K}[y_1, \dots, y_m] \rightarrow \mathbb{K}(X)$ defines the smallest subvariety $Y = \mathcal{V}(J)$ containing $\varphi(X)$, that is, the Zariski closure of $\varphi(X)$. Since $\mathbb{K}(X)$ 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 $\varphi^* : \mathbb{K}[Y] \hookrightarrow \mathbb{K}(X)$. Since Y is irreducible, this map extends to a map of function fields $\varphi^* : \mathbb{K}(Y) \rightarrow \mathbb{K}(X)$. Conversely, given a map $\psi : \mathbb{K}(Y) \rightarrow \mathbb{K}(X)$ of function fields, with $Y \subset \mathbb{K}^m$, we obtain a dominant rational map $\varphi : X \dashrightarrow Y$ given by the rational functions $\psi(x_1), \dots, \psi(x_m) \in \mathbb{K}(X)$ where x_1, \dots, x_m are the coordinate functions on $Y \subset \mathbb{K}^m$.

Suppose we have two rational maps $\varphi : X \dashrightarrow Y$ and $\psi : Y \dashrightarrow Z$ with φ dominant. Then $\varphi(X)$ meets the set of regular points of ψ , and so we may compose these maps $\psi \circ \varphi : X \dashrightarrow Z$. 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{K}^1 and the circle of Figure 3.2 are birationally equivalent. The inverse of stereographic projection from the circle to \mathbb{K}^1 is the map from \mathbb{K}^1 to the circle given by $a \mapsto (\frac{2a}{1+a^2}, -\frac{2}{1+a^2})$.

Most of the next few paragraphs, up to the example, needs strongly revising, as it now appears in Chapter 1. Let us now consider rational functions and maps of projective varieties. Let $X \subset \mathbb{P}^n$ be a projective variety. Recall that a homogeneous polynomial $f \in \mathbb{K}[x_0, \dots, x_n]$ does not define a function on either \mathbb{P}^n or on X , unless it is a constant, but its vanishing set $\mathcal{V}(f)$ is well defined. However, given two homogeneous polynomials f and g in $\mathbb{K}[x_0, \dots, x_n]$ 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 $[a_0, \dots, a_n]$ and $[\lambda a_0, \dots, \lambda a_n]$ are two representatives of the point $a \in \mathbb{P}^n$ and $g(a) \neq 0$, then

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

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 $X - \mathcal{V}(g)$.

Similarly, suppose that $f_0, f_1, \dots, f_m \in \mathbb{K}[X]$ are elements of the same degree d with at least one f_i non-zero on X . These define a *rational map*

$$\begin{aligned} \varphi : X &\dashrightarrow \mathbb{P}^m \\ x &\mapsto [f_0(x), f_1(x), \dots, f_m(x)]. \end{aligned}$$

Indeed, if $a = [a_0, \dots, a_n]$ and $\lambda a = [\lambda a_0, \dots, \lambda a_n]$ are two representatives of a point $x \in X$ where some f_i does not vanish, then

$$[f_0(\lambda a), \dots, f_m(\lambda a)] = [\lambda^d f_0(a), \dots, \lambda^d f_m(a)] = [f_0(a), \dots, f_m(a)].$$

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

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

The map φ is regular at a point $x \in X$ if there is some system of representatives f_0, \dots, f_m for the map φ for which $x \notin \mathcal{V}(f_0, \dots, f_m)$. 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 $\varphi: X \dashrightarrow Y$ is the set of points of X at which φ is not regular.

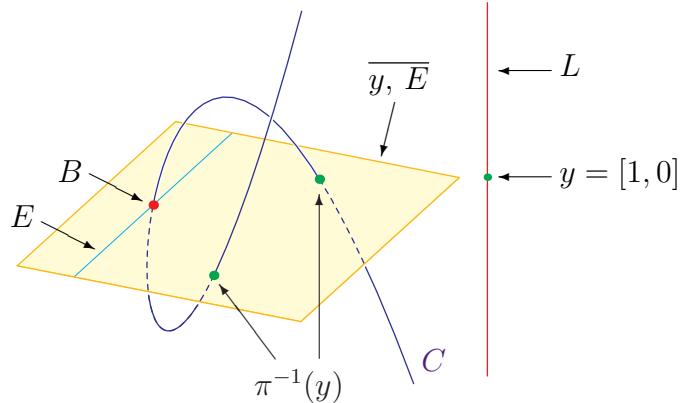
Example 3.3.6. A common example of a rational map is a linear projection. Let $\Lambda_0, \Lambda_1, \dots, \Lambda_m$ be linear forms. These give a rational map φ which is defined at points of $\mathbb{P}^n - E$, where E is the common zero locus of the linear forms $\Lambda_0, \dots, \Lambda_m$, that is $E = \mathbb{P}(\text{kernel}(M))$, where M is the matrix whose columns are the Λ_i . \diamond

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 1.4.3 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 1.4.3 was the map

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

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

Figure 3.3 shows another linear projection. Let C be the cubic space curve with parametrization $[1, t, t^2, 2t^3 - 2t]$ and $\pi: \mathbb{P}^3 \dashrightarrow L \simeq \mathbb{P}^1$ the linear projection defined by

Figure 3.3: A linear projection π with center E .

the last two coordinates, $\pi: [x_0, x_1, x_2, x_3] \mapsto [x_2, x_3]$. We have drawn the image \mathbb{P}^1 in the picture to illustrate that the inverse image of a point under 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 the curve in a point, B .

Exercises

1. Suppose that X is an irreducible affine variety and $0 \neq f \in \mathbb{K}[X]$. Following Exercise 8 of Section 3.1, show that the coordinate ring of the principal open subset X_f (3.3) is $\mathbb{K}[X_f] = \mathbb{K}[X][\frac{1}{f}]$ and that $K[X] \subset \mathbb{K}[X_f]$.
2. Show that irreducible affine varieties X and Y are birationally equivalent if and only if they have isomorphic open sets.
3. We observed that quotients f/g of homogeneous polynomials of the same degree define a function on the principal open set $U_g = \mathbb{P}^n \setminus \mathcal{V}(g)$. The quotient field of the homogeneous coordinate ring of \mathbb{P}^n is graded, and these quotients have degree 0. Show that the degree 0 component of this quotient field is isomorphic to the rational function field $\mathbb{K}(x_1, \dots, x_n)$.
4. 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 $X - \mathcal{V}(g)$.

3.4 Smooth and singular points

Algebraic varieties are not manifolds—the very first example of this book ((1.1) in Section 1.1) included the cubic plane curve $\mathcal{V}(y^2 - x^2 - x^3)$. In a neighborhood of the origin,

this curve is not a manifold; it has two branches crossing at the origin. Many other examples likewise have points that do not have a neighborhood in the Euclidean topology which are manifolds, either differentiable or topological. Algebraic varieties have points at which they are differentiable manifolds (smooth points) and others at which they are not manifolds (singular points). We develop some of the basic properties of these smooth and singular points.

Given a polynomial $f \in \mathbb{K}[x]$ and a point $a = (a_1, \dots, a_n) \in \mathbb{K}^n$, we may write f as a polynomial in new variables $v = (v_1, \dots, v_n)$, with $v_i := x_i - a_i$ to obtain

$$f = f(a) + \sum_{i=1}^n \frac{\partial f}{\partial x_i}(a) \cdot v_i + \dots, \quad (3.6)$$

where the remaining terms have degrees greater than 1 in the variables v . When \mathbb{K} has characteristic zero, this is the usual multivariate Taylor expansion of f at the point a (and the 'derivatives' in (3.6) are derivatives). The coefficient of the monomial v^α in this expansion is the mixed partial derivative of f evaluated at a ,

$$\frac{1}{\alpha_1! \alpha_2! \cdots \alpha_n!} \left(\left(\frac{\partial}{\partial x_1} \right)^{\alpha_1} \left(\frac{\partial}{\partial x_2} \right)^{\alpha_2} \cdots \left(\frac{\partial}{\partial x_n} \right)^{\alpha_n} f \right)(a).$$

In the coordinates v for \mathbb{K}^n , the degree one term in the expansion (3.6) is a linear map

$$d_a f : \mathbb{K}^n \longrightarrow \mathbb{K}$$

called the *differential* of f at the point a . Note that for any constant $c \in \mathbb{K}$, we have $d_a(c) = 0$ and $d_a(f + c) = d_a f$.

Definition 3.4.1. Let $X \subset \mathbb{K}^n$ be an affine variety with ideal $\mathcal{I}(X)$. The (*Zariski*) *tangent space* $T_a X$ to X at the point $a \in X$ is the subspace of \mathbb{K}^n annihilated by the collection $\{d_a f \mid f \in \mathcal{I}(X)\}$ of linear maps. Since

$$\begin{aligned} d_a(f+g) &= d_a f + d_a g \\ d_a(fg) &= f(a)d_a g + g(a)d_a f \end{aligned} \quad (3.7)$$

we do not need all the polynomials in $\mathcal{I}(X)$ to define $T_a X$, but may instead take any finite generating set.

Suppose that $X \subset \mathbb{K}^n$ is an affine variety and $a \in X$. Given a nonzero vector $v \in \mathbb{K}^n$, the map $\ell: t \mapsto a + tv$ parameterizes the line through a with direction v . For $f \in \mathcal{I}(X)$, if we expand the composition $f(\ell(t))$ in powers of t , we obtain

$$f(\ell(t)) = 0 + t(d_a f \cdot v) + t^2(\dots),$$

where we suppress the coefficients of t^2 and of higher powers in t . Here, $d_a f \cdot v$ is the usual dot product. When $d_a f \cdot v = 0$, the function $f(\ell(t))$ of t vanishes to order at least

2 at $t = 0$. Thus the nonzero vectors in $T_a X$ are the directions of lines through a whose algebraic order of contact with every hypersurface containing X is at least 2. If X is a manifold in \mathbb{K}^n (real or complex as $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$), then the Zariski tangent space $T_a X$ is equal to the extrinsic tangent space of the manifold X at a . (Extrinsic as it is a linear subspace of \mathbb{K}^n .)

Example 3.4.2. Consider the cuspidal cubic $C = \mathcal{V}(f) \subset \mathbb{K}^2$, where $f := y^2 - x^3$. This contains the origin $(0, 0)$, and $df_{(0,0)}$ is the zero linear functional, so that $T_{(0,0)} C = \mathbb{K}^2$, which has dimension two. At every other point $a \in C$, we have $d_a f \neq 0$, so that $T_a C$ is one-dimensional. Figure 3.4 shows the cubic, its tangent space at the origin and its

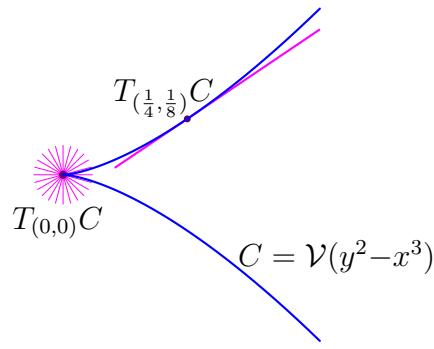


Figure 3.4: Zariski tangent spaces of the cuspidal cubic

tangent space at $(\frac{1}{4}, \frac{1}{8})$. As is customary, we translate the linear subspace $T_a C$ so that its origin is at the point a , to indicate its relation to the variety. \diamond

Theorem 3.4.3. *Let X be an affine variety and suppose that \mathbb{K} is algebraically closed. Then the set of points of X whose tangent space has minimal dimension is a nonempty Zariski open subset of X .*

Proof. Let f_1, \dots, f_m be generators of $\mathcal{I}(X)$. Writing $F = (f_1, \dots, f_m)$, we let $DF \in \text{Mat}_{m \times n}(\mathbb{K}[x])$ be the matrix whose entry in row i and column j is the partial derivative $\partial f_i / \partial x_j$. For $a \in \mathbb{K}^n$, the components of the vector-valued function

$$\begin{aligned} DF : \mathbb{K}^n &\longrightarrow \mathbb{K}^m \\ v &\longmapsto DF(a)v \end{aligned}$$

are the dot products $d_a f_q \cdot v, \dots, d_a f_m \cdot v$ and its kernel is $T_a X$ when $a \in X$.

For each $i = 1, 2, \dots, \min\{n, m\}$, the *degeneracy locus* $\Delta_i \subset \mathbb{K}^m$ is the variety defined by all $i \times i$ subdeterminants (*minors*) of the matrix DF , and set $\Delta_{\min\{n,m\}+1} := \mathbb{K}^n$. Since we may expand any $(i+1) \times (i+1)$ minor along a row or column and express it in terms of $i \times i$ minors, these varieties are nested

$$\Delta_1 \subset \Delta_2 \subset \cdots \subset \Delta_{\min\{n,m\}} \subset \Delta_{\min\{n,m\}+1} = \mathbb{K}^n.$$

By definition, a point $a \in \mathbb{K}^n$ lies in $\Delta_{i+1} - \Delta_i$ if and only if the matrix $DF(a)$ has rank exactly i . In particular, if $a \in \Delta_{i+1} - \Delta_i$, then the kernel of $DF(a)$ 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 . \square

The Zariski tangent space of an affine variety $X \subset \mathbb{K}^n$ is defined extrinsically via a given embedding in affine space. We used this to show that there is a nonempty open subset of X where this has minimal dimension. The Zariski tangent space also has an intrinsic definition. For any point $a \in X$ and polynomial $f \in \mathbb{K}[x]$, the differential $d_a f$ is a linear map on \mathbb{K}^n that we may restrict to the Zariski tangent space $T_a X$ of X at a . By the formulas (3.7) and the definition of $T_a X$, this linear map is well-defined for elements $f \in \mathbb{K}[X]$ of the coordinate ring of X . Recall that \mathfrak{m}_a is the maximal ideal of $\mathbb{K}[X]$ consisting of regular functions that vanish at a . Since $d_a(f) = d_a(f - f(a))$, the formulas (3.7) show that the differential is a linear map from \mathfrak{m}_a to $T_a^* X := \text{Hom}(T_a X, \mathbb{K})$, the space of linear functions on $T_a X$. By the Leibniz formula for d_a (3.7), elements of the square \mathfrak{m}_a^2 of \mathfrak{m}_a have zero differential.

Lemma 3.4.4. *For a point $a \in X$, there is a canonical isomorphism $d_a: \mathfrak{m}_a/\mathfrak{m}_a^2 \xrightarrow{\sim} T_a^* X$.*

Proof. For a linear form Λ on \mathbb{K}^n and $a \in \mathbb{K}^n$, $d_a(\Lambda - \Lambda(a)) = \Lambda$ on $T_a \mathbb{K}^n = \mathbb{K}^n$. Consequently, if $\ell \in \mathbb{K}[X]$ is then image of $\Lambda - \Lambda(a)$, then $\ell \in \mathfrak{m}_a$ and $d_a \ell$ is the restriction of Λ to $T_a X \subset \mathbb{K}^n$. As every linear form on $T_a X$ is the restriction of a linear form on \mathbb{K}^n , we conclude that the map $d_a: \mathfrak{m}_a \rightarrow T_a^* X$ is surjective.

Suppose that $g \in \mathfrak{m}_a$ and $d_a g$ vanishes on $T_a X$. Let h be a polynomial whose image in $\mathbb{K}[X]$ is g , and let f_1, \dots, f_m be polynomials that generate $\mathcal{I}(X)$. Since $T_a X$ is defined by the vanishing of $d_a f_1, \dots, d_a f_m$, and $d_a h$ vanishes on $T_a X$, there are $\lambda_1, \dots, \lambda_m \in \mathbb{K}$ such that

$$d_a h = \lambda_1 d_a f_1 + \lambda_2 d_a f_2 + \cdots + \lambda_m d_a f_m. \quad (3.8)$$

Set $h_1 := h - (\lambda_1 f_1 + \cdots + \lambda_m f_m)$. If we expand h_1 in the parameters v_1, \dots, v_n , where $v_i = x_i - a_i$ (as in (3.6)), then its constant term vanishes (as h and each f_i vanish at a) and its linear terms also vanish, by (3.8). Thus h_1 lies in the ideal $\langle v_1, \dots, v_n \rangle^2$. Since $g \in \mathbb{K}[X]$ is the image of h_1 and $\mathfrak{m}_a \subset \mathbb{K}[X]$ is the image of $\langle v_1, \dots, v_n \rangle$, we conclude that $g \in \mathfrak{m}_a^2$. This completes the proof. \square

We may therefore define the Zariski tangent space $T_a X$ independent of any embedding of X to be the vector space $(\mathfrak{m}_a/\mathfrak{m}_a^2)^*$. Suppose that we have a regular map $\varphi: X \rightarrow Y$ of affine varieties and point $a \in X$. The functorial pullback map $\varphi^*: \mathbb{K}[Y] \rightarrow \mathbb{K}[X]$ sends $\mathfrak{m}_{\varphi(a)}$ to \mathfrak{m}_a as a regular function $g \in \mathbb{K}[Y]$ that vanishes at $\varphi(a)$ has pullback that vanishes

at a . This also induces a map $\varphi^*: \mathfrak{m}_{\varphi(a)}/\mathfrak{m}_{\varphi(a)}^2 \rightarrow \mathfrak{m}_a/\mathfrak{m}_a^2$. Taking linear duals, we obtain a functorial linear map between tangent spaces $d_a\varphi: T_a X \rightarrow T_{\varphi(a)} Y$.

By Exercise 2, tangent spaces of affine varieties are unchanged in passing to principal affine open subsets. We use this to define Zariski tangent spaces for any variety. Given a variety X and a point $a \in X$, define $T_a X$ to be the Zariski tangent space $T_a U$ for any affine open subset $U \subset X$ containing a .

Suppose that X is irreducible and let m be the minimum dimension of a tangent space of X . By Theorem 3.4.3, the points of X whose tangent space has this minimum dimension form a nonempty open and hence dense subset of X . Call these points of X *smooth* points and write X_{sm} for the nonempty open subset of smooth points. The complement $X \setminus X_{\text{sm}}$ is the set X_{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 $\overline{X_{\text{sm}}} \cup X_{\text{sing}}$ of two proper closed subsets.

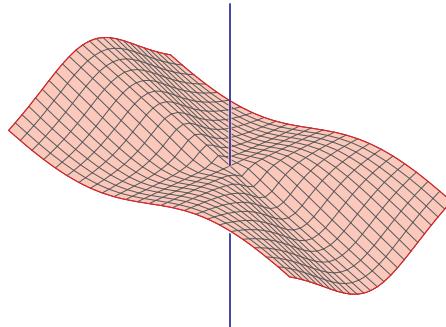
When $\mathbb{K} = \mathbb{C}$, the set of smooth points of X forms a complex manifold of whose dimension at a point $a \in X_{\text{sm}}$ is $\dim_{\mathbb{C}} T_a X$. This is a consequence of standard results about differential manifolds. Similarly, when $\mathbb{K} = \mathbb{R}$, if X has a smooth real point, then the set of smooth points of X is nonempty and forms a manifold whose dimension at a point $a \in X_{\text{sm}}$ is $\dim_{\mathbb{R}} T_a X$. This restriction is necessary, for it is possible that $X_{\text{sm}} = \emptyset$ for a real variety. For example, the real algebraic variety $X = \mathcal{V}(y^2 + x^2)$ has only one real point, the origin, where it is singular as $d_{(0,0)}(y^2 + x^2) = 0$.

Thus when X is smooth and irreducible and $\mathbb{K} = \mathbb{C}$, the dimension of the tangent space to X at a smooth point is equal to its dimension as a manifold. This remains true for any irreducible variety, smooth or not, and for any algebraically closed field. This gives a second definition of dimension which is distinct from the combinatorial definition of Definition 3.2.8.

We have the following facts concerning the locus of smooth and singular points on a real or complex variety

Proposition 3.4.5. *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 3.4.6. 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{R}^3 where the local dimension of its smooth points is either 1 (along the z axis) or 2 (along the ‘canopy’ of the umbrella). \diamond

Use this relation of dimension to tangent space to prove some of the theorems about dimension of varieties and then images under maps. This will be a precursor to the Bertini Theorems. These are needed in later sections in Toric varieties and in numerical algebraic geometry.

Exercises

1. Using that consequence of Lemma 3.4.4 that the Zariski tangent spaces of an affine variety are intrinsic to give another proof that the cuspidal cubic $\mathcal{V}(y^2 - x^3)$ is not isomorphic to either the parabola $\mathcal{V}(y - x^2)$ or to the line \mathbb{K}^1 . (This was shown in Example 1.3.1 of Section 1.3 by other means.)
2. Let $X \subset \mathbb{K}^n$ be an affine variety, $a \in X$, and $f \in \mathbb{K}[X]$ a regular function that does not vanish at a . Using the embedding $X_f \hookrightarrow \mathbb{K}^{n+1}$ given by $x \mapsto (x, f(x))$, show that the two tangent spaces $T_a X$ and $T_{(a, f(a))} X_f$ are isomorphic. We recommend using the definition of $T_a X$ as given in Definition 3.4.1.

3.5 Hilbert functions and dimension

The homogeneous coordinate ring $\mathbb{K}[X]$ of a projective variety $X \subset \mathbb{P}^n$ is an invariant of the variety X which determines it up to a linear automorphisms of \mathbb{P}^n . Basic numerical invariants, such as the dimension of X , are encoded in the combinatorics of $\mathbb{K}[X]$ and expressed through its Hilbert function. The coordinate ring of an affine variety $Y \subset \mathbb{K}^n$ also has a Hilbert function which determines its dimension, and it equals the Hilbert function of its projective closure $\bar{Y} \subset \mathbb{P}^n$.

The homogeneous coordinate ring $\mathbb{K}[X]$ of a projective variety is a graded ring,

$$\mathbb{K}[X] = \bigoplus_{r=0}^{\infty} \mathbb{K}[X]_r,$$

whose degree r component $\mathbb{K}[X]_r$ is equal to the quotient $\mathbb{K}[x_0, \dots, x_n]_r / \mathcal{I}(X)_r$. This is a finite-dimensional \mathbb{K} -vector space as $\dim_{\mathbb{K}} \mathbb{K}[x_0, \dots, x_n]_r = \binom{n+r}{n}$. Graded means that if $f \in \mathbb{K}[X]_r$ and $g \in \mathbb{K}[X]_s$, then $fg \in \mathbb{K}[X]_{r+s}$. The most basic numerical invariant of this ring is the *Hilbert function* of X , whose value at $r \in \mathbb{N}$ is the dimension of the r -th graded component of $\mathbb{K}[X]$,

$$\text{HF}_X(r) := \dim_{\mathbb{K}} (\mathbb{K}[X]_r).$$

This is also the number of linearly independent degree r homogeneous polynomials on X . We may also define the Hilbert function of a homogeneous ideal $I \subset \mathbb{K}[x_0, \dots, x_n]$,

$$\text{HF}_I(r) := \dim_{\mathbb{K}} (\mathbb{K}[x_0, \dots, x_n]_r / I_r).$$

Note that $\text{HF}_X = \text{HF}_{\mathcal{I}(X)}$.

Example 3.5.1. The space curve C of Figure 3.3 is the image of \mathbb{P}^1 under the map

$$\varphi : \mathbb{P}^1 \ni [s, t] \mapsto [s^3, s^2t, st^2, 2t^3 - 2s^2t] \in \mathbb{P}^3.$$

If \mathbb{P}^3 has coordinates $[w, x, y, z]$, 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 r is a homogeneous polynomial of degree $3r$ in the variables s, t , and all homogeneous forms of degree $3r$ in s, t occur as pullbacks. Since there are $3r + 1$ forms of degree $3r$ in s, t , we see that $\text{HF}_C(r) = 3r + 1$. \diamond

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 3.5.2. *Any reduced Gröbner basis for a homogeneous ideal I consists of homogeneous polynomials.*

Proof. Buchberger's algorithm is friendly to homogeneous polynomials. That is, if f and g are homogeneous, then so is $\text{Spol}(f, g)$. Similarly, the reduction of one homogeneous polynomial by another is a homogeneous polynomial. Since Buchberger's algorithm consists of forming S-polynomials and of reductions, 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 . \square

A consequence of Theorem 3.5.2 is that it is no loss of generality to use graded term orders when computing a Gröbner basis of a homogeneous ideal. Theorem 3.5.2 also implies that the linear isomorphism of Theorem 2.3.3 between $\mathbb{K}[x_0, \dots, x_n]/I$ and $\mathbb{K}[x_0, \dots, x_n]/\text{in}(I)$ respects degree and so the Hilbert functions of I and of $\text{in}(I)$ agree.

Corollary 3.5.3. *Let I be a homogeneous ideal. Then for any term order, $\text{HF}_I(r) = \text{HF}_{\text{in}(I)}(r)$, the number of standard monomials of degree r .*

Proof. The image in $\mathbb{K}[x_0, \dots, x_n]/I$ of a standard monomial of degree r lies in the r th graded component. Since the images of standard monomials are linearly independent, we only need to show that they span the degree r graded component of this ring. Let $f \in \mathbb{K}[x_0, \dots, x_n]$ be a homogeneous form of degree r and let G be a reduced Gröbner basis for I . Then the reduction $f \bmod G$ is a linear combination of standard monomials. Each of these will have degree r as G consists of homogeneous polynomials and the division algorithm is homogeneous-friendly. \square

Example 3.5.4. In the degree-reverse lexicographic monomial order with $x \succ y \succ z \succ w$, the polynomials

$$\underline{2y^2} - xz - 2yw, \quad \underline{2xy} - 2xw - zw, \quad \underline{x^2} - yw,$$

form the reduced Gröbner basis for the ideal of the cubic space curve C of Example 3.5.1. The initial terms are underlined, so the initial ideal is the monomial ideal $\langle y^2, xy, x^2 \rangle$.

The standard monomials of degree r are exactly the set

$$\{z^a w^b, xz^c w^d, yz^c w^d \mid a + b = r, c + d = r - 1\}$$

and so there are exactly $r + 1 + r + r = 3r + 1$ standard monomials of degree r . This agrees with the Hilbert function of C , as computed in Example 3.5.1. \diamond

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, which may be answered using combinatorial arguments.

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 3.5.5. Two functions $f, g: \mathbb{N} \rightarrow \mathbb{N}$ are *stably equivalent* if $f(r) = g(r)$ for r sufficiently large.

We prove the following result at the end of this section.

Proposition-Definition 3.5.6. *The Hilbert function of a homogeneous ideal I is stably equivalent to a polynomial, HP_I , called the *Hilbert polynomial* of I .*

The Hilbert polynomial of a projective variety encodes many of its numerical invariants. We explore two such invariants.

Definition 3.5.7. Let $X \subset \mathbb{P}^n$ be a projective variety and suppose that the initial term of its Hilbert polynomial is

$$\text{in}(\text{HP}_X(r)) = d \frac{r^m}{m!}.$$

Then the *dimension* of X is the degree, m , of the Hilbert polynomial and the coefficient d is the *degree* of X .

We computed the Hilbert function of the curve $C \subset \mathbb{P}^3$ of Example 3.5.1 to be $3r + 1$. This is also its Hilbert polynomial, and 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 , using the leading term of its Hilbert polynomial.

Example 3.5.8. In Exercise 3 you are asked to show that if X consists of d distinct points, then the Hilbert polynomial of X is the constant, d . Thus X has dimension 0 and degree d .

Suppose that X is a linear space, $\mathbb{P}(V)$, where $V \subset \mathbb{K}^{n+1}$ has dimension $m+1$. We may choose coordinates x_0, \dots, x_n on \mathbb{P}^n so that V is defined by $x_{m+1} = \dots = x_n = 0$, and so $\mathbb{K}[X] \simeq \mathbb{K}[x_0, \dots, x_m]$. Then $\text{HF}_X(r) = \binom{r+m}{m}$, which has initial term $\frac{r^m}{m!}$ and so X has dimension m and degree 1.

Suppose that $I = \langle f \rangle$, where f is homogeneous of degree d . Then

$$(\mathbb{K}[x_0, \dots, x_n]/I)_r = \frac{\mathbb{K}[x_0, \dots, x_n]_r}{f \cdot \mathbb{K}[x_0, \dots, x_n]_{r-d}}.$$

Since multiplication by f is injective, it follows that $\text{HF}_I(r) = \binom{r+n}{n} - \binom{r-d+n}{n}$, where if $a < n$, then $\binom{a}{n} = 0$. This is a polynomial for $r \geq d$. By Exercise 4, the leading term of the Hilbert polynomial of I is $d \frac{r^{n-1}}{(n-1)!}$, and so I has dimension $n-1$ and degree d . When f is square-free, we have that $I = \mathcal{I}(\mathcal{V}(f))$. This shows that the hypersurface defined by f has dimension $n-1$ and degree equal to the degree of f . \diamond

Remark 3.5.9. Suppose that $Y \subset \mathbb{K}^n$ is an affine variety with coordinate ring $\mathbb{K}[Y]$. This is not a graded ring, but we do have an increasing sequence of finite-dimensional subspaces $\mathbb{K}[Y]_{\leq r}$ for $r \in \mathbb{N}$, where $\mathbb{K}[Y]_{\leq r}$ is image in $\mathbb{K}[Y]$ of the linear span of polynomials in

$\mathbb{K}[x_1, \dots, x_n]$ of degree at most r . We define the Hilbert function HF_Y of the affine variety Y to be the function whose value at $r \in \mathbb{N}$ is $\dim_{\mathbb{K}} \mathbb{K}[Y]_{\leq r}$. Similarly, if $I \subset \mathbb{K}[x_1, \dots, x_n]$ is an ideal, then its Hilbert function is [Fill this in](#).

Also, relate this Hilbert function to that of the projective closure of Y ; perhaps make an exercise. \diamond

We need some additional results on inequalities of Hilbert functions and Hilbert polynomials. You are asked to prove the following lemma in Exercise 5

Lemma 3.5.10. *Suppose that $f, g: \mathbb{N} \rightarrow \mathbb{N}$ are functions that satisfy $f(r) \leq g(r)$ for all $r \in \mathbb{N}$. If, for $r \geq r_0$, f and g are equal to polynomials F and G , respectively, then $\deg F \leq \deg G$. If additionally there is an $a \in \mathbb{N}$ with $g(r-a) \leq f(r)$ for $r \geq a$, then $\deg F = \deg G$ and their leading coefficients are equal.*

Theorem 3.5.11. *Suppose that X is a projective variety of dimension m and degree d . Every subvariety of X has dimension at most m , at least one irreducible component of X has dimension m , and d is the sum of the degrees of the irreducible components of dimension m .*

Proof. Let Y be a subvariety of X . Then the coordinate ring of Y is a quotient of the coordinate ring of X , so $\text{HF}_Y(r) \leq \text{HF}_X(r)$ for all r . By Lemma 3.5.10, the degree of the Hilbert polynomial of Y is at most the degree of the Hilbert polynomial of X , and thus $\dim Y \leq \dim X$.

Suppose that $X = X_1 \cup \dots \cup X_k$ is the decomposition of X into irreducible components. Consider the map of graded vector spaces which is induced by restriction

$$\mathbb{K}[X] \longrightarrow \mathbb{K}[X_1] \oplus \mathbb{K}[X_2] \oplus \dots \oplus \mathbb{K}[X_k].$$

This is injective, which implies the inequality

$$\text{HF}_X(r) \leq \sum_{i=1}^r \text{HF}_{X_i}(r). \quad (3.9)$$

Passing to Hilbert polynomials, by Lemma 3.5.10, there is some component X_i of X whose Hilbert polynomial has degree at least m , which implies that $\dim X_i = \dim X$.

A consequence of Lemma 3.5.12, which is stated below, is that there is a number a such that when $r > a$, we have

$$\sum_{i=1}^r \text{HF}_{X_i}(r-a) \leq \text{HF}_X(r). \quad (3.10)$$

Each Hilbert function is eventually equal to the corresponding Hilbert polynomial, so that the sum in (3.9) is eventually equal to the sum, $\sum_i \text{HP}_{X_i}$ of Hilbert polynomials of the components. By Lemma 3.5.10 and the inequalities (3.9) and (3.10), the polynomial that is the sum of Hilbert polynomials of the components has the same degree and leading

term as does the Hilbert polynomial of X . But this leading term is the sum of leading terms of the Hilbert polynomials of components that have dimension m , which completes the proof. \square

Lemma 3.5.12. *Suppose that $X = X_1 \cup \dots \cup X_k$ is the decomposition of X into irreducible components. There is a positive integer a such that when $r \geq a$, we have*

$$\sum_{i=1}^r \text{HF}_{X_i}(r-a) \leq \text{HF}_X(r).$$

Proof. For each $i = 1, \dots, k$, let $f_i \in \mathbb{K}[X]$ be a nonzero element that vanishes on $X \setminus X_i$. As $0 \neq f_i$, it does not also vanish on X_i , so its image in the domain $\mathbb{K}[X_i]$ is nonzero. Consider the map $\mu: g \mapsto f_i g$ on $\mathbb{K}[X]$. If $g \in \mathcal{I}(X_i)$, then $f_i g$ is identically zero on X , so that $\mathcal{I}(X_i) \subset \ker \mu$, and thus multiplication by f_i is a well-defined map $\mathbb{K}[X_i] \rightarrow \mathbb{K}[X]$.

Consequently, the expression $(g_1, \dots, g_k) \mapsto f_1 g_1 + \dots + f_k g_k$ induces a map

$$\varphi : \mathbb{K}[X_1] \oplus \dots \oplus \mathbb{K}[X_k] \longrightarrow \mathbb{K}[X].$$

This is an injection because if $f_1 g_1 + \dots + f_k g_k = 0$, then $g_i = 0$ for all i . Indeed, the image of $f_1 g_1 + \dots + f_k g_k$ in $\mathbb{K}[X_i]$, which is $f_i g_i$, is also zero. But this implies that $g_i = 0$ as $0 \neq f_i$ as an element of the integral domain $\mathbb{K}[X_i]$.

To complete the proof, let a be any number so that $a \geq \deg(f_i)$ for each i . Then if we replace each f_i by $f_i g_i$ where $0 \neq g_i \in \mathbb{K}[X_i]$ has degree $a - \deg(f_i)$ we may assume that each f_i has degree a , so that the map φ restricts to an injection

$$\varphi : \mathbb{K}[X_1]_{r-a} \oplus \dots \oplus \mathbb{K}[X_k]_{r-a} \longrightarrow \mathbb{K}[X]_r.$$

which proves the inequality of the corollary. \square

We use combinatorics to prove the following at the end of the section.

Theorem 3.5.13. *When I is a monomial ideal, the degree of HP_I is the dimension of the largest linear subspace contained in $\mathcal{V}(I) \subset \mathbb{P}^n$.*

As $\mathcal{V}(I) = \mathcal{V}(\sqrt{I})$, we deduce the following corollary.

Corollary 3.5.14. *If I is a monomial ideal, then the Hilbert polynomials HP_I and $\text{HP}_{\sqrt{I}}$ have the same degree.*

Theorem 3.5.15. *Let X be a subvariety of \mathbb{P}^n and suppose that $f \in \mathbb{K}[X]$ 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 $d \cdot \deg(X)$.*

If X is irreducible, then every proper subvariety has dimension at most $m-1$ and X has a subvariety of dimension $m-1$.

Proof. For $r \geq d$, the degree r component of the quotient ring $\mathbb{K}[x_0, \dots, x_n]/\langle \mathcal{I}(X), f \rangle$ is

$$\mathbb{K}[X]_r/f \cdot \mathbb{K}[X]_{r-d}, \quad (3.11)$$

and so it has dimension $\dim_{\mathbb{K}}(\mathbb{K}[X]_r) - \dim_{\mathbb{K}}(f \cdot \mathbb{K}[X]_{r-d})$.

Suppose that r is large enough so that the Hilbert function of X is equal to its Hilbert polynomial at $r-d$ and all larger integers. Since f is not a zero divisor, multiplication by f is injective. Thus the dimension of (3.11) is

$$\text{HP}_X(r) - \text{HP}_X(r-d).$$

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 4.

Suppose now that X is irreducible, let Y be a proper subvariety of X , and let $0 \neq f \in \mathcal{I}(Y) \subset \mathbb{K}[X]$. Since $\mathbb{K}[X]/\langle f \rangle \twoheadrightarrow \mathbb{K}[X]/\mathcal{I}(Y) = \mathbb{K}[Y]$, we see that the Hilbert polynomial of $\mathbb{K}[Y]$ has degree at most that of $\mathbb{K}[X]/\langle f \rangle$, 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 for 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 observe that we have the chain of inclusions

$$\text{in}(I) \subset \text{in}(\sqrt{I}) \subset \sqrt{\text{in}(I)}. \quad (3.12)$$

Indeed, $I \subset \sqrt{I}$, which implies that $\text{in}(I) \subset \text{in}(\sqrt{I})$. For the other inclusion, let $f \in \sqrt{I}$. Then $f^N \in I$ for some $N \in \mathbb{N}$, which implies that $\text{in}(f)^N = \text{in}(f^N) \in \text{in}(I)$. But then $\text{in}(f) \in \sqrt{\text{in}(I)}$.

This chain of inclusions (3.12) implies corresponding surjections of the coordinate rings, and therefore the chain of inequalities of Hilbert functions, $\text{HF}_{\text{in}(I)}(r) \geq \text{HF}_{\text{in}(\sqrt{I})}(r) \geq \text{HF}_{\sqrt{\text{in}(I)}}(r)$. This implies the chain of inequalities degree of Hilbert polynomials,

$$\deg(\text{HP}_{\text{in}(I)}) \geq \deg(\text{HP}_{\text{in}(\sqrt{I})}) \geq \deg(\text{HP}_{\sqrt{\text{in}(I)}}).$$

By Corollary 3.5.14, $\deg(\text{HP}_{\text{in}(I)}) = \deg(\text{HP}_{\sqrt{\text{in}(I)}})$, so all three degrees are equal. As $\text{HP}_I = \text{HP}_{\text{in}(I)}$, and the same for \sqrt{I} , which is the ideal of $\mathcal{V}(I)$, we conclude that $\mathcal{V}(I)$ is a subvariety of X having dimension $d-1$. \square

We may now show that the combinatorial definition (Definition 3.2.8) of dimension agrees with the definition given in terms of Hilbert function.

Corollary 3.5.16 (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 Theorem 3.5.11, $\dim(X_{i-1}) > \dim(X_i)$ for $i = 1, \dots, m$, 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'_1 with the same dimension. In the same fashion, for each $i = 2, \dots, \dim(X)$, 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. \square

Similarly, the definition of dimension in terms of tangent spaces (Definition ??????) agrees with the definition given in terms of Hilbert function. For this, I think that we should appeal to differential geometry, such as if $d_x f \neq 0$, and $x \in X$ is a smooth point, then $\mathcal{V}(f)$ is smooth in a neighborhood of x in X .

We now turn to the proof of Hilbert's Theorem 3.5.6, that the Hilbert function of a projective variety or homogeneous ideal is stably equivalent to a polynomial. We prove this for Hilbert functions of a more general class of objects, finitely generated graded modules over a polynomial ring.

A *module* over $\mathbb{K}[x] = \mathbb{K}[x_0, x_1, \dots, x_n]$ (*$\mathbb{K}[x]$ -module*) is a vector space M over \mathbb{K} , together with a ring homomorphism $\psi: \mathbb{K}[x] \rightarrow \text{End}(M)$. Here, $\text{End}(M)$ is the set of linear transformations $M \rightarrow M$. This is a \mathbb{K} -vector space with a multiplication induced by composition, and $\text{End}(M)$ is a noncommutative ring. Through ψ , polynomials in $\mathbb{K}[x]$ act as \mathbb{K} -linear transformations of the vector space M . We suppress the map ψ ; simply writing $f.u$ for $\psi(f)(u)$, the image of $u \in M$ under the linear map $\psi(f)$, for $f \in \mathbb{K}[x]$.

The polynomial ring $\mathbb{K}[x]$ is a $\mathbb{K}[x]$ -module, as $\mathbb{K}[x]$ acts linearly on itself by multiplication. An ideal I of $\mathbb{K}[x]$ is a $\mathbb{K}[x]$ -module, and ideals are exactly the $\mathbb{K}[x]$ -submodules of $\mathbb{K}[x]$. Quotients of modules are also modules, so that a quotient ring $\mathbb{K}[x]/I$ is a $\mathbb{K}[x]$ -module. A module M is *finitely generated* if there exist finitely many elements $u_1, \dots, u_k \in M$ such that every element u of M has an expression

$$u = f_1.u_1 + f_2.u_2 + \cdots + f_k.u_k$$

as a $\mathbb{K}[x]$ -linear combination of u_1, \dots, u_k ($f_1, \dots, f_k \in \mathbb{K}[x]$).

A module M is *graded* if it has a decomposition

$$M = \bigoplus_{s \in \mathbb{Z}} M_s,$$

where for each $s \in \mathbb{Z}$, M_s is a vector subspace of M and the $\mathbb{K}[x]$ -action respects this decomposition. That is, for all $r \in \mathbb{N}$ and $s \in \mathbb{Z}$, if $f \in \mathbb{K}[x]_r$ is homogeneous of degree r and $u \in M_s$, then $f.u \in M_{r+s}$.

Lemma 3.5.17. *If M is a finitely generated graded $\mathbb{K}[x]$ -module, then each graded component of M is a finite-dimensional vector space.*

Proof. Let u_1, \dots, u_k be generators of M with $u_i \in M_{s_i}$. For $s \in \mathbb{Z}$, every element $u \in M_s$ has an expression

$$u = f_1 \cdot u_1 + f_2 \cdot u_2 + \cdots + f_k \cdot u_k$$

as a $\mathbb{K}[x]$ -linear combination of u_1, \dots, u_k . Here $f_i \in \mathbb{K}[x]_{s-s_i}$. Thus there is a surjection

$$\bigoplus_{i=1}^k \mathbb{K}[x]_{s-s_i} \longrightarrow M_s .$$

This completes the proof, as the graded components of $\mathbb{K}[x]$ are finite-dimensional. \square

The main consequence of Lemma 3.5.17 is that a finitely generated graded module M has a Hilbert function, defined by $HF_M(s) = \dim_{\mathbb{K}} M_s$.

Theorem 3.5.18. *If M is a finitely generated graded $\mathbb{K}[x_0, x_1, \dots, x_n]$ -module, then its Hilbert function is stably equivalent to a polynomial of degree $m \leq n$. If $a s^m$ is the leading term of this polynomial, then $m!a$ is a nonnegative integer.*

Our proof will use induction on the number of variables, as well as maps of graded modules. A map $\varphi: M \rightarrow N$ of graded modules is a collection of linear maps

$$\varphi_s : M_s \longrightarrow N_s \quad \text{for } s \in \mathbb{Z}$$

such that for every homogeneous polynomial $f \in \mathbb{K}[x]_r$ and $u \in M_s$ we have

$$\varphi_{r+s}(f \cdot u) = f \cdot \varphi_s(u) \quad \text{in } N_{r+s} .$$

That is, the map φ is a map of modules that respects the grading (f has degree 0).

A consequence of this definition is that if $f \in \mathbb{K}[x]_r$ and M is a graded module, then multiplication by f is a linear map that sends M_s to M_{r+s} , but it is not *a priori* a map of graded modules. This is remedied by changing the grading in the domain of this multiplication map. Define a new graded module $M(-r)$ by $M(-r)_s := M_{s-r}$. Then multiplication by $f \in \mathbb{K}[x]_r$ is a map of graded modules $M(-r) \rightarrow M$.

Proof. When there are no variables, M is a finite-dimensional vector space, and so there is an integer s_0 with $M_s = 0$ for all $s \geq s_0$. In this case, HF_M is stably equivalent to 0, a polynomial of degree -1 .

Now suppose that $r \geq 0$ and assume that the theorem holds for r variables. Let M be a finitely generated graded $\mathbb{K}[x_0, \dots, x_r]$ -module, and let $K \subset M$ be the kernel of the

linear map induced by multiplication by x_r . This gives an exact sequence of graded vector spaces,

$$0 \longrightarrow K(-1) \longrightarrow M(-1) \xrightarrow{x_r} M \longrightarrow M/x_r.M \longrightarrow 0.$$

For any $s \in \mathbb{Z}$, if we take the dimension of the s th graded components, then the rank-nullity theorem implies that

$$0 = \dim_{\mathbb{K}} K(-1)_s - \dim_{\mathbb{K}} M(-1)_s + \dim_{\mathbb{K}} M_s - \dim_{\mathbb{K}} (M/x_r.M)_s,$$

or, and using that $K(-1)_s = K_{s-1}$, and the same for $M(-1)$,

$$\dim_{\mathbb{K}} M_s - \dim_{\mathbb{K}} M_{s-1} = \dim_{\mathbb{K}} (M/x_r.M)_s - \dim_{\mathbb{K}} K(-1)_s.$$

Observe that both K and $M/x_r.M$ are finitely generated modules over $\mathbb{K}[x_0, \dots, x_{r-1}]$.

Why K ? By our induction hypothesis, the Hilbert function of each is stably equivalent to a polynomial of degree at most $r-1$. If m is the degree of the polynomial, then the coefficient a of its leading term as^m has $m!a$ a nonnegative integer. The same is true for their difference, (but the integer could be negative). Let $P(s)$ be this polynomial and s_0 the integer such that for $s \geq s_0$, $P(s) = \dim_{\mathbb{K}} (M/x_r.M)_s - \dim_{\mathbb{K}} K(-1)_s$. Suppose that P has degree m and leading coefficient ad^m .

Then, for $s \geq s_0$, $P(s) = \text{HF}_M(s) - \text{HF}_M(s-1)$, and we have

$$\text{HF}_M(s) = \text{HF}_M(s_0) + \sum_{t=s_0}^s P(t).$$

But **Exercise** the right hand side is a polynomial in s of degree $m+1$ and its leading term is $\frac{a}{m+1}s^{m+1}$. This completes the proof. \square

Exercises

1. Show that the dimension of the space $\mathbb{K}[x_0, \dots, x_n]_m$ of homogeneous polynomials of degree m is $\binom{m+n}{n} = \frac{m^n}{n!} + \text{lower order terms in } m$.
2. Let I be a homogeneous ideal. Show that $HF_I \sim HF_{(I : \mathfrak{m}_0)} \sim HF_{I_{\geq d}}$.
3. Show that if $X \subset \mathbb{P}^n$ consists of d points, then, for r sufficiently large, we have $\mathbb{K}[X]_r \simeq \mathbb{K}^d$, and so $HP_X(r) = d$.
4. 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 $mba_0 t^{m-1}$.
5. Prove Lemma 3.5.10
6. 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

$$\langle \underline{x^2+y^2-z^2-w^2}, \underline{y^2z-xz^2-z^3+xyw+yzw-zw^2}, \underline{xyz-y^2w-xzw+yw^2}, \\ \underline{y^3-yz^2-y^2w+z^2w}, \underline{xy^2-xyw-yzw+zw^2} \rangle$$
 - (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.

Draw a picture of these?

7. Show that if $A \subset \mathbb{K}^n$ and $B \subset \mathbb{K}^m$ are subvarieties of degrees a and b , respectively, then $A \times B \subset \mathbb{K}^n \times \mathbb{K}^m$ has degree ab .

3.6 Bertini Theorems

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 $\mathbb{K}[X]$.

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

$$\text{HP}_Y = \text{HP}_{\langle \mathcal{I}(X), \Lambda \rangle},$$

and so by Theorem 3.5.15, $\deg(Y) = \deg(X)$. If $Y \subset \mathbb{P}^n$ has dimension d , then we say that Y has *codimension* $n - d$.

Corollary 3.6.1 (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 3.3 has degree 3, and we see in that figure that it meets the plane $z = 0$ in 3 points.

Does this belong anywhere? Suppose that the ideal I generated by the polynomials f_i of (2.15) is not zero-dimensional, and we still want to count the isolated solutions to (2.15). 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 (2.15). 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.[§]

Exercises

3.7 Notes

Mention about the origin of Zariski topology.

[†]Not formulated here, yet!

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

Chapter 4

Numerical Algebraic Geometry

Outline:

1. Core Numerical Algorithms.	113—122
2. Homotopy continuation.	123—133
3. Numerical Algebraic Geometry	134—142
4. Numerical Irreducible Decomposition	143—145
5. Regeneration	146—
6. Smale’s α -theory	147—
7. Notes	148

Chapter 2 presented fundamental symbolic algorithms, including the classical resultant and general methods based on Gröbner bases. Those algorithms operate on the algebraic side of the algebraic-geometric dictionary underlying algebraic geometry. This chapter develops the fundamental algorithms of numerical algebraic geometry, which uses tools from numerical analysis to represent and study algebraic varieties on a computer. As we will see, this field primarily operates on the geometric side of our dictionary. Since numerical algebraic geometry involves numerical computation, we will work over the complex numbers.

4.1 Core Numerical Algorithms

Numerical algebraic geometry rests on two core numerical algorithms, which go back at least to Newton and Euler. Newton’s method refines approximations to solutions to systems of polynomial equations. A careful study of its convergence leads to methods for certifying numerical output, which we describe in Section 4.5. The other core algorithm comes from Euler’s method for computing a solution to an initial value problem. This is a first-order iterative method, and more sophisticated higher-order methods are used in practice for they have better convergence. These two algorithms are used together for path-tracking in numerical homotopy continuation, which we develop in subsequent sections as a tool for solving systems of polynomial equations and for manipulating varieties

on a computer. While these algorithms are standard in introductory numerical analysis, they are less familiar to algebraists. Our approach is intended to be friendly to algebraists.

By the Fundamental Theorem of Algebra, a univariate polynomial $f(x)$ of degree n has n complex zeroes. When the degree of f is at most four, there are algorithmic formulas for these zeroes that involve arithmetic operations and extracting roots. Zeroes of linear polynomials go back to the earliest mathematical writing, such as the Rhind Papyrus, and the Babylonians had a method for the zeroes of quadratic polynomials that is the precursor of the familiar quadratic formula, which was made explicit by Bramagupta c. 600 CE. The 16th century Italians del Ferro, Tartaglia, Cardano, and Ferrari colorfully gave formulas for the zeroes of cubic and quartic polynomials. It was only in 1823 that Niels Hendrik Abel proved there is no universal such formula for the zeroes of polynomials of degree five and higher.

It was later shown that the zeroes of a general polynomial cannot be expressed in terms of the coefficients using only arithmetic operations on its coefficients and extracting roots. For example, the zeroes of this sextic

$$f := x^6 + 2x^5 + 3x^4 + 5x^3 + 7x^2 + 11x - 13 \quad (4.1)$$

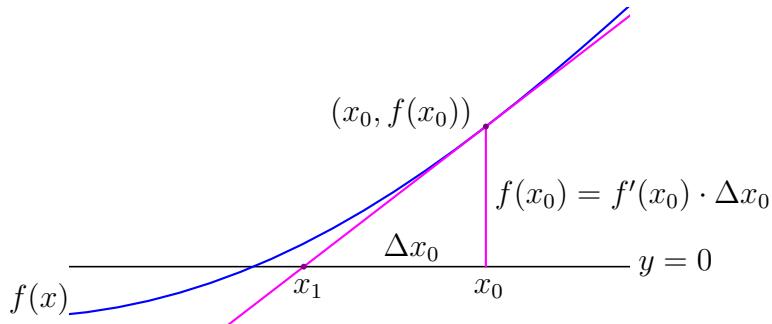
admit no such expression. This is not to say there is no formula for the zeroes of polynomials of degree five or more. For example, there are hypergeometric power series for those zeroes that depend upon the coefficients.

Numerical methods offer another path to the roots of univariate polynomials. While numerical linear algebra may be combined with the eigenvalue approaches to solving of Section 2.5, we will discuss algorithms based on Newton's method.

Newton's method uses the tangent-line approximation to the graph of a differentiable function $f(x)$ to refine an approximation x_0 to a zero of f . The tangent line to the graph of $f(x)$ at the point $(x_0, f(x_0))$ has equation

$$y = f'(x_0)(x - x_0) + f(x_0).$$

If $f'(x_0) \neq 0$, we solve this for $y = 0$ to get the formula $x_1 := x_0 - (f'(x_0))^{-1}f(x_0)$ for the refinement of x_0 . This may also be read from the graph.



Using operator notation Df for the derivative of f , we obtain the expression

$$N_f(x) := x - (Df(x))^{-1}f(x) \quad (4.2)$$

for the passage from x_0 to x_1 above. This *Newton iteration* (4.2) is the basis for Newton's method to compute a zero of a function $f(x)$:

- Start with an initial value x_0 , and
- While $Df(x_i) \neq 0$, compute the sequence $\{x_i \mid i \in \mathbb{N}\}$ using the recurrence $x_{i+1} = N_f(x_i)$.

For $f(x) = x^2 - 2$ with $x_0 = 1$, if we compute the first seven terms of the sequence,

$$\begin{aligned} x_1 &= 1.5 \\ x_2 &= 1.416\bar{6} \\ x_3 &= 1.4142156862745098039\bar{2156862745098039} \\ x_4 &= 1.4142135623746899106262955788901349101165596221157440445849050192000 \\ x_5 &= 1.4142135623730950488016896235025302436149819257761974284982894986231 \\ x_6 &= 1.4142135623730950488016887242096980785696718753772340015610131331132 \\ x_7 &= 1.4142135623730950488016887242096980785696718753769480731766797379907, \end{aligned}$$

then the 58 displayed digits of x_7 are also the first 58 digits of $\sqrt{2}$.

This example suggests that Newton's method may converge rapidly to a solution. It is not always so well-behaved. For example, if $f(x) = x^3 - 2x + 2$, then $N_f(0) = 1$ and $N_f(1) = 0$, and so the sequence $\{x_i\}$ of Newton iterates with $x_0 = 0$ is periodic with period 2, and does not converge to a root of f .

In fact Newton's method is about as badly behaved as it can be, even for polynomials. The *basin of attraction* for a zero x^* of f is the set of all complex numbers x_0 such that Newton's method starting at x_0 converges to x^* . In general, the boundary of a basin of attraction is a fractal Julia set. Figure 4.1 shows basins of attraction for two univariate cubic polynomials. On the left are those for the polynomial $f(x) = x^3 - 1$, in the region $|\Re(x)|, |\Im(x)| \leq 1.3$. This polynomial vanishes at the cubic roots of unity, and each is a fixed point of a Newton iteration. There are three basins, one for each root of f , and their union is dense in the complex plane. Each basin is drawn in a different color.

On the right of Figure 4.1 are basins for the polynomial $f(x) = x^3 - 2x + 2$. The roots of f give three fixed points of a Newton iteration, and we noted there is an orbit of period 2. The roots and the orbit each have a basin of attraction and each basin is drawn in a different color. The basin of attraction for the orbit of period 2 is in red.

Despite this complicated behaviour, Newton's method is a foundation for numerical algebraic geometry, and it may be used to certify the results of numerical computation. To understand why this is so, we investigate its convergence.

Suppose that f is twice continuously differentiable in a neighborhood of a zero ζ of f and that $Df(\zeta)^{-1}$ exists. Differentiating the expression (4.2) for $N_f(x)$ gives

$$\begin{aligned} DN_f(x) &= 1 - Df(x)^{-1}Df(x) + Df(x)^{-1}D^{(2)}f(x)Df(x)^{-1}f(x) \\ &= Df(x)^{-1}D^{(2)}f(x)Df(x)^{-1}f(x). \end{aligned}$$

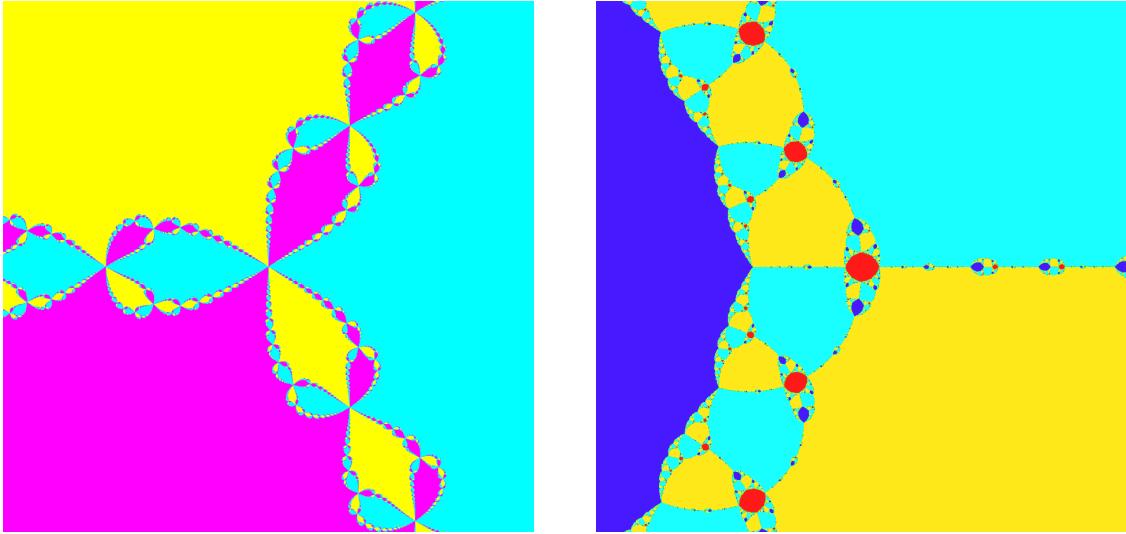


Figure 4.1: Basins of attraction for two cubics

As $f(\zeta) = 0$, we have $DN_f(\zeta) = 0$ and $N_f(\zeta) = \zeta$. The first order Taylor expansion of $N_f(x)$ about the point $x = \zeta$ with remainder is then

$$N_f(x) = \zeta + \frac{1}{2}D^{(2)}N_f(a)(x - \zeta)^2,$$

where a is some point with $|a - \zeta| < |x - \zeta|$, and we used that $N_f(\zeta) = \zeta$. If we set $c := \max\{\frac{1}{2}|D^{(2)}N_f(a)| \mid |a - \zeta| < r\}$, for some $r > 0$, then we have

$$|N_f(x) - \zeta| \leq c|x - \zeta|^2,$$

whenever $|x - \zeta| \leq r$. Suppose now that $|x - \zeta| < \min\{\frac{1}{2c}, r\}$. Then,

$$\begin{aligned} |N_f(x) - \zeta| &< c \cdot \left(\frac{1}{2c}\right)^2 = \frac{1}{4c}, \quad \text{and} \\ |N_f^2(x) - \zeta| &< c \cdot \left(\frac{1}{4c}\right)^2 = \frac{1}{16c}, \end{aligned}$$

and in general, if $N_f^i(x)$ is the i th iteration of N_f applied to z , we have

$$|N_f^i(x) - \zeta| < 2^{1-2^i} \frac{1}{2c}.$$

This implies that the number of digits of ζ that have been computed in x_{i+1} (the number of *significant digits* in x_{i+1}) will be approximately twice the number of significant digits in x_i . We saw this when computing $\sqrt{2}$ using Newton's method as x_1, x_2, \dots, x_6 had 1, 3, 6, 12, 24, and 48 correct decimal digits of $\sqrt{2}$.

This has a straightforward extension to the problem of approximating zeroes to a square system of multivariate polynomials,

$$F : f_1 = f_2 = \dots = f_n = 0, \quad (4.3)$$

where each f_i is a polynomial in n variables. Here *square* means that the number of equations equals the number of variables and (4.3) defines a zero-dimensional variety. It is useful to consider F to be a polynomial map,

$$F = (f_1, \dots, f_n) : \mathbb{C}^n \longrightarrow \mathbb{C}^n,$$

and write the solutions $\mathcal{V}(F)$ as $F^{-1}(0)$. Given a point $x \in \mathbb{C}^n$ where the Jacobian matrix $DF(x)$ of partial derivatives of f_1, \dots, f_n with respect to the components of x is invertible, the *Newton iteration* of F applied to x is

$$NF(x) := x - DF(x)^{-1}F(x).$$

The geometry of this map is the same as for univariate polynomials: $NF(x)$ is the unique zero of the linear approximation of F at x . This is the solution $\zeta \in \mathbb{C}^n$ to

$$0 = F(x) + DF(x)(\zeta - x).$$

The elementary analysis of iterating Newton steps is also the same. As before, Newton steps define a chaotic dynamical system on \mathbb{C}^n , but if x is sufficiently close to a zero ζ of F , then Newton iterations starting at x converge rapidly to ζ .

We quantify this. A sequence $\{x_i \mid i \in \mathbb{N}\} \subset \mathbb{C}^n$ *converges quadratically* to a point $\zeta \in \mathbb{C}^n$ if for all $i \in \mathbb{N}$,

$$\|x_i - \zeta\| \leq 2^{1-2^i} \|x_0 - \zeta\|. \quad (4.4)$$

For example, the sequence of Newton iterations for $x^2 - 2$ beginning with $x_0 = 1$ that we computed converges quadratically to $\sqrt{2}$. A point $x \in \mathbb{C}^n$ is an *approximate zero* of F with *associated zero* $\zeta \in \mathbb{C}^n$ if $F(\zeta) = 0$ and if the sequence of Newton iterates defined by $x_0 := x$ and $x_{i+1} := NF(x_i)$ for $i \geq 0$ converges quadratically to ζ .

Knowing an approximate zero x to a system F is a well-behaved relaxation of the problem of knowing its associated zero ζ . Indeed, the sequence of Newton iterations starting with x reveals as many digits of ζ as desired, in a controlled manner.

At this point, one should ask for methods to determine if x is an approximate zero to F . A heuristic that is used in practice is to treat the length of a Newton step

$$\beta(F, x) := \|x - NF(x)\| = \|DF(x)^{-1}F(x)\|, \quad (4.5)$$

as a proxy for the distance $\|x - \zeta\|$ to the zero ζ of F . If we are in the basin of quadratic convergence to ζ , then we expect that

$$\frac{1}{2}\|x - \zeta\| \lesssim \beta(F, x) \lesssim \|x - \zeta\|.$$

For the heuristic, compute two Newton iterations, $NF(x)$ and $NF^2(x)$. If we have that

$$\beta(F, NF(x)) \leq \beta(F, x)^2, \quad (4.6)$$

with $\beta(F, x)$ below some threshold $\beta \ll 1$, then we replace x by $NF^2(x)$, and declare it to be an approximate zero of F with some certitude. The condition (4.6) implies that the number of digits common to $NF(x)$ and $NF^2(x)$ is at least twice the number of digits common to x and $NF(x)$. See Exercise 5 for potential limitations of this heuristic.

We can do much better than this heuristic. Smale studied the convergence of Newton's method and developed what is now called *α -theory* after a computable constant $\alpha = \alpha(F, x)$ such that if α is sufficiently small, then x is an approximate zero of F . We present this theory in Section 4.5. For now, we define $\alpha(F, x)$ and state Smale's theorem.

The constant $\alpha(F, x)$ is the product of two numbers. The first is the length $\beta(F, x)$ (4.5) of a Newton step at x . For the second, recall the Taylor expansion of F at x , [Relate this to /eqrefEq:Taylor](#).

$$F(w) = F(x) + DF(x)(w - x) + D^2F(x)(w - x)^2 + \cdots + D^N F(x)(w - x)^N,$$

where the polynomial map F has degree N . Let us describe the meaning of the terms in this Taylor expansion. For $v \in \mathbb{C}^n$, v^k is the symmetric tensor indexed by all exponents $a \in \mathbb{N}^n$ of degree k , where

$$(v^k)_a = \frac{1}{a_1!} \frac{1}{a_2!} \cdots \frac{1}{a_n!} v_1^{a_1} v_2^{a_2} \cdots v_n^{a_n} =: \frac{1}{a!} v^a.$$

Let $S^k \mathbb{C}^n$ be this vector space of symmetric tensors. Writing $x = (x_1, \dots, x_n) \in \mathbb{C}^n$, the i th component of $D^k F(x)$ is the vector of partial derivatives of F_i of order k ,

$$(D^k F_i(x))_a = \left(\frac{\partial}{\partial x_1} \right)^{a_1} \left(\frac{\partial}{\partial x_2} \right)^{a_2} \cdots \left(\frac{\partial}{\partial x_n} \right)^{a_n} F_i(x),$$

for $a \in \mathbb{N}^n$ of degree k . Then the i th component of $D^k F(x)(w - x)^k$ is the sum

$$\sum_{|a|=k} D^a F_i(x) \frac{1}{a!} (w - x)^a.$$

Thus $D^k F(x)$ is a linear map from the space $S^k \mathbb{C}^n$ of symmetric tensors to \mathbb{C}^n . The same is true for the composition $DF(x)^{-1} \circ D^k F(x)$. If we use the standard norm for vectors $z \in \mathbb{C}^n$ and $v \in S^k \mathbb{C}^n$,

$$\|z\| := \left(\sum_{i=1}^n |z_i|^2 \right)^{1/2} \quad \text{and} \quad \|v\| := \left(\sum_{|a|=k} |v_a|^2 \right)^{1/2},$$

then the operator norm of this composition is

$$\|DF(x)^{-1} \circ D^k F(x)\| := \max_{\|v\|=1} \|DF(x)^{-1} \circ D^k F(x)(v^k)\|.$$

With these definitions, set

$$\gamma(F, x) := \max_{k \geq 2} \frac{1}{k!} \|DF(x)^{-1} \circ D^k F(x)\|^{\frac{1}{k-1}},$$

and then define

$$\alpha(F, x) := \beta(F, x) \cdot \gamma(F, x).$$

Theorem 4.1.1. *If $\alpha(F, x) < \frac{1}{4}(13 - 3\sqrt{17}) \simeq 0.15767\dots$, then x is an approximate zero of F . The distance from x to its associated zero is at most $2\beta(F, x)$.*

We prove Theorem 4.1.1 and some extensions in Section 4.5. Note the similarity between the formula for $\gamma(F, x)$ and the root test/formula for the radius of convergence of a power series. The shift in the radical from $\frac{1}{k}$ to $\frac{1}{k-1}$ is because this is applied to the expansion of DF .

We apply this to the quadratic polynomial $f(x) = x^2 - 2$. When $x \in \mathbb{C}$ is non-zero, $f'(x) \neq 0$, and we have

$$\beta(f, x) = \left| \frac{f(x)}{f'(x)} \right| \quad \text{and} \quad \gamma(f, x) = \frac{1}{2} \left| \frac{f''(x)}{f'(x)} \right|.$$

Then $\alpha(f, x) = \frac{1}{2} |f(x)f''(x)/(f'(x))^2| = |\frac{x^2-2}{4x^2}|$.

Observe that $\alpha(f, 1) = \frac{1}{4} > \frac{1}{4}(13 - 3\sqrt{17})$. Thus, while Newton iterations starting at $x_0 = 1$ converge quadratically to $\sqrt{2}$, this quadratic convergence is not detected with Smale's α -theory. Note that $\alpha(f, 3/2) = \frac{1}{36} < \frac{1}{4}(13 - 3\sqrt{17})$, so that α -theory certifies the quadratic convergence of Newton iterations starting with $x_0 = 3/2$.

Consider the regions of convergence of Newton's method for the zeroes of $x^2 - 2$ in the complex plane. First, when $\Re(x) > 0$, Newton iterations beginning with x converge to $\sqrt{2}$ and when $\Re(x) < 0$, iterations beginning with x converge to $-\sqrt{2}$. In Figure 4.2, a point $x \in \mathbb{C}$ is yellow if Newton iterations converge quadratically, and magenta otherwise. The zeroes $\pm\sqrt{2}$ are as indicated and the convex regions enclosing them indicate the points whose quadratic convergence is certified using α -theory ($\alpha(f, z) < \frac{1}{4}(13 - 3\sqrt{17})$).

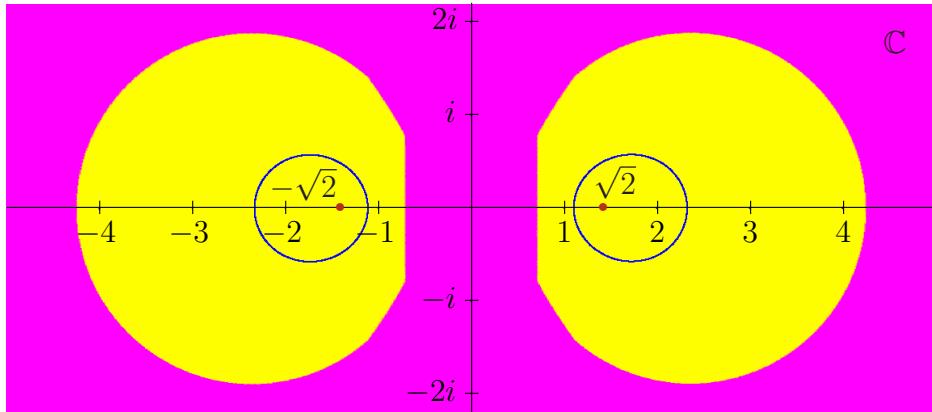
On the positive real line, the interval of quadratic convergence is $(\sqrt{2}/2, 3\sqrt{2})$, while the interval where $\alpha(f, x) < \frac{1}{4}(13 - 3\sqrt{17})$ is $(1.10746, 2.32710)$, which is much smaller.

Euler's method was developed to solve the initial value problem for a first-order ordinary linear differential equation,

$$y' = f(x, y), \quad y(x_0) = y_0,$$

where the function $f(x, y)$ is continuous near (x_0, y_0) in \mathbb{R}^2 . Given a *stepsize* $h > 0$, Euler's method approximates the value of $y(x)$ at $x_1 = x_0 + h$ using the linear approximation given by the differential equation, $y_1 := y_0 + hf(x_0, y_0)$. Euler's method recursively computes a sequence $\{(x_i, y_i) \mid i = 0, \dots, N\}$ of points where

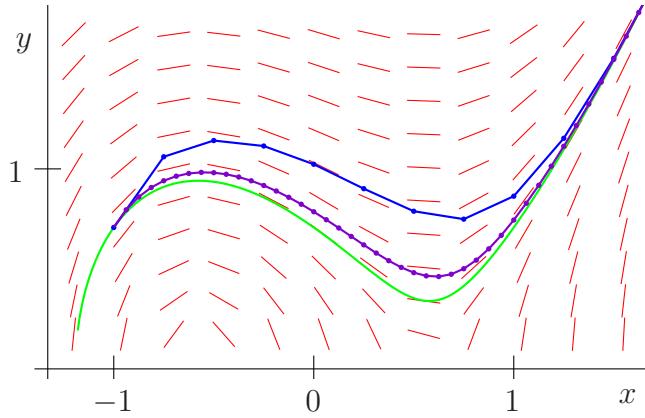
$$x_{i+1} := x_i + h \quad \text{and} \quad y_{i+1} := y_i + hf(x_i, y_i).$$

Figure 4.2: Basins of quadratic convergence for $x^2 - 2$.

Let us consider the initial value problem,

$$y' = \frac{3x^2 - 1}{2y}, \quad y(-1) = \frac{1}{\sqrt{2}}. \quad (4.7)$$

The solution to this initial value problem is $y = \sqrt{x^3 - 3 + \frac{1}{2}}$. The picture below shows the slope field and the solution curve, and two approximations using Euler's method starting at $(-1, \frac{1}{\sqrt{2}})$ with respective stepsizes $h = \frac{1}{4}$ and $h = \frac{1}{16}$.



Like Newton's method, Euler's method extends to solving the intial value problem for a system of first order linear differential equations, so that y is a vector.

Let us investigate the accuracy of Euler's method, assuming that y (and hence $f(x, y)$) has sufficiently many derivatives. The second order Taylor expansion of $y(x)$ at $x = x_0$, together with the differential equation $y'(x) = f(y, x)$ gives

$$y(x+h) = y_0 + hf(x_0, y_0) + \frac{h^2}{2}y''(x_0) + \frac{h^3}{6}y'''(x^*),$$

where x^* lies between x_0 and $x_0 + h$. We estimate

$$|y(x + h) - (y_0 + hf(x_0, y_0))| \leq h^2 \left(\frac{1}{2} |y''(x_0)| + \frac{|h|}{6} |y'''(x^*)| \right).$$

When y''' is bounded near (x_0, y_0) , the error in a single step of Euler's method is at most a constant multiple of h^2 .

To compute $y(x)$ for a fixed x using Euler's method, choose a stepsize h and then perform $\lceil |x - x_0|/h \rceil$ iterations. Each iteration introduces an error at most a constant multiple of h^2 so the difference between $y(x)$ and the computed value will be at most a constant multiple of h . That the global error is at most a constant multiple of the stepsize h , marks Euler's method as a *first-order* iterative method.

The primary value of Euler's method is to illustrate the idea of an iterative solver for initial value problems, as first order accuracy is typically insufficient. A simple higher-order method is the *midpoint rule*, in which the successive values of y are computed using the more complicated formula

$$x_{i+1} = x_i + h, \quad \text{and} \quad y_{i+1} = y_i + hf\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}hf(x_i, y_i)\right),$$

which is second-order.

The classical *Runge-Kutta* method, also called RK4, is a common fourth-order method. For this,

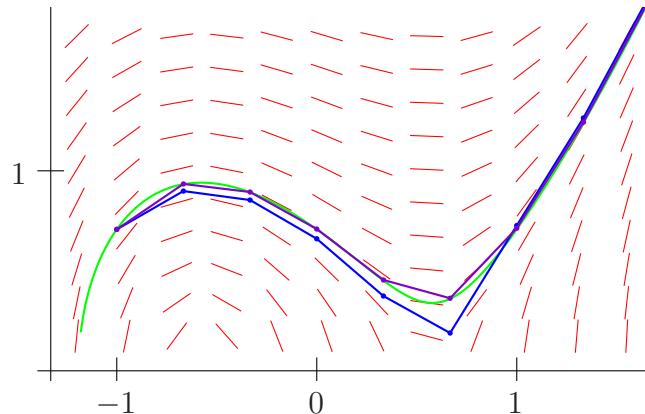
$$x_{i+1} = x_i + h, \quad \text{and} \quad y_{i+1} = y_i + \frac{1}{6}h(z_1 + 2z_2 + 2z_3 + z_4),$$

where

$$\begin{aligned} z_1 &= f(x_i, y_i), & z_2 &= f(x_i + \frac{1}{2}h, y_i + \frac{1}{2}hz_1), \\ z_3 &= f(x_i + \frac{1}{2}h, y_i + \frac{1}{2}hz_2), & \text{and} & z_4 = f(x_i + h, y_i + hz_3). \end{aligned}$$

In Exercise 9 you are asked to relate this to Simpson's rule.

We display the two piecewise linear curves obtained from the midpoint and Runge-Kutta methods with stepsize $\frac{1}{3}$ for the initial value problem (4.7) with solution $y = \sqrt{x^3 - 3 + \frac{1}{2}}$ on the same slope field as we used to illustrate Euler's method.



There is a vast literature on iterative methods for solving ordinary differential equations.

Exercises

1. Consider a depressed cubic equation, one of the form $x^3 + bx = c$. Show that

$$x = \sqrt[3]{\frac{c}{2} + \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}} + \sqrt[3]{\frac{c}{2} - \sqrt{\frac{c^2}{4} + \frac{b^3}{27}}}$$

is a solution. What are the other two solutions? How about the solutions to a general cubic equation $\alpha x^3 + \beta x^2 + \gamma x + \delta = 0$?

2. Prove the assertion about (4.1) using Galois theory. Specifically show that the polynomial f has Galois group the full symmetric group S_6 by factoring f modulo sufficiently many primes, and use lifts of Frobenius elements.
3. Consider the following iterative algorithm used by the Babylonians to compute \sqrt{x} for $x > 0$. Observe that if $x_i > 0$ and $x_i \neq \sqrt{x}$, then the interval with endpoints x_i and x/x_i contains \sqrt{x} in its interior. Set $x_{i+1} = \frac{1}{2}(x_i + \frac{x}{x_i})$ and repeat. Compare this method of computing square roots to Newton's Method.
4. Let $f(x) = x^3 - 2x + 2$ and compute some iterates of Newton's method beginning with the following values of x_0 ,

$$x_0 \in \{0.1, 0.9, \frac{1}{2} + \frac{1}{10}\sqrt{-1}, 1 - \frac{1}{10}\sqrt{-1}, -1\}.$$

5. Newton's method for $f(x) = x^2 - 2$ converges quadratically for x_0 in the interval $[\sqrt{2}/2, 3\sqrt{2}]$. Prove that Newton iterations beginning with these endpoints converge quadratically. Investigate the failure of quadratic convergence for $x > 3\sqrt{2}$: at which step of Newton's method does quadratic convergence fail (the condition (4.4) does not hold) for each of the following starting points for Newton's method.

$$5.2, 4.53, 4.36, 4.298, 4.256, 4.25, 4.246, 4.245, 4.2427.$$

6. Prove that the midpoint rule is a second-order method.
7. Give some examples of Euler's method. If you cannot find something more interesting, start with the exponential function. Have them study its global convergence and the dependence on stepsize.
8. Compare Euler, midpoint, and Runge-Kutta on some examples.
9. Have the students prove the equivalence of Runge-Kutta with Simpson's rule.

4.2 Numerical Homotopy Continuation

The core numerical algorithms introduced in Section 4.1—Newton’s method to refine an approximation to a solution of a system of polynomial equations and iterative methods to solve an initial value problem—are the building blocks of higher-level predictor-corrector methods that track smooth implicitly defined paths. Numerical homotopy continuation uses path tracking to compute the zeroes of a system of polynomials, given the zeroes of a different, but related system. The Bézout Homotopy Algorithm, which is based on numerical homotopy continuation, computes the isolated zeroes of any system of multivariate polynomials and is optimal for generic systems. Algorithms based on numerical homotopy continuation have the added virtue of being inherently parallelizable.

To motivate and illustrate these ideas, consider a toy problem. Suppose we want to compute the (four) solutions to the system of equations

$$z(y - 1) - 1 = y^2 + 2z^2 - 9 = 0. \quad (4.8)$$

Consider instead the system

$$z(y - 1) = y^2 + 2z^2 - 9 = 0, \quad (4.9)$$

whose solutions are found by inspection to be

$$(\pm 3, 0) \quad \text{and} \quad (1, \pm 2). \quad (4.10)$$

Figure 4.3 shows the plane curves defined by the polynomials in these systems. The first system (4.8) seeks the intersection of the hyperbola with the ellipse, while the second, simpler, system (4.9) replaces the hyperbola by the two lines.

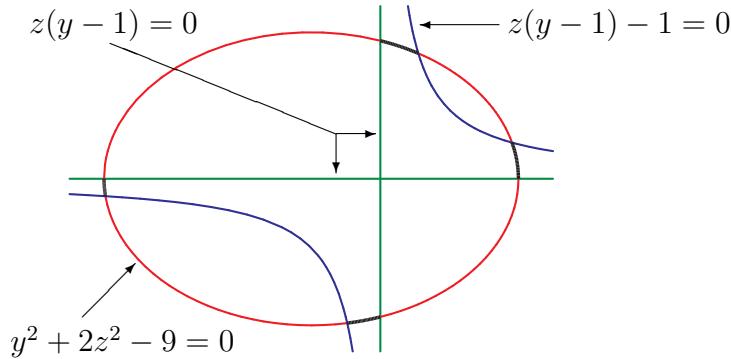


Figure 4.3: The intersection of a hyperbola with an ellipse.

These systems are connected by the one-parameter (in t) family of systems

$$z(y - 1) - (1 - t) = y^2 + 2z^2 - 9 = 0. \quad (4.11)$$

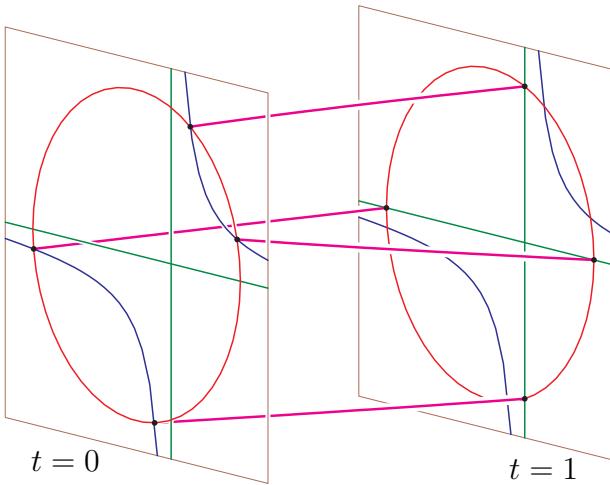


Figure 4.4: Paths connecting solutions.

This defines a space curve C in $\mathbb{C}_{yz}^2 \times \mathbb{C}_t$. Restricting C to $t \in [0, 1]$ gives four paths that connect the known solutions to (4.9) at $t = 1$ to the unknown solutions to (4.8) at $t = 0$. These paths are shown in Figure 4.4. To find the unknown solutions at $t = 0$, we need to track these four paths, starting from the known solutions at $t = 1$.

Let $H(y, z; t)$ be the system of two polynomials in (4.11) that define the space curve C , and let $x(t) := (y(t), z(t))$ for $t \in [0, 1]$ be the projection of one of the paths in Figure 4.4, which is a parametrization of one of the thickened arcs in Figure 4.3.

Then $H(y(t), z(t); t) \equiv 0$ for $t \in [0, 1]$. Differentiating this gives

$$\frac{\partial H}{\partial y} \cdot \frac{dx}{dt} + \frac{\partial H}{\partial z} \cdot \frac{dy}{dt} + \frac{\partial H}{\partial t} = 0.$$

Solving, shows that $x(t)$ satisfies the *Davidenko differential equation*

$$x' = -(D_x H)^{-1} \frac{\partial H}{\partial t}. \quad (4.12)$$

Here, $D_x H$ is the Jacobian matrix of H with respect to its first two (y, z) variables. Thus each of the four paths in Figure 4.4 is a solution of an initial value problem for this differential equation, one for each of the four points (4.10). Using an iterative method to solve these initial value problems computes approximations to the solutions of (4.8).

This is dramatically improved when combined with Newton's method. Fix a sequence of points $1 = t_0 > t_1 > \dots > t_m = 0$ in $[0, 1]$, let z_0 be one of the four solutions (4.10) to the system (4.9) and $x(t)$ the corresponding path. Having computed x_i for some $i < m$, apply one step of any iterative method (Euler, midpoint, RK4, ...) with stepsize $t_{i+1} - t_i$ to get a prediction x_{i+1}^* for $x(t_{i+1})$. Next, perform Newton iterations for $H(x; t_{i+1})$ starting at x_{i+1}^* , until some stopping criterion is reached, obtaining x_{i+1} . If each x_i lies in the basin of attraction of $x(t_i)$ for Newton iterations, then x_m converges to $x(0)$ under

Newton iterations. Better is when the x_i are approximate solutions, for then x_m is an approximate solution to $H(x; 0) = 0$ with associated zero $x(0)$.

This discussion about the system (4.11) is in fact quite general. Let $H(x; t)$ be a system of n polynomials in $n+1$ variables ($x \in \mathbb{C}^n$ and $t \in \mathbb{C}_t$). Then every component of the affine variety $\mathcal{V}(H)$ has dimension at least 1. Let \mathcal{C} be the union of all components of dimension 1 whose projection to \mathbb{C}_t is dense. Then a point $(x; t) \in \mathcal{V}(H)$ is *nondegenerate* if the Jacobian matrix $D_x H$ with respect to its x -variables is invertible at $(x; t)$. A nondegenerate point is isolated from other points of $\mathcal{V}(H)$ in its fiber $\mathbb{C}^n \times \{t\}$ and the nondegenerate points are exactly the points where the Davidenko differential equation (4.12) is defined. You are asked to prove the following lemma in Exercise 2.

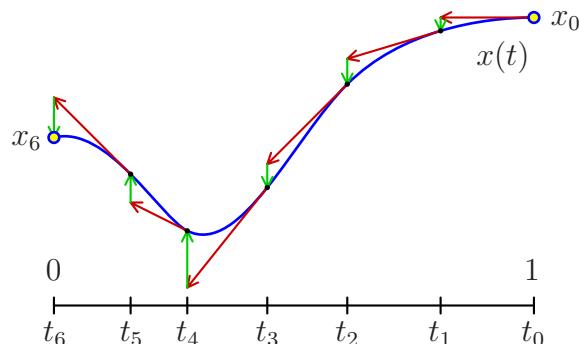
Lemma 4.2.1. *The curve C contains every point $(x; t)$ of $\mathcal{V}(H)$ that is isolated from other points of $\mathcal{V}(H)$ in its fiber. If C is nonempty, then there is a positive integer d and a nonempty Zariski-open subset U of \mathbb{C}_t consisting of all points u such that $H(x; u) = 0$ has d nondegenerate solutions. Above every point b of the complement $B := \mathbb{C}_t \setminus U$ there is some point where either the curve C meets another component of $\mathcal{V}(H)$ or the map $C \rightarrow \mathbb{C}_t$ is ramified or C has a branch tending to infinity near C .*

We call $H(x; t)$ a *homotopy* if C is nonempty and $1 \in U$. Suppose further that

$$\text{the interval } [0, 1] \text{ of } \mathbb{R} \text{ is a subset of } U. \quad (4.13)$$

Then the restriction $\mathcal{C}|_{[0,1]}$ of C to $t \in [0, 1]$ is a collection of d smooth paths. The *start system* is $H(x; 1) = 0$ and the *target system* is $H(x; 0) = 0$, and both have d nondegenerate solutions. Each path in $\mathcal{C}|_{[0,1]}$ connects one nondegenerate solution to the start system to one nondegenerate solution to the target system.

Given a nondegenerate solution x_0 to the start system, let $x(t)$ be the path in $\mathcal{C}|_{[0,1]}$ with $x_0 = x(1)$. It satisfies the Davidenko differential equation (4.12). Choosing a sequence of points $1 = t_0 > t_1 > \dots > t_m = 0$ in $[0, 1]$, a *predictor-corrector method* constructs a sequence x_1, \dots, x_m of approximations to the points $x(t_i)$ along the path alternately applying an iterative method to x_i to get a prediction x_{i+1}^* to $x(t_{i+1})$, which is refined using Newton iterations to get the next point x_{i+1} . Here is a schematic of the predictor-corrector method using Euler predictions to trace a smooth path $x(t)$.



A predictor-corrector method only computes refinable approximations to a sequence of points on an implicitly defined path $x(t)$ for $t \in [0, 1]$. We will largely ignore this distinction and refer to the computed points as lying on the path with $x(1)$ the starting point and $x(0)$ the output, and use the term *path tracking* for this process.

The algorithm of *numerical homotopy continuation* begins with a homotopy $H(x; t)$ satisfying (4.13) and the set of nondegenerate solutions to the start system $H(x; 1) = 0$. By assumption (4.13), each solution x is the starting point $x(1)$ of a smooth path $x(t)$ defined implicitly by $H(x; t) = 0$ for $t \in [0, 1]$. For each solution x to the start system, the algorithm tracks the path $x(t)$ from $t = 1$ to $t = 0$ to obtain $x(0)$, a solution to the target system.

Theorem 4.2.2. *Numerical homotopy continuation under assumption (4.13) computes all nondegenerate solutions to the start system $H(x; 0) = 0$.*

Proof. For each solution x to the start system, the path $x(t)$ is smooth, so that path tracking starting from $x = x(1)$ will compute its endpoint $x(0)$. By assumption (4.13), each nondegenerate solution of the target system is connected to a nondegenerate solution of the start system along one of these paths. This completes the proof. \square

The Bézout homotopy is one of the simplest homotopies. Suppose that $F = (f_1, \dots, f_n)$ is a system of n polynomials in n variables with $\deg f_i = d_i$. By Bézout's Theorem ???, $\mathcal{V}(F)$ has at most $d := d_1 d_2 \cdots d_n$ isolated solutions, and if F is generic, it has exactly d solutions, all nondegenerate. Given the degrees d_1, \dots, d_n , for each $i = 1, \dots, n$, set

$$g_i(x) := x_i^{d_i} - 1. \quad (4.14)$$

Then the system $G = (g_1, \dots, g_n)$ has the d solutions,

$$\{(\zeta_1, \dots, \zeta_n) \mid \zeta_i = e^{\frac{2\pi k}{d_i}\sqrt{-1}} \text{ for } k = 0, \dots, d_i \text{ and } i = 1, \dots, n\}.$$

The *Bézout homotopy* is the convex combination of F and G ,

$$H(x; t) := tG + (1-t)F. \quad (4.15)$$

For any $t \in \mathbb{C}_t$, $H(x; t) = 0$ has at most d isolated solutions. As there are d nondegenerate solutions when $t = 1$, the curve C of Lemma 4.2.1 is nonempty and $1 \in U$. Thus (4.15) is a homotopy with start system G having known solutions and target system F .

Proposition 4.2.3. *If the system F is general, then numerical homotopy continuation using the Bézout homotopy will compute all d solutions to $F = 0$.*

This follows from Theorem 4.2.2, once we see that F general implies that assumption (4.13) holds. While this follows from the discussion below, our goal is to modify the homotopy (4.15) and our path-tracking algorithm to prove the stronger result that the modified Bézout homotopy computes all isolated solutions to the system F .

Let us examine condition (4.13) for the Bézout homotopy. At the endpoints $t = 0, 1$, (4.13) is ensured if the target system is general. To help understand what may happen for $t \in (0, 1)$, consider the Bézout homotopy in one variable

$$H(x; t) := t(x^2 - 1) + (1-t)(x^2 + x + 1) = x^2 + (1-t)x + 1 - 2t. \quad (4.16)$$

The zeroes of $H(x; t) = 0$ as a function of t are found using the quadratic formula

$$x(t) = \frac{t-1}{2} \pm \frac{\sqrt{t^2 + 6t - 3}}{2}.$$

The system $H(x; 2\sqrt{3}-3)$ has a single root $\sqrt{3}-1$ of multiplicity 2. At that point, $\frac{\partial H}{\partial x} = 0$ and so assumption (4.13) fails as $2\sqrt{3}-3 \approx 0.464$ is a branch point in $[0, 1]$.

The Bézout homotopy is a *straight-line homotopy*, which is a convex combination

$$H(x; t) := tG + (1-t)F \quad (4.17)$$

of two polynomial systems that forms a homotopy (defines a curve C with $t = 1$ not a branch point). When both F and G are real as in (4.16), the branch locus likely contains real points that meet the interval $[0, 1]$ even when 0 is not a branch point. A simple modification gives smooth paths above $[0, 1]$. Let γ be any nonzero complex number, and set

$$H_\gamma(x; t) := \gamma t G + (1-t)F. \quad (4.18)$$

The modification (4.18) is called the ‘ γ -trick’.

Theorem 4.2.4. *Let F, G be as above. For any nonzero $\gamma \in \mathbb{C}$, $H_\gamma(x; t)$ is a homotopy with start system G and target system F . When 0 is not a branch point of (4.17), there is a finite set Θ of arguments such that if $\arg(\gamma) \notin \Theta$, then $H_\gamma(x; t)$ satisfies (4.13).*

Proof. For the first statement, substitute $t = 1, 0$ into the formula for $H_\gamma(x; t)$. To understand the modification (4.18) and prove the second statement, note that

$$\frac{\gamma t}{\gamma t + (1-t)}A + \left(1 - \frac{\gamma t}{\gamma t + (1-t)}\right)B = \frac{1}{\gamma t + (1-t)}(\gamma t A + (1-t)B),$$

for indeterminates A, B, t . Consequently, if we define $\tau_\gamma(t) := \gamma t / (\gamma t + (1-t))$ and if $\gamma t + (1-t) \neq 0$ for $t \in [0, 1]$, then for every $t \in [0, 1]$,

$$\gamma t F + (1-t)G \quad \text{and} \quad \tau_\gamma(t)F + (1 - \tau_\gamma(t))G$$

have the same solutions. That is, if $\gamma t + (1-t) \neq 0$ for $t \in [0, 1]$, then the homotopy $H_\gamma(x; t)$ (4.18) for $t \in [0, 1]$ is a straight-line homotopy (4.17), but over the image of $\tau_\gamma: [0, 1] \rightarrow \mathbb{C}$. Solving $\gamma t + (1-t) = 0$ gives $\gamma = 1 - \frac{1}{t}$, so γ cannot be a negative real number, $\arg \gamma \neq \pi$. Identifying \mathbb{C} with \mathbb{R}^2 where $(x, y) \leftrightarrow x + y\sqrt{-1}$ and writing $\gamma = a + b\sqrt{-1}$, the path $\tau_\gamma(t)$ for $t \in [0, 1]$ lies on the circle $x^2 - x + y^2 + \frac{a}{b}y = 0$ with

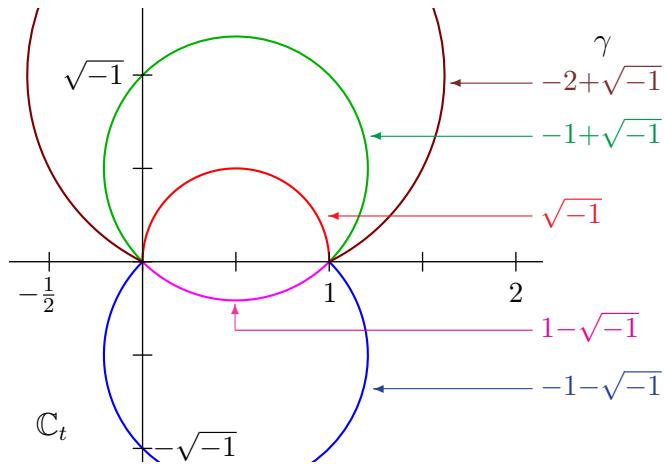


Figure 4.5: Paths τ_γ for $\gamma = -1-\sqrt{-1}, 1-\sqrt{-1}, \sqrt{-1}, -1+\sqrt{-1}, -2+\sqrt{-1}$

center $(\frac{1}{2}, -\frac{a}{2b})$ and radius $\frac{a^2+b^2}{4b^2}$. This circle contains the points 0 and 1, and the path $\tau_\gamma(t)$ for $t \in [0, 1]$ traces the arc of that circle lying in the same half-plane as γ .

The paths defined by $H_\gamma(x; t) = 0$ for $t \in [0, 1]$ are those in the curve C lying above the image $\tau_\gamma[0, 1]$. They will be continuous and satisfy the Davidenko differential equation exactly when $\tau_\gamma[0, 1]$ does not meet the branch locus B . As B is finite, τ_γ depends only upon the argument of γ , and the arcs $\tau_\gamma(0, 1)$ foliate $\mathbb{C}_t \setminus \mathbb{R}$, there are only finitely many arguments for γ such that $\tau_\gamma[0, 1]$ meets B . This completes the proof. \square

For the Bézout homotopy and any other straight-line homotopy, we use the γ -trick for a general $\gamma \in \mathbb{C} \setminus \mathbb{R}$. This γ -trick is a systematic (and easy) way to choose a smooth path in $\mathbb{C}_t \setminus B$ between 0 and 1, but any such path will suffice. For a general homotopy, we will assume that the tracking is done over a general smooth path $\tau \subset \mathbb{C}_t$ between 0 and 1. These assumptions imply that branch points in $\mathbb{C}_t \setminus \{0\}$ may be avoided. This is commonly expressed as “the homotopy defines smooth paths, *with probability one*”. That is, the set of paths between 0 and 1 in \mathbb{C}_t that meet the base locus has measure zero in the collection of all paths considered.

Suppose that we have a homotopy $H(x; t)$ and assume for simplicity that the interval $(0, 1] \subset \mathbb{C}_t$ does not meet the branch locus B . (In general choose a smooth path τ in $(\mathbb{C}_t \setminus B) \cup \{0\}$ between 0 and 1.) Then $C|_{(0,1]}$ is a collection of d half-open smooth paths, and at each point of every path the Jacobian matrix DH_x is invertible. When $0 \notin B$, the homotopy satisfies (4.13) and by Theorem 4.2.2 numerical homotopy continuation computes all solutions to the target system, given the solutions to the start system. When $0 \in B$, by Lemma 4.2.1 there are several cases for a path $x(t)$ in $C|_{(0,1]}$ in the limit as $t \rightarrow 0$.

- (1) The path $x(t)$ does not have a limit as it becomes unbounded as $t \rightarrow 0$.
- (2) The path $x(t)$ has a limit $x(0)$ and $D_x H$ is invertible at $x(0)$.

- (3) The path $x(t)$ has a limit $x(0)$ that lies on another component of $\mathcal{V}(H)$ so that $D_x H$ is not invertible at $x(0)$.
- (4) The path $x(t)$ has a limit $x(0)$ that is a branch point of $C \rightarrow \mathbb{C}_t$, so that $D_x H$ is not invertible at $x(0)$ and at least one other path also ends at $x(0)$.

Figure 4.6 is a schematic showing these four cases.

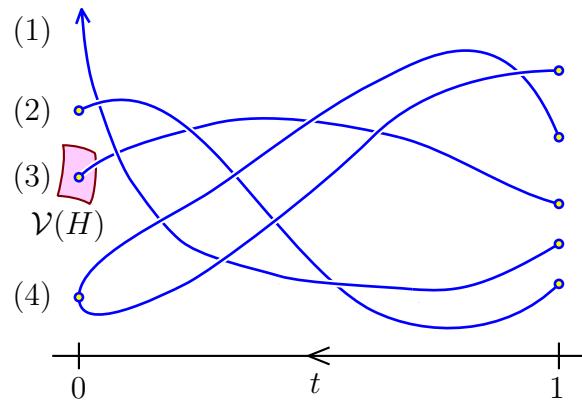


Figure 4.6: Possible behavior of homotopy paths near $t = 0$.

In Case (2), when $D_x H$ is invertible at the endpoint $x(0)$ of the path $x(t)$, this path may be successfully tracked from $x(1)$ to $x(0)$. In all other cases, simple path-tracking will fail, as $D_x H$ is not invertible at $x(0)$. and alternatives to simple path-tracking, called *endgames*, are needed.

Case (1) There are at least two endgames when $x(t)$ becomes unbounded as $t \rightarrow 0$. For one, when $\|x(t)\|$ exceeds a heuristic threshold, tracking is halted and the path is declared to diverge. The other applies, for example, when the homotopy $H(x; t)$ for $x \in \mathbb{C}_x^n$ and $t \in \mathbb{C}_t$ is the restriction of a homotopy $\tilde{H}(z; t)$ for $z \in \mathbb{P}^n$ to an affine patch $\mathbb{C}_x^n \subset \mathbb{P}^n$. Choosing a different affine patch \mathbb{C}_y^n and applying a change of coordinates expresses the homotopy and the computed points on the path $x(t)$ in the coordinates y of \mathbb{C}_y^n . If \mathbb{C}_y^n is chosen propitiously (e.g. at random), then the resulting path $y(t)$ converges in \mathbb{C}_y^n to a point $y(0)$ as $t \rightarrow 0$, and this path falls into one of cases (2), (3), or (4).

Case (3) While geometrically distinct from Case (4), this is treated in the same way as Case (4).

Case (4) The curve C is either singular at $x(0)$ or the map to \mathbb{C}_t is ramified at $x(0)$, or both. Let us examine in detail its geometry near $x(0)$ before describing the Cauchy endgame, which also applies to the other cases (2) and (3).

Let $f: \tilde{C} \rightarrow C$ be the normalization of C , so that \tilde{C} is smooth. The *ramification index* $r = r(x')$ of a preimage $x' \in f^{-1}(x(0))$ is the order of vanishing at x' of the (rational) function t that is the composition $\pi: \tilde{C} \rightarrow C \rightarrow \mathbb{C}_t$, with the second map the projection

onto the t -coordinate. If s is a nonconstant rational function on \tilde{C} that vanishes to order 1 at x' , then $t = s^r g$, for some rational function g with $g(x') \neq 0$.

Suppose that $\Delta \subset \mathbb{C}_t$ is a disc centered at the origin small enough so that 0 is the only branch point in Δ . Then each component of its preimage $\pi^{-1}(\Delta)$ in \tilde{C} contains a unique point $x' \in \pi^{-1}(0)$. On the component Δ' containing x' , the map $\pi: \Delta' \setminus \{x'\} \rightarrow \Delta \setminus \{0\}$ of punctured neighborhoods is an r -fold covering space, analytically isomorphic to the map $s \mapsto s^r$. Over the punctured disc $\Delta \setminus \{0\}$, the two curves \tilde{C} and C agree, and the image in C of a component of $\pi^{-1}(\Delta)$ is a *branch* of C at the corresponding point above 0. Figure 4.7 shows some possibilities near a ramification point. The map $C \rightarrow \mathbb{C}_t$ is the vertical projection and only one vertical real dimension is shown. (The self-intersections,

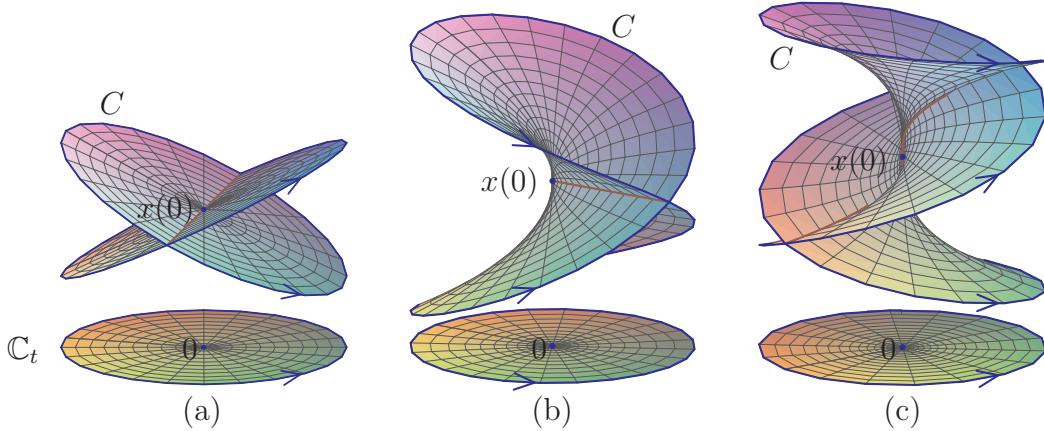


Figure 4.7: Local behaviour near a ramification point.

except at $x(0)$, are artifacts of this.) In (a), C is singular at $x(0)$ with two smooth branches, each with ramification index 1. In (b), C is smooth with one branch at $x(0)$ and ramification index 2. In (c), the ramification index is 3.

The ramification index is also a winding number. Given a point $x(\epsilon) \in C$ for $\epsilon > 0$ in the disc $\Delta \subset \mathbb{C}_t$, analytic continuation on C starting at $x(\epsilon)$ above the circle $\epsilon e^{2\pi\theta\sqrt{-1}}$ for $\theta \geq 0$ gives a closed path $(y(\theta), \epsilon e^{2\pi\theta\sqrt{-1}})$ in C which is parametrized by $\theta \in [0, r]$. The path $(y(\theta), \epsilon e^{2\pi\theta\sqrt{-1}})$ encircling $x(0)$ in C has image in \mathbb{C}_t winding r times around 0. Figure 4.7 shows the circle and these paths. The ramification index may be computed numerically by tracking the path $(y(\theta), \epsilon e^{2\pi\theta\sqrt{-1}})$ for $\theta \geq 0$.

We may compute the endpoint $x(0)$ of the path $x(t)$ in C without tracking the path $x(t)$ to $t = 0$ using Cauchy's integral formula. Recall that if g is a function that is holomorphic in a neighborhood of closed disc D centered at the origin, then

$$g(0) = \frac{1}{2\pi\sqrt{-1}} \oint_{\partial D} \frac{g(z)}{z} dz .$$

This holds also when g is vector-valued. In our case, let D be the unit disc and consider the map $D \rightarrow \Delta$ where $z \mapsto \epsilon z^r =: t$. This lifts to a map $g: D \rightarrow C$ with $g(0) = x(0)$ and

$g(e^{2\pi\alpha\sqrt{-1}}) = (y(r\alpha), e^{2\pi\alpha\sqrt{-1}})$ to which we may apply the Cauchy integral formula. After a change of coordinates, we obtain

$$x(0) = \left(\frac{1}{\sqrt{-1}} \int_0^1 \frac{y(r\alpha)}{e^{2\pi\alpha\sqrt{-1}}} d\alpha, 0 \right). \quad (4.19)$$

The *Cauchy endgame* is a numerical algorithm using this.

Algorithm 4.2.5 (Cauchy Endgame).

INPUT: A homotopy $H(x; t)$, curve C as in Lemma 4.2.1, a path $x(t)$ on C with limit $x(0)$ as $t \rightarrow 0$, and a point $x(\epsilon)$ on the path with $\epsilon > 0$.

OUTPUT: A numerical approximation to $x(0)$ and the ramification index.

1. Starting at $x(\epsilon) = (y(0), \epsilon)$, track the path $y(\theta)$ above the circle $\epsilon e^{2\pi\theta\sqrt{-1}}$ from $\theta = 0$ until the first integer $r > 0$ with $y(r) = y(0)$.
2. Using the computed intermediate values of θ and $y(\theta)$, estimate $x(0)$, using numerical integration for the integral (4.19).
3. Track $x(t)$ from $x(\epsilon)$ to $x(\epsilon/2)$, replace ϵ by $\epsilon/2$, and repeat steps (1) and (2), obtaining another estimate for $x(0)$.

If successive estimates agree up to a tolerance, or do so after it repeating (3), then exit and return the computed value of $x(0)$, along with r .

Remark 4.2.6. The Cauchy endgame applies in each of the cases (1)–(4) above. (In (1), first change coordinates in \mathbb{P}^n so that the path has a finite limit.) Simply stop the tracking of $x(t)$ at some fixed ϵ (e.g. $\epsilon = 0.1$), and then apply the Cauchy endgame. There will be r paths that converge to the endpoint $x(0)$. An additional check verifies that there are indeed $r-1$ other paths with this endpoint.

We deduce a strengthening of Theorem 4.2.2.

Theorem 4.2.2' *Numerical homotopy continuation with endgames computes all isolated solutions to the start system $H(x; 0) = 0$.*

A homotopy is *optimal* if every isolated solution of the target system is connected to a unique solution of the start system along a path of $C|_\tau$, for $\tau \subset \mathbb{C}_t \setminus B$ a path connecting 0 to 1. By Proposition 4.2.3, the Bézout homotopy is optimal, for a general target system F . In a non-optimal homotopy some paths either diverge (Case(1)) or become singular (Cases (3) and (4)), all of which require expensive endgames.

The cost of using numerical homotopy continuation to solve a system of polynomials is dominated by path-tracking and endgames, and is therefore minimized for optimal or near optimal homotopies. A significant advantage is that 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.

Given a specific target system $F = (f_1, \dots, f_n)$ with $\deg F_i = d_i$, using the Bézout homotopy (4.15) to compute its solutions may be problematic as neither the start (4.14) nor the target systems are necessarily generic. In practice the more robust Bézout homotopy algorithm overcomes this by using a two-step process.

Algorithm 4.2.7 (Bézout Homotopy Algorithm).

INPUT: A target system $F = (f_1, \dots, f_n)$ with $\deg f_i = d_i$.

OUTPUT: Numerical approximations to the isolated zeroes of $\mathcal{V}(F)$.

1. Generate a random system $E = (e_1, \dots, e_n)$ of polynomials with $\deg e_i = d_i$.
2. Use the Bézout homotopy (4.15) with start system (4.14) to compute all $d = d_1 \cdots d_n$ solutions to $\mathcal{V}(E)$.
3. Use the straight-line homotopy $tE + (1-t)F$ starting with the solutions to E to compute the isolated solutions to $\mathcal{V}(F)$, possibly employing endgames.

Theorem 4.2.8. *The Bézout homotopy algorithm computes all isolated zeroes of F .*

This is a probability one algorithm, as it requires that $\mathcal{V}(E)$ consist of d isolated solutions, which holds on an open dense set of such systems.

Proof. As E is generic, Proposition 4.2.3 implies that the first homotopy is optimal and it will compute all d solutions to $\mathcal{V}(E)$. For every t , the second homotopy is a system of polynomials in x of degrees d_1, \dots, d_n and so by Bézout's Theorem it has at most d isolated solutions, and when there are d solutions, all are nondegenerate. Since the start system has d solutions, this implies that $C|_{(0,1]}$ consists of d half-open paths beginning at $t = 1$ with $\mathcal{V}(E)$. By our discussion of endgames (and possibly using projective coordinates in Case (1)), all isolated solutions of $\mathcal{V}(F)$ will be found. \square

Example 4.2.9. One source of homotopies are polynomial systems that depend upon parameters. For example, a bivariate polynomial $F_d(x; c)$ of degree d ($x \in \mathbb{C}^2$) has coefficients $c \in \mathbb{C}^{\binom{d+2}{2}}$. Thus a system consisting of a quadratic $F_2(x; a)$ and a cubic $F_3(x; b)$ depends upon $\binom{4}{2} + \binom{5}{2} = 6 + 10 = 16$ parameters. This gives a family

$$\Gamma := \{(x; a, b) \in \mathbb{C}^2 \times \mathbb{C}^6 \times \mathbb{C}^{10} \mid F_2(x; a) = F_3(x; b) = 0\},$$

with a map $\Gamma \rightarrow \mathbb{C}^{16}$ whose fiber over $(a, b) \in \mathbb{C}^{16}$ is the set of common zeroes to $F_2(x; a)$ and $F_3(x; b)$. There is a non-empty Zariski open set $U \subset \mathbb{C}^{16}$ consisting of pairs (a, b) for which the fiber has six solutions and its complement is the branch locus B . We call U the set of regular values of $\Gamma \rightarrow \mathbb{C}^{16}$. For a and b general, this has six solutions by Bézout's Theorem. We obtain a homotopy by parametrizing a line ℓ in the base, $f: \mathbb{C} \rightarrow \ell \subset \mathbb{C}^{16}$ where $f(t) = (a(t), b(t))$ with each component linear, and where $(a(1), b(1))$ gives a system with six solutions. Then $\Gamma|_\ell = f^{-1}(\Gamma)$ is given by the homotopy $H(x; t) := (F_2(x; a(t)), F_3(x; b(t)))$.

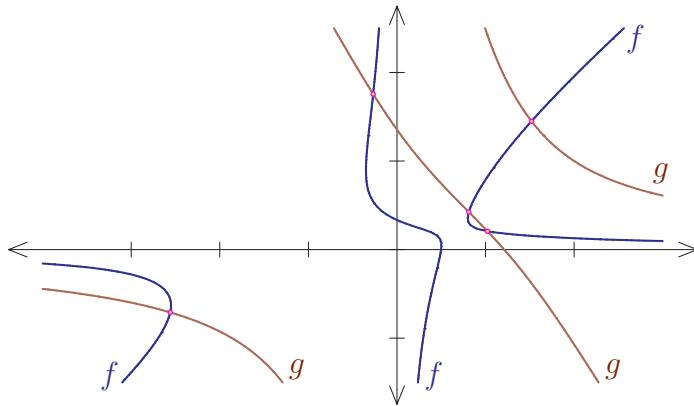
A homotopy arising from such a family where the start and target systems lie in the open set of regular values is a *parameter homotopy*. The Bézout homotopy for a general target system is a parameter homotopy. Nearly every homotopy we use will be a parameter homotopy, often with a propitious choice of line (or a rational curve) in the base.

Polynomial systems arising ‘in nature’ are rarely generic dense systems. The bound from Bézout’s Theorem for the number of isolated solutions is typically not achieved.

For example, consider the system of cubic polynomials,

$$\begin{aligned} f: \quad & 1 - 2x - 3y + 4xy + 5x^2y - 6xy^2 = 0 \\ g: \quad & 23 - 17x - 19y - 13xy + 11x^2y + 7xy^2 = 0 \end{aligned} \quad (4.20)$$

This has the five solutions shown below, and not the 9 predicted by Bézout’s Theorem.



Section ?? describes a method to solve structured systems of equations that are either not square or have fewer solutions than expected from Bézout’s Theorem. Square systems of the form 4.20 which are general given the monomials in each polynomial are called *sparse*, and the polyhedral homotopy, based on ideas from the study of toric varieties, is an optimal homotopy for sparse systems. This will be developed in Section 8.6.

Exercises

1. Find the solutions to the system of equations (4.8) directly, and also by solving the initial value problem for the differential equation 4.12 starting at the solutions (4.10) using any of the iterative methods of Section 4.1.
2. Give a proof of Lemma 4.2.1.
3. Let $\gamma = a + b\sqrt{-1}$ with a, b real and $b \neq 0$. Show that the path in the complex plane $\tau_\gamma(t) := \gamma t / (\gamma t + (1 - t))$ for $t \in [0, 1]$ lies on the circle with centre $(\frac{1}{2}, -\frac{a}{2b})$ and radius $\frac{a^2+b^2}{4b^2}$, which contains the points 0 and 1. Show that the tangent direction of τ_γ at $t = 0$ is γ and that $\tau_\gamma[0, 1]$ lies in the same half-plane as γ .
4. Discuss how to get equations for the branch locus in Example 4.2.9.

5. Explain the ramification of $y^2 = x^3$ in each coordinate projection.
6. Verify the claim in the text that the system (4.20) has exactly five solutions.
7. Needs more exercises.

4.3 Numerical Algebraic Geometry

In Section 4.2 on numerical homotopy continuation, we discussed path-tracking, presented the Bézout homotopy to compute all isolated solutions to a square system of polynomial equations, and mentioned improvements that are covered elsewhere in this text. Numerical algebraic geometry uses this ability to solve systems of polynomial equations to represent and study algebraic varieties on a computer. We first discuss overdetermined and undetermined systems of equations before introducing the notion of a witness set, which is one of the fundamental ideas in numerical algebraic geometry.

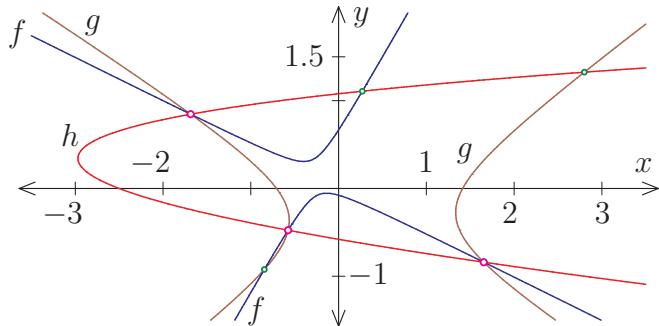
Example 4.3.1. Consider the following problem. For which values of (x, y) does the following matrix of linear polynomials have rank one?

$$M(x, y) := \begin{pmatrix} 4 - 4x + 8y & 3 - 7x - y & 9 - 7x + 8y \\ 6 + 6x + y & 5 + 2x - 5y & 5 + 2x - 8y \end{pmatrix} \quad (4.21)$$

A nonzero 2×3 matrix has rank one if and only if all three of its 2×2 minors vanish. This gives the following system of three polynomial equations,

$$\begin{aligned} f : \quad (4 - 4x + 8y)(5 + 2x - 5y) - (3 - 7x - y)(6 + 6x + y) &= 0 \\ g : \quad (4 - 4x + 8y)(5 + 2x - 8y) - (9 - 7x + 8y)(6 + 6x + y) &= 0 \\ h : \quad (3 - 7x - y)(5 + 2x - 8y) - (9 - 7x + 8y)(5 + 2x - 5y) &= 0 \end{aligned} \quad (4.22)$$

These minors define three curves in the plane



which appear to have three common solutions. We may verify this by computing a lexicographic Gröbner basis with $y < x$ from the minors (4.22),

$$G = \{213y^3 + 100y^2 - 152y - 72, 213y^2 - 136y - 68x - 152\}.$$

The eliminant in y has degree three. Finding its roots, substituting into the second polynomial, and solving for x gives the three solutions

$$(1.654564, -0.8399545), (-0.5754011, -0.4756340), (-1.685073, 0.8461049). \quad (4.23)$$

Each pair of minors vanishes at four points. The fourth point is where the column common to the two minors vanishes. It may be pruned from the other three by evaluating the third minor at all four points and retaining only those where the evaluation is below a predetermined threshold. For example, the minors f and g also vanish at $(-11/13, -12/13)$. Evaluating the third minor h at these four points gives the values $(-4.5 \times 10^{-6}, -1.3 \times 10^{-7}, 9.2 \times 10^{-7}, 45.6)$. The value of h at three of the points is approximately the working precision, so we (correctly) discard the fourth. [†]

As in Example 4.3.1, meaningful systems of equations are not necessarily square; they may have more equations than variables (are *overdetermined*), and yet define a zero-dimensional ideal. There may be no reasonable way to select a square subsystem whose solutions are only those of the original system. When faced with an overdetermined system, one approach is to randomly select a square subsystem ('squaring up' the original system). This square system is solved and polynomials from the original system are used to prune the excess solutions found in the square subsystem.

Algorithm 4.3.2 (Squaring up).

INPUT: An overdetermined system $F: \mathbb{C}^n \rightarrow \mathbb{C}^m$ ($m > n$) of polynomials.

OUTPUT: A square system $G: \mathbb{C}^n \rightarrow \mathbb{C}^n$ whose solutions $\mathcal{V}(G)$ contain the solutions $\mathcal{V}(F)$ of F , with the nondegenerate solutions of F remaining nondegenerate for G . [†]

DO: Select a random linear map $\Lambda: \mathbb{C}^m \rightarrow \mathbb{C}^n$ and return $\textcolor{violet}{G} := \Lambda \circ F$.

Algorithm 4.3.2 is a 'probablility one' algorithm. We give a proof of its correctness.

Theorem 4.3.3. *For any linear map $\Lambda: \mathbb{C}^m \rightarrow \mathbb{C}^n$, we have the containment $(\mathcal{F}) \subset \mathcal{V}(\mathcal{G})$ of solutions in Algorithm 4.3.2. There is a nonempty Zariski open subset U of $n \times m$ complex matrices consisting of linear maps Λ such that the nondegenerate solutions of F remain nondegenerate soutions of $G = \Lambda \circ F$.*

Proof. As Λ is linear, if $F(x) = 0$, then $\Lambda \circ F(x) = 0$, which proves the first statement.

For the second statement, let $x \in \mathcal{V}(F) \subset \mathbb{C}^n$ be nondegenerate. Then the Jacobian matrix $DF(x)$ at x gives an injective linear map from $\mathbb{C}^n = T_x \mathbb{C}^n \rightarrow \mathbb{C}^m$. A point $x \in \mathcal{V}(G)$ is nondegenerate if the composition $DG(x) = \Lambda \circ DF(x): \mathbb{C}^n \rightarrow \mathbb{C}^n$ is injective. Geometrically, this means that the kernel of Λ is transverse to the image of $DF(x)$. This condition defines a nonempty open subset in the space of linear maps. Since $\mathcal{V}(F)$ has finitely many nondegenerate solutions, the set U is the intersection of these nonempty open subsets, one for each nondegenerate solution $x \in \mathcal{V}(F)$. \square

[†]This needs to refer to an algorithm for solving in Chapter 2.

[†]Strengthen this to isolated solutions.

Remark 4.3.4. Choosing a random linear map $\Lambda: \mathbb{C}^m \rightarrow \mathbb{C}^{n-d}$ gives a subsystem $G = \Lambda \circ F$ of F with the following properties: Every component of $\mathcal{V}(G)$ has dimension at least d . These include all irreducible components of $\mathcal{V}(F)$ of dimension d or greater, together with possibly some other components of dimension d .

Example 4.3.5. Let us view Example 4.3.1 in a different light. Suppose that we want a rank one 2×3 matrix M , that is, a solution to the equations

$$M_{1,1}M_{2,2} - M_{1,2}M_{2,1} = M_{1,2}M_{2,3} - M_{1,3}M_{2,2} = M_{1,1}M_{2,3} - M_{1,3}M_{2,1} = 0. \quad (4.24)$$

With three linearly independent equations on a six-dimensional space, this defines a subvariety of either of dimension three or of dimension four. We reduce this to a zero-dimensional problem by slicing the set of solutions to (4.24), adding the (successive) linear equations

$$\begin{aligned} 5M_{1,1} - 2M_{1,3} + 3M_{2,3} &= 17 \\ 40M_{1,2} - 58M_{1,3} - 63M_{2,3} &= -717 \\ 8M_{2,1} + 10M_{1,3} + 11M_{2,3} &= 193 \\ 40M_{2,2} + 6M_{1,3} - 19M_{2,3} &= 159 \end{aligned} \quad (4.25)$$

Only after these four linear equations are added to (4.24) do we obtain a zero-dimensional system with three solutions. This shows that the dimension of the set of rank one 2×3 matrices is four. The local dimension test, which we will describe later, is another way to determine the dimension of a variety, given a point on it and its defining equations.

A system of equations as in (4.24) that defines a positive-dimensional variety V whose points are of interest is an *underdetermined* system. As in Example 4.3.5, adding further equations to reduce its dimension to zero will give a system of polynomials to solve, obtaining points of V . By Bézout's Theorem, this system is expected to have the fewest number of solutions when the additional equations are linear. In this case, the linear equations define a linear subspace L whose codimension equals the dimension of V (L is *complimentary* to V), and the points are the linear section $V \cap L$.

These two techniques of squaring up and slicing down reduce any system of equations to a square system whose solutions may be further processed to obtain solutions to the original problem. When the system is underdetermined and defines a variety $V = \mathcal{V}(F)$, we obtain points of V in the complimentary linear section $V \cap L$. Numerical algebraic geometry uses this to study the algebraic variety V .

Remark 4.3.6. In Examples 4.3.1 and 4.3.5, the variety of rank one 2×3 matrices were sliced by the same linear subspace. The linear equations (4.25) define a four-dimensional linear subspace L of $\text{Mat}_{2 \times 3}\mathbb{C}$, and the family of matrices $M(x, y)$ (4.21) for $x, y \in \mathbb{C}$ is a parametrization of L . You are asked to verify this in Exercise 1.

Any linear subspace of \mathbb{C}^n has an extrinsic description as the vanishing set of some linear equations, and an intrinsic description as the image of a linear map (a parametrization). Either description may be used for slicing. This flexibility may be used to improve the efficiency of an algorithm.

The first question numerical algebraic geometry addresses is how to represent an algebraic variety $V \subset \mathbb{C}^n$ on a computer? In symbolic computation, this is answered by giving a finite set of polynomials $F = \{f_1, \dots, f_m\}$ such that $V = \mathcal{V}(F)$, perhaps with the good algorithmic property of being a Gröbner basis for the ideal of V . In numerical algebraic geometry, if V is zero-dimensional, then the list V of (approximations to) its points, together with the polynomials that define V is a reasonable representation. When the dimension of V is at least one, we will slice V to obtain a collection of points and use these points and the slice as our representation.

Definition 4.3.7. Let $V \subset \mathbb{C}^n$ be an irreducible variety of dimension d . A *witness set* for V is a set W of the form $V \cap L$, where L is a general linear subspace of \mathbb{C}^n complementary to V , so that it has codimension d . The generality of V ensures that the intersection is transverse and by Bézout's Theorem[†], W consists of $\deg V$ points. For computational/algorithmic purposes, we will represent a witness set for V by a triple (W, F, L) . Here, $W = V \cap L$ with V an irreducible component of $\mathcal{V}(F)$ where $F = \{f_1, \dots, f_m\}$ a system of polynomials on \mathbb{C}^n and $L = \{\ell_1, \dots, \ell_d\}$ is d general linear polynomials. (We write L for both the linear subspace and its given equations.)

The same definition makes sense when V is a reducible variety, all of whose components have the same dimension d . While a witness set is certainly a representation of a variety, this definition is justified by its utility. We shall see that a witness sets is the central notion in numerical algebraic geometry and is the input for many of its algorithms. We describe some of the more elementary algorithms that use witness sets. (Needs Examples)

Sampling. A witness set (W, F, L) for a variety $V \subset \mathbb{C}^n$ includes a collection of points of V . If $L' = \{\ell'_1, \dots, \ell'_d\}$ is another collection of d linear polynomials, we may form the straight-line homotopy

$$H(x; t) := (F(x), tL(x) + (1 - t)L'(x)). \quad (4.26)$$

For almost all $t \in \mathbb{C}$, the d linear polynomials $tL(x) + (1 - t)L'(x)$ define a codimension d linear subspace $L_t \subset \mathbb{C}^n$ with $L_1 = L$ and $L_0 = L'$. As $V \cap L$ is transverse, for any point $w \in W = V \cap L$, the homotopy (4.26) defines a path $w(t)$ in V for $t \in (0, 1]$ with $w(1) = w$ and well-defined endpoint $w(0) = \lim_{t \rightarrow 0} w(t)$ (perhaps lying at infinity in V). We may use a witness set as the input for an algorithm to sample points of V .

Algorithm 4.3.8 (Sampling).

INPUT: A witness set (W, F, L) for $V \subset \mathbb{C}^n$.

OUTPUT: Point(s) of V .

Do:

1. Choose d linear polynomials $L' = \{\ell'_1, \dots, \ell'_d\}$ and form the homotopy $H(x; t)$ (4.26).
2. Follow one or more points w of W along the homotopy $H(x; t)$ from $t = 1$ to $t = 0$ and return the endpoints $w(0)$ of the homotopy paths.

[†]Make sure to state it this way in Chapter 3

Proof of correctness. By the generality of L , all homotopy paths $w(t)$ for $w \in W$ are smooth for $t \in (0, 1]$. In particular, each lies in the smooth locus of $\mathcal{V}(F)$. As the initial point $w(1)$ of each path is a point of $W = V \cap L$, the path lies on V , and in particular, its endpoint $w(0)$ is a point of V (possibly singular in $\mathcal{V}(F)$ or at infinity). \square

Moving a witness set. If the linear polynomials $L' = \{\ell'_1, \dots, \ell'_d\}$ in (4.26) are general, then $V \cap L'$ is transverse and consists of $\deg(V)$ points, by Corollary 3.6.1. Thus $W' = (V \cap L', F, L')$ is another witness set for V . This justifies the following algorithm.

Algorithm 4.3.9 (Moving a Witness Set).

INPUT: A witness set (W, F, L) for $V \subset \mathbb{C}^n$.

OUTPUT: A second witness set $(V \cap L', F, L')$ for V .

Do: Choose general linear polynomials $L' = (\ell'_1, \dots, \ell'_d)$ on \mathbb{C}^n . Run Algorithm 4.3.8 on all points $w \in W$, using this choice of L' . Set W' to be the collection of endpoints obtained and output (W, F, L') .

Remark 4.3.10. Note that this algorithm only needs that $W = V \cap L$ is transverse and consists of $\deg(V)$ points, for then (4.26) is a homotopy defining paths that start at points w of W . It is only needed that L' be a general complimentary linear subspace.

Similarly, the Sampling Algorithm 4.3.8 only needs a complimentary linear space L and a point $w \in V \cap L$ where $V \cap L$ is transverse at w to sample points of V .

Membership. Suppose that $x \in \mathbb{C}^n$ is a point of $\mathcal{V}(F)$. We may use a witness set (W, F, L) for a variety V that is a component of $\mathcal{V}(F)$ to determine if $x \in V$.

Algorithm 4.3.11 (Membership Test).

INPUT: A witness set (W, F, L) for $V \subset \mathbb{C}^n$ and a point $x \in \mathcal{V}(F)$.

OUTPUT: True (if $x \in V$) or False (if $x \notin V$).

Do:

1. Choose d linear polynomials $L' = \{\ell'_1, \dots, \ell'_d\}$ that are general given that $\ell'_i(x) = 0$.
2. Call Algorithm 4.3.8 using L' in the homotopy (4.26) and follow homotopy paths from every point $w \in W$.
3. If x is an endpoint of one of these paths, return True, otherwise, return False.

Proof of correctness. By the genericity of L' , the set $W' := V \cap L'$ consists of $\deg(V)$ points, counted with multiplicity. All points of W' , except possibly x (if $x \in W$), are smooth points of V with none at infinity, and all points of W' are endpoints of homotopy paths.[†] Thus $x \in V$ if and only if it is an endpoint of a path given by the homotopy (4.26) that starts at some point of W . \square

[†]There is something to prove here that should have been proved earlier

Inclusion. Suppose that X and V are irreducible subvarieties of \mathbb{C}^n . If $X \not\subset V$, then their set-theoretic difference $X \setminus V$ is open and dense in X . Furthermore, if L is a general linear subspace complimentary to X , then $X \cap L \subset X \setminus V$. This observation leads to the following probability one algorithm to test if $X \subset V$.

Algorithm 4.3.12 (Inclusion).

INPUT: Witness sets (W_X, F_X, L_X) for $X \subset \mathbb{C}^n$ and (W_V, F_V, L_V) for $V \subset \mathbb{C}^n$.

OUTPUT: True (if $X \subset V$) or False (if $X \not\subset V$).

Do:

1. Call the Sampling Algorithm 4.3.8 using the witness set (W_X, F_X, L_X) for X to obtain a point $x \in X \cap L'$, where L' is a general linear subspace complimentary to X .
2. Call the Membership Test Algorithm 4.3.11 to test if $x \in V$.

Proof of correctness. The point $x \in X$ lies in $X \cap L'$, where L' is a general linear subspace complimentary to X . If $X \subset V$, then $x \in V$, and the algorithm returns True. If $X \not\subset V$, then with probability one $(X \cap L') \cap V = \emptyset$, so that $x \notin V$ and the algorithm returns False. \square

The first step in this algorithm is precautionary because L_X may not be sufficiently general to avoid points of $X \cap V$ when $X \not\subset V$.

Witness set of a product. Suppose that $A \subset \mathbb{C}^n$ and $B \subset \mathbb{C}^m$ are irreducible varieties and that (W_A, F_A, L_A) and (W_B, F_B, L_B) are witness sets for A and B , respectively. Here $F_A = F_A(x)$ is a system of polynomials on \mathbb{C}^n and $F_B = F_B(y)$ is a system of polynomials on \mathbb{C}^m . The product $A \times B$ is an irreducible component of the concatenation $F = F(x, y) = (F_A(x), F_B(y))$ of the two systems. Also, the degree of $A \times B$ is the product of the degree of A and the degree of B . Furthermore, $L_A \times L_B$ is a linear subspace of $\mathbb{C}^n \times \mathbb{C}^m$ complimentary to $A \times B$ and $W_A \times W_B = (A \times B) \cap (L_A \times L_B)$ is transverse and consists of $\deg(A) \cdot \deg(B)$ points. While we would like for $(W_A \times W_B, (F_A, F_B), L_A \times L_B)$ to be a witness set for $A \times B$, it is not a witness set as $L_A \times L_B$ is not a general linear subspace of $\mathbb{C}^n \times \mathbb{C}^m$. For example, $L_A \times L_B$ is not in general position with respect to the coordinate projections to \mathbb{C}^n and to \mathbb{C}^m .

Algorithm 4.3.13 (Witness Set of a Product).

INPUT: Witness sets (W_A, F_A, L_A) for $A \subset \mathbb{C}^n$ and (W_B, F_B, L_B) for $B \subset \mathbb{C}^m$.

OUTPUT: A witness set (W, F, L) for the product $A \times B \subset \mathbb{C}^n \times \mathbb{C}^m$.

Do:

1. Set $F := (F_A, F_B)$ and choose general linear forms $L = (\ell_1, \dots, \ell_d)$ on $\mathbb{C}^n \times \mathbb{C}^m$ where $d = \dim(A) + \dim(B)$.
2. Call the Moving Algorithm 4.3.9 with input $(W_A \times W_B, F, (L_A, L_B))$ to move the set $W_A \times W_B$ to the set $W := (A \times B) \cap L$.

Proof of correctness. The collection (L_A, L_B) defines $L_A \times L_B$. We observed that while $W_A \times W_B = (A \times B) \cap (L_A \times L_B)$ is transverse and consists of $\deg(A \times B) = \deg(A) \cdot \deg(B)$ points, it is not a witness set as $L_A \times L_B$ is not a general complimentary linear subspace. By Remark 4.3.10, the Moving Algorithm 4.3.9 only needs its input to be a transverse intersection with a complimentary linear subspace to compute a witness set. This implies that (W, F, L) will be a witness set for the product $A \times B$. \square

Witness sets for projections. Suppose that $X \subset \mathbb{C}^n \times \mathbb{C}^k$ is a variety that is an irreducible component of a system $F = F(x, y)$ of polynomials with $x \in \mathbb{C}^n$ and $y \in \mathbb{C}^k$. Let $\pi: \mathbb{C}^n \times \mathbb{C}^k \rightarrow \mathbb{C}^n$ be the coordinate projection and set $V := \overline{\pi(X)}$, and irreducible subvariety of \mathbb{C}^n . By Lemma 2.1.7, this coordinate projection corresponds to the elimination of the y variables from $F(x, y)$, which may be accomplished by resultants or Gröbner bases. Besides the potential complexity of this computation, there is a very real and practical problem with symbolic elimination: If X' is an irreducible component of $\mathcal{V}(F)$, eliminating y from F gives polynomials that vanish on $\pi(X')$. If it happens that $\overline{\pi(X)} \subsetneq \overline{\pi(X')}$, then we could not study $V = \overline{\pi(X)}$ using an eliminant.

In this setting of a projection, we instead use a variant of a witness set for V . For this, let $M \subset \mathbb{C}^k$ be a general linear subspace complimentary to V . As M is general, $V \cap M$ is transverse and each of its $\deg(V)$ points lies in $\pi(X)$. The intersection $X \cap (\mathbb{C}^n \times M)$ is a collection of $\deg(V)$ fibers of the projection $X \rightarrow V$. Write X_v for the fiber over a point $v \in V \cap M$. The genericity of M implies that these fibers all have the same dimension, $\dim(X) - \dim(V)$, and they all have the same degree.

Let $L \subset \mathbb{C}^n$ be a general linear subspace of codimension $\dim(X) - \dim(V)$. As it is general, L meets each of the fibers X_v for $v \in V \cap M$ transversally in $\deg(X_v)$ points, and we have

$$X \cap (L \times M) = \bigcup \{X_v \cap L \mid v \in V \cap M\}.$$

The quadruple $(X \cap (L \times M), F, L, M)$ is a *pseudowitness set* for $V = \overline{\pi(X)}$. This representation of V does not require knowing polynomials that vanish on V .

A pseudowitness set for the image $\overline{\pi(X)}$ of a projection may be computed from a witness set (W, F, L_X) for X in the same way as Algorithm 4.3.13, but run in reverse. That is, given L and M , we compute $X \cap (L \times M)$ from $X \cap L_X$ using a homotopy between the general linear subspace L_X and the linear subspace $L \times M$.

As the complimentary linear subspace $L \times M$ in a pseudowitness set is not in general position, algorithms that involve manipulating a pseudowitness set for V to obtain a pseudowitness set for another variety $V' = \overline{\pi(X')}$ will have two additional steps (1 and 3 below) as described in the following outline.

1. Use the pseudowitness set $(X \cap (L \times M), F, L, M)$ for $\overline{\pi(X)}$ to compute a witness set (W, F, L_X) for X (using Algorithm 4.3.13).
2. Apply an appropriate construction or algorithm on X using (W, F, L_X) to obtain a witness set (W', F', L'_X) for a variety X' with projection $\overline{\pi(X')} = V'$.

3. Using Algorithm 4.3.13 move the witness set (W', F', L'_X) for X' to a pseudowitness set $(X' \cap (L' \times M'), F', L', M')$ for $V' = \overline{\pi(X')}$.

Studying $V = \overline{\pi(X)}$ using a pseudowitness set is a numerical version of elimination theory.

Local dimension test. Suppose that x is a point lying on a variety V that is an irreducible component of $\mathcal{V}(F)$ where $F = \{f_1, \dots, f_m\}$ is a system of polynomials on \mathbb{C}^n . We would like a numerical method to compute the dimension of V . We discuss one that assumes V is smooth at x . While this does not use a witness set as an input, it has a similar flavor and is needed in subsequent sections.

As explained in Section 3.4, given a point $x \in \mathcal{V}(F)$, the differentials $d_x f_i$ of the polynomials $f_i \in F$ at x define the Zariski tangent space $T_x \mathcal{V}(F)$ of $\mathcal{V}(F)$ at x . Thus $\dim T_x \mathcal{V}(F) = n - \text{rank}(DF(x))$, the corank of the Jacobian matrix of F at x . When x is a smooth point of $\mathcal{V}(F)$, the Zariski tangent space is the ordinary tangent space and its dimension is the dimension of V .

The problem with this calculation is that in practice x is only a numerical approximation to a point of $\mathcal{V}(F)$ and the Jacobian may likely have full rank $\min\{m, n\}$ at x . The notion of numerical rank from numerical analysis suggests a resolution of this problem. We begin with an example.

Example 4.3.14. If we solve the linear equations (4.25) on the set of 2×3 rank one matrices, as in Example 4.3.5, there are three solutions. Here is one,

$$M = \begin{pmatrix} -9.33788 & -7.74194 & -9.30154 \\ 15.0874 & 12.5089 & 15.0288 \end{pmatrix}.$$

This corresponds to the first solution in (4.23), substituted into the matrix $M(x, y)$. The Jacobian matrix of the three 2×2 minors (4.24) is

$$\begin{pmatrix} M_{2,2} & -M_{2,1} & 0 & -M_{1,2} & M_{1,1} & 0 \\ 0 & M_{2,3} & -M_{2,2} & 0 & -M_{1,3} & M_{1,2} \\ M_{2,3} & 0 & -M_{2,1} & -M_{1,3} & 0 & M_{1,1} \end{pmatrix},$$

and evaluating it at the point M gives

$$\begin{pmatrix} 12.5089 & -15.0874 & 0 & 7.74194 & -9.33788 & 0 \\ 0 & 15.0288 & -12.5089 & 0 & 9.30154 & -7.74194 \\ 15.0288 & 0 & -15.0874 & 9.30154 & 0 & -9.33788 \end{pmatrix}. \quad (4.27)$$

This has full rank 3, but the 3×3 -minors are all at most 0.026 in absolute value, so the Jacobian matrix is nearly singular.

Numerical analysis furnishes a method to estimate the rank of such a matrix, by determining a nearby singular matrix. A *singular value decomposition* of a complex $m \times n$ matrix M is a factorization $M = UDV^*$, where U and V are unitary matrices of sizes $m \times m$ and $n \times n$, respectively, and D is a $m \times n$ matrix whose only nonzero entries are

nonnegative real numbers on the diagonal. The diagonal entries of D are the *singular values* of M . The columns of U are orthonormal eigenvectors of MM^* , the columns of V are the orthonormal eigenvectors of M^*M , and the nonzero singular values are the square roots of the non-zero eigenvalues of both M^*M and MM^* . The *numerical rank* of M is the number of singular values that, when divided by the maximum singular value, exceed a pre-determined threshold.

For example, the matrix (4.27) has singular values 29.045, 29.045, 8.35×10^{-5} . As the ratios are 1, 1, and 2.6×10^{-6} with the third about the working precision, we declare its numerical rank to be 2. The singular values for the Jacobian matrices at the other two solutions are 36.12, 36.12, and 3.85×10^{-5} and 15.79, 15.79, and 3.06×10^{-5} , so these likewise have numerical rank 2. Refining the approximate solutions to 12 significant digits does not affect the first two singular values, but the third shrinks to about 10^{-11} , so the ratio is again the working precision.

Algorithm 4.3.15 (Local Dimension Test).

INPUT: A point x on an irreducible component V of $\mathcal{V}(F) \subset \mathbb{C}^n$ and a threshold ϵ .

OUTPUT: An estimate for $\dim(V)$.

Do: Compute the singular value decomposition of the Jacobian $DF(x)$ of F at x . Let σ_{\max} be the maximal singular value and return

$$n = \#\{\sigma \text{ is a singular value of } DF(x) \text{ with } \sigma > \epsilon\sigma_{\max}\}.$$

Explain how this can be used to determine the dimension of an image.

Make an exercise involving 2×4 matrices ?

Exercises

1. Verify the claim in Example 4.3.5 that the system of minors and four linear equations has three solutions. Relate this to Example 4.3.1: Show that the matrix $M(x, y)$ of linear polynomials parametrizes the zero locus of the linear equations (4.25), and that the three rank one matrices obtained in each example are the same.
2. Explain why three linearly independent equations on a six-dimensional space define a subvariety either of dimension three or of dimension four.
3. Verify the claim that in Example 4.3.1 and Example 4.3.5, the variety of rank-one 2×3 matrices was sliced by the same linear subspace.
4. More exercises

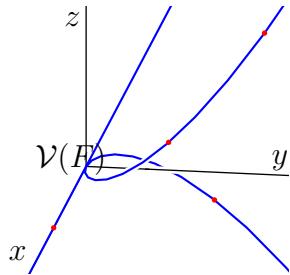
4.4 Numerical Irreducible Decomposition

Section 4.3 introduced witness sets to represent varieties numerically and discussed some algorithms that use witness sets to manipulate varieties. It did not address how to compute a witness set. We offer one method in this section. Here, we explain numerical irreducible decomposition, which begins with a witness set for a (possibly) reducible but equidimensional variety V , decomposing that witness set into witness sets for each irreducible component of V .

Example 4.4.1. Suppose that F is the system of polynomials

$$xy - z = xz - y^2 = 0,$$

in variables x, y, z for \mathbb{C}^3 . Substituting the first into the second gives $x^2y - y^2 = 0$ or $y(x^2 - y) = 0$. If $y = 0$ then $z = 0$ and we see that the x -axis $\mathcal{V}(y, z)$ is a subset of $\mathcal{V}(F)$. If $y \neq 0$, then $y = x^2$ so that $z = x^3$, which shows that the moment curve is also a subset of $\mathcal{V}(F)$. In fact, $\langle F \rangle = \langle y, z \rangle \cap \langle x^2 - y, x^3 - z \rangle$, so that $\mathcal{V}(F)$ is the union of these two curves.



Suppose that we were not able to decompose $\mathcal{V}(F)$ by hand, but only had access to a witness set for it. To be specific, as the two polynomials in F are each irreducible, Exercise 9 of Section 3.2 implies that $\dim(\mathcal{V}(F)) = 1$. If we add the linear equation $x + 2y - 2z = 1$ to F and solve, we obtain the four points

$$(1, 0, 0), (1, 1, 1), \left(\frac{1}{\sqrt{2}}, \frac{1}{2}, \frac{1}{2\sqrt{2}}\right), \left(-\frac{1}{\sqrt{2}}, \frac{1}{2}, -\frac{1}{2\sqrt{2}}\right),$$

which constitute a witness set W for $\mathcal{V}(F)$. A numerical irreducible decomposition of $\mathcal{V}(F)$ is the partition of these points

$$\{(1, 0, 0)\} \sqcup \{(1, 1, 1), \left(\frac{1}{\sqrt{2}}, \frac{1}{2}, \frac{1}{2\sqrt{2}}\right), \left(-\frac{1}{\sqrt{2}}, \frac{1}{2}, -\frac{1}{2\sqrt{2}}\right)\}$$

into two parts with each part being a witness set for one component of $\mathcal{V}(F)$.

Suppose that $V \subset \mathbb{C}^n$ is a (possibly) reducible variety, all of whose irreducible components have the same dimension, and that (W, F, L) is a witness set for V . Let $V = V_1 \cup V_2 \cup \dots \cup V_s$ be the decomposition of V into irreducible components. This induces a partition

$$W = W_1 \sqcup W_2 \sqcup \dots \sqcup W_s \tag{4.28}$$

of the witness set W with each part the witness set of the corresponding component of W , $W_i = V_i \cap L$. We call this partition (4.28) of a witness set W for V a *numerical irreducible decomposition* of V . We will describe methods to compute and verify a numerical irreducible decomposition for V , given a witness set (W, F, L) for V .

Before describing these methods, let us briefly discuss an algorithm to obtain such a witness set. (We will describe more sophisticated methods in the next section.) Let $F = \{f_1, \dots, f_m\}$ be a system of polynomials on \mathbb{C}^n . The variety $\mathcal{V}(F)$ may have many components of different dimensions, and we would like to obtain a witness set for the union V of its components of a given dimension d . The following algorithm furnishes such a method.

Algorithm 4.4.2.

INPUT: A system $F = \{f_1, \dots, f_m\}$ of polynomials on \mathbb{C}^n and a positive integer $d < n$.

OUTPUT: A witness set for the union V of components of $\mathcal{V}(F)$ of dimension d .

Do:

1. Select a random subsystem F' of F consisting of $n-d$ polynomials. This uses a variant of Algorithm 4.3.2.
2. Choose d general linear polynomials, L .
3. Use the Bézout Homotopy Algorithm 4.2.7 (or any other method) to compute all isolated solutions W' to the square system $\mathcal{V}(F', L)$.
4. Let $W \subset W'$ be those points of W' that lie on $\mathcal{V}(F)$ and have local dimension d in $\mathcal{V}(F)$. Return (W, F, L) .

Proof of correctness. Rewrite this As F' consists of $n-d$ polynomials, the components of $\mathcal{V}(F')$ all have dimension d or more. As this is a subsystem of F , $\mathcal{V}(F) \subset \mathcal{V}(F')$, and as it is a random subsystem, with probability one, the components of $\mathcal{V}(F')$ of dimension more than d are components of $\mathcal{V}(F)$ Need a proof of this. Thus every d -dimensional component of $\mathcal{V}(F)$ is a component of $\mathcal{V}(F')$.

The endpoints of the homotopy paths in Step (3) will all be points of $'calV(F') \cap L$, and will include all isolated points. As L is general, it will meet the union of all d -dimensional components of $\mathcal{V}(F')$ transversally in the set of isolated points. Those points that lie in $\mathcal{V}(F)$ and for which $\mathcal{V}(F)$ has local dimension d will thus be the points of a witness set $W = V \cap L$ of the union V of the d -dimensional components of $\mathcal{V}(F)$. \square

- **Monodromy** Explain how to apply to the example. Find a single loop that permutes the three points.

Lemma Homotopy paths remain on the same irreducible component.

Theorem Monodromy on an irreducible component = full symmetric group.

→ The idea is that connectivity of smooth points implies that monodromy is transitive, and the existence of a simple tangent (reduce to plane curves) implies a simple transposition.

Discuss the coarsening algorithm using monodromy, but note that it lacks a stopping criterion when V has two or more components.

- **Trace Test**

Explain the need, apply to the example, and then perhaps to the example from Frank's paper with Anton and Jose.

Prove that trace is linear on a witness set, and it is not linear if the witness set is incomplete.

Discuss how there is currently no way to certify linearity.

Give the numerical irreducible decomposition algorithm, together with a proof of its correctness.

Perhaps give a meatier example ?

- Witness set of an intersection.

Exercises

4.5 Smale's α -theory

There are some notes on this from Frank's last Math 648 class

Need to get Smale's book where this is done carefully.

There is a paper improving Smale's threshold for α

Exercises

4.6 Notes

mention formulas for zeroes of univariate polynomials ? e.g. hypergeometric? Need to mention Davidenko, as well as possible the origins of both Newton and Euler's methods. Do explain that numerical homotopy continuation originated outside of mathematics.

Cut Material. May need to explain parameter homotopies somewhere

Another source of optimal homotopies are *Parameter homotopies* [82],

Parameter homotopies provide a method to solve such a system. We illustrate this with this example. Each polynomial in (4.20) has six monomials, and we may identify the space of polynomial systems consisting of two such polynomials (f, g) with \mathbb{C}^{12} ,

$$\begin{aligned} f &:= f_1 + f_2x + f_3y + f_4xy + f_5x^2y + f_6xy^2, \\ g &:= g_1 + g_2x + g_3y + g_4xy + g_5x^2y + g_6xy^2. \end{aligned}$$

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.

Suppose that we have a system $G := (f^*, g^*)$ in this family whose solutions are known. Given any other system $F = (f, g)$, the straight-line parameter homotopy

$$H(x, t) := tG + (1 - t)F$$

allows us to use the solutions to G to find the isolated solutions to F , as in ????

Chapter 5

Real algebraic and semialgebraic geometry

In the first chapters of the book, we have mostly worked over an algebraically closed fields, such as the complex numbers. Many applications of algebraic geometry deal – at least partially – with real solutions of polynomial equations or of polynomial inequalities. Since the field \mathbb{R} is not algebraically closed, this often poses additional difficulties. To illustrate this situation, let us point out that for univariate complex polynomials, it is often enough to simply know about the Fundamental Theorem of Algebra, which tells us that a non-zero polynomial of degree n always has n roots, counting multiplicity. Over the real numbers, this is drastically different, and makes it worth and necessary to start the chapter with a treatment of univariate polynomials. For example, we discuss methods to count the number of real solutions of a given univariate polynomial.

Some of the basic results presented in this chapter have a long and distinguished history. Indeed, the fact that many facets of real algebraic geometry were quite developed already in the 19th century reflect the situation that algebraic problems over the real numbers often occur in a very natural way, such as in application in the sciences.

In the chapter, we deal with foundational material of real algebraic and semialgebraic geometry, as well as some more recent developments, such as spectrahedra or hyperbolic polynomials.

5.1 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 sequence of polynomials of decreasing degree, defined by

$$p_0 = p, \quad p_1 = p', \quad p_i = -\text{rem}(p_{i-2}, p_{i-1}) \text{ for } i \geq 2,$$

where p' is the derivative of p and rem denotes the remainder of a division with remainder. That is, starting from the polynomials p and p' , the Sturm sequence of x is obtained by

negating the remainder obtained in the Euclidean algorithm. Let p_m be the last non-zero polynomial in the Sturm sequence. As a consequence of the Euclidean algorithm, we observe that p_m is a greatest common divisor of p_0 and p_1 .

Theorem 5.1.1 (Sturm). *Let $p \in \mathbb{R}[x]$ and $\sigma(x)$ denote the number of sign changes in the sequence*

$$p_0(x), p_1(x), p_2(x), \dots, p_m(x). \quad (5.1)$$

Given $a < b$ with $p(a), p(b) \neq 0$, the number of distinct real zeroes of p in the interval $[a, b]$ equals $\sigma(a) - \sigma(b)$.

Here, any zeroes are ignored when counting the number of sign changes in a sequence of real numbers. For example, the sequence $+0+0-+0$ has two sign changes. 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.

We initiate the proof of Sturm's Theorem with the following observation.

Lemma 5.1.2. *If p does not have multiple roots, then for any $x \in \mathbb{R}$, the sequence $p_0(x), p_1(x), \dots, p_m(x)$ cannot have two consecutive zeroes.*

Proof. By our assumption on the multiplicities, p_0 and p_1 cannot simultaneously vanish at x . Moreover, inductively, if p_i and p_{i+1} both vanish at x then the division with remainder

$$p_{i-1} = s_i p_i - p_{i+1} \quad \text{with some } s_i \in \mathbb{R}[x]$$

implies $p_{i-1}(x) = 0$ as well, contradicting the induction hypothesis. \square

Proof of Sturm's Theorem 5.1.1. First we consider the situation of only single roots. In this situation, p and p' do not have a non-constant common factor, and thus, by definition of m , the polynomial p_m is a non-zero constant.

Now imagine a left to right sweep on the real number line. By continuity of polynomial functions, it suffices to show that $\sigma(x)$ decreases by 1 for a root of p and stays constant for a root of p_i , $1 \leq i < m$. Note that several p_i can become zero at the same x , but in this case Lemma 5.1.2 will allow to consider these events separately.

Moving through a root x of p : Before reaching x , the numbers $p_0(x)$ and $p_1(x)$ have different signs, whereas after x , they have identical signs. Hence $\sigma(x)$ decreases by 1.

Moving through an x with $p_i(x) = 0$ for some $1 \leq i < m$: By Lemma 5.1.2, the values $p_{i-1}(x)$ and $p_{i+1}(x)$ are non-zero, and they have opposite signs by definition of p_{i+1} . Hence, the sequence $p_{i-1}(x), \varepsilon, p_{i+1}(x)$ has two sign changes both in the case of $\varepsilon > 0$ and in the case of $\varepsilon < 0$.

In the case of a multiple root x_0 of multiplicity t , the polynomials p_0 and p_1 have a common factor $(x - x_0)^{t-1}$. Then $p_i(x_0) = 0$ for $i \in \{0, \dots, m\}$, and p_m is a greatest common divisor of p_0 and p_1 . Now define the sequence of polynomials $q_i = p_i/p_m$, $0 \leq i \leq m$, in particular q_0 has only simple roots and $q_m = 1$. Note that the sequence q_0, q_1, \dots, q_m

is not a Sturm sequence in the sense of our definition, since $q'_0 \neq q_1$. However, the recursion formula $q_{i-1} = s_i q_i - q_{i+1}$, $1 \leq i \leq m-1$, still holds, so we approach the sequence as if it were a Sturm sequence.

For any $x \in \mathbb{R}$, the sequence $q_0(x), q_1(x), \dots, q_m(x)$ does not have two consecutive zeroes and the values $q_i(x)$ differ from the values $p_i(x)$ just by the factor $p_m(x)$. Hence, outside a multiple zero of p_0 , the number of sign changes in the subsequence $q_1(x), \dots, q_m(x)$ coincides with the number of sign changes in the subsequence $p_1(x), \dots, p_m(x)$. Now it remains to show, that by moving through a multiple zero x_0 of multiplicity t , the number of sign changes of the subsequence $q_0(x), q_1(x)$ decreases by 1. q_0 and q_1 have a common factor $(x - x_0)^{t-1}$. If $t-1$ is even, then $p_m(x)$ has the same sign just before and just after sweeping x_0 , so that the result follows from the earlier consideration of single roots. If $t-1$ is odd, then locally before sweeping x_0 , the values p and p' have different signs, and afterwards both p and p' have the sign which p had before. Hence, since $t-1$ is odd, both q and q' have the opposite sign of the one that q had before. This proves the claim. \square

In order to count all real roots of a polynomial p , 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.

Corollary 5.1.3. *The number of distinct real roots of a polynomial $p \in \mathbb{R}[x]$ is $\sigma_{-\infty} - \sigma_\infty$, where σ_∞ is the number of sign changes between the leading coefficients of the elements of the Sturm sequence and $\sigma_{-\infty}$ is the same, but with the leading coefficients multiplied by -1 whenever the degree is odd.*

Example 5.1.4. The polynomial $p = x^4 - 3x^3 + 3x^2 - 3x + 2$ has the derivative $p' = 4x^3 - 9x^2 + 6x - 3$, which gives the the Sturm sequence

$$p, \quad p', \quad \frac{3}{16}x^2 + \frac{9}{8}x - \frac{23}{16}, \quad -\frac{704}{3}x + 256, \quad -\frac{25}{1936}.$$

Hence, σ_∞ is the number of sign changes in the sequence $1, 1, 1, -1, -1$ and $\sigma_{-\infty}$ is the number of sign changes in the sequence $1, -1, 1, 1, -1$. Since $\sigma_{-\infty} - \sigma_\infty = 3 - 1 = 2$, the polynomial p has exactly two real roots.

Using bisection, Sturm sequences can be transformed into a procedure for isolating the real roots by rational intervals.

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 and $\alpha_1, \dots, \alpha_n$ be the roots of p . In order to study the number of real roots, the number of positive real roots and even more general situations in a unified setting, let $h \in \mathbb{R}[x]$ be a fixed polynomial, with the goal to consider then roots α_j of p which satisfy a sign condition on $h(\alpha_j)$.

Let q_h be the quadratic form

$$q_h(x) = q_h(x_1, \dots, x_n) = h(\alpha_1)y_1^2 + \dots + h(\alpha_n)y_n^2, \quad (5.2)$$

where $y_j = x_1 + \alpha_j x_2 + \dots + \alpha_j^{n-1} x_n$. The coefficients of $q_h(x)$ are symmetric functions in the roots and thus, by the following lemma, they are real numbers.

Lemma 5.1.5. *Let $p \in \mathbb{R}[x]$ of degree n and $\alpha_1, \dots, \alpha_n$ be the roots of p . For any real polynomial $f = \sum_{j=0}^m c_j x^j$, the expression $\sum_{k=1}^n f(\alpha_k)$ is a real number.*

Proof. We can assume that p is monic. Since for any matrix $A \in \mathbb{C}^{n \times n}$ with eigenvalues $\lambda_1, \dots, \lambda_n$, we have $\text{Tr}(A) = \sum_{k=1}^n \lambda_k$ and $\lambda_i(A^j) = \lambda_i(A)^j$, this shows $\sum_{k=1}^n \alpha_k^j = \text{Tr}(C_p^j)$, where C_p is the companion matrix of p from (2.18). Hence,

$$\sum_{k=1}^n f(\alpha_k) = \sum_{k=1}^n \sum_{j=0}^m c_j \alpha_k^j = \sum_{j=0}^m c_j \text{Tr}(C_p^j) = \text{Tr}\left(\sum_{j=0}^m c_j C_p^j\right) = \text{Tr}(f(C_p)),$$

where $f(C_p)$ denotes the component-wise application of f on the entries of C_p . Since C_p has real entries, the statement follows. \square

Recall that for a real quadratic form $q : \mathbb{R}^n \rightarrow \mathbb{R}$, the *signature* $\sigma(q)$ is the number of positive eigenvalues minus the number of negative eigenvalues of its representing matrix. The *rank* $\rho(q)$ is the rank of the representing matrix.

Theorem 5.1.6. *The rank of $q_h(p)$ equals the number of distinct roots α_j of p for which $h(\alpha_j) \neq 0$. The signature of $q_h(p)$ equals the number of distinct real roots α_j of p for which $h(\alpha_j) > 0$ minus the number of real roots α_j of p for which $h(\alpha_j) < 0$.*

Proof. To keep notation simple, we first consider the case of single roots. In this case, we can view the definition of y_j as a variable transformation and can consider q_h as a quadratic form $q_h(y)$ in y_1, \dots, y_n . With any real zero α_k , we associate the term $h(\alpha_k)y_k^2$, where we note that its sign is given by $h(\alpha_k)$. With any non-real conjugate pair $\{\alpha_k, \alpha_l\}$, we associate the terms

$$(q_h)_{\{k,l\}} = h(\alpha_k)y_k^2 + h(\alpha_l)y_l^2$$

and apply a variable transformation. Namely, let $y_k = u + iv$ with $u, v \in \mathbb{R}$. Then $y_l = u - iv$, which gives us the real representation

$$(q_h)_{\{k,l\}} = 2 \begin{pmatrix} u \\ v \end{pmatrix}^T \begin{pmatrix} \text{Re}(h(\alpha_k)) & -\text{Im}(h(\alpha_k)) \\ -\text{Im}(h(\alpha_k)) & -\text{Re}(h(\alpha_k)) \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}.$$

Since the trace of the 2×2 -block is zero, we see that the signature of $(q_h)_{\{k,l\}}$ is zero. And its rank is two unless $h(\alpha_k) \neq 0$. Building together the local view on the terms of the quadratic form q_h , Sylvester's Theorem then implies that the rank and the signature of the quadratic form are invariant under variable transformations. This gives the desired result.

For the case of roots with arbitrary multiplicities, denote by β_1, \dots, β_s the distinct roots with multiplicity $\mu(\beta_j)$. The quadratic form q_h then becomes a quadratic form in the variables y_1, \dots, y_s ,

$$q_h(y_1, \dots, y_s) = \sum_{j=1}^s \mu(\beta_j) h(\beta_j) y_j^2,$$

from which the statement follows similarly. \square

In particular, for counting the number of roots choose $h(x) = 1$ and for counting the number of positive roots choose $h(x) = x$. We obtain.

Corollary 5.1.7. *The number of distinct real roots of p equals the signature of the [Hankel matrix](#)*

$$H_1(p) = \begin{pmatrix} n & s_1 & \cdots & s_{n-1} \\ s_1 & s_2 & \cdots & s_n \\ \vdots & \vdots & \ddots & \vdots \\ s_{n-1} & s_n & \cdots & s_{2n-2} \end{pmatrix}, \quad (5.3)$$

where $s_k = \sum_{i=1}^n \alpha_i^k$ is the k -th [Newton sum](#) of p . The signature of the matrix

$$H_x(p) = \begin{pmatrix} s_1 & s_2 & \cdots & s_n \\ s_2 & s_3 & \cdots & s_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ s_n & s_{n+1} & \cdots & s_{2n-1} \end{pmatrix},$$

equals the number of distinct positive roots of p minus the number of distinct negative roots of p .

The Newton sums can be expressed as polynomials in the coefficients a_i of $p = x^n + \sum_{i=0}^{n-1} a_i x^i$. Namely, the s_i and the a_j are related by [Newton's identities](#)

$$s_k + a_{n-1}s_{k-1} + \cdots + a_0s_{k-n} = 0 \quad (k \geq n), \quad (5.4)$$

$$s_k + a_{n-1}s_{k-1} + \cdots + a_{n-k+1}s_1 = -ka_{n-k} \quad (1 \leq k < n), \quad (5.5)$$

see Exercise 2.

Proof. For $h(x) = 1$, the coefficient of x_j^2 in the quadratic form (5.2) is $\sum_{l=1}^n \alpha_l^{j-1} = s_{j-1}$, and the coefficient of $x_j x_k$, $j < k$ is $2 \sum_{l=1}^n \alpha_l^{j+k-2} = 2s_{j+k-2}$.

This also implies the statement for the case of the number of distinct positive roots minus the number of distinct negative roots. \square

In particular, we obtain:

Corollary 5.1.8. *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. All zeroes are distinct and positive if and only if its associated matrix $H_x(p)$ is positive definite.*

Remark 5.1.9. To prepare for the multivariate version of the Hermite form in Section 5.3, we remark that the quadratic form q_h can also be regarded as a natural quadratic form in the quotient ring $\mathbb{R}[x]/\langle p \rangle$. Namely, for $p = \sum_{j=0}^n a_j x^j \in \mathbb{R}[x]$ and $h \in \mathbb{R}[x]$, q_h can be viewed as the quadratic form

$$\begin{aligned} \mathbb{R}[x]/\langle p \rangle \times \mathbb{R}[x]/\langle p \rangle &\rightarrow \mathbb{R}, \\ (f, g) &\mapsto \text{Tr}(m_{fg}) , \end{aligned}$$

where m_{fgh} denotes the multiplication with fgh in the residue class ring $\mathbb{R}[x]/\langle p \rangle$. This will become apparent from the later discussion of the multivariate setting, see Corollary 5.3.2.

We consider another classical result:

Theorem 5.1.10. (*Descartes' Rule of Signs.*) *The number of positive real roots of a polynomial, counting multiplicities, is at most the number of sign changes in its coefficient sequence.*

Proof. Given $p \in \mathbb{R}[x]$ of degree n , let $n_+(p)$ be the number of positive real roots of p counting multiplicities, and $\sigma(p)$ be the number of sign changes in the coefficient sequence of p .

We show $n_+(p) \leq \sigma(p)$ by induction, where the case $n = 1$ is clear. Now let $p \in \mathbb{R}[x]$ be of degree $n > 1$. We may assume that x does not divide p and is thus of the form

$$p = \sum_{i=k}^n a_i x^i + a_0 \quad \text{with some } k \in \{1, \dots, n\}$$

and $a_n, a_k, a_0 \neq 0$. Since $p' = \sum_{i=k}^n a_i i x^{i-1}$, except for a possible sign change between a_k and a_0 the number of sign changes of p coincides with $\sigma(p')$. The induction hypothesis implies $n_+(p') \leq \sigma(p')$. 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 .

If a_0 and a_k have the same sign, then p cannot have a root in $(0, x_0)$. And since between any two zeroes of p there must be a zero of p' , we obtain

$$n_+(p) \leq n_+(p') \leq \sigma(p') = \sigma(p).$$

Note that this argument also holds true also in the case of multiple roots of p .

Similarly, if a_0 and a_k have opposite signs, then the number of zeroes of p is at most one larger than the number of zeroes of p' , which then shows

$$n_+(p) \leq n_+(p') + 1 \leq \sigma(p') + 1 = \sigma(p).$$

□

By replacing x by $-x$ in Descartes' 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 5.3.4). And we have the following corollary to Descartes' Rule.

Corollary 5.1.11. *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.

Exercises

1. Use a Sturm sequence to analyze the number of real roots of a quadratic polynomial $p = ax^2 + bx + c$ in terms of the real coefficients a, b, c . Why is your result expected?
2. Prove the Newton identities (5.4) and (5.5) based on the following steps.
 - (a) Setting $a_i = 0$ for $i < 0$, it suffices to prove that $P_n^{(k)}(\alpha_1, \dots, \alpha_n) := s_k + \sum_{i=n-k+1}^{n-1} a_i s_{k-n+i} + k a_{n-k} = 0$ for $k \geq 1$, where $\alpha_1, \dots, \alpha_n$ are the roots of p . First show that $P_n^{(k)} = \sum_{j=1}^n \alpha_j^{k-n} p(\alpha_j) = 0$ for any $k \geq n$.
 - (b) For the induction step in an induction over $n - k$, now assume that $n - k \geq 1$, and verify the polynomial identity $P_n^{(k)}(\alpha_1, \dots, \alpha_{n-1}, 0) = P_{n-1}^{(k)}(\alpha_1, \dots, \alpha_{n-1})$.
 - (c) The polynomial $P_n^{(k)} = P_n^{(k)}(\alpha_1, \dots, \alpha_n)$ is divisible by α_n and by symmetry also by $\alpha_2, \dots, \alpha_n$, hence by $\alpha_1 \cdots \alpha_n$. Since $k < n$, the homogeneous polynomial $P_n^{(k)}(\alpha_1, \dots, \alpha_n)$ of degree k must be the zero polynomial.
3. Show that the number of positive roots of a polynomial $p \in \mathbb{R}[x]$ equals the number of sign changes of its coefficient sequence modulo 2.
4. If all roots of a polynomial $p = \sum_{j=0}^n a_j x^j \in \mathbb{R}[x]$ are real, then the coefficients satisfy the concavity condition

$$a_j^2 - \frac{n-j+1}{n-j} \frac{j+1}{j} a_{j-1} a_{j+1} \geq 0, \quad 1 \leq j \leq n-1.$$

5. For $p \in \mathbb{C}[x]$ with roots $\alpha_1, \dots, \alpha_n$, the determinant of the Hankel matrix (5.3) equals the discriminant $\text{disc}(p) = \prod_{1 \leq i < j \leq n} (\alpha_i - \alpha_j)^2$.

5.2 Univariate stable polynomials

The notion of stability can be seen as a generalization of the concept of real rootedness of a real polynomial. It has been and is an important concept in applications. There are various related definitions of stability, and we exhibit some of the relations among them in this section. While stability is defined also for complex polynomials, we see that the notion is strongly tied to real aspects. Recall that $\Re(z)$ and $\Im(z)$ denote the real and the imaginary part of a complex number z .

A polynomial $f \in \mathbb{C}[x]$ is called *stable* if the open half-plane

$$\mathcal{H} = \{z \in \mathbb{C} : \Im(z) > 0\}$$

does not contain a root of f . If a polynomial is real and stable, then we call it *real stable*. Using these definitions, a polynomial $f \in \mathbb{R}[x]$ is real stable if and only if it has only real roots.

Let f and g be real stable, with roots a_1, \dots, a_k of f and b_1, \dots, b_l of g . The roots of f and g are called *interlaced* if

$$a_1 \leq b_1 \leq a_2 \leq b_2 \leq \cdots \quad \text{or} \quad b_1 \leq a_1 \leq b_2 \leq a_2 \leq \cdots$$

If g is a real stable polynomial with simple roots $b_1 < \dots < b_l$, set $\hat{g}_j = \frac{g}{x-b_j}$, $1 \leq j \leq l$, and observe that for $t \neq j$, the polynomials \hat{g}_t vanish at b_j . Hence, for a polynomial f with $\deg(f) \leq \deg(g)$, there exist unique $\alpha, \beta_1, \dots, \beta_l$ with $f = \alpha g + \sum_{j=1}^l \beta_j \hat{g}_j$.

Lemma 5.2.1. *If $f, g \in \mathbb{R}[x]$ are real stable with $\deg(f) \leq \deg(g)$ and fg has only simple roots, then the following statements are equivalent:*

1. *The roots of f and g are interlaced.*
2. *In the representation $f = \alpha g + \sum_{j=1}^l \beta_j \hat{g}_j$, the coefficients β_1, \dots, β_l are nonzero and have all the same sign.*

Proof. If the roots of f and g are interlaced, then the sequence $f(b_1), \dots, f(b_l)$ strictly alternates in sign. Since $f(b_j) = \beta_j \hat{g}_j(b_j)$ for $1 \leq j \leq l$, the coefficients β_j all have the same sign. And the same arguments apply to the converse direction. \square

For $f, g \in \mathbb{R}[x]$ with interlaced roots, the *Wronskian*

$$W_{f,g} = f'g - g'f \tag{5.6}$$

is either nonnegative for all $x \in \mathbb{R}$ or nonpositive for all $x \in \mathbb{R}$. In order to see this, we can assume that fg has only simple roots, the general case then follows, since f and g can be approximated arbitrarily close by polynomials whose product has only simple roots. Assuming without loss of generality $\deg(f) \leq \deg(g)$ and using the representation $f = \alpha g + \sum_{i=1}^l \beta_i \hat{g}_i$ from Lemma 5.2.1, we have

$$W_{f,g} = f'g - g'f = g^2 \frac{d}{dx} \left(\frac{f}{g} \right) = g^2 \sum_{j=1}^l \frac{-\beta_j}{(x-b_j)^2}.$$

Since the backward direction holds as well, we can record the following.

Lemma 5.2.2. *Let $f, g \in \mathbb{R}[x]$ be real stable. Then f and g have interlaced roots if and only if $W_{f,g}$ is non-negative for all $x \in \mathbb{R}$ or non-positive for all $x \in \mathbb{R}$.*

We say that the real polynomials f and g are in *proper position*, written $f \ll g$, if f and g are real stable and $W_{f,g} \leq 0$ for all $x \in \mathbb{R}$.

Theorem 5.2.3 (Hermite-Biehler). *Let f, g be non-constant polynomials in $\mathbb{R}[x]$. Then $g + if$ is stable if and only if $f \ll g$.*

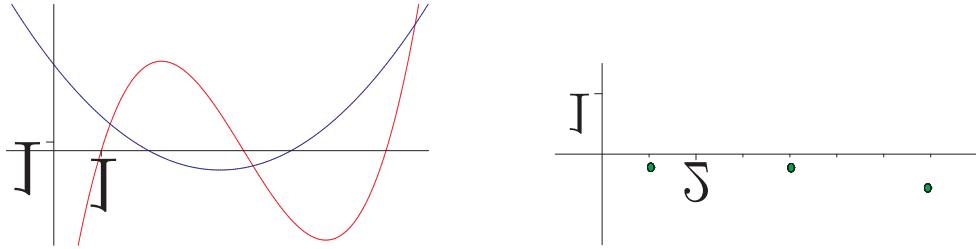


Figure 5.1: The functions $f = (x - 2)(x - 5)$ and $g = (x - 1)(x - 4)(x - 7)$ and the locations of the roots of the stable polynomial $g + if$ in the complex plane

Proof. We can assume that fg has only simple roots, the general case then follows from converging approximations.

First let f and g be real stable with $f \ll g$. For $x \in \mathcal{H}$ and $\tau \in \mathbb{R}$ we have $\Im(\frac{1}{x-\tau}) < 0$, and thus the representation in Lemma 5.2.1,

$$\frac{f}{g} = \alpha + \sum_{j=1}^l \frac{\beta_j}{x - b_j} \text{ with roots } b_j \text{ of } g \text{ and } \alpha, \beta_j \in \mathbb{R},$$

shows that $\Im(f(x)/g(x)) < 0$. Assuming that $g+if$ were not stable would give some $z \in \mathcal{H}$ with $g(z) + if(z) = 0$, and hence $f(z)/g(z) = i$, in contradiction to $\Im(f(x)/g(x)) < 0$. Hence, $g+if$ is stable.

Conversely, let $h = g+if$ be stable. Then any zero α of h satisfies $\Im(\alpha) < 0$, which yields $|z - \alpha| > |\bar{z} - \alpha|$ for every z with $\Im(z) > 0$; hence, $|h(z)| > |h(\bar{z})|$, and therefore

$$0 < h(z)\overline{h(z)} - h(\bar{z})\overline{h(\bar{z})} = 2i(g(\bar{z})f(z) - g(z)f(\bar{z})). \quad (5.7)$$

Thus, f and g have only real roots, because non-real roots come in conjugate pairs and would cause the last expression to vanish and thus would give a contradiction.

It remains to show that the Wronskian $W_{f,g}$ is non-positive for all $x \in \mathbb{R}$. Observe that for every z with $\Im(z) > 0$, we have $-2i(z - \bar{z}) > 0$, and hence, (5.7) implies

$$\frac{(f(z) - f(\bar{z}))}{z - \bar{z}}g(\bar{z}) - \frac{(g(z) - g(\bar{z}))}{z - \bar{z}}f(\bar{z}) < 0.$$

For $\Im(z)$ converging to 0, this gives the non-positivity of the Wronskian. \square

Example 5.2.4. For $f = (x - 2)(x - 5)$ and $g = (x - 1)(x - 4)(x - 7)$, the polynomial $g + if$ is stable, since the roots of f and g interlace and $W_{f,g} \leq 0$ on \mathbb{R} . The polynomial $f + ig$ is not stable however.

The Routh-Hurwitz problem. The Routh-Hurwitz problem asks to decide whether all the roots of a given univariate real polynomial have negative real parts. Solutions of the problem go back until 1876, when Edward John Routh found a recursive algorithm.

The importance of the problem and its solutions comes from systems of linear differential equations, which occur, for instance, in control theory. Here, the question arises whether solutions of the system are stable in the sense that they converge to zero, when the time t goes to infinity.

Example 5.2.5. An ordinary linear differential equation with constant coefficients in a single variable,

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1y' + a_0y = 0$$

with coefficients $a_0, \dots, a_{n-1} \in \mathbb{R}$, leads to solutions of the form

$$y(t) = \sum_{i=1}^n \alpha_i e^{\lambda_i t}$$

with coefficients α_i , where $\lambda_1, \dots, \lambda_n$ are the roots, say, pairwise different, of the characteristic polynomial

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0.$$

For the specific example $y^{(3)} + 4y^{(2)} + 14y' + 20 = 0$, the characteristic polynomial $x^3 + 4x^2 + 14x + 20$ has the roots -2 and $-1 \pm 3i$ and, therefore, it is stable. We obtain real basis solutions

$$e^{-2t}, \quad e^{-t} \sin(3t), \quad e^{-t} \cos(3t),$$

which converge to zero for $t \rightarrow \infty$. In a situation where the characteristic polynomial has a nonzero root with nonnegative real part, the corresponding solution will not converge.

A real polynomial $f \in \mathbb{R}[x]$ is called *Hurwitz stable* if all its roots have negative real parts. There is an immediate necessary condition.

Theorem 5.2.6 (Stodola's Criterion). *All coefficients of a Hurwitz stable polynomial have the same sign.*

Proof. We can assume that $f \in \mathbb{R}[x]$ is a monic stable polynomial. Its roots consist of real numbers $\alpha_1, \dots, \alpha_s < 0$ and of conjugate non-real pairs $\beta_k \pm i\gamma_k$ with $\beta_k < 0$ and $\gamma_k \in \mathbb{R} \setminus \{0\}$, $1 \leq k \leq t$. Hence,

$$f = \prod_{j=1}^s (x - \alpha_j) \cdot \prod_{k=1}^t ((x - \beta_k)^2 + \gamma_k^2).$$

Since $\alpha_s, \beta_k < 0$, this representation of f shows that all coefficients of f are positive. \square

We observe that a polynomial $f = \sum_{j=0}^n a_j x^j \in \mathbb{R}[x]$ of degree n is Hurwitz stable if and only if

$$F(x) = i^{-n} f(ix) = a_n x^n - i a_{n-1} x^{n-1} - a_{n-2} x^{n-2} + i a_{n-3} x^{n-3} + a_{n-4} x^{n-4} + \cdots \quad (5.8)$$

has all its roots in the open upper half-plane. Define

$$\begin{aligned} F_0(x) &= \Re F(x) = a_n x^n - a_{n-2} x^{n-2} + a_{n-4} x^{n-4} - \cdots, \\ F_1(x) &= \Im F(x) = -a_{n-1} x^{n-1} + a_{n-3} x^{n-3} - a_{n-5} x^{n-5} + \cdots, \end{aligned} \quad (5.9)$$

where \Re, \Im denote the real and imaginary part of a polynomial in the sense of the decomposition $F(x) = \Re F(x) + i\Im F(x)$.

Note that if f is Hurwitz stable, then $\deg F_1 = n - 1$, since otherwise $a_{n-1} = 0$, which would imply that the sum of all roots were 0 and thus contradict Hurwitz stability.

Lemma 5.2.7. *For a polynomial $f = \sum_{j=0}^n a_j x^j \in \mathbb{R}[x]$ of degree $n \geq 2$, the following are equivalent:*

1. *f is Hurwitz stable.*
2. *$F = F_0 + iF_1$ has all its zeroes in the open upper half-plane.*
3. *Defining $R \in \mathbb{R}[x]$ as the negative remainder in the division by remainder of F_0 by F_1 , i.e.,*

$$F_0 = QF_1 - R \quad \text{with multiplier } Q \in \mathbb{R}[x],$$

we have $\deg R = n - 2$, the leading coefficients of F_0 and R have the same signs, and $F_1 + iR$ has all its zeroes in the open upper half-plane.

Proof. The equivalence of the first two statement has already been observed. And without loss of generality, we can assume that F has a positive leading coefficient.

If F has all its zeroes in the open upper half-plane, then $F(-x)$ is strictly stable (in the sense that all roots have strictly negative imaginary parts). By the Hermite-Biehler Theorem 5.2.3, the polynomials $F_0(-x)$ and $F_1(-x)$ are real-rooted and have strictly interlaced roots, that is, there does not exist a common root. And therefore also the polynomials F_0 and F_1 are real rooted and have strictly interlaced roots. Hence, F_1 and R have strictly interlaced roots. Since $\deg F_1 = n - 1$ and $\deg R \leq n - 2$, the strict interlacing property also shows $\deg R = n - 2$. By Stodola's Criterion 5.2.6, the leading coefficient of F_1 is negative, and hence the leading coefficient of the linear polynomial Q must be negative. By inspecting the largest root of F_1 , this implies a positive leading coefficient of R .

Since $\deg F_1 = \deg R + 1$, we see that the largest root among the roots of F_1 and R belongs to F_1 . Inspecting the Wronskian (5.6) at this largest root shows that the Wronskian W_{R,F_1} is positive. Thus, by another application of the Hermite-Biehler Theorem, $F_1 + iR$ has all its roots in the open upper half-plane.

Conversely, if the remainder R has these properties, then the same arguments show that $F = F_0 + iF_1$ has all its zeroes in the open upper half-plane. \square

The construction in Lemma 5.2.7 gives rise to a Euclidean algorithm. Identifying F_2 with the negative remainder in the first step, and so on, successively gives a sequence F_0, F_1, F_2, \dots of polynomials. Eventually, some F_k becomes constant. As a consequence of Lemma 5.2.7, if the degree sequence does not coincide with $n, n - 1, n - 2, \dots, 1, 0$, then f has a zero on the imaginary axis and is not Hurwitz stable. And if the signs of the leading coefficients do not alternate, then f cannot be Hurwitz stable either. Now consider the situation that the degree sequence coincides with $n, n - 1, n - 2, \dots, 1, 0$ and that the signs of the leading coefficients alternate. If the linear polynomial $F_{k-1} + iF_k$ has its root in the open upper half-plane, then f is Hurwitz stable, and if $F_{k-1} + iF_k$ has its root in the open lower half-plane, then f is not. If f is stable, the first constant polynomial F_k in the sequence cannot be the zero polynomial by Lemma 5.2.7.

Example 5.2.8. For $f = x^5 + 9x^4 + 41x^3 + 119x^2 + 200x + 150$, we obtain $F_0 = x^5 - 41x^3 + 200x$, $F_1 = -9x^4 + 119x^2 - 150$. Successively dividing with remainder gives $F_2 = \frac{250}{9}x^3 - \frac{550}{3}x$, $F_3 = -\frac{298}{5}x^2 + 150$, $F_4 = \frac{16900}{149}x$, $F_5 = -150$. Since $F_4 + iF_5$ has its root in the open upper half-plane, f is Hurwitz-stable.

Indeed, the zeroes of f are $-1 \pm 3i$, $-2 \pm i$ and -3 .

Corollary 5.2.9. Let $f = \sum_{j=0}^n a_j x^j \in \mathbb{R}[x]$ with $n > 0$ and $a_{n-1}a_n \neq 0$. Then f is Hurwitz stable if and only if the polynomial

$$g = \sum_{j \geq 0} a_{n-1-2j} x^{n-1-2j} + \sum_{j \geq 1} \left(a_{n-2j} - \frac{a_n}{a_{n-1}} a_{n-1-2j} \right) x^{n-2j}$$

of degree $n - 1$ is Hurwitz stable, where all coefficients a_j with $j < 0$ are assumed to be zero.

Note that the precondition $a_{n-1} \neq 0$ is not a real restriction due to Stodola's Criterion 5.2.6.

Proof. Considering the representations (5.8) and (5.9) for the degree $(n - 1)$ -polynomial g , we obtain

$$G(x) = i^{-(n-1)} g(x) = G_0(x) + iG_1(x)$$

with

$$\begin{aligned} G_0(x) &= a_{n-1}x^{n-1} - a_{n-3}x^{n-3} + \dots, \\ G_1(x) &= -\left(a_{n-2} - \frac{a_n}{a_{n-1}}a_{n-3}\right)x^{n-2} + \left(a_{n-4} - \frac{a_n}{a_{n-1}}a_{n-5}\right)x^{n-4} - \dots \end{aligned}$$

Hence, $G_0 = -F_1$ and G_1 is the remainder in the division of F_0 by F_1 , where F_0 and F_1 are defined as in (5.9). By Lemma 5.2.7, $G = G_0 + iG_1 = -(F_1 - iG_1)$ has all its zeroes in the open upper half plane if and only if $F(x)$ has. Thus, also by Theorem (5.9), g is Hurwitz stable if and only if f is. \square

There exists a beautiful characterization of Hurwitz stability in terms of determinantal conditions. For $f = \sum_{j=0}^n a_j x^j \in \mathbb{R}[x]$ define the *Hurwitz determinants* $\delta_0, \dots, \delta_n$ by $\delta_0 = 1$ and

$$\delta_k = \det \begin{pmatrix} a_{n-1} & a_{n-3} & a_{n-5} & \cdots & a_{n+1-2k} \\ a_n & a_{n-2} & a_{n-4} & \cdots & a_{n+2-2k} \\ 0 & a_{n-1} & a_{n-3} & \cdots & a_{n+3-2k} \\ 0 & a_n & a_{n-2} & \cdots & a_{n+4-2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{n-k} \end{pmatrix}, \quad 1 \leq k \leq n,$$

where we set $a_j = 0$ for $j < 0$. The underlying matrices H_k are called *Hurwitz matrices*.

Theorem 5.2.10. *Let $f(x) = \sum_{j=0}^n a_j x^j$ with $n \geq 1$ and $a_n > 0$. Then f is Hurwitz stable if and only if all the Hurwitz determinants $\delta_1, \dots, \delta_n$ are all positive.*

Proof. The proof is by induction, where the case $n = 1$ follows immediately from the Hurwitz matrix $H_1 = (a_{n-1})$. Now let $n > 1$. For $1 \leq k \leq n$, consider the Hurwitz matrix H_k of order k of f . We can assume $a_{n-1} > 0$, since otherwise f is not Hurwitz stable by Stodola's Condition 5.2.6 and $\delta_1 = \det(a_{n-1})$ is not positive.

Subtracting $\frac{a_n}{a_{n-1}}$ times the $(2j-1)$ -st row from the $2j$ -th row, for all j , gives a matrix whose lower right $(n-1) \times (n-1)$ -matrix is the Hurwitz matrix of g from Corollary 5.2.9. The initial column of the new matrix only contains only a single nonzero entry, namely a_{n-1} , and this entry is positive. Hence, the Hurwitz determinants $\delta_1, \dots, \delta_n$ are all positive if and only if the Hurwitz determinants $\delta'_1, \dots, \delta'_{n-1}$ of g are positive. By the induction hypothesis, this holds true if and only if g is Hurwitz stable, that is, by Corollary 5.2.9, if and only if f is Hurwitz stable. \square

Example 5.2.11. Let $f(x) = x^3 + x^2 + 5x + 20$. Not all principal minors of the Hurwitz matrix

$$H_3(f) = \begin{pmatrix} 1 & 20 & 0 \\ 1 & 5 & 0 \\ 0 & 1 & 20 \end{pmatrix}$$

are positive, and hence f is not Hurwitz stable.

Exercises

1. Let f and g be monic of degree $d-1$ and d and with roots a_1, \dots, a_{d-1} and b_1, \dots, b_d . The roots interlace if and only if $W_{f,g} \leq 0$ on \mathbb{R} .
2. If $f \in \mathbb{R}[x]$ is of degree n , then the polynomials $f(x)$ and $x^n f(1/x)$ have the same number of zeros in the open left half-plane.

3. Let all the coefficients of a real polynomial f of degree $n \geq 1$ have the same sign. Show that for $n = 2$ this implies stability of f , but that this statement is not true for $n \geq 3$.
4. The Hurwitz determinants δ_n and δ_{n-1} of a monic polynomial $f = x^n + \sum_{j=0}^{n-1} a_j x^j$ satisfy

$$\delta_n = a_0 \delta_{n-1} = (-1)^{n(n+1)/2} z_1 z_2 \cdots z_n \prod_{j < k} (z_j + z_k),$$

where z_1, \dots, z_n are the roots of f .

5.3 Real roots and the trace form

In this section, we consider multivariate real polynomials $f_1, \dots, f_m \in \mathbb{R}[x] = \mathbb{R}[x_1, \dots, x_n]$ which have only a finite number of common zeroes over the complex numbers. For this situation, the Hermite form studied in Section 5.1 can be generalized to the multivariate case and provides a method to compute the number of real zeroes of f_1, \dots, f_m . In the following, let I be the zero-dimensional ideal in $\mathbb{C}[x_1, \dots, x_n]$ generated by the real polynomials f_1, \dots, f_m .

Recall from Section 2.5 the endomorphism in $\mathbb{K}[x_1, \dots, x_n]$ obtained by multiplying the residue class \bar{x}_i of a fixed variable x_i with the residue class \bar{f} of a polynomial f . This construction can be generalized as follows. For a given polynomial $g \in \mathbb{K}[x_1, \dots, x_n]$, we define the endomorphism m_g by

$$\begin{aligned} m_g : \mathbb{K}[x_1, \dots, x_n]/I &\rightarrow \mathbb{K}[x_1, \dots, x_n]/I, \\ \bar{f} &\mapsto \bar{g} \cdot \bar{f}. \end{aligned}$$

In order to be able to count only real zeroes with certain properties, fix a polynomial $h \in \mathbb{R}[x_1, \dots, x_n]$. We construct the bilinear form q_h by

$$\begin{aligned} q_h : \mathbb{K}[x_1, \dots, x_n]/I \times \mathbb{K}[x_1, \dots, x_n]/I &\rightarrow \mathbb{R}, \\ (r, s) &\mapsto \text{Tr}(m_{qrs}). \end{aligned}$$

q_h 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.

Theorem 5.3.1. *For $h \in \mathbb{R}[x_1, \dots, x_n]$, the signature $\sigma(q_h)$ and the rank $\rho(q_h)$ of the bilinear form q_h satisfy*

$$\begin{aligned} \sigma(q_h) &= \#\{z \in \mathcal{V}(I) \cap \mathbb{R}^n : q(z) > 0\} - \#\{z \in \mathcal{V}(I) \cap \mathbb{R}^n : q(z) < 0\}, \\ \rho(q_h) &= \#\{z \in \mathcal{V}(I) : q(z) \neq 0\}. \end{aligned}$$

Note that these numbers count the number of distinct points, not taking into account multiplicities.

Proof. We generalize the proof of Theorem 5.1.6 to the multivariate setting. To keep notation simple, we assume that all multiplicities are 1. The case of general multiplicities can be done using the same techniques as in the univariate situation.

Since I is zero-dimensional, the dimension d of the vector space $\mathbb{K}[x_1, \dots, x_n]/I$ is finite. Fix a monomial ordering and let $\mathcal{B} = \{x^{\alpha(1)}, \dots, x^{\alpha(d)}\}$ be a set of standard monomials with respect to this ordering. By a slight generalization of Stickelberger's Theorem 2.5.4 (see also Exercise 1), the set of eigenvalues of m_g coincides with the values of g at the points of I .

Let q_h be the quadratic form

$$q_h(x) = q_h(x_1, \dots, x_n) = \sum_{z \in \mathcal{V}(I)} h(z)y_z^2, \quad (5.10)$$

where $y_z = z^{\alpha(1)}x_1 + \dots + z^{\alpha(d)}x_n$ for $z \in \mathcal{V}(I)$. Expressing q_h in the x -variables,

$$q_h(x) = x^T C x = \sum_{i,j} c_{ij} x_i x_j$$

with a symmetric matrix C , the generalization of Stickelberger's Theorem implies

$$\begin{aligned} c_{ij} &= \sum_{z \in \mathcal{V}(I)} h(z)z^{\alpha(i)}z^{\alpha(j)} \\ &= \sum_{z \in \mathcal{V}(I)} \text{eigenvalues of } m_{hx^{\alpha(i)}x^{\alpha(j)}} \\ &= \text{Tr}(m_{hx^{\alpha(i)}x^{\alpha(j)}}). \end{aligned} \quad (5.11)$$

With any real zero $z \in \mathcal{V}(I)$, we associate the term $h(z)y_z^2$, where we note that its sign is given by the sign of $h(z)$. With any non-real conjugate pair $\{z, z'\}$, we associate the terms

$$(q_h)_{\{z, z'\}} = h(z)y_z^2 + h(z')y_{z'}^2,$$

so that, similar to the univariate case, the trace of the resulting 2×2 -block is zero, and thus the signature the signature of $(q_h)_{\{z, z'\}}$ is zero. And its rank is two. Building together the local view on the terms of the quadratic form q_h , the statement follows. \square

For the special case $q = 1$ we obtain:

Corollary 5.3.2. *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 $\mathcal{B} = \{1, x, \dots, x^{d-1}\}$. Then (5.11) implies

$$c_{ij} = \sum_{p \in \mathcal{V}(I)} z^{i-1}z^{j-1}.$$

Thus we have recovered the Hankel matrix $H_1(p)$ from (5.3) containing the Newton sums of p .

Example 5.3.3. We consider the three planar curves given by the real polynomials $f_1 = x^2 + y^2 - 5$, $f_2 = xy - 2$, $f_3 = x^2y + y^2x - 4y + 2$. A Gröbner with respect to the lexicographical ordering $x \succ y$ is $\{y^2 - y - 2, x - y + 1\}$, and the set of standard monomials is $\{1, y\}$. The multiplication matrices are

$$M_x = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad M_y = \begin{pmatrix} 0 & 1 \\ 2 & 1 \end{pmatrix} \quad \text{and} \quad M_{y^2} = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}.$$

Since $\text{Tr}(M_x) = 2$, $\text{Tr}(M_y) = 1$, $\text{Tr}(M_{y^2}) = 5$, we obtain the symmetric matrix of the trace form

$$T_1 = \begin{pmatrix} 2 & 1 \\ 1 & 5 \end{pmatrix}.$$

Since this matrix has two positive eigenvalues, $V(f_1, f_2, f_3)$ has two real zeroes. Indeed, in this example, there no complex zeroes. Similarly, for $q = x$, we obtain

$$T_x = \begin{pmatrix} -1 & 4 \\ 4 & 2 \end{pmatrix}.$$

Since M_x has rank 2 and signature 0, one of the two common zeroes of the three curves has positive x -coordinate and one has negative x -coordinate. In fact, the two real zeroes in this example are $(2, 1)$ and $(-2, -1)$, see Figure 5.2.

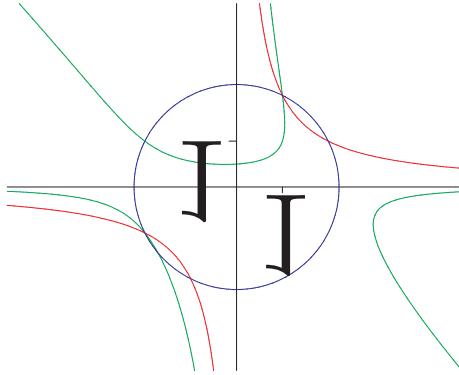


Figure 5.2: Three planar curves with two real intersection points

In fact, the signature can be computed without actually determining the positive and negative eigenvalues.

Theorem 5.3.4. *Let A be a symmetric real matrix. Then the number of positive eigenvalues of A 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 Descartes' 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 Descartes' Rule of Signs holds with equality. \square

Exercises

1. (*Generalized Stickelberger's Theorem.*) Let \mathbb{K} be algebraically closed and $I \subset \mathbb{K}[x_1, \dots, x_n]$ a zero-dimensional ideal. For any $g \in \mathbb{K}[x_1, \dots, x_n]$, a scalar $\lambda \in \mathbb{K}$ is an eigenvalue of the endomorphism m_g if and only if there exists a point $a \in \mathcal{V}(I)$ with $g(a) = \lambda$.
2. In how many real points do the curve $x^2 + 2x - y = 0$ and $x^2 + y^2 + 2x - 1 = 0$ intersect?
3. (*Apollonius' circles.*) Give a construction of three circles in the real plane \mathbb{R}^2 , such that there are eight real circles tangent to all of the three given circles. Use the trace form to prove reality of all these eight solutions.

5.4 Semialgebraic sets and polyhedra

A *semialgebraic set* in \mathbb{R}^n is a subset of \mathbb{R}^n satisfying a Boolean combination (i.e., using finite intersections, finite unions, and complements) of sets of the form

$$\{x \in \mathbb{R}^n : f(x) > 0\}$$

with $f \in \mathbb{R}[x_1, \dots, x_n]$.

Example 5.4.1. i) The shaded area in Figure 5.3 visualizes the semialgebraic set

$$\{(x, y) \in \mathbb{R}^2 : (x^2 + y^2)^3 - 10x^2y^2 \geq 0 \text{ and } (x^2 + y^2 - 1)^3 - x^2y^3 \leq 0\},$$

and the algebraic curves underlying the two inequalities are drawn in blue and red.

ii) Any set of the form $\{x \in \mathbb{R}^n : f_i(x) \geq 0, 1 \leq i \leq m\}$ with $f_1, \dots, f_m \in \mathbb{R}[x_1, \dots, x_n]$ is a semialgebraic set. The sets of this form are called *basic closed semialgebraic sets*.

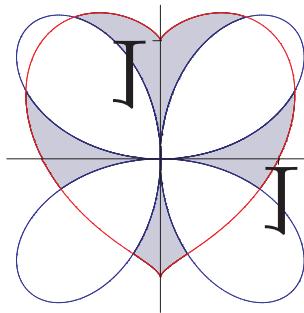


Figure 5.3: A semialgebraic set

The semialgebraic subsets of \mathbb{R} are the unions of finitely many points and open intervals. Also note that every algebraic variety in \mathbb{R}^n is a semialgebraic set. The following statement collects some further elementary properties, whose proofs do not require paper.

- Lemma 5.4.2.** 1. Let $f : \mathbb{R}^m \rightarrow \mathbb{R}^n$ be a polynomial mapping : $f = (f_1, \dots, f_n)$, where $f_i \in \mathbb{R}[x_1, \dots, x_n]$. For any semialgebraic subset A of \mathbb{R}^n , the preimage $f^{-1}(A)$ is a semialgebraic subset of \mathbb{R}^m .
2. 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.
3. If $A \subset \mathbb{R}^m$ and $B \subset \mathbb{R}^n$ are semialgebraic sets, then $A \times B$ is a semialgebraic subset of $\mathbb{R}^m \times \mathbb{R}^n$.

There are various normal forms for the Boolean combinations in the definition of a semialgebraic sets. One of them is discussed in the following theorem:

- Theorem 5.4.3.** 1. Every semialgebraic set $S \subset \mathbb{R}^n$ can be written as a finite union of sets of the form

$$\{x \in \mathbb{R}^n : g(x) = 0 \text{ and } f_i(x) > 0, 1 \leq i \leq m\}. \quad (5.12)$$

2. Every semialgebraic set can be written as the projection of a real algebraic variety.

Before the proof, observe that every real algebraic variety can be written in the form (5.12), since $\{x \in \mathbb{R}^n : h_i(x) = 0, 1 \leq i \leq m\} = \{x \in \mathbb{R}^n : \sum_{i=1}^m h_i(x)^2 = 0\}$ for any $h_1, \dots, h_m \in \mathbb{R}[x]$.

Proof. Denote by \mathcal{S} the class of finite unions of the form (5.12). Clearly, \mathcal{S} is contained in the class of semialgebraic sets. Hence, it suffices to show that \mathcal{S} is closed under intersection, finite union and taking the complement. For the intersection this follows from the observation prior to the proof, and for the finite union it follows from the definition of \mathcal{S} . Further, if T is a set of the form (5.12), then

$$x \notin T \iff g(x) > 0 \text{ or } g(x) < 0 \text{ or } f_i(x) \leq 0 \text{ for some } i \in \{1, \dots, m\}.$$

Since $f_i(x) \leq 0$ if and only if $f_i(x) < 0$ or $f_i(x) = 0$, we see that the complement of T is contained in \mathcal{S} , and further the complement of any set in \mathcal{S} is contained in \mathcal{S} as well.

Then, for the second statement of the theorem, note that any semialgebraic set T of the form (5.12) is the projection of the real algebraic variety

$$\{(x, y) \in \mathbb{R}^{n+m} : g(x) = 0, y_1^2 f_1(x) = 1, \dots, y_m^2 f_m(x) = 1\} \quad (5.13)$$

onto the x -coordinates. To extend this argument to the whose class \mathcal{S} , we simply consider unions of sets of the form (5.13) within a space that contains enough y -variables. \square

Before further studying general semialgebraic sets, it is worth considering the prominent subclass of polyhedra. A *Polyhedron* in \mathbb{R}^n is a subset of \mathbb{R}^n which can be written as

$$\{x \in \mathbb{R}^n : \ell_i(x) \geq 0\} \quad (5.14)$$

with some affine-linear polynomials $\ell_i \in \mathbb{R}[x]$. Clearly, every polyhedron is a semialgebraic set. The class of polyhedra is closed under projections.

Theorem 5.4.4. *Let $n \geq 2$, P be a polyhedron in \mathbb{R}^n and $\pi : \mathbb{R}^n \rightarrow \mathbb{R}^{n-1}$ be the natural projection onto the first $n - 1$ coordinates. Then $\pi(P)$ is a polyhedron.*

Proof. There is a constructive proof for the statement, which is named as Fourier-Motzkin elimination. Let P be given by the linear inequalities $\sum_{j=1}^n a_{ij}x_j \leq b_i$, $1 \leq i \leq m$. Since multiplying an inequality with a non-zero real number does not change the polyhedron P , we can assume that the linear inequalities are of the form

$$\begin{aligned} \sum_{j=1}^{n-1} a_{ij}x_j + x_n &\leq b_i, \quad 1 \leq i \leq m', \\ \sum_{j=1}^{n-1} a_{ij}x_j - x_n &\leq b_i, \quad m' < i \leq m'', \\ \sum_{j=1}^{n-1} a_{ij}x_j &\leq b_i, \quad m'' < i \leq m. \end{aligned} \tag{5.15}$$

The first two rows are equivalent to

$$\max_{m' < i \leq m''} \left(\sum_{j=1}^{n-1} a_{ij}x_j - b_i \right) \leq x_n \leq \min_{1 \leq k \leq m'} \left(b_k - \sum_{j=1}^{n-1} a_{kj}x_j \right),$$

and thus x_n can be eliminated. An equivalent system without x_n is

$$\sum_{j=1}^{n-1} a_{ij}x_j - b_i \leq b_k - \sum_{i=1}^{n-1} a_{kj}x_j, \quad m' < i \leq m'', \quad 1 \leq k \leq m'. \tag{5.16}$$

Any solution for (x_1, \dots, x_{n-1}) of this system can be extended to a solution $x = (x_1, \dots, x_n)$ of (5.15). \square

From this projection statement and the Fourier-Motzkin elimination technique employed in the proof, the following characterization of emptiness of a polyhedron can be deduced.

Theorem 5.4.5 (Farkas' Lemma). *Let $Ax \leq b$ be a system of linear inequalities with $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Either the system has a solution x , or there exists a non-negative vector $y \in \mathbb{R}^m$ with $A^T y = 0$ and $b^T y < 0$.*

Proof. We prove the statement by induction over n . It is convenient to establish the induction base for $n = 0$. In that case, the two alternatives are $b \geq 0$ and the existence of a non-negative vector $y \in \mathbb{R}^m$ with $b^T y < 0$. Since the latter condition merely means that b is not a non-negative vector, exactly one of the two alternatives hold.

Assume now that the system $Ax \leq b$ in $n \geq 1$ variables does not have a solution, where we can use the notation from (5.15). Then the Fourier-Motzkin elimination step (5.16) shows that the system $A'x' \leq b'$ defined by $x' = (x'_1, \dots, x'_n)$ and

$$\begin{aligned} \sum_{j=1}^{n-1} (a_{ij} + a_{kj})x'_j &\leq b_k - b_i, \quad m' < i \leq m'', \quad 1 \leq k \leq m', \\ \sum_{j=1}^{n-1} a_{ij}x'_j &\leq b_i, \quad m'' < i \leq m \end{aligned}$$

does not have a solution either. By induction hypothesis, the vector $(0, \dots, 0, -1)$ is a non-negative combination of the rows of the matrix (A', b') . Since each row of the matrix $(A', \mathbf{0}, b')$ (where $\mathbf{0}$ denotes a zero column) is a sum of two rows of (A, b) , the vector $(0, \dots, 0, -1)$ is a non-negative combination of the rows of (A, b) . \square

The much more general class of semialgebraic sets is also closed under projections.

Theorem 5.4.6 (Projection theorem.). *Let $n \geq 2$, S be a semialgebraic set in \mathbb{R}^n and $\pi : \mathbb{R}^n \rightarrow \mathbb{R}^{n-1}$ be the natural projection onto the first $n - 1$ coordinates. Then $\pi(S)$ is semialgebraic.*

Note that this statement fails for real algebraic sets. For example, the projection π of the circle $C = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ onto the x -variable gives the closed interval $\pi(C) = [-1, 1]$, which is not an algebraic set.

Theorem 5.4.6 can be deduced from the general principle of quantifier elimination (see Theorem A.1.13 in Appendix A.1.5).

Proof. We can write $\pi(S)$ in the form

$$\pi(S) = \{x \in \mathbb{R}^{n-1} : \exists x_n \text{ with } (x_1, \dots, x_n) \in S\}.$$

By Theorem 5.4.3, we can assume that S can be represented as $S = S_1 \cup \dots \cup S_m$ 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 $g, f_1, \dots, f_m \in \mathbb{R}[x]$. First we consider the special case $i = 1$, i.e., $S = S_1$. Let $c = (c_1, \dots, c_N)$ (for some N) be the sequence of all coefficients of the polynomials g, f_1, \dots, f_m . Let $G, F_1, \dots, F_m \in \mathbb{Z}[C_1, \dots, C_N, x_1, \dots, x_n]$ be the polynomials which result from g, f_1, \dots, 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 $G_i, F_{ij} \in \mathbb{Z}[C_1, \dots, C_N, x_1, \dots, x_n]$ such that $\pi(S)$ is of the form T_1, \dots, 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 $G_i(c, x_1, \dots, x_{n-1})$ and $F_{ij}(c, x_1, \dots, x_{n-1})$.

In the case $S = S_1 \cup \dots \cup S_m$ with $m > 1$, we can apply quantifier elimination to each of the S_i separately, then taking the union of the results. \square

Let $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}^m$ be semialgebraic sets. A map $f : X \rightarrow Y$ 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 5.4.7. *Let $f : X \rightarrow Y$ 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 repeated application of the Projection Theorem 5.4.6. \square

Cylindrical algebraic decomposition We sketch some basic ideas on the cylindrical algebraic decomposition of semialgebraic sets. This technique provides a systematic, but often computationally costly, access for deciding algorithmic questions on semialgebraic sets. The idea is to partition the spaces $\mathbb{R}^n, \mathbb{R}^{n-1}, \dots, \mathbb{R}^1$ into finitely many semialgebraic subsets called *cells*, with the idea that in each of the cells evaluation of the input polynomials gives uniform sign patterns. The decomposition of the space \mathbb{R}^n is induced by successive projections to the smaller-dimensional spaces.

As the technique is directed to exact algorithmic computation, we usually start from rational rather than general real polynomials, since this allows to work with algebraic numbers.

A *cylindrical algebraic decomposition (CAD)* of \mathbb{R}^n is a sequence $\mathcal{C}_1, \dots, \mathcal{C}_n$, where \mathcal{C}_k is a partition of \mathbb{R}^k into finitely many cells, and the following inductive conditions are satisfied.

If $k = 1$: \mathcal{C}_1 partitions \mathbb{R}^1 into finitely many algebraic numbers and the finite and infinite open intervals bounded by these numbers.

If $k > 1$: Assume inductively, that $\mathcal{C}_1, \dots, \mathcal{C}_{k-1}$ is a CAD of \mathbb{R}^{k-1} . And for each cell $C \in \mathcal{C}_{k-1}$, let $g_C(x) = g_C(x', x_n)$ be a polynomial in $\mathbb{Q}[x] = (\mathbb{Q}[x_1, \dots, x_{n-1}])[x_n]$. For a fixed g_C , let $\alpha_1(x'), \dots, \alpha_m(x')$ be the roots of g_C in increasing order, considered as a continuous function in x' .

Each set in \mathcal{C}_k is required to be of the form

$$\begin{aligned} C_i &= \{(x', x_n) : x_n = \alpha_i(x')\}, \quad 1 \leq i \leq m, \\ \text{or } C'_i &= \{(x', x_n) : x_n \in (\alpha_i(x'), \alpha_{i+1}(x'))\}, \quad 1 \leq i \leq m-1, \end{aligned}$$

where we set $\alpha_0(x') = -\infty$ and $\alpha_{m+1}(x') = \infty$.

Definition 5.4.8. Let $f_1, \dots, f_r \in \mathbb{R}[x]$.

1. A subset $C \subset \mathbb{R}^n$ is called *(f_1, \dots, f_r) -invariant* if every polynomial f_i has a uniform sign on C , that is, either $<$ or $=$ or $>$.
2. A cylindrical algebraic decomposition $\mathcal{C}_1, \dots, \mathcal{C}_n$ is *adapted to f_1, \dots, f_r* if every cell $C \in \mathcal{C}_n$ is (f_1, \dots, f_r) -invariant.

By Collins' work, an adapted CAD can be effectively computed for given rational polynomials f_1, \dots, f_r . Carrying this out requires a substantial amount of technical details concerning resultants and discriminants, which we omit here.

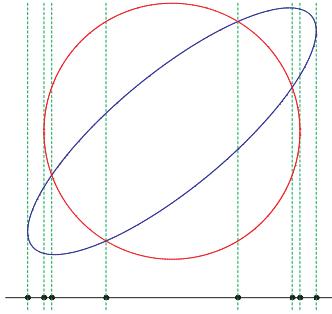


Figure 5.4: Illustration of an adapted CAD of two polynomials in \mathbb{R}^2 .

Theorem 5.4.9 (Collins). *For $f_1, \dots, f_r \in \mathbb{Q}[x_1, \dots, x_n]$, there exists an adapted CAD of \mathbb{R}^n , and it can be algorithmically computed.*

Figure 5.4 visualizes an adapted CAD of two polynomials in \mathbb{R}^2 .

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 $g = (g_1, \dots, g_m) : X \rightarrow Y$ is semialgebraic if and only if all the functions g_i are semialgebraic.
 - (b) If $h : X \rightarrow Y$ is semialgebraic then $h^{-1}(Y)$ is a semialgebraic set.
3. Show that the set $\{(x, y) \in \mathbb{R}^2 : y = \lfloor x \rfloor \text{ or } (x \in \mathbb{Z} \text{ and } x \leq y \leq x + 1)\}$ (“infinite staircase”) is not semialgebraic.
4. Let $Q = \{(x_1, \dots, x_n) \in \mathbb{R}^n : x_1 > 0, \dots, x_n > 0\}$ be the interior of the positive orthant in \mathbb{R}^n . Show that Q cannot be written in the form

$$Q = \{x \in \mathbb{R}^n : p_1(x) > 0, \dots, p_{n-1}(x) > 0\} \text{ with } p_1, \dots, p_{n-1} \in \mathbb{R}[x_1, \dots, x_n].$$

Note: By a Theorem of Bröcker, every semialgebraic set of the form $S = \{x \in \mathbb{R}^n : p_1(x) > 0, \dots, p_k(x) > 0\}$ with $k \in \mathbb{N}$ can be written using at most n strict inequalities, i.e., in the form

$$S = \{x \in \mathbb{R}^n : q_1(x) > 0, \dots, q_n(x) > 0\}.$$

5. Construct an adapted CAD for the sequence of polynomials $f_1 = (x^2 + y^2)^3 - 10x^2y^2$, $f_2 = (x^2 + y^2 - 1)^3 - x^2y^3$ from Example 5.4.1.
6. Give an adapted CAD for the unit sphere in \mathbb{R}^3 given by the polynomial $f = x^2 + y^2 + z^2 - 1$. What is the minimum number of cells required?

5.5 Spectrahedra

Let $\text{Sym}_k(\mathbb{R})$ denote the set of real symmetric $k \times k$ -matrices. Given $A_0, \dots, A_n \in \text{Sym}_k(\mathbb{R})$, we consider the *linear matrix polynomial*

$$A(x) := A_0 + \sum_{i=1}^n x_i A_i.$$

Then the semialgebraic set

$$S := \{x \in \mathbb{R}^n : A(x) \succeq 0\}$$

is called a *spectrahedron*, where \succeq denotes positive semidefiniteness of a matrix. The inequality $A_0 + \sum_{i=1}^n x_i A_i \succeq 0$ is called a *linear matrix inequality (LMI)*.

For instance, 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 5.5.1. 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 $1 - x_1^2 - x_2^2 \geq 0$.

Example 5.5.2. Figure 5.5 shows the example of the ellotope

$$S_A = \{x \in \mathbb{R}^3 : \begin{pmatrix} 1 & x_1 & x_2 \\ x_1 & 1 & x_3 \\ x_2 & x_3 & 1 \end{pmatrix} \succeq 0\}$$

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 $A(x)$ indexed by the set $I = 2^{\{1, \dots, k\}} \setminus \emptyset$. A slightly more concise representation is given by the following statement.

Theorem 5.5.3. *Any spectrahedron S is a basic closed semialgebraic set. In particular, given the modified characteristic polynomial*

$$t \mapsto \det(A(x) + tI_k) =: t^k + \sum_{i=0}^{k-1} p_i(x)t^i,$$

S has the representation $S = \{x \in \mathbb{R}^n : p_i(x) \geq 0, 1 \leq i \leq k\}$.

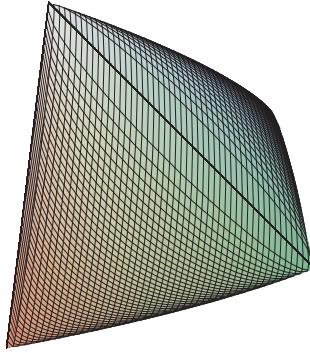


Figure 5.5: Visualization of an ellipope.

Proof. Denoting by $\lambda_1(x), \dots, \lambda_k(x)$ the eigenvalues of the linear pencil $A(x)$, we observe

$$\det(A(x) + tI_k) = (t + \lambda_1(x)) \cdots (t + \lambda_k(x)).$$

Since $A(x)$ is symmetric all $\lambda_i(x)$ are real, for any $x \in \mathbb{R}^n$. Comparing the coefficients we have

$$p_{k-i} = \sum_{t_1 < \dots < t_i} \lambda_{t_1}(x) \cdots \lambda_{t_i}(x), \quad 1 \leq i \leq k$$

with $p_m(x) = 1$.

Now “ \subset ” of the desired representation follows from the fact that positive semidefiniteness of $A(x)$ implies non-negativity of all eigenvalues. Conversely, if $p_i(x) \geq 0$ for all i , the determinant polynomial $\det(A(x) + tI_k)$ has no sign changes. By Descartes’ Rule of Signs this implies that it has no positive roots, and therefore $A(x)$ is positive semidefinite. \square

We investigate some geometric properties of spectrahedra.

Definition 5.5.4. A closed subset $C \subset \mathbb{R}^n$ is called an *algebraic interior* if there exists a polynomial $p \in \mathbb{R}[x_1, \dots, x_n]$ such that C is the closure of a connected component of the positivity domain in $\{x \in \mathbb{R}^n : p(x) > 0\}$ of p .

For a given algebraic interior C the defining polynomial of C of minimal degree is unique up to a positive multiple.

Definition 5.5.5. An algebraic interior C is called *rigidly convex* if 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.

Theorem 5.5.6. *Every spectrahedron is rigidly convex.*

Proof. Let $A(x)$ be a monic linear pencil, i.e., $A(x) = I_n + \sum_{i=1}^n A_i x_i$. Then the spectrahedron S_A is an algebraic interior with defining polynomial $p(x) = \det A(x)$. The determinant $p(x) = \det A(x)$ has the property that for every $u \neq \mathbb{R}^n \setminus \{0\}$ the univariate

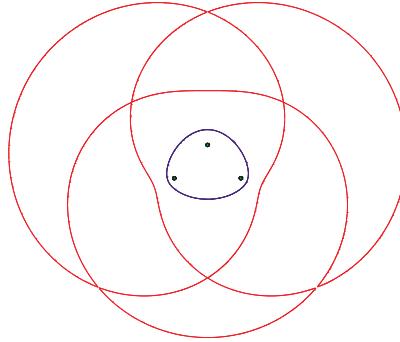


Figure 5.6: The closed curve in the center is a 3-ellipse of the three depicted points.

polynomial $p(tw)$ has only real zeros (cf. Section 5.6, where we will call p a real-zero polynomial then). This is true because the roots of $p(tw)$ can be interpreted as eigenvalues of a symmetric matrices, cf. Theorem 5.5.3. Hence, S_A is rigidly convex. \square

For the case of dimension 2 the converse is true as well.

Theorem 5.5.7 (Helton, Vinnikov). *Every rigidly convex set in \mathbb{R}^2 is a spectrahedron.*

We do not prove this theorem, but illustrate it by two examples.

Example 5.5.8. For given k points $(a_1, b_1)^T, \dots, (a_k, b_k)^T \in \mathbb{R}^2$ the *k -ellipse* with focal points $(a_i, b_i)^T$ and radius d is the plane curve \mathcal{E}_k given by

$$\left\{ (x, y)^T \in \mathbb{R}^2 : \sum_{i=1}^k \sqrt{(x - a_i)^2 + (y - b_i)^2} = d \right\} \quad (5.17)$$

(see Figure 5.6). For the special case $k = 2$ we obtain usual ellipses. The convex hull C of a k -ellipse \mathcal{E}_k is a spectrahedron in \mathbb{R}^2 . In order to see this for the example in Figure 5.6, consider the Zariski closure \mathcal{E}'_3 of the set defined by (5.17); its real points 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 line (not passing through a point of higher multiplicity of the curve) through z contains exactly 8 points of \mathcal{E}'_3 . By Theorem 5.5.7, this property implies that C is a spectrahedron.

Example 5.5.9. Let p be the irreducible polynomial

$$p(x, y) = x^3 - 3xy^2 - (x^2 + y^2)^2$$

(see Figure 5.7). The positivity domain consists of three bounded connected components, as illustrated by 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 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 a spectrahedron.

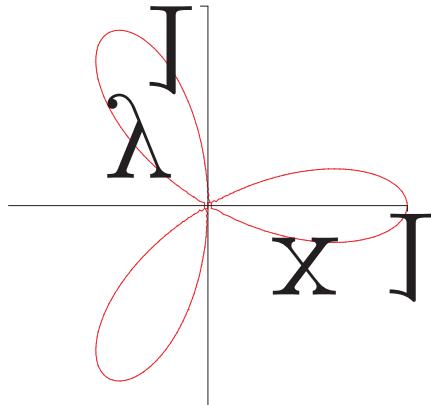


Figure 5.7: The real variety $V_{\mathbb{R}}(p)$ of the polynomial p .

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. It is an open question whether every rigidly convex sets in \mathbb{R}^n is a spectrahedron.

In the following we discuss a geometric property 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 hyperplane H of S such that $H \cap S = F$. We also say that the hyperplane H *exposes* F . See Figure 5.8 for an example. We record the following geometric result.

Theorem 5.5.10 (Ramana and Goldman; Netzer, Plaumann, Schweighofer). *The faces of a rigidly convex set are exposed. Since every spectrahedron is rigidly convex, every face of a spectrahedron set is exposed.*

It is an open question to provide good effective criteria to test whether a given convex semialgebraic set is a spectrahedron or the linear projection of a spectrahedron. An earlier conjecture that every convex semialgebraic set would be the linear projection of a spectrahedron (“Helton-Nie conjecture”) has been disproven by Scheiderer.

Exercises

1. Show that the intersection of two spectrahedra is again a spectrahedron.
2. If $A(x)$ is a linear matrix polynomial and V is a nonsingular matrix, then $G = \{x : V^T A(x) V \succeq 0\}$ is a spectrahedron.

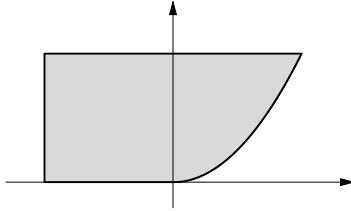


Figure 5.8: The origin is a non-exposed face of the shaded set S , since there does not exist a supporting hyperplane to S which intersects S only in the origin.

3. Convex quadratic inequalities give spectrahedra: Let $f(x) = x^T L^T Lx + b^T x + c$. 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. Show that the Lorentz cone $\mathcal{L} = \{x \in \mathbb{R}^n : x_1 > \sqrt{x_2^2 + \dots + x_n^2}\}$ is a spectrahedron.
 5. Show Theorem 5.5.10 for the special case of spectrahedra, that is, every face of a spectrahedron is exposed.

5.6 Stable and hyperbolic polynomials

Building upon the study of univariate stable polynomials in Section 5.2, we now have a look at stable polynomials in several variables as well as at the more general hyperbolic polynomials. As in Section 5.2, denote by \mathcal{H} the open halfplane $\{x \in \mathbb{C} : \Im(x) > 0\}$.

Definition 5.6.1. A polynomial $p \in \mathbb{C}[x] = \mathbb{C}[x_1, \dots, x_n]$ is called *stable* if it has no root x in \mathcal{H}^n . If $p \in \mathbb{R}[x]$ and p is stable, then we call p *real stable*.

Clearly, a univariate polynomial $p \in \mathbb{R}[x]$ is real stable if and only if it has only real zeroes, because non-real zeroes occur in conjugate pairs.

Example 5.6.2. i) The polynomial $p(x) = x_1 \cdots x_n$ is real stable.

ii) Affine-linear polynomials $p(x) = \sum_{i=1}^n a_i x_i + b$ with $a_1, \dots, a_n, b \in \mathbb{R}$ and a_1, \dots, a_n all positive (or all negative) are real stable.

iii) The polynomial $p(x_1, x_2) = 1 - x_1 x_2$ is real stable. Namely, if $(a + bi, c + di)$ with $b, d > 0$ were a zero of p , then it evaluates to $1 - ac + bd - i(ad + bc)$. Its real part implies that either a and c are both positive or a and c are both negative. However, then the imaginary part $ad + bc$ cannot vanish.

We record an elementary property, that will become the starting point for the later viewpoint in terms of hyperbolic polynomials.

Lemma 5.6.3. *A polynomial $p \in \mathbb{C}[x]$ is stable if and only if for all $a, b \in \mathbb{R}^n$ with $b > 0$ the univariate polynomial $t \mapsto p(a + bt)$ is stable.*

Proof. If p is not stable, then there exists a point $z = a + bi$ with $b > 0$ and $p(z) = 0$. Hence, $t = i$ is zero of the univariate polynomial $t \mapsto p(a + bt)$, and thus that univariate polynomial is not stable.

Conversely, if for some $a, b \in \mathbb{R}^n$ with $b > 0$, the univariate polynomial $t \mapsto p(a + bt)$ is not stable, then it has a non-real root t^* with positive imaginary part. Hence, $p((a + b\Re(t^*)) + ib\Im(t^*)) = 0$, so that p is not stable. \square

An important class of stable polynomials is provided by determinantal polynomials of the following form.

Theorem 5.6.4 (Borcea, Bränden). *For positive semidefinite $d \times d$ -matrices A_1, \dots, A_n and a Hermitian $d \times d$ -matrix B , the polynomial*

$$p(x) = \det(x_1 A_1 + \dots + x_n A_n + B)$$

is real stable.

Note that the special case $n = 1$, $A_1 = I$ gives the normalized characteristic polynomial of $-B$.

Before the proof, recall the spectral decomposition of a Hermitian $n \times n$ matrix A ,

$$A = S^T D S = \sum_{i=1}^n \lambda_i v^{(i)} (v^{(i)})^T$$

with eigenvalues $\lambda_1, \dots, \lambda_n$, corresponding orthonormal eigenvectors $v^{(1)}, \dots, v^{(n)}$, $D = \text{diag}(\lambda_1, \dots, \lambda_n)$, and $S = (v^{(1)}, \dots, v^{(n)})$. If A is positive semidefinite, then it has a square root

$$A^{1/2} = \sum_{i=1}^n \sqrt{\lambda_i} v^{(i)} (v^{(i)})^T.$$

Observe that $A^{1/2} A^{1/2} = A$, and indeed $A^{1/2}$ is the unique matrix with this property. The definition of $A^{1/2}$ also implies that $A^{1/2}$ is Hermitian and has eigenvalues $\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}$, hence $A^{1/2}$ is positive definite.

Proof. To see that p has real coefficients, observe that the complex conjugate polynomial \bar{f} of the polynomial f (coefficientwise) satisfies $\bar{f} = f$. This follows from $\bar{A}_i = A_i^T$ and $\bar{B} = B^T$.

For stability, first consider the situation that A_1, \dots, A_n are positive definite, and let $a, b \in \mathbb{R}^n$ with $b > 0$. The positive definite matrix $P = \sum_{j=1}^n b_j A_j$ has a square root $P^{1/2}$. $P^{1/2}$ is positive definite and thus invertible. We obtain

$$p(a + tb) = \det P \det(tI + P^{-1/2} H P^{-1/2}), \quad (5.18)$$

where $H := B + \sum_{j=1}^n a_j A_j$ is Hermitian. The polynomial (5.18) is a constant multiple of a characteristic polynomial of a Hermitian matrix, which shows that all its roots are real. Hence, the univariate polynomial $t \mapsto p(a + tb)$ is stable for all $a, b \in \mathbb{R}^n$ with $b > 0$, so that p is stable by Theorem 5.6.3.

The general case of positive semidefinite matrices A_1, \dots, A_n follows from the complex-analytical Hurwitz' Theorem (with $U = \mathbb{R}_{>0}^n$) which is recorded without proof below. \square

Theorem 5.6.5 (Hurwitz). *Let $\{f_j\}_{j \in \mathbb{N}} \subset \mathbb{C}[x]$ be a sequence of polynomials, non-vanishing in a connected open set $U \subset \mathbb{C}^n$, and assume it converges to a function f uniformly on compact subsets of U . Then f is either non-vanishing on U or it is identically 0.*

Multivariate stable polynomials are often adequately approached through the more general hyperbolic polynomials. A homogeneous polynomial $p \in \mathbb{R}[x]$ is called *hyperbolic* in direction $e \in \mathbb{R}^n \setminus \{0\}$, if $p(e) \neq 0$ and for every $x \in \mathbb{R}^n$ the univariate function $t \mapsto p(x + te)$ has only real roots.

Example 5.6.6. (i) The polynomial $p(x) = x_1 \cdots x_n$ is hyperbolic in direction $(1, \dots, 1)$.

(ii) Denoting by X the real symmetric $n \times n$ -matrix (x_{ij}) and identifying it with a vector of length $n(n - 1)/2$, the polynomial $p(X) = \det X$ is hyperbolic in the direction of the unit matrix I_n . Namely, observe that the roots of $\det(X + tI)$ are just the additive inverses of the eigenvalues of the symmetric real matrix X .

(iii) The polynomial

$$p(x) = x_1^2 - \sum_{j=2}^n x_j^2$$

is hyperbolic in direction $(1, 0, \dots, 0) \in \mathbb{R}^n$.

The concept of hyperbolic polynomials is motivated by its applications in linear hyperbolic differential equations.

Example 5.6.7. Let $C_\infty(\mathbb{R}^n)$ be the class of infinitely-differentiable complex-valued functions in n real variables. For homogeneous $p \in \mathbb{C}[x] = \mathbb{C}[x_1, \dots, x_n]$, let $D(p)$ be the linear differential operator which substitutes a variable x_i by $(\partial/\partial x_i)$. For example, for $p(x_1, x_2) = x_1^2 - x_2^2$, we obtain

$$D[p](f) = \frac{\partial^2}{\partial x_1^2} f - \frac{\partial^2}{\partial x_2^2} f = 0. \quad (5.19)$$

Consider the topology of uniform convergence of all derivatives on compact sets. The differential equation $D[p](f) = 0$ is called *stable under perturbations* (in direction e) if for all sequences (f_k) in $C_\infty(\mathbb{R}^n)$ with $D[p](f_k) = 0$, (f_k) converges to zero on \mathbb{R}^n whenever the restriction of f_k to the halfspace $H_e = \{x \in \mathbb{R}^n : x^T e \geq 0\}$ converges to zero. By a result of Gårding, the differential equation $D[p](f) = 0$ is stable under perturbation in direction e if and only if the polynomial p is hyperbolic.

Since the specific example polynomial p is hyperbolic in direction $(1, 0)$ (see Example 5.6.6(iii)), the differential equation (5.19) is stable under perturbations in direction $(1, 0)$.

Theorem 5.6.8. *A homogeneous polynomial $p \in \mathbb{R}[x]$ is real stable if and only if p is hyperbolic with respect to every point in the positive orthant $\mathbb{R}_{>0}^n$.*

Note that this is a homogeneous version of Lemma 5.6.3.

Proof. Let $p \in \mathbb{R}[x]$ be a homogeneous polynomial which is not hyperbolic with respect to some $e \in \mathbb{R}_{>0}^n$. Then there exists $x \in \mathbb{R}^n$ such that the univariate polynomial $f(x + te) \in \mathbb{R}[t]$ has a non-real root t^* . Since non-real roots of real polynomials come in conjugate pairs, we can assume that $\Im(t^*) > 0$. Hence, $x + t^*e$ is a non-real root of p whose components have positive imaginary parts, and thus p is not stable.

Conversely, consider a homogeneous polynomial $p \in \mathbb{R}[x]$ which is not stable. Then it has a root $a \in \mathcal{H}^n$. Setting $x_j = \Re(a_j)$ and $e_j = \Im(a_j)$ for all j gives a polynomial $p(x + te) \in \mathbb{R}[t]$ which has the root $t = i$. Hence, p is not hyperbolic with respect to $e = (e_1, \dots, e_n) \in \mathbb{R}_{>0}^n$. \square

Definition 5.6.9. Let $p \in \mathbb{R}[x]$ be hyperbolic in direction e . Then the *hyperbolicity cone of p* with respect to e is

$$C(p, e) = \{x \in \mathbb{R}^n : p(x + te) \in \mathbb{R}[t] \text{ has only negative roots}\}.$$

We will see below that $C(p, e)$ is indeed an “open” convex cone, which then justifies the notion (or, more precisely, the topological closure $\overline{C(p, e)}$ is a convex cone). Also note that e is contained in $C(p, e)$.

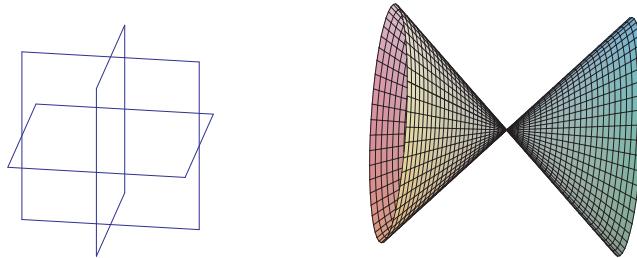


Figure 5.9: Any of the eight orthants is a hyperbolicity cone of $p(x, y, z) = xyz$, and the polynomial $p(x, y, z) = x^2 - y^2 - z^2$ has two hyperbolicity cones.

Example 5.6.10. We consider the polynomials from Example 5.6.6, see also the illustrations for (i) and (iii) in Figure 5.9.

(i) For $p(x) = \prod_{i=1}^n x_i$ and $e = (1, \dots, 1)$, the hyperbolicity cone $C(p, e)$ is the positive orthant.

(ii) For $p(X) = \det X$ and $e = I_n$, the hyperbolicity cone $C(p, e)$ is the set of positive definite $n \times n$ -matrices.

(iii) For $p(x) = x_1^2 - \sum_{j=2}^n x_j^2$ and $e = (1, 0, \dots, 0)$, note that the polynomial $(x_1 - t)^2 - \sum_{j=2}^n x_j^2$ has only negative roots if and only if $x_1 > 0$ and $x_1^2 > \sum_{j=2}^n x_j^2$. Hence, $C(p, e)$ is the *Lorentz cone*

$$\left\{ x \in \mathbb{R}^n : x_1 > \sqrt{x_2^2 + \dots + x_n^2} \right\}.$$

Theorem 5.6.11. *If $p \in \mathbb{R}[x]$ is hyperbolic in direction e , then $C(p, e)$ is the connected component of $\{x \in \mathbb{R}^n : p(x) \neq 0\}$ which contains e .*

Proof. Let C be the connected component of $\{x : p(x) \neq 0\}$ which contains e . For a given point $x \in C$, consider a path $\gamma : [0, 1] \rightarrow C$ with $\gamma(0) = e$ and $\gamma(1) = x$. The roots of the polynomial $p(\gamma(s) + te)$ vary continuously with s , and since the path γ is contained in C , the zeroes remain all negative. Hence, the polynomial p is also hyperbolic in direction e .

Conversely, if p is hyperbolic in direction e and $x \in C(p, e)$, it suffices to construct a path from e to x inside $C(p, e)$. Writing

$$p(sx + (1-s)e + te) = p(sx + ((1-s) + t)e)$$

shows that, for $s \in [0, 1]$, the roots of $t \mapsto p(sx + (1-s)e + te)$ are just by the value of $1-s$ smaller than the roots of $t \mapsto p(sx + te)$. Hence, for every $s \in [0, 1]$, the polynomial $p(sx + (1-s)e + te)$ has only negative roots, and thus $sx + (1-s)e \in C(p, e)$. \square

The notion of a hyperbolic polynomial is tied to the following convexity result.

Theorem 5.6.12. *Let $p \in \mathbb{R}[x]$ is hyperbolic in direction e .*

1. *For any $e' \in C(p, e)$, the polynomial p is hyperbolic in direction e' and $C(p, e) = C(p, e')$.*
2. *The set $C(p, e)$ is open, and its topological closure $\overline{C(p, e)}$ is a convex cone.*

Proof. Let $x \in \mathbb{R}^n$. We have to show that $t \mapsto p(x + te')$ is real stable.

We claim that for fixed $\beta \geq 0$, any root $t \in \mathbb{C}$ of

$$g : t \mapsto p(\beta x + te' + ie) \tag{5.20}$$

satisfies $\Im(t) < 0$.

Case $\beta = 0$: First observe that $t = 0$ is not a root of g (since $p(e) \neq 0$). And, by homogeneity, any non-zero root t of g gives $p(e' + \frac{1}{t}ie) = 0$, so that $e' \in C(p, e)$ implies that $\frac{1}{t}i$ is real and negative, that is, $t = \gamma i$ for some $\gamma < 0$.

Case $\beta > 0$: Assume that for some $\beta_0 > 0$, g has a root t with $\Im(t) \geq 0$. By continuity, there exists some $\beta \in (0, \beta_0]$ such that (5.20) has a real root t . Hence, i is a root of

$s \mapsto p(\beta x + te' + se) = p((\beta x + te') + se)$, which contradicts the hyperbolicity of p in direction s .

By the claim and writing p as polynomial $p(x, t)$ in x and t , the roots of the polynomial $t \mapsto \varepsilon^{\deg g} \cdot p(x/\varepsilon, t/\varepsilon) = p(\beta x + te' + \varepsilon ie)$ have negative imaginary parts (for any $\varepsilon > 0$). Using the continuous dependence of the roots of a polynomial on its coefficients, the roots of the real polynomial $t \mapsto p(\beta x + te')$ have non-positive imaginary parts. Since non-real roots of univariate real polynomials occur in conjugate pairs, all roots of $t \mapsto p(\beta x + te')$ must be real.

Now, since $e' \in C(p, e)$ and $e \in C(p, e')$, Theorem 5.6.11 implies that $C(p, e) = C(p, e')$.

In the proof of Theorem 5.6.11, we have seen that for any $x \in C(p, e)$, the line segment connecting e and x is contained in $C(p, e)$ as well. Now let y be another point in $C(p, e)$. By the first part, we have $C(p, e) = C(p, y)$ and hence $x \in C(p, y)$. Therefore the line segment connecting x and y is contained in $C(p, y) = C(p, e)$. Furthermore, $C(p, e)$ is clearly an open set. It is closed under multiplication with a positive scalars and hence $C(p, e)$ is a convex cone. \square

Exercises

1. Is the polynomial $p(x, y) = x^2 + y^2$ stable?
2. For a stable polynomial $p \in \mathbb{C}[x]$, the following operations preserve stability:
 - (a) Diagonalization: $p \mapsto p(x)|_{x_j=x_k}$ for $\{j, k\} \subseteq \{1, \dots, n\}$.
 - (b) Specialization: $p \mapsto p(a, x_2, \dots, x_n)$, $a \in \mathbb{C}$, $\Im(a) \geq 0$.
 - (c) Inversion: $p \mapsto x_1^d f(-x_1^{-1}, x_2, \dots, x_n)$, if $\deg_{x_1} f = d$.
 - (d) Differentiation: $p \mapsto \partial_{x_1} p(x_1, \dots, x_n)$.
3. Show that a polynomial $p \in \mathbb{R}[x] = \mathbb{R}[x_1, \dots, x_n]$ is stable if and only if its homogenization $p_h \in \mathbb{R}[x_0, x_1, \dots, x_n]$ is hyperbolic in direction $(0, e)$ for all $e \in \mathbb{R}_{>0}^n$.
4. A polynomial $q \in \mathbb{R}[x] = \mathbb{R}[x_1, \dots, x_n]$ is called a *real zero polynomial* if for any $x \in \mathbb{R}^n$ the univariate polynomial $t \mapsto q(tx) = q(tx_1, \dots, tx_n)$ has only real roots.
 - (a) Let $p \in \mathbb{R}[x_1, \dots, x_n]$ be a hyperbolic polynomial of degree d with respect to $e = (1, 0, \dots, 0)$, and let $p(e) = 1$. Then the polynomial $q(x_2, \dots, x_n) = p(1, x_2, \dots, x_n)$ is a real zero polynomial of degree at most d and which satisfies $q(0) = 1$.
 - (b) Let $q \in \mathbb{R}[x_1, \dots, x_n]$ be a real zero polynomial of degree d and $q(0) = 1$. Then the polynomial (defined for $x \neq 0$ and extended continuously to \mathbb{R}^n)

$$p(x_1, \dots, x_n) = x_1^d q\left(\frac{x_2}{x_1}, \dots, \frac{x_n}{x_1}\right) \quad (x_1 \neq 0)$$

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

5.7 Notes

As standard references in real algebraic and semialgebraic geometry, we list the monograph of Bochnak, Coste, and Roy [16] as well as the computationally-oriented one of Basu, Pollack, and Roy [5].

Sturm sequences have been discovered by the French mathematician Jacques Charles François Sturm in 1829 [137], and Hermite introduced the Hermite form in 1853 [58]. René Descartes formulated the Rule of Signs as early as in 1637 [30]. For further information on eigenvalue techniques see [26, 100]. The lecture notes of Sturmfels [138] capture many of the state-of-the-art techniques and results.

The inductive proof of Farkas' Lemma goes back to Kuhn [74], Bröcker's Theorem mentioned in the Exercises of Section 5.4 is published in [19]. The algorithm for the cylindrical algebraic decomposition is due to Collins [24], a detailed treatment can be found in the monograph of Basu, Pollack and Roy [5].

For a comprehensive treatment of stability of polynomials and related topics we refer to Rahman and Schmeisser [111]. The term spectrahedron was introduced by Ramana and Goldman [112] who also showed exposedness of the faces. Much material on spectrahedra can be found in [14]. Theorem 5.5.7 was shown in [53], and it was applied by Lewis, Parrilo and Ramana to prove the Lax conjecture that provides a determinantal representation of ternary hyperbolic polynomials [81]. The case of k -ellipses is studied in [94]. The facial structure of LMI-representable sets was studied by Netzer, Plaumann and Schweighofer [93] and provides Theorem 5.5.10, the special case of spectrahedra also treated in Exercise 5 of Section 5.5 was already given by Ramana and Goldman [112]. The disproof of the Helton-Nie conjecture was given by Scheiderer [119].

Hyperbolic polynomials have been pioneered by Gårding [41, 42]. See Wagner [142] for a survey on recent developments in stable polynomials and Pemantle [101] for the use of hyperbolic and stable polynomials in combinatorics and probability. Theorem 5.6.4 was proven by Borcea and Brändén [18, Proposition 2.4]. Hurwitz' Theorem 5.6.5 with a proof can be found in Theorem 1.3.8 in Rahman and Schmeisser [111].

Chapter 6

Positive polynomials

A polynomial in $\mathbb{R}[x_1, \dots, x_n]$ that is a sum of squares of other polynomials is nonnegative on \mathbb{R}^n . During Minkowski's 1885 public defense of his Ph.D. thesis on quadratic forms in Königsberg, he conjectured that there exist nonnegative real forms (real homogeneous polynomials) which cannot be written as a sum of squares of real forms. His ‘opponent’ at the defense was Hilbert. By the end of the defense, Hilbert declared that he was convinced by Minkowski's exposition that already when $n = 3$, there may well be remarkable ternary forms “which are so stubborn as to remain positive without allowing to put up with a representation as sums of squares of forms”:

“Es fiel mir als Opponent die Aufgabe zu, bei der öffentlichen Promotion diese These anzugreifen. Die Disputation schloss mit der Erklärung, ich sei durch seine Ausführungen überzeugt, dass es wohl schon im ternären Gebiete solche merkwürdigen Formen geben möchte, die so eigensinnig seien, positiv zu bleiben, ohne sich doch eine Darstellung als Summe von Formenquadraten gefallen zu lassen.” [59]

Hilbert proved Minkowski's conjecture in 1888 (see Theorem 6.2.3).

Among the 23 problems which Hilbert presented during¹ his legendary lecture at the 1900 International Congress of Mathematicians in Paris, he asked whether every nonnegative polynomial $p \in \mathbb{R}[x_1, \dots, x_n]$ can be written as a finite sum of squares of real rational functions. This problem is widely known as *Hilbert's 17th problem*.

This classical theory and its more recent developments are cornerstones of modern treatments of optimization of polynomial functions. Algebraic certificates are of particular importance—they serve to “witness” that a certain polynomial is positive on a given set. A common set of witnesses for global positivity is the set $\Sigma[x] = \Sigma[x_1, \dots, x_n]$ of polynomials which can be written as a sum of squares of polynomials

$$\Sigma[x] := \left\{ \sum_{i=1}^k h_i^2 \text{ with } h_1, \dots, h_k \in \mathbb{R}[x], k \in \mathbb{N} \right\}. \quad (6.1)$$

¹Hilbert did not present all 23 in his lecture, only in the published version.

In this chapter we develop the key elements of the classical and of the modern results.

6.1 Nonnegative univariate polynomials

Let $\mathbb{R}[x]$ be the ring of real univariate polynomials. Theorem 2.5.1 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 6.1.1. *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 the set $\Sigma[x]$ of sums of squares of univariate polynomials.

Theorem 6.1.2. *A univariate polynomial of even degree is nonnegative if and only if it may be written as a sum of squares of univariate polynomials.*

Proof. We only need to show that a nonnegative univariate polynomial $p \in \mathbb{R}[x]$ is a sum of squares. The non-real roots of p occur in conjugate pairs, and as p is nonnegative, its real roots have even multiplicity. Thus $p = r^2 \cdot q \cdot \bar{q}$, 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. \square

We may characterize real polynomials that are nonnegative on an interval $I \subsetneq \mathbb{R}$. 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

$$q(x) := p\left(\frac{(b-a)x + (b+a)}{2}\right) \quad (6.2)$$

is nonnegative (respectively strictly positive) on $[-1, 1]$, and the same is true for half-open or open bounded intervals. 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. The (*d-th degree*) *Goursat transform* \tilde{p} of a polynomial p of degree at most d is

$$\tilde{p}(x) := (1+x)^d p\left(\frac{1-x}{1+x}\right) \in \mathbb{R}[x].$$

Then \tilde{p} is a polynomial of degree at most d . Applying the same Goursat transform to \tilde{p} yields the original polynomial p , multiplied by 2^d ,

$$(1+x)^d \tilde{p}\left(\frac{1-x}{1+x}\right) = (1+x)^d \left(1 + \frac{1-x}{1+x}\right)^d p\left(\frac{1 - \frac{1-x}{1+x}}{1 + \frac{1-x}{1+x}}\right) = 2^d p(x). \quad (6.3)$$

If $\deg \tilde{p} < d$ then the $\deg \tilde{p}$ -th degree Goursat transform of the polynomial p is a polynomial as well. The formula (6.3) implies that $(1+x)^{d-\deg \tilde{p}}$ divides p and that no higher power of $(1+x)$ 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 $\tilde{p} = 4x$, whose degree two Goursat transform is $4(1-x^2)$.

Lemma 6.1.3 (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 \tilde{p} is nonnegative on $[0, \infty)$.*
2. *p is strictly positive on $[-1, 1]$ if and only if \tilde{p} is strictly positive on $[0, \infty)$ and $\deg \tilde{p} = d$.*

Proof. Let $p \in \mathbb{R}[x]$ and consider the bijection $\varphi : (-1, 1] \rightarrow [0, \infty)$, $x \mapsto \frac{1-x}{1+x}$. Since $(1+x)^d$ is strictly positive on $(-1, 1]$, p is strictly positive on $(-1, 1]$ if and only if \tilde{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 $\deg \tilde{p} = d$ if and only if $1+x$ does not divide p , so that p does not vanish at -1 . \square

Given polynomials $f_1, \dots, f_m \in \mathbb{R}[x]$ and $I \subset \{1, \dots, m\}$ set $\underline{f}_I := \prod_{i \in I} f_i$, with the convention that $\underline{f}_\emptyset = 1$. The *preorder generated by f_1, \dots, f_m* is

$$P(f_1, \dots, f_m) := \left\{ \sum_{I \subset \{1, \dots, m\}} s_I \underline{f}_I \mid s_I \in \Sigma[x] \right\}.$$

Elements of this preorder are nonnegative on the basic semialgebraic set

$$S(f_1, \dots, f_m) = \{x \in \mathbb{R}^n : f_1(x) \geq 0, \dots, f_m(x) \geq 0\}. \quad (6.4)$$

Theorem 6.1.4 (Pólya and Szegő). *If a univariate polynomial $p \in \mathbb{R}[x]$ is nonnegative on $[0, \infty)$ then we have*

$$p = f + xg \quad \text{for some } f, g \in \Sigma[x], \quad (6.5)$$

with $\deg f, \deg xg \leq \deg p$. In particular, p lies in the preorder $P(x)$ generated by x .

Section 6.7 gives a multivariate version of the Pólya–Szegő Theorem.

Proof. Dividing p by its (necessary positive) leading coefficient and by any square factors (as in the proof of Theorem 6.1.2), we may assume that p is monic and square-free.

Suppose that p is irreducible. If p is linear, then, as its root is nonpositive, $p = \alpha + x$ with $\alpha \geq 0$, an expression of the form (6.5). If p is quadratic, it has no real roots and $p \in \Sigma[x]$ by Theorem 6.1.2, so $p = p + x \cdot 0$, again an expression of the form (6.5).

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 $f := f_1f_2 + x^2g_1g_2$ and $g := f_1g_2 + g_1f_2$ gives the desired expression (6.5) for p . \square

Example 6.1.5. The polynomial $p = x^5 + x^4 - 2x^3 - x^2 - x + 2 = (x+2)(x-1)^2(x^2+x+1)$ is nonnegative on $[0, \infty)$. The proof of the Pólya-Szegő Theorem gives the expression

$$p = (x-1)^2(2(x^2+1) + x^2 + x(2+x^2+1)),$$

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

Corollary 6.1.6. A univariate polynomial $p \in \mathbb{R}[x]$ is nonnegative on an interval $[a, b]$ if and only if it lies in the preorder $P(x-a, b-x)$. Moreover, there is an expression

$$p = f + (x-a)g + (b-x)h + (x-a)(b-x)k, \quad (6.6)$$

where $f, g, h, k \in \Sigma[x]$ and each term in this expression has degree at most $\deg(p)$.

An expression (6.6) for a polynomial p shows it lies in the preorder $P(x-a, b-x)$, and is called a *certificate of nonnegativity* for p on $[a, b]$.

Proof. Using the change of variables (6.2), we may assume that $[a, b] = [-1, 1]$. Since elements of the preorder $P(x+1, 1-x)$ are nonnegative on $[-1, 1]$, we only need to show that polynomials which are nonnegative on $[-1, 1]$ lie in this preorder.

Let p be a degree d polynomial that is nonnegative on $[-1, 1]$. By Goursat's Lemma, \tilde{p} is nonnegative on $[0, \infty)$ and by the Pólya-Szegő Theorem, there are polynomials $f, g \in \Sigma[x]$ with

$$\tilde{p} = f + xg,$$

where both f and xg have degree at most $x \deg \tilde{p}$. Since $\tilde{x} = 1 - x$, if we apply the d -th degree Goursat transform to this expression, we obtain

$$2^d p = (x+1)^{d-\deg f} \tilde{f} + (1-x)(x+1)^{d-1-\deg g} \tilde{g}. \quad (6.7)$$

The corollary follows as the Goursat transform is multiplicative, $\widetilde{hk} = \widetilde{h}\widetilde{k}$, so that if $f, g \in \Sigma[x]$ then $\tilde{f}, \tilde{g} \in \Sigma[x]$, and it is additive, if $\deg(h) \geq \deg(k)$ with $\deg(h) = \deg(h+k)$, then

$$\widetilde{h+k} = \widetilde{h} + (x+1)^{\deg(h)-\deg(k)} \widetilde{k}.$$

Absorbing even powers of $(x+1)^{d-\deg f}$ into \tilde{f} , and the same for \tilde{g} , the expression (6.7) shows that $p \in P(x+1, 1-x)$. \square

Example 6.1.7. The polynomial $p = 4 - 4x + 8x^2 - 4x^4$ is nonnegative on $[-1, 1]$. Indeed,

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

Exercises

1. Use the Goursat transform and the algorithm in the proof of the Pólya and Szegő Theorem 6.1.4 to compute a certificate that

$$(x^2 - 9)(x^2 - 4) = x^4 - 13x^2 + 36$$

is nonnegative 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 , $0 \leq k \leq d$ 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 $x \in [-1, 1]$, $\sum_{k=0}^d B_k^d(x) \geq 0$ 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 nonnegative on $[-1, 1]$ is 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 a_k nonnegative 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) $p = 3(1+x)^2 - 3(1+x)(1-x) + (1-x)^2$.

6.2 Positive polynomials and sums of squares

As we saw in Section 6.1, 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 $n > 1$ variables.

For example, as we will see in more detail in Chapter 7, deciding the nonnegativity of a given polynomial $p \in \mathbb{R}[x_1, \dots, x_n]$ 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 6.2.1. *The polynomial $p = 1 - 3x^2y^2 + x^2y^4 + x^4y^2 \in \mathbb{R}[x, y]$ is nonnegative on \mathbb{R}^2 , but it cannot be written as a sum of squares.*

See Figure 6.1 for an illustration of the graph $z = p(x, y)$ of the Motzkin polynomial.

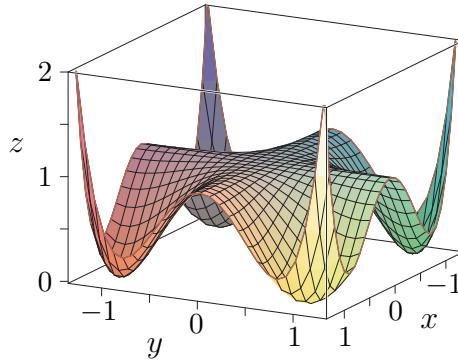


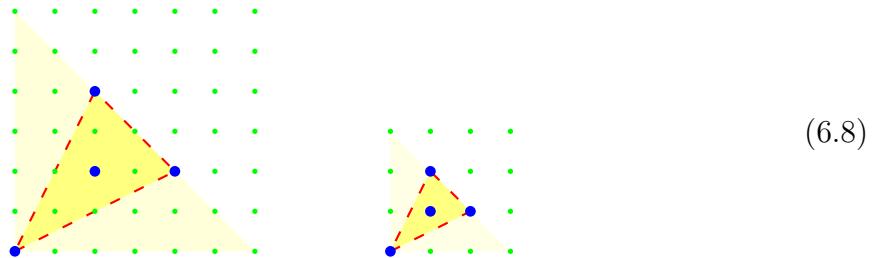
Figure 6.1: The Motzkin polynomial. Its zeroes are exactly at the four points $(\pm 1, \pm 1)$.

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 show that p is not a sum of squares by considering its Newton polygon $\text{New}(p)$ and support, shown below on the left. (Newton polytopes are defined in A.1.1 of the Appendix.)



Suppose that p is a sum of squares, so there exist polynomials $p_1, \dots, p_k \in \mathbb{R}[x, y]$ with

$$p = p_1^2 + p_2^2 + \cdots + p_k^2. \quad (6.9)$$

By Exercise 2, we have that $2 \text{New}(p_i) \subset \text{New}(p)$ for each $i = 1, \dots, k$. Thus each summand p_i has Newton polytope contained in the polygon on the right of (6.8). 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 (6.9) implies that $-3 = \sum_i b_i^2$, which is impossible. \square

For $d \geq 0$, let $\mathbb{R}_d[x_1, \dots, x_n]$ be the space of polynomials in x_1, \dots, x_n that are homogeneous of degree d , called *d-forms*. For $n \geq 1$ and $d \geq 2$ let

$$\mathcal{P}_{n,d} := \{p \in \mathbb{R}_d[x_1, \dots, x_n] \mid p \geq 0\}$$

denote the set of nonnegative polynomials of degree d in x_1, \dots, x_n and let

$$\Sigma_{n,d} := \{p \in \mathbb{R}_d[x_1, \dots, x_n] \mid p \text{ is a sum of squares}\} \subset \mathcal{P}_{n,d}$$

be its subset of polynomials that are sums of squares. Both 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 implies that if $p, q \in \mathcal{P}_{n,d}$ and $\alpha, \beta \in \mathbb{R}_{>0}$, then $\alpha p + \beta q \in \mathcal{P}_{n,d}$. For $\Sigma_{n,d}$, this is evident from its definition. Both cones $\mathcal{P}_{n,d}$ and $\Sigma_{n,d}$ are closed. For $\mathcal{P}_{n,d}$ this follows from its inequality description, and for $\Sigma_{n,d}$, this will be explained in Chapter 7, see Theorem 7.2.4.

For any polynomial $p \in \mathbb{R}[x_1, \dots, x_n]$ of degree at most d , its degree d homogenization,

$$\bar{p} := x_0^d p\left(\frac{x_1}{x_0}, \dots, \frac{x_n}{x_0}\right),$$

is a form of degree d . Homogenization $p \mapsto \bar{p}$ is a vector space isomorphism from the space of all polynomials in $\mathbb{R}[x_1, \dots, x_n]$ of degree at most d to the space of homogeneous polynomials in $\mathbb{R}[x_0, x_1, \dots, x_n]$ of degree d . The *dehomogenization* of a homogeneous polynomial $p \in \mathbb{R}[x_0, x_1, \dots, x_n]$ is the polynomial $p(1, x_1, \dots, x_n) \in \mathbb{R}[x_1, \dots, x_n]$. For $p \in \mathbb{R}[x_1, \dots, x_n]$ the dehomogenization of \bar{p} is p .

Lemma 6.2.2. *Let d be even and $p \in \mathbb{R}[x_1, \dots, x_n]$ be a polynomial of degree d .*

1. *p is nonnegative on \mathbb{R}^n if and only if \bar{p} is nonnegative on \mathbb{R}^{n+1} .*
2. *p is a sum of squares of polynomials if and only if \bar{p} is a sum of squares of forms of degree $\frac{d}{2}$.*

Proof. For the first statement, suppose that p is nonnegative on \mathbb{R}^n . For $a = (a_0, \dots, a_n) \in \mathbb{R}^{n+1}$ with $a_0 \neq 0$ we have $\bar{p}(a) = a_0^d p\left(\frac{a_1}{a_0}, \dots, \frac{a_n}{a_0}\right) > 0$, and the continuity of \bar{p} implies that if $a_0 = 0$ then $\bar{p}(a) \geq 0$. The converse follows by 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 $\bar{p} = \sum_{i=1}^k (x_0^{d/2} \cdot p_i(\frac{x_1}{x_0}, \dots, \frac{x_n}{x_0}))^2$ is a representation as sum of squares of forms of degree $\frac{d}{2}$. The converse follows by dehomogenizing. \square

The following classical theorem of 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 6.2.3. *Let $n \geq 2$ and d even. We have that $\Sigma_{n,d} = \mathcal{P}_{n,d}$ in exactly the following cases.*

1. $n = 2$ (binary forms).
2. $d = 2$ (quadratic forms).
3. $n = 3, d = 4$ (ternary quartics).

Proof. The first assertion is a consequence of Theorem 6.1.2, as dehomogenizing a nonnegative binary form ($n = 2$) gives a nonnegative univariate polynomial.

For $d = 2$, note that every quadratic form f can be written as

$$f(x) = f(x_1, \dots, x_n) = x^T A x$$

with A a symmetric matrix. Then f is nonnegative if and only if A is positive semidefinite. Employing a Choleski decomposition, A can be written as $A = V^T V$ and thus

$$f(x) = x^T A x = x^T V^T V x = (Vx)^T (Vx) = \|Vx\|^2,$$

which is a sum of squares of linear forms. Hence, $\Sigma_{n,2} = \mathcal{P}_{n,2}$.

The case $(n, d) = (3, 4)$ is Theorem 6.2.4 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. Homogenizing the Motzkin polynomial $p \in \mathbb{R}[x, y]$ from Theorem 6.2.1 to degree $d \geq 6$,

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

we have that $\bar{p} \in \mathcal{P}_{n,d} \setminus \Sigma_{n,d}$, which shows that $\mathcal{P}_{n,d} \setminus \Sigma_{n,d} \neq \emptyset$ for $n \geq 3$ and $d \geq 6$. For the cases $(n, 4)$ with $n \geq 4$ the difference follows from the nonnegative quartic form $w^4 + x^2y^2 + x^2z^2 + y^2z^2 - 4wxyz$ of Exercise 3. \square

Theorem 6.2.4. *Every nonnegative ternary quartic can be written as a sum of squares of quadratic forms.*

The proof uses the following slightly technical lemma.

Lemma 6.2.5. *For any nonzero nonnegative ternary quartic form p , there is a nonzero quadratic form q such that $p - q^2$ remains nonnegative.*

Proof. Suppose first that p does not vanish on the real projective plane $\mathbb{P}_{\mathbb{R}}^2$, so that if $(x, y, z) \neq (0, 0, 0)$, then $p(x, y, z) > 0$. Let α^2 be the minimum value of p on the unit sphere $x^2 + y^2 + z^2 = 1$. As p is a homogeneous quartic, $p(x, y, z) \geq \alpha^2(x^2 + y^2 + z^2)^2$ for any $(x, y, z) \in \mathbb{R}^3$, and we may set $q := \alpha(x^2 + y^2 + z^2)$.

If p vanishes on $\mathbb{P}_{\mathbb{R}}^2$, we may assume that $p(1, 0, 0) = 0$. Expanding p as a polynomial in x ,

$$p = a_0 x^4 + a_1 x^3 + a_2 x^2 + a_3 x + a_4,$$

the coefficient $a_i \in \mathbb{R}[y, z]$ is a form of degree i . Then $a_0 = p(1, 0, 0) = 0$ and also $a_1 = 0$, for otherwise p takes negative values near $(1, 0, 0)$. Thus p has the form,

$$p = x^2 f(y, z) + 2xg(y, z) + h(y, z), \quad (6.10)$$

where f, g, h are homogeneous of degrees 2, 3, 4. As p is nonnegative near $(1, 0, 0)$ and when $x = 0$, both f and h are nonnegative. If $f = 0$, then as before $g = 0$ and $p = h$ is a nonnegative binary quartic, which is a sum of squares of quadratics, and we let q be one of those quadratics.

Suppose now that $f \neq 0$. For $(y, z) \in \mathbb{R}^2$, the quadratic (6.10) has either two complex zeroes or a single real zero of multiplicity two. Thus its discriminant $g^2 - fh$ is nonpositive and vanishes only if p has a real zero. Note that $pf = (xf + g)^2 + (fh - g^2)$ with both summands nonnegative.

Suppose that $(1, 0, 0)$ is the only zero of p in $\mathbb{P}_{\mathbb{R}}^2$. Then $fh - g^2$ is strictly positive on $\mathbb{R}^2 \setminus \{(0, 0)\}$. If f is irreducible over \mathbb{R} , then it is also strictly positive on $\mathbb{R}^2 \setminus \{(0, 0)\}$. Let α^2 be the minimum on the unit circle of the positive homogeneous function $(fh - g^2)/f^3$ of degree zero. Then $fp \geq (xf + g)^2 \geq \alpha^2 f^3$ and so $p \geq (\alpha f)^2$, and we may set $q := \alpha f$.

If f is reducible over \mathbb{R} , then it is the square of a linear form f_1 . Since $fh - g^2 = f_1^2 h - g^2$ is nonnegative, g vanishes at the real zeroes of f_1 , and so it is divisible by f_1 . Writing $g = f_1 g_1$, where $g_1 \in \mathbb{R}[y, z]$ is a quadratic form, we have $pf \geq (xf + g)^2 = f(xf_1 + g_1)^2$, and thus $p \geq (xf_1 + g_1)^2$, so we may take $q = xf_1 + g_1$.

Now suppose that p has another zero in $\mathbb{P}_{\mathbb{R}}^2$, which we may assume is at the point $(0, 1, 0)$. Then the decomposition (6.10) simplifies to

$$p = x^2 f(y, z) + 2xzg(y, z) + z^2 h(y, z), \quad (6.11)$$

where f, g, h are quadratic forms in $\mathbb{R}[y, z]$. As before, both f and h are nonnegative, as is $fh - g^2$. If f (or h) has a zero, then it is the square of a linear form, and the arguments of the previous paragraph give the desired quadratic form q .

We are left now with the case when both f and h are irreducible nonnegative quadratic forms, and therefore are strictly positive on $\mathbb{R}^2 \setminus \{(0, 0)\}$. Suppose in addition that $fh - g^2$ is similarly strictly positive. Let α^2 be the minimum of the positive rational function $\frac{fh-g^2}{f(y^2+z^2)}$ of degree zero on the unit circle. The decomposition $fp = (xf + zg)^2 + z^2(fh - g^2)$ implies that

$$fp \geq z^2(fh - g^2) \geq \alpha^2 z^2(y^2 + z^2)f,$$

and therefore $p \geq \alpha^2 z^4$, and so we may set $q := \alpha z^2$.

Finally, suppose that $fh - g^2$ vanishes at $(b, c) \neq (0, 0)$. Let $a := -g(b, c)/f(b, c)$ and define

$$p^*(x, y, z) := p(x + az, y, z) = x^2 f + 2xz(g + af) + z^2(h + 2ag + a^2 f). \quad (6.12)$$

Observe that $f(h + 2ag + a^2 f)$ vanishes at the point $(y, z) = (b, c)$. Thus the coefficient of z^2 in (6.12) has a zero, and previous arguments give the desired quadratic form q . \square

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 6.2.4. Given a form $p \in \mathcal{P}_{3,4}$, write $p = s_1 + \cdots + s_k$ as a sum of finitely many extremal forms s_1, \dots, s_k . Lemma 6.2.5 gives 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 p is a sum of squares. \square

Hilbert in fact showed that every ternary quartic is a sum of at most *three* squares, but all known proofs of this refinement are substantially more involved.

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 ,

$$\text{disc}(A) = \text{disc}(\chi_A(t)) = (-1)^{\binom{n}{2}} \text{Res}_t(\chi_A, \chi'_A). \quad (6.13)$$

Hence, by Example 2.1.5 in Chapter 2,

$$\text{disc}(A) = \prod_{i < j} (\lambda_i - \lambda_j)^2,$$

where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of A . If A is real and symmetric then its eigenvalues are all real and so $\text{disc}(A)$ is nonnegative. Note that (6.13) expresses $\text{disc}(A)$ as a homogeneous polynomial of degree $n(n-1)$ in the coefficients of A . In light of Theorem 6.2.3 it is remarkable that the nonnegative polynomial $\text{disc}(A)$ can be written as a sum of squares in the entries of A .

Theorem 6.2.6 (Ilyushechkin). *Let $A = (a_{ij})$ be a symmetric $n \times n$ -matrix with indeterminates a_{ij} . Then $\text{disc}(A)$ is a sum of squares of polynomials in the a_{ij} .*

For a matrix $A \in \mathbb{R}^{n \times n}$ write $\text{vec}(A)$ for the vector in \mathbb{R}^{n^2} obtained by arranging the elements of A in a single column. Let $A^* \in \mathbb{R}^{n^2 \times n}$ be the matrix whose i -th column is $\text{vec}(A^{i-1})$, for $1 \leq i \leq n$. When $n \geq 2$, the matrix A^* is not square. Theorem 6.2.6 follows from an explicit representation of $\text{disc}(A)$ as a sum of squares.

Lemma 6.2.7. *For a symmetric $n \times n$ -matrix we have*

$$\text{disc}(A) = \sum_{I \in \binom{[n^2]}{n}} (\det A_I^*)^2,$$

where A_I^* is the submatrix of A^* formed by the rows with indices I .

We remark that this notation will reappear in the treatment of Plücker coordinates in Section 10.1.

If $A = \text{diag}(\lambda_1, \dots, \lambda_n)$ 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 $(\det(A_I^*))^2 = \text{disc}(A)$, 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 $\lambda_1, \dots, \lambda_n$ of A , we have

$$\begin{aligned} \text{disc}(A) &= \det \begin{pmatrix} 1 & \cdots & 1 \\ \lambda_1 & \cdots & \lambda_n \\ \vdots & \ddots & \vdots \\ \lambda_1^{n-1} & \cdots & \lambda_n^{n-1} \end{pmatrix} \det \begin{pmatrix} 1 & \lambda_1 & \cdots & \lambda_1^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_n & \cdots & \lambda_n^{n-1} \end{pmatrix} \\ &= \det \begin{pmatrix} 1 & p_1 & p_2 & \cdots & p_{n-1} \\ p_1 & p_2 & \ddots & \ddots & p_n \\ p_2 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & p_{n-3} \\ p_{n-1} & p_n & \cdots & p_{n-3} & p_{n-2} \end{pmatrix}, \end{aligned} \tag{6.14}$$

where $p_k = \lambda_1^k + \cdots + \lambda_n^k$ is the k th Newton power sum of the eigenvalues of A . This is the trace, $\text{Tr}(A^k)$ 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

$$\text{Tr}(BC) = \sum_{k,l=1}^n B_{k,l} C_{l,k} = \sum_{k,l=1}^n B_{k,l} C_{k,l}.$$

Letting $B = C = A$, we see that the (i, j) -entry of $(A^*)^T A^*$ is

$$\sum_{k,l=1}^n A_{k,l}^{i-1} A_{k,l}^{j-1} = \text{Tr}(A^{i-1} A^{j-1}) = p_{i+j-2},$$

which is the (i, j) -entry in the Hankel matrix (6.14). Thus $\text{disc}(A) = \det((A^*)^T A^*)$, and the formula of the lemma follows from the Cauchy-Binet formula for the expansion of the determinant $\det((A^*)^T A^*)$. \square

Example 6.2.8. For the symmetric matrix

$$A = \begin{pmatrix} a & c \\ c & b \end{pmatrix} \text{ we have } A^* = \begin{pmatrix} 1 & a \\ 0 & c \\ 0 & c \\ 1 & b \end{pmatrix}$$

and thus from Lemma 6.2.7 we have the sum of squares representation

$$\text{disc}(A) = c^2 + c^2 + (b-a)^2 + 0 + c^2 + c^2 = 4c^2 + (b-a)^2.$$

Exercises

1. Suppose that $p = \sum_{i=1}^k p_i^2$ is a sum of squares of univariate polynomials $p_1, \dots, p_k \in \mathbb{R}[x]$. Show that if at least one of the polynomials p_i is non-zero then p is non-zero and $\deg p = 2 \max\{\deg p_i \mid 1 \leq i \leq k\}$.
2. Suppose that $p = \sum_{i=1}^k p_i^2$ is a sum of squares of multivariate polynomials $p_1, \dots, p_k \in \mathbb{R}[x_1, \dots, x_n]$. 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 $2w \cdot \beta_i \leq w \cdot \alpha$, and there is some i for which this is an equality. Conclude that

$$\text{New}(p) = \text{conv}\left(\bigcup_{i=1}^k 2 \cdot \text{New}(p_i)\right),$$

Where $\text{New}(p)$ is the Newton polytope of the polynomial f .

3. Show that $w^4 + x^2y^2 + x^2z^2 + y^2z^2 - 4wxyz$ is nonnegative but not a sum of squares.
4. Let $d \geq 2$ be even, $p = \sum_{i=1}^n c_i x_i^d + bx^\beta$ with positive coefficients c_i and β a strictly positive integer vector satisfying $\sum_{i=1}^n \beta_i = d$. Show that p is non-negative if and only if

$$\begin{cases} b \geq -\Theta_f, & \text{if all coordinates of } \beta \text{ are even,} \\ |b| \leq \Theta_f, & \text{else,} \end{cases}$$

where $\Theta = \prod_{i=1}^n \left(\frac{2\beta_i}{c_i}\right)^{c_i/2}$ is the *circuit number* of p .

5. 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

6.3 Hilbert's 17th problem

At the beginning of the chapter, we have already stated Hilbert's 17th problem, which has strongly influenced the developments in real algebraic geometry since then. In Artin's and Schreiers's solution of the problem, the theory of ordered fields plays a central role. In this section, we develop some central aspects of the theory of ordered fields and will use them to present the solution to Hilbert's 17th problem. The concepts developed for the solution to Hilbert's 17th problem will also be crucial for later sections.

A field \mathbb{K} together with a total order \leq is called an *ordered field* if it satisfies the two conditions

1. $a \leq b$ implies $a + c \leq b + c$.
2. $a \leq b$ and $0 \leq c$ imply $ac \leq bc$.

Example 6.3.1. The fields \mathbb{Q} and \mathbb{R} with their natural order provide ordered fields.

As usual we can write shortly $a < b$ if $a \leq b$ and $a \neq b$. The first property in the definition of an ordered fields implies

$$a \leq b \iff b - a \geq 0.$$

As a consequence, any order relation \leq on a field \mathbb{K} can be unambiguously expressed in terms of its nonnegative elements

$$P = \{a \in \mathbb{K} : a \geq 0\}.$$

Clearly, P satisfies the conditions

1. $P + P \subset P$ and $PP \subset P$,
2. $P \cap -P = \{0\}$,
3. $P \cup -P = \mathbb{K}$.

Conversely, if a given subset P of a field \mathbb{K} satisfies these three conditions, then the relation

$$a \leq_P b : \iff b - a \in P$$

defines an order on \mathbb{K} . As a consequence of this one-one-correspondence, we denote any subset P of a field \mathbb{K} satisfying the three conditions an *order* on \mathbb{K} .

As we have already encountered at the beginning of the section, sums of squares play a central role in the context of non-negativity. A basic building block is provided by the sums of squares of elements in \mathbb{K} . In any ordered field \mathbb{K} , the set of sums of squares clearly satisfies closure under addition and under multiplication as well. Moreover, the set \mathbb{K}^2 is contained in the set of sums of squares. This suggests the following definition.

Definition 6.3.2. A subset P of a field \mathbb{K} is called a *(quadratic) preorder* if it satisfies the conditions

1. $P + P \subset P$ and $PP \subset P$,
2. $a^2 \in P$ for every $a \in \mathbb{K}$.

Theorem 6.3.3. Let P be a preorder on a field \mathbb{K} which does not contain -1 . Then there exists an order P' of \mathbb{K} such that $P \subset P'$.

Proof. Let P' be a maximal preorder of \mathbb{K} containing P but not containing -1 ; the existence of P' follows from Zorn's Lemma. We show that P' is an order on \mathbb{K} . By Exercise 5, we have $P' \cap -P' = \{0\}$, hence it remains to prove $\mathbb{K} = P' \cup -P'$. For $a \in \mathbb{K} \setminus (-P')$, Exercise 6 implies that $P'' = P' + aP'$ is a preorder of \mathbb{K} with $-1 \notin P''$. The maximality of P' gives $P' = P'' = P' + aP'$, whence $a \in P'$ and altogether $\mathbb{K} = P' \cup -P'$. \square

A field \mathbb{K} is called *real* if $-1 \notin \sum \mathbb{K}^2$.

Theorem 6.3.4 (Artin-Schreier). *A field \mathbb{K} is real if and only if it has an order. And \mathbb{K} has a unique order if and only $\sum \mathbb{K}^2$ is an order.*

Proof. If \mathbb{K} is real, then Theorem 6.3.3 allows to extend the preorder $\sum \mathbb{K}^2$ to an order. For the converse direction, assume that \mathbb{K} has an order P . Since P is closed under multiplication, all squares and thus all elements in $\sum \mathbb{K}^2$ are contained in P . Since $P \cup -P = \{0\}$, the element -1 cannot be contained in $\sum \mathbb{K}^2$, which means that \mathbb{K} is real.

Since $\sum \mathbb{K}^2$ is a preorder, for the second statement it suffices to show that any preorder P of \mathbb{K} with $-1 \notin P$ satisfies

$$P = \bigcap \{P^* : P^* \text{ order on } \mathbb{K} \text{ with } P \subseteq P^*\}. \quad (6.15)$$

The inclusion “ \subset ” is clear. For the converse one, let $a \in \mathbb{K} \setminus P$. By Exercise 6, $P' = P - aP$ is a preorder with $-1 \notin P'$. Since $-a \in P'$, Lemma 6.3.3 allows to extend P' to an order P^* on \mathbb{K} with $-a \in P^*$. Observing $a \neq 0$, we can conclude $a \notin P^*$.

Since the set $\sum \mathbb{K}^2$ is contained in every order on \mathbb{K} , the overall statement can be deduced from (6.15). \square

An ordered field (\mathbb{L}, \leq') is called an *order extension* of an ordered field (\mathbb{K}, \leq) if \mathbb{L} is a field extension of \mathbb{K} and \leq' coincides with \leq on \mathbb{K} .

We use a weak form of Tarski's Transfer Principle (see Appendix A.1.12), which states that over an ordered field extension (\mathbb{K}, \leq) of (\mathbb{R}, \leq) , there exists a solution in \mathbb{K} to a system of inequalities with coefficients in \mathbb{R} exactly when there exists a solution over \mathbb{R} .

Theorem 6.3.5 (Artin-Schreier, Solution to Hilbert's 17th problem). *Any nonnegative multivariate polynomial $f \in \mathbb{R}[x_1, \dots, x_n]$ can be written as a sum of squares of rational functions.*

Proof. Let $P = \sum \mathbb{R}(x_1, \dots, x_n)^2$ be the sums of squares of rational functions. P defines a preorder on $\mathbb{R}(x_1, \dots, x_n)$ with $-1 \notin P$.

Assume that there exists a nonnegative polynomial $f \in \mathbb{R}[x_1, \dots, x_n]$ which is not contained in P . By Theorem 6.3.3, the preorder P can be extended to an order P' on $\mathbb{R}(x_1, \dots, x_n)$. In particular, we have $f <_{P'} 0$. Over the field $\mathbb{R}(x_1, \dots, x_n)$, there exist $a_1, \dots, a_n \in \mathbb{R}(x_1, \dots, x_n)$ with

$$f(a_1, \dots, a_n) <_{P'} 0. \quad (6.16)$$

Namely, just choose a_i as the variable x_i , which can be viewed as an element of $\mathbb{R}(x_1, \dots, x_n)$. Clearly, $\mathbb{R}(x_1, \dots, x_n)$ is a field extension of \mathbb{R} . Since \mathbb{R} has only one order, $\leq_{P'}$ coincides on \mathbb{R} with the usual order. Hence, Tarski's Transfer Principle implies that there exist $a_1, \dots, a_n \in \mathbb{R}^n$ with

$$f(a_1, \dots, a_n) < 0, \quad (6.17)$$

where $<$ refers to the natural order on \mathbb{R} . This is a contradiction to the nonnegativity of f . \square

Exercises

1. If \mathbb{K} is an ordered field then $\sum_{i=1}^k a_i^2 > 0$ for any $a_1, \dots, a_k \in \mathbb{K}^\times$. Conclude that every ordered field has characteristic zero.
2. Show that the natural orders on the fields \mathbb{R} and \mathbb{Q} are the unique orders on these fields.
3. Why is the field \mathbb{C} of complex numbers not an ordered field?
4. The field $\mathbb{Q}(\sqrt{2}) = \{a+b\sqrt{2} : a, b \in \mathbb{Q}\}$ has more than one order. For this, consider the natural order $P = \{a+b\sqrt{2} : a+b\sqrt{2} \geq 0\}$ as well as $P' = \{a+b\sqrt{2} : a-b\sqrt{2} \geq 0\}$, where " \geq " is the usual \geq -relation on \mathbb{R} .
5. If P is a preorder of a field \mathbb{K} , 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) $-1 \notin P$.

If $\text{char } \mathbb{K} \neq 2$, then the statement $P \neq \mathbb{K}$ is equivalent as well.

6. Let P be a preorder of a field \mathbb{K} with $-1 \notin P$. For every element $a \in \mathbb{K}$ with $a \notin -P$, the set $P' = P + aP$ is a preorder of \mathbb{K} with $-1 \notin P'$.

6.4 Systems of polynomial inequalities

Hilbert's Nullstellensatz (Theorem 1.2.9) leads to the dictionary between algebraic subvarieties of \mathbb{C}^n and ideals in $\mathbb{C}[x_1, \dots, x_n]$. It is equivalent to the weak Nullstellensatz (Theorem 1.2.6), which gives a certificate for the nonexistence of solution to a system of polynomial equations, that is, of emptiness of the corresponding variety. To see this, we give a reformulation of the Weak Nullstellensatz.

Theorem 6.4.1. *Given $f_1, \dots, f_m \in \mathbb{C}[x]$, the following two statements are equivalent.*

1. *The set $\{x \in \mathbb{C}^n \mid f_i(x) = 0 \text{ for } 1 \leq i \leq m\}$ is empty.*
2. *$1 \in \langle f_1, \dots, f_m \rangle$. That is, there exist $g_1, \dots, g_m \in \mathbb{C}[x]$ with*

$$f_1g_1 + \cdots + f_mg_m = 1. \quad (6.18)$$

However, as pointed out in Section 1.6, the inherent difficulty is that the degrees of the polynomials in the representation (6.18) can grow doubly exponentially in the number n of variables.

Our goal in this section is to derive, for given $g_1, \dots, g_m \in \mathbb{R}[x]$ a characterization of the emptiness of a basic semialgebraic set

$$S(g_1, \dots, g_m) = \{x \in \mathbb{R}^n : g_1(x) \geq 0, \dots, g_m(x) \geq 0\}.$$

Now our main goal in this chapter is to provide the following characterization for the emptiness of a semialgebraic set given in terms of inequalities $g_j(x) \geq 0$, $1 \leq j \leq m$. This characterization provides the prominent step towards the Positivstellensatz 6.5.1, which will be further detailed in the next section.

To write down the statement and to develop the proof techniques, recall that in Section 6.3, we have used preorders on a field to prove Hilbert's 17th problem. Now we study preorders in the more general context of a ring, specifically of the ring $\mathbb{R}[x]$ of multivariate real polynomials. We also discuss the related notion of a quadratic module. Let us begin with two examples.

Example 6.4.2. The set $\Sigma[x] = \Sigma[x_1, \dots, x_n]$ 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 $g_1, \dots, g_m \in \mathbb{R}[x_1, \dots, x_n]$, let $P = P(g_1, \dots, g_m)$ 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 6.4.3. Given polynomials $g_1, \dots, g_m \in \mathbb{R}[x_1, \dots, x_n]$, let $M = \text{QM}(g_1, \dots, g_m)$ be the set of polynomials of the form

$$\sigma_0 + \sigma_1 g_1 + \dots + \sigma_m g_m,$$

where 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 $a \in \mathbb{R}[x]$.

For polynomials g_1, \dots, g_m , we have $\text{QM}(g_1, \dots, g_m) \subset P(g_1, \dots, g_m)$, and any polynomial in $P(g_1, \dots, g_m)$ (and hence also any polynomial in $\text{QM}(g_1, \dots, g_m)$) is nonnegative on the basic semialgebraic set $S(g_1, \dots, g_m) = \{x \in \mathbb{R}^n : g_j(x) \geq 0, 1 \leq j \leq m\}$ given by g_1, \dots, g_m .

Example 6.4.2 leads to the notion of a preorder of $\mathbb{R}[x_1, \dots, x_n]$ and Example 6.4.3 leads to the notion of a quadratic module of $\mathbb{R}[x_1, \dots, x_n]$. While our preorders are typically subsets of $\mathbb{R}[x_1, \dots, x_n]$, it is useful to define them over an arbitrary commutative ring R containing \mathbb{R} .

Definition 6.4.4. 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

$$M + M \subset M, \quad 1 \in M, \quad \text{and } a^2 M \subset M \text{ for all } a \in R.$$

Every preorder of R is a quadratic module of R . Also, the intersection of preorders is again a preorder, and the same for quadratic modules. Call the set $P(g_1, \dots, g_m)$ the *preorder generated by* g_1, \dots, g_m . This is the smallest preorder containing g_1, \dots, g_m . Similarly, $\text{QM}(g_1, \dots, g_m)$, the *quadratic module generated by* g_1, \dots, g_m , is the smallest quadratic module containing g_1, \dots, g_m . We will write $-M$ for $\{-h \mid h \in M\}$, where M is a set of polynomials.

Theorem 6.4.5 (Infeasibility certificate for polynomial inequalities). *Given $g_1, \dots, g_m \in \mathbb{R}[x]$, the following statements are equivalent.*

1. *The set $S(g_1, \dots, g_m) = \{x \in \mathbb{R}^n : g_j(x) \geq 0 \text{ for } 1 \leq j \leq m\}$ is empty.*
2. *$-1 \in P(g_1, \dots, g_m)$. That is, for all $J \subset \{1, \dots, m\}$ there exist sum of squares polynomials $\sigma_J \in \Sigma[x]$ with*

$$\sum_{J \subset [m]} \sigma_J \prod_{j \in J} g_j = -1.$$

Example 6.4.6. For $g_1 = 1 - x^2 - y^2$, $g_2 = x - 2 \in \mathbb{R}[x, y]$, the semialgebraic set

$$S(g_1, g_2) = \{(x, y) \in \mathbb{R}^2 : 1 - x^2 - y^2 \geq 0, x - 2 \geq 0\} = \emptyset$$

is the intersection of the unit disc with an affine halfplane. Since $\frac{1}{3}(x-2)^2 + y^2 \in \Sigma[x, y]$, the identity

$$\frac{1}{3}(1-x^2-y^2) + \frac{4}{3}(x-2) + \frac{1}{3}(x-2)^2 + \frac{1}{3}y^2 = -1$$

shows that $-1 \in P(g_1, g_2)$, as predicted by Theorem 6.4.5. Note that, in this example -1 is also contained in the quadratic module $QM(g_1, g_2)$.

Example 6.4.7. An example which shows that in Theorem 6.4.5 the preorder $P(g_1, \dots, g_m)$ cannot simply be replaced by the quadratic module $QM(g_1, \dots, g_m)$ is given by $g_1 = x$, $g_2 = y$, $g_3 = -xy - 1 \in \mathbb{R}[x, y]$. We have $S(g_1, g_2, g_3) = \emptyset$ and $-1 \in P(g_1, g_2, g_3)$ through the representation

$$-1 = g_1g_2 + g_3.$$

However, $-1 \notin QM(g_1, g_2, g_3)$. To see this, we consider a specific preorder P' on $\mathbb{R}[x, y]$ defined as follows. For any given polynomial $p \in \mathbb{R}[x, y]$, let $c_{st}x^sy^t$ be its leading term with respect to the lexicographical ordering of the monomials, and set

$$p \in P' \iff \begin{cases} c_{st} > 0 \text{ in the case } (s, t) \not\equiv (1, 1) \pmod{2}, \\ c_{st} < 0 \text{ in the case } (s, t) \equiv (1, 1) \pmod{2}. \end{cases}$$

It is easy to check that P' is a preorder on $\mathbb{R}[x]$ with $g_1, g_2, g_3 \in P'$, as well as $-1 \notin P'$. However, then -1 cannot be contained in $QM(g_1, g_2, g_3)$.

The characterization in Theorem 6.4.5 constitutes the core of the Positivstellensatz, whose more comprehensive formulations and their implications will be treated in the subsequent Section 6.5 then.

Note that if $S(g_1, \dots, g_m)$ is a polyhedron given by the affine polynomials $g_j = b_j - \sum_i a_{ij}y_j$, then Farkas' Lemma 5.4.5 shows the existence of non-negative real vector λ with

$$\sum_{j=1}^m \lambda_j g_j = -1.$$

Since $\sum_{j=1}^m \lambda_j g_j$ is contained in the preorder $P(g_1, \dots, g_m)$, Theorem 6.4.5 generalizes Farkas' Lemma.

In contrast to an ordered field, for a quadratic module $M \cap -M = \{0\}$ does not hold in general. Yet, the set $M \cap -M$ is quite relevant. Our next lemma shows that $M \cap -M$ constitutes an ideal.

Lemma 6.4.8. *Let M be a quadratic module of $\mathbb{R}[x]$. Then $I := M \cap -M$ is an ideal of $\mathbb{R}[x]$. Moreover, $-1 \in M$ if and only if $M = \mathbb{R}[x]$.*

Proof. The properties $I + I \subset I$ and $a^2I \subset I$ for all $a \in \mathbb{R}[x]$ are clear. For $p \in I = M \cap -M$, writing a in the form $a = \frac{1}{4}((a+1)^2 - (a-1)^2)$ shows $ap \in I$ and thus $aI \subset I$. Hence, I is an ideal.

For the second statement, suppose that $-1 \in M$. Since $1 \in M$ by definition, $1 \in M \cap -M = I$. Since I is an ideal, we have $M \cap -M = \mathbb{R}[x]$; hence $M = \mathbb{R}[x]$. \square

A quadratic module M in $\mathbb{R}[x]$ is *proper* if $-1 \notin M$ so that $M \neq \mathbb{R}[x]$. A proper quadratic module M is *maximal* if it is not contained in any strictly larger proper quadratic module. When considering quadratic modules in more general rings R , the notion of properness is vacuous if -1 is a sum of squares in R , for then there are no proper quadratic modules. This occurs, for example, when $R = \mathbb{R}[x]/\langle x^2 + 1 \rangle \simeq \mathbb{C}$.

As already mentioned, for a preorder $P \subset \mathbb{R}[x]$, the ideal $I = P \cap -P$ will be crucial. Given a preorder P and a suitable prime ideal $J \supset I$, we will be interested in extending a given preorder to an order in the quotient field of $\mathbb{R}[x]/J$.

Lemma 6.4.9. *Let P be a proper preorder of $\mathbb{R}[x]$. If $p, q \in \mathbb{R}[x]$ with $pq \in P$ then $P + pP$ or $P - qP$ is a proper preorder of $\mathbb{R}[x]$.*

Proof. If neither of $P + pP$ and $P - qP$ were a proper preorder of $\mathbb{R}[x]$ then there exist $p_1, p_2 \in P$ with $p_1 + pp_2 = -1$ as well as $q_1, q_2 \in P$ with $q_1 - qq_2 = -1$. Hence,

$$(-pp_2)(qq_2) = (1 + p_1)(1 + q_1) = 1 + r$$

for some $r \in P$. This implies $-1 = pp_2qq_2 + r \in P$, which contradicts the precondition. \square

Lemma 6.4.10. *Let P be a proper preorder of $\mathbb{R}[x]$. Then P can be extended to a preorder P' of $\mathbb{R}[x]$ such that $P' \cup -P' = \mathbb{R}[x]$ and $P' \cap -P'$ is a prime ideal of $\mathbb{R}[x]$.*

Proof. Consider the set of proper preorders which contain P . Since P itself belongs to this set, the set is not empty. And since for any chain $(P_t) \subset P$ of preorders containing P , the union $\bigcup_t P_t$ is a proper preorder as well, there exists a maximal proper preorder $P' \supset P$ by Zorn's Lemma.

We claim that P' satisfies $P' \cup -P' = \mathbb{R}[x]$ and that $P' \cap -P'$ is a prime ideal of $\mathbb{R}[x]$.

Let $a \in \mathbb{R}[x]$. Since $a^2 \in P'$, Lemma 6.4.9 implies that $P + aP'$ or $P - aP'$ is a proper preorder of $\mathbb{R}[x]$. By the maximality precondition, we obtain $a \in P'$ or $-a \in P'$, i.e., $a \in P' \cup -P'$.

By Lemma 6.4.8, $I' := P' \cap -P'$ is an ideal of $\mathbb{R}[x]$. We claim that I is prime. Let $ab \in I$. By Lemma 6.4.9, since $ab \in P'$, we have that $P' + aP'$ or $P' - bP'$ is a proper quadratic preorder, hence $a \in P'$ or $-b \in P'$ by the maximality of P' . And similarly, since $-ab = a(-b) \in P'$, we have that $P' + aP'$ or $P' + bP'$ is a proper quadratic preorder, hence $a \in P'$ or $b \in P'$. Consequently $a \in P'$ or $b = 0$. By writing $ab = (-a)(-b)$, we can deduce similarly $-a \in P'$ or $b = 0$. Thus $a \in I$ or $b = 0$, and hence, I is a prime ideal. \square

Suppose that P is a proper preorder in $\mathbb{R}[x]$. For an ideal $J \subset \mathbb{R}[x]$, P naturally carries over to $\mathbb{R}[x]/J$ and, in case of a prime ideal J , also then extends to the quotient field \mathbb{F} of $\mathbb{R}[x]/J$. More precisely, the preorder P^* on \mathbb{F} is given by

$$\begin{aligned} P^* &= \left\{ p \in \mathbb{F} : p = \sum_i \left(\frac{a_i}{b_i} \right)^2 c_i \text{ with } a_i, b_i \in \mathbb{R}[x]/J, b_i \neq 0, c_i \in P/J \right\} \\ &= \left\{ p \in \mathbb{F} : p = \frac{c}{b^2} \text{ with } b \in \mathbb{R}[x]/J, b \neq 0, c \in P/J \right\}, \end{aligned}$$

where the latter reformulation results from taking a common denominator.

Now we can provide the proof of Theorem 6.4.5. Like the proof of Hilbert's 17th problem in Theorem 6.3.5, it is based on Tarski's Transfer Principle.

Proof of Theorem 6.4.5. Write $P = P(g_1, \dots, g_m)$, for short. If $-1 \in P$, then $S(g_1, \dots, g_m) = \emptyset$, for if $x \in S(g_1, \dots, g_m)$, then evaluating the expression for $-1 \in P$ at x gives the contradiction that $-1 > 0$.

Conversely, assume that $-1 \notin P$, that is, P is proper. By Lemma 6.4.10, there exists a preorder $Q \supset P$ such that $Q \cup -Q = \mathbb{R}[x]$ and $J = Q \cap -Q$ is a prime ideal.

The image \overline{Q} of Q in $\mathbb{R}[x]/J$ is a preorder in $\mathbb{R}[x]/J$ with $\overline{Q} \cup -\overline{Q} = \mathbb{R}[x]/J$ and $\overline{Q} \cap -\overline{Q} = \{0\}$. In particular, \overline{Q} is proper. \overline{Q} can be extended in a natural way to a proper preorder Q^* of the quotient field \mathbb{F} of $\mathbb{R}[x]/J$. As we are now in the situation of a proper preorder over a field, Theorem 6.3.3 then gives an order P' on \mathbb{F} with $Q^* \subset P'$. Let $\leq_{P'}$ be the associated order relation. As \mathbb{F} is a field extension of \mathbb{R} , if we restrict $\leq_{P'}$ to \mathbb{R} , we obtain the unique order on \mathbb{R} .

Now we show that there exists an element $a \in \mathbb{F}^n$ with $g_j(x) \geq 0$, $1 \leq j \leq m$. Namely, let $a_i = \bar{x}_i$, where \bar{x}_i is the residue class of x_i in $\mathbb{R}[x]/J$. For any $g = \sum_\alpha c_\alpha x^\alpha \in \mathbb{R}[x]$, the image \bar{g} of g in \mathbb{F} is

$$\bar{g} = \sum_\alpha \bar{c}_\alpha \bar{x}^\alpha = g(\bar{x}) = g(a) \in \mathbb{F}.$$

Since $g_j \in P$, the definitions of \overline{Q} , Q^* and P' imply $g_j(a) \geq_{P'} 0$ with regard to the order P' , $1 \leq j \leq m$.

Thus, by Tarski's Transfer Principle A.1.12, there exists some $z \in \mathbb{R}^n$ with $g_j(z) \geq 0$, $1 \leq j \leq m$. Therefore $S(g_1, \dots, g_m) \neq \emptyset$. \square

Exercises

1. Generalizing Example 6.4.6, for $f_1 = 1 - x^2 - y^2$, $f_2 = x - \alpha \in \mathbb{R}[x, y]$ with $\alpha > 1$, give an identity that shows $-1 \in P(f_1, f_2)$.
2. Let $g_1, \dots, g_m \in \mathbb{R}[x_1, \dots, x_n]$ and $S = S(g_1, \dots, g_m)$, and suppose that p is strictly positive on S . Then for each compact subset $C \subset \mathbb{R}^n$ there exists some $q \in \text{QM}(g_1, \dots, g_m)$ such that $p - q$ is strictly positive on C . Proceed as follows:
 - (a) There exists some $\delta > 0$ with $g_j(x) \leq 2\delta$ for all $j \in \{1, \dots, m\}$ and all $x \in C$.
 - (b) There exists some $\epsilon > 0$ such that for all $x \in C$ with $f(x) \leq 0$ there exists some $j \in \{1, \dots, m\}$ with $g_j(x) < -2\epsilon$ or $g_j(x) > \delta$.
 - (c) For any $N \geq 1$, the polynomial $h(t) = t \left(\frac{t-\delta}{\delta+\epsilon} \right)^{2N}$ satisfies

$$h(t) \in \begin{cases} \left[0, \left(\frac{\delta}{\delta+\epsilon} \right)^{2N} \right] & \text{for } 0 \leq t \leq 2\gamma, \\ \left(-\infty, 2\epsilon \left(\frac{\gamma+2\epsilon}{\gamma+\epsilon} \right)^{2N} \right] & \text{for } t \leq -2\epsilon. \end{cases}$$

- (d) Show that $g(x) = \sum_{j=1}^m h(g_j(x))$ satisfies the desired property for sufficiently large N .

6.5 The Positivstellensatz

Extending the infeasibility certificate for polynomial inequalities in Theorem 6.4.5, we study comprehensive formulations and implications of the Positivstellensatz. This theorem can be viewed as an analog of Hilbert's Nullstellensatz for semialgebraic sets. It gives the existence of a certificate for the nonnegativity of a polynomial on a semialgebraic set. Let $g_1, \dots, g_s, h_1, \dots, h_t \in \mathbb{R}[x_1, \dots, x_n]$ be polynomials. As before $P(g_1, \dots, g_s)$ is the preorder generated by the polynomials g_1, \dots, g_s , and we let $\text{Mon}(h_1, \dots, h_t)$ be the multiplicative monoid defined by the polynomials h_1, \dots, h_t . This is the set of (finite) products of the h_i including the empty product, 1.

Theorem 6.5.1 (Positivstellensatz). *For polynomials $f_1, \dots, f_r, g_1, \dots, g_s, h_1, \dots, h_t \in \mathbb{R}[x_1, \dots, x_n]$ the following statements are equivalent.*

1. *The set $\{x \in \mathbb{R}^n \mid f_i(x) = 0, g_j(x) \geq 0, h_k(x) \neq 0 \quad \forall i, j, k\}$ is empty.*
2. *There exist $F \in \langle f_1, \dots, f_r \rangle$, $G \in P(g_1, \dots, g_s)$ and $H \in \text{Mon}(h_1, \dots, h_t)$ with $F + G + H^2 = 0$.*

Note that (2) \Rightarrow (1), for if x lies in the set of (1) and F, G, H are as in (2), then $0 = F(x) + G(x) + H^2(x) > 0$, a contradiction, as $F(x) = 0$, $G(x) \geq 0$, and $H^2(x) > 0$. And observe that the theorem has several important special cases. The specific case where $r = 0$ and $t = 0$ has already been treated in Theorem 6.4.5. And in case $s = t = 0$, we have the Real Nullstellensatz.

Corollary 6.5.2 ((Weak) Real Nullstellensatz). *Let $f_1, \dots, f_r \in \mathbb{R}[x_1, \dots, x_n]$. Then the real variety $\mathcal{V}_{\mathbb{R}}(f_1, \dots, f_r)$ is empty if and only if there exists $F \in \langle f_1, \dots, f_r \rangle$ and a sum of squares $G \in \Sigma[x]$ with*

$$F + G + 1 = 0. \tag{6.19}$$

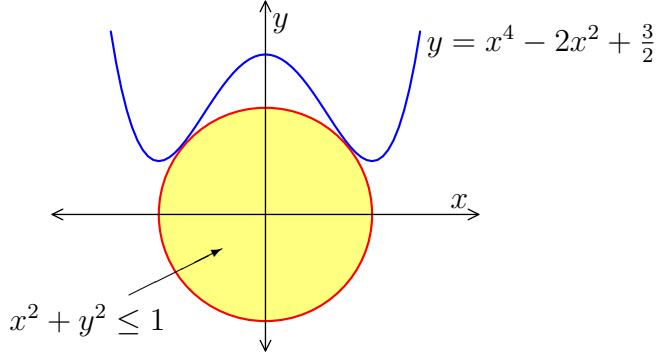
Example 6.5.3. The polynomial $f(x) = x^2 + 1$ does not have a real zero. The polynomials $F = -(x^2 + 1) \in \langle f \rangle$ and $G = x^2 \in \Sigma[x]$ satisfy (6.19) and thus provide a certificate that $\mathcal{V}_{\mathbb{R}}(f)$ is empty.

The general quadratic equation $x^2 + ax + b = 0$ with coefficients $a, b \in \mathbb{R}$ has a real solution unless the discriminant $D := \frac{a^2}{4} - b$ is negative. If $D < 0$, then

$$F := \frac{1}{D} (x^2 + ax + b) \quad \text{and} \quad G := \left(\frac{1}{\sqrt{-D}} \left(x + \frac{a}{2} \right) \right)^2$$

provide a certificate of the form (6.19) that $\mathcal{V}_{\mathbb{R}}(F)$ is empty.

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.



Indeed, set $f := y - (x^4 - 2x^2 + \frac{3}{2})$, $g := 1 - x^2 - y^2$, and $a := (\frac{2}{3})^{1/4}$. Then

$$f + (ay - \frac{1}{2a})^2 + (x^2 + \frac{a^2}{2} - 1)^2 + a^2g + \beta = 0,$$

with $\beta := \frac{8-3\sqrt{6}}{24} < 0$, so that scaling the equation with $\frac{1}{\beta}$ shows $\emptyset = \{(x, y) \in \mathbb{R}^2 \mid f(x, y) = 0 \text{ and } g(x, y) \geq 0\}$.

Proof of the Positivstellensatz 6.5.1. First we consider the situation $t = 0$, i.e., there are no constraints of the form $h_i \neq 0$.

Given $f_1, \dots, f_r, g_1, \dots, g_s$, let the set $\{x \in \mathbb{R}^n \mid f_i(x) = 0, g_j(x) \geq 0 \quad \forall i, j\}$ be empty. Setting $f'_i = -f_i$, the set $\{x \in \mathbb{R}^n \mid f_i(x) \geq 0, f'_i(x) \geq 0, g_j(x) \geq 0 \quad \forall i, j\}$ is empty. Hence, by Theorem 6.4.5, we have

$$-1 \in P(f_1, \dots, f_r, f'_1, \dots, f'_r, g_1, \dots, g_s).$$

Since every polynomial in $P(f_1, \dots, f_r, f'_1, \dots, f'_r, g_1, \dots, g_s)$ is of the form $F + G$ with $F \in \langle f_1, \dots, f_r \rangle$ and $G \in P(g_1, \dots, g_m)$, we have obtained the desired certificate.

Now we reduce the case of general t to the situation $t = 0$ using a Rabinowitsch-type trick as in the proof of Hilbert's Nullstellensatz 1.2.9. The precondition $\{x \in \mathbb{R}^n \mid f_i(x) = 0, g_j(x) \geq 0, h_k(x) \neq 0 \quad \forall i, j, k\} = \emptyset$ holds if and only if

$$\{x \in \mathbb{R}^n \mid f_i(x) = 0, g_j(x) \geq 0, 1 - y_k h_k(x) = 0 \quad \forall i, j, k\} = \emptyset,$$

where y_1, \dots, y_k are new variables. Hence, by the already known case of $t = 0$, there exist $F \in \langle f_1, \dots, f_r, 1 - y_1 h_1, \dots, 1 - y_t h_t \rangle$ and $G \in P(g_1, \dots, g_m) \subset \Sigma[x, y]$ with $F + G + 1 = 0$. Substituting $z_k = \frac{1}{h_k}$, $1 \leq k \leq t$, and clearing denominators gives an identity

$$F' + G' + H' = 1$$

in the polynomial ring $\mathbb{R}[x]$, where $F' \in \langle f_1, \dots, f_r \rangle \subseteq \mathbb{R}[x]$, $G' \in P(g_1, \dots, g_m) \subset \Sigma[x]$ and H' is a product of powers of h_1, \dots, h_t . Moreover, since G and G' are sums of squares, every of the powers of h_k in H must be even. This proves the claim. \square

We record more special cases of the Positivstellensatz.

Corollary 6.5.4. *Let $g_1, \dots, g_m \in \mathbb{R}[x_1, \dots, x_n]$. Set $S = S(g_1, \dots, g_m)$ and let $P = P(g_1, \dots, g_m)$ be the preorder generated by g_1, \dots, g_m . For a polynomial f , we have*

1. $f > 0$ on S if and only if there exist $G, H \in P$ with $fG = 1 + H$.
2. $f \geq 0$ on S if and only if there exist $G, H \in P$ and $k \geq 0$ with $fG = f^{2k} + H$.
3. $f = 0$ on S if and only if there exists a $G \in P$ and $k \geq 0$ with $f^{2k} + G = 0$.

Proof. For the first statement, consider the set

$$\{x \in \mathbb{R}^n \mid -f(x) \geq 0, g_j(x) \geq 0, 1 \leq j \leq m\}.$$

This is empty if and only if there exist $G, H \in P$ with $H - fG + 1 = 0$.

For the second statement, apply the Positivstellensatz 6.5.1 to the set

$$\{x \in \mathbb{R}^n \mid -f(x) \geq 0, f(x) \neq 0 \text{ and } g_j(x) \geq 0 \text{ for } 1 \leq j \leq m\}.$$

This is empty if and only if there exist $G, H \in P$ and $k \geq 0$ with $H - fG + f^{2k} = 0$.

For the third, the set

$$\{x \in \mathbb{R}^n \mid f(x) \neq 0 \text{ and } g_j(x) \geq 0 \text{ for } 1 \leq j \leq m\}$$

is empty if and only if there exist $G \in P$ and $k \geq 0$ with $G + f^{2k} = 0$. □

Remark 6.5.5. A slight variant of the first statement in Corollary 6.5.4 states that $f > 0$ on S if and only if there exist $G, H \in P$ with $f(1 + G) = 1 + H$. Clearly, that condition implies positivity of f . And in the case of strictly positive f , there exist $G', H' \in P$ with $fG' = 1 + H'$. G' cannot have a zero, so it must contain a constant term $\alpha \geq 0$. For $\alpha \geq 1$ observe $f(1 + G'') = (1 + H')$ with $G'' \in P$, and for $0 \leq \alpha < 1$ consider $f(\alpha + G') = (\alpha + H'')$ with $H'' \in P$ and normalize the fraction.

The Positivstellensatz implies the Artin–Schreier solution to Hilbert’s 17th Problem.

Corollary 6.5.6 (Solution to Hilbert’s 17th problem). *Any nonnegative multivariate polynomial $f \in \mathbb{R}[x_1, \dots, x_n]$ can be written as a sum of squares of rational functions.*

Proof. When $m = 0$, Corollary 6.5.4(2) implies that for any nonnegative polynomial f , there is an identity $fg = f^{2k} + h$ with g, h sums of squares and k a nonnegative integer. We may assume that $f \neq 0$. Then $f^{2k} + h \neq 0$, and so $g \neq 0$. Hence, we have

$$f = \frac{1}{g}(f^{2k} + h) = \left(\frac{1}{g}\right)^2 g(f^{2k} + h),$$

a representation of f as a sum of squares of rational functions. □

This explains why a treatment of the Positivstellensatz is closely connected to a treatment of Hilbert's 17th problem.

We close the section by pointing out that similar to the situation over an algebraic closed field, the Weak Real Nullstellensatz can be transformed into a strong version. For an ideal $I \subset \mathbb{R}[x]$ define the *real radical* by

$$\sqrt{\mathbb{R}I} = \{p \in \mathbb{R}[x] : p^{2k} + q \in I \text{ for some } k > 0 \text{ and } q \in \Sigma[x]\}.$$

The set $\sqrt{\mathbb{R}I}$ is an ideal in $\mathbb{R}[x]$. Namely, if $p^{2k} + q \in I$ and $r^{2j} + s \in I$ with $q, s \in \Sigma[x]$, then set $h = (p+r)^{2(k+j)} + (p-r)^{2(k+j)} = p^{2k}t + r^{2j}t'$ for some $t, t' \in \Sigma[x]$. We obtain $h + tq + t's \in I$, and thus $p+r \in \sqrt{\mathbb{R}I}$. And closure under multiplication with arbitrary elements in $\mathbb{R}[x]$ is clear.

Theorem 6.5.7 (Strong Real Nullstellensatz, by Dubois and Risler). *Let I be an ideal in $\mathbb{R}[x]$. Then $\mathcal{I}(\mathcal{V}_{\mathbb{R}}(I)) = \sqrt{\mathbb{R}I}$.*

Here, recall that $\mathcal{I}(\mathcal{V}_{\mathbb{R}}(I)) = \{f \in \mathbb{R}[x] : f(a) = 0 \text{ on all points of } \mathcal{V}_{\mathbb{R}}(I)\}$ denotes the vanishing ideal of $\mathcal{V}_{\mathbb{R}}(I)$.

Proof. We begin with the easy direction $\mathcal{I}(\mathcal{V}_{\mathbb{R}}(I)) \supset \sqrt{\mathbb{R}I}$. If $p^{2k} + q \in I$ with $q \in \Sigma[x]$, then p vanishes on $\mathcal{V}_{\mathbb{R}}(I)$, which implies $p \in \mathcal{I}(\mathcal{V}_{\mathbb{R}}(I))$. For the converse direction, similar to the proof of Theorem 1.2.9, Rabinowitsch's trick can be used. Let $p \in \mathcal{I}(\mathcal{V}_{\mathbb{R}}(I))$. We introduce a new variable t and set $R' = \mathbb{R}[x, t]$. Then define the ideal

$$I' = IR' + (1-pt)R'.$$

Specializing R' in the second term to the zero polynomial shows that any zero of I' is of the form (a, β) with $a \in \mathcal{V}_{\mathbb{R}}(I)$. However, then it cannot be a zero of $1-pt$. Hence, $\mathcal{V}_{R'}(I') = \emptyset$, which implies $1 \in \sqrt{\mathbb{R}I'}$ by the weak form of the real Nullstellensatz. Thus there exist polynomials $g_1, \dots, g_s, h_1, \dots, h_r, h \in R'$ with

$$1 + \sum_{i=1}^s g_i^2 = \sum_{j=1}^r f_j h_j + (1-pt)h.$$

Substituting $t = 1/p$ and multiplying by a sufficiently large power of p then gives

$$p^{2k} + \sum_{i=1}^s (g'_i)^2 = \sum_{j=1}^r f_j h'_j \in I$$

with $g'_i, h'_j \in \mathbb{R}[x]$. This shows $p \in \sqrt{\mathbb{R}I}$. □

Exercises

1. Write the Motzkin polynomial 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.
3. Derive the following affine version of Farkas' Lemma from Theorem 5.4.5, and show that the certificates therein are special cases of the certificates in Corollary 6.5.4.
Let p and g_1, \dots, g_m be affine-linear functions. If $S = S(g_1, \dots, g_m)$ is nonempty and p is nonnegative on S , then there exist scalars $\lambda_0, \dots, \lambda_m \geq 0$ with

$$p = \lambda_0 + \sum_{j=1}^m \lambda_j g_j.$$

4. An ideal $I \subset \mathbb{R}[x_1, \dots, x_n]$ is called *real* if $q_1^2 + \dots + q_s^2 \in I$ implies $q_1, \dots, q_s \in I$. Show that for an ideal $J \subset \mathbb{R}[x_1, \dots, x_n]$, $\sqrt{\mathbb{R}J}$ is the smallest real ideal in $\mathbb{R}[x_1, \dots, x_n]$ containing J .
5. For an ideal $I \subset \mathbb{R}[x_1, \dots, x_n]$, show that $\sqrt{\mathbb{R}I}$ is the intersection of all real prime ideals containing I .
6. For an ideal $I \subset \mathbb{R}[x_1, \dots, x_n]$, show that

$$I' = \{f \in \mathbb{R}[x_1, \dots, x_n] : f^2 + \sigma \in I \text{ for some } \sigma \in \Sigma[x_1, \dots, x_n]\}$$

is an ideal with $\sqrt{\mathbb{R}I'} = \sqrt{I'}$, where $\sqrt{I'}$ denotes the usual ideal of I' .

6.6 Theorems of Pólya and Handelman

In 1927 Pólya discovered a fundamental theorem for positive polynomials on a simplex, which will also be an ingredient for important representation theorems of positive polynomials. After discussing Pólya's Theorem, we derive Handelman's Theorem from it, which provides a distinguished representation of a positive polynomial on a polytope.

Theorem 6.6.1 (Pólya). *Let $p \in \mathbb{R}[x_1, \dots, x_n]$ be a homogeneous polynomial which is positive on $\mathbb{R}_{\geq 0}^n \setminus \{0\}$. Then for all sufficiently large $N \in \mathbb{N}$, the polynomial*

$$(x_1 + \dots + x_n)^N p$$

has only nonnegative coefficients.

As p is homogeneous, the positivity of p on $\mathbb{R}_{\geq 0}^n \setminus \{0\}$ is equivalent to its positivity on the unit simplex $\Delta_n := \{y \in \mathbb{R}_{\geq 0}^n \mid \sum_{i=1}^n y_i = 1\}$. Further note that when $n = 1$, Pólya's Theorem is trivial, since a homogeneous univariate polynomial is a single term.

Proof. Let $p = \sum_{|\alpha|=d} c_\alpha x^\alpha$ be a homogeneous polynomial of degree d . Define the new polynomial

$$g_p = g_p(x_1, \dots, x_n, t) := \sum_{|\alpha|=d} c_\alpha \prod_{i=1}^n x_i(x_i - t) \cdots (x_i - (\alpha_i - 1)t)$$

in the ring extension $\mathbb{R}[x_1, \dots, x_n, t]$. Then $p = g_p(x_1, \dots, x_n, 0)$. This polynomial g_P arises in the expansion of $(x_1 + \cdots + x_n)^N p$ for any $N \in \mathbb{N}$,

$$(x_1 + \cdots + x_n)^N p = \sum_{|\beta|=d+N} \frac{N!(d+N)^d}{\beta_1! \cdots \beta_n!} g_p\left(\frac{\beta_1}{d+N}, \dots, \frac{\beta_n}{d+N}, \frac{1}{d+N}\right) x^\beta. \quad (6.20)$$

To see this, first expand $(x_1 + \cdots + x_n)^N$ using the multinomial theorem and then collect terms with the same monomial x^β to obtain

$$\begin{aligned} (x_1 + \cdots + x_n)^N p &= \sum_{|\alpha|=d} \sum_{|\gamma|=N} c_\alpha \binom{N}{\gamma_1 \dots \gamma_n} x_1^{\alpha_1+\gamma_1} \cdots x_n^{\alpha_n+\gamma_n} \\ &= \sum_{|\beta|=d+N} \sum_{|\alpha|=d, \alpha \leq \beta} c_\alpha \binom{N}{\beta_1-\alpha_1 \dots \beta_n-\alpha_n} x^\beta. \end{aligned} \quad (6.21)$$

Equality of the coefficients in (6.20) and (6.21) follows from the identity

$$\sum_{|\alpha|=d, \alpha \leq \beta} c_\alpha \binom{N}{\beta_1-\alpha_1 \dots \beta_n-\alpha_n} = \frac{N!}{\beta_1! \cdots \beta_n!} \sum_{|\alpha|=d} c_\alpha \prod_{i=1}^d \beta_i(\beta_i - 1) \cdots (\beta_i - (\alpha_i - 1)).$$

Since $(\frac{\beta_1}{d+N}, \dots, \frac{\beta_n}{d+N})$ lies in the compact set Δ_n , and since $\lim_{N \rightarrow \infty} \frac{1}{d+N} = 0$, 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 in \mathbb{R}^n and U_x is a neighborhood of 0 in \mathbb{R} .

The family of open sets $\{S_x \mid x \in \Delta_n\}$ covers the compact set Δ_n , and hence there exists finite covering $\{S_x \mid x \in X\}$ for a finite subset X of Δ_n . Choosing $U = \bigcap_{x \in X} U_x$ yields the desired neighborhood of 0 . \square

By Pólya's Theorem, any homogeneous polynomial p which is positive on the simplex Δ_n can be written as a quotient of two polynomials with nonnegative coefficients, because $p = \frac{f}{(x_1 + \cdots + x_n)^N}$ with some polynomial f that has nonnegative coefficients. This gives a constructive solution to a special case of Hilbert's 17th problem. A polynomial q is *even* if $q(x) = q(-x)$. This is equivalent to every monomial that appears in q having even exponents, so that there is a polynomial p with $q(x) = p(x_1^2, \dots, x_n^2)$. If an even polynomial

q is homogeneous and positive on $\mathbb{R}^n \setminus \{0\}$, then the polynomial p is homogeneous and positive on the simplex Δ_n . Then substituting x_i^2 for x_i in the expression $p = \frac{f}{(x_1 + \dots + x_n)^N}$ given by Polya's Theorem gives an expression for q as a quotient of polynomials that are sums of squares of monomials.

Example 6.6.2. Let p be defined by

$$p = (x - y)^2 + (x + y)^2 + (x - z)^2 = 3x^2 - 2xz + 2y^2 + z^2.$$

This homogeneous quadratic is positive on $\mathbb{R}^3 \setminus \{0\}$, and hence on the simplex Δ_3 . 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 $(x_1 + x_2 + x_3)^2 p$ also has some negative coefficients. All coefficients of $(x_1 + \dots + x_n)^N p$ are nonnegative, and therefore all coefficients of $(x_1 + x_2 + x_3)^N p$ are nonnegative, for any integer $N \geq 3$.

For a polynomial $p \in \mathbb{R}[x_1, \dots, x_n]$ that is positive on the simplex Δ_n , the smallest positive integer N so that all coefficients of $(x_1 + \dots + x_n)^N p$ are positive is 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 6.5.1) a representation of the polynomial p itself, rather than only a product of a sum of squares polynomial with p . For polynomials g_1, \dots, g_m and an integer vector $\beta \in \mathbb{N}^m$, write g^β for the product $g_1^{\beta_1} \cdots g_m^{\beta_m}$.

Theorem 6.6.3 (Handelman). *Let $g_1, \dots, g_m \in \mathbb{R}[x]$ be affine-linear polynomials such that the polyhedron $S = S(g_1, \dots, g_m)$ is non-empty and bounded. Any polynomial $p \in \mathbb{R}[x]$ which is positive on S can be written as a finite sum*

$$p = \sum_{\beta \in \mathbb{N}^m} c_\beta \prod_{j=1}^m g_j^{\beta_j} = \sum_{\beta \in \mathbb{N}^m} c_\beta g^\beta \quad (6.22)$$

of monomials in the g_i with nonnegative coefficients c_β .

In terms of the semiring

$$\mathbb{R}_{\geq 0}[g_1, \dots, g_m] = \left\{ \sum_{\beta \in \mathbb{N}^m} c_\beta g^\beta : c_\beta \geq 0 \text{ for all } \beta \in \mathbb{N}^m \right\},$$

Handelman's Theorem states that $p \in \mathbb{R}_{\geq 0}[g_1, \dots, g_m]$.

By Farkas' Lemma 5.4.5, any linear form which is strictly positive on S lies in $\mathbb{R}_{\geq 0}[g_1, \dots, g_m]$. Thus it suffices to show that $p \in \mathbb{R}_{\geq 0}[g_1, \dots, g_m, \ell_1, \dots, \ell_r]$ where ℓ_1, \dots, ℓ_r are linear forms 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 6.6.3 will follow from a suitable substitution.

Proof. For $1 \leq i \leq n$ let ℓ_i be an affine linear form $\ell_i = x_i + \tau_i$, such that $\tau_i > -\min\{x_i \mid x \in S\}$. Then ℓ_i is strictly positive on S . Further set

$$\ell_{n+1} := \tau - \sum_{j=1}^m g_j - \sum_{i=1}^n \ell_i$$

with $\tau \in \mathbb{R}$ chosen so that ℓ_{n+1} is strictly positive on S . We show $p \in \mathbb{R}_{\geq 0}[g_1, \dots, g_m, \ell_1, \dots, \ell_{n+1}]$, from which the desired statement follows.

Applying a suitable transformation we may assume $\ell_i = x_i$, $1 \leq i \leq n$. Now extend the variable set x_1, \dots, x_n to x_1, \dots, x_{n+m+1} and consider the polynomial

$$h(x) := p(x) + c \sum_{j=1}^m (x_{n+j} - g_j(x)),$$

where c is a positive constant. Note that $h(x_1, \dots, x_n, g_1(x), \dots, g_m(x), x_{n+m+1}) = p(x)$. Let $\Delta = \{x \in \mathbb{R}_+^{n+m+1} \mid \sum_{i=1}^{n+m+1} x_i = \tau\}$ be the scaled standard simplex and

$$\Delta' = \{x \in \mathbb{R}_+^{n+m+1} \mid x_{n+1} = g_1(x), \dots, x_{n+m} = g_m(x)\}.$$

For any $x \in \Delta'$ and $j \in \{1, \dots, m\}$ 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 . Let

$$\bar{h} = \left(\frac{1}{\tau} \sum x_i \right)^{\deg p} h \left(\frac{x_1}{\frac{1}{\tau} \sum x_i}, \dots, \frac{x_{n+m}}{\frac{1}{\tau} \sum x_i} \right)$$

be the homogenization of h with respect to $\frac{1}{\tau} \sum x_i$. (Note that h and p have the same degree.) Since \bar{h} is strictly positive on Δ , Pólya's Theorem 6.6.1 gives an $N \in \mathbb{N}_0$ such that all coefficients of $H(x) := (\frac{1}{\tau} \sum x_i)^N \bar{h}(x)$ are nonnegative. Substituting successively $x_{n+m+1} = \tau - \sum_{i=1}^{n+m} x_i$ and $x_{n+j} = g_j(x)$, $1 \leq j \leq m$, we see that

$$p(x_1, \dots, x_n) = H(x_1, \dots, x_n, g_1(x), \dots, g_m(x), \ell_{n+m+1}(x)),$$

which gives $p \in \mathbb{R}_{\geq 0}[g_1, \dots, g_m, \ell_1, \dots, \ell_n, \ell_{n+1}]$ as needed. \square

Example 6.6.4. It can happen that the degree of the right hand side of any Handelman representation (6.22) of p exceeds the degree of p . This occurs even for univariate polynomials. Indeed, suppose that $g_1(x) = 1 + x$ and $g_2(x) = 1 - x$. The parabola $x^2 + 1$ is strictly positive on $[-1, 1] = S(g_1, g_2)$, but it cannot be represented as a nonnegative linear combination of the monomials 1 , $1 + x$, $1 - x$, and $(1 + x)(1 - x)$ in the g_i . This phenomenon has already been mentioned in connection with the Lorentz degree in the exercises to Section 6.1.

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. Show that the polynomial $p = x^2zw + y^2zw + xyz^2 + xyw^2 - xyzw$ is not strictly positive, but there exists some $N > 0$ such that $(x+y+z+w)p$ has only non-negative coefficients. What is the smallest N ?
4. Determine a Handelman representation for the polynomial $p = 3/4(x^2+y^2)-xy+1/2$ over the simplex $\Delta = \{x \in \mathbb{R}^n : x \geq 0, y \geq 0, 1 - x - y \geq 0\}$ such that each term in the Handelman representation is of degree at most 2.

6.7 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 exhibits some techniques which will also be applied in the subsequent Theorems of Schmüdgen and Putinar.

In the following let $g_1, \dots, g_m \in \mathbb{R}[x] = \mathbb{R}[x_1, \dots, x_n]$ and $S = S(g_1, \dots, g_m)$. Recall from Example 6.4.3 that $\text{QM}(g_1, \dots, g_m)$ is the quadratic module defined by g_1, \dots, g_m .

Theorem 6.7.1 (Jacobi-Prestel). *Suppose that S is nonempty and bounded, and that $\text{QM}(g_1, \dots, g_m)$ contains linear polynomials ℓ_1, \dots, ℓ_k with $k \geq 1$ such that the polyhedron $S(\ell_1, \dots, \ell_k)$ is bounded. If p is strictly positive on S , then $p \in \text{QM}(g_1, \dots, g_m)$.*

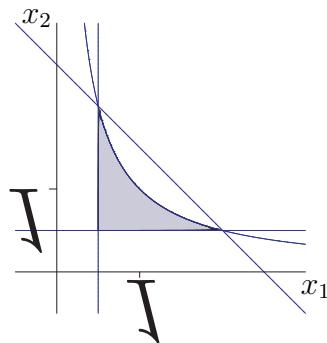


Figure 6.2: The feasible region $S(g_1, \dots, g_4)$ for $g_1 = x_1 - 1/2$, $g_2 = x_2 - 1/2$, $g_3 = 1 - x_1 - x_2$, $g_4 = 5/2 - x_1 - x_2$.

Example 6.7.2. Let $g_1 = x_1 - 1/2$, $g_2 = x_2 - 1/2$, $g_3 = 1 - x_1 x_2$, $g_4 = 5/2 - x_1 - x_2$, see Figure 6.2 for an illustration of $S = S(g_1, \dots, g_4)$. The polynomial $p = 8 - x_1^2 - x_2^2$ is strictly positive on the set $S = S(g_1, g_2, g_3, g_4)$, and since $S(g_1, g_2, g_4)$ is a bounded polygon, Theorem 6.7.1 guarantees that $p \in \text{QM}(g_1, \dots, g_4)$. To write down one such a representation, first observe that the identities

$$\begin{aligned} 2 - x_i &= \left(\frac{5}{2} - x_1 - x_2 \right) + \left(x_{3-i} - \frac{1}{2} \right), \\ 2 + x_i &= \left(x_i - \frac{1}{2} \right) + \frac{5}{2} \end{aligned}$$

($i \in \{1, 2\}$) allow that we may additionally use the box constraints $2 - x_i \geq 0$ and $2 + x_i \geq 0$ for our representation. Then, clearly, one possible Jacobi-Prestel representation for p is provided by

$$p = \frac{1}{4} \sum_{i=1}^2 ((2 + x_i)^2 (2 - x_i) + (2 - x_i)^2 (2 + x_i)).$$

Note that the feasible S does not change if we omit the linear constraint $g_4 \geq 0$. Interestingly, then the precondition of Theorem 6.7.1 is no longer satisfied, and, as we will see in Exercise 1, although p is strictly positive on $S(g_1, g_2, g_3)$, it is not contained $\text{QM}(g_1, g_2, g_3)$.

Proof. Without loss of generality we can assume that the polytope $L := S(\ell_1, \dots, \ell_k)$ is contained in the cube $[-1, 1]^n$.

We start by showing that $p \in \text{QM}(g_1, \dots, g_m, n - \|x\|^2 + 1)$. For this, first note that

$$n + 2 \pm x_i = \frac{1}{2} \left((n+2) + (1 \pm x_i)^2 + \sum_{j \in \{1, \dots, n\} \setminus \{i\}} x_j^2 + (n - \|x\|^2 + 1) \right) \quad (6.23)$$

$$\in \text{QM}(n - \|x\|^2 + 1). \quad (6.24)$$

Hence, by taking linear combinations, $\tau + \ell_i \in \text{QM}(n - \|x\|^2 + 1)$ for sufficiently large $\tau \in \mathbb{R}$. Since L is bounded, the set

$$P = S(\tau + \ell_1, \dots, \tau + \ell_k)$$

is bounded as well and thus a polytope. By Exercise 2 in Section 6.4 there exists a polynomial $q \in \text{QM}(g_1, \dots, g_m)$ such that $p - q$ is strictly positive on P . Applying Handelman's Theorem 6.6.3, we can deduce

$$p - q \in \mathbb{R}_{\geq 0}[\tau + \ell_1, \dots, \tau + \ell_k] \subset \text{QM}(n - \|x\|^2 + 1),$$

whence $p \in \text{QM}(g_1, \dots, g_m, n - \|x\|^2 + 1)$.

To deduce $p \in \text{QM}(g_1, \dots, g_m)$ from $p \in \text{QM}(g_1, \dots, g_m, n - \|x\|^2 + 1)$, observe

$$\begin{aligned} n - \|x\|^2 &= \frac{1}{2} \sum_{i=1}^n ((1 + x_i)^2 (1 - x_i) + (1 - x_i)^2 (1 + x_i)) \\ &\in \text{QM}(1 - x_1, \dots, 1 - x_n, 1 + x_1, \dots, 1 + x_n) \end{aligned}$$

Since by Farkas' Lemma we have $1 \pm x_i \in \mathbb{R}_{\geq 0} + \sum \mathbb{R}_{\geq 0} \ell_i$ and by assumption $\ell_1, \dots, \ell_k \in \text{QM}(g_1, \dots, g_m)$, we can conclude $n - \|x\|^2 \in \text{QM}(g_1, \dots, g_m)$. Hence, the polynomial $n - \|x\|^2 + 1$ is contained in $\text{QM}(g_1, \dots, g_m)$ as well, so that $p \in \text{QM}(g_1, \dots, g_m, n - \|x\|^2 + 1) \subset \text{QM}(g_1, \dots, g_m)$. \square

We now derive the fundamental Theorem of Schmüdgen. Under the condition of compactness of the feasible set, it characterizes (in contrast to Theorem 6.5.1) a representation of the polynomial p itself in terms of the preorder of the polynomials defining the feasible set.

Theorem 6.7.3 (Schmüdgen). *If a polynomial $f \in \mathbb{R}[x]$ is strictly positive on a compact set $S = S(g_1, \dots, g_m)$ then $f \in P(g_1, \dots, g_m)$. That is, f can be written in the form*

$$f = \sum_{e \in \{0,1\}^m} \sigma_e g^e = \sum_{e \in \{0,1\}^m} \sigma_e g_1^{e_1} \cdots g_m^{e_m} \quad (6.25)$$

with $\sigma_e \in \Sigma[x]$ for all $e \in \{0,1\}^m$.

We use the following auxiliary result.

Lemma 6.7.4 (Berr, Wörmann). *For every $g = \sum_\alpha c_\alpha x^\alpha \in \mathbb{R}[x]$ and $\gamma > 0$ we have:*

1. *There exists some $\tau > 0$ with $\tau - g \in P(\gamma - \|x\|^2)$.*
2. *There exists some $\gamma' > 0$ with $\gamma' - \|x\|^2 \in \text{QM}(g, (1+g)(\gamma - \|x\|^2))$.*

Proof. Setting $\tau = \sum_\alpha |c_\alpha|(\gamma + 1)^{|\alpha|}$, we obtain

$$\tau - g = \sum_\alpha (|c_\alpha|(\gamma + 1)^{|\alpha|} - c_\alpha x^\alpha). \quad (6.26)$$

Now the general formula

$$x_1 \cdots x_n - y_1 \cdots y_n = \frac{1}{2^{n-1}} \sum_{\substack{\beta \in \{0,1\}^n \\ |\beta| \text{ odd}}} \prod_{i=1}^n (x_i + \beta_i y_i)$$

implies that for each α , the term in (6.26) with index α is contained in the preorder P generated by the polynomials $\gamma + 1 \pm x_i$, $1 \leq i \leq n$. Since, in slight variation of (6.23),

$$\begin{aligned} \gamma + 1 \pm x_i &= \frac{1}{2} \left((\gamma + 1) + (1 \pm x_i)^2 + \sum_{j \in \{1, \dots, n\} \setminus \{i\}} x_j^2 + (\gamma - \|x\|^2) \right) \\ &\in \text{QM}(\gamma - \|x\|^2), \end{aligned}$$

we see that $\tau - g \in P(\gamma - \|x\|^2)$. Setting $\gamma' = \gamma(1 + t/2)^2$, we then write

$$\begin{aligned} \gamma' - \|x\|^2 &= \gamma(1 + t/2)^2 - \|x\|^2 \\ &= (1+g)(\gamma - \|x\|^2) + g\|x\|^2 + \gamma(1+g)(\tau - g) + \gamma(\tau/2 - g)^2, \end{aligned}$$

and thus $\gamma' - \|x\|^2 \in \text{QM}(g, (1+g)(\gamma - \|x\|^2), \tau - g) \subset \text{QM}(g, (1+g)(\gamma - \|x\|^2))$. \square

Proof of Schmüdgen's Theorem. Let $\gamma > 0$ such that $\gamma - \|x\|^2$ is strictly positive on S . By Remark 6.5.5, there exist $G, H \in P(g_1, \dots, g_m)$ with $(1 + G)(\gamma - \|x\|^2) = (1 + H)$. Hence, $(1 + G)(\gamma - \|x\|^2) \in P(g_1, \dots, g_m)$. By Lemma 6.7.4, there exists some $\gamma' > 0$ with $\gamma' - \|x\|^2 \in P(g_1, \dots, g_m, (1 + G)(\gamma' - \|x\|^2))$, whence $\gamma' - \|x\|^2 \in P(g_1, \dots, g_m)$.

In view of this, it suffices to show that $p \in \text{QM}(g_1, \dots, g_m, \gamma' - \|x\|^2)$.

Using Lemma 6.7.4 again, there exists some $\tau > 0$ such that all the polynomials $\tau - x_i$ and $\tau - (-x_i) = \tau + x_i$, $1 \leq i \leq n$ are contained in $P(\gamma' - \|x\|^2)$. Since the set $C = \{x \in \mathbb{R}^n : \tau - x_i \geq 0, \tau + x_i \geq 0\}$ is the cube $[-\tau, \tau]^n$, it is bounded, so that Exercise 2 in Section 6.4 gives some $q \in \text{QM}(g_1, \dots, g_m)$ such that $p - q$ is strictly positive on C . Handelman's Theorem 6.6.3 then allows to write $p - q$ as a non-negative sum

$$p - q = \sum_{\beta} c_{\beta} g_1^{\beta_1} \cdots g_m^{\beta_m} \in P(\gamma' - \|x\|^2) = \text{QM}(\gamma' - \|x\|^2).$$

Hence, $p \in \text{QM}(g_1, \dots, g_m, \gamma' - \|x\|^2)$. □

Example 6.7.5. Let $g_1 = 1 - \sum_{i=1}^n x_i^2$, $g_2 = x_n$ and $S = S(g_1, g_2)$. If x_n is the vertical direction, the 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 $\sigma_0, \dots, \sigma_3 \in \Sigma[x]$.

Example 6.7.6. Let $g_1 = x_1 - 1/2$, $g_2 = x_2 - 1/2$, $g_3 = 1 - x_1 x_2$ and $p = 8 - x_1^2 - x_2^2$ as in Example 6.7.2. Since p is positive on $S(g_1, g_2, g_3)$, Theorem 6.7.3 asserts the existence of a Schmüdgen representation. To obtain one, first note that we can add the polynomial

$$g_4 = \frac{1}{2}g_1g_2 + 2g_3 = -x_1 - x_2 + 5/2$$

to the set of constraints, and thus, as in Example 6.7.2 also the box constraints $g_5 = 2 + x_1$, $g_6 = 2 - x_1$, $g_7 = 2 + x_2$, $g_8 = 2 - x_2$. Now the representation

$$p = 2g_1^2g_2 + 2g_1g_2^2 + 2g_1g_2 + 2g_1g_3 + 2g_2g_3 + 5/2g_6 + 5/2g_8$$

shows $p \in P(g_1, \dots, g_8) = P(g_1, g_2, g_3)$. Although p is strictly positive on $S(g_1, g_2, g_3)$ and thus contained in $P(g_1, g_2, g_3)$, p is not contained $\text{QM}(g_1, g_2, g_3)$, see Exercise 1.

Schmüdgen's Representation Theorem 6.7.3 is fundamental, but there could be 2^m terms in the sum (6.25). Putinar's Theorem gives a representation of p with a simpler structure, when we have more information about the representation of the compact basic semialgebraic set S .

Again let $g_1, \dots, g_m \in \mathbb{R}[x] = \mathbb{R}[x_1, \dots, x_n]$ and $S := S(g_1, \dots, g_m)$. 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 *Archimedean* if for every $h \in \mathbb{R}[x]$ there is some $N \in \mathbb{N}$ such that $N \pm h \in M$.

Theorem 6.7.7. *The following conditions are equivalent:*

1. *The quadratic module $\text{QM}(g_1, \dots, g_m)$ is Archimedean.*
2. *There exists an $N \in \mathbb{N}$ such that $N - \sum_{i=1}^n x_i^2 \in \text{QM}(g_1, \dots, g_m)$.*
3. *There exists an $h \in \text{QM}(g_1, \dots, g_m)$ such that $S(h)$ is compact.*
4. *There exist finitely many polynomials $h_1, \dots, h_r \in \text{QM}(g_1, \dots, g_m)$ such that $S(h_1, \dots, h_r)$ is compact and $\prod_{i \in I} h_i \in \text{QM}(g_1, \dots, g_m)$ for all $I \subseteq \{1, \dots, r\}$.*

Observe that if $h_1, \dots, h_r \in \text{QM}(g_1, \dots, g_m)$, then $S(g_1, \dots, g_m) \subset S(h_1, \dots, h_r)$.

Proof. The implications $1 \implies 2 \implies 3 \implies 4$ are obvious.

To show $4 \implies 1$, let $h_1, \dots, h_r \in \text{QM}(g_1, \dots, g_m)$ such that $S' := S(h_1, \dots, h_r)$ is compact and $\prod_{i \in I} h_i \in \text{QM}(g_1, \dots, g_m)$ for all $I \subseteq \{1, \dots, r\}$. By the compactness of S' , for any $h \in \mathbb{R}[x_1, \dots, x_n]$ there exists an $N \in \mathbb{N}$ such that $N + h > 0$ on S' and $N - h > 0$ on S' . Then Schmüdgen's Theorem 6.7.3 implies that both $N + h$ and $N - h$ have representations of the form $\sum_{e \in \{0,1\}^m} \sigma_e h_1^{e_1} \cdots h_r^{e_m}$ with sums of squares σ_e , $e \in \{0,1\}^m$. Since $\prod_{i \in I} h_i \in \text{QM}(g_1, \dots, g_m)$ for all $I \subset \{1, \dots, r\}$, we obtain $h \in \text{QM}(g_1, \dots, g_m)$. \square

The conditions in Theorem 6.7.7 are actually not conditions on the compact set S , but on its representation in terms of the polynomials g_1, \dots, g_m . See Exercise 1 for an example which shows that the conditions are stronger than just requiring that S is compact. In many practical applications, the precondition in Theorem 6.7.7 can be imposed by adding a witness of compactness, $N - \sum_{i=1}^n x_i^2 \geq 0$ for some $N > 0$.

Theorem 6.7.8 (Putinar). *Let $S = S(g_1, \dots, g_m)$ and suppose that $\text{QM}(g_1, \dots, g_m)$ is Archimedean. If a polynomial $f \in \mathbb{R}[x]$ is positive on S then $f \in \text{QM}(g_1, \dots, g_m)$. That is, there exist sums of squares $\sigma_0, \dots, \sigma_m \in \Sigma[x]$ with*

$$f = \sigma_0 + \sum_{i=1}^m \sigma_i g_i. \quad (6.27)$$

It is evident that each polynomial of the form (6.27) is nonnegative on S .

Proof. By Theorem 6.7.7, there exists some $h \in \text{QM}(g_1, \dots, g_m)$ such that the set $C := \{x \in \mathbb{R}^n : h(x) \geq 0\}$ is compact. Hence, by Exercise 2 in Section 6.4, there exists some $q \in \mathbb{R}[x_1, \dots, x_n]$ such that $p - q$ is strictly positive on C . Applying Schmüdgen's Theorem 6.7.3 yields $p - q \in P(h) = \text{QM}(h) \subset \text{QM}(g_1, \dots, g_m)$. \square

Example 6.7.9. The strict positivity in Putinar's Theorem is essential, even for univariate polynomials. This can be seen in the example $p = 1 - x^2$, $g = g_1 = (1 - x^2)^3$ (see Figure 6.3).

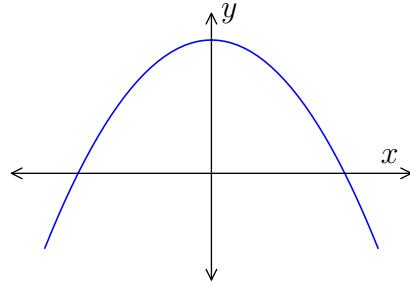


Figure 6.3: Graph of $p(x) = 1 - x^2$.

The feasible set S is the interval $S = [-1, 1]$, and hence the minima of the function $p(x)$ are at $x = -1$ and $x = 1$, both with function value 0. The precondition of Putinar's theorem satisfied since

$$\frac{2}{3} + \frac{4}{3}(x^3 - \frac{3}{2}x)^2 + \frac{4}{3}(1 - x^2)^3 = 2 - x^2.$$

If a representation of the form (6.27) existed, i.e.,

$$1 - x^2 = \sigma_0(x) + \sigma_1(x)(1 - x^2)^3 \quad \text{with } \sigma_0, \sigma_1 \in \Sigma[x], \quad (6.28)$$

then the right hand side of (6.28) must vanish at $x = 1$ as well. The second term has at 1 a zero of at least third order, so that σ_0 vanishes at 1 as well; by the SOS-condition this zero of σ_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 (6.28).

When p is nonnegative on a compact set $S = \{x \in \mathbb{R}^n : g_1(x) \geq 0, \dots, g_m(x) \geq 0\}$ then by Schmüdgen's Theorem the polynomial $p + \varepsilon$ is contained in $P(g_1, \dots, g_m)$. And similarly, if the module $\text{QM}(g_1, \dots, g_m)$ is Archimedean, $p + \varepsilon \in \text{QM}(g_1, \dots, g_m)$. However, for $\varepsilon \rightarrow 0$, the smallest degrees of those representations may be unbounded.

Exercises

- Let $g_1 = 2x_1 - 1$, $g_2 = 2x_2 - 1$, $g_3 = 1 - x_1 x_2$. Show that $S := \{x \in \mathbb{R}^2 : g_i(x) \geq 0, 1 \leq i \leq 3\}$ is compact but that $\text{QM}(g_1, g_2, g_3)$ is not Archimedean. (Hint: Example 6.4.7.)
- For $g_1 = x$, $g_2 = y$, $g_3 = 1 - x - y$, determine a Schmüdgen representation for the polynomial

$$p = -3x^2y - 3xy^2 + x^2 + 2xy + y^2$$

over the set $S = S(g_1, g_2, g_3)$.

6.8 Notes

Much material in this chapter is classical. Some standard references include the books of Basu, Pollack, and Roy [5], of Bochnak, Coste, and Roy [16] and of Prestel and Delzell [110].

The treatment on univariate polynomials is based on the exposition by Powers and Reznick [108]. Goursat's Lemma is due to Édouard Jean-Baptiste Goursat (1858–1936). It was shown by Bernstein [9] that every nonnegative polynomial on $[0, 1]$ has a non-negative representation in terms of the Bernstein polynomials (see also [105, vol. II, p. 83, Exercise 49]). The Pólya-Szegő Theorem 6.1.4 is given in [105, VI.45].

The elementary proof of Hilbert's Classification Theorem 6.2.3 is due to Choi and Lam [23], and likewise the polynomial in Exercise 3 in Section 6.2. In fact, Hilbert also showed that every ternary quartic can be written as the sum of at most three squares. Powers, Reznick, Scheiderer, and Ottlie have refined this by showing that for a general ternary quartic, there are 63 inequivalent representations as a sum of three squares over \mathbb{C} and 8 inequivalent representations as a sum of three squares over \mathbb{R} (Powers, Reznick, Scheiderer and Ottlie [109], see also its generalization with regard to varieties of minimal degree by Blekherman, Ottlie, Sinn and Ottlie [15]). Robinson showed that the cone $\Sigma_{n,d}$ is closed [116].

Theorem 6.2.6 on the discriminant of a symmetric matrix was shown by Ilyushechkin [65]. The version of Farkas' Lemma 5.4.5 in Exercise 3 can be found, for example, in Schrijver's book [123, Corollary 7.1h]. Non-negativity certificates based on sums of non-negative circuit polynomials have been studied by Ilman and de Wolff [64], the circuit number in Exercise 4 of Section 6.2 originates from that work.

We have only glimpsed into the theory of ordered fields. Theorem 6.3.3 on the extension of proper preorders to orders was shown by Artin and Schreier [3]. Textbook references are, for example, [16, Lemma 1.1.7], [83, § 20, Theorem 1], [88, Theorem 1.4.4], [110, Theorem 1.1.9].

Our treatment of quadratic modules and the Positivstellensatz is based on Marshall's book [88]. The Positivstellensatz is due to Krivine [73] and Stengle [135], for the historical development see [110]. The statement in Exercise 2 in Section 6.4 can be found in the work of Schweighofer [125] or Averkov [4].

Pólya's Theorem was proven by Pólya ([104], see also [47]). The special cases $n = 2$ and $n = 3$ were shown before by Poincaré [103] and Meissner [91]. For the case of strictly positive polynomials, Habicht has shown how to derive a solution to Hilbert's 17th problem from Pólya's Theorem [45].

Handelman's Theorem was proven in [46], our proof follows Schweighofer's and Averkov's derivation from Pólya's Theorem [4, 124, 125].

Theorem 6.7.1 was proven by Jacobi and Prestel in [66] and Theorem 6.7.3 was proven by Schmüdgen [121]. Lemma 6.7.4 goes back to Berr and Wörmann [10]. The equivalences on the Archimedean property in Theorem 6.7.7 were shown by Schmüdgen [121].

Chapter 7

Polynomial optimization

We consider the general problem of finding the infimum of a polynomial function p on a set S . When $S = \mathbb{R}^n$ this is unconstrained optimization. When it is constrained, so that $S \neq \mathbb{R}^n$, we will assume that S is a basic semialgebraic set, $S = S(g_1, \dots, g_m)$ given by polynomials $g_1, \dots, g_m \in \mathbb{R}[x] = \mathbb{R}[x_1, \dots, x_n]$.

The representation theorems for positivity from Chapter 6 offer an approach to this optimization problem. By the special case of Stengle's Positivstellensatz given in Corollary 6.5.4 (2), a polynomial p is nonnegative on the set $S(g_1, \dots, g_m)$ if and only if there exist a nonnegative integer k and polynomials G, H in the preorder generated by g_1, \dots, g_m that satisfy $pG = p^{2k} + H$. Minimizing a polynomial on such a set S is therefore equivalent to determining the largest number $\gamma \in \mathbb{R}$ such that the polynomial $p - \gamma$ has such a certificate. In this way, algebraic certificates for the nonnegativity of polynomials on a semialgebraic set S are linked to polynomial optimization.

A main issue in this approach is that most proofs of the Positivstellensatz are nonconstructive, in that they do not lead to a practical algorithm for a certificate. For example, the degrees of the required polynomials G , and H above can be infeasibly large. The best published bound is n -fold exponential.

Since the beginning of the 2000's, powerful interactions between positivity theorems and techniques from optimization have been revealed, which can be effectively applied to polynomial optimization. As we will see, under certain restrictions on the semialgebraic set S (such as polyhedrality or compactness), some of the positivity theorems are particularly suited for optimization problems.

We start by discussing linear programming relaxations to polynomial optimization based on Handelman's Theorem. This reveals the duality between positive polynomials and moments. Many approaches to polynomial optimization rely on semidefinite programming. We then present methods for unconstrained and constrained optimization based on semidefinite programming. This includes the Lasserre hierarchy of semidefinite relaxations, which builds upon Putinar's Positivstellensatz.

7.1 Linear programming relaxations

We consider the problem of finding the minimum of a polynomial function p over a compact basic semialgebraic set defined by affine polynomials g_1, \dots, g_m ,

$$p^* = \min\{p(x) \mid x \in S(g_1, \dots, g_m)\}.$$

In this case, the feasible set $S = S(g_1, \dots, g_m)$ is a polytope.

We may use linear programming to solve this problem. The minimal value p^* of p on S is the largest number λ such that $p - \lambda$ is nonnegative on S . For any $\varepsilon > 0$, the polynomial $p - p^* + \varepsilon$ is positive on the polytope S and by Handelman's Theorem 6.6.3, it has a representation of the form

$$p - p^* + \varepsilon = \sum_{\beta \in \mathbb{N}^m} c_\beta g^\beta \quad (7.1)$$

with nonnegative coefficients $c_\beta \geq 0$. If we knew the maximum degree of the monomials g^β in the polynomials g_i needed for this representation, then this becomes a linear program. *A priori*, we do not know this maximum degree.

We instead consider a sequence of relaxations obtained by degree truncations of (7.1). That is, for each $t \geq \deg(p)$ we consider the optimization problem

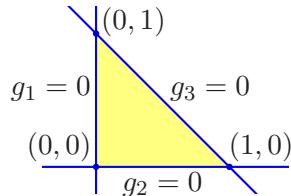
$$p_t^{\text{Han}} := \sup_{|\beta| \leq t} \left\{ \lambda \mid p - \lambda = \sum_{\beta} c_\beta g^\beta \right\}. \quad (7.2)$$

Comparing the coefficients in the representation for $p - \lambda$ gives linear conditions on the coefficient c_β and λ . With the positivity constraints that $c_\beta \geq 0$, the optimization problem (7.2) becomes a linear program. Linear programs are known to be solvable fast by Dantzig's simplex algorithm or by using interior-point methods (cf. Section 7.3).

Example 7.1.1. Consider minimizing a bivariate quadratic polynomial

$$p(x_1, x_2) = a_1 x_1^2 + a_2 x_2^2 + a_3 x_1 x_2 + a_4 x_1 + a_5 x_2 + a_6$$

over the triangle $S(g_1, g_2, g_3)$, where $g_1 = x_1$, $g_2 = x_2$, and $g_3 = 1 - x_1 - x_2$.



At order $t = 2$, comparing the coefficients in the degree truncation gives the linear program

to maximize λ under the constraints

$$\begin{aligned} a_1 &= c_{200} - c_{101} + c_{002}, \\ a_2 &= c_{020} - c_{011} + c_{002}, \\ a_3 &= c_{110} - c_{101} - c_{011} + 2c_{002}, \\ a_4 &= c_{100} - c_{001} + c_{101} - 2c_{002}, \\ a_5 &= c_{010} - c_{001} + c_{011} - 2c_{002}, \\ a_6 &= c_{000} + c_{001} + c_{002} + \lambda, \end{aligned}$$

where $c_\beta \geq 0$ for $|\beta| \leq 2$. Note that maximizing λ and the condition that $c_{000} \geq 0$ forces $c_{000} = 0$, so the constant term in the Handelman representation is redundant.

This sequence of linear programs converges.

Theorem 7.1.2. *Let $p \in \mathbb{R}[x] = \mathbb{R}[x_1, \dots, x_n]$ and $S = S(g_1, \dots, g_m)$ be a non-empty polytope given by affine polynomials g_1, \dots, g_m . Then the sequence $\{p_t^{\text{Han}} \mid t \in \mathbb{N}\}$ is increasing with limit the minimum p^* of p on S .*

Proof. On the compact set S , the polynomial p attains its infimum p^* . By construction of the truncations, for each $t \in \mathbb{N}$, $p_t^{\text{Han}} \in [-\infty, \infty)$ with $p_t^{\text{Han}} \leq p^*$, and the sequence $\{p_t^{\text{Han}} \mid t \in \mathbb{N}\}$ is monotone increasing. Since $p - p^* + \varepsilon$ is positive on S for any $\varepsilon > 0$, it has a Handelman representation of the form (6.22). Hence, there exists a truncation t with $p_t^{\text{Han}} \geq p^* - \varepsilon$, which proves the convergence. \square

We now consider a second series of linear programming relaxations. First observe that

$$p^* = \min_{x \in S} p(x) = \min_{\mu \in \mathcal{P}(S)} \int p(x) d\mu, \quad (7.3)$$

where $\mathcal{P}(S)$ is the set of all probability measures μ supported on the set S . Let $\mu \in \mathcal{P}(S)$. Writing $p = \sum_\alpha p_\alpha x^\alpha$, we have

$$\int p(x) d\mu = \sum_\alpha p_\alpha \int x^\alpha d\mu = \sum_\alpha p_\alpha y_\alpha,$$

where $y = (y_\alpha \mid \alpha \in \mathbb{N}^n)$ is the *moment sequence* of μ defined by

$$y_\alpha := \int x^\alpha d\mu.$$

Note that, since S is compact, any representing measure of a given moment sequence $(y_\alpha)_{\alpha \in \mathbb{N}^n}$ on S is determinate, that is, unique (see Appendix A.5). For some further background on moment sequences, see Appendix A.5.

We characterize the possible moment sequences. Let $L_y : \mathbb{R}[x_1, \dots, x_n] \rightarrow \mathbb{R}[y_\alpha \mid \alpha \in \mathbb{N}^n]$ denote the *Riesz functional* which maps a given polynomial to the linear form obtained by replacing every monomial with its associated moment variable,

$$L_y : p = \sum_\alpha p_\alpha x^\alpha \mapsto \sum_\alpha p_\alpha y_\alpha. \quad (7.4)$$

Theorem 7.1.3. Suppose that $g_1, \dots, g_m \in \mathbb{R}[x]$ are affine polynomials such that the polyhedron $S = S(g_1, \dots, g_m)$ is non-empty and bounded. An infinite sequence $(y_\alpha \mid \alpha \in \mathbb{N}^n)$ of real numbers is the moment sequence of a probability measure μ on the polytope S if and only if $y_0 = 1$ and

$$L_y(g^\beta) \geq 0 \quad \text{for all } \beta \in \mathbb{N}^m. \quad (7.5)$$

Proof. Let μ be a probability measure on S with moment sequence $(y_\alpha \mid \alpha \in \mathbb{N}^n)$. For any $\beta \in \mathbb{N}^m$, we have

$$L_y(g^\beta) = \int_S g^\beta d\mu \geq 0,$$

as g^β is nonnegative on S .

For the other direction, for $t \geq 0$ let $\Lambda_n(t)$ denote the index set $\Lambda_n(t) = \{\alpha \in \mathbb{N}^n \mid |\alpha| \leq t\}$. It is well-known (see Appendix A.5) that an infinite sequence $(y_\alpha \mid \alpha \in \mathbb{N}^n) \subset \mathbb{R}$ is the moment sequence of a Borel measure on the simplex $\Delta := \{x \in \mathbb{R}^n \mid x_i \geq 0, \sum_{i=1}^n x_i \leq 1\}$ if and only if for $t \geq 0$

$$\sum_{\alpha \in \Lambda_n(t)} (-1)^{|\alpha|} \begin{bmatrix} t \\ \alpha \end{bmatrix} y_{\alpha+\beta} \geq 0 \quad \text{for all } \beta \in \mathbb{N}^n, \quad (7.6)$$

where

$$\begin{bmatrix} t \\ \alpha \end{bmatrix} = \begin{bmatrix} t \\ \alpha_1 \dots \alpha_n \end{bmatrix} = \frac{t!}{\alpha_1! \alpha_2! \dots \alpha_n! (t - |\alpha|)!} = \binom{|\alpha|}{\alpha_1 \dots \alpha_n} \binom{t}{|\alpha|}$$

is the pseudo multinomial coefficient of dimension n and order t .

This is equivalent to the statement of the theorem when $S = \Delta$. Indeed, setting $g_i = x_i$, $1 \leq i \leq n$, $g_{n+1} = 1 - \sum_{i=1}^n x_i$, and $\beta = (\beta', \beta_{n+1})$, the quantity $L_y(g^\beta)$ becomes

$$L_y(g^\beta) = \sum_{\alpha \in \Lambda_n(\beta_{n+1})} \binom{|\alpha|}{\alpha_1 \dots \alpha_n} \binom{\beta_{n+1}}{|\alpha|} (-1)^{|\alpha|} y_{\alpha+\beta'},$$

where $|\alpha|$ records how many choices of a single term from $1 - x_1 - \dots - x_n$ fall into the subset of the last n terms in the expansion of $(1 - x_1 - \dots - x_n)^{\beta_{n+1}}$. A simple affine change of coordinates transforms this into the case of an arbitrary full-dimensional simplex.

For the general case, assume for simplicity that any $n+1$ of the polynomials g_1, \dots, g_m are affinely independent (otherwise consider suitable linear combinations of the g_i for the subsequent considerations). Then the polytope S can be written as $S = \bigcap_{i=1}^k \Delta_i$, where each Δ_i is a full-dimensional simplex defined by $n+1$ of the polynomials g_1, \dots, g_m , written as $S = \bigcap_{i=1}^k S(g_{i,1}, \dots, g_{i,n+1})$.

Let $y = (y_\alpha \mid \alpha \in \mathbb{N}^n)$ be a sequence of real numbers with $y_0 = 1$ and $L_y(g^\beta) \geq 0$ for all β . Then, for each $i \in \{1, \dots, k\}$, we have in particular $L_y(g_{i,1}^{\beta_{i,1}} \cdots g_{i,n+1}^{\beta_{i,n+1}}) \geq 0$ for each vector $(\beta_{i,1}, \dots, \beta_{i,n+1}) \in \mathbb{N}^{n+1}$. By the case already shown, (y_α) is the moment sequence of some probability measure μ_i on the simplex Δ_i . Now set $Q = \bigcup_{j=1}^k \Delta_j$, and

extend μ_i formally to a probability measure μ'_i on Q by setting $\mu_i(Q \setminus \Delta_i) = 0$. All the probability measures μ'_1, \dots, μ'_k have the same support. Since, by Appendix A.5, the moments determine the measure uniquely on the compact set Q , we have $\mu'_1 = \dots = \mu'_k =: \mu$. Moreover, μ has its support in each Δ_i and thus in the intersection $\bigcap_{i=1}^k \Delta_i = S$.

Hence, (y_α) is the moment sequence of the probability measure μ on the intersection S . \square

This method of moments also leads to a sequence of linear programming relaxations to this problem of finding p^* , based on degree truncations. For each $t \geq \deg(p)$, set

$$p_t^{\text{mom}} := \inf\{L_y(p) \mid y_0 = 1, L_y(g^\beta) \geq 0 \text{ for all } \beta \in \Lambda_m(t)\}. \quad (7.7)$$

This sequence of moment relaxations converges.

Theorem 7.1.4. *Let $p \in \mathbb{R}[x]$ be a polynomial and $g_1, \dots, g_m \in \mathbb{R}[x]$ be affine polynomials that define a nonempty polytope $S = S(g_1, \dots, g_m)$. Then $p_t^{\text{Han}} \leq p_t^{\text{mom}}$ for every t and*

$$\lim_{t \rightarrow \infty} p_t^{\text{Han}} = \lim_{t \rightarrow \infty} p_t^{\text{mom}} = p^*.$$

Proof. We show that the Handelman linear program relaxation of order t and the moment relaxation of order t are a primal-dual pair of linear programs. Recall that $\Lambda_n(t)$ is the set of exponents $\alpha \in \mathbb{N}^n$ of total degree at most t . Let $\Lambda_n(t)^\circ$ be those exponents that are not 0. Then the Handelman linear program of order t may be written as

$$\begin{aligned} p_t^{\text{Han}} &= \sup \lambda \\ \text{s.t. } & p - \lambda = \sum_{\beta \in \Lambda_m(t)^\circ} c_\beta g^\beta, \\ & c_\beta \geq 0 \text{ for } \beta \in \Lambda_m(t)^\circ. \end{aligned}$$

Using the moment variables y_α for $\alpha \in \Lambda_n(t)$, the moment linear program becomes

$$\begin{aligned} p_t^{\text{mom}} &= \inf \sum_{\alpha \in \Lambda_n(t)} p_\alpha y_\alpha \\ \text{s.t. } & L_y(g^\beta) \geq 0 \text{ for } \beta \in \Lambda_m(t)^\circ, \\ & y_0 = 1, \text{ and} \\ & y_\alpha \in \mathbb{R} \text{ for } \alpha \in \Lambda_n(t). \end{aligned}$$

Substituting $y_\alpha = -z_\alpha$ with new variables z_α for $\alpha \in \Lambda_n(t)$ and writing $\inf \sum_{\alpha \in \Lambda_n(t)} p_\alpha y_\alpha = -\sup \sum_{\alpha \in \Lambda_n(t)} p_\alpha z_\alpha$ shows that the linear program for p_t^{mom} is the dual of the linear program for p_t^{Han} . \square

The most important aspect of Theorem 7.1.4 is the duality of linear programs that is at the core of its proof.

Corollary 7.1.5. *The Handelman linear program at order t and the moment relaxation at order t form a pair of dual linear programs.*

We note that if p_t^{Han} is finite, then by strong duality of linear programming we have the equality $p_t^{\text{Han}} = p_t^{\text{mom}}$.

Example 7.1.6. As in Example 7.1.1, we minimize a quadratic polynomial

$$p(x_1, x_2) = a_1x_1^2 + a_2x_2^2 + a_3x_1x_2 + a_4x_1 + a_5x_2 + a_6$$

over the unit simplex. At order $t = 2$, the linear program for the moment relaxation is

$$\min a_1y_{20} + a_2y_{02} + a_3y_{11} + a_4y_{10} + a_5y_{01} + a_6,$$

with $y_{00} = 1$, the linear constraints for degree 1 are

$$y_{10} \geq 0, \quad y_{01} \geq 0, \quad 1 - y_{10} - y_{01} \geq 0,$$

and the linear constraints for degree 2 are

$$\begin{aligned} y_{20} &\geq 0, \quad y_{11} \geq 0, \quad y_{02} \geq 0, \\ y_{10} - y_{20} - y_{11} &\geq 0, \quad y_{01} - y_{11} - y_{02} \geq 0, \quad \text{and} \\ 1 + y_{20} + y_{02} - 2y_{11} - 2y_{10} - 2y_{01} &\geq 0. \end{aligned}$$

That moments are a dual to non-negative polynomials is a major theme in this chapter.

Exercises

1. Show that if p achieves its minimum at an interior point x^* of a polytope $S(g_1, \dots, g_m)$, so that $g_j(x^*) > 0$ for all j , then the Handelman hierarchy does not converge in finitely many steps.
2. Let $p_N := N(\sum_{i=1}^n x_i^n) - \prod_{i=1}^n x_i$ with $N > \frac{1}{n}$. The Handelman hierarchy of p_N over the feasible set $\Delta = \{x \in \mathbb{R}^n : x \geq 0, \sum_{i=1}^n x_i \leq 1\}$ does not converge in finitely many steps, although p_N attains its minimum on the boundary.
3. Show that the conditions (7.6) are necessary for a sequence $(y_\alpha)_{\alpha \in \mathbb{N}^n}$ to be the sequence of moments of some probability distribution on the standard simplex $\Delta_n = \{x \in \mathbb{R}_{\geq 0}^n \mid \sum_{i=1}^n x_i \leq 1\}$.
4. Show that for every $p \in \mathbb{R}[x_1, \dots, x_n]$ and for every $g_1, \dots, g_m \in \mathbb{R}[x_1, \dots, x_n]$ with $S(g_1, \dots, g_m)$ nonempty and bounded, there exists some $t_0 \geq \deg(p)$ such that for $t \geq t_0$, strong duality holds in Theorem 7.1.5.

7.2 Unconstrained optimization and sums of squares

We consider unconstrained polynomial optimization. That is, given a polynomial $p \in \mathbb{R}[x] = \mathbb{R}[x_1, \dots, x_n]$, determine its infimum on \mathbb{R}^n ,

$$p^* = \inf_{x \in \mathbb{R}^n} p(x).$$

The decision version of this problem asks whether $p(x) \geq \lambda$ for all $x \in \mathbb{R}^n$, where λ is a given constant. As in Section 7.1, the representation theorems for positivity in Chapter 6 lead to a useful series of relaxations. The idea is to replace the nonnegativity of $p - \lambda$ by the condition that $p - \lambda$ lies in the set $\Sigma[x]$ of sums of squares of polynomials. This gives the *SOS relaxation*,

$$\begin{aligned} p^{\text{sos}} &= \sup \lambda \\ \text{s.t. } p(x) - \lambda &\in \Sigma[x]. \end{aligned} \tag{7.8}$$

The optimal value p^{sos} is a lower bound for the global minimum of p (where we usually assume that this minimum is finite). And we will see in the next section that this relaxation (7.8) is computationally tractable.

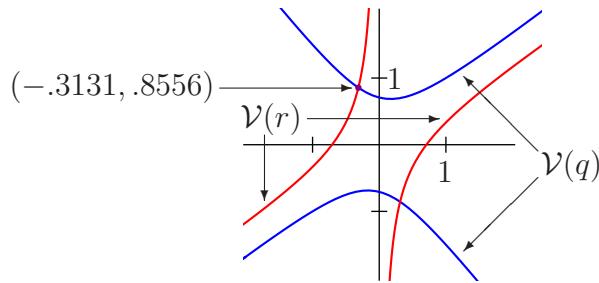
Example 7.2.1. To determine the infimum over \mathbb{R}^2 of the bivariate quartic polynomial

$$p(x, y) = 4x^4 - 8x^3y + x^2y^2 + 2xy^3 + 2y^4 + 2xy - 2y^2 + 2$$

we need only compute $p^{\text{sos}} = \sup\{\lambda \mid p(x, y) - \lambda \in \Sigma[x, y]\}$. Indeed, Hilbert showed (Theorem 6.2.3) that every nonnegative binary quartic may be written as a sum of at most three squares. Since we have

$$p(x, y) = \frac{1}{2}(2x^2 - 2y^2 - xy + 1)^2 + \frac{1}{2}(2x^2 - 3xy - 1)^2 + 1, \tag{7.9}$$

and the two quadratics $q = 2x^2 - 2y^2 - xy + 1$ and $r = 2x^2 - 3xy - 1$ have a common zero at $(0.3131, -0.8556)$, this infimum is 1.



In many instances in practical applications, $p^{\text{sos}} = p^*$, the global infimum. In particular, this holds for all the polynomials covered by Hilbert's Classification Theorem 6.2.3. We call the (nonnegative) difference $p^* - p^{\text{sos}}$ the *gap* of this SOS relaxation.

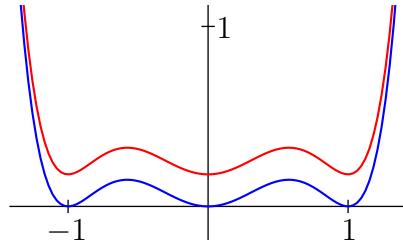
Example 7.2.2. Bivariate sextics with a nonzero gap may be constructed from versions of the Motzkin polynomial of Theorem 6.2.1. The homogeneous Motzkin polynomial is $M(x, y, z) := x^4y + x^2y^4 + z^6 - 2x^2y^2z^2$. Consider the dehomogenization

$$f(x, z) := M(x, 1, z) = x^4 + x^2 + z^6 - 3x^2z^2.$$

The global minimum of f is 0, which is attained at $(x, z) = (\pm 1, \pm 1)$. The sum of squares relaxation gives the lower bound $p^{\text{sos}} = -\frac{729}{4096} \approx 0.17798$, and a corresponding sum of squares decomposition is

$$f(x, z) + \frac{729}{4096} = \left(-\frac{9}{8}z + z^3\right)^2 + \left(\frac{27}{64} + x^2 - \frac{3}{2}z^2\right)^2 + \frac{5}{32}x^2.$$

We compare the graphs of f with $f + \frac{729}{4096}$ over the line $x = z$.



The dehomogenization back to the bivariate polynomial of Theorem 6.2.1,

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

has an infinite gap: There is no real number λ so that $p - \lambda$ is a sum of squares, see Exercise 2

For this relaxation to be useful in practice, we need a method to compute sums of squares decompositions. Let $p \in \mathbb{R}[x_1, \dots, x_n]$ be a polynomial of even degree $2d$. Let Y be the vector of all monomials in x_1, \dots, x_n of degree at most d —this has $\binom{n+d}{d}$ 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

$$p = Y^T \left(\sum_j s_j s_j^T \right) Y.$$

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. We record this observation.

Lemma 7.2.3. *A polynomial $p \in \mathbb{R}[x_1, \dots, x_n]$ 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 linear equations in the matrix variables Y of order $\binom{n+d}{d}$, with $\binom{n+2d}{2d}$ equations. For fixed d or n this size is polynomial.

Corollary 7.2.4. *For $n, d \geq 1$, the cone $\Sigma_{n,d}$ of homogeneous sum of squares polynomials of degree d in n variables is closed.*

Proof. Restricting the monomials in Y to monomials of degree exactly d , the statement is immediate as the cone S_n^+ of positive semidefinite $n \times n$ -matrices is closed. \square

Example 7.2.5. Continuing Example 7.2.1, we have

$$p(x, y) = (x^2, y^2, xy, 1) Q \begin{pmatrix} x^2 \\ y^2 \\ xy \\ 1 \end{pmatrix}$$

with a symmetric matrix $Q \in \mathbb{R}^{4 \times 4}$. Since Q must be positive semidefinite, there exists a decomposition $Q = LL^T$. One specific solution is

$$L = \frac{1}{\sqrt{2}} \begin{pmatrix} 2 & 2 & 0 \\ -2 & 0 & 0 \\ -1 & -3 & 0 \\ 1 & -1 & \sqrt{2} \end{pmatrix}, \quad \text{hence } Q = \begin{pmatrix} 4 & -2 & -4 & 0 \\ -2 & 2 & 1 & -1 \\ -4 & 1 & 5 & 1 \\ 0 & -1 & 1 & 2 \end{pmatrix}.$$

This implies the SOS decomposition (7.9).

For a homogeneous polynomial of degree $2d$, it suffices to consider homogeneous polynomials of degree d for the decomposition. See Exercise 3 for a more precise quantitative statement, taking into account the Newton polytope and thus sparsity.

Apparently, the SOS relaxation of an unconstrained polynomial optimization problem does not always give the infimum of the original optimization problem. In the subsequent sections of this chapter, we will intensively look at optimization aspects based on the general SOS idea, both in the situation of unconstrained and constrained optimization. Since the SOS approaches are strongly tied to the use of semidefinite programming, the next section will equip the reader with the basic concepts of semidefinite programming.

Exercises

1. Show that the polynomial $p(x, y) = (1 - xy)^2 + y^2$ has a finite infimum $\inf_{x \in \mathbb{R}^n} p(x, y)$, but that the infimum is not attained.

2. Show that for the inhomogeneous Motzkin polynomial $p(x, y) = x^4y^2 + x^2y^4 + 1 - 3x^2y^2$, there is no $\lambda \in \mathbb{R}$ so that $p - \lambda$ is a sum of squares.
3. If a polynomial $p \in \mathbb{R}[x]$ with Newton polytope $\text{New}(p)$ can be written as $\sum_{i=1}^m q_i^2$ then $\text{New}(q_i) \subset \frac{1}{2}\text{New}(p)$ for $1 \leq i \leq m$.
4. Show that the Robinson polynomial

$$R(x, y, z) = x^6 + y^6 + z^6 - (x^4y^2 + x^2y^4 + x^4z^2 + x^2z^4 + y^4z^2 + y^2z^4) + 3x^2y^2z^2$$

has minimal value 0, and determine its minimal points. Hint: Non-negativity of R can be verified quickly via the identity

$$(x^2 + y^2)R(x, y, z) = (x^4 - y^4 - x^2z^2 + y^2z^2)^2 + y^2z^2(y^2 - z^2)^2 + x^2z^2(x^2 - z^2)^2.$$

7.3 Semidefinite programming

We present some background on semidefinite programming. Our point of departure is linear programming which we have already met in Section 7.1. In the investigation of linear programs one usually starts from a normal form, such as

$$\begin{aligned} & \inf c^T x \\ \text{s.t. } & Ax = b, \\ & x \geq 0 \quad (x \in \mathbb{R}^n) \end{aligned} \tag{7.10}$$

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 different form, such as $\inf\{c^T x : Ax \leq b\}$, then by introducing additional variables it can be transformed into a normal form of type (7.10). Each linear program (7.10) has a *dual program*

$$\begin{aligned} & \sup b^T y \\ \text{s.t. } & A^T y + s = c, \\ & s \geq 0. \end{aligned} \tag{7.11}$$

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 are often solved in practice using Dantzig's simplex algorithm, despite its exponential time complexity in the worst case. Khachiyan's ellipsoid algorithm is a theoretically efficient method, having polynomial complexity. An alternative is Karmarkar's *interior point method*—this is efficient both theoretically and practically. Currently, interior point methods compete with the simplex algorithm for large problems.

The basic idea of primal-dual interior point methods can be explained as follows. Rather than aiming directly on a primal-dual pair $(x, (y, s))$ with Hadamard product 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 steps up to a given precision $\varepsilon > 0$. By the assumption of rational input data, these bounds then also imply the exact solvability in polynomial time.

From an abstract viewpoint the last inequality in (7.10) defines a cone. Thus linear programs can be seen as a special class of a *conic optimization problem*

$$\begin{aligned} & \inf c^T x \\ \text{s.t. } & Ax = b, \\ & x \in K \end{aligned} \tag{7.12}$$

with a cone K . Analogous to linear programs one can associate a dual program to (7.12):

$$\begin{aligned} & \sup b^T y \\ \text{s.t. } & A^T y + s = c, \\ & s \in K^*, \end{aligned} \tag{7.13}$$

where K^* denotes the dual cone to K . If K satisfies certain properties (closed, convex, pointed, nonempty interior), then for conic optimization problems a strong duality theorem holds as well, under the technical condition of strict feasibility. The cone \mathbb{R}_+^n is *self-dual*, that is, it is dual to itself.

Semidefinite programming which has been studied intensively since the 1990s, is a generalization of linear programming to matrix-based variables. First we choose a scalar product $\langle \cdot, \cdot \rangle$; usually the choice is $\langle A, B \rangle = \text{Tr}(A \cdot B)$, on which the Frobenius norm $\|A\| = (\sum_{i,j} a_{ij}^2)^{1/2}$ is defined. 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. Let $\text{Sym}_n(\mathbb{R})$ be the set of symmetric real $n \times n$ -matrices and S_n^+ the subset of positive semidefinite $n \times n$ -matrices.

A normal form of a *semidefinite program* is

$$\begin{aligned} & \inf \langle C, X \rangle \\ \text{s.t. } & \langle A_i, X \rangle = b_i, \quad 1 \leq i \leq m, \\ & X \succeq 0 \quad (X \in \text{Sym}_n(\mathbb{R})). \end{aligned} \tag{7.14}$$

A matrix $X \in \text{Sym}_n(\mathbb{R})$ is (*primal*) *feasible* if it satisfies 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.

Duality of semidefinite programs. To every semidefinite program of the form (7.14) one can associate a dual semidefinite program via

$$(D) \quad \begin{aligned} & \sup_{\substack{y, S \\ y \in \mathbb{R}^m \\ S \in S_n^+}} b^T y \\ & \sum_{i=1}^m y_i A_i + S = C, \\ & S \succeq 0, \quad y \in \mathbb{R}^m, \end{aligned}$$

whose optimal value is denoted by $\sup D^*$. The primal and dual feasible regions are denoted by \mathcal{P} and \mathcal{D} . The set of sets of optimal solutions are

$$\begin{aligned} \mathcal{P}^* &= \{X \in \mathcal{P} : \text{Tr}(CX) = p^*\}, \\ \mathcal{D}^* &= \{(S, y) \in \mathcal{D} : b^T y = d^*\}. \end{aligned}$$

One often makes the assumptions that A_1, \dots, 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 $S \in \mathcal{D}$ with $X \succ 0$ and $S \succ 0$. In particular, then Slater's condition from nonlinear programming is satisfied.

For $X \in \mathcal{P}$ and $(y, S) \in \mathcal{D}$, the difference

$$\text{Tr}(CX) - b^T y$$

is the *duality gap* of (\mathcal{P}) and (\mathcal{D}) in (X, y, S) .

Theorem 7.3.1. (Weak duality theorem for semidefinite program.) *Let $X \in \mathcal{P}$ and $(y, S) \in \mathcal{D}$. Then*

$$\text{Tr}(CX) - b^T y = \text{Tr}(SX) \geq 0.$$

Besides the weak duality statement, this theorem also gives an explicit description of the duality gap.

Proof. For any feasible solutions X and (y, S) of the primal and the dual program we evaluate the difference

$$\text{Tr}(CX) - b^T y = \text{Tr}\left(\left(\sum_{i=1}^m y_i A_i + S\right)X\right) - \sum_{i=1}^m y_i \text{Tr}(A_i X) = \text{Tr}(SX),$$

which is nonnegative by Féjer's Theorem A.3.11 on the positive semidefinite matrices S and X . \square

Theorem 7.3.2. (Strong duality theorem for semidefinite program). *Let $d^* < \infty$, and let the dual problem be strictly feasible. Then we have $\mathcal{P}^* \neq \emptyset$ and $p^* = d^*$.*

Analogously, if $p^ > -\infty$ and the primal problem is strictly feasible, then $\mathcal{D}^* \neq \emptyset$ and $p^* = d^*$.*

The strict feasibility condition in the strong duality theorem is a specialization of Slater's condition in convex programming.

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

$$M := \{S \in \text{Sym}_n(\mathbb{R}) \mid S = C - \sum_{i=1}^m y_i A_i, b^T y \geq d^*, y \in \mathbb{R}^m\}.$$

The idea is to separate this convex set from the set of positive semidefinite matrices. The proof is carried out in three steps.

We show that there exists a nonzero $Z \in \text{Sym}_n(\mathbb{R})$ with $\sup_{S \in M} \text{Tr}(SZ) \leq \inf_{U \in S_n^+} \text{Tr}(UZ)$. We first observe that $\text{relint}(M) \cap \text{relint}(S_n^+) = \emptyset$, because the existence of some $S \in M \cap S_n^{++}$ would contradict the optimal value d^* of (D).

Identify $\text{Sym}_n(\mathbb{R})$ with $\mathbb{R}^{\frac{1}{2}n(n+1)}$, where the scalar product is induced from the scalar product on $\text{Sym}_n(\mathbb{R})$ 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., [117, Cor. 11.4.1]) there exists a nonzero $Z \in \text{Sym}_n(\mathbb{R})$ with $\sup_{S \in M} \text{Tr}(SZ) \leq \inf_{U \in S_n^+} \text{Tr}(UZ)$. Since S_n^+ 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 $\inf_{U \in S_n^+} \text{Tr}(UZ) = 0$ (which by Féjer implies $Z \succeq 0$) yields $\sup_{S \in M} \text{Tr}(SZ) \leq 0$.

We show that there exists some $\beta > 0$ with $\text{Tr}(A_i Z) = \beta b_i$ for all $i \in \{1, \dots, m\}$. First observe that on the halfspace $\{y \in \mathbb{R}^m : b^T y \geq d^*\}$, the linear function $f(y) := \sum_{i=1}^m y_i \text{Tr}(A_i Z)$ is bounded from below (by $\text{Tr}(CZ)$).

Let $y \in \mathbb{R}^m$ (where y uniquely determines an $S \in M$) with $b^T y \geq d^*$. Then

$$f(y) = \sum_{i=1}^m y_i \text{Tr}(A_i Z) = -\text{Tr}((S - C)Z) = -\text{Tr}(SZ) + \text{Tr}(CZ) \geq \text{Tr}(CZ).$$

Hence there exists a $\beta \geq 0$ such that $\text{Tr}(A_i Z) = \beta b_i$ for all $i \in \{1, \dots, m\}$ (since otherwise one can make f smaller on the halfspace.)

Assuming $\beta = 0$ would imply $\text{Tr}(A_i Z) = 0$, $1 \leq i \leq m$, and therefore $\text{Tr}(CZ) \leq 0$. By assumption there exist a $(y^\circ, S^\circ) \in \mathcal{D}$ with $S^\circ \succ 0$. Hence,

$$\text{Tr}(S^\circ Z) = \text{Tr}(CZ) - \sum_{i=1}^m y_i^\circ \text{Tr}(A_i Z) = \text{Tr}(CZ) \leq 0.$$

This is a contradiction, since $Z \succeq 0$ and $S^\circ \succ 0$ imply that $\text{Tr}(S^\circ Z) > 0$ (due to Féjer, continuity, $Z \neq 0$). Hence, $\beta > 0$.

Finally, the goal is to show that for $X^* := \frac{1}{\beta}Z$ we have $X^* \in \mathcal{P}$ and $\text{Tr}(CX^*) = d^*$. We have $\text{Tr}(A_i X^*) = b_i$ for $1 \leq i \leq m$ i.e., $X^* \in \mathcal{P}$. Hence, $\text{Tr}(CX^*) \leq b^T y$ for all

$y \in \mathbb{R}^m$ with $b^T y \geq d^*$, and further $\text{Tr}(CX^*) \leq d^*$. The weak Duality Theorem implies $\text{Tr}(CX^*) = d^*$, 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. \square

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 $\begin{pmatrix} x_1 & 1 \\ 1 & x_2 \end{pmatrix} \succeq 0$, 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.

Algorithms. The interior point methods for linear programming can be efficiently extended to semidefinite programming. In the 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 $\mu > 0$.

Remark 7.3.3. We remark that the complexity of the (“exact”) semidefinite feasibility problem SDFP in the Turing machine model is still open, that is, it is not known whether SDFP is contained in the class **P** of problems which can be decided in polynomial time. This is one of the most important open problems concerning the complexity of semidefinite program. If the dimension n or the number of constraints m are constants, then SDFP is decidable in polynomial time. Hence, if n or m is fixed, then deciding an SOS decomposition can be done in polynomial time.

Exercises

1. Let $A \in \text{Sym}_n(\mathbb{R})$ be positive definite. Show that if $B \in \text{Sym}_n(\mathbb{R})$ is positive semidefinite with $\langle A, B \rangle = 0$ then $B = 0$.
2. 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.

3. Show that the semidefinite program 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 optimal primal value and its optimal dual value, but has a duality gap of 1.

4. Show that the semidefinite program 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 attained.

7.4 Unconstrained optimization and semidefinite programming

We continue our discussion of the unconstrained optimization problem

$$\inf\{p(x) : x \in \mathbb{R}^n\}$$

for a given polynomial $p \in \mathbb{R}[x]$ of even degree. Combining the insights of Sections 7.2 and 7.3, the relaxation $p^{\text{sos}} = \sup\{\gamma : p(x) - \gamma \in \Sigma[x]\}$ can be formulated as a semidefinite program.

Example 7.4.1. For the polynomial $p(x, y) = 4x^4 - 8x^3y + x^2y^2 + 2xy^3 + 2y^4 + 2xy - 2y^2 + 2$ of Example 7.2.1, the semidefinite program to compute p^{sos} is

$$\begin{aligned} & \sup \gamma \\ \text{s.t. } & q_{11} = 4, 2q_{13} = -8, 2q_{12} + q_{33} = 1, 2q_{23} = 2, q_{22} + \gamma = 2, \\ & 2q_{14} = 0, 2q_{34} = 2, 2q_{24} = -2, q_{44} = 2, \\ & Q \succeq 0 \quad (Q \in \text{Sym}_4(\mathbb{R})), \end{aligned}$$

where the rows columns of Q are indexed by the vector of monomials $(x^2, y^2, xy, 1)$.

Applying the duality theory of semidefinite programming, we can look at the unconstrained optimization problem from the dual point of view. As in Section 7.1, we consider the moments

$$y_\alpha = \int x^\alpha d\mu$$

for a probability measure μ on \mathbb{R}^n , and interpret them as the images of the monomial basis under a linear map $L: \mathbb{R}[x_1, \dots, x_n] \rightarrow \mathbb{R}$. That is, $y_\alpha = L(x^\alpha) = \int x^\alpha d\mu$. We

say that L is the *integration with respect to μ* on \mathbb{R}^n and observe that the linear map L coincides with the linearization operator L_y introduced in (7.4).

We record a straightforward necessary condition that a given linear map $L: \mathbb{R}[x_1, \dots, x_n] \rightarrow \mathbb{R}$ arises from a probability measure μ .

Theorem 7.4.2. *If a linear map $L: \mathbb{R}[x_1, \dots, x_n] \rightarrow \mathbb{R}$ is the integration with respect to a probability measure μ on \mathbb{R}^n then $L(0) = 1$ and the bilinear form*

$$\mathcal{L}: \mathbb{R}[x] \times \mathbb{R}[x] \rightarrow \mathbb{R}, \quad (p, q) \mapsto L(p, q)$$

is positive semidefinite.

We call the bilinear form \mathcal{L} the *moment form associated* with L .

Proof. If there exists a probability measure μ with $L(p) = \int p(x)d\mu$ for every $p \in \mathbb{R}[x]$ then

$$\mathcal{L}(p, p) = L(p^2) = \int p(x)^2 d\mu \geq 0,$$

since $p(x)^2 \geq 0$ for every $x \in \mathbb{R}^n$. □

There are various viewpoints on the quadratic form \mathcal{L} . Let $\mathcal{L}_{\leq t}$ be the restriction of \mathcal{L} to polynomials of degree at most t , and recall the notation $\Lambda_n(t) = \{\alpha \in \mathbb{N}^n : |\alpha| \leq t\}$. Then the representation matrix M_t of $\mathcal{L}_{\leq t}$ with respect to the monomial basis is given by

$$M_t = \mathcal{L}_{\leq t}(x^\alpha, x^\beta) = \mathcal{L}(x^\alpha, x^\beta) = L(x^{\alpha+\beta})$$

for $(\alpha, \beta) \in \Lambda_n(t) \times \Lambda_n(t)$. If the precondition of Theorem 7.4.2 is satisfied, then M_t is a positive semidefinite matrix for every $t \geq 0$. Similarly, rather than working with the quadratic form \mathcal{L} , we can think of its infinite representation matrix M whose rows and columns are indexed by \mathbb{N}^n and which is defined by $M_{\alpha, \beta} = \mathcal{L}(x^\alpha, x^\beta) = L(x^{\alpha+\beta})$. This matrix is a *moment matrix*, and the truncated version M_t is called a *truncated moment matrix*.

For a given polynomial $p = \sum_\alpha p_\alpha x^\alpha$ of even degree $2d$, observe that in the SOS condition for p^{sos} it suffices to consider sum of squares polynomials of degree at most $2d$. Moreover, define

$$p^{\text{mom}} := \inf \left\{ \sum_\alpha p_\alpha y_\alpha \mid M_d(y) \succeq 0 \right\}.$$

As shown by the following theorem, the semidefinite programs for p^{sos} and for p^{mom} can be viewed as a primal-dual pair, and there is no duality gap.

Theorem 7.4.3. *Given $p \in \mathbb{R}[x]$ of even degree $2d$, we have $p^{\text{sos}} = p^{\text{mom}}$. Moreover, if $p^{\text{mom}} > -\infty$ then the SOS relaxation has an optimal solution.*

Proof. Let $(A_\alpha)_{\beta,\gamma}$ be the coefficient of x^α in $x^{\beta+\gamma}$. The optimization problem for the sum of squares relaxation may be rephrased as

$$\begin{aligned} p^{\text{sos}} &= \sup \gamma \\ \text{s.t. } &\sum_{\alpha \in \Lambda_n(2d)} x^\alpha \langle A_\alpha, G \rangle = \sum_{\alpha \in \Lambda_n(2d)} p_\alpha x^\alpha - \gamma, \\ &G \in S_{\Lambda_n(d)}^+, \gamma \in \mathbb{R} \\ &= p_0 + \sup \langle -A_0, G \rangle \\ \text{s.t. } &\langle A_\alpha, G \rangle = p_\alpha, \quad 0 \neq \alpha \in \Lambda_n(2d), \\ &G \in S_{\Lambda_n(d)}^+, \gamma \in \mathbb{R}. \end{aligned}$$

Using the moment variables y_α for $0 \neq \alpha \in \Lambda_n(2d)$, the semidefinite program for p^{mom} can be written as

$$\begin{aligned} p^{\text{mom}} &= p_0 + \inf \sum_{\emptyset \neq \alpha \in \Lambda_n(2d)} p_\alpha y_\alpha \\ \text{s.t. } &A_0 + \sum_{0 \neq \alpha \in \Lambda_n(d)} y_\alpha A_\alpha \succeq 0, \quad 0 \leq j \leq m, \\ &y_\alpha \in \mathbb{R} \text{ for } 0 \neq \alpha \in \Lambda_n(d). \end{aligned}$$

Substituting $y_\alpha = -z_\alpha$ shows that the semidefinite program for $p^{\text{mom}} - p_0$ is the dual of the semidefinite program for $p^{\text{sos}} - p_0$.

We now exhibit a positive definite solution for the semidefinite program of p^{mom} , which then implies $p^{\text{sos}} = p^{\text{mom}}$ by the strong duality theorem for semidefinite programming. Let μ be a measure on \mathbb{R}^n with a strictly positive density function and with all moments finite. Then the moments given by

$$y_\alpha = \int x^\alpha d\mu > 0$$

provide a positive definite solution for the semidefinite program of p^{mom} .

Moreover, if $p^{\text{mom}} > -\infty$ then the Strong Duality Theorem 7.3.2 for semidefinite programming implies that the semidefinite program for p^{sos} has an optimal solution. \square

For the case that $p - p^*$ is SOS, the following investigations explain how to extract a minimizer for the polynomial optimization problem.

Theorem 7.4.4. *Let $p \in \mathbb{R}[x]$ be of degree at most $2d$ with global minimum p^* . If the non-negative polynomial $p - p^*$ is SOS, then $p^* = p^{\text{sos}} = p^{\text{mom}}$, and if x^* is a minimizer for p on \mathbb{R}^n , then the moment vector*

$$y^* = (x_\alpha^*)_{\alpha \in \Lambda_n(2d)}$$

is a minimizer of the moment relaxation.

Proof. If $p - p^*$ is SOS then $p^* = p^{\text{sos}}$, and by Theorem 7.4.2, we have $p^{\text{sos}} = p^{\text{mom}}$. By considering the Dirac measure δ_{x^*} , we see that y^* is a feasible point for the moment relaxation, and its objective value coincides with p^* . By the weak duality theorem for semidefinite programming, y^* is a minimizer of the moment relaxations. \square

The easiest case in extracting a minimizer is when the semidefinite program for p^{mom} has a minimal solution $M_d(y^*)$ with rank $M_d(y) = 1$ and such that $M_d(y^*)$ is of the form $y^* = v^*(v^*)^T$ for some moment sequence v^* of order up to d with regard to a Dirac measure δ_{x^*} . Then y^* is the sequence of moments up to order $2d$ of the Dirac measure δ_{v^*} of v^* .

More general cases of extracting the minimizer are related to flat extension conditions, see the treatment of constrained optimization in Section 7.7.

For the case of unconstrained optimization of a polynomial p , let us note that a principle improvement of the SOS relaxation would be to use representations based on rational functions, as exhibited in the solution to Hilbert's 17th problem via Theorem 6.5.6. For example, for the Motzkin polynomial $p(x, y) = M(x, y, 1)$, the rational representation

$$\begin{aligned} p(x, y) &= M(x, y, 1) \\ &= \frac{(x^2y - y)^2 + (xy^2 - x)^2 + (x^2y^2 - 1)^2 + \frac{1}{4}(xy^3 - x^3y)^2 + \frac{3}{4}(xy^3 + x^3y - 2xy)^2}{x^2 + y^2 + 1} \end{aligned} \tag{7.15}$$

provides a certificate for the non-negativity. In Exercise 3, the reader will write down a semidefinite program which could have been used to compute such a representation for $M(x, y, 1)$ as a sum of squares of rational functions.

However, to use the solution of Hilbert's problem to decide algorithmically the non-negativity of an arbitrarily given polynomial p , we have to know a degree bound on the numerator and the denominator in a possible representation of p as a sum of squares of rational functions. And, of course, the computational costs for solving the SDP will generally increase with the size of the degree bound.

Exercises

1. Is the polynomial $p(x, y) = 4x^2 - \frac{21}{10}x^4 + \frac{1}{3}x^6 + xy - 4y^2 + 4y^4$ non-negative?
2. Determine a good lower bound for the polynomial function $p : \mathbb{R}^{11} \rightarrow \mathbb{R}$,

$$\begin{aligned} p(x) &= \sum_{i=1}^{11} x_i^4 - 59x_9 + 45x_2x_4 - 8x_3x_{11} - 93x_1^2x_3 + 92x_1x_2x_7 \\ &\quad + 43x_1x_4x_7 - 62x_2x_4x_{11} + 77x_4x_5x_8 + 66x_4x_5x_{10} + 54x_4x_{10}^2 - 5x_7x_9x_{11}. \end{aligned}$$

3. Formulate a semidefinite program which computes a representation for (7.15) as a sum of squares of rational functions.

7.5 Duality and the moment problem

We have already seen in Sections 7.1–7.3 that the concept of duality is fundamental for optimization and that in case of polynomial optimization, it connects to the classical

moment problem. Before continuing with methods for constrained polynomial optimization, it is useful to have a geometric view on the dual cones of the cone of non-negative polynomials and the cone of sums of squares.

Throughout the section, assume that K is a closed set in \mathbb{R}^n , possibly $K = \mathbb{R}^n$. Given a sequence $y = (y_\alpha)_{\alpha \in \mathbb{N}^n}$, the moment problem asks for necessary and sufficient conditions on the sequence (y_α) concerning 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) is a convex cone. For fixed number of variables n and fixed degree d , the ambient space of this cone is finite-dimensional, while when the degree is unbounded, the ambient space has infinite dimension. Let

$$\begin{aligned} \mathcal{P}[x] &= \{p \in \mathbb{R}[x] \mid p(x) \geq 0 \text{ for all } x \in \mathbb{R}^n\}, \\ \Sigma[x] &= \{p \in \mathbb{R}[x] \mid p \text{ is SOS}\} \end{aligned}$$

be the set non-negative polynomials on \mathbb{R}^n and the set of sums of squares. These are convex cones in the infinite-dimensional vector space $\mathbb{R}[x]$.

We can identify an element $\sum_\alpha c_\alpha x^\alpha$ in the vector space $\mathbb{R}[x]$ with its coefficient vector (c_α) . The dual space of $\mathbb{R}[x]$ consists of the set of linear mappings on $\mathbb{R}[x]$ and each vector in the dual space can be identified with a vector in the infinite dimensional space $\mathbb{R}^{\mathbb{N}^n}$. Topologically, $\mathbb{R}^{\mathbb{N}^n}$ is a locally convex space in the topology of pointwise convergence. We identify the dual space of a space $X \subset \mathbb{R}^{\mathbb{N}^n}$ with a subspace of $\mathbb{R}^{\mathbb{N}^n}$.

To characterize the dual cone $\mathcal{P}[x]^*$, let \mathcal{M}_n denote the set of sequences $y = (y_\alpha)_{\alpha \in \mathbb{N}^n}$ admitting a representing Borel measure.

Theorem 7.5.1. *For $n \geq 1$, the cones $\mathcal{P}[x]$ and \mathcal{M}_n are dual to each other, i.e.,*

$$\mathcal{P}[x]^* = \mathcal{M}_n, \quad \mathcal{M}_n^* = \mathcal{P}[x].$$

A main goal of this section is to enlighten Theorem 7.5.1, and to prove most inclusions of it. As a warm-up example, it is very instructive to consider for $n = 1$ the cone

$$\mathcal{P}[x]_{\leq d} = \{p \in \mathcal{P}[x] : \deg p \leq d\}, \quad d \text{ even},$$

of non-negative univariate polynomials of even degree at most d . The ambient space of this cone is finite-dimensional. Let

$$\mathcal{M}_{1,d} = \{(y_0, \dots, y_d)^T \mid y_i = \int_{\mathbb{R}} x^i d\mu \text{ for some Borel measure } \mu \text{ on } \mathbb{R}\}$$

be the *truncated moment cone of order d*.

For simplicity, we identify a univariate polynomial p of degree d with its coefficient vector (p_0, \dots, p_d) . Denote by

$$\pi_d : \mathbb{R} \rightarrow \mathbb{R}^d, \quad t \mapsto (1, t, t^2, \dots, t^d)^T \in \mathbb{R}^{d+1}$$

the *moment mapping of order d*. We claim that $\mathcal{M}_{1,d} = \text{pos}\{\pi_d(t) \mid t \in \mathbb{R}\}$ for even d , where pos denotes the positive hull. Namely, both sets are convex cones and the inclusion from right to left is clear. And if there were a vector $(p_0, \dots, p_d) \in \mathcal{M}_{1,d} \setminus \text{pos}\{\pi_d(t) \mid t \in \mathbb{R}\}$, then then separating (p_0, \dots, p_d) from $\text{pos}\{\pi_d(t) \mid t \in \mathbb{R}\}$ gives some $(a_0, \dots, a_d) \in \mathbb{R}^{d+1} \setminus \{0\}$ with $\sum_{i=0}^d a_i t^i \geq 0$ for all t , but $\sum_{i=0}^d a_i p_i < 0$. But then $\sum_{i=0}^d a_i x^i$ would be a non-negative polynomial with $\int_{\mathbb{R}} a_i x^i d\mu < 0$ for some measure μ , which is a contradiction.

Lemma 7.5.2. *For $n = 1$ and even d , the dual cone $(\mathcal{P}[x]_{\leq d})^*$ satisfies $(\mathcal{P}[x]_{\leq d})^* = \text{cl } \mathcal{M}_{1,d}$ and $\mathcal{P}[x]_{\leq d} = \mathcal{M}_{1,d}^*$, where cl denotes the topological closure.*

Proof. By our prior considerations, for the first equality it suffices to prove

$$(\mathcal{P}[x]_{\leq d})^* = \text{cl pos}\{\pi_d(t) \mid t \in \mathbb{R}\}. \quad (7.16)$$

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 (7.16) by C , this implies $C \subset (\mathcal{P}[x]_{\leq d})^*$. Conversely, assume that there exists a $z \in (\mathcal{P}[x]_{\leq d})^* \setminus C$. By the Separation Theorem, there exists some $b \in \mathbb{R}^{d+1}$ 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 $\mathcal{P}[x]_{\leq d} = \mathcal{M}_{1,d}^*$, first observe that any $p \in \mathcal{M}_{1,d}^*$ satisfies $\sum_{i=0}^d p_i y_i \geq 0$ for all $\mathcal{M}_{1,d}$. 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 $t \in \mathbb{R}$, that is, p is non-negative. Conversely, let $p \in \mathcal{P}[x]_{\leq d}$ and $y \in \mathcal{M}_{1,d}$. Then

$$p^T y = \sum_{i=0}^d p_i y_i = \int_{\mathbb{R}} p(x) d\mu \geq 0,$$

which shows $p \in \mathcal{M}_{1,d}^*$. □

To provide an inequality characterization of the dual cone $(\mathcal{P}[x]_{\leq d})^*$, consider, for a given $z = (z_0, z_1, \dots, z_d)^T$, 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 7.5.3. *For even d , we have*

$$(\mathcal{P}[x]_{\leq d})^* = \{z \in \mathbb{R}^{d+1} : H_d(z) \succeq 0\}.$$

Proof. Let $C = \{z \in \mathbb{R}^{d+1} : H_d(z) \succeq 0\}$. In order to show $(\mathcal{P}[x]_{\leq d})^* \subset C$, we start from $(\mathcal{P}[x]_{\leq d})^* \subset \text{cl pos}\{\pi_d(t) : t \in \mathbb{R}\}$ stated in Theorem 7.5.2. 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 $(\mathcal{P}[x]_{\leq d})^* \subset C$.

Conversely, let $z \in C$ and p be an arbitrary non-negative 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

$$p^T z = \sum_j (q^{(j)})^T H_d(z) q^{(j)} \geq 0,$$

as $H_d(z)$ is positive semidefinite. □

We return to the general multivariate situation of Theorem 7.5.1.

Proof (of three out of the four inclusions in Theorem 7.5.1). We first consider the equality $\mathcal{M}_n^* = \mathcal{P}[x_1, \dots, x_n]$. 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, $p \in \mathcal{P}[x_1, \dots, x_n]$.

Conversely, let $p = \sum_\alpha c_\alpha x^\alpha \in \mathcal{P}[x_1, \dots, x_n]$. For each $y \in \mathcal{M}_n$ there exists a Borel measure μ with $y_\alpha = \int x^\alpha d\mu$, which implies

$$\sum_\alpha c_\alpha y_\alpha = \int p(x) d\mu \geq 0,$$

so that $p \in \mathcal{M}_n^*$.

Concerning the equality $\mathcal{M}_n = \mathcal{P}[x_1, \dots, x_n]^*$, 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 a proof. □

If for a Borel measure μ on a set K , a function $L : \mathbb{R}[x] \rightarrow \mathbb{R}$ is defined as $L(p) = \int_K p d\mu$ for all $p \in \mathbb{R}[x]$, then $L(p) \geq 0$ for all polynomials p which are non-negative on K . Haviland's Theorem establishes the converse, characterizing linear maps coming from a measure.

Theorem 7.5.4 (Haviland). *Let $K \subset \mathbb{R}^n$ be a measurable set. For a linear map $L: \mathbb{R}[x] \rightarrow \mathbb{R}$, 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 .*

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 $\Sigma[x]$. In order to characterize the dual cone $(\Sigma[x])^*$, consider a real sequence $(y_\alpha)_{\alpha \in \mathbb{N}^n}$ indexed by non-negative integer vectors. Let

$$\mathcal{M}_n^+ := \{y = (y_\alpha)_{\alpha \in \mathbb{N}^n} \mid M(y) \succeq 0\},$$

where $M(y)$ is the (infinite) moment matrix $(M(y))_{\mathbb{N}^n \times \mathbb{N}^n}$ with $(M(y))_{\alpha, \beta} = y_{\alpha+\beta}$. Observe that for univariate polynomials, $M(y)$ is an infinite Hankel matrix,

$$M(y) = \begin{pmatrix} y_0 & y_1 & y_2 & \cdots \\ y_1 & y_2 & & \\ y_2 & & \ddots & \\ \vdots & & & \end{pmatrix}.$$

Lemma 7.5.5. *For $n \geq 1$, it holds $\mathcal{M}_n \subset \mathcal{M}_n^+$.*

For $t \in \mathbb{N}^n$, recall that $(y_\alpha)_{\alpha \in \Lambda_n(t)}$ denotes the sequence of moments of μ up to order t .

Proof. Let μ be a representing measure for $y \in \mathcal{M}_n$. Since for any $t \in \mathbb{N}$ and for any $p(x) \in \mathbb{R}[x]$ with degree $\leq t$, we have

$$\begin{aligned} \mathcal{L}_{\leq t}(p, p) &= \sum_{\alpha, \beta \in \Lambda_n(t)} p_\alpha p_\beta y_{\alpha+\beta} = \sum_{\alpha, \beta \in \Lambda_n(t)} p_\alpha p_\beta \int x^{\alpha+\beta} d\mu \\ &= \int p(x)^2 d\mu \geq 0, \end{aligned}$$

the statement follows. \square

We record the following result, whose proof is partially covered in the exercises.

Theorem 7.5.6. *The cones $\Sigma[x]$ and \mathcal{M}_n^+ are dual to each other, i.e.*

$$\Sigma[x]^* = \mathcal{M}_n, \quad (\mathcal{M}_n^*)^* = \Sigma[x].$$

We obtain the following corollary.

Corollary 7.5.7 (Hamburger). *The cones \mathcal{M}_1 and \mathcal{M}_1^+ coincide. For $n \geq 2$, we have $\mathcal{M}_n \neq \mathcal{M}_n^+$.*

Proof. By Hilbert's Classification 6.2.3, the inclusion $\Sigma[x_1, \dots, x_n] = \mathcal{P}[x_1, \dots, x_n]$ is strict exactly for $n \geq 2$. Hence, Theorem 7.5.1 and 7.5.6 imply

$$\mathcal{M}_1 = \mathcal{P}[x_1]^* = \Sigma[x_1]^* = \mathcal{M}_1^+$$

and for $n \geq 2$

$$\mathcal{M}_n = \mathcal{P}[x_1, \dots, x_n]^* \subsetneq \Sigma[x_1, \dots, x_n]^* = \mathcal{M}_n^+.$$

□

Exercises

1. For even d , show that $\mathcal{M}_{1,d}$ is not closed. Hint: Consider the moment sequence $(1, \varepsilon, 1/\varepsilon)$ coming from a normal distribution with mean ε and variance $\varepsilon - 1/\varepsilon^2$.
2. 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 $\mathcal{M}_{1,d}$.
3. Prove the equality $\Sigma[x]^* = \mathcal{M}_n$ and the inclusion $\Sigma[x] \subset (\mathcal{M}_n^*)^*$.

7.6 Optimization over compact sets

Consider constrained polynomial optimization problems of the form

$$\begin{aligned} p^* &= \inf p(x) \\ \text{s.t. } g_j(x) &\geq 0, \quad 1 \leq j \leq m, \\ x &\in \mathbb{R}^n. \end{aligned} \tag{7.17}$$

For convenience we set $g_0 = 1$ and usually, we will assume that the basic semialgebraic set $K = S(g_1(x), \dots, g_m)$ is compact. From the viewpoint of computational complexity these problems are in general **NP-hard**. Namely, for example, the partition problem belongs to this class: Given $a_1, \dots, a_m \in \mathbb{N}$, does there exist an $x \in \{-1, 1\}^n$ with $\sum x_i a_i = 0$?

We present Lasserre's hierarchy for approaching these problems via semidefinite programs. This technique can both be approached from the viewpoint of non-negative polynomials and from the viewpoint of moments, which turn out to be dual to each other.

Assume that the quadratic module $\text{QM}(g_1, \dots, g_m)$ is Archimedean as introduced in Section 6.7. Hence, there exists an $N \in \mathbb{N}$ with $N - \sum_{i=1}^n x_i^2 \in \text{QM}(g_1, \dots, g_m)$. From an applied viewpoint this is usually 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. In case of an Archimedean module, the feasible set K is compact, and Putinar's Positivstellensatz implies that every polynomial p which is strictly positive on K is contained in $\text{QM}(g_1, \dots, g_m)$. The nice point about working with $\text{QM}(g_1, \dots, g_m)$ is that it introduces some convexity structure into the constrained optimization problem.

However, the infinite-dimensional cone $\text{QM}(g_1, \dots, g_m)$ cannot be handled easily from a practical point of view. By restricting the degrees we replace it by a hierarchy of finite-dimensional cones.

Namely, denote by $\text{QM}_{2t}(g_1, \dots, g_m)$ the *truncated quadratic module*

$$\text{QM}_{2t}(g_1, \dots, g_m) = \left\{ \sum_{j=0}^m s_j g_j \mid s_0, \dots, s_m \in \Sigma[x], \deg(s_j g_j) \leq 2t \text{ for } 0 \leq j \leq m \right\}.$$

For

$$t \geq \max \left\{ \left\lceil \frac{\deg(f)}{2} \right\rceil, \left\lceil \frac{\deg(g_1)}{2} \right\rceil, \dots, \left\lceil \frac{\deg(g_m)}{2} \right\rceil \right\}, \quad (7.18)$$

we consider the sequence of *SOS relaxations of order t* defined by

$$\begin{aligned} p_t^{\text{sos}} &= \sup \gamma \\ &\text{s.t. } p - \gamma \in \text{QM}_{2t}(g_1, \dots, g_m) \\ &= \sup \gamma \\ &\text{s.t. } p - \gamma = \sum_{j=0}^m s_j g_j \text{ for some } s_j \in \Sigma[x] \text{ with } \deg(s_j g_j) \leq 2t. \end{aligned} \quad (7.19)$$

As we will see below, this problem can be phrased as a semidefinite program.

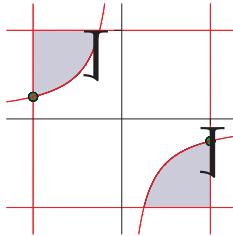


Figure 7.1: Feasible region and optimal points of a two-dimensional polynomial optimization problem

Example 7.6.1. Let $g_1 = x + 1$, $g_2 = 1 - x$, $g_3 = y + 1$, $g_4 = 1 - y$ and $g_5 = -xy - \frac{1}{4}$, and the goal is to minimize $p = y^2$ over $S(g_1, \dots, g_5)$. See Figure 7.1. By the proof of Theorem 6.7.1, the quadratic module $\text{QM}(g_1, \dots, g_5)$ is Archimedean. Hence, for $t \geq 1$, (7.19) will give a sequence of lower bounds for the optimal value p^* . Specifically, for $t = 1$, the corresponding truncated quadratic module is

$$\begin{aligned} \text{QM}_{2t}(g_1, \dots, g_5) &= \text{QM}_2(g_1, \dots, g_5) \\ &= \Sigma[x, y] + [0, \infty)(x + 1) + [0, \infty)(1 - x) + \\ &\quad + [0, \infty)(y + 1) + [0, \infty)(1 - y) + [0, \infty) \left(-xy - \frac{1}{4} \right). \end{aligned}$$

Similar to the discussion of the LP relaxation based on Handelman's Theorem, there is also a dual point of view to convexify the optimization problem. Using the language of moments, consider

$$p^* = \min_{x \in K} p(x) = \min_{\mu \in \mathcal{P}(K)} \int p(x) d\mu, \quad (7.20)$$

where $\mathcal{P}(K)$ denotes the set of all probability measures μ supported on the set K .

As earlier, we identify a sequence $y = (y_\alpha)_{\alpha \in \mathbb{N}^n}$ with a linear map $L : \mathbb{R}[x] \rightarrow \mathbb{R}$. In order to characterize those measures whose support is contained in some given set K , the following lemma will be helpful.

Lemma 7.6.2. *For a linear map $L : \mathbb{R}[x] \rightarrow \mathbb{R}$ 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 \mathcal{L}_g defined by

$$\mathcal{L}_g : \mathbb{R}[x] \times \mathbb{R}[x] \rightarrow \mathbb{R}, \quad (q, r) \mapsto L(qrg)$$

is positive semidefinite.

The symmetric bilinear form \mathcal{L}_g is called the *localization form* with respect to g .

Proof. The first two conditions are equivalent, by linearity. For the equivalence of (2) and (3), observe that $\mathcal{L}_{g,g} = L(h^2 g)$. \square

We are interested in necessary conditions that a given sequence is the sequence of moments supported on the feasible set K . To illuminate the localization form in this context, take a measure μ whose support set is contained in $\{x \in \mathbb{R}^n : g(x) \geq 0\}$, with corresponding moment sequence y . Then, for any $q \in \mathbb{R}[x]$, we have $\mathcal{L}_g(q, q) = \int g(x)q(x)^2 d\mu \geq 0$. We call the restriction $\mathcal{L}_{g,\leq t}$ of \mathcal{L}_g to $\mathbb{R}[x]_{\leq t} \times \mathbb{R}[x]_{\leq t}$ the *truncated localization form*.

Theorem 7.6.3. *Let $g \in \mathbb{R}[x]$.*

1. *If a real sequence $(y_\alpha)_{\alpha \in \mathbb{N}^n}$ is the sequence of moments of a measure μ supported on the set $S = \{x \in \mathbb{R}^n : g(x) \geq 0\}$ then \mathcal{L}_g is positive semidefinite.*
2. *If a real sequence $(y_\alpha)_{\alpha \in \Lambda_n(2t)}$ (with $t \geq \lceil \deg(g)/2 \rceil$) is the sequence of moments up to order $2t$ of a measure μ supported on the set $S = \{x \in \mathbb{R}^n : g(x) \geq 0\}$ then $\mathcal{L}_{g,\leq t-\lceil \deg(g)/2 \rceil}$ is positive semidefinite.*

Proof. Let $g \in \mathbb{R}[x]$. Then for any polynomial $q \in \mathbb{R}[x]$ of degree at most $t - \lceil \deg(g)/2 \rceil$, we have

$$\mathcal{L}_g(q, q) = \int g(x)q(x)^2 d\mu \geq 0.$$

The second statement is just the truncated version, where only polynomials $q \in \mathbb{R}[x]$ of degree at most $t - \lceil \deg(g)/2 \rceil$ are considered. \square

Primal-dual semidefinite relaxations. We now deduce the sequence of primal-dual semidefinite relaxations for the optimization problem (7.17). The SOS relaxation of order t and the moment relaxation of order t can be interpreted as a primal-dual pair of semidefinite programs. To see this, recall that $\Lambda_n(t) = \{\alpha \in \mathbb{N}^n : |\alpha| \leq t\}$ and set $d_j = \max\{d \in \mathbb{N} : 2d + \deg g_j \leq t\}$, $1 \leq j \leq m$. Further let $(A_{\alpha j})_{\beta, \gamma}$ be the coefficient of x^α in $x^{\beta+\gamma}g_j$ and $A = (A_{\alpha j})_{\beta, \gamma \in \Lambda_n(d_j)}$. The primal problem can be rephrased as

$$\begin{aligned} p_t^{\text{sos}} &= \sup \gamma \\ \text{s.t. } &\sum_{\alpha \in \Lambda_n(2t)} x^\alpha \sum_{j=0}^m \langle A_{\alpha j}, G_j \rangle, = \sum_{\alpha \in \Lambda_n(2t)} p_\alpha x^\alpha - \gamma, \\ &G_0, \dots, G_m \in S_{\Lambda_n(d_j)}^+, \gamma \in \mathbb{R} \\ &= p_0 + \sup \sum_{j=0}^m \langle -A_{0j}, G_j \rangle \\ \text{s.t. } &\sum_{i=0}^m \langle A_{\alpha i}, G_i \rangle = p_\alpha, \quad 0 \neq \alpha \in \Lambda_n(2t), \\ &G_0, \dots, G_m \in S_{\Lambda_n(d_j)}^+, \gamma \in \mathbb{R}. \end{aligned}$$

On the dual side, we consider the sequence of *moment relaxations of order t* defined by

$$\begin{aligned} p_t^{\text{mom}} &= \inf L(p) \\ \text{s.t. } &L \in (\mathbb{R}[x]_{2t})^* \text{ with } L(1) = 1, L(p) \geq 0 \text{ for all } f \in \text{QM}_{2t}(g_1, \dots, g_m) \\ &= \inf \sum_\alpha p_\alpha y_\alpha \\ \text{s.t. } &L(1) = 1, \mathcal{L}_{t-\deg[g_j/2]} \text{ positive semidefinite, } \quad 0 \leq j \leq m. \end{aligned} \tag{7.21}$$

And, using the moment variables y_α for $0 \neq \alpha \in \Lambda_n(2t)$, the moment relaxation can be written as

$$\begin{aligned} p_t^{\text{mom}} &= p_0 + \inf \sum_{\emptyset \neq \alpha \in \Lambda_n(2t)} p_\alpha y_\alpha \\ \text{s.t. } &A_{0j} + \sum_{0 \neq \alpha \in \Lambda_n(t)} y_\alpha A_{\alpha j} \succeq 0, \quad 0 \leq j \leq m, \\ &y_\alpha \in \mathbb{R} \text{ for } 0 \neq \alpha \in \Lambda_n(2t). \end{aligned}$$

Substituting $y_\alpha = -z_\alpha$ then shows that the semidefinite program for $p_t^{\text{mom}} - p_0$ is precisely the dual of the semidefinite program for $p_t^{\text{sos}} - p_0$.

Example 7.6.4. For $n = 2$ the following truncated piece of the moment matrix $M(y)$

refers to the monomials $\{x^{(0,0)}, x^{(1,0)}, x^{(0,1)}, x^{(2,0)}, x^{(1,1)}, x^{(0,2)}\}$ of total degree at most 2,

$$\left(\begin{array}{c|cc|ccc} 1 & y_{10} & y_{01} & y_{20} & y_{11} & y_{02} \\ \hline y_{10} & y_{20} & y_{11} & y_{30} & y_{21} & y_{12} \\ y_{01} & y_{11} & y_{02} & y_{21} & y_{12} & y_{03} \\ \hline y_{20} & y_{30} & y_{21} & y_{40} & y_{31} & y_{22} \\ y_{11} & y_{21} & y_{12} & y_{31} & y_{22} & y_{13} \\ y_{02} & y_{12} & y_{03} & y_{22} & y_{13} & y_{04} \end{array} \right).$$

Example 7.6.5. We continue Example 7.6.1. For $t = 1$, the moment relaxation gives

$$\begin{aligned} p_t^{\text{mom}} &= \inf y_{02} \\ &\left(\begin{array}{c|cc} 1 & y_{10} & y_{01} \\ \hline y_{10} & y_{20} & y_{11} \\ y_{01} & y_{11} & y_{02} \end{array} \right) \succeq 0, \\ &-1 \leq y_{10} \leq 1, \\ &-1 \leq y_{01} \leq 1, \\ &y_{11} \leq -\frac{1}{4}. \end{aligned} \tag{7.22}$$

The positive semidefiniteness condition on the truncated moment matrix implies that the diagonal element y_{02} is non-negative in any feasible solution. Hence, $p_t^{\text{sos}} \geq 0$. Considering the moments

$$y_{10} = 0, \quad y_{01} = 0, \quad y_{11} = -\frac{1}{4}, \quad y_{20} = \frac{1}{4}\varepsilon^2, \quad y_{02} = \frac{1}{4}\frac{1}{\varepsilon^2}$$

for $\varepsilon > 0$ gives feasible solutions, which shows that $p_t^{\text{mom}} = 0$.

Example 7.6.6. With the notation $y_{ij} = L(x_1^i x_2^j)$ introduced above and $g_0(x_1, x_2) = 1$, the polynomial $g_1 = -4x_1^2 + 7x_1 \geq 0$ has a localizing form \mathcal{L}_{g_1} whose initial 3×3 -submatrix of the representation matrix is

$$\left(\begin{array}{c|cc} -4y_{20} + 7y_{10} & -4y_{30} + 7y_{20} & -4y_{21} + 7y_{11} \\ \hline -4y_{30} + 7y_{20} & -4y_{40} + 7y_{30} & -4y_{31} + 7y_{21} \\ -4y_{21} + 7y_{11} & -4y_{31} + 7y_{21} & -4y_{22} + 7y_{12} \end{array} \right).$$

Although each of the relaxation values might not be optimal for the original problem, one has the following convergence result.

Theorem 7.6.7. Let $p, g_1, \dots, g_m \in \mathbb{R}[x]$ and $K = \{x \in \mathbb{R}^n : g_j(x) \geq 0\}$.

1. For each admissible t we have $p_t^{\text{sos}} \leq p_t^{\text{mom}}$.
2. If $\text{QM}(g_1, \dots, g_m)$ is Archimedean, then the sequences (p_t^{sos}) and (p_t^{mom}) are monotone non-decreasing with

$$\lim_{t \rightarrow \infty} p_t^{\text{sos}} = \lim_{t \rightarrow \infty} p_t^{\text{mom}} = p^*.$$

Proof. The first statement immediately follows from weak duality.

For the second statement, 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 (6.27). Hence, there exists a t with $p_t^{\text{sos}} \geq p^* - \varepsilon$. Passing over to the limit $\varepsilon \downarrow 0$, this shows the claim. \square

For t satisfying (7.18) this defines a hierarchy of semidefinite programs whose optimal values converges monotonically 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 p_t^{mom} obtained in the t -th relaxation is the optimal value is not easy. We will discuss a sufficient condition in Section 7.7.

Example 7.6.8. For $n \geq 2$ we consider the (parametric) optimization problem

$$\min \sum_{i=1}^{n+1} x_i^4 \quad \text{s.t. } \sum_{i=1}^{n+1} x_i^3 = 0, \quad \sum_{i=1}^{n+1} x_i^2 = 1, \quad \sum_{i=1}^{n+1} x_i = 0 \quad (7.23)$$

in the n variables x_1, \dots, 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 (7.23), 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 (7.23) there exists a simple polynomial identity

$$\sum_{i=1}^{n+1} x_i^4 - \frac{1}{n+1} = \frac{2}{n+1} \left(\sum_{i=1}^{n+1} x_i^2 - 1 \right) + \sum_{i=1}^{n+1} \left(x_i^2 - \frac{1}{n+1} \right)^2, \quad (7.24)$$

which shows that the minimum is bounded from below by $1/(n+1)$; and since this value is attained at $x_1 = \dots = x_{(n+1)/2} = -x_{(n+3)/2} = \dots = -x_{n+1} = 1/\sqrt{n+1}$, the minimum is $1/(n+1)$. For each $\varepsilon > 0$ adding ε on both sides of (7.24) yields a representation of the positive polynomial in the quadratic module $\text{QM}(g_1, \dots, g_6)$. 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.

We remark that Putinar's Positivstellensatz has the following direct counterpart in the dual setting of moments.

Theorem 7.6.9 (Putinar). *Suppose the quadratic module $\text{QM}(g_1, \dots, g_m)$ is Archimedean. A linear map $L : \mathbb{R}[x] \rightarrow \mathbb{R}$ is the integration with respect to a probability measure μ on K , i.e.,*

$$\exists \mu \forall p \in \mathbb{R}[x] : L(p) = \int_K p d\mu, \quad (7.25)$$

if and only if $L(1) = 1$ and all the bilinear forms

$$\mathcal{L}_{g_i} : \mathbb{R}[x] \times \mathbb{R}[x] \rightarrow \mathbb{R}, \quad (q, r) \mapsto L(q \cdot r \cdot g_i) \quad (7.26)$$

$(0 \leq j \leq m)$ are positive semidefinite.

Applying Theorem 7.6.9, we can restate (7.3) as

$$p^* = \inf\{L(p) : L : \mathbb{R}[x] \rightarrow \mathbb{R} \text{ linear, } L(1) = 1 \text{ and each } \mathcal{L}_{g_i} \text{ is psd}\}. \quad (7.27)$$

Since every linear map $L : \mathbb{R}[x] \rightarrow \mathbb{R}$ is given by the values $L(x^\alpha)$ on the monomial basis $(x^\alpha)_{\alpha \in \mathbb{N}^n}$, Theorem 7.6.9 characterizes the families $(y_\alpha)_{\alpha \in \mathbb{N}^n}$ which arise as the sequences of moments of a probability measure on K , i.e., $y_\alpha = \int_K x^\alpha d\mu$ for every $\alpha \in \mathbb{N}^n$. Therefore Theorem 7.6.9 is also said to solve the *moment problem* on K . The proof of this theorem can be deduced from Putinar's Positivstellensatz 6.7.8. However, for the proof of that untruncated version, tools from functional analysis are required, such as Riesz's Representation Theorem.

Exercises

1. Show that the semidefinite condition (7.26) in Theorem 7.6.9 can be equivalently replaced by the condition $L(M) \subset [0, \infty)$.
2. Show the following moment matrix version of Schmüdgen's Theorem 6.7.3.

An infinite sequence $y = (y_\alpha)_{\alpha \in \mathbb{N}^n}$ of real numbers is the moment sequence of some measure μ supported on K if and only if the moment forms $\mathcal{L}_{g_1^{e_1} \dots g_m^{e_m}, \leq t}$ are positive semidefinite for $e_1, \dots, e_m \in \{0, 1\}$ and all t .

3. If the moment and localization matrices are indexed by the monomials $\alpha \in \mathbb{N}^n$ of the monomial bases $(x^\alpha)_{\alpha \in \mathbb{N}^n}$ then the localizing matrix $M(g \cdot y)$ of a polynomial $g = \sum_\alpha c_\alpha x^\alpha$ is given by

$$M(g \cdot y)_{\alpha, \beta} = \sum_{\gamma \in \mathbb{N}^n} c_\gamma y_{\alpha+\beta+\gamma}.$$

4. Show that the infimum in the semidefinite program (7.22) is not attained.

7.7 Finite convergence and detecting optimality

By Theorem 7.6.7, convergence of the semidefinite relaxation scheme is guaranteed under mild preconditions. Moreover, in certain situations finite convergence can be guaranteed, that is, the optimal value will be attained at a finite relaxation order. From the viewpoint of the representation theorems, this translates to the question when the precondition of

strict positivity (such as in Putinar's Positivstellensatz) can be replaced by non-negativity. A related issue is the question how to decide if in a certain relaxation order the optimal value has already been reached. We study some aspects of these questions in the current section.

We first consider a specific situation where the feasible set involves also equality constraints,

$$\begin{aligned} p^* = \inf & \quad p(x) \\ \text{s.t. } & g_j(x) \geq 0, \quad 1 \leq j \leq m, \\ & h_j(x) = 0, \quad 1 \leq j \leq l, \\ & x \in \mathbb{R}^n, \end{aligned} \tag{7.28}$$

and assume h_1, \dots, h_l generate a zero-dimensional radical ideal. Set

$$K = \{x \in \mathbb{R}^n : g_1(x) \geq 0, \dots, g_m(x) \geq 0, h_1(x) = 0, \dots, h_l(x) = 0\}. \tag{7.29}$$

In this situation, a rather explicit representation for any non-negative polynomial on K is explicitly available. Clearly, the hierarchy of relaxations can be adapted to the problem (7.28).

Recall from Section 1.2 that over an arbitrary field \mathbb{F} the radical ideal \sqrt{I} is defined as $\sqrt{I} = \{q \in \mathbb{F}[x_1, \dots, x_n] : q^k \in I \text{ for some } k \geq 1\}$ and that an ideal is called radical if $\sqrt{I} = I$.

Theorem 7.7.1 (Parrilo). *Let $m \geq 0$, $l \geq 1$, and $g_1, \dots, g_m, h_1, \dots, h_l \in \mathbb{R}[x_1, \dots, x_n]$. If the ideal $I = \langle h_1, \dots, h_l \rangle$ is zero-dimensional and radical then any non-negative polynomial p on K can be written in the form*

$$p = s_0 + \sum_{j=1}^m s_j g_j + q$$

with $s_0, \dots, s_m \in \Sigma[x_1, \dots, x_n]$ and $q \in I$.

Example 7.7.2. Let $m = 0$, $h_1 = x^2 + y^2 - 1$ and $h_2 = y$. The polynomial $p = 3x^2y + 5y^2$ is non-negative on the variety $\mathcal{V}_{\mathbb{C}}(I) = \{(-1, 0), (1, 0)\}$. A certificate for this non-negativity in the light of Theorem 7.7.1 is

$$p = 5y^2 + (3x^2) \cdot y + 0 \cdot (x^2 + y^2 - 1),$$

since $5y^2 \in \Sigma[x, y]$.

For the proof of Theorem 7.7.1, we first deal with the case without inequality constraints which was illustrated in Example 7.7.2. Recall that $\mathcal{V}_{\mathbb{C}}(I)$ denotes the variety 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 conjugate pairs, thus giving a decomposition

$$\mathcal{V}_{\mathbb{C}}(I) = \mathcal{V}_{\mathbb{R}}(I) \cup U \cup \overline{U} \tag{7.30}$$

with $U \subset \mathbb{C}^n \setminus \mathbb{R}^n$.

Lemma 7.7.3. *Let $h_1, \dots, h_l \in \mathbb{R}[x]$ and let $I = \langle h_1, \dots, h_l \rangle$ be zero-dimensional and radical. Then any polynomial $p \in \mathbb{R}[x]$ which is non-negative on $\mathcal{V}_{\mathbb{R}}(I)$ can be written as $p = s + q$ with $s \in \Sigma[x]$ and $q \in I$.*

Proof. Using Lagrange interpolation, we can find (complex) polynomials p_a , $a \in \mathcal{V}_{\mathbb{C}}(I)$, satisfying $p_a(a) = 1$ and $p_a(b) = 0$ for $b \in \mathcal{V}_{\mathbb{C}}(I) \setminus \{a\}$. Whenever $a \in \mathcal{V}_{\mathbb{R}}(I)$ we can clearly choose p_a with real coefficients, say, by choosing the real part of the complex polynomial. With respect to the decomposition (7.30), for any $a \in V_{\mathbb{R}}(I) \cup U$ let γ_a be a square root of 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 $p^\circ = p - \sum_{a \in \mathcal{V}_{\mathbb{R}}(I) \cup U} q_a^2$ satisfies $p^\circ(a) = 0$ for all $a \in \mathcal{V}_{\mathbb{C}}(I)$, it is contained in the radical ideal I . This provides the desired representation. \square

Proof of Theorem 7.7.1. Let $p \in \mathbb{R}[x]$ be nonnegative on K . Via Lagrange interpolation, we construct polynomials $r_0, \dots, r_m \in \mathbb{R}[x]$ with the following properties. Whenever a is a non-real zero of I or whenever a is a real point in I with $p(a) \geq 0$, then we enforce $r_0(a) = f(a)$ and $r_j(a) = 0$, $1 \leq j \leq m$. Otherwise, we have $a \notin K$ and there exists some $j_a \in \{1, \dots, m\}$ with $g_{j_a}(a) < 0$. In that situation, we enforce $r_{j_a}(a) = \frac{p(a)}{g_{j_a}(a)}$ as well as $r_0(a) = r_j(a) = 0$ for all $j \neq j_a$. Note that each of the polynomials r_0, \dots, r_m is non-negative on $V_{\mathbb{R}}(I)$.

By Lemma 7.7.3, there exist $s_0, \dots, s_m \in \Sigma[x_1, \dots, x_n]$ and $q_0, \dots, q_m \in I$ with $r_j = s_j + q_j$, $1 \leq j \leq m$. The polynomial

$$p^\circ = p^\circ - r_0 - \sum_{j=1}^m r_j g_j \tag{7.31}$$

satisfies $p^\circ(a) = 0$ for all $a \in \mathcal{V}_{\mathbb{C}}(I)$ and is therefore contained in the radical ideal I . Solving (7.31) for f and substituting r_j by $s_j + q_j$ provides the representation we were aiming for. \square

Adapting the hierarchy to the equality constraints, we obtain for that variant of the semidefinite hierarchy:

Corollary 7.7.4. *If the set K is defined as in (7.29) and the ideal $I = \langle h_1, \dots, h_l \rangle$ is zero-dimensional and radical then there is some t with $p_t^{\text{sos}} = p_t^{\text{mom}} = p^*$.*

We mention the following extension of the result to the setting of (7.17) which requires that the ideal $\langle g_1, \dots, g_m \rangle$ is zero-dimensional but which does not require that it is radical.

Theorem 7.7.5 (Laurent). *If the ideal generated by g_1, \dots, g_m is zero-dimensional then there is some t with $p_t^{\text{mom}} = p^*$.*

Moment forms, ideals and detecting optimality. We now provide some connections between the viewpoint of moments and some ideal-theoretic concepts which were

considered earlier. These investigations also form the basis for criteria to detect whether a certain relaxation step in the relaxation scheme already provides the optimal value.

Let $L : \mathbb{R}[x] \rightarrow \mathbb{R}$ be a linear map and consider the associated bilinear form $\mathcal{L} : \mathbb{R}[x] \times \mathbb{R}[x] \rightarrow \mathbb{R}$, $\mathcal{L}(q, r) = L(qr)$ (note that in contrast to Theorem 7.4.2, L is not assumed to come from a measure).

For a positive semidefinite form \mathcal{L} , we consider its *kernel*

$$\begin{aligned} I &= \{q \in \mathbb{R}[x] : \mathcal{L}(q, r) = 0 \text{ for all } r \in \mathbb{R}[x]\} \\ &= \{q \in \mathbb{R}[x] : \mathcal{L}(q, x^\alpha) = 0 \text{ for all } \alpha \in \mathbb{N}^n\}. \end{aligned}$$

Observe, that I is an ideal. Clearly, $I + I \subset I$, and in order to show that for $q \in I$ and $s \in \mathbb{R}[x]$ we have $qs \in I$ observe that for any $r \in \mathbb{R}[x]$ we have $\mathcal{L}(qs, r) = \mathcal{L}(q, sr) = 0$, since $q \in I$.

Recall from Chapter 6.5 that the real radical of an ideal $I \subset \mathbb{R}[x]$ is defined by

$$\sqrt[{\mathbb{R}}]{I} = \left\{ p \in \mathbb{R}[x] : p^{2m} + \sum_j s_j^2 \in I \text{ for polynomials } s_j \in \mathbb{R}[x] \text{ and } m \geq 1 \right\}$$

and that an ideal is real radical if $I = \sqrt[{\mathbb{R}}]{I}$. If an ideal I of $\mathbb{R}[x]$ is real radical then it is also radical, and a real radical ideal I with $|V_{\mathbb{R}}(I)| < \infty$ gives $V_{\mathbb{C}}(I) = V_{\mathbb{R}}(I)$ (see Exercise 4).

Theorem 7.7.6. *Let \mathcal{L} be a positive semidefinite form and I be the kernel of \mathcal{L} . Then*

1. *I is a real radical ideal in $\mathbb{R}[x]$.*
2. *If \mathcal{L} has finite rank then $|V_{\mathbb{C}}(I)| = \text{rank } \mathcal{L}$.*

Proof. (a) Let $p_1, \dots, p_k \in \mathbb{R}[x]$ with $\sum_{i=1}^k p_i^2 \in I$. Then

$$0 = \mathcal{L}(1, \sum_{i=1}^k p_i^2) = \sum_{i=1}^k \mathcal{L}(p_i, p_i).$$

Since \mathcal{L} is positive semidefinite, we have $\mathcal{L}(p_i, p_i) = 0$ and therefore $p_i \in I$ for all i . Hence, by Exercise 3, I is real radical.

(b) By the proof of Theorem 2.4.1, for a radical ideal I the cardinality $|V_{\mathbb{C}}(I)|$ coincides with the dimension of the vector space $\mathbb{R}[x]/I$. Therefore it suffices to show $\text{rank } \mathcal{L} = \dim \mathbb{R}[x]/I$.

By definition of the kernel of \mathcal{L} , polynomials in I do not contribute to the rank and thus the quadratic form \mathcal{L} can be viewed as a quadratic form $\mathcal{L}' : \mathbb{R}[x]/I \times \mathbb{R}[x]/I \rightarrow \mathbb{R}$. Moreover, we can think of indexing the variables of \mathcal{L}' by a set of standard monomials \mathcal{B} with respect to the ideal I . If $\text{rank } \mathcal{L}$ is finite then there exists a finite symmetric representation matrix M for \mathcal{L}' whose rows and columns are indexed by \mathcal{B} . The matrix M has full rank, since otherwise the zero vector can be represented as a non-trivial linear combination of the columns. This would imply a non-trivial combination of standard monomials that is contained in I , a contradiction. Hence, $\text{rank } \mathcal{L} = |\mathcal{B}| = \dim \mathbb{R}[x]/I$. \square

A probability measure μ on K is called r -atomic if can be written as $\mu = \sum_{i=1}^r \lambda_i \delta_{x^{(i)}}$ with $\lambda_1, \dots, \lambda_r \geq 0$, where $x^{(1)}, \dots, x^{(r)} \in \mathbb{R}^n$ and $\delta_{x^{(i)}}$ is the Dirac measure at $x^{(i)}$. And recall that a symmetric bilinear form f on a vector space V (possibly infinite-dimensional) is of finite rank t if t is the smallest number such that there exists an ordered basis \mathcal{B} of V of cardinality t such that relative to \mathcal{B} , f can be written as $\sum_{i=1}^t \pm x_i^2$.

In the following, let p_1, \dots, p_r be interpolation polynomials for $V_{\mathbb{C}}(I) = \{v^{(1)}, \dots, v^{(r)}\}$, as defined by $p_i(v^{(i)}) = 1$ and $p_i(v^{(j)}) = 0$ for $j \neq i$.

Corollary 7.7.7. *Let \mathcal{L} be a positive semidefinite form of rank r , I be the kernel of \mathcal{L} and $V_{\mathbb{R}}(I) = V_{\mathbb{C}}(I) = \{v^{(1)}, \dots, v^{(r)}\}$. Then*

$$\mathcal{L}(s, t) = \sum_{k=1}^r \mathcal{L}(p_k, p_k) s(v^{(k)}) t(v^{(k)}),$$

and

$$\mu = \sum_{k=1}^r \mathcal{L}(p_k, p_k) \delta_{v^{(k)}} \quad (7.32)$$

is the unique measure representing \mathcal{L} , where $\delta_{v^{(k)}}$ denotes the Dirac measure at $v^{(k)}$.

Proof. The interpolation polynomials p_1, \dots, p_r form a basis of $\mathbb{R}[x]/I$, and it suffices to prove the statement on a basis of $\mathbb{R}[x]/I$. Defining $\mathcal{L}'(s, t) = \sum_{k=1}^r \mathcal{L}(s, t) s(v^{(k)}) t(v^{(k)})$, we observe $\mathcal{L}'(p_i, p_i) = \sum_{k=1}^r \mathcal{L}(p_k, p_k) p_i(v^{(k)}) p_i(v^{(k)}) = \mathcal{L}(p_i, p_i)$ and $\mathcal{L}'(p_i, p_j) = 0$ for $j \neq i$. This shows $\mathcal{L} = \mathcal{L}'$ and also that the underlying linear map L is the integration with respect to the measure μ defined in (7.32).

Now assume that there is another measure μ' representing the form \mathcal{L} . For any polynomial $p \in I$, the positive semidefiniteness of \mathcal{L} implies that μ be supported on the zeroes of p . Hence, μ is supported on $V_{\mathbb{R}}(I)$, and therefore μ' is a sum of Dirac measures of the form $\mu' = \sum_{k=1}^r \lambda_k \delta_{v^{(k)}}$. The rank condition on \mathcal{L} gives $\lambda_k \neq 0$ for all k . Finally, evaluating \mathcal{L} at the points $v^{(k)}$ shows $\mathcal{L}(p_k, p_k) = \lambda_k$ and thus $\mu' = \mu$. \square

Theorem 7.7.8. *For a form \mathcal{L} , the following statements are equivalent:*

1. \mathcal{L} is positive semidefinite and has finite rank r .
2. There exists a unique probability measure μ representing \mathcal{L} , and μ is r -atomic.

Proof. If \mathcal{L} is positive semidefinite and has finite rank r , then Corollary 7.7.7 gives the existence of a unique measure μ representing \mathcal{L} , and μ is r -atomic.

Conversely, let $\mu = \sum_{k=1}^r \lambda_k \delta_{v^{(k)}}$ with $\lambda_k > 0$ be an r -atomic measure representing a form \mathcal{L} . Then, for any polynomial $q \in \mathbb{R}[x]$, we have $\mathcal{L}(q, q) = \int q^2 d\mu \geq 0$. Further, set $\mathcal{L}(p_k, p_k) = L(p_k)^2 = \lambda_k > 0$. Then the form $\mathcal{L}(s, t)$ from Corollary 7.7.7 is the form represented by μ . Hence, $\text{rank } \mathcal{L} = r$. \square

For a constrained optimization problem

$$p^* = \inf\{p(x) : g_j(x) \geq 0\}$$

with $p, g_1, \dots, g_m \in \mathbb{R}[x]$, $K = \{x \in \mathbb{R}^n : g_j(x) \geq 0\}$ with an Archimedean quadratic module $\text{QM}(g_1, \dots, g_m)$, these characterizations can be used to provide a sufficient criterion whether at some relaxation step t the optimal value p_t^{mom} of a moment relaxation already coincides with p^* .

Theorem 7.7.9 (Henrion, Lasserre). *Let $L : \mathbb{R}[x]_{\leq 2t} \rightarrow \mathbb{R}$ be an optimal solution to the truncated moment relaxation (7.21) and $d = \max\{\lceil \deg(g_j)/2 \rceil : 1 \leq j \leq m\}$. If*

$$\text{rank } \mathcal{L}_{\leq t} = \text{rank } \mathcal{L}_{\leq t-d}$$

then $p_t^{\text{mom}} = p^$.*

We omit the proof. As main technical tool to treat truncated moment matrices, it uses the Flat Extension Theorem, which we state here without proof as well. A symmetric bilinear form $\mathcal{L} : \mathbb{R}[x]_{\leq d} \rightarrow \mathbb{R}$ of the form $\mathcal{L} = \mathcal{L}' \oplus \mathcal{L}''$ is called a *flat extension* of \mathcal{L}' if $\text{rank } \mathcal{L} = \text{rank } \mathcal{L}'$.

Theorem 7.7.10 (Flat Extension Theorem of Curto and Fialkow). *Let $L : \mathbb{R}[x]_{\leq 2t} \rightarrow \mathbb{R}$, $\mathcal{L}_{\leq t}$ positive semidefinite and $\mathcal{L}_{\leq t}$ be a flat extension of \mathcal{L}_{t-1} . Then L can be extended to a mapping $L : \mathbb{R}[x] \rightarrow \mathbb{R}$ which is the integration of a $(\text{rank } \mathcal{L}_{\leq t})$ -atomic representing measure.*

Exercises

1. Construct a representation to certify that the polynomial $x_1^2 + x_2^2 - x_1x_2$ is non-negative over the three-point set $\{(0,0), (1,0), (0,1)\}$.
2. Deduce from Lemma 7.7.3 a statement to certify the non-existence of real zeroes of a zero-dimensional radical ideal.
3. An ideal $I \subset \mathbb{R}[x]$ is real radical if and only if for any $p_1, \dots, p_k \in \mathbb{R}[x]$ the property $\sum_{i=1}^k p_i^2 \in I$ implies $p_1, \dots, p_k \in I$.
4. If $I \subset \mathbb{R}[x]$ is real radical then it is also radical, and a real radical I with $|V_{\mathbb{R}}(I)| < \infty$ implies $V_{\mathbb{C}}(I) = V_{\mathbb{R}}(I)$
5. 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 $f = s + q$.

7.8 Notes

The use of sums of squares in optimization has mainly been initiated by N.Z. Shor [128], Lasserre [75] and Parrilo [95, 96]. See also the comprehensive treatments of Laurent [79] and Lasserre [76]. The sparsity result in Exercise 3 of Section 7.2 is due to Reznick [113].

The relaxation scheme for constrained global optimization has been introduced by Lasserre [75]. Theorem 7.6.9 is due to Putinar, where his proof of Theorem 7.6.9 was using methods from functional analysis and where he used this theorem to deduce then Theorem 6.7.8 from it. The link for the converse direction, from Theorem 6.7.8 to Theorem 7.6.9, as utilized in our treatment, goes back to Krivine [73], provided with a modern treatment in the book of Prestel and Delzell [110].

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 [70].

Since around the year 2000, there have been intense developments to approximate the constrained polynomial optimization problems in a hierarchical way using semidefinite programming and real algebraic geometry. The roots of this development go back to N.Z. Shor [128], and the main developments of the semidefinite program hierarchies have been initiated by J. Lasserre [75] and P. Parrilo [95]. More information on the moment-based approach for optimization can be found in Lasserre's book [76] or in Laurent's extensive survey [79].

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 an introduction to semidefinite programming see the book of De Klerk [29] or the survey article by Vandenberghe and Boyd [140]. The polynomial time decidability of SDP for fixed dimension or number of constraints is due to Porkolab, Khachiyan [106].

Haviland's proof of his Theorem 7.5.4 is contained in [50] and in [88]. Indeed, the statement is implicitly contained in Riesz' work in 1923 [115], see also the historical description in [54]. The univariate special cases $K = [0, \infty)$, $K = \mathbb{R}$, $K = [0, 1]$ (cf. Theorem 7.5.2) were proven earlier by Stieltjes, Hamburger, and Hausdorff. Among the four inclusions, $(\mathcal{M}_n^+)^* \subset \Sigma[x]$ is the most difficult one. Berg, Christensen, and Jensen [7] and independently Schmüdgen ([120]) have shown that the infinite-dimensional cone $\Sigma[x]$ is closed (see also [8, Ch. 6, Thm. 3.2]), which then implies $(\mathcal{M}_n^+)^* \subset (\Sigma[x])^{**} = \Sigma[x]$.

A proof of Theorem 7.6.9 from Putinar's Positivstellensatz 6.7.8 can be found in Schweighofer [125].

Theorem 7.7.1 was proven by Parrilo in [97], and the finite convergence result 7.7.5 of Laurent appears in [78]. The rank criterion 7.7.9 to detect optimality is due to Henrion and Lasserre [56]. They build on the flat extension results of Curto and Fialkow who used functional-analytic methods [27, 28]. The transfer to a rather algebraic derivation has been achieved by Laurent [77, 78]. In case the criterion in Theorem 7.7.9 is satisfied, an

optimal point can then be characterized similar to Theorem 7.4.4 and then be extracted using the eigenvalue methods from Section 2.5, see [76].

Chapter 8

Toric Varieties

Outline:

1. Toric Ideals and Affine Toric Varieties.
2. Projective toric varieties
3. Kushnirenko's Theorem
4. Toric and Gröbner degenerations. *Maybe not*
5. Real Toric Varieties *Maybe not* Birch's Theorem and the moment map, Gluing construction of real toric varieties, Viro construction, Sturmfels' reality theorem.
6. Bernstein's Theorem and Polyhedral Homotopies *Maybe two sections?* Give two proofs of Bernstein's Theorem, Describe polyhedral homotopies

Algebraic varieties often arise from parametrizations, that is as the closure of the image of a polynomial map. In general, it is a challenging problem to find the ideal of such a parameterized variety. This is the implicitization problem of Chapter 2. When the polynomials have degree 1, the variety is a linear or affine space, and it is given by polynomials of degree 1. Perhaps the next simplest parametrization, which is common in applications, is when the polynomials in the parametrization are monomials. Such a parameterized variety is a toric variety. Toric varieties are well-understood, as their structure is controlled by objects from geometric combinatorics. We develop some elementary aspects of toric varieties. A goal will be the Bernstein-Kushnirenko theorem, which gives the number of solutions to a general system of sparse polynomials, and to present a numerical homotopy algorithm to find their solutions.

8.1 Toric Ideals and Affine Toric Varieties

Consider the map $\mathbb{K}^m \times \mathbb{K}^n \rightarrow \mathbb{K}^{m \times n} = \text{Mat}_{m \times n}(\mathbb{K})$, the space of $m \times n$ matrices, which sends a pair of vectors $x \in \mathbb{K}^m$ and $y \in \mathbb{K}^n$ to their outer product $xy^T \in \mathbb{K}^{m \times n}$. If $z_{i,j}$ for $i = 1, \dots, m$ and $j = 1, \dots, n$ are the coordinates for $\mathbb{K}^{m \times n}$, then the polynomials

$$\underline{z_{i,j} z_{k,l}} - z_{i,l} z_{k,j} = \det \begin{pmatrix} z_{i,j} & z_{i,l} \\ z_{k,l} & z_{k,j} \end{pmatrix} \quad 1 \leq i < k \leq m \text{ and } 1 \leq j < l \leq n, \quad (8.1)$$

vanish on the image. These determinants cut out the image of this map, which is the variety of $m \times n$ matrices of rank one. This *Segre variety* is a prototypical example of a toric variety.

Write \mathbb{K}^\times for the multiplicative group of nonzero elements of the field \mathbb{K} and $(\mathbb{K}^\times)^n$ for the group of invertible diagonal $n \times n$ matrices over \mathbb{K} , equivalently, of ordered n -tuples of nonzero elements of \mathbb{K} . The free abelian group \mathbb{Z}^n of rank n is associated to $(\mathbb{K}^\times)^n$ in two distinct ways. It is isomorphic to the lattice of *cocharacters* (group homomorphisms) from \mathbb{K}^\times to $(\mathbb{K}^\times)^n$. An integer vector $w = (w_1, \dots, w_n) \in \mathbb{Z}^n$ gives the map which sends $t \in \mathbb{K}^\times$ to the diagonal matrix $t^w := \text{diag}(t^{w_1}, \dots, t^{w_n}) \in (\mathbb{K}^\times)^n$. The group of characters, equivalently of *Laurent monomials*, is also isomorphic to \mathbb{Z}^n . Here, an integer vector $\alpha = (\alpha_1, \dots, \alpha_n)^T \in \mathbb{Z}^n$ gives the Laurent monomial $x^\alpha := x_1^{\alpha_1} \cdots x_n^{\alpha_n}$, which is also a group homomorphism $(\mathbb{K}^\times)^n \ni x \mapsto x^\alpha \in \mathbb{K}^\times$, where $x = \text{diag}(x_1, \dots, x_n)$.

A linear combination of Laurent monomials is a Laurent polynomial. The coordinate ring of $(\mathbb{K}^\times)^n$ is the ring $\mathbb{K}[x_1, x_1^{-1}, \dots, x_n, x_n^{-1}]$ of *Laurent polynomials*. This is also the group algebra $\mathbb{K}[\mathbb{Z}^n]$ and we write $\mathbb{K}[x^\pm]$ for this Laurent ring.

Let $\mathcal{A} \subset \mathbb{Z}^n$ be a finite subset of monomials. It is convenient to represent \mathcal{A} as the set of column vectors of an integer matrix with n rows. We will also write \mathcal{A} for this matrix. We will use this set \mathcal{A} to index coordinates, variables, etc. For example $(\mathbb{K}^\times)^\mathcal{A}$ is the set of functions from \mathcal{A} to \mathbb{K}^\times . It is the algebraic torus $(\mathbb{K}^\times)^{|\mathcal{A}|}$ whose coordinates are indexed by the elements of \mathcal{A} . Likewise $\mathbb{K}^\mathcal{A}$ is the vector space of functions from \mathcal{A} to \mathbb{K} . These all have coordinates $(z_a \mid a \in \mathcal{A})$. If \mathcal{A} is represented by the matrix $(\begin{smallmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{smallmatrix})$, then $\mathbb{K}^\mathcal{A} \simeq \mathbb{K}^4$ has coordinates $z_{(0,0)}, z_{(0,1)}, z_{(1,0)}, z_{(1,1)}$.

A finite subset $\mathcal{A} \subset \mathbb{Z}^n$ may be used to define a map $\varphi_\mathcal{A}: (\mathbb{K}^\times)^n \rightarrow \mathbb{K}^\mathcal{A}$, where

$$\varphi_\mathcal{A}(x) := (x^a \mid a \in \mathcal{A}). \quad (8.2)$$

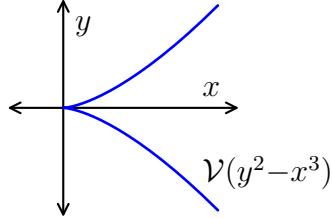
When $\mathcal{A} = (\begin{smallmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{smallmatrix})$, for $(x, y) \in (\mathbb{K}^\times)^2$ we have $\varphi_\mathcal{A}(x, y) = (1, x, y, xy) \in \mathbb{K}^\mathcal{A}$. Notice that the map $\varphi_\mathcal{A}$ (8.2) is a group homomorphism $\varphi_\mathcal{A}: (\mathbb{K}^\times)^n \rightarrow (\mathbb{K}^\times)^\mathcal{A}$ followed by the inclusion $(\mathbb{K}^\times)^\mathcal{A} \subset \mathbb{K}^\mathcal{A}$. The Zariski closure of the image $\varphi_\mathcal{A}(\mathbb{K}^\times)^n$ in $\mathbb{K}^\mathcal{A}$ is the *affine toric variety* $X_\mathcal{A}$.

If the subgroup $\mathbb{Z}\mathcal{A} \subset \mathbb{Z}^n$ generated by \mathcal{A} is proper, then the homomorphism $\varphi_\mathcal{A}$ has a nontrivial kernel $T := \ker \varphi_\mathcal{A}$. In this case, $\varphi_\mathcal{A}$ (8.2) induces an injective map $(\mathbb{K}^\times)^n / \ker \varphi_\mathcal{A} \rightarrow \mathbb{K}^\mathcal{A}$. Then the quotient $\mathbb{Z}\mathcal{A}$ is the group of characters of $(\mathbb{K}^\times)^n / \ker \varphi_\mathcal{A}$ and so $\dim X_\mathcal{A} = \dim(\mathbb{K}^\times)^n / \ker \varphi_\mathcal{A} = \text{rank } \mathbb{Z}\mathcal{A}$. **Maybe we need to discuss the rank of a free abelian group?**

We deduce two characterizations of affine toric varieties from this definition. First, by (8.2), they are varieties that are parameterized by monomials. Also, affine toric varieties are varieties that arise as the closure in the affine space \mathbb{K}^m of a subtorus of $(\mathbb{K}^\times)^m$. Here, $m = |\mathcal{A}|$ and the subtorus is the image $\varphi_\mathcal{A}(\mathbb{K}^\times)^n$. Observe that the torus $(\mathbb{K}^\times)^n$ acts on $X_\mathcal{A}$ with a dense orbit and this action extends to the ambient affine space $\mathbb{K}^\mathcal{A}$.

Example 8.1.1. Suppose that $n = 1$ and $\mathcal{A} = \{2, 3\} \subset \mathbb{Z}$. For $s \in \mathbb{K}^\times$, $\varphi_\mathcal{A}(s) = (s^2, s^3) \in \mathbb{K}^2$. The closure of $\varphi_\mathcal{A}(\mathbb{K}^\times)^n$ is the cuspidal cubic, $\mathcal{V}(y^2 - x^3)$, where \mathbb{K}^2 has coordinates

(x, y) .



Since $\mathbb{Z}\mathcal{A} = \mathbb{Z}$, $\varphi_{\mathcal{A}}$ is injective. Indeed, if $(x, y) = \varphi_{\mathcal{A}}(s)$, then $s = y/x$.

Suppose that $m, n \geq 1$ are integers. Let e_1, \dots, e_m and f_1, \dots, f_n be the standard unit basis vectors for \mathbb{Z}^m and \mathbb{Z}^n , respectively, and set

$$\mathcal{A} := \{e_i + f_j \mid i = 1, \dots, m \text{ and } j = 1, \dots, n\} \subset \mathbb{Z}^m \times \mathbb{Z}^n,$$

which has mn elements. The map $\varphi_{\mathcal{A}}: \mathbb{K}^m \times \mathbb{K}^n \rightarrow \mathbb{K}^{mn}$ is

$$(x_1, \dots, x_m, y_1, \dots, y_n) \mapsto (x_i y_j \mid i = 1, \dots, m \text{ and } j = 1, \dots, n).$$

Identifying \mathbb{K}^{mn} with $\text{Mat}_{m \times n}(\mathbb{K})$, this map is $(x, y) \mapsto xy^T$, with which we started this section, and thus $X_{\mathcal{A}}$ is the Segre variety of $m \times n$ matrices of rank at most 1. [Relate this to Theorem 1.5.4 in Section 1.5.](#)

Finally, suppose that $d \geq 1$ is an integer. The d th *Veronese map* $\varphi_{\mathcal{A}}: \mathbb{K}^n \rightarrow \mathbb{K}^{\binom{d+n}{n}}$, is when \mathcal{A} is the set of all exponent vectors in \mathbb{N}^n of degree at most d . When $n = 1$ and $d = 3$, we have $\mathcal{A} = \{0, 1, 2, 3\} \subset \mathbb{Z}$ and $\varphi_{\mathcal{A}}(x) = (1, x, x^2, x^3)$. Ignoring the first coordinate which is a constant, $X_{\mathcal{A}}$ is the moment (rational normal) curve in \mathbb{K}^3 . \diamond

The ideal $I_{\mathcal{A}}$ of $X_{\mathcal{A}}$ is a *toric ideal*. It lies in the coordinate ring $\mathbb{K}[z_{\alpha} \mid \alpha \in \mathcal{A}]$ of the affine space $\mathbb{K}^{\mathcal{A}}$. To understand $I_{\mathcal{A}}$, consider the pull back map of Section 1.3 induced by $\varphi_{\mathcal{A}}$ on coordinate rings,

$$\begin{aligned} \varphi_{\mathcal{A}}^*: \mathbb{K}[z_{\alpha} \mid \alpha \in \mathcal{A}] &\longrightarrow \mathbb{K}[x^{\pm}] \\ z_{\alpha} &\longmapsto x^{\alpha}. \end{aligned}$$

The toric ideal $I_{\mathcal{A}}$ is the kernel of $\varphi_{\mathcal{A}}^*$. The exponent of a monomial z^u in $\mathbb{K}[z_{\alpha} \mid \alpha \in \mathcal{A}]$ is the vector $u = (u_{\alpha} \mid \alpha \in \mathcal{A}) \in \mathbb{N}^{\mathcal{A}}$, and the image of z^u under $\varphi_{\mathcal{A}}^*$ is

$$\varphi_{\mathcal{A}}^*(z^u) = \prod_{\alpha \in \mathcal{A}} (x^{\alpha})^{u_{\alpha}} = x^{\sum \alpha u_{\alpha}}.$$

Let us write the sum $\sum_{\alpha \in \mathcal{A}} \alpha u_{\alpha}$ in this exponent as $\mathcal{A}u$. When \mathcal{A} is represented by an integer matrix \mathcal{A} , this is the usual matrix-vector product. The kernel $I_{\mathcal{A}}$ of $\varphi_{\mathcal{A}}^*$ contains the following set of binomials

$$\{z^u - z^v \mid \mathcal{A}u = \mathcal{A}v\}. \tag{8.3}$$

Indeed, $\varphi_{\mathcal{A}}^*(z^u) = \varphi_{\mathcal{A}}^*(z^v)$ if and only if $\mathcal{A}u = \mathcal{A}v$.

Suppose that \mathcal{A} is represented by the matrix $(\begin{smallmatrix} 0 & 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 & 0 \end{smallmatrix})$. If $u = (0, 1, 1, 1, 0)^T$ and $v = (1, 0, 1, 0, 1)^T$, then $\mathcal{A}u = \mathcal{A}v$, which gives the binomial in $I_{\mathcal{A}}$,

$$z_{(1)} z_{(2)} z_{(1)} - z_{(4)} z_{(2)} z_{(0)}.$$

Theorem 8.1.2. *The toric ideal $I_{\mathcal{A}}$ is a prime ideal. As a vector space, it is spanned by the binomials (8.3).*

Proof. The image of $\varphi_{\mathcal{A}}^*$ is the subalgebra of $\mathbb{K}[x^{\pm}]$ generated by the monomials $\{x^{\alpha} \mid \alpha \in \mathcal{A}\}$. Since $\mathbb{K}[x^{\pm}]$ is an integral domain, this subalgebra is also a domain and so the kernel $I_{\mathcal{A}}$ is a prime ideal. Equivalently, note that $X_{\mathcal{A}}$ is irreducible as $X_{\mathcal{A}}$ is the closure of the image of the irreducible variety $(\mathbb{K}^*)^n$ under the map $\varphi_{\mathcal{A}}$. By Theorem 3.2.4, its defining ideal $I_{\mathcal{A}}$ is prime.

For the second statement, let \prec be any term order on $\mathbb{K}[z_{\alpha} \mid \alpha \in \mathcal{A}]$. Let $f \in I_{\mathcal{A}}$. We may write f as

$$f = c_u z^u + \sum_{v \prec u} c_v z^v \quad c_u \neq 0,$$

so that $\text{in}_{\prec}(f) = c_u z^u$ is the initial term of f . 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 $\mathcal{A}v = \mathcal{A}u$, for otherwise the term $c_u x^{\mathcal{A}u}$ is not canceled in $\varphi_{\mathcal{A}}^*(f)$ and $\varphi_{\mathcal{A}}^*(f) \neq 0$. Set $\bar{f} := f - c_u(z^u - z^v)$. Then $\varphi_{\mathcal{A}}^*(\bar{f}) = 0$ and $\text{in}_{\prec}(\bar{f}) \prec \text{in}_{\prec}(f)$.

If the leading term of f were \prec -minimal in the initial ideal $\text{in}_{\prec}(I_{\mathcal{A}})$, then \bar{f} would be zero, and so f is a scalar multiple of a binomial of the form (8.3). Suppose now that $\text{in}_{\prec} f$ is not minimal in $\text{in}_{\prec}(I_{\mathcal{A}})$ and that every polynomial in $I_{\mathcal{A}}$ all of whose terms are \prec -less than the initial term of f is a linear combination of binomials of the form (8.3). Then \bar{f} is a linear combination of binomials of the form (8.3), which implies that f is as well, by our induction hypothesis. This completes the proof. \square

The coordinate ring of the affine toric variety $X_{\mathcal{A}}$ is the quotient $\mathbb{K}[z_{\alpha} \mid \alpha \in \mathcal{A}]/I_{\mathcal{A}}$. As $X_{\mathcal{A}}$ is parameterized by monomials in $(\mathbb{K}^\times)^n$ through the map $\varphi_{\mathcal{A}}$, Corollary 1.3.13 identifies its coordinate ring as the image of $\varphi_{\mathcal{A}}^*$ in the coordinate ring of $(\mathbb{K}^\times)^n$. As observed in the proof of Theorem 8.1.2, this is the subring of the ring of Laurent polynomials $\mathbb{K}[x^{\pm}]$ generated by the monomials $\{x^{\alpha} \mid \alpha \in \mathcal{A}\}$.

Corollary 8.1.3. *The coordinate ring of the affine toric variety $X_{\mathcal{A}}$ is $\mathbb{K}[x^{\alpha} \mid \alpha \in \mathcal{A}]$.*

Theorem 8.1.2 gives a spanning set for $I_{\mathcal{A}}$. We seek an economical generating set. Suppose that $\mathcal{A}u = \mathcal{A}v$ with $u, v \in \mathbb{N}^{\mathcal{A}}$. We define vectors $r, w^{\pm} \in \mathbb{N}^{\mathcal{A}}$. For $\alpha \in \mathcal{A}$, set

$$\begin{aligned} r_{\alpha} &:= \min(u_{\alpha}, v_{\alpha}), \\ w_{\alpha}^+ &:= \max(u_{\alpha} - v_{\alpha}, 0), \quad \text{and} \\ w_{\alpha}^- &:= \max(v_{\alpha} - u_{\alpha}, 0). \end{aligned}$$

Then we have $u - v = w^+ - w^-$ with $u = r + w^+$ and $v = r + w^-$, and so

$$z^r(z^{w^+} - z^{w^-}) = z^u - z^v \in I_{\mathcal{A}}, \quad (8.4)$$

with $z^r = \text{lcm}\{z^u, z^v\}$. Note also that $\mathcal{A}w^+ = \mathcal{A}w^-$ as $0 = \mathcal{A}(u - v) = \mathcal{A}(w^+ - w^-)$, so that $z^{w^+} - z^{w^-} \in I_{\mathcal{A}}$. When $\mathcal{A} = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 & 0 \end{pmatrix}$, with $u = (0, 1, 1, 1, 0)^T$ and $v = (1, 0, 1, 0, 1)^T$, we have $r = (0, 0, 1, 0, 0)^T$, $w^+ = (1, 0, 0, 0, 1)^T$ and $w^- = (0, 1, 0, 1, 0)^T$, and

$$z_{(1)} z_{(2)} z_{(3)} - z_{(4)} z_{(2)} z_{(0)} = z_{(2)} (z_{(1)} z_{(3)} - z_{(4)} z_{(0)}) \in I_{\mathcal{A}}.$$

For $u \in \mathbb{Z}^{\mathcal{A}}$, let u^+ be the coordinatewise maximum of u and the 0-vector, and let u^- be the coordinatewise maximum of $-u$ and the 0-vector. We see that $I_{\mathcal{A}}$ is generated by binomials coming from the integer kernel of the matrix \mathcal{A} .

Corollary 8.1.4. $I_{\mathcal{A}} = \langle z^{u^+} - z^{u^-} \mid \mathcal{A}u = 0 \rangle$.

Theorem 8.1.5. *Any reduced Gröbner basis of $I_{\mathcal{A}}$ consists of binomials.*

The point is that Buchberger's algorithm 2.3.6 is binomial-friendly, in the same sense as it was homogeneous-friendly in the proof of Theorem 3.5.2. That is, if f and g are binomials, then their S -polynomial is again a binomial, and the reduction of one binomial by another is again a binomial.

It is an important problem to compute or to find relatively small Gröbner bases for toric ideals. By Corollary 8.1.4, these are given by special subsets of the integer kernel $\{u \in \mathbb{Z}^{\mathcal{A}} \mid \mathcal{A}u = 0\}$ of \mathcal{A} . For example, a reduced Gröbner basis for the ideal of $m \times n$ matrices of rank 1 is given by the binomials in (8.1). Here the term order is the degree reverse lexicographic order with the variables ordered by $z_{i,l} \prec z_{k,j}$ if $i < k$ or $i = k$ and $l > i$ (the leading term is underlined in (8.1)).

By Corollary 8.1.3, the coordinate ring of the affine toric variety $X_{\mathcal{A}}$ is the algebra $\mathbb{K}[x^\alpha \mid \alpha \in \mathcal{A}]$. This subalgebra of the ring of Laurent polynomials is spanned by monomials. Its set of monomials is identified with \mathbb{NA} , which is the subset of \mathbb{Z}^n consisting of nonnegative integer combinations of elements of \mathcal{A} under the map

$$\sum_{\alpha \in \mathcal{A}} \alpha u_\alpha \longmapsto \prod_{\alpha \in \mathcal{A}} (x^\alpha)^{u_\alpha}.$$

This subset \mathbb{NA} of the group \mathbb{Z}^n is closed under addition and contains the identity.

Let us generalize this. A *monoid* is a set with an associative binary operation with an identity, but it does not necessarily have inverses. For example, a field \mathbb{K} under multiplication forms a monoid with zero an absorbing element, $a \cdot 0 = 0$ for any $a \in \mathbb{K}$. Any group is a monoid. Given a finitely generated submonoid S of \mathbb{Z}^n , write $\mathbb{K}[S] \subset \mathbb{K}[x^\pm]$ for the *monoid algebra* of S . This is the set of all \mathbb{K} -linear combinations of elements of S , where the multiplication is distributive and induced by the monoid operation on S . Thus Corollary 8.1.3 asserts that the coordinate ring of the toric variety $X_{\mathcal{A}}$ is the monoid algebra $\mathbb{K}[\mathbb{NA}]$.

Given a finitely generated submonoid $S \subset \mathbb{Z}^n$, if we choose a finite generating set \mathcal{A} of S , so that $S = \mathbb{N}\mathcal{A}$, then $\mathbb{K}[S]$ is isomorphic to the coordinate ring of the affine toric variety $X_{\mathcal{A}} \subset \mathbb{K}^{\mathcal{A}}$. By a result in Section 1.1, the points of $X_{\mathcal{A}}$ are identified with the maximal ideals of $\mathbb{K}[S]$, when \mathbb{K} is algebraically closed. There is a second perspective on these points, via monoid homomorphisms. By Exercise 8, maximal ideals in $\mathbb{K}[S]$ correspond to monoid homomorphisms $\text{Hom}_{\text{mon}}(S, \mathbb{K})$, from S to \mathbb{K} (additive on S and multiplicative on \mathbb{K}). More generally, given any commutative monoid M , we may define $X_{\mathcal{A}}(M) := \text{Hom}_{\text{mon}}(S, M)$.

The map $\varphi_{\mathcal{A}}$ (8.2) commutes with the inclusions $(\mathbb{R}^{\times})^n \subset (\mathbb{C}^{\times})^n$ and $\mathbb{R}^{\mathcal{A}} \subset \mathbb{C}^{\mathcal{A}}$. We define the real affine toric variety to be the Zariski closure of the image of the map $\varphi_{\mathcal{A}}: (\mathbb{R}^{\times})^n \rightarrow \mathbb{R}^{\mathcal{A}}$. This may differ from the closure in the usual topology on $\mathbb{R}^{\mathcal{A}}$, as we see in Exercise 10. Similarly, if we write $\mathbb{R}_>$ for the positive real numbers and \mathbb{R}_{\geq} for the nonnegative real numbers, $\varphi_{\mathcal{A}}$ commutes with the inclusions $\mathbb{R}_>^n \subset (\mathbb{C}^{\times})^n$ and $\mathbb{R}_{\geq}^{\mathcal{A}} \subset \mathbb{C}^{\mathcal{A}}$, where $\mathbb{R}_{\geq}^{\mathcal{A}}$ is the positive orthant in $\mathbb{R}^{\mathcal{A}}$. Define the *nonnegative affine toric variety* $X_{\mathcal{A}, \geq}$ to be the closure of the image $\varphi_{\mathcal{A}}(\mathbb{R}_>^n)$ in the positive orthant $\mathbb{R}_{\geq}^{\mathcal{A}}$. We have the following maps

$$X_{\mathcal{A}, \geq} \hookrightarrow X_{\mathcal{A}}(\mathbb{R}) \hookrightarrow X_{\mathcal{A}}(\mathbb{C}) \twoheadrightarrow X_{\mathcal{A}, \geq}. \quad (8.5)$$

These are induced by the maps

$$\mathbb{R}_{\geq} \hookrightarrow \mathbb{R} \hookrightarrow \mathbb{C} \twoheadrightarrow \mathbb{R}_{\geq}, \quad (8.6)$$

with the last map $z \mapsto |z|$. The maps (8.5) also come from the identification of affine toric varieties with monoid homomorphisms and the sequence of maps of monoids (8.6). You are asked to investigate this in Exercise 9. As the composition of these monoid maps is the identity, the composition (8.5) is also the identity.

Exercises

1. Use the embedding $x \mapsto (x, x^{-1})$ of the torus \mathbb{K}^{\times} into \mathbb{K}^2 , or any other method, to identify the coordinate ring of the torus \mathbb{K}^{\times} with $\mathbb{K}[x, x^{-1}]$. Deduce that the coordinate ring of $(\mathbb{K}^{\times})^n$ may be identified with $\mathbb{K}[x_1, x_1^{-1}, \dots, x_n, x_n^{-1}]$. Show that this is the group ring $\mathbb{K}[\mathbb{Z}^n]$ of the group \mathbb{Z}^n of characters of the torus $(\mathbb{K}^{\times})^n$. Harder: What algebraic structure of $\mathbb{K}[\mathbb{Z}^n]$ induces the group structure on $(\mathbb{K}^{\times})^n$?
2. Show that the monomial map $\varphi_{\mathcal{A}}: (\mathbb{K}^{\times})^n \rightarrow (\mathbb{K}^{\times})^{\mathcal{A}}$ (8.2) is injective if and only if \mathcal{A} generates \mathbb{Z}^n . Show that we have the identification $\ker \varphi_{\mathcal{A}} = \text{Hom}_{\text{groups}}(\mathbb{Z}^n / \mathbb{Z}\mathcal{A}, \mathbb{K}^{\times})$.
3. Suppose that $\mathcal{A} \subset \mathbb{Z}^6 = (\mathbb{Z}^2)^3$ is represented by the 4×8 matrix

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}.$$

Find a set of nine linearly independent quadratic generators of the toric ideal $I_{\mathcal{A}}$.

4. Let $\mathcal{A} \subset \mathbb{Z}^6 = (\mathbb{Z}^2)^3$ be represented by the 6×8 matrix

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Find linearly independent quadratic generators of the toric ideal $I_{\mathcal{A}}$. Hint: The even (respectively odd) numbered rows give the vertices of the 3-cube. What is $X_{\mathcal{A}}$? For this, consider the map $\varphi_{\mathcal{A}}$ where the rows correspond to the variables $x_0, x_1, y_0, y_1, z_0, z_1$ and the columns to the variables $p_{000}, p_{100}, \dots, p_{111}$.

5. Show that the collection of 2×2 minors of the $m \times n$ matrix $(z_{i,j})_{i=1,\dots,m}^{j=1,\dots,n}$ of indeterminates forms a reduced Gröbner basis for the toric ideal of the variety $X_{\mathcal{A}}$ of matrices of rank 1, where the term order is degree reverse lexicographic with the variables ordered by $z_{i,l} \prec z_{k,j}$ if $i < k$ or $i = k$ and $l > j$. What about other term orders?
6. Show that in Corollary 8.1.4 we need only consider u with $\mathcal{A}u = 0$, where the coordinates of u do not have a common factor.
7. Complete the suggested proof of Theorem 8.1.5. Why is ‘reduced’ is necessary for the conclusion?
8. Let S be a finitely generated submonoid of \mathbb{Z}^n . Show that every maximal ideal \mathfrak{m} of $\mathbb{K}[S]$ restricts to a monoid homomorphism from S to \mathbb{K} , and vice-versa. Hint: use that maximal ideals correspond to algebra maps $\mathbb{K}[S] \rightarrow \mathbb{K}$.
9. For $\mathcal{A} \subset M$ finite, show that $X_{\mathcal{A}}(\mathbb{R}) = \text{Hom}_{\text{mon}}(\mathbb{N}\mathcal{A}, \mathbb{R})$ and $X_{\mathcal{A}, \geq} = \text{Hom}_{\text{mon}}(\mathbb{N}\mathcal{A}, \mathbb{R}_{\geq})$.
10. Describe the image of $(\mathbb{R}^{\times})^2$ under the map $\varphi_{\mathcal{A}}$ for \mathcal{A} the columns of the matrix $\begin{pmatrix} 1 & 1 & 0 \\ -1 & 1 & 2 \end{pmatrix}$. Show that $\varphi(\mathbb{R}^{\times})^2$ is not equal to the real points of the image of $(\mathbb{C}^{\times})^2$ under $\varphi_{\mathcal{A}}$.

8.2 Projective Toric Varieties

When an affine toric variety $X_{\mathcal{A}}$ is stable under multiplication by scalars in $\mathbb{K}^{\mathcal{A}}$ it may be considered to be a projective variety. Such projective toric varieties have a very close relation to the convex hull of the set \mathcal{A} and they provide a means to prove one of the signature results related to toric varieties; Kushnirenko’s Theorem about the number of solutions to a system of sparse polynomial equations, which we give in the next section.

For a finite set $\mathcal{A} \subset \mathbb{Z}^n$, we will write $\mathbb{P}^{\mathcal{A}}$ for the projective space $\mathbb{P}(\mathbb{K}^{\mathcal{A}})$. As explained in Section 1.4 it has homogeneous coordinates $[z_{\alpha} \mid \alpha \in \mathcal{A}]$. We claim that an affine toric

variety $X_{\mathcal{A}} \subset \mathbb{K}^{\mathcal{A}}$ is a stable under scalar multiplication when the set \mathcal{A} lies on an affine hyperplane. By this, we mean that there is some $w \in \mathbb{Z}^n$ with

$$w \cdot \alpha = w \cdot \beta \quad \text{for all } \alpha, \beta \in \mathcal{A},$$

and this common value c is nonzero. (An affine hyperplane does not contain the origin.) Then, under the composition of the cocharacter, $t \mapsto t^w$, with the map $\varphi_{\mathcal{A}}$, $t \in \mathbb{K}^\times$ acts as multiplication by the scalar t^c on $\mathbb{K}^{\mathcal{A}}$ as $\varphi_{\mathcal{A}}(t^w) = (t^{w \cdot \alpha} = t^c \mid \alpha \in \mathcal{A})$, and thus $X_{\mathcal{A}} \subset \mathbb{K}^{\mathcal{A}}$ is stable under scalar multiplication. For another way to see this, suppose that $u, v \in \mathbb{N}^{\mathcal{A}}$ are integer vectors with $\mathcal{A}u = \mathcal{A}v$. Then $w \cdot \mathcal{A}u = w \cdot \mathcal{A}v$, which implies that

$$c \sum_{a \in \mathcal{A}} u_a = c \sum_{a \in \mathcal{A}} v_a,$$

and thus $z^u - z^v \in I_{\mathcal{A}}$ is homogeneous and therefore $X_{\mathcal{A}} = \mathcal{V}(I_{\mathcal{A}})$ is stable under scalar multiplication. By Theorem 8.1.2, we obtain the following.

Corollary 8.2.1. *If \mathcal{A} lies on an affine hyperplane, then $I_{\mathcal{A}}$ is a homogeneous ideal and $X_{\mathcal{A}}$ is a projective subvariety of $\mathbb{P}^{\mathcal{A}}$.*

Example 8.2.2. Suppose that \mathcal{A} is represented by the matrix $\begin{pmatrix} 3 & 2 & 1 & 0 \\ 0 & 1 & 2 & 3 \end{pmatrix}$, which are the points a of \mathbb{N}^2 where $(1, 1) \cdot a = 3$. Then $\varphi_{\mathcal{A}}(x, y) = (x^3, x^2y, xy^2, y^3) \in \mathbb{K}^{\mathcal{A}} \simeq \mathbb{K}^4$, and the

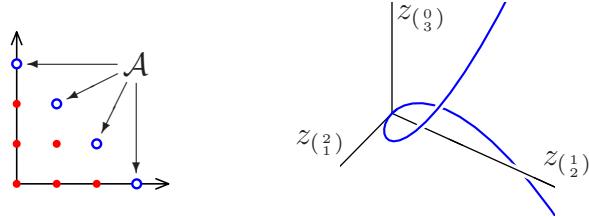


Figure 8.1: Exponents \mathcal{A} and the twisted cubic.

closure $X_{\mathcal{A}}$ of its image in $\mathbb{P}^{\mathcal{A}} = \mathbb{P}^3$ is the twisted cubic. If $[z_{(3)} : z_{(2)} : z_{(1)} : z_{(0)}]$ are the coordinates of $\mathbb{P}^{\mathcal{A}}$, then the homogeneous toric ideal $I_{\mathcal{A}}$ is generated by

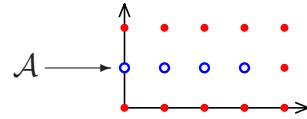
$$z_{(3)}z_{(1)} - z_{(2)}^2, \quad z_{(3)}z_{(0)} - z_{(2)}z_{(1)}, \quad \text{and} \quad z_{(2)}z_{(0)} - z_{(1)}^2, \tag{8.7}$$

which correspond to the vectors $(1, -2, 1, 0)^T$, $(1, 1, -1, -1)^T$, and $(0, 1, -2, 1)^T$ in $\ker \mathcal{A}$, which are also the primitive relations among the elements of \mathcal{A} ,

$$\binom{3}{0} + \binom{1}{2} = 2 \binom{2}{1}, \quad \binom{3}{0} + \binom{0}{3} = \binom{2}{1} + \binom{1}{2}, \quad \text{and} \quad \binom{2}{1} + \binom{0}{3} = 2 \binom{1}{2}.$$

Here, $\mathbb{Z}\mathcal{A}$ is a full rank sublattice of index 3 in \mathbb{Z}^2 , which you are asked to show in Exercise 1. When \mathbb{K} is algebraically closed, the kernel of $\varphi_{\mathcal{A}}$ is $\{(1), (\zeta), (\zeta^2)\}$, where

$\zeta := -\frac{1}{2} + \frac{\sqrt{-3}}{2}$ is a cube root of 1, and thus $\ker \varphi_{\mathcal{A}} \simeq \text{Hom}_g(\mathbb{Z}^2/\mathbb{Z}\mathcal{A}, \mathbb{K}^\times)$. Choosing $(\begin{smallmatrix} 3 \\ 0 \end{smallmatrix})$ and $(\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}) = (\begin{smallmatrix} 2 \\ 1 \end{smallmatrix}) - (\begin{smallmatrix} 3 \\ 0 \end{smallmatrix})$ as a basis for $\mathbb{Z}\mathcal{A}$ (and identifying it with \mathbb{Z}^2), the set \mathcal{A} becomes the columns of the matrix $(\begin{smallmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{smallmatrix})$, which we draw with the first coordinate vertical.



This is the set \mathcal{A} for the affine rational normal curve of Example 8.1.1 lifted to an affine hyperplane in \mathbb{Z}^2 by prepending a new first coordinate of 1 to each element of \mathcal{A} . \diamond

We turn to a geometric description of the generators of a homogeneous toric ideal $I_{\mathcal{A}}$. Suppose that $\mathcal{A} \subset \mathbb{Z}^n$ lies on an affine hyperplane. Let $u, v \in \mathbb{N}^{\mathcal{A}}$ be nonzero vectors with $u \neq v$ and $\mathcal{A}u = \mathcal{A}v$, so that $z^u - z^v$ is a binomial in $I_{\mathcal{A}}$. By our assumption on \mathcal{A} , the toric ideal $I_{\mathcal{A}}$ is homogeneous, so that $\deg z^u = \deg z^v$. Let $d := \sum_{\alpha} u_{\alpha} = \sum_{\alpha} v_{\alpha}$ be this degree. Writing $\lambda_{\alpha} := \frac{1}{d}u_{\alpha}$ and $\mu_{\alpha} := \frac{1}{d}v_{\alpha}$ for $\alpha \in \mathcal{A}$, we have

$$\sum_{\alpha \in \mathcal{A}} \alpha \lambda_{\alpha} = \sum_{\alpha \in \mathcal{A}} \alpha \mu_{\alpha}.$$

As $\lambda_{\alpha}, \mu_{\alpha} \geq 0$ are rational numbers and $\sum_{\alpha} \lambda_{\alpha} = \sum_{\alpha} \mu_{\alpha} = 1$, this is a point in $\text{conv}(\mathcal{A})$ having two distinct representations as a rational convex combination of points of \mathcal{A} .

Suppose that $\mathcal{A}u = 0$. Then $z^{u^+} - z^{u^-}$ is a generator for $I_{\mathcal{A}}$, by Corollary 8.1.4. Note that $\text{supp}(u^+)$ is disjoint from $\text{supp}(u^-)$. (The support *support* of v is the set of indices of its non-zero coordinates, $\text{supp}(v) := \{\alpha \in \mathcal{A} \mid v_{\alpha} \neq 0\}$.) Then the above construction (applied to u^+ and u^-) gives

$$\sum_{\alpha \in \text{supp}(u^+)} \alpha \lambda_{\alpha} = \sum_{\alpha \in \text{supp}(u^-)} \alpha \mu_{\alpha},$$

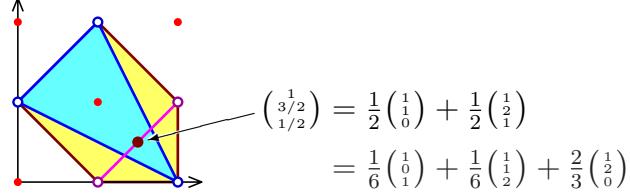
which is a rational point common to the convex hulls of two disjoint subsets of \mathcal{A} .

Conversely, suppose that $\mathcal{F}, \mathcal{F}' \subset \mathcal{A}$ are two disjoint subsets of \mathcal{A} whose convex hulls have a point in common. A rational point in the intersection of these convex hulls has two convex combinations,

$$\sum_{\alpha \in \mathcal{F}} \alpha \lambda_{\alpha} = \sum_{\alpha \in \mathcal{F}'} \alpha \mu_{\alpha}, \tag{8.8}$$

where λ_{α} and μ_{α} are rational numbers whose sum is 1. Extending these vectors $(\lambda_{\alpha} \mid \alpha \in \mathcal{F})$ and $(\mu_{\alpha} \mid \alpha \in \mathcal{F}')$ by zero to all of \mathcal{A} and then multiplying both by a common denominator gives vectors v and w in $\mathbb{N}^{\mathcal{A}}$ that satisfy $\mathcal{A}v = \mathcal{A}w$ with $\text{supp}(v) = \mathcal{F}$ disjoint from $\text{supp}(w) = \mathcal{F}'$. Setting $u = v - w$, we have $u^+ = v$ and $u^- = w$ and so we obtain a binomial $z^v - z^w \in I_{\mathcal{A}}$.

Example 8.2.3. Suppose that \mathcal{A} is represented by the matrix $\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 2 & 2 \\ 1 & 0 & 2 & 0 & 1 \end{pmatrix}$. The point $(1, \frac{3}{2}, \frac{1}{2})^T$ lies in the convex hull of two disjoint subsets of \mathcal{A} , $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 1 & 2 & 0 \end{pmatrix}$, respectively.



Clearing denominators in these two coincident convex combinations give the binomial $z_{\begin{pmatrix} 1 \\ 0 \end{pmatrix}}^3 z_{\begin{pmatrix} 1 \\ 1 \end{pmatrix}}^3 - z_{\begin{pmatrix} 1 \\ 1 \end{pmatrix}} z_{\begin{pmatrix} 1 \\ 2 \end{pmatrix}} z_{\begin{pmatrix} 1 \\ 0 \end{pmatrix}}^4 \in I_{\mathcal{A}}$. \square

We summarize this discussion.

Theorem 8.2.4. Suppose that \mathcal{A} lies on an affine hyperplane. Homogeneous generators $z^{u^+} - z^{u^-}$ of $I_{\mathcal{A}}$ of Corollary 8.1.4 correspond to rational points of $\text{conv}(\mathcal{A})$ lying in the intersection of convex hulls of two disjoint subsets of \mathcal{A} .

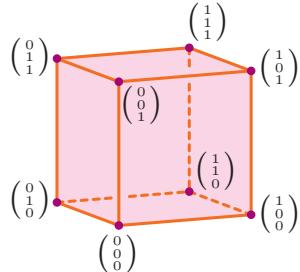
For a finite subset $\mathcal{A} \subset \mathbb{Z}^n$, let $\text{conv}(\mathcal{A})$ be its convex hull. For any $w \in \mathbb{R}^n$ the support function $h_{\mathcal{A}}(w)$ is the minimum value that the function $a \mapsto w \cdot a$ takes on points of \mathcal{A} ; this equals the support function of $\text{conv}(\mathcal{A})$, as given in Section A.3.1 of the Appendix. The subset of \mathcal{A} where the function $\alpha \mapsto w \cdot \alpha$ attains its minimum,

$$\mathcal{A}_w := \{\alpha \in \mathcal{A} \mid w \cdot \alpha = h_{\mathcal{A}}(w)\}, \quad (8.9)$$

is the face of \mathcal{A} *exposed* by w . This is the intersection of \mathcal{A} with the face $\text{conv}(\mathcal{A})_w$ of its convex hull exposed by w , and we have $\text{conv}(\mathcal{A}_w) = \text{conv}(\mathcal{A})_w$.

A *face* of \mathcal{A} is any subset of the form \mathcal{A}_w . As $\mathcal{A} \subset \mathbb{Z}^n$, the set of $w \in \mathbb{R}^n$ that expose a given face \mathcal{F} of \mathcal{A} in that $\mathcal{F} = \mathcal{A}_w$ is given by linear equations and linear inequalities with integer coefficients (coming from the elements of \mathcal{A} and \mathcal{F}). Thus there is a $w \in \mathbb{Z}^n$ that exposes \mathcal{F} . As \mathcal{A} is considered to be a set of characters, it will be natural to consider such an integral w as a cocharacter.

When \mathcal{A} is the set of column vectors of $\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$, which are the vertices of the lattice cube, the face exposed by the vector $(-1, -2, -3)$ is the vertex $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$, the face exposed by the vector $(-1, 2, 0)$ is the subset $\{\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}\}$ spanning an edge of the cube, and the face exposed by the vector $(0, 0, 1)$ is the subset $\{\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}\}$, which spans the downward-pointing facet.



For any subset \mathcal{F} of \mathcal{A} , the inclusion $\mathcal{F} \subset \mathcal{A}$ of subsets induces an inclusion of projective spaces $\mathbb{P}^{\mathcal{F}} \subset \mathbb{P}^{\mathcal{A}}$, where $\mathbb{P}^{\mathcal{F}}$ is identified with the coordinate subspace $\{z \in \mathbb{P}^{\mathcal{A}} \mid z_{\alpha} = 0 \text{ if } \alpha \notin \mathcal{F}\}$ of $\mathbb{P}^{\mathcal{A}}$. We state a relation between faces \mathcal{F} of \mathcal{A} and toric subvarieties of $X_{\mathcal{A}}$.

Lemma 8.2.5. *Let $X_{\mathcal{A}} \subset \mathbb{P}^{\mathcal{A}}$ be the projective toric variety given by finite set $\mathcal{A} \subset \mathbb{Z}^n$ lying on an affine hyperplane. For any point $z \in X_{\mathcal{A}}$, its support is a face of \mathcal{A} . For every face \mathcal{F} of \mathcal{A} , the intersection $X_{\mathcal{A}} \cap \mathbb{P}^{\mathcal{F}}$ is naturally identified with $X_{\mathcal{F}}$.*

Call a toric subvariety $X_{\mathcal{F}} \subset X_{\mathcal{A}}$ for \mathcal{F} a face of \mathcal{A} a *facial subvariety* of $X_{\mathcal{A}}$. By Lemma 8.2.5, the complement of $\varphi_{\mathcal{A}}(\mathbb{K}^{\times})^n$ in $X_{\mathcal{A}}$ is the union of all proper facial subvarieties.

Proof. Let $z \in X_{\mathcal{A}}$ and set $\mathcal{F} := \text{supp}(z)$. We claim that \mathcal{F} is a face of \mathcal{A} . Let \mathcal{E} be the smallest face of \mathcal{A} containing \mathcal{F} and suppose that $\alpha \in \mathcal{E} \setminus \mathcal{F}$. Then \mathcal{E} is a subset of the affine span of \mathcal{F} , so there exist rational numbers $\{c_f \mid f \in \mathcal{F}\}$ with sum 1 such that

$$\alpha = \sum_{f \in \mathcal{F}} c_f f \quad \text{or} \quad \alpha - \sum_{f \in \mathcal{F}} c_f f = 0.$$

Multiplying by the common denominator d of the numbers c_f , the second equation gives an integer relation in the kernel of \mathcal{A} , and thus a homogeneous generator $z^{u+} - z^{u-}$ in $I_{\mathcal{A}}$, by Corollary 8.1.4.

This binomial vanishes at the point z and only involves coordinates z_{α} and z_f for $f \in \mathcal{F}$. Using that the coordinates $z_f \neq 0$ for $f \in \mathcal{F}$, we rewrite $0 = z^{u+} - z^{u-}$, solving for z_{α} , to get

$$z_{\alpha}^d = \prod_{f \in \mathcal{F}} z_f^{d c_f}.$$

Since $z_f \neq 0$ for $f \in \mathcal{F}$, this implies that $z_{\alpha} \neq 0$. But then $\alpha \in \text{supp}(z) = \mathcal{F}$, which contradicts our assumption that $\alpha \notin \mathcal{F}$. Thus \mathcal{F} is a face of \mathcal{A} .

Suppose now that \mathcal{F} is a face of \mathcal{A} . Let $z \in X_{\mathcal{A}} \cap \mathbb{P}^{\mathcal{F}}$. By the geometric description of generators of $I_{\mathcal{A}}$ of Theorem 8.2.4 in terms of coincident convex combinations of elements of $\text{conv}(\mathcal{A})$, we have that $I_{\mathcal{F}} \subset I_{\mathcal{A}}$, as $\mathcal{F} \subset \mathcal{A}$. Thus $z \in \mathcal{V}(I_{\mathcal{F}}) = X_{\mathcal{F}}$ as $z \in \mathbb{P}^{\mathcal{F}}$ and $I_{\mathcal{F}}$ is the homogeneous ideal of $X_{\mathcal{F}}$. Thus shows that $X_{\mathcal{A}} \cap \mathbb{P}^{\mathcal{F}} \subset X_{\mathcal{F}}$.

For the other inclusion, let $w \in \mathbb{Z}^n$ be a cocharacter that exposes \mathcal{F} , in that $\mathcal{A}_w = \mathcal{F}$. Let $x \in (\mathbb{K}^{\times})^n$ and $t \in \mathbb{K}^{\times}$, and let us compute $\varphi_{\mathcal{A}}(t^w \cdot x)$, which is

$$[(t^w \cdot x)^{\alpha} \mid \alpha \in \mathcal{A}] = [t^{w \cdot \alpha} x^{\alpha} \mid \alpha \in \mathcal{A}] = [t^{w \cdot \alpha - h_{\mathcal{A}}(w)} x^{\alpha} \mid \alpha \in \mathcal{A}]. \quad (8.10)$$

The last equality is because this is a point in the projective space $\mathbb{P}^{\mathcal{A}}$. Consider the α -coordinate of $\varphi_{\mathcal{A}}(t^w \cdot x)$ in this representation. If $\alpha \in \mathcal{F}$, then $w \cdot \alpha = h_{\mathcal{A}}(w)$, so that the α -coordinate is x^{α} . If $\alpha \notin \mathcal{F}$, then $w \cdot \alpha > h_{\mathcal{A}}(w)$, so that t divides the α -coordinate.

Fix $x \in (\mathbb{K}^{\times})^n$. Then $t \mapsto \varphi_{\mathcal{A}}(t^w \cdot x)$ is a map from \mathbb{K}^{\times} to $X_{\mathcal{A}}$. The last expression for $\varphi_{\mathcal{A}}(t^w \cdot x)$ in (8.10) allows us to extend this to a map f from \mathbb{K} to $X_{\mathcal{A}}$. Indeed, as t only occurs to nonnegative powers, it gives a map from \mathbb{K} to $\mathbb{K}^{\mathcal{A}}$. As t does not occur

in the coordinates indexed by elements of \mathcal{F} , its value $f(0)$ at $t = 0$ is nonzero, and therefore $f(0)$ is a well-defined point of $\mathbb{P}^{\mathcal{A}}$. Since $f(\mathbb{K}^\times) \subset X_{\mathcal{A}}$ and $X_{\mathcal{A}}$ is Zariski-closed, $f(0) \in X_{\mathcal{A}}$. Finally, observe that $f(0) = \varphi_{\mathcal{F}}(x)$ as t divides coordinates of $f(t)$ indexed by elements $\alpha \notin \mathcal{F}$. Since this is true for all $x \in (\mathbb{K}^\times)^n$, we have shown that $\varphi_{\mathcal{F}}(\mathbb{K}^\times)^n \subset X_{\mathcal{A}}$. Taking Zariski closures shows that $X_{\mathcal{F}} \subset X_{\mathcal{A}}$, and completes the proof. \square

We defined a projective toric variety as the image in projective space $\mathbb{P}^{\mathcal{A}}$ of a homogeneous affine toric variety $X_{\mathcal{A}}$; this required that \mathcal{A} lie on an affine hyperplane. Projective varieties also arise as closures of affine varieties via homogenization. We explain a useful version of this construction for affine toric varieties.

Given a finite set $\mathcal{A} \subset \mathbb{Z}^n$, the corresponding toric variety $X_{\mathcal{A}} \subset \mathbb{K}^{\mathcal{A}}$ is not necessarily homogeneous. This occurs for example, if $0 \in \mathcal{A}$. The *lift* of \mathcal{A} is the set

$$\mathcal{A}^+ := \{(1, \alpha) \mid \alpha \in \mathcal{A}\} \subset \mathbb{Z}^{1+n},$$

which lies on an affine hyperplane in \mathbb{Z}^{1+n} . The map $\varphi_{\mathcal{A}^+}: \mathbb{K}^\times \times (\mathbb{K}^\times)^n \rightarrow \mathbb{P}^{\mathcal{A}}$ is given by

$$\varphi_{\mathcal{A}^+}(t, x) = [tx^\alpha \mid \alpha \in \mathcal{A}]. \quad (8.11)$$

In Exercise 2 you are asked to show that the set of differences $\{\alpha - \beta \mid \alpha, \beta \in \mathcal{A}\}$ spans \mathbb{Z}^n if and only if the set \mathcal{A}^+ of vectors spans \mathbb{Z}^{1+n} .

Example 8.2.6. Suppose that \mathcal{A} is represented by the matrix $\begin{pmatrix} 0 & 1 & 1 & 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & 1 & 0 & -1 & -1 \end{pmatrix}$. Then \mathcal{A} consists of the integer points of the hexagon in Figure 8.2. Figure 8.2 also shows the lift of this hexagon, where the first coordinate is vertical in the lift. \diamond

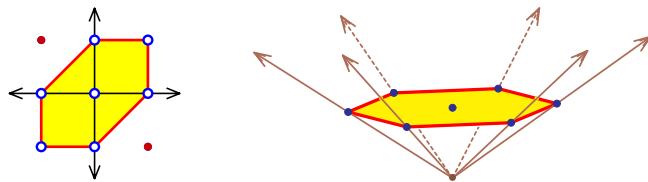


Figure 8.2: The hexagon and its lift.

In Exercise 3, you are asked to show that any finite set $\mathcal{A} \subset \mathbb{Z}^n$ lying on an affine hyperplane has the form \mathcal{B}^+ in appropriate coordinates for $\mathbb{Z}\mathcal{A}$. Exercise 4 gives another (equivalent) characterization of a set \mathcal{A} lying on an affine hyperplane.

The map $\varphi_{\mathcal{A}}$ parameterizes $X_{\mathcal{A}^+}$, and we first study the injectivity of this parametrization. Recall that the affine span (A.4) of a finite set \mathcal{A} is the collection of all affine combinations of elements of \mathcal{A} —these are linear combinations where the sum of the coefficients is 1. This differs from the convex hull in that the coefficients λ_α may be negative. When the coefficients are integers, this is the *integral affine span* $\text{Aff}_{\mathbb{Z}} \mathcal{A}$. For any $\alpha \in \mathcal{A}$, the affine span is the coset

$$\text{Aff } \mathcal{A} = \alpha + \mathbb{R}\{\beta - \alpha \mid \beta \in \mathcal{A}\}, \quad (8.12)$$

and the same expression (replacing \mathbb{Z} for \mathbb{R}) gives the integral affine span.

The map $\varphi_{\mathcal{A}}: (\mathbb{K}^\times)^n \rightarrow \mathbb{P}^{\mathcal{A}}$ is the restriction of $\varphi_{\mathcal{A}^+}$ (8.11) to the subtorus $\{1\} \times (\mathbb{K}^\times)^n$ of the torus $\mathbb{K}^\times \times (\mathbb{K}^\times)^n$ where $t = 1$. In Exercise 5 you are asked to show that $\varphi_{\mathcal{A}}$ is injective if and only if $\text{Aff}_{\mathbb{Z}} \mathcal{A} = \mathbb{Z}^n$. **Reconcile this with the following remark.**

Remark 8.2.7. Suppose that $0 \in \mathcal{A}$. Then the z_0 -coordinate of $\varphi_{\mathcal{A}}(x)$ is $x^0 = 1$, for any $x \in (\mathbb{K}^\times)^n$. Thus the affine variety $X_{\mathcal{A}} \subset \mathbb{K}^{\mathcal{A}}$ may be considered to lie in the principal affine set $U_{z_0} \subset \mathbb{P}^{\mathcal{A}}$ consisting of points $z \in \mathbb{P}^{\mathcal{A}}$ with $z_0 \neq 0$. This is identified with $\mathbb{K}^{\mathcal{A} \setminus \{0\}}$ and $X_{\mathcal{A}}$ is equal to $X_{\mathcal{A} \setminus \{0\}}$. We also have that $X_{\mathcal{A} \setminus \{0\}} = X_{\mathcal{A}^+} \cap U_{z_0}$, so that the projective toric variety $X_{\mathcal{A}^+}$ is the projective closure of the affine toric variety $X_{\mathcal{A}} = X_{\mathcal{A} \setminus \{0\}}$.

Now let $\mathcal{A} \subset \mathbb{Z}^n$ be any finite set and let $\alpha \in \mathcal{A}$. Let us consider $X_{\mathcal{A}^+} \cap U_{z_\alpha}$, the points of $X_{\mathcal{A}^+}$ with non-vanishing z_α -coordinate. For any $x \in (\mathbb{K}^\times)^n$, the z_α -coordinate of $x^{-\alpha} \varphi_{\mathcal{A}}(x)$ is 1. Thus if we set $\mathcal{B} := \{\beta - \alpha \mid \beta \in \mathcal{A}\}$, then for $x \in (\mathbb{K}^\times)^n$, we have that $\varphi_{\mathcal{B}}(x) = x^{-\alpha} \varphi_{\mathcal{A}}(x)$, up to a shift in their indices. This shows that $X_{\mathcal{A}^+}$ is the projective closure of the affine toric variety $X_{\mathcal{B} \setminus \{0\}}$. Since for $x \in (\mathbb{K}^\times)^n$ we have the equality $x^{-\alpha} \varphi_{\mathcal{A}}(x) = \varphi_{\mathcal{A}}(x)$ as points in projective space, we see that $X_{\mathcal{A}^+}$ is the projective closure of the image $\varphi_{\mathcal{A}}(\mathbb{K}^\times)^n$. \diamond

Exercises

1. Verify that the subgroup $\mathbb{Z}\mathcal{A}$ for the set $\mathcal{A} = \begin{pmatrix} 3 & 2 & 1 & 0 \\ 0 & 1 & 2 & 3 \end{pmatrix}$ of Example 8.2.2 is a rank 2 subgroup of index 3 in \mathbb{Z}^2 . You may find the map $\mathbb{Z}^2 \rightarrow \mathbb{Z}$ given by $(p, q) \mapsto p + q$ useful; consider its kernel, image, and cokernel, and its restriction to $\mathbb{Z}\mathcal{A}$.
2. Let $\mathcal{A} \subset \mathbb{Z}^n$ be a finite set of points. Show that its lift $\mathcal{A}^+ \subset \mathbb{Z}^{1+n}$ spans \mathbb{Z}^{1+n} if and only if the set of differences $\{\alpha - \beta \mid \alpha, \beta \in \mathcal{A}\}$ spans \mathbb{Z}^n .
3. Prove that if a finite set $\mathcal{A} \subset \mathbb{Z}^n$ lies on an affine hyperplane and $\text{rank}(\mathbb{Z}\mathcal{A}) = 1 + m$, then there is a basis for $\mathbb{Z}\mathcal{A}$ identifying it with \mathbb{Z}^{1+m} and a subset $\mathcal{B} \subset \mathbb{Z}^m$ such that $\mathcal{A} = \mathcal{B}^+$.
4. Suppose that $\mathcal{A} \subset \mathbb{Z}^n$ is represented by an integer matrix, A . Show that \mathcal{A} lies on an affine hyperplane if and only if the row space of A in $\mathbb{R}^{\mathcal{A}}$ has a vector with every coordinate 1.
5. Let $\mathcal{A} \subset \mathbb{Z}^n$ be a finite collection of exponent vectors for Laurent monomials. Show that the map $\varphi_{\mathcal{A}}: (\mathbb{K}^\times)^n \rightarrow \mathbb{P}^{\mathcal{A}}$ 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 .)

8.3 Kushnirenko's Theorem

We turn to one of the most celebrated applications of toric varieties, understanding the number of solutions to a system of polynomial equations. A Laurent polynomial $f \in \mathbb{K}[x^\pm]$

is a finite linear combination of monomials. That is, there are coefficients $c_\alpha \in \mathbb{K}$ for $\alpha \in \mathbb{Z}^n$ such that

$$f = \sum_{\alpha} c_\alpha x^\alpha,$$

with at most finitely many coefficients c_α nonzero. The set \mathcal{A} of indices of nonzero coefficients is called the *support* of f and its convex hull is the Newton polytope $\text{New}(f)$ of f . We consider the number of solutions in $(\mathbb{K}^\times)^n$ to a system

$$f_1(x) = f_2(x) = \cdots = f_n(x) = 0 \quad (8.13)$$

of (Laurent) polynomial equations, where each polynomial has the same support \mathcal{A} . The coefficients of a polynomial identify $\mathbb{K}^\mathcal{A}$ with the set of polynomials whose support is a subset of \mathcal{A} , and $(\mathbb{K}^\mathcal{A})^n$ is identified with set of polynomial systems (8.13) with support \mathcal{A} . When \mathbb{K} is algebraically closed, Kushnirenko proved the following count for the number of solutions to a system of polynomial equations (8.13) with support \mathcal{A} . For a polytope $P \subset \mathbb{R}^n$, we let $\text{vol}(P)$ be its usual volume, in which the unit cube $[0, 1]^n$ has volume 1.

Theorem 8.3.1 (Kushnirenko). *A system (8.13) of n polynomials in n variables with support \mathcal{A} has at most $n! \text{vol}(\text{conv}(\mathcal{A}))$ isolated solutions in $(\mathbb{K}^\times)^n$. When \mathbb{K} is algebraically closed, there is a dense open subset of $(\mathbb{K}^\mathcal{A})^n$ consisting of systems with support \mathcal{A} having exactly $n! \text{vol}(\text{conv}(\mathcal{A}))$ simple solutions in $(\mathbb{K}^\times)^n$.*

Remark 8.3.2. Suppose that $\mathcal{A} := \{\alpha \in \mathbb{N}^n \mid \alpha_1 + \cdots + \alpha_n \leq d\}$, the set of exponents of monomials in x_1, \dots, x_n of total degree at most d . A polynomial system (8.13) with support \mathcal{A} is a system of n polynomials, each of degree d . As all monomials of degree d or less may occur, this is called a *dense system*. The Newton polytope $\text{conv}(\mathcal{A})$ is $d\Delta$, where Δ is the convex hull of the origin and the standard unit vectors. Since this unit simplex has volume $\frac{1}{n!}$ (see Exercise 1), the volume of $d\Delta$ is $\frac{d^n}{n!}$. We deduce the following special case of Bézout's Theorem 2.4.5 from Kushnirenko's Theorem: the number of solutions to a general dense system of degree d in n variables is d^n . \diamond

When the support \mathcal{A} of a system of polynomials does not have $\text{conv}(\mathcal{A}) = d\Delta$, the system is said to be *sparse*.

Remark 8.3.3. Lemma 8.6.4 of Section 8.6 establishes the claim that there is an open set in $(\mathbb{K}^\mathcal{A})^n$ of systems with support \mathcal{A} all of which have the same number of isolated solutions and where each solution is simple. Write $d(\mathcal{A})$ for this number. Theorem Baby-Bertini implies the existence of a dense open subset $U \subset (\mathbb{K}^\mathcal{A})^n$ consisting of systems $F = (f_1, \dots, f_n)$ which have only simple solutions. Theorem 8.6.5 gives conditions on such systems F so that the number of points in $\mathcal{V}(F)$ equals $d(\mathcal{A})$.

For a face \mathcal{A}_w of \mathcal{A} exposed by a non zero vector w and a Laurent polynomial f with support a subset of \mathcal{A} , write f_w for the restriction of f to \mathcal{A}_w , the sum of terms in f whose exponents lie in \mathcal{A}_w . This is called the *facial form* of f . In the terminology of Gröbner bases from Chapter 2, this is the initial form of f in the weighted partial order \succ_{-w} of Example 2.2.5. The *facial system* F_w of $F = (f_1, \dots, f_n) \in (\mathbb{K}^\mathcal{A})^n$ is the system

consisting of the facial forms $(f_{1,w}, \dots, f_{n,w})$. Theorem 8.6.5 states that for $F \in U$, we have $\#\mathcal{V}(F) \leq d(\mathcal{A})$ with equality exactly when $\mathcal{V}(F_w) = \emptyset$ for all non zero w . We will deduce these facts for the unmixed systems of Kushnirenko's Theorem 8.3.1 using the geometry of the projective toric variety $X_{\mathcal{A}^+}$.

Note that the facial form f_w depends not only on f but also on the set \mathcal{A} . If $\mathcal{A} \subset \mathcal{B}$, then a Laurent polynomial f with support \mathcal{A} lies in $\mathbb{K}^{\mathcal{B}}$, but the restriction of f to the face \mathcal{A}_w of \mathcal{A} exposed by w may not equal its restriction to \mathcal{B}_w . \diamond

Remark 8.2.7 explained that $X_{\mathcal{A}^+}$ is the projective closure of the image of $(\mathbb{K}^\times)^n$ under the map $\varphi_{\mathcal{A}}$. We will use this to relate linear sections of $X_{\mathcal{A}^+}$ to systems of polynomials with support \mathcal{A} , and then use the Hilbert function of the projective toric variety $X_{\mathcal{A}^+}$ to prove Kushnirenko's Theorem.

Given a linear form Λ on $\mathbb{P}^{\mathcal{A}}$,

$$\Lambda = \sum_{\alpha \in \mathcal{A}} c_\alpha z_\alpha,$$

its pullback $\varphi_{\mathcal{A}}^*(\Lambda)$ along $\varphi_{\mathcal{A}}$ is a polynomial with support \mathcal{A} ,

$$\varphi_{\mathcal{A}}^*(\Lambda) = \sum_{\alpha \in \mathcal{A}} c_\alpha x^\alpha.$$

Consequently, a system of n polynomials (8.13) with support \mathcal{A} is the pullback along $\varphi_{\mathcal{A}}$ of a system of n linear forms on $\mathbb{P}^{\mathcal{A}}$. Note that a linear form Λ on $\mathbb{P}^{\mathcal{A}}$ defines a hyperplane $H \subset \mathbb{P}^{\mathcal{A}}$ and n general linear forms define a linear subspace L of codimension n .

Lemma 8.3.4. *The solution set of a system of polynomials (8.13) with support \mathcal{A} is the pullback $\varphi_{\mathcal{A}}^{-1}(L) = \varphi_{\mathcal{A}}^{-1}(L \cap \varphi_{\mathcal{A}}(\mathbb{K}^\times)^n)$ of a linear section of $\varphi_{\mathcal{A}}(\mathbb{K}^\times)^n$, where L has codimension equal to the dimension of the linear span of the polynomials f_i .*

Since $\varphi_{\mathcal{A}}$ is a homomorphism, the solutions are cosets of the kernel of $\varphi_{\mathcal{A}}$, one coset for each point in the linear section $L \cap \varphi_{\mathcal{A}}(\mathbb{K}^\times)^n$.

Example 8.3.5. Consider the polynomial system

$$f := x^2y + 2xy^2 - 1 + xy = 0 \quad \text{and} \quad g := x^2y - xy^2 + 2 - xy = 0. \quad (8.14)$$

These polynomials define two plane curves which have one real point of intersection at $(1.53277, -0.90655)$ and are displayed in Figure 8.3. The exponent vectors \mathcal{A} are the columns of the matrix $(\begin{smallmatrix} 2 & 1 & 0 & 1 \\ 1 & 2 & 0 & 1 \end{smallmatrix})$. The map $\varphi_{\mathcal{A}}$ is

$$(x, y) \mapsto [x^2y : xy^2 : 1 : xy] \in \mathbb{P}^{\mathcal{A}} \simeq \mathbb{P}^3.$$

This is injective as the exponent vectors \mathcal{A} span \mathbb{Z}^2 . Its image consists of those points $[z_{(2)} : z_{(1)} : z_{(0)} : z_{(1)}]$ with $z_{(2)}z_{(1)}z_{(0)} = z_{(1)}^3 \neq 0$, which is part of a cubic surface. The polynomial system (8.14) corresponds to the two linear forms

$$z_{(2)} + 2z_{(1)} - z_{(0)} + z_{(1)} = z_{(2)} - z_{(1)} + 2z_{(0)} - z_{(1)} = 0.$$

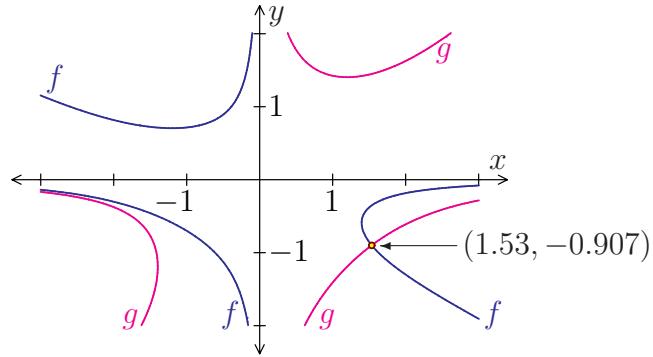


Figure 8.3: Curves of polynomial system (8.14).

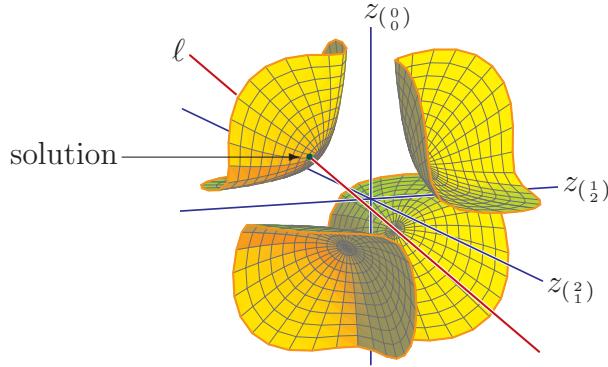


Figure 8.4: Linear section of cubic surface.

These define a line ℓ in \mathbb{P}^3 . Figure 8.4 shows ℓ and (part of) the cubic surface. This is in the affine part of \mathbb{P}^A where $z_{(1)} \neq 0$ in the box $[-4, 4]^3$. The best view is from the $+-+$ -orthant. From this, we see that there is one real solution to the system (8.14). \diamond

By Remark 8.2.7, the projective toric variety $X_{\mathcal{A}^+}$ is the Zariski closure of $\varphi_{\mathcal{A}}(\mathbb{K}^\times)^n$ in \mathbb{P}^A . Thus for a linear subspace L , we have $L \cap \varphi_{\mathcal{A}}(\mathbb{K}^\times)^n \subset L \cap X_{\mathcal{A}^+}$. By Corollary 3.6.1 if L is general and has codimension n , then $L \cap X_{\mathcal{A}^+}$ is zero-dimensional and every point of the intersection is smooth. Thus $L \cap \varphi_{\mathcal{A}}(\mathbb{K}^\times)^n$ is a finite set and smooth. By Lemma 8.3.4, the solutions are cosets of the kernel of $\varphi_{\mathcal{A}}$. This implies that all solutions to the system F corresponding to L are simple.

This gives another explanation for the set U of Remark 8.3.3 given by Theorem 8.6.5. By Lemma 8.2.5, the difference $X_{\mathcal{A}^+} \setminus \varphi_{\mathcal{A}}(\mathbb{K}^\times)^n$ is the union of facial subvarieties $X_{\mathcal{F}^+}$ corresponding to proper faces \mathcal{F} of \mathcal{A} . Thus $L \cap \varphi_{\mathcal{A}}(\mathbb{K}^\times)^n = L \cap X_{\mathcal{A}^+}$ and thus $\#\mathcal{V}(F) = d(\mathcal{A})$ if and only if L is disjoint from all such proper facial subvarieties $X_{\mathcal{F}^+}$ of $X_{\mathcal{A}^+}$.

In Exercise 4 you are asked to show that if $\mathcal{F} = \mathcal{A}_w$, then under the map $\varphi_{\mathcal{F}}: (\mathbb{K}^\times)^n \rightarrow \mathbb{P}^{\mathcal{F}} \subset \mathbb{P}^A$, the pullback $\varphi_{\mathcal{F}}^{-1}(\Lambda)$ of a linear form Λ is the variety $\mathcal{V}(f_w)$ of the facial form f_w , where f is the Laurent polynomial corresponding to Λ . Thus the condition of Theorem 8.6.5 that the facial systems F_w have no solutions for all non-zero $w \in \mathbb{R}^n$ is

equivalent to $L \cap X_{\mathcal{A}^+} \subset \varphi_{\mathcal{A}}(\mathbb{K}^\times)^n$.

Recall that a map, such as $\varphi_{\mathcal{A}}$ with finite fibers has a degree, which is the cardinality of a fiber. In this case, $\varphi_{\mathcal{A}}$ is a homomorphism so its degree is the cardinality of its kernel. The degree of a projective variety is the cardinality of a general linear section.

Lemma 8.3.6. *We have $d(\mathcal{A}) = \deg(\varphi_{\mathcal{A}}) \cdot \deg(X_{\mathcal{A}^+})$. When the affine span of \mathcal{A} is \mathbb{Z}^n , then $d(\mathcal{A}) = \deg(X_{\mathcal{A}^+})$.*

Proof. Let F be a system with support \mathcal{A} having finitely many simple solutions that corresponds to a linear subspace $L \subset \mathbb{P}^{\mathcal{A}}$. By Lemma 8.3.4, the number of solutions of F is the product of $\deg(\varphi_{\mathcal{A}})$ and the number of points in $L \cap \varphi_{\mathcal{A}}(\mathbb{K}^{\text{times}})^n$. By the discussion preceding the lemma, the number of points in this linear section is maximized for general L for which $L \cap \varphi_{\mathcal{A}}(\mathbb{K}^{\text{times}})^n = L \cap X_{\mathcal{A}^+}$, and in that case, this number is $\deg(X_{\mathcal{A}^+})$. This proves the first statement.

The second statement follows from the first, for by Exercise 5 of Section 8.2 the map $\varphi_{\mathcal{A}}$ is injective, and therefore has degree 1, when the affine span of \mathcal{A} is \mathbb{Z}^n . \square

We will prove Kushnirenko's theorem in the case when $\text{Aff}_{\mathbb{Z}} \mathcal{A} = \mathbb{Z}^n$ by showing that

$$n! \cdot \text{vol}(\text{conv}(\mathcal{A})) = \deg(X_{\mathcal{A}^+}).$$

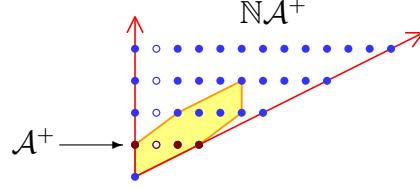
We use this to deduce the general case of Kushnirenko's Theorem at the end of this section.

This will need to be rewritten once Chapter 3 is complete. Recall that the homogeneous coordinate ring $\mathbb{K}[X]$ of a projective variety $X \subset \mathbb{P}^{\mathcal{A}}$ is the quotient of the homogeneous coordinate ring $\mathbb{K}[z_\alpha \mid \alpha \in \mathcal{A}]$ of $\mathbb{P}^{\mathcal{A}}$ by the ideal I_X of homogeneous polynomials vanishing on X . These rings and ideals are graded by the total degree of the polynomials. Writing $\mathbb{K}_d[X]$ for the d th graded piece of $\mathbb{K}[X]$, the Hilbert function $\text{HF}_X(d)$ is the function $d \mapsto \dim_{\mathbb{K}} \mathbb{K}_d[X]$.

Hilbert proved that the Hilbert function for $d \gg 0$ is equal to a polynomial, which is now called the Hilbert polynomial $\text{HP}_X(d)$ of X . This encodes many numerical invariants of X . For example, the degree of the Hilbert polynomial is the dimension n of X and its leading coefficient is $\deg(X)/n!$.

We determine the Hilbert polynomial of the toric variety $X_{\mathcal{A}^+}$. Its homogeneous coordinate ring is the coordinate ring of $X_{\mathcal{A}^+} \subset \mathbb{K}^{\mathcal{A}}$. By Corollary 8.1.3, this is the coordinate ring $\mathbb{K}[x^\alpha \mid \alpha \in \mathcal{A}^+]$, which is isomorphic to $\mathbb{K}[\mathbb{N}\mathcal{A}^+]$. As the first coordinate 1 of points of \mathcal{A}^+ corresponds to the homogenizing parameter t in (8.11), $\mathbb{K}[\mathbb{N}\mathcal{A}^+]$ is graded by the first component of elements of $\mathbb{N}\mathcal{A}^+$. Thus $\mathbb{K}_d[\mathbb{N}\mathcal{A}^+]$ has a basis $\{(d, \alpha) \in \mathbb{N}\mathcal{A}^+ \mid d \geq 0\}$. Projection to the last n coordinates identifies this with $\mathbb{N}\mathcal{A}^+ \cap d \text{conv}(\mathcal{A}^+)$, which is $d\mathcal{A}^+$, the set of d -fold sums of vectors in \mathcal{A}^+ .

Example 8.3.7. Consider this for the projectivization $X_{\mathcal{A}^+}$ of the cuspidal cubic of Example 8.1.1. Here, $\mathcal{A} = \{0, 2, 3\}$ and $\varphi_{\mathcal{A}}(s) = [1, s^2, s^3] \in \mathbb{P}^{\mathcal{A}}$. Figure 8.5 shows its lift $\mathcal{A}^+ = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 3 \end{pmatrix}$ and the submonoid $\mathbb{N}\mathcal{A}^+$. The open circles are points that do not lie in $\mathbb{N}\mathcal{A}^+$. The Hilbert function of $X_{\mathcal{A}^+}$ has values $(1, 3, 6, 9, 12, \dots)$, so its Hilbert polynomial is $3d$. \square

Figure 8.5: Submonoid generated by the lift \mathcal{A}^+ of $\{0, 2, 3\}$.

Projecting the set $d\mathcal{A}^+$ to the last n coordinates is a bijection with the set $d\mathcal{A}$ of d -fold sums of vectors in \mathcal{A} . These arguments show that

$$\text{HF}_{X_{\mathcal{A}}}(d) = |d\mathcal{A}|.$$

Thus an upper bound on $\text{HF}_{X_{\mathcal{A}}}(d)$ is given by $|d \text{conv}(\mathcal{A}) \cap \mathbb{Z}^n|$, as $d\mathcal{A} \subset d \text{conv}(\mathcal{A}) \cap \mathbb{Z}^n$.

Need Ehrhart polynomials. Appendix ? For an integer polytope, the counting function

$$E_P : \mathbb{N} \ni d \longmapsto |dP \cap \mathbb{Z}^n|$$

for the integer points contained in positive integer multiples of P is a polynomial in d , called the *Ehrhart polynomial* of P . The degree of E_P is the dimension of the affine span of P . When P has dimension n , its leading coefficient is the volume of P . For example, the Ehrhart polynomial of the interval $[0, 3] = \text{conv}\{0, 2, 3\}$ of length 3 is $3d + 1$.

Now suppose that $P = \text{conv}(\mathcal{A})$, the convex hull of \mathcal{A} . Since $d\mathcal{A} \subset d \text{conv}(\mathcal{A}) \cap \mathbb{Z}^n$, we have the upper bound for $\text{HF}_{X_{\mathcal{A}}}(d)$,

$$\text{HF}_{X_{\mathcal{A}}}(d) \leq E_{\text{conv}(\mathcal{A})}(d). \quad (8.15)$$

Note that we have this inequality for the cubic of Example 8.3.7.

A lower bound for $\text{HF}_{X_{\mathcal{A}}}(d)$ is best expressed in terms of an inclusion of semigroups, which are more general than monoids, as they do not necessarily have an identity. Let $S_{\mathcal{A}} := \mathbb{R}_{\geq 0} \mathcal{A}^+ \cap \mathbb{Z}^{1+n}$ be the monoid/semigroup of all integer points that are in the positive span of \mathcal{A}^+ . The inequality (8.15) arises from the inclusion $\mathbb{N}\mathcal{A}^+ \subset S_{\mathcal{A}}$ by considering points with first coordinate d . We will produce a vector $v \in \mathbb{N}\mathcal{A}^+$ and show that $v + S_{\mathcal{A}} \subset \mathbb{N}\mathcal{A}^+$, which will imply our lower bound.

Let $\mathcal{B} \subset S_{\mathcal{A}}$ be the set of points $\beta \in \mathbb{Z}^{1+n}$ which may be written as

$$\beta = \sum_{\alpha \in \mathcal{A}} \lambda_{\alpha}(1, \alpha),$$

where λ_{α} is a rational number in $[0, 1)$. For the set $\mathcal{A} = \{0, 2, 3\}$ of Example 8.3.7, \mathcal{B} consists of the origin, together with the four points in the interior of the hexagonal shaded region (a zonotope). These are the columns of the matrix $(\begin{smallmatrix} 0 & 1 & 1 & 2 & 2 \\ 0 & 1 & 2 & 3 & 4 \end{smallmatrix})$.

For each $b \in \mathcal{B}$, fix an expression

$$\beta = \sum_{\alpha \in \mathcal{A}} b_{\alpha}(\beta)(1, \alpha) \quad (b_{\alpha}(\beta) \in \mathbb{Z}) \quad (8.16)$$

as an integer linear combination of elements of \mathcal{A}^+ . Let $-\nu$ with $\nu \geq 0$ be an integer lower bound for the coefficients $b_\alpha(\beta)$ in these expressions for the finitely many elements $\beta \in \mathcal{B}$. For the set $\mathcal{A} = \{0, 2, 3\}$, we may take these expressions to be $(\frac{1}{1}) = (\frac{1}{0}) - (\frac{1}{2}) + (\frac{1}{3})$, $(\frac{1}{2}) = (\frac{1}{2})$, $(\frac{2}{3}) = (\frac{1}{0}) + (\frac{1}{3})$, and $(\frac{2}{4}) = 2(\frac{1}{2})$, so that $\nu = 1$. Finally, define

$$\textcolor{violet}{v} := \nu \cdot \sum_{\alpha \in \mathcal{A}} (1, \alpha).$$

Its first coordinate is $\nu|\mathcal{A}|$. For the set $\mathcal{A} = \{0, 2, 3\}$, this vector is $(\frac{3}{5}) = (\frac{1}{0}) + (\frac{1}{2}) + (\frac{1}{3})$.

We claim that we have the inclusion of sets of integer vectors,

$$v + S_{\mathcal{A}} \subset \mathbb{N}\mathcal{A}^+ \subset S_{\mathcal{A}}. \quad (8.17)$$

Comparing these sets at any level $d \geq \nu|\mathcal{A}|$ gives the inequality

$$E_{\text{conv}(\mathcal{A})}(d - \nu|\mathcal{A}|) \leq \text{HF}_{X_{\mathcal{A}}}(d) \leq E_{\text{conv}(\mathcal{A})}(d),$$

Since both the lower bound and the upper bound are polynomials in d of the same degree and leading term, we deduce that the Hilbert polynomial $\text{HP}_{X_{\mathcal{A}}}(d)$ has the same degree and leading term as the Ehrhart polynomial $E_{\text{conv}(\mathcal{A})}$.

Thus the Hilbert polynomial has degree n and its leading coefficient is the volume of $\text{conv}(\mathcal{A})$. Since the degree of $X_{\mathcal{A}^+}$ is $n!$ times this leading coefficient, we conclude that the degree of $X_{\mathcal{A}^+}$ is

$$n! \cdot \text{vol}(\text{conv}(\mathcal{A})),$$

which proves Kushnirenko's Theorem when $\text{Aff}_{\mathbb{Z}} \mathcal{A} = \mathbb{Z}^n$, given the inclusions (8.17).

We establish the first inclusion in (8.17). (The second is immediate.) Let $u \in v + S_{\mathcal{A}}$. Then $u - v \in S_{\mathcal{A}}$ and so it has an expression

$$u - v = \sum_{\alpha \in \mathcal{A}} a_\alpha (1, \alpha) \quad \text{with} \quad a_\alpha \in \mathbb{Q}_{\geq}.$$

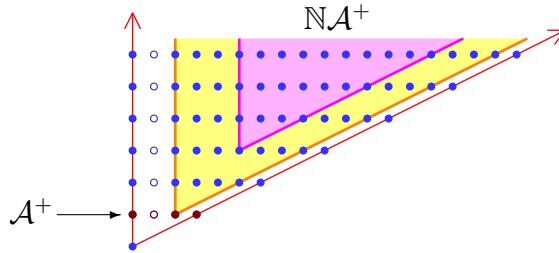
Writing each coefficient a_α in terms of its fractional and integral parts gives $a_\alpha = \lambda_\alpha + \gamma_\alpha$ where $\lambda_\alpha \in [0, 1] \cap \mathbb{Q}$ and $\gamma_\alpha \in \mathbb{N}$. Then

$$u - v = \sum_{\alpha \in \mathcal{A}} \lambda_\alpha (1, \alpha) + \sum_{\alpha \in \mathcal{A}} \gamma_\alpha (1, \alpha) = \beta + c,$$

Using the fixed expression (8.16) for β , we have

$$w = v + \sum_{\alpha \in \mathcal{A}} b_\alpha(\beta) (1, \alpha) + c = \sum_{\alpha \in \mathcal{A}} (b_\alpha(\beta) + \nu) (1, \alpha) + c,$$

which lies in $\mathbb{N}\mathcal{A}^+$ as $-\nu \leq \beta_\alpha$. This establishes the inclusion of semigroups (8.17) and completes the proof of Kushnirenko's Theorem in when $\text{Aff}_{\mathbb{Z}} \mathcal{A} = \mathbb{Z}^n$. \square

Figure 8.6: Inclusions of cones for $\mathcal{A} = \{0, 2, 3\}$.

The vector v used to establish the inclusion (8.17) may be replaced by one that is more economical. For each $\alpha \in A$, set $\nu_\alpha := \max\{0, -b_\alpha(\beta) \mid \beta \in \mathcal{B}\}$. If we set $v' := \sum_{\alpha \in A} \nu_\alpha(1, \alpha)$, then the same argument shows that we still have an inclusion $v' + S_A \subset \mathbb{N}\mathcal{A}^+$. For $\mathcal{A} = \{0, 2, 3\}$ with $\mathcal{A}^+ = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 3 \end{pmatrix}$ the new vector v' is $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$. Figure 8.6 shows the subsemigroup S_A (all the circles, filled and unfilled), the semigroup $\mathbb{N}\mathcal{A}^+$ (the filled circles), the semigroup $\begin{pmatrix} 1 \\ 2 \end{pmatrix} + S_A$ (larger shaded region), and finally the semigroup $\begin{pmatrix} 3 \\ 5 \end{pmatrix} + S_A$ (smaller shaded region). Observe that the semigroup $\begin{pmatrix} 1 \\ 2 \end{pmatrix} + S_A$ is the least shift of S_A that lies in $\mathbb{N}\mathcal{A}^+$.

The expression $n! \cdot \text{vol}(\text{conv}(\mathcal{A}))$ in Kushnirenko's Theorem is often called the *normalized volume* of $\text{conv}(\mathcal{A})$.

We deduce the general case of Kushnirenko's Theorem from the special case of $\text{Aff}_{\mathbb{Z}} \mathcal{A} = \mathbb{Z}^n$. First observe that we may assume that $0 \in \mathcal{A}$ as multiplying a polynomial f by a monomial does not change its set of zeroes in $(\mathbb{K}^\times)^n$. Then the affine span of \mathcal{A} equals its linear span. By Exercise 2 of Section 8.1, the kernel of $\varphi_{\mathcal{A}}$ is identified with $\text{Hom}_{\text{groups}}(\mathbb{Z}^n / \mathbb{Z}\mathcal{A}, \mathbb{K}^\times)$. For similar reasons Exercise?, the characters of the torus $\mathbb{T}_{\mathcal{A}} := (\mathbb{K}^\times)^n / \ker \varphi_{\mathcal{A}}$ are naturally identified with $\mathbb{Z}\mathcal{A}$.

Suppose that $\text{vol conv}(\mathcal{A}) = 0$. Then the affine span of \mathcal{A} does not have full dimension n . Since $0 \in \mathcal{A}$, we have that the rank of $\mathbb{Z}\mathcal{A}$ is less than n . In this case, $\mathbb{T}_{\mathcal{A}}$, and thus $X_{\mathcal{A}^+}$ has dimension less than n . But then a general codimension n linear subspace of $\mathbb{P}^{\mathcal{A}}$ is disjoint from $X_{\mathcal{A}^+}$ and thus by Lemma 8.3.4, there are no solutions to a general system (8.13) of polynomials with support \mathcal{A} . Some of this dimension stuff is done in the first section of this chapter

Suppose now that $\text{vol conv}(\mathcal{A}) \neq 0$ so that $\mathbb{Z}\mathcal{A}$ has rank n . Then the kernel of $\varphi_{\mathcal{A}}$ is a finite group of order $|\mathbb{Z}^n / \mathbb{Z}\mathcal{A}| = [\mathbb{Z}^n : \mathbb{Z}\mathcal{A}]$, the index of $\mathbb{Z}\mathcal{A}$ in \mathbb{Z}^n , and so the map Replacing \mathbb{Z}^n by $\mathbb{Z}\mathcal{A}$ and $(\mathbb{K}^\times)^n$ by $\mathbb{T}_{\mathcal{A}}$ in our derivation of the degree of $X_{\mathcal{A}^+}$, we obtain that $\deg X_{\mathcal{A}^+} = \text{vol}^* \text{conv}(\mathcal{A})$, where vol^* is the translation-invariant volume on \mathbb{R}^n normalized so that a fundamental parallelepiped¹ $\Pi_{\mathcal{A}}$ of $\mathbb{Z}\mathcal{A}$ has volume 1. But this fundamental parallelepiped $\Pi_{\mathcal{A}}$ has volume $\text{vol } \Pi_{\mathcal{A}} = [\mathbb{Z}^n : \mathbb{Z}\mathcal{A}]$. Thus we have

$$n! \text{vol conv}(\mathcal{A}) = [\mathbb{Z}^n : \mathbb{Z}\mathcal{A}] \cdot \deg X_{\mathcal{A}^+} = |\ker \varphi_{\mathcal{A}}| \deg X_{\mathcal{A}^+} = d(\mathcal{A}),$$

by Lemma 8.3.4 Check this reference

¹Define somewhere, make an exercise

Exercises

1. As in Remark (8.3.2), let $\Delta \subset \mathbb{R}^n$ be the convex hull of the origin and the standard unit vectors. Show that the volume of Δ is $\frac{1}{n!}$.
2. Let $\mathcal{A} \subset \mathbb{Z}^n$ be a finite set and suppose that f is a Laurent polynomial whose support is a subset of \mathcal{A} . For any $w \in \mathbb{R}^n$, show that the facial form f_w of f equals the initial form of f in the weighted partial order \succ_{-w} of Example 2.2.5 when $\text{supp}(f) \cap \mathcal{A}_w \neq \emptyset$ and is zero otherwise.
3. For a permutation $\pi \in S_n$ of the set $\{1, \dots, n\}$ let $\Delta_\pi \subset \mathbb{R}^n$ be the convex hull of $n+1$ points

$$0, e_{\pi(1)}, e_{\pi(1)} + e_{\pi(2)}, \dots, e_{\pi(1)} + \dots + e_{\pi(n)}.$$

(The last point is $(1, \dots, 1)$). Show that the union of these simplices Δ_π for $\pi \in S_n$ is the unit cube and that any two are isomorphic. Deduce that $n! \cdot \text{vol } \Delta_\pi = 1$.

4. Let f be a polynomial with support contained in a finite set $\mathcal{A} \subset \mathbb{Z}^n$ and let Λ be the linear form on $\mathbb{P}^{\mathcal{A}}$ corresponding to f in that $f = \varphi_{\mathcal{A}}^{-1}(\Lambda)$. Let $w \in \mathbb{R}^n$ and let $\mathcal{F} = \mathcal{A}_w$ be the face of \mathcal{A} exposed by w . Show that the facial form f_w of f is $\varphi_{\mathcal{F}}^{-1}(\Lambda)$.
5. A polynomial $f \in \mathbb{K}[x_1, \dots, x_n]$ is *multi affine* if it is a sum of square free monomials. Show that the support of a multi affine polynomial is a subset of the vertices $\{0, 1\}^n$ of the unit cube. Deduce that a general system of multi affine polynomials has $n!$ solutions.
6. Give some exercises to compute the volume of a polytope using Gröbner bases/elimination.

8.4 Toric Degenerations and regular subdivisions

Preamble:

Start with Gröbner degenerations. Let $X \subset \mathbb{P}^m$ be a projective variety with homogeneous ideal I . For a weight/cocharacter $w \in \mathbb{Z}^{m+1}$ we have a partial term order \succ_w and a corresponding initial ideal $\text{in}_w I$. This weight gives rise to a family $\mathcal{X}_w \subset \mathbb{C}_t \times \mathbb{P}^m$ with fibers X_t for $t \in \mathbb{K}^\times$ translates of X , and we define the limit $\lim_{t \rightarrow 0} X_t$ to be the scheme-theoretic fiber at $t = 0$. **May need to figure out a way to finesse schemes.** The main result here will be that $\mathcal{V}(\text{in}_w I)$ is this limit scheme. **For this, reread the relevant chapters in GBCP and RSEG**

The second part will be to consider this when $X = X_{\mathcal{A}}$ is a toric variety, and relate to regular subdivisions of the point set \mathcal{A} .

A possible topic will be to further explore these limits as facial subvarieties of toric varieties.

8.5 Real Toric Varieties

Preamble: Only write this after the others, if at all ??? Explain the importance of reality in applications ??

Perhaps we will have already explained about the real and positive parts of a toric variety.

Do Birch's Theorem, both cones and polytopes. If this does not get written, then do this at the end of Section 8.2. Prove a homeomorphism with the nonnegative part. Explain the implication for the topology of a toric variety, both over \mathbb{C} and over \mathbb{R} . Mention moment map. Maybe relate to expectation in algebraic statistics.

Discuss the gluing of projective real toric varieties.

Do Viro's construction with examples.

Give Sturmfels' reality theorem.

If 3-4 pages, deduce fewnomial bound.

There is much more relating the structure of the polytope $\text{conv}(\mathcal{A})$ and the toric variety. We state another such result without proof. Consider the map $\mu_{\mathcal{A}}: \mathbb{P}^{\mathcal{A}} \rightarrow \text{conv}(\mathcal{A})$ given by

$$\mathbb{P}^{\mathcal{A}} \ni z = [z_a \mid a \in \mathcal{A}] \mapsto \frac{\sum_{a \in \mathcal{A}} a|z_a|}{\sum_{a \in \mathcal{A}} |z_a|} \in \text{conv}(\mathcal{A}).$$

Lemma 8.5.1. *The map $\mu_{\mathcal{A}}: X_{\mathcal{A}^+} \rightarrow \text{conv}(\mathcal{A})$ is surjective. The inverse image of a face F of $\text{conv}(\mathcal{A})$ is $X_{\mathcal{F}}$, where $\mathcal{F} = F \cap \mathcal{A}$. The map $\mu_{\mathcal{A}}$ remains surjective when restricted to $X_{\mathcal{A}^+}(\mathbb{R}) = X_{\mathcal{A}^+} \cap \mathbb{P}^{\mathcal{A}}(\mathbb{R})$, where $\mathbb{P}^{\mathcal{A}}(\mathbb{R}) = \mathbb{P}(\mathbb{R}^{\mathcal{A}})$, and also to*

$$X_{\mathcal{A}^+}(\mathbb{R}_{\geq}) := \{x = [x_a \mid a \in \mathcal{A}] \in X_{\mathcal{A}^+} \mid x_a \geq 0 \text{ for all } a \in \mathcal{A}\},$$

where it is a homeomorphism. This map identifies $X_{\mathcal{A}^+}(\mathbb{R})$ with 2^n copies of $\text{conv}(\mathcal{A})$ glued along facets.

Exercises

8.6 Bernstein's Theorem and Polyhedral Homotopies

A theme in this book has been solving systems of polynomial equations. While this includes algorithms—symbolic in Chapter 2 and numerical in Chapter 4—it also includes qualitative aspects of this topic, including Bertini's Theorem ??. The fundamental qualitative result is Bézout's Theorem 2.4.5, which bounds the number of solutions to a system of polynomials by the product of their degrees. In Section 8.3 we presented Kushnirenko's refinement in which the polynomials all have the same support. Here we treat the general case of a system of polynomials in which every polynomial has a possibly different support. When the polynomials are general given their supports, Bernstein's Theorem asserts that the number of solutions is the mixed volume of the convex hulls of the supports.

We present two proofs of Bernstein's Theorem. The first is in the spirit of Bernstein's original proof, which uses the properties of mixed volume as developed in Section A.3.2 of the Appendix. The second is algorithmic; it is the polyhedral homotopy algorithm, which computes the solutions to systems of polynomials and is optimal for general polynomials with fixed support.

In this section we work over the complex numbers as we use analytic methods. We begin with an example.

Example 8.6.1. The system $f = g = 0$ of cubic sparse polynomials on $(\mathbb{C}^\times)^2$, where

$$f := \textcolor{blue}{x + 2y + 3xy + 5x^2y + 7y^2 + 11xy^2} \quad \text{and} \quad g := \textcolor{magenta}{1 + 3xy + 9x^2y + 27xy^2}, \quad (8.18)$$

has six solutions

$$(-0.21013, -0.44087), (0.94037, -0.13693), (-0.62796, 0.29688), (-1.1747, 0.36649), \\ (0.85566 \pm 0.55260\sqrt{-1}, -0.36620 \pm 0.25941\sqrt{-1}),$$

and not $9 = 3 \cdot 3$, which is the number predicted by Bézout's Theorem 2.1.17. Figure 8.7 shows the curves defined by f and g in \mathbb{R}^2 . The Newton polytopes for f and g are the

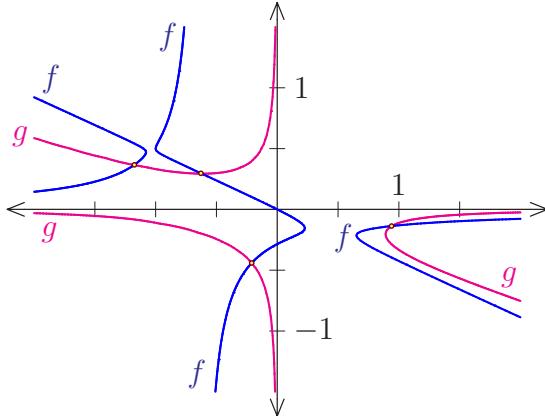


Figure 8.7: Curves of the polynomial system (8.18).

lattice polygons P and Q in Figure 8.8, respectively. Observe that the number of solutions is twice the mixed volume of P and Q . By Theorem A.3.7, this is $\text{vol}(P+Q) - \text{vol}(P) - \text{vol}(Q)$, which is the area of the four parallelograms in the decomposition of $P+Q$ in Figure 8.8. Exercise 1 asks you to compute the number of solutions for different pairs of polynomials with the same support as f and g (8.18). \diamond

Bernstein's Theorem generalizes this observation. As in Section 8.3, for a finite set $\mathcal{A} \subset \mathbb{Z}^n$, we identify the set of polynomials whose support is a subset of \mathcal{A} with the vector space $\mathbb{C}^{\mathcal{A}}$ of the coefficients of such polynomials. Thus given finite subsets $\mathcal{A}_1, \dots, \mathcal{A}_n \subset \mathbb{Z}^n$, we identify $\mathbb{C}^{\mathcal{A}_1} \oplus \dots \oplus \mathbb{C}^{\mathcal{A}_n}$ with the set of systems of polynomials with support

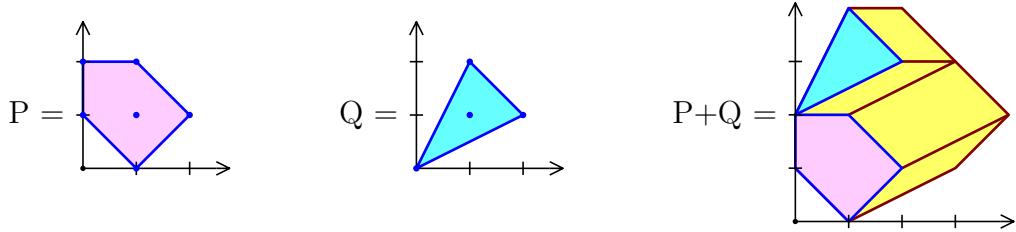


Figure 8.8: Minkowski sum of two polygons.

$(\mathcal{A}_1, \dots, \mathcal{A}_n)$. Each point of $\mathbb{C}^{\mathcal{A}_1} \oplus \dots \oplus \mathbb{C}^{\mathcal{A}_n}$ is a list of polynomials (f_1, \dots, f_n) , which then corresponds to a system

$$f_1(x) = f_2(x) = \dots = f_n(x) = 0, \quad (8.19)$$

of polynomial equations where $\text{supp}(f_i) = \mathcal{A}_i$.

Theorem 8.6.2 (Bernstein). *The number of isolated solutions in $(\mathbb{C}^\times)^n$, counted with multiplicity, of a system*

$$f_1(x) = f_2(x) = \dots = f_n(x) = 0$$

of n polynomials is at most $n! \text{MV}(P_1, \dots, P_n)$, where P_i is the Newton polytope of f_i . There is a dense open subset of $\mathbb{C}^{\mathcal{A}_1} \oplus \dots \oplus \mathbb{C}^{\mathcal{A}_n}$ consisting of systems with support $(\mathcal{A}_1, \dots, \mathcal{A}_n)$ having exactly $n! \text{MV}(P_1, \dots, P_n)$ solutions in $(\mathbb{C}^\times)^n$, each isolated and occurring with multiplicity one.

Example 8.6.3. Suppose that for each $i = 1, \dots, n$, we have that f_i is a polynomial of total degree d_i , so that its Newton polytope is $d_i \Delta$, where Δ is the convex hull of the origin and the standard unit vectors. By Bernstein's Theorem 8.6.2, the number of solutions to the system (8.19) is $n! \text{MV}(d_1 \Delta, \dots, d_n \Delta)$. Since $\Delta + \dots + \Delta$ (d times) equals $d \Delta$, the multilinearity of mixed volume (Lemma A.3.5) implies that

$$n! \text{MV}(d_1 \Delta, \dots, d_n \Delta) = d_1 \cdots d_n n! \text{MV}(\Delta, \dots, \Delta) = d_1 \cdots d_n,$$

as $n! \text{MV}(\Delta, \dots, \Delta) = 1$. This last equality follows also from Bernstein's Theorem; a polynomial with support a subset of Δ is linear and a general system of linear polynomials has a single solution.

Thus we deduce the general case of Bézout's Theorem 2.4.5 from Bernstein's Theorem.

◊

Given the properties of mixed volume as developed in Section A.3.2, particularly its characterization in Theorem A.3.7, we prove Bernstein's Theorem 8.6.2 by first showing that the number of solutions to a generic system with given support depends only on the convex hulls of the support, and then that this number is symmetric in its arguments, that it is normalized, and that it is multilinear under Minkowski sum (as in Theorem A.3.7).

Then by Theorem A.3.7 this number (suitably normalized) equals the mixed volume of the convex hulls of the supports. We first prove a lemma about this number for a generic system.

Lemma 8.6.4. *Let $\mathcal{A}_1, \dots, \mathcal{A}_n$ be finite subsets of \mathbb{Z}^n . Then there a nonempty open subset U of $\mathbb{C}^{\mathcal{A}_1} \oplus \dots \oplus \mathbb{C}^{\mathcal{A}_n}$ consisting of polynomial systems having only simple solutions. Furthermore, a nonnegative integer d and there is a nonempty open subset $V \subset U$ consisting of polynomial systems having d solutions.*

When $d = 0$, no system has any isolated solutions.

Write $d(\mathcal{A}_1, \dots, \mathcal{A}_n)$ for the number d of solutions from the lemma.

Proof. Consider the incidence variety of solutions to systems of polynomials with support $(\mathcal{A}_1, \dots, \mathcal{A}_n)$,

$$\Gamma := \{(x, f_1, \dots, f_n) \in (\mathbb{C}^\times)^n \oplus \mathbb{C}^{\mathcal{A}_1} \oplus \dots \oplus \mathbb{C}^{\mathcal{A}_n} \mid f_1(x) = \dots = f_n(x) = 0\}.$$

For $x \in (\mathbb{C}^\times)^n$, the set $\{f_i \in \mathbb{C}^{\mathcal{A}_i} \mid f_i(x) = 0\}$ is a hyperplane in $\mathbb{C}^{\mathcal{A}_i}$, as $f_i(x) = 0$ is a linear equation on the space $\mathbb{C}^{\mathcal{A}_i}$ of coefficients of f_i . Thus the map $\Gamma \rightarrow (\mathbb{C}^\times)^n$ has fibers that are linear spaces of dimension $\sum_{i=1}^n |\mathcal{A}_i| - n$, and so Γ is irreducible of dimension $\sum_{i=1}^n |\mathcal{A}_i|$, by the Dimension Theorem ???.

The projection π of Γ to the other factor $\mathbb{C}^{\mathcal{A}_1} \oplus \dots \oplus \mathbb{C}^{\mathcal{A}_n}$ has fiber over a point $F = (f_1, \dots, f_n)$ equal to the set of solutions $\mathcal{V}(F)$. Suppose that this projection is dominant, then by Baby Bertini there is a nonempty Zariski open subset $U \subset \pi(\Gamma)$ consisting of systems so that the fiber is smooth and of dimension zero. Consequently if $F \in U$ is such a system, $\mathcal{V}(F)$ is a finite set of smooth points so that each solution to the system is simple. Furthermore, there is a positive integer d and a nonempty open subset $V \subset U$ such that every system has exactly d preimages, and systems in $U \setminus V$ have at most d solutions. Systems in V are sparse systems with exactly d simple solutions in $(\mathbb{C}^\times)^n$.

If the projection is not dominant, then the complement of its image contains an open subset U , and we set $V = U$. Polynomial systems $F \in U$ have no solutions, $\mathcal{V}(F) = \emptyset$, and so $d = 0$. This completes the proof of the first statement. Since the image of Γ has dimension less than that of Γ , every component of every fiber has positive dimension, proving the second statement. \square

Observe that the function $d(\mathcal{A}_1, \dots, \mathcal{A}_n)$ is symmetric in its arguments. An *unmixed system* is one in which each polynomial f_i has the same support, \mathcal{A} . For unmixed systems, Bernstein's Theorem reduces to Kushnirenko's Theorem 8.3.1. Thus $d(\mathcal{A}, \dots, \mathcal{A}) = n! \operatorname{vol}_n(\operatorname{conv}(\mathcal{A}))$, which is the normalization condition. To prove Bernstein's Theorem, we show that $d(\mathcal{A}_1, \dots, \mathcal{A}_n)$ depends only on the convex hulls of the \mathcal{A}_i and that it is multilinear under Minkowski sum, as in Theorem A.3.7.

To understand multilinearity, consider the variety $\mathcal{V}(f_1, \dots, f_n) \subset (\mathbb{C}^\times)^n$ consisting of solutions to a system (8.13) of polynomials. Let us write $d(f_1, \dots, f_n)$ for the number of isolated points in this variety. It is clear that

$$d(f \cdot g, f_2, \dots, f_n) \leq d(f, f_2, \dots, f_n) + d(g, f_2, \dots, f_n). \quad (8.20)$$

This is an equality when the systems have disjoint varieties, when $\mathcal{V}(f, g, f_2, \dots, f_n) = \emptyset$. We will show that this can be achieved. Note that multilinearity of $d(\mathcal{A}_1, \dots, \mathcal{A}_n)$ would follow from equality in (8.20), if we could show that

$$d(f, f_2, \dots, f_n) = d(\mathcal{A}, \mathcal{A}_2, \dots, \mathcal{A}_n) \quad \text{and} \quad d(g, f_2, \dots, f_n) = d(\mathcal{B}, \mathcal{A}_2, \dots, \mathcal{A}_n),$$

where f has support \mathcal{A} , g has support \mathcal{B} , and f_i has support \mathcal{A}_i together imply that

$$d(f \cdot g, f_2, \dots, f_n) = d(\mathcal{A} + \mathcal{B}, \mathcal{A}_2, \dots, \mathcal{A}_n).$$

In Exercise 4, you show that $\text{supp}(f \cdot g) \subset \mathcal{A} + \mathcal{B}$ and they have the same convex hulls. The problem with this implication is that $(f \cdot g, f_2, \dots, f_n)$ is not a general system with support $(\mathcal{A} + \mathcal{B}, \mathcal{A}_2, \dots, \mathcal{A}_n)$, because the first polynomial factors. Our next theorem allows us to overcome this problem by characterizes the discriminant condition when all points of $\mathcal{V}(f_1, \dots, f_n)$ are simple and $d(f_1, \dots, f_n) < d(\mathcal{A}_1, \dots, \mathcal{A}_n)$, where (f_1, \dots, f_n) has support $(\mathcal{A}_1, \dots, \mathcal{A}_n)$.

Let $w \in \mathbb{Z}^n = N$ be a cocharacter. Given a Laurent monomial x^α and $t \in \mathbb{C}^\times$, we have $(t^w \cdot x)^\alpha = t^{w \cdot \alpha} x^\alpha$. Given a Laurent polynomial $f = \sum_\alpha c_\alpha x^\alpha$ with support \mathcal{A} , we have

$$f(t^w \cdot x) = \sum_{\alpha \in \mathcal{A}} t^{w \cdot \alpha} c_\alpha x^\alpha = t^{h_{\mathcal{A}}(w)} \cdot \sum_{\alpha \in \mathcal{A}} t^{w \cdot \alpha - h_{\mathcal{A}}(w)} c_\alpha x^\alpha.$$

Note that the exponents of t in the sum are all nonnegative and are 0 for $\alpha \in \mathcal{A}_w$, the face of \mathcal{A} exposed by w . The sum of those terms of f in which t does not appear is its facial form f_w . Then

$$t^{-h_{\mathcal{A}}(w)} f(t^w \cdot x) = f_w(x) + \sum_{\alpha \in \mathcal{A} \setminus \mathcal{A}_w} t^{w \cdot \alpha - h_{\mathcal{A}}(w)} c_\alpha x^\alpha, \quad (8.21)$$

and all terms in the sum are divisible by a positive power of t .

Given a system $F = (f_1, \dots, f_n) \in \mathbb{C}^{\mathcal{A}_1} \oplus \dots \oplus \mathbb{C}^{\mathcal{A}_n}$ of polynomials and $w \in \mathbb{Q}^n$, we have the facial system $F_w := (f_{1,w}, \dots, f_{n,w})$. Note that the facial system depends upon the sets $(\mathcal{A}_1, \dots, \mathcal{A}_n)$. Scaling w by a positive integer, we may assume that $w \in \mathbb{Z}^n$ and obtain the same facial system. Note that by (8.21), $f_w(t^w \cdot x) = t^{h_{\mathcal{A}}(w)} f_w(x)$. Thus if $a \in (\mathbb{C}^\times)^n$ is a solution of the facial system F_w , then so is $t^w \cdot a$ for all $t \in \mathbb{C}^\times$. Since the solutions of the facial system are not isolated, Lemma 8.6.4 implies that there is a nonempty Zariski open subset of facial systems with no solutions.

Theorem 8.6.5. *Let $\mathcal{A}_1, \dots, \mathcal{A}_n$ be finite subsets of \mathbb{Z}^n , and let $U \subset \mathbb{C}^{\mathcal{A}_1} \oplus \dots \oplus \mathbb{C}^{\mathcal{A}_n}$ be a non-empty Zariski open set consisting of systems whose solutions are simple. For $F \in U$, we have $dF = d(\mathcal{A}_1, \dots, \mathcal{A}_n)$ if and only if $\mathcal{V}(F_w) = \emptyset$ for all $w \in \mathbb{Z}^n \setminus \{0\}$.*

Theorem 8.6.5 will be used in the proof of Bernstein's Theorem to imply multilinearity. We also use it to show that the number $d(\mathcal{A}_1, \dots, \mathcal{A}_n)$ of solutions depends only upon the convex hulls of the \mathcal{A}_i . To that end, for a finite subset $\mathcal{A} \subset \mathbb{Z}^n$, let $\overline{\mathcal{A}} := \text{conv}(\mathcal{A}) \cap \mathbb{Z}^n$ be the set of all integer points in the convex hull of \mathcal{A} .

Corollary 8.6.6. *For any collection of finite subsets $\mathcal{A}_1, \dots, \mathcal{A}_n$ of \mathbb{Z}^n , we have*

$$d(\mathcal{A}_1, \dots, \mathcal{A}_n) = d(\overline{\mathcal{A}_1}, \dots, \overline{\mathcal{A}_n}).$$

We prove Theorem 8.6.5 and Corollary 8.6.6 after deducing Bernstein's Theorem.

Proof of Bernstein's Theorem. Let f_1, \dots, f_n be Laurent polynomials in n variables. For each $i = 1, \dots, n$ let $\mathcal{A}_i \subset \mathbb{Z}^n$ be the support of f_i . Note first that $d(f_1, \dots, f_n) \leq d(\mathcal{A}_1, \dots, \mathcal{A}_n)$, as $d(\mathcal{A}_1, \dots, \mathcal{A}_n)$ is the maximum number of isolated solutions in $(\mathbb{C}^\times)^n$ to a system with the given supports. **Need to prove this statement.**

By Corollary 8.6.6, the number $d(\mathcal{A}_1, \dots, \mathcal{A}_n)$ is a function of the convex hulls $\text{conv}(\mathcal{A}_i)$ for $i = 1, \dots, n$. We earlier noted that this function is symmetric in its arguments, and that Kushnirenko's Theorem 8.3.1 gives the normalization, $d(\mathcal{A}, \dots, \mathcal{A}) = n! \text{vol}_n \text{conv}(\mathcal{A})$. By Theorem A.3.7, all that remains to show is that

$$d(\mathcal{A}, \mathcal{A}_2, \dots, \mathcal{A}_n) + d(\mathcal{B}, \mathcal{A}_2, \dots, \mathcal{A}_n) = d(\mathcal{A} + \mathcal{B}, \mathcal{A}_2, \dots, \mathcal{A}_n), \quad (8.22)$$

for any finite subsets $\mathcal{A}, \mathcal{B}, \mathcal{A}_2, \dots, \mathcal{A}_n$ of \mathbb{Z}^n . Since the subset of systems with support $(\mathcal{A}, \mathcal{A}_2, \dots, \mathcal{A}_n)$ having $d(\mathcal{A}, \mathcal{A}_2, \dots, \mathcal{A}_n)$ isolated simple solutions in $(\mathbb{C}^\times)^n$ contains an open dense set, and the same for those with support $(\mathcal{B}, \mathcal{A}_2, \dots, \mathcal{A}_n)$, there are Laurent polynomials f, g, f_2, \dots, f_n having respective supports $\mathcal{A}, \mathcal{B}, \mathcal{A}_2, \dots, \mathcal{A}_n$ such that

$$d(f, f_2, \dots, f_n) = d(\mathcal{A}, \mathcal{A}_2, \dots, \mathcal{A}_n) \quad \text{and} \quad d(g, f_2, \dots, f_n) = d(\mathcal{B}, \mathcal{A}_2, \dots, \mathcal{A}_n) \quad (8.23)$$

Suppose that $\mathcal{V}(f, g, f_2, \dots, f_n) = \emptyset$. Then we have

$$d(f \cdot g, f_2, \dots, f_n) = d(f, f_2, \dots, f_n) + d(g, f_2, \dots, f_n), \quad (8.24)$$

as the two systems have disjoint solution sets. Furthermore, all solutions to the system $(f \cdot g, f_2, \dots, f_n)$ are simple. By Theorem 8.6.5, no facial system of either (f, f_2, \dots, f_n) or (g, f_2, \dots, f_n) has solutions, as we have the equalities in (8.23). But then no facial system of $(f \cdot g, f_2, \dots, f_n)$ has any solutions. By Theorem 8.6.5 we then have

$$d(f \cdot g, f_2, \dots, f_n) = d(\mathcal{A} + \mathcal{B}, \mathcal{A}_2, \dots, \mathcal{A}_n),$$

which implies the multilinearity (8.22), and thus the conclusion of Bernstein's Theorem.

Suppose now that $\mathcal{V}(f, g, f_2, \dots, f_n) \neq \emptyset$. Then g vanishes at some point of $S := \mathcal{V}(f, f_2, \dots, f_n)$. We replace g by another polynomial with support \mathcal{B} that does not vanish at any point of S and with the property that the new system has the same number of solutions. This will imply the sum (8.24) and therefore Bernstein's Theorem.

Let $w \in \mathbb{Z}^n$ be a cocharacter that exposes a single point of \mathcal{B} (e.g. a vertex of its convex hull) and for $t \in \mathbb{C}^\times$ set $g_t(x) := t^{-h_{\mathcal{B}}(w)} g(t^w x)$. As the facial form g_w is a single term, the computation (8.21) shows that for any $x \in (\mathbb{C}^\times)^n$, g_t is a polynomial in t with a nonzero constant term, and thus it vanishes at only finitely many $t \in \mathbb{C}^\times$. Since S is a finite set, there are only finitely many $t \in \mathbb{C}^\times$ for which g_t vanishes at some point of S . Since $g = g_1$ and the discriminant conditions of Theorem 8.6.5 (only simple solutions, no facial system has a solution) hold on Zariski-open sets, they hold for (g_t, f_2, \dots, f_n) for all except finitely many t . Thus there is some $t \in \mathbb{C}^\times$ for which $d(g_t, f_2, \dots, f_n) = d(\mathcal{B}, \mathcal{A}_2, \dots, \mathcal{A}_n)$ and g_t does not vanish at any point of S . This completes our proof of Bernstein's Theorem. \square

Proof of Corollary 8.6.6. Let $(f_1, \dots, f_n) \in \mathbb{C}^{\mathcal{A}_1} \oplus \dots \oplus \mathbb{C}^{\mathcal{A}_n}$ be a system of polynomials having only simple solutions with $d(f_1, \dots, f_n) = d(\mathcal{A}_1, \dots, \mathcal{A}_n)$. Then no facial system of (f_1, \dots, f_n) has any solutions, by Theorem 8.6.5.

As $\mathcal{A}_i \subset \overline{\mathcal{A}_i}$, the polynomial $f_i \in \mathbb{C}^{\overline{\mathcal{A}_i}}$. Also, as both \mathcal{A}_i and $\overline{\mathcal{A}_i}$ have the same convex hull, the same is true for their faces exposed by any $w \in \mathbb{Z}^n$. This implies that the facial systems of (f_1, \dots, f_n) for the two different supports $(\mathcal{A}_1, \dots, \mathcal{A}_n)$ and $(\overline{\mathcal{A}_1}, \dots, \overline{\mathcal{A}_n})$ coincide. Since no facial system of (f_1, \dots, f_n) has any solutions, Theorem 8.6.5 for the support $(\overline{\mathcal{A}_1}, \dots, \overline{\mathcal{A}_n})$, implies that $d(f_1, \dots, f_n) = d(\overline{\mathcal{A}_1}, \dots, \overline{\mathcal{A}_n})$, which completes the proof. \square

Proof of Theorem 8.6.5. Let $F = (f_1, \dots, f_n)$ and $G = (g_1, \dots, g_n)$ be systems with support $(\mathcal{A}_1, \dots, \mathcal{A}_n)$ having only simple solutions, where G has the maximum number $d := d(\mathcal{A}_1, \dots, \mathcal{A}_n)$ solutions. Form the straight-line homotopy (4.17)

$$H(x; t) := t \cdot G + (1 - t) \cdot F$$

between these two systems. As in Section 4.2, this defines a curve C in $(\mathbb{C}^\times)_x^n \times \mathbb{C}_t$ with an open set $U \subset \mathbb{C}_t$ (containing 0 and 1) of points t where $H(x; t)$ has only simple solutions, and a further subset $V \subset U$ of points where $H(x; t)$ has d solutions. By our choice of F , only the cases (1) and (2) of Figure 4.6 may occur near $t = 0$. In particular $\mathcal{V}(F)$ consists of d points unless C is unbounded near $t = 0$.

To investigate this, we consider $H(x; t)$ as a system of polynomials in x with coefficients in the field of rational functions $\mathbb{C}(t)$. Then it has solutions in the algebraic torus $(\overline{\mathbb{C}(t)})^\times$ over the algebraic closure $\overline{\mathbb{C}(t)}$ of $\mathbb{C}(t)$. In fact, it, has d solutions. To see this take a point $\tau \in V$, and then by the implicit function theorem, there is a disc D containing τ and for each point $p \in C$ above τ an analytic map $z_p: D \rightarrow C$ with $z_p(\tau) = p$. Then for points $t \in D$, $H(z_p(t), t) = 0$ so that these d functions z_p are solutions.

Embedding $\mathbb{C}(t)$ into the field of formal Laurent series, its algebraic closure $\overline{\mathbb{C}(t)}$ is a subfield of the field $\mathbb{C}\{\{t\}\}$ of Puiseaux series, by the Newton-Puiseaux Theorem 9.4.5. Thus a nonzero point $z(t) \in \overline{\mathbb{C}(t)}^\times$ is represented by a Puiseaux series. There is a positive integer k , an integer N , and complex numbers c_j for $j \geq N$ with $c_N \neq 0$ such that

$$z(t) = \sum_{j \geq N} c_j t^{j/k} = c_N t^{N/k} + \sum_{j > N} c_j t^{j/k}. \quad (8.25)$$

The *initial coefficient* of $z(t)$ is c_N and its *order* is N/k . We will write a Puiseaux series as $z(t) = c_N t^{N/k} + \text{h.o.t.}$, where **h.o.t.** refers to “terms of higher order in t ”.

A point of the torus over $\mathbb{C}\{\{t\}\}$ is a vector $z(t) = (z_1(t), \dots, z_n(t))$ of nonzero Puiseaux series (8.25). Let $w := (w_1, \dots, w_n) \in \mathbb{Q}^n$ be the vector of orders of the $z_i(t)$ and $c := (c_1, \dots, c_n) \in \mathbb{C}^\times$ be the vector of initial coefficients of the $z_i(t)$. Then we have

$$z(t) = (z_1(t), \dots, z_n(t)) = t^w \cdot c + \text{h.o.t.}$$

A calculation similar to that leading up to (8.21) shows that for a Laurent polynomial $f \in \mathbb{C}[x^\pm]$ with support \mathcal{A} ,

$$f(z(t)) = t^{h_{\mathcal{A}}(w)} (f_w(c) + \text{h.o.t.}),$$

where the higher-order terms involve positive powers of t . We expand $H(z(t); t)$ to get

$$H(z(t); t) = t^{h_{\mathcal{A}_\bullet}(w)} F_w(c) + \text{h.o.t.},$$

where $h_{\mathcal{A}_\bullet}(w) = (h_{\mathcal{A}_1}(w), \dots, h_{\mathcal{A}_n}(w)) \in \mathbb{Q}^n$. For a solution $z(t)$ to $H(x; t) = 0$ every term vanishes. In particular, $F_w(c) = 0$, so that c is a solution to the facial system F_w .

The curve C is unbounded in $(\mathbb{C}^\times)^n$ near $t = 0$ if and only if there is a solution $z(t)$ to $H(x; t) = 0$ with order $w \neq 0$. Thus $d(F) < d$ implies that there is a solution with a nonzero order, and thus a facial system F_w having a solution.

To prove the converse, given a solution $c \in (\mathbb{C}^\times)^n$ to a facial system F_w with $w \neq 0$, we construct a homotopy $H(x, t)$ between F and a system G having d solutions, where $H(x; t) = 0$ has a solution $z(t) \in \mathbb{C}\{\{t\}\}$ of order w . As the subset of $\mathbb{C}^{\mathcal{A}_1} \oplus \dots \oplus \mathbb{C}^{\mathcal{A}_n}$ consisting of systems with d simple solutions is open and dense, there is a system G in this set and a point $b \in (\mathbb{C}^\times)^n$ such that $G(b) = 0$ but both $F(b)$ and $G_w(c)$ have no nonzero components. We may assume that $w \in \mathbb{Z}^n$. Define $z(t) := t^w \cdot (c + t(b - c))$ and set

$$H(x; t) := t^{-h_{\mathcal{A}_\bullet}(w)} (G(z(t)) \cdot F(x) - F(z(t)) \cdot G(x)).$$

Observe that $z(t)$ is a solution to $H(x; t)$ of order w as an element of $\mathbb{C}\{\{t\}\}$. What remains is to show that $H(x; t)$ is a homotopy between F and G .

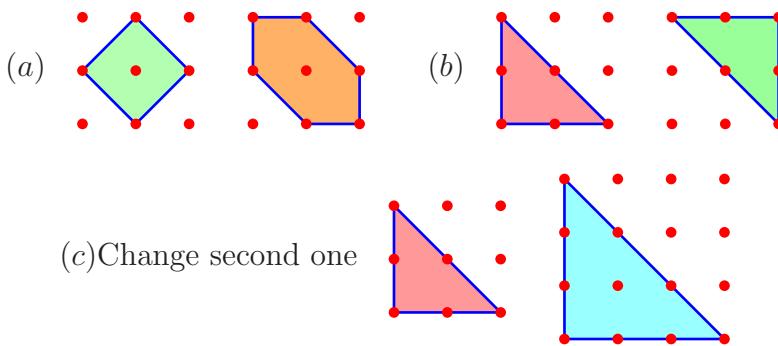
When $t = 1$, note that $z(1) = b$ and so $H(x; 1) = F(b) \cdot G(x)$, which is essentially the system G as no component of $F(b)$ vanishes. We only need to investigate when $t = 0$. If we expand H in powers of t , we obtain

$$H(x; t) = (G_w(c) + \text{h.o.t.}) \cdot F(x) - (F_w(c) + \text{h.o.t.}) \cdot G(x),$$

Since no component of $G_w(c)$ vanishes and $F_w(c) = 0$, we see that $H(x; t)$ reduces to the system F at $t = 0$. This completes the proof. \square

Exercises

1. Generate other pairs of polynomials with the same support as the polynomials in (8.18). For each pair, compute the degree of the ideal they generate. Can you prove this degree is six for generic coefficients?
2. Have the reader prove the number of solutions to some mixed volume problem in the plane using resultants.
3. Show that for any sets $\mathcal{A}, \mathcal{B} \subset \mathbb{R}^n$, we have $\text{conv}(\mathcal{A}) + \text{conv}(\mathcal{B}) = \text{conv}(\mathcal{A} + \mathcal{B})$.
4. Let f and g be Laurent polynomials. Prove that $\text{New}(f \cdot g) = \text{New}(f) + \text{New}(g)$. Note that if f and g have support \mathcal{A} and \mathcal{B} , respectively, then we only have $\text{supp}(f \cdot g) \subset \mathcal{A} + \mathcal{B}$, as there may be cancellation. (Suppose that $f = 1+x$ and $g = 1-x$.) However, there is no cancellation in the extreme points of \mathcal{A} and \mathcal{B} , and this equality can be shown by considering the support functions.
5. Determine the Newton polytope of each polynomial, and the mixed volume of the Newton polytopes of each polynomial system. Check the conclusion of Bernstein's Theorem using a computer algebra system such as Macaulay2 or Singular.
 - (a) $1 + 2x + 3y + 4xy = 1 - 2xy + 3x^2y - 5xy^2 = 0$.
 - (b) $1 + 2x + 3y - 5xy + 7x^2y^2 = 0$
 $1 - 2xy + 4x^2y + 8xy^2 - 16x^3y + 32xy^3 - 64x^2y^2 = 0$.
 - (c) $2 + 5xy - x^2y - 6xy^2 + 4xy^3 = 0$
 $2x - y - 2y^2 - xy^2 + 2x^2y + x^2 - 5xy = 0$.
 - (d) $1 + x + y + z + xy + xz + yz + xyz = 0$
 $xy + 2xyz + 3xyz^2 + 5xz + 7xy^2z + 11yz + 13x^2yz = 0$.
 $4 - x^2y + 2x^2z - xz^2 + 2yz^2 - y^2z + 2y^2x - 8xyz = 0$
6. Compute the mixed volume of the following pairs of lattice polygons.



8.7 Notes

In practice, Buchberger's algorithm is not the best way to compute a Gröbner basis for a toric ideal. We remark that there are other, often superior algorithms available, for example the project-and-lift algorithm of Hemmecke and Malkin [55] which is implemented in the software **4ti2**.

[144]

Chapter 9

Tropical geometry

Tropical geometry denotes the 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 are replaced by the real space \mathbb{R}^n and complex algebraic varieties by polyhedral cell complexes. Thus, enhancing the discussion from Chapters 8, tropical geometry provides another important link between algebraic geometry and combinatorics.

The origins of the tropical degeneration ideas go back to Viro’s patchworking method (in the 1970’s), to the Bergman complex (in the 1970’s), to Maslov’s dequantization of positive real numbers (in the 1980’s) and to the concept of an amoeba introduced by I. Gelfand, M. Kapranov and A. Zelevinsky (in the early 90’s) as the logarithmic image of a complex variety.

In the chapter, we discuss some of the central ideas and concepts from tropical geometry. To fix notation, let $(\mathbb{R}, \oplus, \odot)$ denote the *tropical semiring*, where we usually choose to work with the min-plus version, that is,

$$x \oplus y = \min\{x, y\} \quad \text{and} \quad x \odot y = x + y.$$

Sometimes the underlying set \mathbb{R} of real numbers is augmented by ∞ .

9.1 Tropical hypersurfaces

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

$$\tilde{f} : \mathbb{R}^n \rightarrow \mathbb{R}, \quad (x_1, \dots, x_n) \mapsto \alpha_1 x_1 + \cdots + \alpha_n x_n + c.$$

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,

$$\tilde{f} : \mathbb{R}^n \rightarrow \mathbb{R}, \quad (x_1, \dots, x_n) \mapsto \min_{\alpha \in A} \left\{ c_\alpha + \sum_{i=1}^n \alpha_i x_i \right\}.$$

The support of f is defined by $\text{supp } f = \{\alpha \in A : \alpha \neq \infty\}$. And let $\mathbb{R}_{\text{trop}}[x_1, \dots, x_n]$ denote the semiring of tropical polynomials.

Theorem 9.1.1. *Every tropical polynomial $f \in \mathbb{R}_{\text{trop}}[x_1, \dots, x_n]$ defines a continuous, concave and piecewise-linear function $\tilde{f} : \mathbb{R}^n \rightarrow \mathbb{R}$.*

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 $\beta, \gamma, \delta \in \text{supp } f$ with $f(b) = f_\beta(b)$ and $f(c) = f_\gamma(c)$, and $f(b + (1 - \lambda)c) = f_\delta(b + (1 - \lambda)c)$. Hence, $f_\beta(b) \leq f_\delta(b)$, $f_\gamma(c) \leq f_\delta(c)$ and therefore

$$\begin{aligned} \lambda f(b) + (1 - \lambda)f(c) &= \lambda f_\beta(b) + (1 - \lambda)f_\gamma(c) \\ &\leq \lambda f_\delta(b) + (1 - \lambda)f_\delta(c) = f_\delta(\lambda b + (1 - \lambda)c) \\ &= f(\lambda b + (1 - \lambda)c). \end{aligned}$$

□

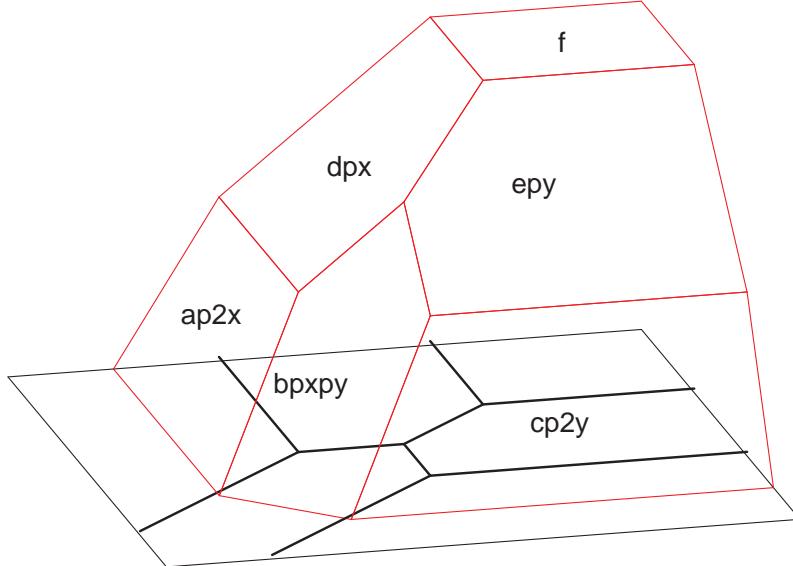


Figure 9.1: The graph of $\tilde{f} : \mathbb{R}^2 \rightarrow \mathbb{R}$ for a tropical polynomial $f = a \odot x^2 \oplus b \odot x \odot y \oplus c \odot y^2 \oplus d \odot x \oplus e \odot y \oplus f$.

At each given point $x \in \mathbb{R}^n$ the minimum of the piecewise linear function $\tilde{f} : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 the 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 the *corner locus* of the piecewise linear function \tilde{f} , i.e., the set of points where \tilde{f} is not differentiable. Considering f as a concave, piecewise linear function, Figure 9.1 visualizes the graph of \tilde{f} and the resulting curve $\mathcal{T}(f) \subset \mathbb{R}^2$ for a quadratic tropical polynomial.

In the following, 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.3.1 on polyhedral geometry.

Let $\mathcal{A} \subset \mathbb{N}_0^n$ be finite and $f(x_1, \dots, x_n) = \bigoplus_{\alpha \in \mathcal{A}} c_\alpha \cdot x^\alpha$ be a tropical polynomial with $c_\alpha \in \mathbb{R}$ for all $\alpha \in \mathcal{A}$. The *extended Newton polytope* of f is the convex hull

$$\text{New}^e(f) = \text{conv}\{(\alpha, c_\alpha) : \alpha \in \mathcal{A}\} \subset \mathbb{R}^{n+1}.$$

Definition 9.1.2. Let f be a tropical polynomial in x_1, \dots, x_n and $\mathcal{N}(\text{New}^e(f))$ be the inner normal fan of its extended Newton polytope. The set

$$\mathcal{L}_f = \{F \text{ face of } \text{New}^e(f) : \text{there exists some } w \in C(F) \text{ with } w_{n+1} > 0\}$$

is called the *lower complex* of f , where $C(F)$ denotes the inner 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 $\mathcal{N}_k(\text{New}^e(f))$ denote the set of faces of the normal fan $\mathcal{N}(\text{New}^e(f))$ of dimension at most k .

Theorem 9.1.3. *Let f be a tropical polynomial in x_1, \dots, x_n . Then*

$$\mathcal{T}(f) = \mathcal{N}_n(\text{New}^e(f)) \cap \{x \in \mathbb{R}^{n+1} : x_{n+1} = 1\}. \quad (9.1)$$

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 $(w_1, \dots, w_n, 1)$ attains its minimum at more than one of the points $(\alpha_1, \dots, \alpha_n, c)$ representing a term of f . Hence, a point w is in $\mathcal{T}(f)$ if and only if the face of $\text{New}^e(f)$ minimizing the linear function $x \mapsto (w_1, \dots, w_n, 1)^T \cdot x$ has dimension at least 1. Hence, $\mathcal{T}(f)$ coincides with the right hand set in (9.1).

The fact that $\mathcal{T}(f)$ is a polyhedral complex then follows from the first statement. The dimension statement follows from the observation that none of the maximal cells of $\mathcal{N}(\text{New}^e(f))$ is contained in the hyperplane $\{x \in \mathbb{R}^{n+1} : x_{n+1} = 1\}$. \square

The Newton polytope $\text{New}(f)$ of a tropical polynomial f comes with a *privileged subdivision* $\text{subdiv}(f)$. This privileged subdivision is defined by projecting down the faces of the lower complex \mathcal{L}_f by forgetting the last coordinate. 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 $\text{subdiv}(f)$. Explicitly, the dual face F^\vee of a face F of $\mathcal{T}(f)$ is the maximal face G of $\text{subdiv}(f)$ such that the set $F \times \{1\}$ is contained in the inner normal cone of the lifted face $\hat{G} \subset \text{New}^e(f)$.

Theorem 9.1.4. *The tropical hypersurface $\mathcal{T}(f)$ and the subdivision $\text{subdiv}(f)$ 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 $\mathcal{N}_n(\text{New}^e(f))$. 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 $\text{New}^e(f)$. 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 . \square

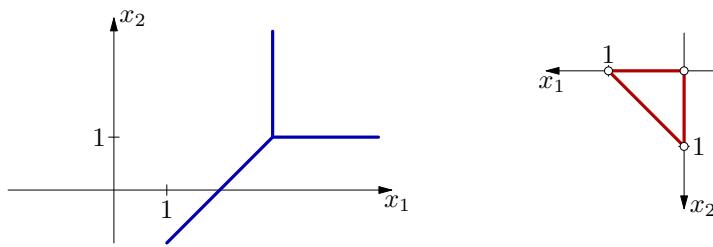


Figure 9.2: The tropical curve of a linear polynomial f in two variables and the Newton polygon of f .

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 see Figure 9.3 for an example of a tropical cubic curve, as well as their dual subdivisions. Note that the coordinate axes are directed to the left and to the bottom to visualize the duality.

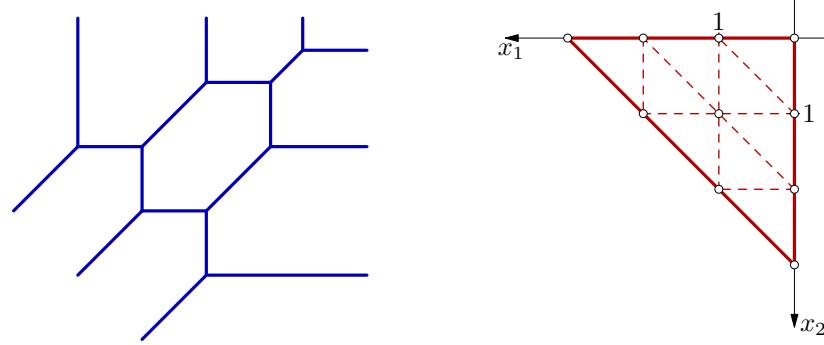


Figure 9.3: An example of a tropical cubic curve $\mathcal{T}(f)$ and the dual subdivision of the Newton polygon of f .

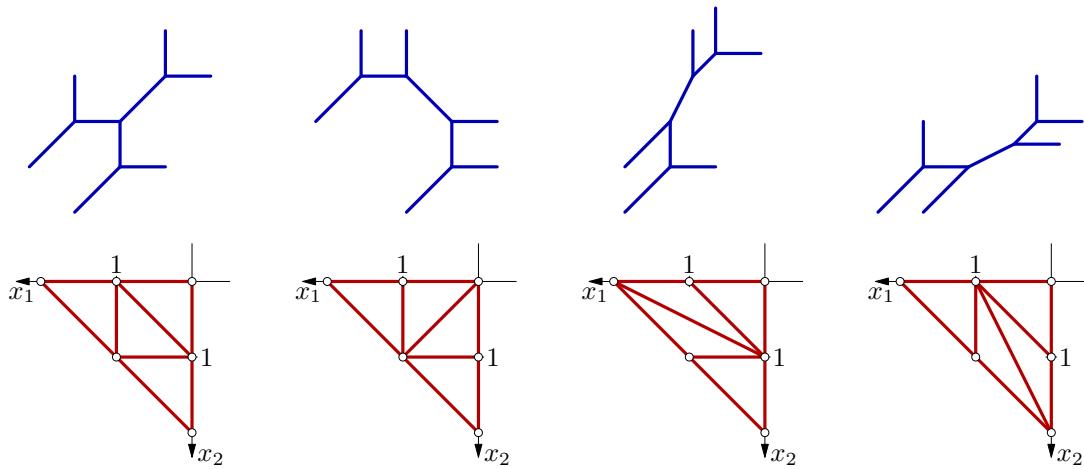


Figure 9.4: Types of tropical conics with a unimodular dual triangulation.

Example 9.1.5. Tropical polynomials $f \in \mathbb{R}_{\text{trop}}[x, y]$ of degree 2,

$$f = a_1 \odot x^2 \oplus a_2 \odot x \odot y \oplus a_3 \odot y^2 \oplus a_4 \odot x \oplus a_5 \odot y \oplus a_6,$$

define *tropical quadratic curves in the plane*. 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 2, the reader will investigate the dual subdivision in dependence of the coefficients a_1, \dots, a_6 .

1

¹Maybe: picture tropical cubic surface (Polymake)

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

$$m_F = (n - j)! \operatorname{vol}'_{n-j}(F),$$

where vol' denotes the volume in the lattice $\mathbb{Z}(F)$ affinely spanned by the integer vectors of F . In particular, if $\mathcal{T}(f)$ is a planar tropical curve, then the multiplicity is the lattice length of the corresponding edge in the dual subdivision $\operatorname{subdiv}(f)$ (see Figure 9.5)

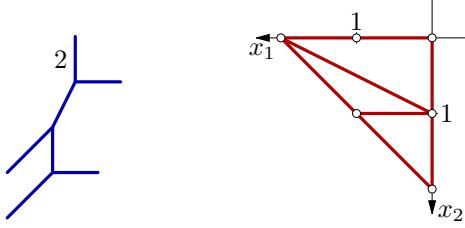


Figure 9.5: A tropical curve with a segment of weight 2 and its dual subdivision.

If f is a tropical polynomial whose Newton polytope Δ affinely spans \mathbb{Z}^n , then the number of vertices N (counting multiplicities) results to

$$\sum_{C \text{ } n\text{-cell in } \Delta} n! \operatorname{vol}'_n(C) = \sum_{C \text{ } n\text{-cell in } \Delta} n! \operatorname{vol}_n(C) = \operatorname{vol}_n(\Delta).$$

Similarly, the number of j -faces (counting multiplicity) is

$$\sum_{C \text{ } (n-j)\text{-cell in } \Delta} \operatorname{vol}'_{n-j}(C).$$

The balancing condition. As explained in the following, tropical hypersurfaces satisfy a local *balancing condition* (or *equilibrium condition*). The intrinsic relevance of this balancing condition is reflected by the fact that it can even be used to provide an alternative, synthetic definition of tropical hypersurfaces.

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 in \mathbb{R}^2 , it is easy to see that in each vertex a balancing condition is satisfied.

Lemma 9.1.6. *Let p be a vertex of a tropical curve $\mathcal{T}(f)$ in the plane, let $v^{(1)}, \dots, v^{(r)}$ be the primitive lattice vectors in the directions of the edges emanating from p , and let m_1, \dots, 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 $\operatorname{subdiv}(f)$. 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 to the zero vector, the claim follows. \square

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, \dots, 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 9.1.7. *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, \dots, 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, $\sum_{i=1}^k v_{F_i} = 0$. \square

In the following we discuss the converse and see that the balancing condition can be used to characterize tropical hypersurfaces.

Theorem 9.1.8. *The pure rational weighted balanced polyhedral complexes in \mathbb{R}^n are exactly the tropical hypersurfaces in \mathbb{R}^n .*

Proof. By Lemma 9.1.7 any tropical hypersurface 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 piecewise linear function $f : \mathbb{R}^n \rightarrow \mathbb{R}$.

We define f inductively. First choose a connected component C_0 of $\mathbb{R}^n \setminus \mathcal{P}$ 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 exists 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_0} to \mathbb{R}^n . Now define $f_{C'}$ as follows. By definition of F , the affine-linear extension of $f|_{C'}$ is fixed up to the rotation around the lifted $(n - 1)$ -face \hat{F} .

Let $\mu(F)$ be the weight of F and $n(F)$ its primitive normal vector, and set $u^* = \mu(F) \cdot n(F)$. Further let

$$\ell'_C = \ell_C + (u^*)^T x + \gamma$$

with γ a suitable constant such that ℓ_C and ℓ'_C coincide on F . Note that the term $(u^*)^T x$ is constant along F . Let $f_{C'}$ be the restriction of $\ell'_{C'}$ to C' . Since the graph of f is simply connected, the construction is well-defined. \square

An embedded graph G (allowing infinite half-rays as edges) in the plane \mathbb{R}^2 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, where $v^{(1)}, \dots, v^{(r)}$ are the primitive lattice vectors in the directions of the edges emanating from p and m_1, \dots, m_r are the multiplicities of these edges. Since the direction vectors and the normal vectors of the edges of an embedded planar are related by a 90 degree rotation, we obtain the following corollary of Theorem 9.1.8 for the planar case.

Corollary 9.1.9. *The rational, weighted, embedded graphs without isolated vertices are exactly the tropical curves in the plane.*

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. In order to explain 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 R_t . Indeed, note that for $x, y, z \in \mathbb{R}$ we have the distributive law $(x \oplus_t y) \odot z = x \odot y \oplus_t x \odot z$. Moreover, there exists a semiring isomorphism ϕ_t from the ordinary semiring $(\mathbb{R}_{>0}, +, \cdot)$ to each semiring R_t via $x \mapsto \log_t x$, because

$$\phi_t(x + y) = \phi_t(x) \oplus_t \phi_t(y) \text{ and } \phi_t(x \cdot y) = \phi_t(x) \odot \phi_t(y).$$

In the limit case for $t \downarrow 0$, we obtain

$$x \oplus_0 y = \min\{x, y\},$$

so that the tropical semiring can be regarded as the limit of R_t for $t \downarrow 0$. Exercise 4, however, will show that the tropical semiring is not isomorphic to any $(\mathbb{R}_t, +, \cdot)$ and thus it is not isomorphic to $(\mathbb{R}, +, \cdot)$ either.

The following inequality holds for $k \in \mathbb{N}$ and $x_1, \dots, x_k \in \mathbb{R}$:

$$\min\{x_1, \dots, x_k\} + \underbrace{\log_t k}_{<0} \leq x_1 \oplus_t \cdots \oplus_t x_k \leq \min\{x_1, \dots, x_k\}.$$

Given a polynomial $f \in \mathbb{R}[x_1, \dots, x_n]$, let f_t be the polynomial obtained from using the operations \oplus_t, \odot . Then we consider the function $g_t : (\mathbb{R}_{>0})^n \rightarrow \mathbb{R}_{>0}$,

$$g_t(z) = t^{f_t(\log_t z)}, \tag{9.2}$$

where \log_t is take component-wise, $\log_t z = (\log_t z_1, \dots, \log_t z_n)$.

Theorem 9.1.10 (Maslov, Viro). *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*

$$g_t(z) = \sum_{\alpha} t^{c_{\alpha}} z^{\alpha}.$$

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 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 corresponding upper convex hull defines a polyhedral subdivision of Δ into cells Δ_k . Denoting by $F = \sum_{\alpha \in \Delta} c_{\alpha} z^{\alpha}$ a generic real polynomial with regard to the Newton polytope Δ , one considers the *truncations* $F_{\Delta_k} = \sum_{\alpha \in \Delta_k} c_{\alpha} z^{\alpha}$ and the patchworking polynomials $f_t^v : \mathbb{R}^n \rightarrow \mathbb{R}$,

$$f_t^v = \sum_{\alpha} c_{\alpha} t^{v(\alpha)} z^{\alpha}$$

with a parameter $t > 1$ (which generalizes the polynomial in Theorem 9.1.10).

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 $\mathbb{R}V_t$ emerges from $\mathbb{R}V_{\Delta_k}$.

Exercises

1. Show that a face F of $\mathcal{T}(f)$ is unbounded if and only if F^{\vee} is contained in the boundary of $\text{New } f$.
2. Let

$$f = a_1 \odot x^2 \oplus a_2 \odot x \odot y \oplus a_3 \odot y^2 \oplus a_4 \odot y \odot z \oplus a_5 \odot z^2 \oplus a_6 \odot x \odot z$$

be a homogeneous tropical quadratic form of degree 2. Determine conditions on a_1, \dots, a_6 such that $\mathcal{T}(f)$ has six distinct half-rays.

3. 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 $(0, 1)$.
4. Show that the tropical semiring is not semiring isomorphic to any of the rings $(\mathbb{R}, \oplus_t, \odot)$, $0 < t < 1$.

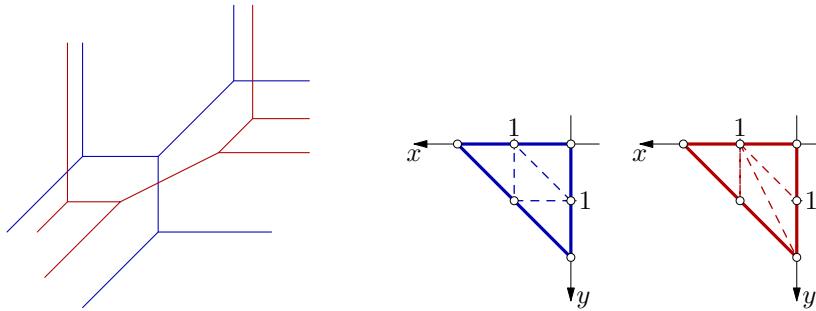


Figure 9.6: The intersection of two tropical conics in the plane. The two right pictures show the Newton polygons of these two conics with their corresponding dual triangulations.

9.2 Tropical prevarieties and stable intersections

An intersection of finitely many tropical hypersurfaces is called a *tropical prevariety*. In this section, we will see that these intersections obey certain intersection principles which we know from the intersection of projective hypersurfaces over algebraically closed fields. In particular, we present tropical versions of Bézout's and of Bernstein Theorem.

Given tropical polynomials $f_1, \dots, f_k \in \mathbb{R}_{\text{trop}}[x_1, \dots, x_n]$ with $k \leq n$, our goal is to study the intersection $\mathcal{T}(f_1) \cap \dots \cap \mathcal{T}(f_k)$. See Figure 9.6 for an example of the intersection of two tropical curves in the plane.

To approach this situation, we observe that $\mathcal{T}(\bigodot_{i=1}^k f_i) = \bigcup_{i=1}^k \mathcal{T}(f_i)$ can be regarded as a tropical hypersurface in tropical n -space, with Newton polytope $\text{New}(f) = \text{New}(f_1) + \dots + \text{New}(f_k)$. The *privileged subdivision* of the Newton polytope $\text{New}(f)$ is given by projecting down the lower hull of the Minkowski sum $\text{New}^e(f_1) + \dots + \text{New}^e(f_k)$. For a generic choice of coefficients in the system f_1, \dots, f_k , this subdivision is mixed (as defined in Appendix A.3.2).

By the duality between the privileged subdivision of $\text{New}(f)$ and the union $\bigcup_{i=1}^k \mathcal{T}(f_i)$, each cell C in the privileged subdivision corresponds to a cell A in the union $\bigcup_{i=1}^k \mathcal{T}(f_i)$ such that $\dim(C) + \dim(A) = n$, and C and A span orthogonal real affine spaces. A cell A of $\bigcup_{i=1}^k \mathcal{T}(f_i)$ is in the intersection $\mathcal{I} = \bigcap_{i=1}^k \mathcal{T}(f_i)$ if and only if the corresponding dual cell C in the privileged subdivision is mixed.

To provide quantitative statements on the number of points in an intersection, we first consider a well-behaved situation. For an intersection $\mathcal{I} = \bigcap_{i=1}^k \mathcal{T}(f_i)$, \mathcal{I} is *transversal along a cell* A of this complex if the dual cell $C = F_1 + \dots + F_k$ in the privileged subdivision of $P_1 + \dots + P_k$ satisfies

$$\dim(C) = \dim(F_1) + \dots + \dim(F_k) .$$

We call the intersection *transversal* if for each subset $J \subset \{1, \dots, k\}$ the intersection is transversal along each cell of the complex.

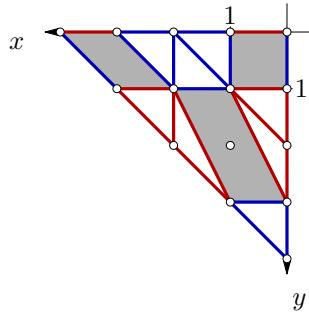


Figure 9.7: The privileged subdivision of $\text{New}(f_1 \odot f_2)$ for the two polynomials from Figure 9.6

In the following theorems, we focus on the case $k = n$. Every point in a transversal tropical intersection $\mathcal{I} = \bigcap_{i=1}^n \mathcal{T}(f_i)$ naturally comes with a multiplicity.

Definition 9.2.1. The *multiplicity* (or *weight*) m_p of a point p in a tropical transversal intersection $\bigcap_{i=1}^n \mathcal{T}(f_i)$ is given as follows. If $C = F_1 + \dots + F_n$ is the dual cell in $P_1 + \dots + P_n$, then m_p is the mixed volume $m_p = \text{MV}_n(P_1, \dots, P_n)$.

Theorem 9.2.2 (Tropical Bernstein Theorem (Weak version)). *Let f_1, \dots, f_n be tropical polynomials with Newton polytopes P_1, \dots, P_n , and suppose that the intersection $\mathcal{I} = \bigcap_{i=1}^n \mathcal{T}(f_i)$ is transversal. Then \mathcal{I} consists of finitely many points, and their number is $\text{MV}_n(P_1, \dots, P_n)$ counting multiplicity.*

Proof. Finiteness follows from the precondition that the dual cell of every point in $\bigcap_{i=1}^n \mathcal{T}(f_i)$ is of dimension n . For every point $p \in \bigcap_{i=1}^n \mathcal{T}(f_i)$, there is a mixed cell A_p in the dual subdivision. Since, by ...², the sum of the volumes of the mixed cells equals $\text{MV}_n(P_1, \dots, P_n)$, the claim follows. \square

Corollary 9.2.3 (Tropical Bézout Theorem (Weak version)). *Let f_1, \dots, f_n be tropical polynomials of degrees d_1, \dots, d_n . If the intersection $\bigcap_{i=1}^n \mathcal{T}(f_i)$ is transversal, then it consists of $\prod_{i=1}^n d_i$ points counting multiplicity.*

Proof. Let $\Delta = \text{conv}\{0, e^{(1)}, \dots, e^{(n)}\}$ with the unit vectors $e^{(1)}, \dots, e^{(n)}$ and $\Delta_i = d_i \Delta$. Then it suffices to observe that

$$\text{MV}(\Delta_1, \dots, \Delta_n) = \prod_{i=1}^n d_i.$$

\square

In the case of a non-transversal intersection $\mathcal{I} = \bigcap_{i=1}^k X_i$ with tropical hypersurfaces $X_i = \mathcal{T}(f_i)$ we can perturb the hypersurfaces by small parameters $\varepsilon_1, \dots, \varepsilon_k$ to obtain

²[the characterization of the mixed cell in terms of the sum of the volumes of the mixed cells should be located somewhere else; and CHECK consistency w.r.t. factor in the MV-definition!]

again a transversal intersection \mathcal{I}_ε . The *stable intersection* \mathcal{I}_{st} is defined as the limit of these transversal intersections when ε goes to 0,

$$\mathcal{I}_{\text{st}} = X_1 \cap_{\text{st}} \cdots \cap_{\text{st}} X_k = \lim_{\varepsilon \rightarrow 0} X_1^{(\varepsilon_1)} \cap \cdots \cap X_k^{(\varepsilon_k)}.$$

See Figure 9.15 further below. Since for any tropical polynomial $g \in \mathbb{R}_{\text{trop}}[x_1, \dots, x_n]$ the 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.

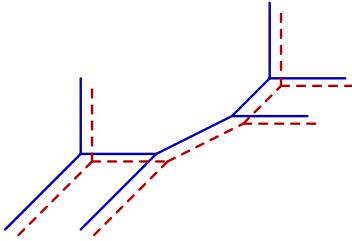


Figure 9.8: The stable intersection of a tropical quadratic curve with itself yields the vertices of the curve.

In the non-transversal case, the intersection multiplicities are induced from the corresponding stable intersection.

Theorem 9.2.4 (Tropical Bernstein Theorem). *Given tropical polynomials f_1, \dots, f_n with Newton polytopes P_1, \dots, P_n , their stable intersection consists of exactly $\text{MV}_n(P_1, \dots, P_n)$ points, counting multiplicities.*

Example 9.2.5. In the case of two tropical curves in the plane, the intersection multiplicity of two intersecting 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 in a spirit related to the treatment of numerical algebraic geometry in Chapter 4. The statement clearly holds for curves 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 9.9).

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

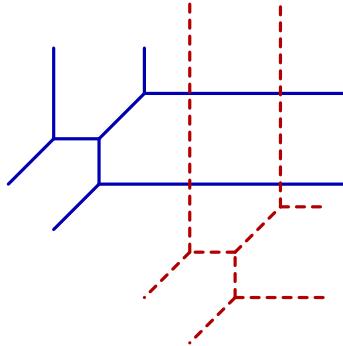


Figure 9.9: Two quadratic tropical curves in special position with points of intersection.

locally constant. 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 $v^{(1)}, \dots, v^{(k)}$ and $w^{(1)}, \dots, w^{(l)}$ 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.

Using the notion of stable intersection, we can provide the following variant of Bézout's Theorem as follows:

Corollary 9.2.6 (Tropical Bézout Theorem). *Given n tropical hypersurfaces of degrees d_1, \dots, d_n in n -space, their stable intersection consists of exactly $\prod_{i=1}^n d_i$ points, counting multiplicities.*

Exercises

1. Let $f \in \mathbb{R}_{\text{trop}}[x_1, \dots, x_n]$. Show that the unbounded cells of $\mathcal{T}(f)$ are associated to those cells of $\text{subdiv}(f)$ which are contained in the boundary of the Newton polytope of f .
2. Let p_1, \dots, p_5 be distinct points in \mathbb{R}^2 . If there are finitely many tropical conics passing through p_1, \dots, p_5 , then there is at most one.

3. Let C be a tropical curve in three-space given as the intersection of $X_1 \cap X_2$ of two tropical hypersurfaces X_1 and X_2 of degrees d and e . How many vertices can C at most have?
4. Generalize Exercise 3 to the intersection $\bigcap_{i=1}^{n-1} X_i$ of $n-1$ tropical hypersurfaces of degrees d_1, \dots, d_{n-1} in tropical n -space.

9.3 Amoebas

One way to access further concepts in tropical geometry is through amoebas. In the current section, we provide some basic insights on classical amoebas, which are logarithmic images of complex algebraic varieties. A main purpose is to observe several connections between amoebas and the tropical concepts developed in the previous subsections. Building upon this, the subsequent sections will then develop the view of tropical varieties as non-archimedean amoebas.

Throughout the section, rather than restricting to the ring of polynomials, it is usually convenient to consider the ring $\mathbb{C}[x^{\pm 1}] = \mathbb{C}[x_1^{\pm 1}, \dots, x_n^{\pm 1}]$ of *Laurent polynomials* and the subvariety of a Laurent polynomial within the algebraic torus $(\mathbb{C}^*)^n = (\mathbb{C} \setminus \{0\})^n$.

Definition 9.3.1. For a polynomial $f \in \mathbb{C}[x]$, the amoeba \mathcal{A}_f of f is the image set of its variety $\mathcal{V}(f) \subset (\mathbb{C}^*)^n$ under the log-absolute-map

$$\begin{aligned} \text{Log } |\cdot| : (\mathbb{C}^*)^n &\rightarrow \mathbb{R}^n, \\ x = (x_1, \dots, x_n) &\mapsto (\log |x_1|, \dots, \log |x_n|). \end{aligned}$$

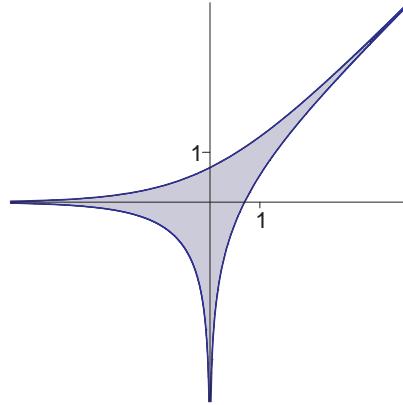
While in this section, we stay with the field of complex numbers and the usual absolute value function, it is useful to recall that an absolute value over an arbitrarily given field \mathbb{K} is a function $|\cdot|_{\mathbb{K}} : \mathbb{K} \rightarrow \mathbb{R}_{\geq 0}$ such that for any $a, b \in \mathbb{K}$ we have

1. $|a|_{\mathbb{K}} = 0$ if and only if $a = 0$,
2. $|ab|_{\mathbb{K}} = |a|_{\mathbb{K}}|b|_{\mathbb{K}}$,
3. $|a + b|_{\mathbb{K}} \leq |a|_{\mathbb{K}} + |b|_{\mathbb{K}}$.

Example 9.3.2. Let $f = ax + by + c \in \mathbb{C}[x, y] \subset \mathbb{C}[x^{\pm 1}, y^{\pm 1}]$ with $a, b, c \neq 0$. Since the variable transformations $x \mapsto cx'/a$, $y \mapsto cy'/a$ just translate the amoeba by $(\log c - \log a, \log c - \log b)$, it suffices to consider $a = 1$, $b = 1$, $c = 1$. Then $\mathcal{V}(f) = \{(v, -v - 1) : v \in \mathbb{C}^*\}$. If $|v| > 1$ then $|-v - 1| = |v + 1|$ can take any value in $[|v| - 1, |v| + 1]$. Similarly, if $|v| = 1$ then $|-v - 1|$ can take any value in $(0, 2]$, and if $0 < |v| < 1$ then v can take any value in $[1 - |v|, |v| + 1]$. Hence, the amoeba of f is bounded by the three curves

$$\{\log(\gamma, \gamma+1) : \gamma \in (0, \infty)\}, \quad \{\log(\gamma, 1-\gamma) : \gamma \in (0, 1)\} \text{ and } \{\log(\gamma, \gamma-1) : \gamma \in (1, \infty)\}.$$

See Figure 9.10.

Figure 9.10: The amoeba of $f = x + y + 1$

Lemma 9.3.3. *The amoeba of a Laurent polynomial $f \in \mathbb{C}[x^{\pm 1}]$ is closed. If f is irreducible, then \mathcal{A}_f is connected.*

Proof. Since irreducible complex varieties are connected, \mathcal{A}_f is connected. And since (\mathbb{C}^*) is locally compact, \mathcal{A}_f is closed. \square

The following properties are the reason why it is often convenient to look at $\log|x_i|$ rather than $|x_i|$ itself.

Theorem 9.3.4. *The complement of a hypersurface amoeba \mathcal{A}_f consists of finitely many convex regions, and these regions are in bijective correspondence with the different Laurent expansions of the rational function $1/f$ centered at the origin.*

In order to prove the theorem, we use the well-known result from multivariate complex analysis that the domains of convergence of Laurent series centered at the origin are of the form $\text{Log}^{-1}|B|$ for a convex open subset $B \subset \mathbb{R}^n$. And conversely, holomorphic functions f on a domain of the form $\text{Log}^{-1}|B|$ with an open and connected subset $B \subset \mathbb{R}^n$ have a unique Laurent expansion centered at the origin, that converges to f on $\text{Log}^{-1}|B|$.

Proof. Let C be a component in the complement of \mathcal{A}_f . Then $\frac{1}{f}$ is a holomorphic function on the set $\text{Log}^{-1}|C|$ and there is a unique Laurent series which converges to $\frac{1}{f}$ on $\text{Log}^{-1}|C|$. Since the domains of convergence of Laurent series are of the form $\text{Log}^{-1}|B|$ with an open and convex set $B \subset \mathbb{R}^n$, we can conclude $C \subset B$, and then $C \cap \mathcal{A}_f = \emptyset$ gives $C = B$.

As a consequence, the components of the complement of \mathcal{A}_f are convex, and each of them is associated to a unique Laurent series.

To show that the map from the complement components to the Laurent expansions is surjective, let L be a Laurent series of $\frac{1}{f}$ centered at the origin. Its domain of convergence is of the form $\text{Log}^{-1}|B|$ for some convex, open set $B \subset \mathbb{R}^n$. Since $B \cap \mathcal{A}_f = \emptyset$, B must intersect one of the components C in the complement of A . Since the Laurent expansion L' associated to C is unique and converges to $1/f$, we obtain $L = L'$. \square

Definition 9.3.5. The *order* ν of a point $a \in \mathbb{R}^n \setminus \mathcal{A}_f$ is defined by

$$\begin{aligned}\nu_j &= \frac{1}{(2\pi i)^n} \int_{\text{Log}^{-1}(a)} \frac{x_j \partial_j f(x)}{f(x)} \frac{dx_1 \cdots dx_n}{x_1 \cdots x_n} \\ &= \frac{1}{2\pi i} \int_{|\zeta|=1} d\log f(w_1, \dots, \zeta w_j, \dots, w_n) \text{ with } \text{Log}|w| = a, \quad 1 \leq j \leq n.\end{aligned}$$

The agreement of the two expressions follows from the argument principle of complex variables in one variable, see Appendix A.4. Using further methods from multidimensional complex analysis one can show that the order is always an integer vector within the Newton polytope $\text{New}(f)$. The order is invariant within the same component of the complement, and the resulting map

$$\{\text{components in the complement of } \mathcal{A}_f\} \longrightarrow \text{New}(f) \cap \mathbb{Z}^n$$

is injective.

Example 9.3.6. Figure 9.11 depicts the amoeba \mathcal{A}_f of a product $f = \prod_{i=1}^4 f_i$ of four linear functions linear polynomials $f_i \in \mathbb{C}[x, y]$. 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) several times independently a small value. See the right picture in Figure 9.11.

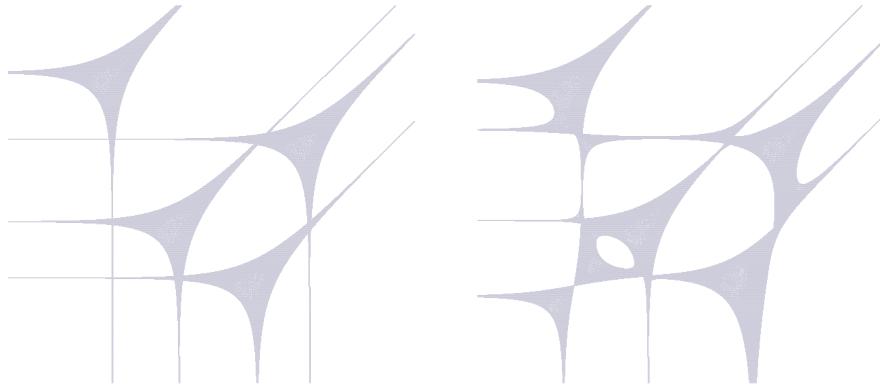


Figure 9.11: Two amoebas in two variables. The left picture shows the amoeba of $\mathcal{V}(f_1 \cdot f_2 \cdot f_3 \cdot f_4)$, where $f_1(x_1, x_2) = (\frac{1}{30}x_1 + \frac{1}{30}x_2 - 1)$, $f_2(x_1, x_2) = (\frac{1}{5}x_1 + 4x_2 - 1)$, $f_3(x_1, x_2) = (3x_1 + \frac{4}{7}x_2 - 1)$, $f_4(x_1, x_2) = (30x_1 + \frac{1}{300}x_2 - 1)$.

Lemma 9.3.7. Let $f \in \mathbb{C}[x^{\pm 1}]$. Then all components associated to the vertices of $\text{New}(f)$ exist in \mathcal{A}_f .

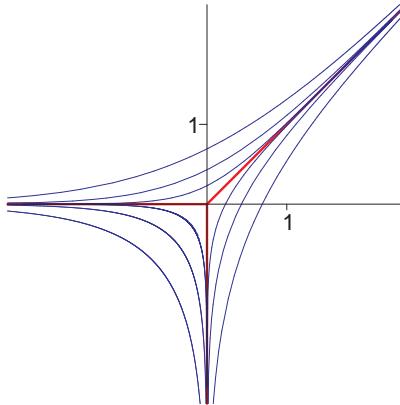


Figure 9.12: The amoebas $\mathcal{A}_f^{(t)}$ for $f = x + y + 1$ and $t \in \{1/e, 1/5, 1/20\}$. For $t = 1/e$, $\mathcal{A}_f^{(t)} = \mathcal{A}_f$, and for the smaller values, $\mathcal{A}_f^{(t)}$ converges to the tropical curve depicted in red.

Hence, the number of components in the complement is at least equal to the number of vertices in $\text{New}(f) \cap \mathbb{Z}^n$ and it is at most $|\text{New}(f) \cap \mathbb{Z}^n|$.

Proof. Let $f = \sum_{\alpha} c_{\alpha} x^{\alpha}$, and fix a vertex vertex v of $\text{New}(f)$. We can pick a $z \in \mathbb{C}^n$ such that $|c_v z^v| > |\sum_{\alpha \neq v} c_{\alpha} z^{\alpha}|$. Then the point $a = \text{Log}|z|$ is not contained in \mathcal{A}_f , and now it suffices to show that the order ν of a is v .

By (9.3) and the argument principle from complex analysis (see Appendix A.4), the entry ν_j equals the number of zeroes of the univariate function

$$\zeta \mapsto f(z_1, \dots, z_j \zeta, \dots, z_n)$$

inside the unit circle. Set $g(\zeta) = f(z_1, \dots, z_j \zeta, \dots, z_n) - c_v z^v \zeta^{v_j}$. Since $|g(\zeta)| < |f(z_1, \dots, z_j \zeta, \dots, z_n)|$ on the unit circle, Rouché's Theorem implies that ν_j equals the number of zeroes of the function

$$\zeta \mapsto f(z_1, \dots, z_j \zeta, \dots, z_n) - g(\zeta) = c_v z^v \zeta^{v_j}$$

inside the unit circle, that is, $\nu_j = v_j$. Hence, the point a lies in a component of the complement of \mathcal{A}_f with order v . \square

Maslov dequantization of amoebas. In Section 9.1, we have regarded the tropical semiring as the limit of dequantizations. Given $f \in \mathbb{R}[x]$, recall the functions f_t and $g_t(z) = t^{f_t(\log z)}$ from (9.2). Now we study the resulting deformation on the amoeba of a polynomial $f \in \mathbb{C}[x]$. For $0 < t < 1$, let $\text{Log}_t |\cdot|$ be the mapping $x \mapsto (\log|x_1|, \dots, \log|x_n|)$ and set

$$\mathcal{A}_f^{(t)} = \{\text{Log}_t |z| : z \in (\mathbb{C}^*)^n, g_t(z) = 0\}.$$

Note that $\mathcal{A}_f^{(1/e)} = \mathcal{A}_f$.

Figure 9.12 depicts some examples of dequantized amoebas. In the next section, we will consider the polynomials g_t from (9.2) as a family of polynomials and interpret the limit situation as the non-archimedean amoeba of g_t .

Lemma 9.3.8. *If a point $w \in \mathbb{R}^n$ belongs to the amoeba $\mathcal{A}_f^{(t)}$, then for each multiindex α we have*

$$c_\alpha \odot w^\alpha \geq \bigoplus_{\beta \neq \alpha} c_\beta \odot w^\beta.$$

Here, the index t in \bigoplus is omitted for notational convenience.

Proof. If $w = \text{Log}_t |x|$ with $g_t(x) = 0$ then for each α

$$t^{c_\alpha} x^\alpha = \sum_{\beta \neq \alpha} t^{c_\beta} x^\beta.$$

Passing over to the absolute value and applying the triangle inequality yields

$$t^{c_\alpha} |x|^\alpha \leq \sum_{\beta \neq \alpha} t^{c_\beta} |x|^\beta.$$

Now applying \log_t (for $0 < t < 1$) on both sides gives

$$c_\alpha \odot w^\alpha \geq \bigoplus_{\beta \neq \alpha} c_\beta \odot w^\beta.$$

□

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 $\mathcal{A}_t = \text{Log}_t |V_t|$ and $\mathcal{A}_{\text{trop}}$ be the tropical hypersurface of the tropical polynomial with the coefficients of f . The following convergence can be shown, whose proof we omit.

Theorem 9.3.9 (Mikhalkin). *For $t \downarrow 0$, the sequence of \mathcal{A}_t converges in the Hausdorff metric to the tropical hypersurface $\mathcal{A}_{\text{trop}}$.*

The spine. Among the various connections between amoebas and tropical geometry we sketch a second one.

Definition 9.3.10. For $f \in \mathbb{C}[x^{\pm 1}]$, the *spine* \mathcal{S}_f of the amoeba \mathcal{A}_f is the tropical hyperplane of the max-plus tropical polynomial

$$\max_{\alpha \in A'}\{r_\alpha + \langle \alpha, x \rangle\},$$

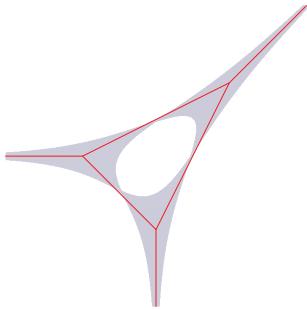


Figure 9.13: The amoeba of $f = x^3 + y^3 + 4xy + 1$ and its spine. The lines underlying the three infinity rays of the spine meet in the origin of the coordinate system.

where $A' \subset \text{New}(f) \cap \mathbb{Z}^n$ is the set of vectors α such that f has a component of order α in the complement. And for $\alpha \in A'$, the *Ronkin coefficients* r_α are defined by

$$r_\alpha = \frac{1}{(2\pi i)^n} \int_{\text{Log}^{-1}(a)} \log \left| \frac{f(x_1, \dots, x_n)}{x^\alpha} \right| \frac{dx_1 \cdots dx_n}{x_1 \cdots x_n},$$

where the integration is taken over any torus $\text{Log}^{-1}(a)$ with a in the component of order α .

The importance of the spine comes from the following statement, which we do not prove here, but only visualize it in Figure 9.13.

Theorem 9.3.11 (Passare, Rullgård). *For $f \in \mathbb{C}[x]$, the spine \mathcal{S}_f of the amoeba \mathcal{A}_f is a strong deformation retract of \mathcal{A}_f .*

Hence, 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 complement $\mathbb{R}^n \setminus \mathcal{A}_f$ of the amoeba.

Exercises

1. Give an example of a univariate polynomial f with pairwise distinct zeroes such that the amoeba $\mathcal{A}(f)$ consists of a single point.
2. Given a binomial $f(z) = x^\alpha - x^\beta \in \mathbb{C}[x]$ with $\alpha, \beta \in \mathbb{Z}^n$, determine that amoeba \mathcal{A}_f .
3. If α is a vertex of the Newton polytope of $f = \sum c_\alpha x^\alpha$, then the Ronkin coefficients r_α satisfy $r_\alpha = \log |c_\alpha|$.

4. The amoeba \mathcal{A}_f of $f = \sum c_\alpha x^\alpha$ is called *solid* if it has only the complement components corresponding to the vertices V of the Newton polytope of f . The spine of a solid amoeba is the tropical hypersurface of the max-plus tropical polynomial

$$\max_{\alpha \in V} \{\log |r_\alpha| + \langle \alpha, x \rangle\}.$$

9.4 Valuations and Kapranov's Theorem

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 and to prove Kapranov's Theorem, which provides a central link between the elementary viewpoint of tropical geometry and the advanced viewpoint in terms of valuations.

For a field \mathbb{K} , a *real valuation* is a map $\text{val} : \mathbb{K} \rightarrow \mathbb{R}_\infty = \mathbb{R} \cup \{\infty\}$ with

1. $\text{val}(x) = \infty \iff x = 0$,
2. $\text{val}(xy) = \text{val}(x) + \text{val}(y)$ and
3. $\text{val}(x+y) \geq \min\{\text{val}(x), \text{val}(y)\}$.

Example 9.4.1. 1.) The *trivial valuation* is given by $\text{val}(0) = \infty$ and $\text{val}(x) = 0$ for $x \neq 0$.

2) The *p-adic valuation* $v_p(\cdot)$ on the field $\mathbb{K} = \mathbb{Q}$ is defined as follows. If q has the form

$$q = p^s \frac{m}{n}, \quad s \in \mathbb{Z}, m \in \mathbb{Z}, n \in \mathbb{N}$$

where p neither divides m nor n , then $v_p(q) = s$.

3) If \mathbb{K} is a field and x a single indeterminate, then

$$v_\infty \left(\frac{f}{g} \right) = \deg g - \deg f$$

defines a valuation $v_\infty : \mathbb{K}(x) \rightarrow \mathbb{Z} \cup \{\infty\}$ on the quotient field $\mathbb{K}(x)$.

Remark 9.4.2. Let val be a real valuation on a field \mathbb{K} . Then for $c \in (0, 1)$ the function

$$|x| := c^{-\text{val}(x)}$$

defines a non-archimedean norm on \mathbb{K} , i.e., $|\cdot|$ is a map $\mathbb{K} \rightarrow \mathbb{R}_+$ satisfying the three properties $|x| = 0 \iff x = 0$, $|xy| = |x| \cdot |y|$, and $|x+y| \leq \max\{|x|, |y|\}$ for all $x, y \in \mathbb{K}$.

Let val be a real valuation on a field \mathbb{K} . Then the *valuation ring* of val is defined by

$$R_{\text{val}} = \{x \in \mathbb{K} : \text{val}(x) \geq 0\}.$$

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 \mathbb{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 $\text{val}(a) = \text{val}(-a)$, as will be shown in Exercise 1.

The set

$$M_{\text{val}} = \{x \in \mathbb{K} : \text{val}(x) > 0\} \subset R_{\text{val}}. \quad (9.3)$$

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 $x \in R_{\text{val}}$ with $\text{val}(x) = 0$, the inverse x^{-1} in \mathbb{K} satisfies $\text{val}(x^{-1}) = \text{val}(x) = 0$ by Exercise 1, 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, assume that I is another maximal ideal and that there is an element $x \in I \setminus M_{\text{val}}$. Since $\text{val}(x) = 0$, I contains a unit and thus $I = R_{\text{val}}$. Hence, M_{val} is a maximal ideal. \square

The unique maximal ideal M_{val} from (9.3) 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 \mathbb{K}_{val} .

Theorem 9.4.4. Let val be a real valuation on a field \mathbb{K} . If \mathbb{K} is algebraically closed then the residue class field \mathbb{K}_{val} is algebraically closed as well.

Proof. Denote by $\varphi : R_{\text{val}} \rightarrow \mathbb{K}_{\text{val}}$, $x \mapsto x + M_{\text{val}}$ be the canonical residue class mapping.

Let $f \in \mathbb{K}_{\text{val}}[x]$ 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 \mathbb{K}_{val} has a φ -preimage which is not contained in M_{val} .

Since $R_{\text{val}} \subset \mathbb{K}$ and \mathbb{K} is algebraically closed, there exists a $z \in \mathbb{K}$ with $q(z) = 0$. We show that z is a unit in R_{val} which implies the desired statement, since then $\varphi(z) \in \mathbb{K}_{\text{val}}$ 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 2 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., $z \in R_{\text{val}}^*$. \square

We can extend val to a mapping $\mathbb{K}^n \rightarrow \mathbb{R}^n$ by applying the valuation componentwise.

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 (Newton-Puiseux Theorem). $\mathbb{C}\{\{t\}\}$ 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 $\mathbb{K} = \mathbb{C}\{\{t\}\}$, 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 \mathbb{K} under this valuation map gives the set of rational numbers.

Alternatively, 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.

Rather than giving here a complete proof of Theorem 9.4.5, we restrict ourselves here to present the constructive method behind it. Given a polynomial

$$p(x) = c_n x^n + \cdots + c_1 x + c_0$$

with coefficients $c_i = c_i(t) \in \mathbb{K}$, this method does not only compute one of the zeros of p , but all of them.

Example 9.4.6. Let $p = x^2 + x - t^2 \in \mathbb{K}$. The zeroes of p in \mathbb{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

$$\gamma_{00}, \gamma_{10} + \mu_0, \gamma_{20} + 2\mu_0, \dots, \gamma_{n0} + n\mu_0$$

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 a modified Newton polygon. For a univariate polynomial $p \in \mathbb{K}[x]$, let the *support* and the *extended support* of p be defined by

$$\begin{aligned} \text{supp}(p) &= \{i : c_i \neq 0, 0 \leq i \leq n\} \subset \mathbb{N}_0, \\ \text{supp}^e(p) &= \{(i, \text{val } c_i) : c_i \neq 0, 0 \leq i \leq n\} \subset \mathbb{N}_0 \times \mathbb{R}, \end{aligned}$$

and let

$$\text{New}^m(p) = \text{conv}\{(\alpha, u) \in \text{supp}(p) \times \mathbb{R} : u \geq \text{val}(c_\alpha)\} \subset \mathbb{R}^2,$$

be the *modified Newton polygon* of p . Figure 9.14 shows the modified Newton polygon for our running example.

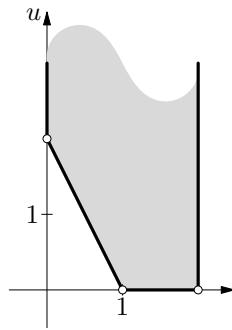


Figure 9.14: Modified Newton polygon.

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 \mathbb{K}) are exactly the negative slopes of the lower edges of $\text{New}^m(p)$. 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 modified 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

$$p_1 = t^{-\beta_0} p(t^{\mu_0}(x + b_0)) \in \mathbb{K}[x]$$

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 modified 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 completing a formal proof of Theorem 9.4.5 is that in each iteration there exists an edge with negative slope and that the fractional exponents that occur have a common denominator.

Kapranov's Theorem. For a polynomial $f = \sum_{\alpha \in \mathcal{A}} c_\alpha(t)x^\alpha \in \mathbb{K}[x_1, \dots, x_n]$ with 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

$$\text{trop } f = \bigoplus_{\alpha \in \mathcal{A}} (c_\alpha(t)) \odot x^\alpha.$$

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 \mathbb{K} .

Theorem 9.4.9 (Kapranov). *Let \mathbb{K} be algebraically closed. Then for $f \in \mathbb{K}[x]$ we have*

$$\mathcal{T}(\text{trop } f) \cap \text{val}(\mathbb{K}^*)^n = \text{val } \mathcal{V}(f), \quad (9.4)$$

where $\mathcal{V}(f)$ denotes the zero set in $(\mathbb{K}^*)^n$. In particular, the topological closure of the set $\text{val } \mathcal{V}(f)$ is contained in $\mathcal{T}(\text{trop } f)$. If the valuation val is surjective, then the equality

$$\mathcal{T}(\text{trop } f) = \text{val } \mathcal{V}(f) \quad (9.5)$$

holds.

Example 9.4.10. Let $\mathbb{K} = \mathbb{C}\{\{t\}\}$ and f be the linear polynomial $f = (3t^2 + t^4)x^2 + txy + 2t^7y^3 + 5 \in \mathbb{K}[x, y]$. Clearly, for any point $a \in \mathcal{V}(f)$ the valuation $(\text{val } x, \text{val } y) \in (\mathbb{R}_\infty)^2$ is contained in the tropical hypersurface of the tropical polynomial $2 \odot x^2 \oplus 1 \odot x \odot y \oplus 7 \odot y^3 \oplus 5$. The interesting direction of Kapranov's Theorem guarantees that for any point $w \in \mathcal{T}(2 \odot x^2 \oplus 1 \odot x \odot y \oplus 7 \odot y^3 \oplus 5)$, there is indeed a point $a \in \mathcal{V}(f)$ whose valuation coincides with w .

Before proving the statement, we provide some auxiliary lemmas:

Lemma 9.4.11. *Let $f \in \mathbb{K}[x]$ be a univariate polynomial and $w \in \mathcal{T}(\text{trop } f)$. Then there exists some $z \neq 0$ in the algebraic closure of \mathbb{K} such that $f(z) = 0$ and $\text{val}(z) = w$.*

Proof. Let $f \in \mathbb{K}[x]$ be a univariate polynomial. By the Fundamental Theorem of Algebra, f factors into $f(x) = \prod_{i=1}^n (x - a_i)$ with $a_1, \dots, a_n \in \mathbb{K}$. Then, by Exercise 5,

$$\begin{aligned}\operatorname{trop} f &= \operatorname{trop}((x - a_1) \cdots (x - a_n)) \\ &= (x \oplus \operatorname{val}(a_1)) \odot \cdots \odot (x \oplus \operatorname{val}(a_n))\end{aligned}$$

and $\mathcal{T}(\operatorname{trop} f) = \bigcup_{i=1}^n \mathcal{T}(x \oplus \operatorname{val}(a_i))$. Hence, 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. \square

Lemma 9.4.12. *Let $f_1, \dots, f_k \in R_{\operatorname{val}}[x_1, \dots, x_n]$, where R_{val} denotes the valuation ring. Assume that for $1 \leq i \leq n$ there exists a coefficient in f_i which is a unit in the valuation ring R_{val} . Then there exists some $y \in (R_{\operatorname{val}}^*)^n$ with $f_i(y) \in R_{\operatorname{val}}^*$ for $1 \leq i \leq k$.*

Proof. Denoting by $\varphi : R_{\operatorname{val}} \rightarrow \mathbb{K}_{\operatorname{val}} = R_{\operatorname{val}}/M_{\operatorname{val}}$ the canonical residue class homomorphism, φ extends to a ring homomorphism $\varphi : R_{\operatorname{val}}[x_1, \dots, x_n] \rightarrow \mathbb{K}_{\operatorname{val}}[x_1, \dots, x_n]$. By our assumptions on f_1, \dots, f_k , none of the polynomials $\bar{f}_i := \varphi(f_i)$ is the zero polynomial. Since \mathbb{K} is algebraically closed, $\mathbb{K}_{\operatorname{val}}$ is algebraically closed by Theorem 9.4.4. And since any algebraically closed field has infinitely many elements, we can choose an $r \in (\mathbb{K}_{\operatorname{val}}^*)^n$ with $\bar{f}_i(r) \neq 0$. Let $y \in R_{\operatorname{val}}^n$ be a (componentwise) preimage of r . Since $r \in (\mathbb{K}_{\operatorname{val}}^*)^n$, we have $y \in (R_{\operatorname{val}}^*)^n$. Hence

$$\varphi(f_i(y)) = \bar{f}_i(r) \neq 0,$$

and consequently $f_i(y) \in R_{\operatorname{val}}^*$, $1 \leq i \leq k$. \square

Lemma 9.4.13. *Let $f_1, \dots, f_k \in \mathbb{K}[x_1, \dots, x_n]$ and $w \in \operatorname{val}(\mathbb{K}^*)^n$. Then there exists an $a \in \mathbb{K}^n$ with $\operatorname{val}(a) = w$ and $\operatorname{val}(f_i(a)) = (\operatorname{trop} f_i)(w)$ for $1 \leq i \leq k$.*

Proof. Let $y \in \mathbb{K}^n$ with $\operatorname{val}(y) = w$. Further, for $i \in \{1, \dots, n\}$ there exists $z_i \in \mathbb{K}^*$ with $\operatorname{val}(z_i) = (\operatorname{trop} f_i)(w)$. For each i define the polynomial

$$g_i(x_1, \dots, x_n) = \frac{1}{z_i} f_i(y_1 x_1, \dots, y_n x_n).$$

Writing g_i in the form $g_i = \sum_{\alpha} c_{i\alpha} x^{\alpha}$, we see from the choice of z_i that $(\operatorname{trop} g_i)(0, \dots, 0) = \bigoplus_{\alpha} \operatorname{val}(c_{i\alpha}) = 0$. Hence, for all α we have $\operatorname{val}(c_{i\alpha}) \geq 0$, and there exists an α with $\operatorname{val}(c_{i\alpha}) = 0$. In other words, $g_i \in R_{\operatorname{val}}[x_1, \dots, x_n]$, and there exists a coefficient of g_i which is unit in R_{val} . By Lemma 9.4.12, there exists an $u \in (R_{\operatorname{val}}^*)^n$ with $\operatorname{val}(g_i(u)) = 0$. We show that the point $a = (y_1 u_1, \dots, y_n u_n)$ satisfies the desired properties. The condition $\operatorname{val}(f_i(a)) = (\operatorname{trop} f_i)(w)$ can be verified straightforwardly. Moreover, since each u_i is a unit in the valuation ring, we observe

$$\operatorname{val}(a) = (\operatorname{val}(y_1 u_1), \dots, \operatorname{val}(y_n u_n)) = (\operatorname{val}(y_1), \dots, \operatorname{val}(y_n)) = w.$$

\square

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 $\text{val}(a) \in \mathcal{T}(f_{\text{trop}})$, we observe

$$\text{val}\left(\sum_{\alpha} c_{\alpha}a^{\alpha}\right) = \text{val}(0) = \infty > \min_{\alpha} \text{val}(c_{\alpha}a^{\alpha}).$$

If that minimum were attained only for a single multi-exponent α^* , then

$$\begin{aligned} \text{val}(c_{\alpha^*}a^{\alpha^*}) &= \text{val}\left(\sum_{\alpha} c_{\alpha}a^{\alpha} - \sum_{\alpha \neq \alpha^*} c_{\alpha}a^{\alpha}\right) \geq \min\left\{\infty, \text{val}\sum_{\alpha \neq \alpha^*} c_{\alpha}a^{\alpha}\right\} \\ &> \text{val}(c_{\alpha^*}a^{\alpha^*}), \end{aligned}$$

a contradiction. Hence, $\text{val}(a) \in \mathcal{T}(\text{trop } f)$.

Conversely, let $w \in \mathcal{T}(\text{trop } f) \cap (\text{val } \mathbb{K}^*)^n$. Choose $y \in (\mathbb{K}^*)^n$ with $\text{val}(y) = w$. By the definition of a tropical hypersurface, 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

$$f(x_1, \dots, x_n) = \sum_i h_i(x_2, \dots, x_n)x_1^i.$$

By Lemma 9.4.13 we can pick a $y^* \in \mathbb{K}^{n-1}$ with $\text{val}(y^*) = (w_2, \dots, w_n)$ and $\text{val}(h_i(y^*)) = (\text{trop } h_i)((w_2, \dots, w_n))$ for all i . For the univariate polynomial

$$g(x_1) = \sum_i h_i(y^*)x_1^i,$$

we have $w_1 \in \mathcal{T}(\text{trop } g)$. By Lemma 9.4.11, there exists a $z \in \mathbb{K}^*$ with $z = \text{val}(w_1)$ and $g(z) = 0$. Hence, $g(z) = f(y^*, z) = 0$.

Since $\mathcal{T}(f)$ is topologically closed, the closure of $\text{val } V(f)$ is contained in $\mathcal{T}(f)$. Moreover, if val is surjective then clearly (9.5) holds. \square

Exercises

1. Let val be a real valuation on a field \mathbb{K} . Show for $a \in \mathbb{K}$, $n \in \mathbb{N}$:

- (a) $\text{val}(1) = 0$;
- (b) $\text{val}(a) = \text{val}(-a)$;
- (c) $\text{val}(a^{-1}) = -\text{val}(a)$;
- (d) $\text{val}(a^n) = n \text{val}(a)$.

2. Let val be a real valuation on a field \mathbb{K} . Show

- (a) $\text{val}(x_1 + \dots + x_n) \geq \min(\text{val}(x_1), \dots, \text{val}(x_n))$;
- (b) If there exists exactly one $k \in \{1, \dots, n\}$ with $\text{val}(x_k) = \min\{\text{val}(x_1), \dots, \text{val}(x_n)\}$, then $\text{val}(x_1 + \dots + x_n) = \text{val}(x_k)$.

- (c) If $x_1 + \cdots + x_n = 0$ for $n \geq 2$, then there exist $j \neq k \in \{1, \dots, n\}$ with $\text{val}(x_j) = \text{val}(x_k) = \min\{\text{val}(x_1), \dots, \text{val}(x_n)\}$.
3. Let $\mathbb{K} = \mathbb{Q}$ and val be the p -adic valuation. Then:
- the valuation ring is $\left\{ \frac{a}{b} \in \mathbb{Q} : p \text{ does not divide } b \right\}$.
 - the valuation ideal is $\left\{ \frac{a}{b} \in \mathbb{Q} : p \text{ does not divide } b \text{ and } p|a \right\}$.
 - the residue class field is \mathbb{F}_p (the field with p elements).
4. Let val be a real valuation on a field \mathbb{K} and L an algebraic field extension of \mathbb{K} . Show that val extends uniquely to L .
5. If $f, g \in \mathbb{K}[x]$ then $\text{trop}(fg) = \text{trop}(f) \odot \text{trop}(g)$.

9.5 Tropical varieties

Now we discuss general tropical varieties. In the following, let $\mathbb{K} = \mathbb{C}\{\{t\}\}$ denote the field of Puiseux series with the natural valuation. We remark that many statements generalize to arbitrary nontrivial valuations.

For an ideal $I \subset \mathbb{K}[x_1, \dots, x_n]$, the *tropical variety* of I is defined by the topological closure $\mathcal{T}(I) = \overline{\text{val}(\mathcal{V}(I))}$ where $\mathcal{V}(I)$ is the subvariety of I in $(\mathbb{K}^*)^n$.

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 9.4.9.

We introduce the following notation for an ideal I in $\mathbb{K}[x_1, \dots, x_n]$. 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 $\text{val}(ct^b) + w \cdot \alpha = b + w \cdot \alpha$. The t -initial form $\text{t-in}_w(f) \in \mathbb{C}[x_1, \dots, x_n]$ of a polynomial $f \in \mathbb{K}[x_1, \dots, x_n]$ 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:

$$\text{t-in}_w(I) = \langle \text{t-in}_w(f) : f \in I \rangle \subset \mathbb{C}[x_1, \dots, x_n].$$

Example 9.5.1. Let $f = (t+t^2) \cdot x_1^2 + (t^2+t^3) \cdot x_2 + t^3$. 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 9.5.2 (Fundamental Theorem of Tropical Geometry). *For an ideal $I \subset \mathbb{K}[x_1, \dots, x_n]$ 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 $t\text{-in}_w(I)$ does not contain a monomial.

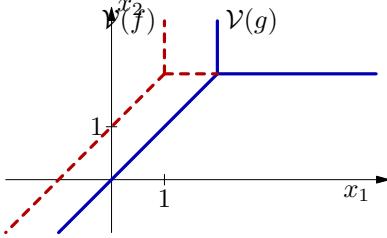


Figure 9.15: A non-transversal tropical intersection $\mathcal{T}(f) \cap \mathcal{T}(g)$, where the tropical variety $\mathcal{T}(I) = \mathcal{T}(\langle f, g \rangle)$ is the singleton $(2, 2)$.

Example 9.5.3. Let $f = t^2 \cdot x_1 + t^1 \cdot x_2 + t^3$, $g = t^2 \cdot x_1 + t^0 \cdot x_2 + (t^2 + t^3) \in \mathbb{K}[x_1, x_2]$. The set-theoretic intersection of $\mathcal{T}(f)$ and $\mathcal{T}(g)$ is the ray $\{(1, \alpha) : \alpha \geq 2\}$, and hence the intersection of the two tropical hypersurfaces is not transversal. See Figure 9.15

Setting $I = \langle f, g \rangle$, the tropical variety $\mathcal{T}(I)$ equals $\{(1, 2)\}$. To see this, note that $f - g = (t^1 - t^0)x_2 - t^2 \in I$ implies $\mathcal{T}(f - g) = \mathbb{R} \times \{2\}$ and $t \cdot g - f = (t^3 - t^2)x_1 + t^4$ implies $\mathcal{T}(t \cdot g - f) = \{2\}$. Hence, $\mathcal{T}(\langle f, g \rangle) \subset \mathcal{T}(f - g) \cap \mathcal{T}(g - t \cdot f) = \{(2, 2)\}$. And since the rational functions $x_1 = -\frac{t^4}{t^3 - t^2}$, $x_2 = -(t^3 + t^2) + \frac{t^6}{t^3 - t^2}$ give a zero of both f and g , we see that the subset relation holds with equality.

Remark 9.5.4. If all coefficients of the generators are contained in $\mathbb{C} \subset \mathbb{K}$ (“constant coefficient case”) then the third set in Theorem (9.5.2) can also be replaced by the set of all vectors $w \in \mathbb{R}^n$ such that $\text{in}_w(I)$ does not contain a monomial, where $\text{in}_w(I)$ is the well known initial ideal from Gröbner basis theory. Namely, recall that the initial form $\text{in}_w(f)$ of a polynomial $f \in \mathbb{C}[x]$ is the sum of all terms of f whose exponents maximize the linear form $\langle \omega, \cdot \rangle$. And $\text{in}_w(I) = \{\text{in}_w(f) : f \in I\}$.

For the proof of Theorem 9.5.2, we first record the following auxiliary statement. For notational convenience, let $\mathcal{T}'(I) = \{w \in \mathbb{R}^n : t\text{-in}_w(I) \text{ does not contain a monomial}\}$.

Lemma 9.5.5. Let I be an ideal in $\mathbb{K}[x_1, \dots, x_n]$ 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 $t\text{-in}_w(f)$ is not a monomial.

Proof. The direction “ \implies ” is clear. For the converse direction, we show that all elements in $t\text{-in}_w(I)$ have the form $t\text{-in}_w(f)$ 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 polynomial $f \in I$. \square

Based on this lemma, we can now prove the equivalence of the last two statements in Theorem 9.5.2.

Proof. (Equivalence of last two statements in Theorem 9.5.2). Choose a fixed polynomial $f \in I$. By Lemma 9.5.5 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 $\text{trop } f$ is attained at least twice in w if and only $\text{t-in}_w(f)$ is not a monomial. \square

The following lemma captures the equivalence of the first and the third statement in Theorem 9.5.2.

Lemma 9.5.6. *Let $I \subset \mathbb{K}[x_1, \dots, x_n]$ 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 9.5.7. *Let $I, J \subset \mathbb{K}[x_1, \dots, x_n]$ 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 9.5.5 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 $\text{t-in}_w(f) \cdot \text{t-in}_w(g)$. 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 $w \notin \mathcal{T}'(I)$. \square

Proof. (of Lemma 9.5.6). Let $p \in \mathcal{V}(I) \cap \mathbb{K}^*$ 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 9.5.5 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 9.5.7, we have

$$\mathcal{T}'(I) = \mathcal{T}'\left(\bigcap_i Q_i\right) = \bigcup_i \mathcal{T}'(Q_i) = \bigcup_i \mathcal{T}'(\sqrt{Q_i}).$$

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 $\sqrt{Q_1} = \langle x_1 - p_1, \dots, x_n - p_n \rangle$ for some $p \in \mathbb{K}^n$. Since $w \in \mathcal{T}'(I)$, we can conclude $p \in (\mathbb{K}^*)^n$ and $w = \text{val } p$ for $1 \leq i \leq n$. This means $w \in \mathcal{T}(I)$. We omit the case of general dimension. \square

Exercises

1. Let

$$I = \langle a_1(t) \cdot x_1 + \dots + a_4(t) \cdot x_4, b_1(t) \cdot x_1 + \dots + b_4(t) \cdot x_4 \rangle$$

with coefficients $a_i(t), b_j(t) \in \mathbb{C}\{\{t\}\}$, $1 \leq i, j \leq 4$. Show that if $\text{val}(a_i(t)b_j(t)) \neq \text{val}(a_j(t)b_i(t))$ for all i, j then $\mathcal{T}(I)$ consists of a line segment and four rays, where from each endpoint from a ray two rays emanate in coordinate directions.

2. Setting $p_{ij}(t) = a_i(t)b_j(t) - a_j(t)b_i(t)$ in the situation of the previous exercise, determine the coordinates of the endpoints. Distinguish the three cases $p_{14} + p_{23} = p_{13} + p_{24} \leq a_{12} + a_{34}$, $p_{14} + p_{23} = p_{12} + p_{34} \leq a_{13} + a_{24}$ and $p_{13} + p_{24} = p_{12} + p_{34} \leq a_{14} + a_{23}$.
3. Let I be an ideal in $\mathbb{C}[x]$. Considering I as a subset of $\mathbb{C}\{\{t\}\}[x]$, set $J = \langle I \rangle \subseteq \mathbb{C}\{\{t\}\}[x]$ and show $\text{in}_w(I) = \text{t-in}_w(J)$ for all $w \in \mathbb{R}^n$.
4. Given

$$I = \langle t^0 \cdot x^2y^2 - t^2 \cdot x^2y + t \cdot xy^2 + t^0 \cdot y^0 + t^1, t^4 \cdot xyz^2 + t^0 \cdot xyz + t^0 \cdot z^2 - t^1 \cdot z - t^1 \rangle,$$

how many segment and rays does the tropical variety $\mathcal{T}(I)$ have?

9.6 Tropical bases

In this section, we present the concept of tropical bases. As before, we focus on the field $\mathbb{K} = \mathbb{C}\{\{t\}\}$ of Puiseux series with its natural valuation.

Let I be an ideal in $\mathbb{K}[x]$. A finite subset $\{f_1, \dots, f_t\} \subset I$ is called a *tropical basis* of I if

$$\mathcal{T}(I) = \mathcal{T}(f_1) \cap \cdots \cap \mathcal{T}(f_t).$$

We will see that every tropical variety has a tropical basis.

Let $I \subset \mathbb{K}[x]$ be an m -dimensional prime ideal. If $m = n - 1$, then, by Krull's Principal Ideal Theorem, I is a principal ideal $I = \langle f \rangle$ and thus $\mathcal{T}(I)$ is a tropical hypersurface, $\mathcal{T}(I) = \mathcal{T}(f)$.

Following a strategy developed by Bieri and Groves, consideration of general tropical varieties can be reduced to the hypersurface situation via rational projections. Given n and d , a *rational projection* is a linear map $\pi : \mathbb{R}^n \rightarrow \mathbb{R}^{d+1}$ described by a matrix $A = (a_{ij})$ with rational entries. The main geometric idea in the reduction is to consider $n - m + 1$ different (rational) projections $\pi_0, \dots, \pi_{n-m} : \mathbb{R}^n \rightarrow \mathbb{R}^{m+1}$. We will show the following theorems.

Proposition 9.6.1 (Bieri, Groves). *Let $I \subset \mathbb{K}[x]$ be a prime ideal. For any dense set \mathcal{D} of projections there exist $\text{codim } I + 1$ projections $\pi_0, \dots, \pi_{n-\dim I} \in \mathcal{D}$ such that*

$$\mathcal{T}(I) = \bigcap_{i=0}^{n-\dim I} \pi_i^{-1} \pi_i(\mathcal{T}(I)).$$

The set $\mathcal{T}(I)$ defines a pure polyhedral complex of dimension $\dim I$.

In particular, this works for any choice of sufficiently generic projections. The projection idea is visualized in Figure 9.16. Further note, that the face structure of the polyhedral complex is inherited in the canonical way from the set-theoretic description via a finite number of polyhedra. Namely, by adding all lower-dimensional faces of any of these polyhedra.

This technique can be used to obtain a tropical basis:

Theorem 9.6.2 (Hept, Theobald). *Let $I \subset \mathbb{K}[x]$ be a prime ideal generated by the polynomials f_1, \dots, f_r . Then there exist $g_0, \dots, g_{n-\dim I} \in I$ with*

$$\mathcal{T}(I) = \bigcap_{i=0}^{n-\dim I} \mathcal{T}(g_i) \tag{9.6}$$

and thus $\mathcal{G} := \{f_1, \dots, f_r, g_0, \dots, g_{n-\dim I}\}$ is a tropical basis for I of cardinality $r + \text{codim } I + 1$.

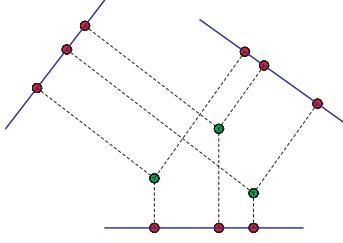


Figure 9.16: The green points show tropical variety of dimension $m = 0$ in dimension $n = 2$. In Theorem 9.6.1, this tropical variety is reconstructed from $n - m + 1 = 3$ projections, which are visualized by the points on the outer lines.

Consider the image of the tropical variety $\mathcal{T}(I)$ under a single (rational) projection

$$\begin{aligned}\pi : \mathbb{R}^n &\rightarrow \mathbb{R}^{m+1}, \\ x &\mapsto Ax\end{aligned}$$

with a non-singular rational matrix A whose rows are denoted by $a^{(1)}, \dots, a^{(m+1)}$. Let $u^{(1)}, \dots, u^{(l)} \in \mathbb{Z}^n$ with $l := n - (m + 1)$ be a basis of the orthogonal complement of $\text{span}\{a^{(1)}, \dots, a^{(m+1)}\}$.

Set $R = \mathbb{K}[x, \lambda] = \mathbb{K}[x_1, \dots, x_n, \lambda_1, \dots, \lambda_l]$, and for any polynomial $f \in \mathbb{K}[x]$ 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.,

$$\hat{f}(x_1, \dots, x_n, \lambda_1, \dots, \lambda_l) = f(x_1 \prod_{j=1}^l \lambda_j^{u_1^{(j)}}, \dots, x_n \prod_{j=1}^l \lambda_j^{u_n^{(j)}}) \in R.$$

Define the ideal $J \subset R$ by

$$J = \langle \hat{f} \in R : f \in I \rangle.$$

We show the following characterization of $\pi^{-1}(\pi(\mathcal{T}(I)))$ in terms of elimination.

Theorem 9.6.3. *Let $I \subset \mathbb{K}[x]$ be an m -dimensional prime ideal and $\pi : \mathbb{R}^n \rightarrow \mathbb{R}^{m+1}$ be a rational projection. Then $\pi^{-1}(\pi(\mathcal{T}(I)))$ is a tropical variety with*

$$\pi^{-1}(\pi(\mathcal{T}(I))) = \mathcal{T}(J \cap \mathbb{K}[x]). \quad (9.7)$$

If $\dim \pi(\mathcal{T}(I)) = m$, then $\pi^{-1}(\pi(\mathcal{T}(I)))$ is a tropical hypersurface. In particular, this happens if π is sufficiently general.

Note that this statement also implies that $\mathcal{T}(I)$ is polyhedral set of dimension d . We start with two auxiliary statements.

Lemma 9.6.4. *For any $w \in \mathcal{T}(J \cap \mathbb{K}[x])$ and $u \in \text{span}\{u^{(1)}, \dots, u^{(l)}\}$ we have $w + u \in \mathcal{T}(J \cap \mathbb{K}[x])$.*

Proof. Let $u = \sum_{j=1}^l \mu_j u^{(j)}$ with $\mu_1, \dots, \mu_l \in \mathbb{Q}$. The case of real μ_i then follows as well.

Let $w \in \mathcal{T}(J \cap \mathbb{K}[x])$. Since $\mathcal{T}(J \cap \mathbb{K}[x])$ is closed, we can assume without loss of generality that there exists $z \in \mathcal{V}(J \cap \mathbb{K}[x])$ with $\text{ord } z = w$. Define $y = (y', y'') \in (\mathbb{K}^*)^{n+l}$ by

$$y = (y', y'') = \left(z_1 t^{\sum_{j=1}^l \mu_j u_1^{(j)}}, \dots, z_n t^{\sum_{j=1}^l \mu_j u_n^{(j)}}, t^{-\mu_1}, \dots, t^{-\mu_l} \right).$$

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, $y' \in \mathcal{V}(J \cap \mathbb{K}[x])$. Moreover,

$$\text{ord } y' = (w_1 + \sum_{j=1}^l \mu_j u_1^{(j)}, \dots, w_n + \sum_{j=1}^l \mu_j u_n^{(j)}) = w + \sum_{j=1}^l \mu_j u^{(j)} = w + u,$$

which proves our claim. \square

Lemma 9.6.5. *Let $I \subset \mathbb{K}[x]$ be an ideal. Then $J \cap \mathbb{K}[x] \subset I$.*

Proof. Let $p = \sum_i h_i \hat{f}_i$ be a polynomial in $J \cap \mathbb{K}[x]$ with $f_i \in I$. Since p is independent of $\lambda_1, \dots, \lambda_l$ we have

$$p = p|_{\lambda_1=1, \dots, \lambda_l=1} = \sum_i h_i|_{\lambda_1=1, \dots, \lambda_l=1} f_i \in I.$$

\square

Theorem 9.6.6. *Let $I \subset \mathbb{K}[x]$ be a prime ideal and $\pi : \mathbb{R}^n \rightarrow \mathbb{R}^{m+1}$ be a sufficiently general projection. Then $\pi^{-1}\pi(\mathcal{T}(I))$ is a tropical variety with*

$$\pi^{-1}\pi(\mathcal{T}(I)) = \mathcal{T}(J \cap \mathbb{K}[x]). \quad (9.8)$$

See Exercise 9.6.6 for a formal definition ensuring the general position.

Proof. Let $w \in \pi^{-1}\pi(\mathcal{T}(I))$. Since the right hand set of (9.8) is closed, we can assume without loss of generality that there exists $z' \in \mathcal{V}(I)$ and $u \in \text{span}\{u^{(1)}, \dots, u^{(l)}\}$ with $\text{ord } z' = w + u$. 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, $z' \in \mathcal{V}(J \cap \mathbb{K}[x])$. By Lemma 9.6.4, $w \in \mathcal{T}(J \cap \mathbb{K}[x])$ as well.

Let now $w \in \mathcal{T}(J \cap \mathbb{K}[x])$. Again we can assume that there is a $z \in \mathcal{V}(J \cap \mathbb{K}[x]) \subset (\mathbb{K}^*)^n$ with $w = \text{ord}(z)$. Since the projection is assumed to be sufficiently general, by the Extension Theorem, 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

$$z' := (z_1 \tilde{z}_{n+1}^{u_1^{(1)}} \cdots \tilde{z}_{n+l}^{u_l^{(1)}}, \dots, z_n \tilde{z}_{n+1}^{u_n^{(1)}} \cdots \tilde{z}_{n+l}^{u_l^{(1)}})$$

is a root of I . Then

$$\text{ord}(z') = \text{ord}(z) + \sum_{i=1}^l \text{ord}(\tilde{z}_{n+i})u^{(i)}$$

which means that $\text{ord}(z) = w \in \pi^{-1}\pi(\mathcal{T}(I))$. \square

Proof of Theorem 9.6.3 (Sketch). We restrict here to treat the case of sufficiently general projections. Then the first statement is clear from Theorem 9.6.6.

For the second statement, observe that the primality of I implies the primality of J and of $J \cap \mathbb{K}[x]$. Moreover,

$$\text{codim } J = \text{codim } I = n - m,$$

so that $\dim J = (n + (n + m - 1)) - (n - m) = n - 1$. Hence, for a sufficiently generic projection, $\dim(J \cap \mathbb{K}[x]) = n - 1$ as well. Hence, by Krull's Principal Ideal Theorem, $J \cap \mathbb{K}[x]$ is a principal ideal, and thus $\mathcal{T}(J \cap \mathbb{K}[x])$ is a tropical hypersurface. \square

We can now complete the proof of Theorem 9.6.1.

Proof of Theorem 9.6.1. Let $I \subset \mathbb{K}[x]$ be a prime ideal of dimension m . The case of a hypersurface, $m = n - 1$ is clear.

Now let $\text{codim } I = n - m > 1$. For sufficiently general projections $\pi : \mathbb{R}^n \rightarrow \mathbb{R}^{m+1}$, Theorem 9.6.6 implies that $\pi^{-1}\pi(\mathcal{T}(I))$ is a tropical hypersurface, the set $\pi(\mathcal{T}(I))$ provides a polyhedral complex of dimension at most $n - 1$. Since this holds for all sufficiently general projections $\pi : \mathbb{R}^n \rightarrow \mathbb{R}^{m+1}$, the set $\pi^{-1}(\mathcal{T}(I))$ defines a polyhedral complex as well. \square

The second part of the proofs of Theorems 9.6.1 and reftheo:tropbasis is pure combinatorial. To prepare for this, 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. We will show:

Corollary 9.6.7. *Let \mathcal{C} be a pure d -dimensional polyhedral complex in \mathbb{R}^n . Then there are $n - m + 1$ projections $\pi_0, \dots, \pi_{n-m} : \mathbb{R}^n \rightarrow \mathbb{R}^{d+1}$ such that*

$$\mathcal{C} = \bigcap_{i=0}^{n-m} \pi_i^{-1}(\pi_i(\mathcal{C})). \quad (9.9)$$

Proof. By construction, \mathcal{C} is contained in the right hand side. To show the converse, we observe that for sufficiently general hyperplanes H_0, \dots, H_k , the set $\mathbb{R}^n \setminus \bigcap_{i=0}^k H_i$ is of dimension $n - 1 - k$. Since for sufficiently general π , each set $\bigcap_{i=0}^n \pi_i^{-1}(\pi_i(\mathcal{C}))$ is a finite union of hyperplanes, a simple dimension count shows that the minimal number k^* of projections to achieve $\mathcal{C} = \bigcap_{i=0}^k \pi_i^{-1}(\pi_i(\mathcal{C}))$ is the minimal k satisfying $n - 1 - k < m - 1$. Hence, $k^* = n - m$. \square

The set of regular projections needed in Theorem 9.6.3 is dense in the space of projections. Hence, combining Proposition 9.6.1 with Theorem 9.6.6 yields Theorem 9.6.2. Note that by Lemma 9.6.5 the generators g_i are actually contained in I .

The following statement tells us that any ideal, even non-prime, can be written as the set-theoretic intersection of at most $n + 1$ tropical hypersurfaces. Note that here we do not require the polynomials of the hypersurfaces to generate the ideal.

Theorem 9.6.8. *Let $I \subset \mathbb{K}[x]$ be an arbitrary ideal. Then there exists $t \leq n$ and $g_0, \dots, g_t \in I$ with*

$$\mathcal{T}(I) = \bigcap_{i=0}^n \mathcal{T}(g_i) \quad (9.10)$$

Proof. First assume that I is radical. Then it is well-known that there exists a decomposition $I = P_1 \cap \dots \cap P_r$ with prime ideals P_i . By Theorem 9.6.2, for each i there exists $n_i \leq n$ and $g_{i,0}, \dots, g_{i,n_i}$ with $\mathcal{T}(P_i) = \bigcap_{j=0}^{n_i} \mathcal{T}(g_{i,j})$. Without loss of generality, we can assume that $n_i = n$ for all i . Since

$$\mathcal{T}(I) = \bigcup_{i=1}^r \mathcal{T}(P_i) = \bigcup_{i=1}^r \bigcap_{j=0}^n \mathcal{T}(g_{i,j}) = \bigcap_{j=0}^n \bigcup_{i=1}^r \mathcal{T}(g_{i,j}) = \bigcup_{j=0}^n \mathcal{T}(g_{1,j} \cdots g_{r,j}),$$

setting $g_j = g_{1,j} \cdots g_{r,j}$ proves the claim.

If I is not radical, then there exist $g_0, \dots, g_n \in \sqrt{I}$ with

$$\mathcal{T}(I) = \mathcal{T}(\sqrt{I}) = \bigcap_{j=0}^n \mathcal{T}(g_j),$$

and passing over to appropriate powers of g_j shows the claim. \square

Tropical linear spaces. Only in particular situations, tropical bases are explicitly known. Among those situations subclasses of ideals generated by linear forms. If $I \subset \mathbb{K}[x_1, \dots, x_n]$ is an ideal generated by (homogeneous) linear forms, then $\mathcal{T}(I)$ is called a *tropical linear space*.

Example 9.6.9. 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 $\mathbb{K}[x_1, x_2, x_3, x_4]$. A tropical basis of such an ideal I consists of four linear forms,

$$\begin{aligned} U = \{ & p_{12}(t) \cdot x_2 + p_{13}(t) \cdot x_3 + p_{14}(t) \cdot x_4, \\ & -p_{12}(t) \cdot x_1 + p_{23}(t) \cdot x_3 + p_{24}(t) \cdot x_4, \\ & -p_{13}(t) \cdot x_1 - p_{23}(t) \cdot x_2 + p_{34}(t) \cdot x_4, \\ & -p_{14}(t) \cdot x_1 - p_{24}(t) \cdot x_2 - p_{34}(t) \cdot x_3 \}, \end{aligned}$$

where the coefficients of the linear forms satisfy the *Grassmann-Plücker relation*

$$p_{12}(t) \cdot p_{34}(t) - p_{13}(t) \cdot p_{24}(t) + p_{14}(t) \cdot p_{23}(t) = 0, \quad (9.11)$$

which will be discussed in more detail in Chapter 10.

As explained in the following, if an ideal is generated by linear forms with coefficients in \mathbb{C} (rather than the more general $\mathbb{C}\{\{t\}\}$), then a tropical basis can be described rather explicitly (“constant coefficient case”). We identify two (homogeneous) linear forms in $\mathbb{C}[x]$, if their coefficient vectors agree as points in \mathbb{P}_C^n . The *support* $\text{supp}(\ell)$ of a linear form $\ell = \sum_i a_i x_i \in I$ is $\{i \in \{1, \dots, n\} : a_i \neq 0\}$. A *circuit* of I is a subset $C \subset \{1, \dots, n\} \setminus \{0\}$ such that $C = \text{supp}(\ell)$ for some linear form $\ell \in I$ and C is inclusion-minimal with this property.

Theorem 9.6.10. *Let $I \subset \mathbb{K}[x]$ be generated by linear forms, whose coefficients are contained in \mathbb{C} . Then the set of linear forms in $I \cap \mathbb{C}[x]$ whose supports are circuits of I constitute a tropical bases for I .*

First we establish the following connection between reduced Gröbner bases for I and the circuits of I .

Lemma 9.6.11. *Let $I \subset \mathbb{K}[x]$ be generated by linear forms, whose coefficients are contained in \mathbb{C} . Then the set of linear forms in $I \cap \mathbb{C}[x]$ whose supports are circuits of I is the union of all reduced Gröbner basis for $I \cap \mathbb{C}[x]$.*

We shortly write $I_{\mathbb{C}} = I \cap \mathbb{C}[x]$ and remind the reader that this is nothing else but the ideal I considered in the ring $\mathbb{C}[x]$.

Proof. Let $\ell = \sum_{i \in C} a_i x_i$, where $a_i \neq 0$ for $i \in C$, be a linear form in the reduced Gröbner basis G for $I_{\mathbb{C}}$ with respect to some term order \prec . If the support C of ℓ were not a circuit of I , there would be coefficients b_i , $i \in C$, not all zero, and some $j \in C$ with $b_j = 0$ such that $\sum_{i \in C} b_i x_i \in I_{\mathbb{C}}$. Let $s \in \{1, \dots, n\}$ such that $\text{in}_{\prec}(\ell) = x_s$, where the corresponding coefficient a_s is 1, because G is reduced. Also by the reducedness, there cannot be a linear form in $I_{\mathbb{C}}$ with support contained in $C \setminus \{l\}$. Hence, $s \neq j$. However, since $b_l \ell - \sum_{i \in C} b_i x_i \in I_{\mathbb{C}}$ and this polynomial does not depend on x_i , we obtain a contradiction to the precondition that ℓ comes from a reduced Gröbner basis.

Conversely, let C be a circuit of I . Let j be an arbitrary index in C , and let ℓ be the corresponding linear form in which the coefficient of x_j is 1. Choose a term order \prec such that $\text{in}_{\prec}(\ell) = x_j$. The linear form ℓ is an element of the Gröbner basis G of $I_{\mathbb{C}}$ with respect to \prec , since otherwise ℓ can be reduced by another element of G , which contradicts the property that C is a circuit of I . \square

Proof of Theorem 9.6.10. Let ℓ_1, \dots, ℓ_k be the linear forms in $I_{\mathbb{C}}$ whose supports are circuits of I . We have to show that $\mathcal{T}(I) = \bigcap_{i=1}^k \mathcal{T}(\ell_i)$. Since each ℓ_i is contained in I , the Fundamental Theorem 9.5.2 immediately implies $\mathcal{T}(I) \subset \bigcap_{i=1}^k \mathcal{T}(\ell_i)$.

Conversely, let $w \notin \mathcal{T}(I)$. Then, by the Fundamental Theorem 9.5.2 and Exercise 3, $t\text{-in}_w(I)$ and $\text{in}_w(I)$ contain a monomial. Let \prec be a term order which comes from the weight vector w and some tie-breaking rule. And let $G = \{g_1, \dots, g_t\}$ be a reduced Gröbner basis of $I_{\mathbb{C}}$ with respect to \prec . By Lemma 9.6.11, every polynomial in G is a linear form in $I_{\mathbb{C}}$ whose support is a circuit of I .

Since G is a Gröbner basis for $I_{\mathbb{C}}$ with respect to \prec , the set $\{\text{in}_w(g_1), \dots, \text{in}_w(g_t)\}$ is a Gröbner basis for $\text{in}_w(I_{\mathbb{C}})$. In particular, $\text{in}_w(I_{\mathbb{C}})$ is generated by linear forms.

Moreover, $\text{in}_w(I_{\mathbb{C}})$ is a prime ideal. To see this, assume there are polynomials $f, g \in I_{\mathbb{C}}$ with $f, g \notin I_{\mathbb{C}}$ with $fg \in I_{\mathbb{C}}$. Then there exist polynomials $h_i \in I_{\mathbb{C}}$ with $fg = \sum h_i \ell_i$. We can further assume that $\mathbb{C}[x]$ contains exactly the variables, which are present in f, g , or in the decomposition $fg = \sum h_i \ell_i$. Using linear combinations of the polynomials ℓ_i and linear variable transformations, we can reduce the situation to the case where $\ell_i = x_i$, for which the contradiction is apparent.

Since $\text{in}_w(I_{\mathbb{C}})$ is prime, the containment of a monomial implies the containment of a variable x_j . Hence, there is a polynomial $g_s \in G$ with $\text{in}_{\prec} g_s = x_j$. Since the Gröbner basis G is reduced, g_s must coincide with x_j . Hence, $w \notin \mathcal{T}(g_s) \subset \bigcap_{i=1}^k \mathcal{T}(\ell_i)$. \square

Exercises

1. Let $w \geq 0$ and G be a Gröbner basis of an ideal $I \subset \mathbb{C}[x]$ with respect to some term order \prec which refines the weighted order given by w . Then $\{\text{in}_w(g) : g \in G\}$ is a Gröbner basis for the initial ideal $\text{in}_w(I)$ with respect to \prec .
2. Show that the following formal genericity condition can be used within Theorem 9.6.6: For each $i \in \{1, \dots, l\}$ the elimination ideal $J \cap \mathbb{K}[x_1, \dots, x_n, \lambda_1, \dots, \lambda_i]$ 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 $x_1, \dots, x_n, \lambda_1, \dots, \lambda_{i-1}$.

9.7 Notes

For an introductory text to tropical geometry covering also some aspects not treated here, we refer to Richter-Gebert, Sturmfels and Theobald [114]. A comprehensive treatment of tropical geometry can be found in the book of Maclagan and Sturmfels [86]. For the dequantization see Maslov [89] and Viro's work on patchworking [141]. The tropical Bernstein theorem was first stated by Sturmfels [138], see also [11, 114, 138]. The stable intersection was introduced in [114].

For the notion of intersection multiplicity there are various equivalent approaches, see [11, 71, 92, 139].

A proof of the Newton-Puiseux Theorem can be found, for example, in [143]. For a precise treatment of variants of Puiseux series with real exponents see Markwig [87]. Kapranov's Theorem is contained in [34], our presentation is based on Aroca's exposition [2].

The Fundamental Theorem of Tropical Geometry was presented by Speyer and Sturmfels [134], however, their proof was not complete. A complete proof was given by Draisma [33], our presentation is based on the ones by Jensen [67] and Jensen, Markwig and Markwig [68]. We also refer to these papers as well as to Payne [98, 99] and Maclagan and Sturmfels [86] for the case of a positive dimensional ideal.

Our discussion of tropical bases is mostly based on Hept and Theobald [57], the treatment of linear spaces follows [86].

Chapter 10

Non-Linear Computational Geometry

Methods and ideas from algebraic geometry can be fruitfully applied to solve problems in the ordinary geometry of three- and n -dimensional space. To illustrate this, we consider some concepts and problems from non-linear computational geometry. First we study in detail the Grassmann variety, which serves to parameterize k -dimensional linear or affine subspaces in n -dimensional real or complex space. Then we illustrate some natural occurrences of the Grassmannian, by considering the geometric problem of lines which are tangent to four general spheres. We shall see that this seemingly elementary problem contains quite a lot of interesting geometry, which can also be used to illustrate some of the techniques developed in earlier chapters.

Then, as an outlook on the use of algebraic geometry in kinematics and robotics, we consider the Stewart platform.

10.1 Grassmann varieties and Plücker coordinates

In many geometric applications in real space, such as in computer graphics or computer vision, lines in real space \mathbb{R}^3 or \mathbb{R}^n are of prominent importance. For example, a point a can be seen from a point b if there is no line segment from a to b intersecting any other object in the scene.

While a line is an affine subspace of the ambient space, intersection conditions with other lines are inherently nonlinear. To illustrate this, consider the problem to determine the lines intersecting four given lines $\ell_1, \dots, \ell_4 \subseteq \mathbb{R}^3$. If the problem were of linear or of affine-linear nature, then one would always obtain either 0, 1 or infinitely many solutions. As we will see, for lines in general position, there will always be two (in general complex) lines with this property.

In this section, we explain the nonlinear geometry behind these questions, which is provided by Grassmann varieties and Plücker coordinates. We will need some multilinear algebra, see Appendix A.1.4.

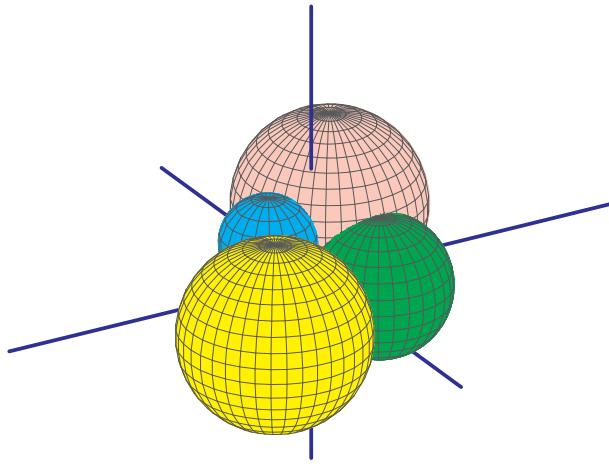


Figure 10.1: Four spheres with centers $(0, 1, 1)^T$, $(-1, -1, 0)^T$, $(1, -1, 1)^T$, $(-2, 2, 0)^T$, and respective radii 1, $3/2$, 2, and 2.

We abbreviate k -dimensional linear subspaces as k -*planes*. Let U be a k -plane of \mathbb{C}^n with a basis $\{u^{(1)}, \dots, u^{(k)}\}$. Setting $N = \binom{n}{k} - 1$, we associate with U the point in the projective N -space given by the k -th exterior power $u^{(1)} \wedge \dots \wedge u^{(k)}$. Considering the equivalence class $[u^{(1)} \wedge \dots \wedge u^{(k)}]$ in projective space yields the map

$$\begin{aligned} \text{Pl} : \{k\text{-planes in } \mathbb{C}^n\} &\longrightarrow \mathbb{P}^N \cong \mathbb{P}(\bigwedge^k \mathbb{C}^n), \\ U &\longmapsto [u^{(1)} \wedge \dots \wedge u^{(k)}] \end{aligned} \quad (10.1)$$

called the *Plücker embedding*. $\mathbb{P}(\bigwedge^k \mathbb{C}^n)$ is called *Plücker space*, and the image of a subspace U under the Plücker embedding are called the *Plücker coordinate* of U . Here, recall from multilinear algebra, that the map (10.1) is well-defined, since two different bases $\{u^{(1)}, \dots, u^{(k)}\}$ and $\{v^{(1)}, \dots, v^{(k)}\}$ of the same k -plane are related by $(u^{(1)}, \dots, u^{(k)}) = (v^{(1)}, \dots, v^{(k)})A$ with a regular matrix $A = (a_{ij})_{k \times k}$ and thus satisfy

$$u^{(1)} \wedge \dots \wedge u^{(k)} = \det(A) \cdot v^{(1)} \wedge \dots \wedge v^{(k)}.$$

Example 10.1.1. For $n = 4$, the unit vectors $e^{(0)}, \dots, e^{(3)}$ for \mathbb{C}^4 give the basis

$$e^{(0)} \wedge e^{(1)}, \quad e^{(0)} \wedge e^{(2)}, \quad e^{(0)} \wedge e^{(3)}, \quad e^{(1)} \wedge e^{(2)}, \quad e^{(1)} \wedge e^{(3)}, \quad e^{(2)} \wedge e^{(3)}$$

for $\bigwedge^2 \mathbb{C}^4$, showing that the Plücker space $\mathbb{P}(\bigwedge^2 \mathbb{C}^4)$ is a five-dimensional projective space. Note that here and in the following, the numbering of the indices begins with zero, in order to support also well the transition from \mathbb{C}^n to \mathbb{P}^{n-1} . If

$$u = u_0 e^{(0)} + u_1 e^{(1)} + u_2 e^{(2)} + u_3 e^{(3)} \quad \text{and} \quad v = v_0 e^{(0)} + v_1 e^{(1)} + v_2 e^{(2)} + v_3 e^{(3)}$$

are vectors in \mathbb{C}^4 , their exterior product $u \wedge v$ is

$$\begin{aligned} & (u_0 v_1 - u_1 v_0) e^{(0)} \wedge e^{(1)} + (u_0 v_2 - u_2 v_0) e^{(0)} \wedge e^{(2)} + (u_0 v_3 - u_3 v_0) e^{(0)} \wedge e^{(3)} \\ & + (u_1 v_2 - u_2 v_1) e^{(1)} \wedge e^{(2)} + (u_1 v_3 - u_3 v_1) e^{(1)} \wedge e^{(3)} + (u_2 v_3 - u_3 v_2) e^{(2)} \wedge e^{(3)}. \end{aligned}$$

A tensor in $\bigwedge^k \mathbb{C}^n$ is *decomposable* if it has the form $v^{(1)} \wedge \cdots \wedge v^{(k)}$ for some vectors $v^{(1)}, \dots, v^{(k)} \in \mathbb{C}^n$. The image of the Plücker embedding consists exactly of the decomposable vectors in $\bigwedge^k \mathbb{C}^n$. And the coordinates of a decomposable tensor $v^{(1)} \wedge \cdots \wedge v^{(k)}$ are nothing else but the determinants

$$p_I := \det \begin{pmatrix} v_{i_1}^{(1)} & \cdots & v_{i_1}^{(k)} \\ \vdots & \ddots & \vdots \\ v_{i_k}^{(1)} & \cdots & v_{i_k}^{(k)} \end{pmatrix}, \quad I = \{i_1, \dots, i_k\} \text{ with } 0 \leq i_1 < \cdots < i_k < n. \quad (10.2)$$

The coordinates $(p_I)_{I \subset [k]}$ of this decomposable tensor give the Plücker coordinates of the k -plane $\langle v^{(1)}, \dots, v^{(k)} \rangle$. We also write $(p_{i_1 \dots i_k})_{0 < i_1 < \dots < i_k < n}$.

The next lemma shows that the Plücker embedding is injective.

Lemma 10.1.2. *If $u^{(1)} \wedge \cdots \wedge u^{(k)}$ and $v^{(1)} \wedge \cdots \wedge v^{(k)}$ are two decomposable tensors representing the same point in $\mathbb{P}(\bigwedge^k \mathbb{C}^n)$, then they come from the same k -plane in \mathbb{C}^n . That is, $\langle u^{(1)}, \dots, u^{(k)} \rangle = \langle v^{(1)}, \dots, v^{(k)} \rangle$.*

Proof. Let $U = \langle u^{(1)}, \dots, u^{(k)} \rangle$ and $V = \langle v^{(1)}, \dots, v^{(k)} \rangle$. Since $u^{(1)}, \dots, u^{(k)}$ are linearly independent, for any $u \in U$ the wedge product $u^{(1)} \wedge \cdots \wedge u^{(k)} \wedge v$ is zero if and only if $v \in U$. Using the same argument for V , we can conclude that $v \in U$ if and only if $v \in V$. \square

Example 10.1.3. In the case $k = 2$ and $n = 2$, observe that $p_{03}p_{12}$ equals

$$(u_0v_3 - u_3v_0)(u_1v_2 - u_2v_1) = \color{red}{u_0u_1v_2v_3} - u_0u_2v_1v_3 - u_1u_3v_0v_2 + \color{purple}{u_2u_3v_0v_1}.$$

Another product which has the same leading term, $\color{red}{u_0u_1v_2v_3}$, in the lexicographic monomial order where $u_0 > u_1 > \cdots > v_2 > v_3$ is $p_{02}p_{13}$, which is

$$(u_0v_2 - u_2v_0)(u_1v_3 - u_3v_1) = \color{red}{u_0u_1v_2v_3} - u_0u_3v_1v_2 - u_1u_2v_0v_3 + \color{purple}{u_2u_3v_0v_1}.$$

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*

$$p_{03}p_{12} - p_{02}p_{13} + p_{01}p_{23} = 0, \quad (10.3)$$

which holds on the Plücker coordinates of decomposable tensors. Exercise 1 will show that any p_{01}, \dots, p_{23} satisfying this relation is a Plücker coordinate of a 2-plane in \mathbb{C}^4 , and this property will be substantially generalized in the subsequent section.

Definition 10.1.4. The *Grassmannian of k -planes in \mathbb{C}^n* , $G(k, n)$, is the image of the set of k -planes in \mathbb{C}^n under the Plücker embedding.

Equivalently, $G(k, n)$ is the set of decomposable vectors in $\mathbb{P}(\bigwedge^k \mathbb{C}^n)$. It will be a main goal of the next section to determine the equations, which characterize the Grassmannian as a subvariety of \mathbb{P}^N .

Exercises

1. Show that if $[p_{01}, \dots, p_{23}]$ is the homogeneous coordinate 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 .
2. The cylinder $x^2 + y^2 = 1$ in \mathbb{R}^3 is generated by a ruling of lines. Determine this curve in Plücker coordinates.
3. Determine the curves of the two rulings of lines of a hyperboloid $x^2 + y^2 - z^2 = 1$ in \mathbb{R}^3 in Plücker coordinates.

10.2 Duality and the Plücker relations

In the definition of the Plücker embedding, we employed used a representation of a k -plane $U \subset \mathbb{C}^n$ as the span of k linearly independent vectors. If one starts from a representation of U as the intersection of $n-k$ hyperplanes, then one arrives at the dual Plücker embedding, which is defined as follows.

Let U be a k -dimensional subspace in \mathbb{C}^n , given as the intersection of $n-k$ hyperplanes

$$\sum_{i=1}^n w_i^{(1)} x_i = 0, \dots, \sum_{i=1}^n w_i^{(n-k)} x_i = 0,$$

with coefficient vectors $w^{(1)}, \dots, w^{(n-k)}$. The dual basis to the standard basis of \mathbb{C} is $(e^{(1)})^*, \dots, (e^{(n)})^*$, which consists of the linear forms $x \mapsto x_1, \dots, x \mapsto x_n$. We consider the exterior algebra $\wedge^k (\mathbb{C}^n)^*$ on the dual vector space $(\mathbb{C}^n)^*$. The dual Plücker embedding of U is the map

$$\begin{aligned} \{k\text{-planes in } \mathbb{C}^n\} &\longrightarrow (\mathbb{P}^N)^* \cong \mathbb{P}(\wedge^k (\mathbb{C}^n)^*), \\ U &\longmapsto [(w^{(1)})^* \wedge \cdots \wedge (w^{(k)})^*]. \end{aligned}$$

The coordinates of the dual Plücker embedding are called *dual Plücker coordinates*. For $k = n-1$, the dual Plücker coordinates are exactly the homogeneous coordinates of the hyperplanes in \mathbb{P}^{n-1} .

In order to capture the relationship between the primal Plücker coordinates and the dual ones, we define a linear operator $* : \wedge^k \mathbb{C}^n \rightarrow \wedge^{n-k} (\mathbb{C}^n)^*$. The exterior power $\wedge^n \mathbb{C}^n$ is a one-dimensional space spanned by $e^{(1)} \wedge \cdots \wedge e^{(n)}$. Hence, for given $a \in \wedge^k \mathbb{C}^n$, the linear form

$$\phi_a : \wedge^{n-k} \mathbb{C}^n \rightarrow \wedge^n \mathbb{C}^n \cong \mathbb{C}, \quad x \mapsto a \wedge x$$

can be uniquely written in the form

$$(*a(x)) \cdot (e^{(1)} \wedge \cdots \wedge e^{(n)}) \tag{10.4}$$

with some $*a$ in the dual space $(\wedge^k (\mathbb{C}^n))^* \cong \wedge^k (\mathbb{C}^n)^*$. Clearly, the resulting operator $*$ is linear.

Example 10.2.1. For $n = 4$, $k = 2$, the $*$ -operator gives $*(1) = e^{(1234)}$, $*(e^{(1234)}) = 1$ as well as

$$\begin{aligned} *e^{(1)} &= (e^{(234)})^*, & *e^{(12)} &= (e^{(34)})^*, & *e^{(123)} &= (e^{(4)})^*, \\ *e^{(2)} &= -(e^{(134)})^*, & *e^{(13)} &= -(e^{(24)})^*, & *e^{(124)} &= -(e^{(3)})^*, \\ *e^{(3)} &= (e^{(124)})^*, & *e^{(14)} &= (e^{(23)})^*, & *e^{(134)} &= (e^{(2)})^*, \\ *e^{(4)} &= -(e^{(123)})^*, & *e^{(23)} &= (e^{(14)})^*, & *e^{(234)} &= -(e^{(1)})^*, \\ && *e^{(24)} &= -(e^{(13)})^*, && \\ && *e^{(34)} &= (e^{(12)})^*. && \end{aligned}$$

We show that the dual Plücker embedding of a k -subspace U coincides with the image of the primal Plücker embedding under the $*$ -operator.

Theorem 10.2.2. *Given a k -dimensional subspace U of \mathbb{C}^n , its primal and dual Plücker embeddings $p \in \mathbb{P}(\bigwedge^k(\mathbb{C}^n))$ and $q \in \mathbb{P}(\bigwedge^k(\mathbb{C}^{n-k})^*)$ satisfy $*p = q$.*

In the proof, we will use the following determinantal identity of Jacobi.

Lemma 10.2.3. *Let $A \in \mathbb{C}^{n \times n}$ be invertible and of the form*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad B := A^{-1} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

with $k \times k$ -matrices A_{11} , B_{11} . Then

$$\det B_{22} \cdot \det A = \det A_{11}.$$

Proof. Since $A \cdot A^{-1} = \text{Id}$, we have

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \cdot \begin{pmatrix} \text{Id} & B_{12} \\ 0 & B_{22} \end{pmatrix} = \begin{pmatrix} A_{11} & 0 \\ A_{12} & \text{Id} \end{pmatrix}.$$

Taking the determinant on both sides immediately shows the claim. \square

Proof of Theorem 10.2.2. We begin with the special case of the k -plane defined by $x_{k+1} = \dots = x_n = 0$. This subspace is spanned by the unit vectors $e^{(1)}, \dots, e^{(k)}$, and the primal Plücker embedding is $p = e^{(1)} \wedge \dots \wedge e^{(k)}$ considered as a point in \mathbb{P}^N . In order to determine $*p$, first observe that for basis elements $x = e^{(j_1)} \wedge \dots \wedge e^{(j_{n-k})}$, the right hand side of condition (10.4) is zero whenever $\{j_1, \dots, j_{n-k}\} \cap \{1, \dots, k\} \neq \emptyset$, and hence these coefficients of $(e^{(j_{n-k})})^*$ in $*p$ must be zero. In the remaining case, we can assume that j_1, \dots, j_k are ordered, so that $(j_1, \dots, j_{n-k}) = (k+1, \dots, n)$. This shows $*p = e^{(k+1)} \wedge \dots \wedge e^{(n)}$, and this is clearly the dual Plücker embedding of the k -plane.

For the general case, we assume that the k -plane U results from a regular linear transformation from the special k -plane. By (10.2), the primal and similarly the dual Plücker coordinates can be expressed as determinants. Hence, Jacobi's determinantal

identity from Lemma 10.2.3 and the linearity of the star operator imply that the property $*p = q$ is preserved under the linear transformation. \square

For a fixed $\omega \in \bigwedge^k \mathbb{C}^n$, consider the linear map

$$\begin{aligned}\wedge_\omega : \mathbb{C}^n &\rightarrow \bigwedge^{k+1} \mathbb{C}^n, \\ v &\mapsto v \wedge \omega.\end{aligned}$$

Lemma 10.2.4. *For $\omega \in \bigwedge^k \mathbb{C}^n \setminus \{0\}$ the following statements are equivalent.*

1. *There exist $v^{(1)}, \dots, v^{(k)} \in \mathbb{C}^n$ with $\omega = v^{(1)} \wedge \dots \wedge v^{(k)}$.*
2. $\dim \ker \wedge_\omega = k$.

Proof. Recall that for linearly independent $v^{(1)}, \dots, v^{(k)} \in \mathbb{C}^n$, a vector v gives $v \wedge v^{(1)} \wedge \dots \wedge v^{(k)} = 0$ if and only if $v \in \langle v^{(1)}, \dots, v^{(k)} \rangle$.

If $0 \neq \omega = v^{(1)} \wedge \dots \wedge v^{(k)}$, then the vectors $v^{(1)}, \dots, v^{(k)}$ are linearly independent, and we have $\ker \wedge_\omega = \langle v^{(1)}, \dots, v^{(k)} \rangle$, that is, $\dim \ker \wedge_\omega = k$.

Conversely, let $v^{(1)}, \dots, v^{(n)}$ be a basis of \mathbb{C}^n , such that the first k vectors $v^{(1)}, \dots, v^{(k)}$ form a basis of the kernel of \wedge_ω . The set of vectors $v^{(I)} = v^{(i_1)} \wedge \dots \wedge v^{(i_k)}$ with $I = \{i_1, \dots, i_k\}$ und $1 \leq i_1 < \dots < i_k \leq n$ is a basis for $\bigwedge^k \mathbb{C}^n$. Hence, there exists a unique representation of ω as linear combination of these basis vectors

$$\omega = \sum_I \omega_I v^{(I)}$$

with coefficients $\omega_I \in \mathbb{C}$. For each $i \in \{1, \dots, k\}$, the construction implies $v^{(i)} \wedge \omega = 0$, and hence, by definition of the exterior product, all ω_I with $i \notin I$ vanish. Consequently, only the coefficient $\omega_{\{1, \dots, k\}}$ can be non-zero. \square

We obtain the immediate corollary.

Corollary 10.2.5. *For $\omega \in \bigwedge^k \mathbb{C}^n \setminus \{0\}$, we have*

$$\dim \ker \wedge_\omega = k \iff \dim \ker \wedge_\omega \geq k.$$

Example 10.2.6. Choosing the canonical bases for \mathbb{C}^n and $\bigwedge^{k+1} \mathbb{C}^n$ in lexicographical order, we obtain the corresponding representation matrix

$$M_\omega \in \mathbb{C}^{\binom{n}{k+1} \times n}.$$

Recall that the columns of the representation matrix M_ω of \wedge_ω contain the coordinate vectors of the image of the canonical basis vectors. For the order $e^{(1)}, \dots, e^{(4)}$, the canonical

basis vectors of \mathbb{C}^n and the order $e^{(123)}, e^{(124)}, e^{(134)}, e^{(234)}$ of the canonical basis vectors of $\Lambda^3 \mathbb{C}^n$ gives the representation matrix M_ω of \wedge_ω ,

$$M_\omega = \begin{pmatrix} p_{23} & -p_{13} & p_{12} & 0 \\ p_{24} & -p_{14} & 0 & p_{12} \\ p_{34} & 0 & -p_{14} & p_{13} \\ 0 & p_{34} & -p_{24} & p_{23} \end{pmatrix}.$$

By Lemma 10.2.4, the vector ω defines a Plücker vector of a line in \mathbb{P}^3 if and only if this matrix has rank 2.

Similar to \wedge_ω , we define the mapping

$$\begin{aligned} \wedge_{*\omega} : (\mathbb{C}^n)^* &\rightarrow \wedge^{n-k+1}(\mathbb{C}^n)^*, \\ \phi &\mapsto \phi \wedge *\omega. \end{aligned}$$

The representation matrix of this mapping (with respect to lexicographically ordered canonical bases) is denoted by $M_\omega^* \in K^{\binom{n}{n-k+1} \times n}$.

Example 10.2.7. In case $n = 4, k = 2$, considering $\omega = \sum_{1 \leq i < j \leq 4} p_{ij}(e_i \wedge e_j)$ gives

$$*\omega = p_{12}(e^{(34)})^* - p_{13}(e^{(24)})^* + p_{14}(e^{(23)})^* + p_{23}(e^{(14)})^* - p_{24}(e^{(13)})^* + p_{34}(e^{(12)})^*.$$

This implies, for instance,

$$(e^{(1)})^* \wedge *\omega = p_{14}(e^{(123)})^* - p_{13}(e^{(124)})^* + p_{12}(e^{(134)})^*,$$

which gives in particular the first column of the representation matrix

$$M_\omega^* = \begin{pmatrix} p_{14} & p_{24} & p_{34} & 0 \\ -p_{13} & -p_{23} & 0 & p_{34} \\ p_{12} & 0 & -p_{23} & -p_{24} \\ 0 & p_{12} & p_{13} & p_{14} \end{pmatrix}.$$

Theorem 10.2.8. A vector $v \in \Lambda^k \mathbb{C}^n \setminus \{0\}$ is decomposable if and only if

$$\ker \wedge_{*\omega} = (\ker \wedge_\omega)^\circ, \tag{10.5}$$

where $(\ker \wedge_\omega)^\circ \subset (\mathbb{C}^n)^*$ denotes the annihilator of the kernel of \wedge_ω .

Proof. Theorem 10.2.2 implies that ω is the Plücker embedding of some of a k -plane if and only if $*\omega$ is a dual Plücker embedding of some k -plane. Hence, in this case there exists a basis $v^{(1)}, \dots, v^{(n)}$ of \mathbb{C}^n with

$$\omega = v^{(1)} \wedge \cdots \wedge v^{(k)} \quad \text{und} \quad *\omega = v^{(k+1)} \wedge \cdots \wedge v^{(n)}.$$

Thus, for each $v \in \mathbb{C}^n$ and each $u \in (\mathbb{C}^n)^*$, the linear form $u \wedge * \omega$ vanishes on $v \wedge \omega$, which shows the claim.

Now consider the converse direction. By Lemma 10.2.4 and Remark 10.2.5, for any $\omega \in \bigwedge^k \mathbb{C}^n \setminus \{0\}$ we have $\dim \ker \wedge_\omega \leq k$ und similarly $\dim \ker \wedge_{*\omega} \leq n - k$. Hence, if (10.5) is satisfied, then in equality must hold within both inequalities. Lemma 10.2.4 then implies that ω is the Plücker vector of some k -plane. \square

Corollary 10.2.9. *A vector $v \in \bigwedge^k \mathbb{C}^n \setminus \{0\}$ is decomposable if and only if*

$$M_\omega \cdot (M_\omega^*)^T = 0. \quad (10.6)$$

Proof. The condition (10.5) is equivalent $\text{im } \wedge_{*\omega} = (\text{im } \wedge_\omega)^\circ$. In terms of the representing matrices, this means that the dot products of the rows of M_ω and of M_ω^* vanish. \square

As a consequence we obtain the desired characterization of those points in $p \in \mathbb{P}^N$, which are the Plücker coordinates of some k -dimensional subspace of \mathbb{C}^n . Corollary 10.2.9 gives quadratic conditions. To write them down in coordinates, see Exercises 3 and 4.

For the case case $k = 2$, that is, for the case of lines in projective space, we obtain the following corollary.

Corollary 10.2.10. *The Plücker coordinates of a line ℓ in \mathbb{P}^{n-1} satisfies the conditions*

$$p_{ij}p_{rs} - p_{ir}p_{js} + p_{is}p_{jr} = 0 \quad \text{for } 1 \leq i < j < r < s \leq n.$$

For $n = 4$ these conditions specializes further to a single quadratic equation

$$p_{12}p_{34} - p_{13}p_{24} + p_{14}p_{23} = 0. \quad (10.7)$$

The quadric in \mathbb{P}^5 defined by (10.7) is called *Klein's quadric*. Wir summarize the statements of this section as follows.

Corollary 10.2.11. *The image of the set of k -planes in \mathbb{C}^n under the Plücker embedding is exactly the set given by the conditions (10.6). Hence, the Plücker embedding bijectively maps the set of k -planes in \mathbb{C}^n to that image.*

Computing with Plücker coordinates. Using Plücker coordinates in primal and in dual form, intersections of subspaces can be computed comfortably. We can identify projective subspace of \mathbb{P}^{n-1} with linear subspaces of \mathbb{C}^n .

Theorem 10.2.12. *A $(k-1)$ -dimensional projective subspace U of \mathbb{P}^{n-1} intersects an $(n-k-1)$ -dimensional projective subspace W of \mathbb{P}^{n-1} if and only if the Plücker embedding $p \in \mathbb{P}(\bigwedge^k \mathbb{C}^n)$ of U and the dual Plücker coordinate $q \in \mathbb{P}(\bigwedge^k (\mathbb{C}^n)^*)$ of W satisfy $q(p) = 0$.*

Proof. Using the identification of projective subspaces of \mathbb{P}^{n-1} with linear subspaces of \mathbb{C}^n , let $u^{(1)}, \dots, u^{(k)}$ be a basis of the linear subspace U of \mathbb{C}^n and $w^{(1)}, \dots, w^{(k)}$ be the coefficient vectors of the equations for the linear subspace W . A point $\sum_{i=1}^k \lambda_i u^{(i)} \in U$ with coefficients $\lambda_1, \dots, \lambda_k$ is contained in W if and only if

$$\sum_{i=1}^k \sum_{l=1}^k \lambda_i u_l^{(i)} w_l^{(j)} = 0, \quad \text{for all } j \in \{1, \dots, k\}.$$

This system of equations has a non-trivial solution in $\lambda_1, \dots, \lambda_k$ if and only if

$$\det \left(\sum_{l=1}^n u_l^{(i)} w_l^{(j)} \right)_{\substack{1 \leq i \leq k \\ 1 \leq j \leq k}} = 0. \quad (10.8)$$

This determinant can also be interpreted as the determinant of the product of the matrices $(u_l^{(i)})_{l,i}$ und $(w_l^{(j)})_{j,l}$.

By the Cauchy-Binet multiplication formula for determinants (cf. Appendix A.1.4), for any two matrices $A \in \mathbb{K}^{n \times k}$, $B \in \mathbb{K}^{k \times n}$ over a field \mathbb{K} we have the property

$$\det AB = \sum_{I \subseteq \{1, \dots, n\}, |I|=k} \det A_I \det B_I,$$

where A_I and B_I are the submatrices of A and B , which has only the columns in A and rows of B which indexed by I . Using the Cauchy-Binet formula, (10.8) can also be written as $q(p) = 0$, which implies the claim. \square

For lines in the three-dimensional projective space \mathbb{P}^3 , this intersection condition specializes as follows.

Corollary 10.2.13. *A line ℓ intersects with a line ℓ' in \mathbb{P}^3 if and only if their Plücker coordinates p and p' satisfy*

$$p_{12}p'_{34} - p_{13}p'_{24} + p_{14}p'_{23} + p_{23}p'_{14} - p_{24}p'_{13} + p_{34}p'_{12} = 0. \quad (10.9)$$

With the elimination techniques developed in Section 2, it is possible to compute the Plücker relation from (10.7) comfortably via a computer algebra system. For this, consider the polynomials $p_{ij} = x_i y_j - x_j y_i$ in the polynomial ring over \mathbb{C} with variables $x_1, \dots, x_4, y_1, \dots, y_4$ and p_{ij} , $1 \leq i < j \leq 4$. Using the lexicographic ordering

$$x_1 > \dots > x_4 > y_1 > \dots > y_4 > p_{12} > p_{13} > \dots > p_{34},$$

eliminating all x - and y -variables from the ideal I generated by the polynomials

$$p_{ij} - (x_i y_j - x_j y_i), \quad 1 \leq i < j \leq 4,$$

gives the polynomial $p_{12}p_{34} - p_{13}p_{24} + p_{14}p_{23}$ Plücker relation. The lexicographic Gröbner basis for the ideal I consists of 17 polynomials. In consistency with Theorem 2.4.8 for the elimination, one of these polynomials is the polynomial of the Plücker relation.

Schubert calculus. We close our discussion on the Grassmannian of k -planes in general n -dimensional space by mentioning the viewpoint 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 $1 \leq k \leq n - 2$ there are $d_{k,n} := (k+1)(n-k)$ real $(n-k-1)$ -planes $U_1, \dots, U_{d_{k,n}}$ 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 $U_1, \dots, U_{d_{k,n}}$. 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, which has already been mentioned in the introduction the current chapter. 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 ℓ_1, \dots, ℓ_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 10.2). This is either (i) a hyperboloid of

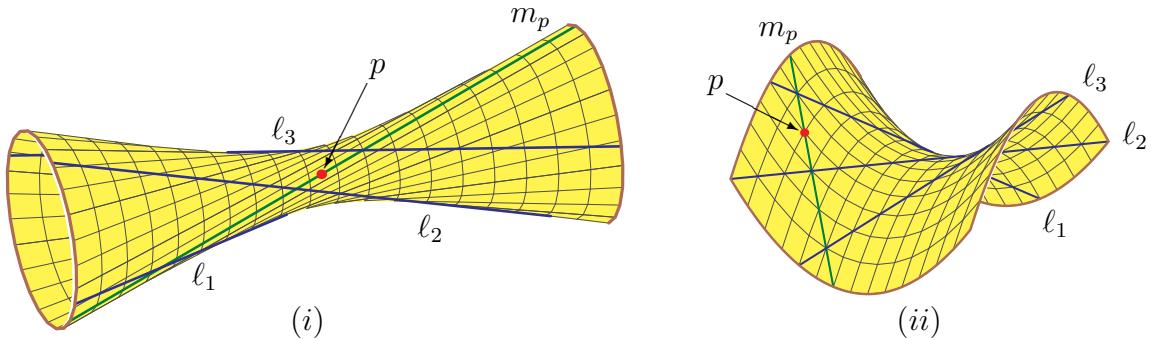


Figure 10.2: Hyperboloids through 3 lines.

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.

Exercises

1. Show that the star operator gives

$$*(e^{(j_1)} \wedge \cdots \wedge e^{(j_n)}) = \text{sgn}(i_1, \dots, i_k, j_1, \dots, j_{n-k}) \cdot (e^{(j_1)})^* \wedge \cdots \wedge (e^{(j_{n-k})})^*$$

where $I = \{i_1, \dots, i_k\}$ with $1 \leq i_1 < \cdots < i_k \leq n$ and $J = \{j_1, \dots, j_{n-k}\} := \{1, \dots, n\} \setminus I$ with increasing indices $j_1 < \cdots < j_{n-k}$ is the complement of I .

2. Show that for $a \in \bigwedge^k \mathbb{C}^n$, we have $*(*(\omega)) = (-1)^{k(n-k)}a$.

3. For the representation matrix $M_\omega \in K^{\binom{n}{k+1} \times n}$ of the map \wedge_ω with respect to the canonical bases in lexicographical order, show that

$$(M_\omega)_{Ij} = \begin{cases} 0 & \text{if } j \notin I, \\ \epsilon p_{I \setminus \{j\}} & \text{if } j \in I, \end{cases}$$

where $I \setminus \{j\} = \{i_1, \dots, i_k\}$ with $i_1 \leq \cdots \leq i_k$ and $\epsilon = \text{sgn}(j, i_1, \dots, i_k)$. And for $\wedge_{*(\omega)}$ show that

$$(M_\omega^*)_{I'j} = \begin{cases} 0 & \text{if } j \notin I', \\ \epsilon' p_J & \text{if } j \in I' \end{cases}$$

with $I' = \{i'_1, \dots, i'_{n-k+1}\}$, $i'_1 \leq \cdots \leq i'_{n-k+1}$, $J = \{1, \dots, n\} \setminus I' = \{j_1, \dots, j_{k-1}\}$, $j_1 \leq \cdots \leq j_{k-1}$ and $\epsilon' = \text{sgn}(i'_1, \dots, i'_{n-k+1}, j_1, \dots, j_{k-1})$.

4. Conclude that $v \in \mathbb{P}(\bigwedge^k \mathbb{C}^n)$ is decomposable if and only if for all $i_1, \dots, i_{k+1}, j_1, \dots, j_{k-1} \in \{0, \dots, n\}$ it satisfies

$$\sum_{l=1}^{k+1} (-1)^l p_{i_1, \dots, \hat{i}_l, \dots, i_{k+1}} p_{j_1, \dots, j_{k-1}, i_l} = 0, \quad (10.10)$$

where \hat{i}_l means that this index is omitted.

10.3 Case study: Lines tangent to spheres

Now we illustrate some natural occurrences of the Grassmannian, by considering the following geometric problem:

How many lines are tangent to four general spheres?

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 .

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

$$Q = \begin{bmatrix} a^2 + b^2 + c^2 - r^2 & -a & -b & -c \\ -a & 1 & 0 & 0 \\ -b & 0 & 1 & 0 \\ -c & 0 & 0 & 1 \end{bmatrix}. \quad (10.11)$$

Indeed, $X^T Q X$ is $(x_1 - ax_0)^2 + (x_2 - bx_0)^2 + (x_3 - cx_0)^2 - r^2 x_0^2$. 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 1). 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

$$V \hookrightarrow \mathbb{C}^4 \xrightarrow{Q} (\mathbb{C}^4)^\vee \twoheadrightarrow V^\vee, \quad (10.12)$$

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 (10.12) 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 $0 \leq i < j \leq 3$, 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 (10.11), this is

$$\wedge^2 Q = \begin{pmatrix} b^2 + c^2 - r^2 & -ab & -ac & b & c & 0 \\ -ab & a^2 + c^2 - r^2 & -bc & -a & 0 & c \\ -ac & -bc & a^2 + b^2 - r^2 & 0 & -a & -b \\ b & -a & 0 & 1 & 0 & 0 \\ c & 0 & -a & 0 & 1 & 0 \\ 0 & c & -b & 0 & 0 & 1 \end{pmatrix} \begin{matrix} 01 \\ 02 \\ 03 \\ 12 \\ 13 \\ 23 \end{matrix} \quad (10.13)$$

We remark that there is nothing special about spheres in this discussion.

Theorem 10.3.1. *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 $\wedge^2 Q$.*

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 (10.3), 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 [44] 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 $u > v > \dots > z$. 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 P[6][1] = u,v,w,x,y,z;
```

We give a procedure to compute $\wedge^2 Q$ (10.13),

```

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 $(0, 1, 1)^T$, $(-1, -1, 0)^T$, $(1, -1, 1)^T$, and $(-2, 2, 0)^T$.

```

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-known 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 ,

$$(x_1 - ax_0)^2 + (x_2 - bx_0)^2 = r^2 x_0^2.$$

For the same reason, spheres are the quadrics in \mathbb{P}^3 which contain the imaginary *circular conic at infinity*, which is defined by

$$x_0 = 0 \quad \text{and} \quad x_1^2 + x_2^2 + x_3^2 = 0.$$

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 (10.13), becomes

$$p_{12}^2 + p_{13}^2 + p_{23}^2 = 0. \quad (10.14)$$

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 (10.14), 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 1.4.11, the saturation $(I : J^\infty)$ 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 $(I : J)$. 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
```

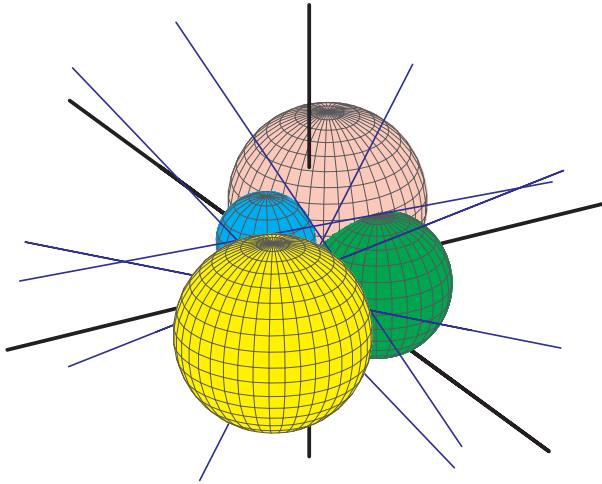
While the degree of $(I : J)$ is less than that of I , it is still 1-dimensional, so we take the quotient ideal again.

```
I = std(quotient(I,J));
degree(I);
// dimension (proj.) = 0
// degree (proj.) = 12
```

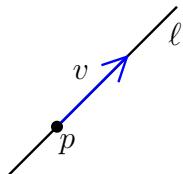
The dimension is now zero and we have removed the excess component from $\mathcal{V}(I)$.

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 any four spheres in general position. In Exercise 4 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.



Twelve lines tangent to four general spheres. We remark that the computation above, 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.

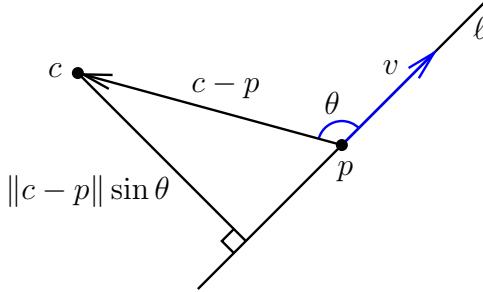
Lemma 10.3.2. *The set of direction vectors $v \in \mathbb{RP}^2$ of lines tangent to four spheres with affinely independent centers consists of the common solutions to a cubic and a quartic equation on \mathbb{RP}^2 . Each direction vector gives one common tangent.*

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

$$p \cdot v = 0. \quad (10.15)$$

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

$$[(c - p) \times v]^2 = r^2 v^2. \quad (10.16)$$

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

$$p^2 = r^2. \quad (10.17)$$

For the other spheres, we expand (10.16), use vector product identities, and the equations (10.15) and (10.17) to obtain the vector equation

$$2v^2 \begin{pmatrix} c_1^T \\ c_2^T \\ c_3^T \end{pmatrix} \cdot p = - \begin{pmatrix} (c_1 \cdot v)^2 \\ (c_2 \cdot v)^2 \\ (c_3 \cdot v)^2 \end{pmatrix} + v^2 \begin{pmatrix} c_1^2 + r^2 - r_1^2 \\ c_2^2 + r^2 - r_2^2 \\ c_3^2 + r^2 - r_3^2 \end{pmatrix}. \quad (10.18)$$

Now suppose that the spheres have affinely independent centers. Then the matrix $(c_1, c_2, c_3)^T$ appearing in (10.18) is invertible. Assuming $v^2 \neq 0$, we may use (10.18) to write p as a quadratic function of v . Substituting this expression into equations (10.15) and (10.17), we obtain a cubic and a quartic equation for $v \in \mathbb{RP}^2$. The lemma now follows from Bézout's Theorem. \square

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 10.3.3. Suppose that the spheres have equal radii, r , and have centers at the vertices of a regular tetrahedron with side length $2\sqrt{2}$,

$$(2, 2, 0)^T, \quad (2, 0, 2)^T, \quad (0, 2, 2)^T, \quad \text{and} \quad (0, 0, 0)^T.$$

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 10.3 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

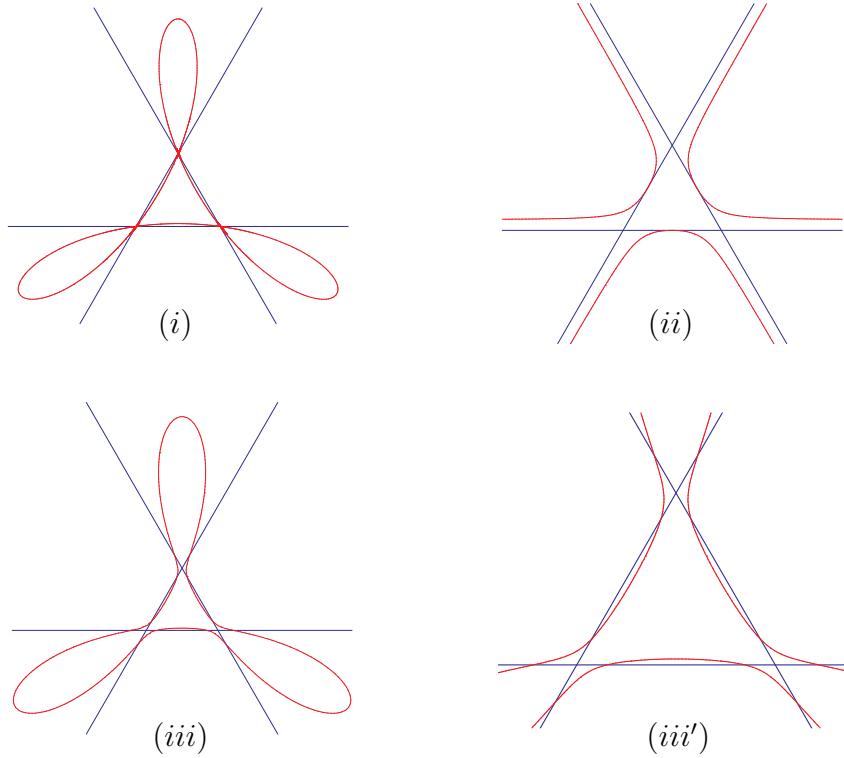


Figure 10.3: The cubic and quartic for symmetric configurations.

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 10.3.2 is in fact attained.

Theorem 10.3.4. *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 10.3.5. We give an example when the radii are distinct, namely 1.4, 1.42, 1.45, and 1.474. Figure 10.4 shows the quartic and cubic and the configuration of 4 spheres and their 12 common tangents.

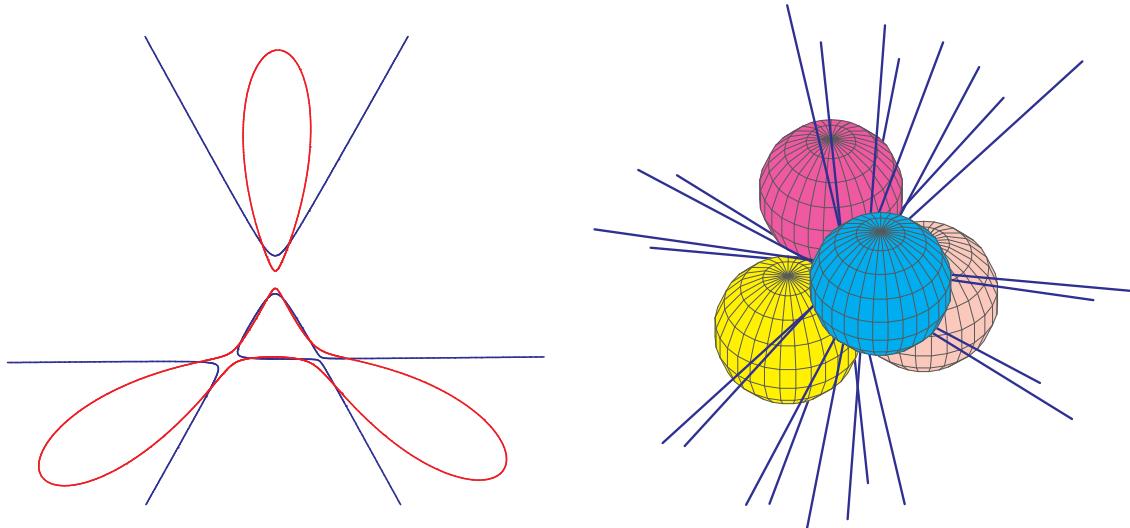


Figure 10.4: Spheres with 12 common tangents.

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 10.5. 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 10.3.3 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 10.4 and 10.5 not all pairs of spheres meet. However, the union of the spheres is connected. This turns out to be

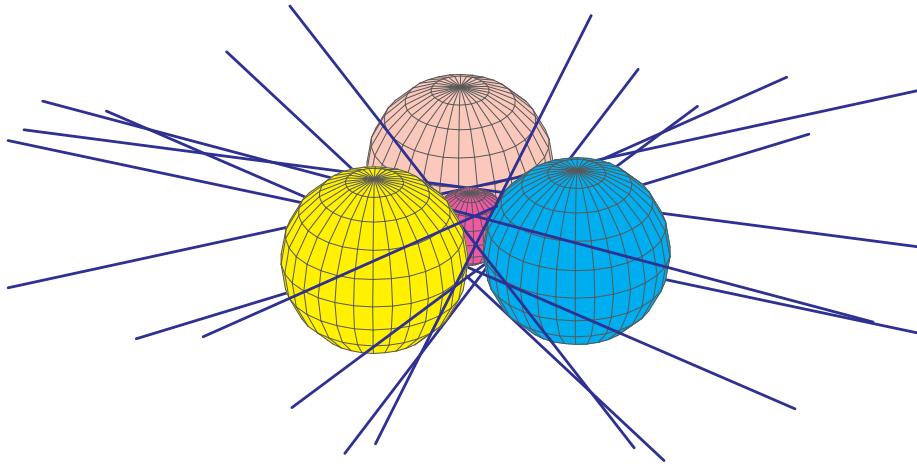


Figure 10.5: Spheres with coplanar centers and 12 common tangents.

unnecessary. 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 10.6 displays such a configuration. The three large spheres have radius

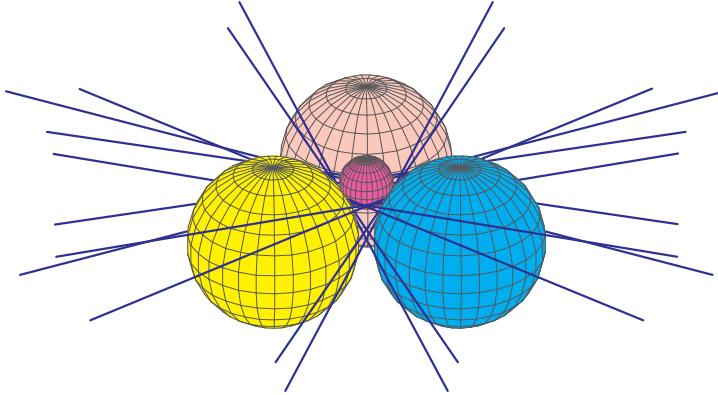


Figure 10.6: Four disjoint spheres with 12 common tangents.

$4/5$ 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.

If the spheres are unit spheres and the centers are coplanar, then Megyesi showed that the maximal number of solutions goes down to 8.

For the question of degenerate configurations of spheres, the following characterization can be given.

Theorem 10.3.6 (Macdonald, Pach, Theobald). *Four degenerate spheres in \mathbb{R}^3 of equal radii have colinear centers.*

This characterization can be extended to the case of general radii.

Theorem 10.3.7 (Borcea, Goaoc, Lazard, Petitjean). *Four degenerate spheres in \mathbb{R}^3 have colinear centers.*

Exercises

1. Let Q be a symmetric 2×2 matrix. Show that the quadratic form $X^T Q X$ has distinct factors if and only if Q is invertible. If Q has rank 1, then show that $X^T Q X$ is the square of a linear form, and that $X^T Q X$ is the zero polynomial only when Q has rank zero.
2. Verify that there are indeed 12 lines tangent (four real and 8 complex) to the four spheres of the example in Section 10.3.
3. 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 .
4. Verify the claim that there are 12 complex tangents to the four spheres discussed in the computer algebra example. Show that six of these are real.

10.4 The Stewart platform and robotics

As an outlook, we consider some robot mechanisms, which, in a simplified way, we can think of as systems of bars, joints and axes, whose parameters (for example, the length of an element or the angle between two elements) are variable. In particular, we concentrate on so-called *manipulators*, that is, mechanisms which are used at a fixed place and which are not mobile.

The discipline of *kinematics* is concerned with the geometry and the time-dependent aspects of the movements of such a mechanism. The forces, which cause the movements, are not taken into these considerations.

Specifically, we consider a particular three-dimensional manipulator called the *Stewart platform*. It is a manipulator, in which six points $p^{(1)}, \dots, p^{(6)}$ are fixed in space (often in the ground plane) and six points $q^{(1)}, \dots, q^{(6)}$ lie on a rigid body K , which can be moved in space (by translation and rotation). The points $p^{(i)}$ and $q^{(i)}$ are connected by segments ("legs") of variable lengths ℓ_i , and these legs are fixed at the points $p^{(i)}$ and $q^{(i)}$ via spherical joints (see Figure 10.7). Mechanisms of this type are used, for instance, in special vehicles and in flight simulators.

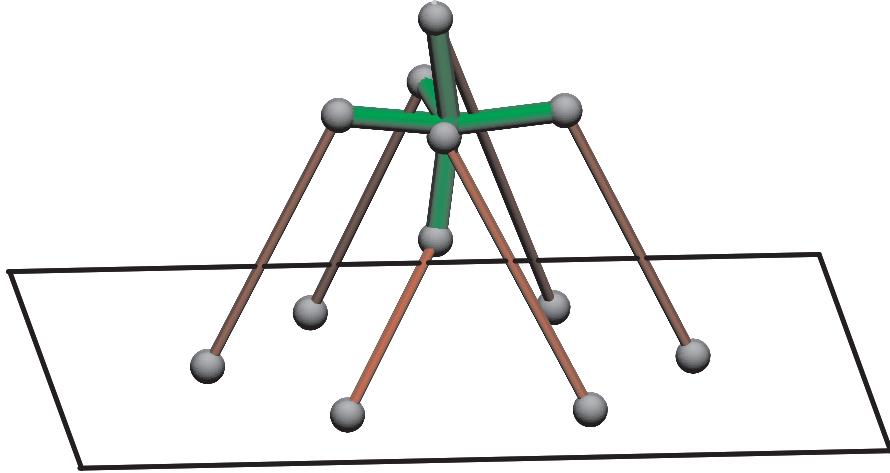


Figure 10.7: A Stewart platform

In the direct kinematic problem for the Stewart platform, the task is to determine the position and the orientation of K from the lengths of the six connecting segments. For each leg, the distance condition is given by an equation. Usually, one initially considers the base points and the platform points in separated coordinate systems Σ_1 and Σ_2 . Let $p^{(i)}$ und $q^{(j)}$ be the base points and the platform points in the individual coordinate systems. Denote by $x = (x_1, x_2, x_3)$ the coordinates of the origin of Σ_2 in Σ_1 . Moreover, let R be the orthogonal 3×3 -matrix, which describes the orientation (that is, the rotation) of K in the outer coordinate systems Σ_1 . The equation for the i -th leg can then be stated in the form

$$(x + Rq^{(i)} - p^{(i)})^T(x + Rq^{(i)} - p^{(i)}) = \ell_i^2. \quad (10.19)$$

Denote by $\|\cdot\|$ the Euclidean norm. Choosing, say, $p^{(6)}$ and $q^{(6)}$ in the origin then gives the system of equations

$$\begin{aligned} \|x\|^2 &= \ell_i^2, \\ x^T R q^{(i)} - (p^{(i)})^T R q^{(i)} - (p^{(i)})^T x &= \gamma_i, \quad 1 \leq i \leq 5, \end{aligned} \quad (10.20)$$

where $\gamma_i = \ell_i^2 - \ell_1^2 - \frac{1}{2}R^2\|q^{(i)}\|^2 - \frac{1}{2}\|p^{(i)}\|^2$. Since we are interested in real solutions, we can write a norm $\|x\|^2$ as $x^T x$, so that (10.20) becomes the system of polynomial equations

$$\begin{aligned} x^T x^2 &= \ell_i^2, \\ x^T R q^{(i)} - (p^{(i)})^T R q^{(i)} - (p^{(i)})^T x &= \gamma_i, \quad 1 \leq i \leq 5. \end{aligned} \quad (10.21)$$

Using the computer-algebraic methods developed in Chapter 1, we now show that whenever the system (10.20) has finitely many solutions, then it has at most 40, even over \mathbb{C} .

We substitute $x^T R$ by u^T and express the equations for $i = 1, \dots, 5$ in the set of x -, R - and u -variables. Altogether these are $3 + 9 + 3 = 15$ variables. The substitution of

$x^T R$ by u^T makes the equations for $i = 1, \dots, 5$ linear in the u -variables and hence in all the variables. Since R is required to be an orthogonal matrix, the equation $u^T = x^T R$ also gives the equation $x^T = u^T R^T$, i.e., $x = Ru$

We consider the following equations in the 15 variables given by x , u and R .

$$\begin{aligned} RR^T &= I_3 && (9 \text{ scalar equations}), \\ \det(R) &= 1 && (\text{one scalar equation}), \\ u^T &= x^T R && (3 \text{ scalar equations}), \\ x &= Ru && (3 \text{ scalar equations}), \end{aligned} \tag{10.22}$$

where I_3 is the 3×3 -identity matrix.

Let I be the ideal defined by the 16 corresponding polynomials. We consider the graded reverse lexicographic ordering on the set of variables

$$x_1, x_2, x_3, u_1, u_2, u_3, r_{11}, r_{12}, r_{13}, r_{21}, r_{22}, r_{23}, r_{31}, r_{32}, r_{33}$$

Using a computer algebra systems, such as **Singular**, we obtain a Gröbner bases for I consisting of 41 polynomials. Moreover, the (projective) dimension of the ideal I is 5 and the (projective) degree is 20. Now consider the ideal J which results from I by adding the five linear polynomials and one quadratic polynomial from (10.21) as generators. By Bézout's Theorem, if I has finitely many zeroes over \mathbb{C} , then their number is bounded by 40. Any of these zeroes gives a solution to the direct kinematic problem for a Stewart platform. Hence, the computer calculation has shown the following result.

Theorem 10.4.1 (Ronga, Vust). *If the direct kinematic problem for a Stewart platform has a finite number of solutions over \mathbb{C} , then there are at most 40 solutions.*

Indeed, the direct kinematic problem for a Stewart platform has 40 solutions over \mathbb{C} if the lengths are chosen in general position. And there exist edge lengths, for which all of the 40 solutions are real.

Exercises

1. Consider the subsystem of (10.22) given by $RR^T = I_3$ and $\det(R)$, where $R = (r_{ij})_{1 \leq i,j \leq 3}$. Verify that the ten polynomials underlying this system generate an ideal whose Gröbner basis with respect to the degree lexicographic ordering for r_{11}, \dots, r_{33} consists of the 20 polynomials

$$\begin{aligned} r_{31}^2 + r_{32}^2 + r_{33}^2 - 1, & r_{22}r_{33} - r_{23}r_{32} - r_{11}, & r_{21}r_{33} - r_{23}r_{31} + r_{12}, \\ r_{21}r_{32} - r_{22}r_{31} - r_{13}, & r_{21}r_{31} + r_{22}r_{32} + r_{23}r_{33}, & r_{21}^2 + r_{22}^2 + r_{23}^2 - 1, \\ r_{13}^2 + r_{23}^2 + r_{33}^2 - 1, & r_{12}r_{33} - r_{13}r_{32} + r_{21}, & r_{12}r_{23} - r_{13}r_{22} - r_{31}, \\ r_{12}r_{13} + r_{22}r_{23} + r_{32}r_{33}, & r_{12}^2 + r_{22}^2 + r_{32}^2 - 1, & r_{11}r_{33} - r_{13}r_{31} - r_{22}, \\ r_{11}r_{32} - r_{12}r_{31} + r_{23}, & r_{11}r_{31} + r_{12}r_{32} + r_{13}r_{33}, & r_{11}r_{23} - r_{13}r_{21} + r_{32}, \\ r_{11}r_{22} - r_{12}r_{21} - r_{33}, & r_{11}r_{21} + r_{12}r_{22} + r_{13}r_{23}, & r_{11}r_{13} + r_{21}r_{23} + r_{31}r_{33}, \\ r_{11}r_{12} + r_{21}r_{22} + r_{31}r_{32}, & r_{11}^2 + r_{12}^2 + r_{13}^2 - 1. \end{aligned}$$

2. Determine experimentally configurations of the Stewart platform which have many real solutions.

10.5 Notes

Plücker coordinates and the Grassmannian go back to Julius Plücker (1808–1868) and to Herrmann Grassmann (1809–1877). Our treatment follows the presentations in Hodge and Pedoe [61], Pottmann and Wallner [107], Fischer and Piontkowski [40] and Joswig and Theobald [69].

For the lines tangent to spheres see [17, 90, 85, 132], some of the material presented here comes from the survey [133].

Ronga and Vust [118] showed that the Stewart platform generically has 40 solutions. The Gröbner basis calculation presented here to obtain the bound of 40 was found by Lazard [80]. Dietmaier has given a configuration for which all 40 solutions are real [32].

For further applications and connections of enumerative real algebraic geometry see the survey [130] and the book [131].

Appendix A

Appendix

Frank thinks that basic material should go here, with references. Complete explanations are not necessary, but care should be taken to not abuse the reader.

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

Field of rational functions, $\mathbb{C}(t)$? 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{K} equipped with operations of addition and multiplication and distinguished elements 0 and 1 (the additive and multiplicative identities). Every number $a \in \mathbb{K}$ has an additive inverse $-a$ and every non-zero number $a \in \mathbb{K}^\times := \mathbb{K} - \{0\}$ 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{K} 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 $v_1, \dots, v_n \in V$ is any vector of the form

$$a_1v_1 + a_2v_2 + \cdots + a_nv_n,$$

where $a_1, \dots, a_n \in \mathbb{K}$. 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 a 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

$$\langle S \rangle := \{r_1s_1 + r_2s_2 + \cdots + r_ms_m \mid r_1, \dots, r_m \in R \text{ and } s_1, \dots, s_m \in S\}.$$

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

$$\begin{aligned} [r] + [s] &= \{r' + s' \mid r' \in [r] \text{ and } s' \in [s]\} \stackrel{!}{=} [r + s], \quad \text{and} \\ [r] \cdot [s] &= \{r' \cdot s' \mid r' \in [r] \text{ and } s' \in [s]\} \stackrel{!}{=} [rs]. \end{aligned}$$

The last equality 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* $\varphi: R \rightarrow S$ 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 $\varphi: R \rightarrow S$,

$$\ker \varphi := \{r \in R \mid \varphi(r) = 0\}$$

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 $\varphi: R \rightarrow R/I$ whose kernel is I . Dually, given a homomorphism $\varphi: R \rightarrow S$, the image of R in S is identified with $R/\ker \varphi$. More generally, if $\varphi: R \rightarrow S$ 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 $\varphi: R/I \rightarrow S$.

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

$$I =: I_0 \subsetneq I_1 \subsetneq I_2 \subsetneq \dots$$

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

$$p = p(x) := a_0 + a_1x + a_2x^2 + \cdots + a_mx^m, \quad (\text{A.1})$$

where m is a nonnegative integer and the coefficients a_0, a_1, \dots, a_m lie in \mathbb{K} . Write $\mathbb{K}[x]$ for the set of all polynomials in the variable x with coefficients in \mathbb{K} . We may add, subtract, and multiply polynomials and $\mathbb{K}[x]$ is a ring.

While a polynomial p may be regarded as a formal expression (A.1), evaluation of a polynomial defines a function $p: \mathbb{K} \rightarrow \mathbb{K}$: The value of the function p at a point $a \in \mathbb{K}$ is simply $p(a)$. When \mathbb{K} is infinite, the polynomial and the function determine each other, but this is not the case when \mathbb{K} is finite.

For polynomials with more than one variable, we begin with multivariate monomials.

Definition A.1.2. A *monomial* in the variables x_1, \dots, x_n is a product of the form

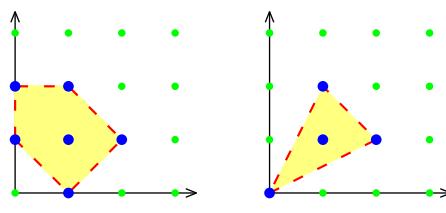
$$x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n},$$

where the exponents $\alpha_1, \dots, \alpha_n$ are nonnegative integers. For notational convenience, set $\alpha := (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^{n\dagger}$ 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* $f = f(x_1, \dots, x_n)$ in the variables x_1, \dots, x_n is a linear combination of monomials, that is, a sum of the form

$$f = \sum_{\alpha \in \mathbb{N}^n} a_\alpha x^\alpha,$$

where each *coefficient* a_α lies in \mathbb{K} and all but finitely many coefficients vanish. The product $a_\alpha x^\alpha$ of an element a_α of \mathbb{K} and a monomial x^α is 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. For example, the bivariate polynomial $p := 1 - 2xy + 4xy^2 - 8x^2y$ has support $\{(0,0), (1,1), (1,2), (2,1)\}$. Writing the elements of the support as the columns of a matrix, this is $(\begin{smallmatrix} 0 & 1 & 1 & 2 \\ 0 & 1 & 2 & 1 \end{smallmatrix})$. The polynomial $q := x + 2y + 3xy + 5y^2 + 7xy^2 + 11x^2y$ has support $(\begin{smallmatrix} 1 & 0 & 1 & 0 & 1 & 2 & 2 & 1 \\ 0 & 1 & 1 & 2 & 2 & 1 \end{smallmatrix})$. The *Newton polytope*, $\text{New}(f)$ of a polynomial f is the convex hull of its support. Here are the Newton polytopes of p and q (which are polygons), respectively.



[†]Where have we defined \mathbb{N} ?

We will say that f has support \mathcal{A} to mean that the support of f is a subset of \mathcal{A} . With this definition, the set of all polynomials with support a finite set $\mathcal{A} \subset \mathbb{N}^n$ is the vector space $\mathbb{K}^{\mathcal{A}}$, consisting of all coefficient vectors $(a_{\alpha} \mid \alpha \in \mathcal{A})$ for polynomials with support \mathcal{A} . Here, as elsewhere, we take the finite set \mathcal{A} as a natural index set.

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, ... Every integer $n > 1$ has a unique factorization into prime numbers

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_n^{\alpha_n},$$

where $p_1 < \cdots < p_n$ are distinct primes, and $\alpha_1, \dots, \alpha_n$ are (strictly) positive integers. For example, $999 = 3^3 \cdot 37$. Polynomials also have unique factorization.

Definition A.1.3. A nonconstant polynomial $f \in \mathbb{K}[x_1, \dots, x_n]$ 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.

Theorem A.1.4. *Every polynomial $f \in \mathbb{K}[x_1, \dots, x_n]$ is a product of irreducible polynomials*

$$f = p_1 \cdot p_2 \cdots p_m,$$

where the polynomials p_1, \dots, 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 $f(x) \in \mathbb{K}[x]$ 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 $f_m \in \mathbb{K}^{\times}$ is the leading coefficient of f , the polynomials p_1, \dots, 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 $f, g \in \mathbb{K}[x]$ (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 $\mathbb{K}[x]$ with $\deg(g) \geq \deg(f)$,

$$\begin{aligned} f &= f_0 + f_1x + f_2x^2 + \cdots + f_mx^m \\ g &= g_0 + g_1x + g_2x^2 + \cdots + g_nx^n, \end{aligned}$$

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 $f, g \in \mathbb{K}[x]$.

OUTPUT: Polynomials $q, r \in \mathbb{K}[x]$ 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 $f, g \in \mathbb{K}[x]$.

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

$$\langle f \rangle = \{h \cdot f \mid h \in \mathbb{K}[x]\},$$

for some $f \in \mathbb{K}[x]$. We say that f *generates* $\langle f \rangle$. Since $\langle f \rangle = \langle \alpha f \rangle$ for any $\alpha \in \mathbb{K}$, the principal ideal has a unique monic generator.

Theorem A.1.8. *Every ideal I of $\mathbb{K}[x]$ is principal.*

Proof. Suppose that I is a nonzero ideal of $\mathbb{K}[x]$, 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 $q, r \in \mathbb{K}[x]$ with

$$g = qf + r \quad \text{with} \quad \deg(r) < \deg(f).$$

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 $I = \langle f \rangle$. \square

The ideal generated by univariate polynomials f_1, \dots, f_s is the principal ideal $\langle p \rangle$, where p is the greatest common divisor of f_1, \dots, f_s .

For univariate polynomials p the quotient ring $\mathbb{K}[x]/\langle p \rangle$ has a concrete interpretation. Given $f \in \mathbb{K}[x]$, we may call the Division Algorithm to obtain polynomials q, r with

$$f = q \cdot p + r, \text{ where } \deg(r) < \deg(p).$$

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 $\mathbb{K}[x]/\langle p \rangle$.

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 $1, x, x^2, \dots, x^{\deg(p)-1}$ form a basis for the \mathbb{K} -vector space $\mathbb{K}[x]/\langle p \rangle$. This describes the additive structure on $\mathbb{K}[x]/\langle p \rangle$.

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.

[†]This is poorly written!

- 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 $R = \mathbb{K}[x]/p(x)$ 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 $\mathbb{C}[x_1, \dots, x_n]$ all have the form \mathfrak{m}_a for some $a \in \mathbb{A}^n$.*

A.1.4 Multilinear algebra

Let \mathbb{K} be a field and V be the n -dimensional vector space \mathbb{K}^n with a basis $e^{(1)}, \dots, e^{(n)}$. For $k \in \{1, \dots, n\}$ and indices $1 \leq i_1 < \dots < i_k \leq n$, the formal symbol

$$e^{(i_1)} \wedge \cdots \wedge e^{(i_k)}$$

is called the *exterior product* of the basis vectors $e^{(i_1)}, \dots, e^{(i_k)}$. The k -th exterior power $\bigwedge^k V$ is the set of formal \mathbb{K} -linear combinations of the symbols $e^{(i_1)} \wedge \cdots \wedge e^{(i_k)}$, $1 \leq i_1 < \dots < i_k \leq n$. By convention, $\bigwedge^0 V = \mathbb{K}$. The direct sum

$$\bigwedge V = \bigwedge^0 V \oplus \bigwedge^1 V \oplus \cdots \oplus \bigwedge^n V$$

is called the *exterior algebra* of V .

The exterior multiplication is associative and anti-commutative, that is, for any $x, y \in V$ we have $x \wedge y = -y \wedge x$. For $x^{(1)}, \dots, x^{(k)} \in V$, we have $x^{(1)} \wedge \cdots \wedge x^{(k)} = 0$ if and only if $x^{(1)}, \dots, x^{(k)}$ are linearly dependent over \mathbb{K} .

The *Cauchy–Binet formula* states that for two matrices $A \in \mathbb{K}^{n \times k}$, $B \in \mathbb{K}^{k \times n}$ we have

$$\det AB = \sum_{I \subseteq \{1, \dots, n\}, |I|=k} \det A_I \det B_I, \quad (\text{A.2})$$

where A_I and B_I are the submatrices of A and B in which only those columns of A and rows of B are used whose indices are in I .

A.1.5 Real algebra

In real algebra, the concept of an ordered field is central.

Definition A.1.10. A field \mathbb{K} together with a relation \leq on \mathbb{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* $P = \{a \in \mathbb{K}, a \geq 0\}$. Note that $P + P \subset P$, $PP \subset P$, $P \cap -P = \{0\}$ as well as $P \cup -P = \mathbb{K}$. Conversely, any subset P of \mathbb{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.

Theorem A.1.11. For a field \mathbb{K} , the following statements are equivalent.

1. \mathbb{K} is real closed, that is, \mathbb{K} has no proper, real, algebraic extension.
2. $\mathbb{K}[i] = \mathbb{K}[y]/(y^2 + 1)$ is algebraically closed.
3. \mathbb{K} is an ordered field, whose set of non-negative elements is exactly the set \mathbb{K}^2 , and such that every polynomial in $\mathbb{K}[x]$ of odd degree has a root in \mathbb{K} .

For a proof, see, e.g., [5].

We also use the following version of Tarski's Transfer Principle. For proofs, see, for instance, [5, Theorem 2.80] or [110, Theorem 2.1.10].

Theorem A.1.12 (Tarski's transfer principle). *Let (\mathbb{K}, \leq) be an ordered field extension of (\mathbb{R}, \leq) . If there exists an $a \in \mathbb{K}^n$ 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.1.13 (Quantifier elimination). *Let $g, f_1, \dots, f_m \in \mathbb{Z}[x_1, \dots, x_n, y]$. Then there are $g_i, f_{ij} \in \mathbb{Z}[x_1, \dots, x_n]$, $1 \leq i \leq l$, $1 \leq j \leq r_i$ (with $l, r_i \in \mathbb{N}$) such that for every real closed field R and for all $a \in R^n$*

$$\exists b \in R \left(g(a, b) = 0 \wedge \bigwedge_{j=1}^m f_j(a, b) > 0 \right) \iff \bigvee_{i=1}^l \left(g_i(a) = 0 \wedge \bigwedge_{j=1}^{r_i} f_{ij}(a) > 0 \right).$$

A.2 Topology

Topology concerns the most basic properties of shape and space. A fundamental notion is that of continuity that many of us first encountered in calculus. It begins quite formally. A topology on a set X is a collection of subsets of X , called *open sets*, that satisfy the following properties.

1. Both the empty set and X itself are open.
2. Any union of open sets are open.
3. Any finite intersection of open sets is open.

A set X with a topology is called a *topological space*. The complement of an open set is a *closed set*. A topology is dually specified by its collection of closed sets. Closed sets satisfy corresponding properties: Both X and are closed, any intersection of closed sets is closed, and any finite union of closed sets is closed.

The *closure* \overline{Z} of a subset $Z \subset X$ of a topoloical space is the intersection of all closed subsets that contain Z . It is the smallest closed subset of X that contains Z . A topology on X can be specified by giving a collection of subsets of X called *basic open sets*, and then taking the open sets to be the smallest collection of subsets containing these basic open sets and that satisfies the three given properties.

The standard examples of topological spaces are \mathbb{R}^n and \mathbb{C}^n with what we call their *Euclidean topology*. Here, the basic open sets are Euclidean balls. For $x \in \mathbb{R}^n$ and $\epsilon > 0$

$$B(x, \epsilon) := \{a \in \mathbb{R}^n \mid \sum |a_i - x_i|^2 < \epsilon\}.$$

The same formula gives a Euclidean ball in \mathbb{C}^n , where for a complex number $z \in \mathbb{C}$, we have $|z|^2 = z\bar{z}$, where \bar{z} is the complex conjugate of z .

A subset $Y \subset X$ of a topological space has an induced *subspace topology*. A subset of Y is open if it is the intersection of Y with an open subset of X . A subset Y of X with this topology is a (topological) subspace of X .

Contiuous functions are the functions between topological spaces that preserve their structures. More formally, a function $f: X \rightarrow Y$ between topological spaces X and Y is continous if for any open subset U of Y , $f^{-1}(U)$ is an open subset of X . Dually, for any closed subset Z of Y , $f^{-1}(Z)$ is closed in X .

The standard topology on \mathbb{R}^n or \mathbb{C}^n , along with the subspace topology on their subsets, is common in mathematics. Algebraic geometry uses a weaker topology, called the *Zariski topology*. Its basis of open sets have the form $U_f := \{x \in \mathbb{C}^n \mid f(x) \neq 0\}$, where $f \in \mathbb{C}[x_1, \dots, x_n]$ is a polynomial, and its closed subsets are varieties $\mathcal{V}(S) := \{x \in \mathbb{C}^n \mid f(x) = 0 \forall f \in S\}$, where $S \subset \mathbb{C}[x_1, \dots, x_n]$ is a set of polynomials. This is introduced in Section 3.1.

A.3 Convex geometry

A.3.1 Polytopes and polyhedra

Polytopes and polyhedra are important semi-algebraic sets that are fundamental to toric varieties and tropical geometry. For more information see the books of Ewald [36] and Ziegler [144]. Let $\{v_1, \dots, v_m\} \subset \mathbb{R}^n$ be a finite set of points. A sum

$$\sum_{i=1}^m \lambda_i v_i \quad \text{where} \quad \lambda_1, \dots, \lambda_m \geq 0 \quad \text{and} \quad \sum_{i=1}^m \lambda_i = 1$$

is a *convex combination* of the points v_1, \dots, v_m . The *convex hull* of $\{v_1, \dots, v_m\}$ is the set of all their convex combinations,

$$\text{conv}\{v_1, \dots, v_m\} := \left\{ \sum_{i=1}^m \lambda_i v_i \mid \lambda_1, \dots, \lambda_m \geq 0 \quad \text{and} \quad \sum_{i=1}^m \lambda_i = 1 \right\}. \quad (\text{A.3})$$

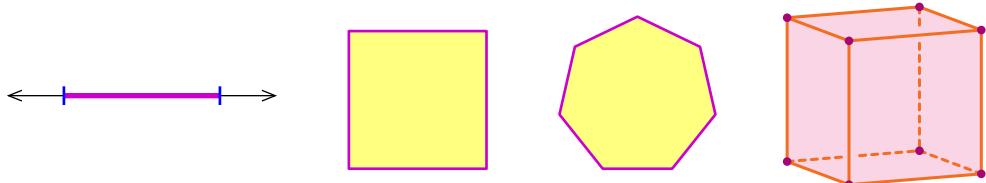
A convex hull of a finite set of points is a *polytope*. If this representation is irredundant in that no point v_i lies in the convex hull of the others, then each point v_i is a *vertex* of the polytope.

A translate $x + L$ of a linear subspace L by a vector $x \in \mathbb{R}^n$ is an affine subspace. The dimension of $x + L$ is the dimension of L . The *affine span*, $\text{Aff}(X)$ of a set X is the intersection of all affine subspaces that contain X . For any $x \in X$, it has the form $x + \text{span}\{y - x \mid y \in X\}$. It is also the set of all affine combinations of points of X ,

$$\text{Aff } X = \left\{ \sum_{i=1}^m \lambda_i x_i \mid x_1, \dots, x_m \in X \quad \text{and} \quad \sum_{i=1}^m \lambda_i = 1 \right\}. \quad (\text{A.4})$$

(This differs from the convex hull in that the coefficients λ_i may be negative.) The dimension of a polytope P is the dimension of its affine span.

There is only one polytope (a point) of dimension 0. Polytopes of dimension 1 are line segments, two-dimensional polytopes are polygons, and three-dimensional polytopes are familiar objects such as the cube in \mathbb{R}^3 .



A polytope P with m vertices has dimension at most $m-1$. When it has dimension $m-1$, it is a *simplex*. For example, the *standard*, or probability, n -dimensional simplex Δ^n is the convex hull of the $n+1$ linearly independent standard basis vectors e_1, \dots, e_{n+1}

in \mathbb{R}^{n+1} . It is the intersection of the affine hyperplane $\{\lambda \in \mathbb{R}^{n+1} \mid \lambda_1 + \dots + \lambda_{n+1} = 1\}$ and the positive orthant,

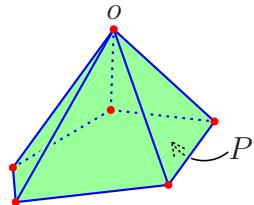
$$\Delta^n = \{\lambda \in \mathbb{R}^{n+1} \mid \lambda_i \geq 0 \text{ and } \lambda_1 + \dots + \lambda_{n+1} = 1\}.$$

A polytope with m vertices $\text{conv}\{v_1, \dots, v_m\}$ (A.3) is the image of the standard $m-1$ simplex under the linear map that sends $e_i \in \mathbb{R}^m$ to $v_i \in \mathbb{R}^n$, for $i = 1, \dots, m$. This is evident as points in the simplex parametrize convex combinations under this map. More generally, the image of any polytope under a linear or an affine map is another polytope—simply take the convex hull of the images of the vertices.

As a polytope P is closed and bounded, for $w \in \mathbb{R}^n$, the linear function $x \mapsto w^T x$ is bounded below and achieves its minimum value, $h_P(w)$, on P . The function $w \mapsto h_P(w)$ is the *support function* of P , and the subset $P_w := \{x \in P \mid w^T x = h_P(w)\}$ of P where this minimum is attained is the face of P *exposed* by w . A *face* of P is a set of the form P_w for some $w \in \mathbb{R}^n$. Faces are themselves polytopes; if $P = \text{conv}\{v_1, \dots, v_m\}$, then $P_w = \text{conv}\{v_i \mid w^T v_i = h_P(w)\}$. The polytope P is itself a face; it is exposed by the zero vector $0 \in \mathbb{R}^n$. Vertices are faces of dimension zero and edges are faces of dimension one. A *facet* of P is a face F of codimension one, $\dim F = \dim P - 1$. As its faces are convex hulls of subsets of its vertices, a polytope P has only finitely many faces.

Perhaps this is where it would be useful to point out that the set of w which expose a face of a polytope is a cone, giving equations and inequations, and note that there are integer points in this cone if the polytope is integral. This is used in Chapter 8. It would also be used in defining the Gröbner fan, which might be useful in tropical geometry ?

Example A.3.1. A useful construction of one polytope from another is a pyramid. Suppose that a polytope $P \subset \mathbb{R}^n$ has dimension $n-1$. Then its affine span is a hyperplane H . For any point $o \in \mathbb{R}^n \setminus H$, the *pyramid* with base P and *apex* o is the convex hull of the polytope P and the point o .



◊

By the definition of the support functions, for every $w \in \mathbb{R}^n$ we have that

$$P \subset \{x \in \mathbb{R}^n \mid w^T x \geq h_P(w)\}.$$

When $w \neq 0$, the set $\{x \in \mathbb{R}^n \mid w^T x \geq h_P(w)\}$ is a *half space* and its boundary $H := \{x \in \mathbb{R}^n \mid w^T x = h_P(w)\}$ is a *supporting hyperplane* of P . The intersection of P with this supporting hyperplane is the face P_w of P exposed by w . Also note that a face P_w of a polytope is the intersection of P with the affine span of P_w .

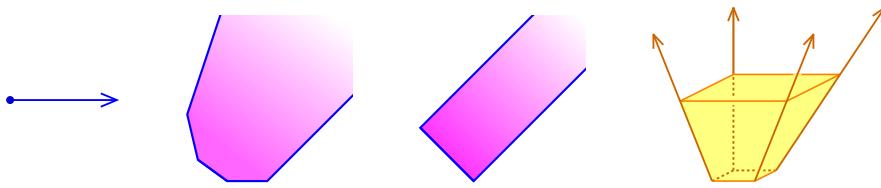
As a closed, convex body, P is the intersection of all half-spaces that contain it,

$$P = \bigcap_{w \in \mathbb{R}^n} \{x \in \mathbb{R}^n \mid w^T x \geq h_P(w)\}. \quad (\text{A.5})$$

This intersection may be taken to be finite, there is a finite set $w_1, \dots, w_d \in \mathbb{R}^n$ such that

$$P = \{x \in \mathbb{R}^n \mid w_i^T x \geq h_P(w_i) \text{ for } i = 1, \dots, d\}. \quad (\text{A.6})$$

A *polyhedron* is the intersection of finitely many half spaces. In general, a polyhedron may be unbounded. Here are four unbounded polyhedra in \mathbb{R} , \mathbb{R}^2 , \mathbb{R}^2 , and \mathbb{R}^3 , respectively.



A polyhedron P has a support function $h_P(w)$ that takes values in $\mathbb{R} \cup \{-\infty\}$. When P is unbounded in the direction opposite w , then $h_P(w) = -\infty$. With this definition, the description (A.5) holds for P . Polytopes are exactly the bounded polyhedra.

Given the facet description (A.6) of a polyhedron, let Λ be the $d \times n$ matrix whose rows are the facet normals w_i^T for $i = 1, \dots, d$ and let $b \in \mathbb{R}^d$ be the column vector with i th entry $-h_P(w_i)$. Then (A.6) becomes

$$P = \{x \in \mathbb{R}^n \mid Ax + b \geq 0\}, \quad (\text{A.7})$$

where \geq is coordinatewise comparison.

This leads to another description. The affine map $\Lambda: x \mapsto Ax + b$ sends \mathbb{R}^n to the affine subspace $L := A\mathbb{R}^n + b$ of \mathbb{R}^d , and by (A.7) the image $\Lambda(P)$ of P under this map is the intersection $L \cap \mathbb{R}_+^d$ of L with the nonnegative orthant $\mathbb{R}_+^d := \{y \in \mathbb{R}^d \mid y_i \geq 0\}$. Thus P is the inverse image of the polyhedron \mathbb{R}_+^d under the affine map Λ .

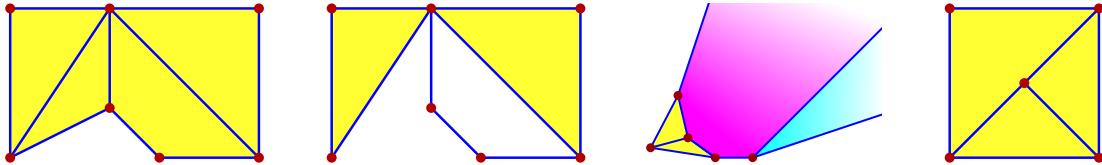
Suppose that an affine subspace $L \subset \mathbb{R}^n$ is defined by the affine equations $Ax = b$ where $A: \mathbb{R}^n \rightarrow \mathbb{R}^d$ is a linear map and $b \in \mathbb{R}^d$. Then the polyhedron $L \cap \mathbb{R}_+^n$ is

$$\{x \in \mathbb{R}^n \mid Ax = b \text{ and } x \geq 0\}.$$

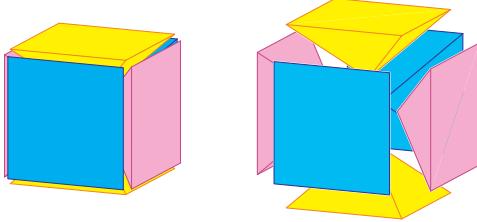
More generally, any section $L \cap P$ of a polyhedron P by an affine subspace L is again a polyhedron, as is the inverse image of a polyhedron under an affine map. By Fourier-Motzkin elimination (Theorem 5.4.4) the image of a polyhedron under an affine map is again a polyhedron.

A *polyhedral complex* is a collection \mathcal{P} of polyhedra in \mathbb{R}^n such that every face of a polyhedron P in \mathcal{P} is another polyhedron in \mathcal{P} and the intersection of any two polyhedra P, P' in \mathcal{P} is a common face of each. For example, of the four collections of vertices, line

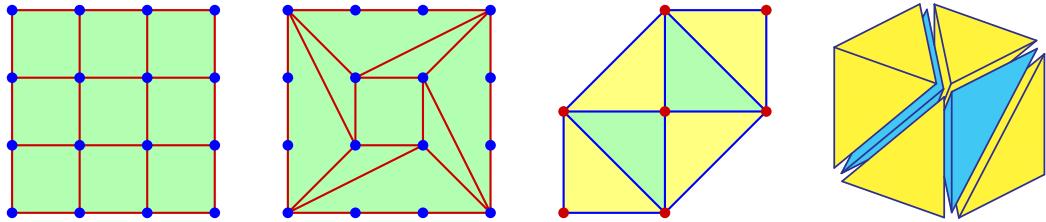
segments and polyhedra shown below, the first three are polyhedral complexes, while the last is not; the large triangle does not meet either smaller triangle in one of its faces.



The collection of all faces of a polyhedron is a polyhedral complex, as is the collection of its proper faces. For a less trivial example, suppose that $o \in P$ is any point of a polytope P . For every face F of P that does not contain o we may consider the pyramid with base F and apex o . This collection of pyramids, their bases, and the apex forms a polyhedral complex. For example, consider the cube in \mathbb{R}^3 with o its center. The resulting polyhedral complex has six square pyramids, which we show in both slightly and exaggerated exploded views.



The *support* of a polyhedral complex \mathcal{P} is the union of the polyhedra in \mathcal{P} . When the support of a polyhedral complex is a polyhedron P , the complex is a *subdivision* of P . When every polytope in a polyhedral complex \mathcal{P} is a simplex, \mathcal{P} is a *triangulation* of its support. Of the four polyhedral subdivisions below, the last two are triangulations.



A.3.2 Minkowski sum and mixed volumes

Polytopes in the vector space \mathbb{R}^n inherit two operations of sum and scalar multiplication, and they also have an intrinsic metric invariant, their volume. The interplay of these structures leads to the notion of mixed volume, which is important in the application of algebraic geometry. A standard reference is [122], see also [36].

Addition of vectors in \mathbb{R}^n induces the operation of *Minkowski sum* on polytopes, where

$$P + Q := \{x + y \mid x \in P, y \in Q\}.$$

We may similarly multiply a polytope P by a positive scalar λ to get λP . When λ is an integer, these operations coincide, for example $P + P = 2P$. We may combine them. Given polytopes $P_1, \dots, P_r \subset \mathbb{R}^n$ and nonnegative real numbers $\lambda_1, \dots, \lambda_r$, define

$$P(\lambda) := \lambda_1 P_1 + \dots + \lambda_r P_r. \quad (\text{A.8})$$

Lemma A.3.2. *The Minkowski sum $P(\lambda)$ (A.8) is a polytope. For any vector $w \in \mathbb{R}^n$, its support function $h_{P(\lambda)}(w)$ is the sum $\lambda_1 h_{P_1}(w) + \dots + \lambda_r h_{P_r}(w)$, and*

$$P(\lambda)_w = \lambda_1 P_{1,w} + \dots + \lambda_r P_{r,w}.$$

If $P(\lambda)_w$ is a facet of $P(\lambda)$ for one choice of $\lambda_1, \dots, \lambda_r$ with all $\lambda_i > 0$, then $P(\lambda)_w$ is a facet of $P(\lambda)$ for any $\lambda_1, \dots, \lambda_r$ with all $\lambda_i > 0$.

Example A.3.3. Suppose that $P \subset \mathbb{R}^2$ is the convex hull of the column vectors of the matrix $\begin{pmatrix} 1 & 0 & 1 & 2 & 1 \\ 0 & 1 & 1 & 1 & 2 \end{pmatrix}$ and Q is the convex hull of the column vectors of the matrix $\begin{pmatrix} 0 & 1 & 2 & 1 & 1 \\ 0 & 0 & 0 & 1 & 2 \end{pmatrix}$. Then $P + Q$ is the convex hull of the column vectors of the matrix $\begin{pmatrix} 1 & 3 & 0 & 4 & 1 & 3 & 2 \\ 0 & 0 & 1 & 1 & 3 & 3 & 4 \end{pmatrix}$. We display these polygons and their Minkowski sum in Figure A.1.

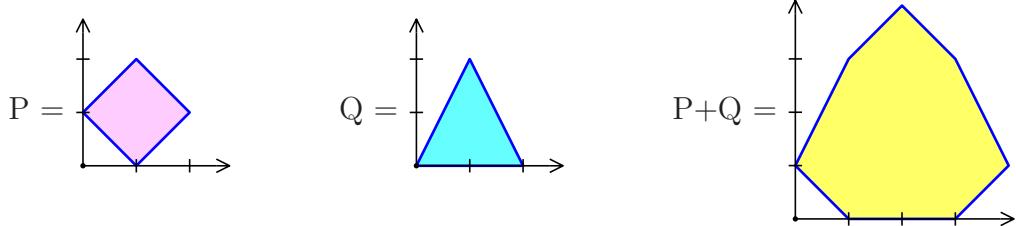


Figure A.1: Minkowski sum of two polygons.

Proof. Let $\{v_1, \dots, v_m\}$ and $\{u_1, \dots, u_d\}$ be finite subsets of \mathbb{R}^n . The sum of a convex combination of each is a convex combination of their sum,

$$\sum_{i=1}^m \lambda_i v_i + \sum_{j=1}^d \mu_j u_j = \sum_{i,j} \lambda_i \mu_j (v_i + u_j), \quad (\text{A.9})$$

as $\sum_i \lambda_i = \sum_j \mu_j = 1$, which implies that $\sum_{i,j} \lambda_i \mu_j = 1$. The same is true for a scalar multiple of a convex combination. Thus the Minkowski sum of polytopes is a polytope and a scalar multiple of a polytope is a polytope.

If λ, μ are nonnegative, then

$$\min\{(\lambda v_i + \mu u_j) w^T\}_{i,j} = \lambda \min\{v_i w^T\}_i + \mu \min\{u_j w^T\}_j,$$

which implies the linearity of the support function, and that the face of $P(\lambda)$ exposed by w is the scaled sum of faces of the P_i exposed by w . Finally, the statement about facets follows as the affine span of a nonnegative scalar multiple μX of a set X is the scalar multiple of the affine span of X , and the affine span of a Minkowski sum is the Minkowski sum of the affine span, again by (A.9). \square

We prove Minkowski's result about the volume of the scaled sum (A.8).

Theorem A.3.4. *Let $P_1, \dots, P_r \subset \mathbb{R}^n$ be polytopes. For nonnegative $\lambda_1, \dots, \lambda_r$, $\text{vol}_n(P(\lambda))$ is a homogeneous polynomial of degree n in $\lambda_1, \dots, \lambda_r$.*

Proof. Suppose first that $n = 1$. Then each P_i is an interval $[a_i, b_i]$ with $a_i \leq b_i$ so that $P(\lambda) = [\lambda_1 a_1 + \dots + \lambda_r a_r, \lambda_1 b_1 + \dots + \lambda_r b_r]$, and we have

$$\text{vol}_1(P(\lambda)) = \sum_{i=1}^r \lambda_i b_i - \sum_{i=1}^r \lambda_i a_i = \sum_{i=1}^r \lambda_i (b_i - a_i) = \sum_{i=1}^r \lambda_i \text{vol}_1(P_i),$$

which is homogeneous of degree 1 in $\lambda_1, \dots, \lambda_r$.

Now suppose that $n > 1$. As volume is invariant under translation, we make some assumptions for the purpose of computation. For a given $w \in \mathbb{R}^n$ and all i , we may assume that 0 lies in the face $P_{i,w}$ of P_i exposed by w . Then each $P_{i,w}$ as well as $P(\lambda)_w$ lies in the hyperplane annihilated by w , which is isomorphic to \mathbb{R}^{n-1} . By induction on dimension, we may assume that $\text{vol}_{n-1}(P(\lambda)_w) = \text{vol}_{n-1}(\lambda_1 P_{1,w} + \dots + \lambda_r P_{r,w})$ is a homogeneous polynomial of degree $n-1$ in $\lambda_1, \dots, \lambda_r$. This conclusion about $\text{vol}_{n-1}(P(\lambda)_w)$ remains true even if 0 does not lie in any face $P_{i,w}$.

Again translating $P(\lambda)$ if necessary, we may assume that $h_{P(\lambda)}(w) > 0$. Then the pyramid C_w with apex $0 \in \mathbb{R}^n$ over the facet $P(\lambda)_w$ of $P(\lambda)$ has height $\frac{1}{\|w\|} h_{P(\lambda)}(w)$ and therefore has volume

$$\frac{1}{n} \cdot \frac{1}{\|w\|} h_{P(\lambda)}(w) \cdot \text{vol}_{n-1}(P(\lambda)_w)$$

which is a homogeneous polynomial of degree n in $\lambda_1, \dots, \lambda_r$, as $h_{P(\lambda)}(w)$ is linear in $\lambda_1, \dots, \lambda_r$. Again using that volume is invariant under translation, now suppose that $0 \in P(\lambda)$, and thus the support function of $P(\lambda)$ is nonnegative for all $w \in \mathbb{R}^n$. Then the pyramids over facets of $P(\lambda)$ form a polyhedral subdivision of $P(\lambda)$, so that $\text{vol}(P(\lambda))$ is the sum of the volumes of these pyramids. This completes the proof. \square

Let us write the polynomial $\text{vol}(P(\lambda))$ as a tensor (nonsymmetric in $\lambda_1, \dots, \lambda_r$),

$$\text{vol}(P(\lambda)) = \sum_{a_1, \dots, a_n=1}^r \text{MV}(P_{a_1}, P_{a_2}, \dots, P_{a_n}) \lambda_{a_1} \lambda_{a_2} \cdots \lambda_{a_n}, \quad (\text{A.10})$$

where the coefficients are chosen to be symmetric—for any permutation $\pi \in S_n$, we have

$$\text{MV}(P_{a_1}, P_{a_2}, \dots, P_{a_n}) = \text{MV}(P_{\pi(a_1)}, P_{\pi(a_2)}, \dots, P_{\pi(a_n)}).$$

The coefficient $\text{MV}(P_{a_1} \dots, P_{a_n})$ is the *mixed volume* of the polytopes P_{a_1}, \dots, P_{a_n} .

Lemma A.3.5. *Mixed volumes satisfy the following properties. Let $P, Q, P_1, \dots, P_n \subset \mathbb{R}^n$ be polytopes.*

1. *Symmetry.* $\text{MV}(P_{a_1}, \dots, P_{a_n}) = \text{MV}(P_{\pi(a_1)}, \dots, P_{\pi(a_n)})$ for any permutation $\pi \in S_n$.

2. *Multilinearity.* For any nonnegative λ, μ , we have

$$\text{MV}(\lambda P + \mu Q, P_2, \dots, P_n) = \lambda \text{MV}(P, P_2, \dots, P_n) + \mu \text{MV}(Q, P_2, \dots, P_n).$$

3. *Normalization.* $\text{MV}(P, \dots, P) = \text{vol}_n(P)$.

Proof. Symmetry follows from the definition of mixed volume. For multilinearity, equate the coefficient of $\lambda_1 \dots \lambda_n$ in the nonsymmetric expansions (A.10) of

$$\text{vol}(\lambda_1(\lambda P + \mu Q) + P_2 + \dots + P_n) = \text{vol}(\lambda_1 \lambda P + \lambda_1 \mu Q + P_2 + \dots + P_n).$$

(For the first, $r = n$ and for the second, $r = n+1$ in (A.10).) Finally, for normalization, note that for $\lambda \geq 0$, $\lambda^n \text{vol}(P) = \text{vol}(\lambda P) = \lambda^n \text{MV}(P, \dots, P)$, with the first equality coming from the definition of volume and the second from the expansion (A.10) defining mixed volume. \square

These three properties characterize mixed volumes.

Corollary A.3.6. *Mixed volume is the unique function of n -tuples of polytopes in \mathbb{R}^n that satisfies the properties of symmetry, multilinearity, and normalization of Lemma A.3.5.*

Proof. Let L be a function of n -tuples of polytopes in \mathbb{R}^n that satisfies the three properties of symmetry, multilinearity, and normalization of Lemma A.3.5. For any polytopes $P_1, \dots, P_n \subset \mathbb{R}^n$ and nonnegative $\lambda_1, \dots, \lambda_n$, we have $\text{vol}(P(\lambda)) = L(P(\lambda), \dots, P(\lambda))$ by normalization. Expanding this using (A.8) and the multilinearity of L , we obtain

$$L(P(\lambda), \dots, P(\lambda)) = \sum_{a_1, \dots, a_n=1}^n L(P_{a_1}, P_{a_2}, \dots, P_{a_n}) \lambda_{a_1} \lambda_{a_2} \dots \lambda_{a_n}.$$

The equality of this sum with the sum (A.10) and the symmetry of both L and MV in their arguments completes the proof. \square

We give another formula for mixed volume and prove a stronger version of Corollary A.3.6. Given polytopes P_1, \dots, P_n and $\emptyset \neq A \subset [n]$ write $P(A)$ for the Minkowski sum $\sum_{i \in A} P_i$.

Theorem A.3.7. *Let \mathcal{P} be a collection of polytopes in \mathbb{R}^n that is closed under Minkowski sum. Suppose that L is a function of n -tuples of polytopes in \mathcal{P} that is symmetric in its arguments and normalized (as in Lemma A.3.5), and that L is multilinear under Minkowski sum ($\lambda = \mu = 1$ in Lemma A.3.5). Then for any polytopes $P_1, \dots, P_n \in \mathcal{P}$, we have*

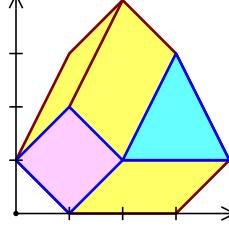
$$n! L(P_1, \dots, P_n) = \sum_{\emptyset \neq A \subset [n]} (-1)^{n-|A|} \text{vol}(P(A)). \quad (\text{A.11})$$

In particular, L equals mixed volume, $L(P_1, \dots, P_n) = \text{MV}(P_1, \dots, P_n)$.

Example A.3.8. If P, Q, R are polytopes in \mathbb{R}^3 , then $6 \text{MV}(P, Q, R)$ equals

$$\text{vol}(P + Q + R) - \text{vol}(P + Q) - \text{vol}(P + R) - \text{vol}(Q + R) + \text{vol}(P) + \text{vol}(Q) + \text{vol}(R).$$

For polygons P, Q , we have $2 \text{MV}(P, Q) = \text{vol}(P + Q) - \text{vol}(P) - \text{vol}(Q)$. For the polygons in Figure A.1, if we subdivide $P + Q$ as shown,



then $2 \text{MV}(P, Q)$ equals the combined areas of the four parallelograms, which is six.

Proof of Theorem A.3.7. Let $\emptyset \neq A \subset [n]$. Since L is normalized, $L(P(A), \dots, P(A))$ equals $\text{vol}(P(A))$. Expand $L(P(A), \dots, P(A))$ using the multilinearity of L to obtain

$$\text{vol}(P(A)) = \sum_{a_1, \dots, a_n \in A} L(P_{a_1}, \dots, P_{a_n}). \quad (\text{A.12})$$

Let b_1, \dots, b_n be any sequence with $b_i \in [n]$ and set $B := \{b_1, \dots, b_n\}$. Then $L(P_{b_1}, \dots, P_{b_n})$ occurs in the sum (A.12) if and only if $B \subset A$, and in that case, it appears with coefficient 1.

Expand the right hand side of (A.11) in terms of the function L using (A.12). Then for $b_1, \dots, b_n \in [n]$ the term $L(P_{b_1}, \dots, P_{b_n})$ occurs with coefficient

$$\sum_{B \subset A \subset [n]} (-1)^{n-|A|} = (1-1)^{n-|B|} = \begin{cases} 0 & \text{if } B \neq [n] \\ 1 & \text{if } B = [n] \end{cases}.$$

Thus the right hand side of (A.11) reduces to the sum of $L(P_{b_1}, \dots, P_{b_n})$ for b_1, \dots, b_n distinct. Each of these $n!$ terms are equal by symmetry, which completes the proof. \square

A.3.3 Positive semidefinite matrices

In the text we denote by $\text{Sym}_n(\mathbb{R})$ the set of all symmetric $n \times n$ -matrices. A matrix $A \in \text{Sym}_n(\mathbb{R})$ 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.3.9. 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.3.10. 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 submatrices $A_{\{1,\dots,k\},\{1,\dots,k\}}$) 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 $v^{(1)}, \dots, v^{(n)}$ be an orthonormal system of eigenvectors with respect to the eigenvalues $\lambda_1, \dots, \lambda_n$. Then

$$A = SDS^T \text{ with } S := (v^{(1)}, \dots, v^{(n)}), \quad D = \text{diag}(\lambda_1, \dots, \lambda_n).$$

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

$$\begin{aligned} \langle A, B \rangle &:= \text{Tr}(A^T B) = \text{Tr}(B^T A) = \text{Tr}(AB^T) = \text{Tr}(BA^T) \\ &= \text{vec}(A)^T \text{vec}(B), \end{aligned}$$

where $\text{vec}(A) := (a_{11}, a_{21}, \dots, a_{n1}, a_{12}, a_{22}, \dots, a_{nn})^T$ and Tr denotes the trace.

For $A \in \mathbb{R}^{n \times n}$ the definition $\|A\|_F^2 := \langle A, A \rangle = \text{Tr}(A^T A) = \sum_{i,j=1}^n a_{ij}^2$ defines the *Frobenius norm* on $\mathbb{R}^{n \times n}$. If $A \in \text{Sym}_n$ with eigenvalues $\lambda_1, \dots, \lambda_n$, then $\|A\|_F^2 = \sum_{i=1}^n \lambda_i^2$.

Theorem A.3.11. (Féjer.) A matrix $A \in S_n$ is positive semidefinite if and only if $\text{Tr}(AB) \geq 0$ 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 $\text{Tr}(AB) = \text{Tr}(A^{1/2} A^{1/2} B^{1/2} B^{1/2}) = \text{Tr}(A^{1/2} B^{1/2} B^{1/2} A^{1/2})$. Since A and B are symmetric, this implies $\text{Tr}(AB) = \|A^{1/2} B^{1/2}\|_F^2 \geq 0$.

Conversely, let $A \in S_n$ and $\text{Tr}(AB) \geq 0$ for all $B \in S_n^+$. Moreover, let $x \in \mathbb{R}^n$. For $B := xx^T \in S_n^+$ this implies $0 \leq \text{Tr}(AB) = \text{Tr}(Axx^T) = \sum_{i,j=1}^n a_{ij}x_i x_j = x^T Ax$. \square

Theorem A.3.12. (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 $C - B^T A^{-1} B$ is positive (semi-)definite. The matrix $C - B^T A^{-1} B$ is called the Schur complement of A in M .

Proof. For $D := -A^{-1}B$ we have

$$\begin{pmatrix} I & 0 \\ D^T & I \end{pmatrix} \underbrace{\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}}_{=M^T=M} \begin{pmatrix} I & D \\ 0 & I \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & C - B^T A^{-1} B \end{pmatrix}.$$

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

$$X \succeq 0 \iff C^T X C \succeq 0 \text{ for all } C \in \mathbb{R}^{n \times n}.$$

□

A.4 Complex analysis

Let f be a univariate meromorphic function inside and on some closed contour $C \subset \mathbb{C}$. Then the *argument principle* states that the number Z of zeroes and the number P of poles inside the contour (counting multiplicities) is

$$\int_C \frac{f'(x)}{f(x)} dx = 2\pi i(Z - P).$$

If f and g are univariate holomorphic functions with $|f(z)| < |g(z)|$ on a circle C , then *Rouche's Theorem* states that f and $f + g$ have the same number of zeroes in the interior of C , counting multiplicity.

A.5 Moments

We collect some ideas and results from the theory of moments, see, e.g., [39] or [127]. Let $(y_k)_{k \in \mathbb{N}}$ be a sequence of real numbers. The *Hausdorff moment problem* asks to characterize when there exists a probability measure μ on the unit interval $[0, 1]$ such that y_k is the k -th moment of μ for all $k \geq 0$, that is,

$$y_k = \int_0^1 x^k d\mu \quad \text{for } k \geq 0.$$

For any sequence (a_k) of real numbers, let Δ be the difference operator defined by $\Delta a_k = a_{k+1} - a_k$. Applying this operator to the sequence (Δa_k) gives another sequence $(\Delta^2 a_k)$ and, recursively, define the r -th iterated difference by $\Delta^r = \Delta(\Delta^{r-1})$ for $r \geq 2$, and set $\Delta^1 = \Delta$ as well as $\Delta^0 a_k = a_k$. Inductively, this gives

$$\Delta^r a_j = \sum_{k=0}^r \binom{r}{k} (-1)^{r+k} c_{j+k}$$

and the inversion formula

$$a_j = \sum_{k=0}^r \binom{r}{k} (-1)^{r-k} \Delta^{r-k} a_{j+k}. \quad (\text{A.13})$$

Theorem A.5.1 (Hausdorff). *A sequence $(y_k)_{k \in \mathbb{N}}$ of real numbers is the sequence of moments of some probability measure on $[0, 1]$ if and only if for all $r \geq 1$*

$$(-1)^{-r} \Delta^r y_k \geq 0 \quad \text{and} \quad y_0 = 1,$$

or, equivalently, for all $r \geq 1$

$$\sum_{k=0}^r \binom{r}{k} (-1)^k y_{j+k} \geq 0 \quad \text{and} \quad y_0 = 1.$$

The necessity of this condition can be immediately seen as follows. By taking differences, we obtain $-y_k = \int_0^1 x^k (1-x)^r d\mu$, as well as inductively,

$$(-1)^r \Delta^r y_k = \int_0^1 x^k (1-x)^r d\mu,$$

and this integral is clearly non-negative. We sketch the sufficiency. Setting

$$p_k^{(n)} = \binom{n}{k} (-1)^{n-k} \Delta^{n-k} y_k, \quad (\text{A.14})$$

we obtain

$$\sum_{k=0}^n \binom{k}{j} p_k^{(n)} = \sum_{k=j}^n \binom{k}{j} p_k^{(n)} = \sum_{k=0}^{n-j} \binom{j+k}{j} \binom{n}{j+k} (-1)^{(n-j)-k} \Delta^{(n-j)-k} y_{j+k},$$

and thus the inversion formula (A.13) with $r = n - j$ gives

$$\sum_{k=0}^n \binom{k}{j} p_k^{(n)} = \binom{n}{j} y_j. \quad (\text{A.15})$$

For $j = 0$, this shows that $\sum_{k=0}^n p_k^{(n)} = y_0 = 1$, so that we can interpret the $p_k^{(n)}$ as a discrete probability distribution D_n assigning weight $p_k^{(n)}$ to the point $\frac{k}{n}$. The left hand side of (A.15) gives the expected value of the random variable $\binom{n}{j} X^j$ with respect to this distribution D_n .

Now consider the moments of the distribution F_n . We have the expectation

$$\mathbb{E}[X^j] = \sum_{k=0}^n \left(\frac{k}{n}\right)^j p_k(n) \quad \text{for } j \geq 0.$$

In an elementary way, we see that $\mathbb{E}[X^j]$ converges for $n \rightarrow \infty$ to the j -th moment of the left-hand side of (A.15) multiplied by $j!n^{-j}$ and thus, due to (A.15), to y_j . Moreover, it can be shown that the sequence (y_j) thus occurs as the sequence of moments of the limit of the distributions D_n .

Theorem A.5.1 has the following multivariate version, where for $\alpha, \gamma \in \mathbb{N}^n$ we write

$$\binom{\beta}{\gamma} = \prod_{i=1}^n \binom{\beta_i}{\gamma_i}$$

Theorem A.5.2 ([52]). *A real sequence $(y_\alpha)_{\alpha \in \mathbb{N}^n}$ is the sequence of moments of some probability measure on $[0, 1]^n$ if and only if for all $\alpha, \beta \in \mathbb{N}^n$ we have*

$$\sum_{\gamma \in \mathbb{N}^n, |\gamma| \leq \beta} \binom{\beta}{\gamma} (-1)^{|\gamma|} y_{\alpha+\gamma} \geq 0 \quad \text{and } y_0 = 1.$$

A variation of this gives the characterization of the moments for the standard simplex $\Delta_n = \{x \in \mathbb{R}_{\geq 0}^n : \sum_{i=1}^n x_i \leq 1\}$.

Theorem A.5.3 ([51, 136]). *A real sequence $(y_\alpha)_{\alpha \in \mathbb{N}^n}$ is the sequence of moments of some probability measure on Δ_n if and only if for all $t \geq 0$*

$$\sum_{\alpha \in \mathbb{N}^n, |\alpha| \leq t} (-1)^{|\alpha|} \begin{bmatrix} t \\ \alpha \end{bmatrix} y_{\alpha+\beta} \geq 0 \quad \text{for all } \beta \in \mathbb{N}^n \quad \text{and } y_0 = 1,$$

where

$$\begin{bmatrix} t \\ \alpha \end{bmatrix} = \begin{bmatrix} t \\ \alpha_1 \cdots \alpha_n \end{bmatrix} = \frac{t!}{\alpha_1! \alpha_2! \cdots \alpha_n! (t - |\alpha|)!} = \binom{|\alpha|}{\alpha_1 \cdots \alpha_n} \binom{t}{|\alpha|}$$

is the pseudo multinomial coefficient of dimension n and order t .

If $(y_\alpha)_{\alpha \in \mathbb{N}^n}$ is a sequence of moments for some measure μ , then μ is called **determinate** if it is the unique representing for $(y_\alpha)_{\alpha \in \mathbb{N}^n}$. Every measure with compact support is determinate.

Bibliography

- [1] W. W. Adams and P. Loustaunau, *An introduction to Gröbner bases*, Graduate Studies in Mathematics, vol. 3, American Mathematical Society, Providence, RI, 1994. MR 1287608 (95g:13025)
- [2] F. Aroca, *Krull-tropical hypersurfaces*, Ann. Fac. Sci. Toulouse Math. (6) **19** (2010), no. 3-4, 525–538. MR 2790807
- [3] E. Artin and O. Schreier, *Algebraische Konstruktion reeller Körper*, Abh. Math. Sem. Univ. Hamburg **5** (1927), no. 1, 85–99. MR 3069467
- [4] G. Averkov, *Constructive proofs of some Positivstellensätze for compact semialgebraic subsets of \mathbb{R}^d* , J. Optim. Theory Appl. **158** (2013), no. 2, 410–418. MR 3084384
- [5] S. Basu, R. Pollack, and M.-F. Roy, *Algorithms in real algebraic geometry*, second ed., Algorithms and Computation in Mathematics, vol. 10, Springer-Verlag, Berlin, 2006. MR 2248869 (2007b:14125)
- [6] M.C. Beltrametti, E. Carletti, D. Gallarati, and G. Monti Bragadin, *Lectures on curves, surfaces and projective varieties*, EMS Textbooks in Mathematics, European Mathematical Society (EMS), Zürich, 2009. MR 2549804 (2010k:14001)
- [7] 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)
- [8] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic analysis on semigroups*, Graduate Texts in Mathematics, vol. 100, Springer-Verlag, New York, 1984. MR 747302 (86b:43001)
- [9] S. Bernštein, *Sur la représentation des polynomes positifs*, Charikov, Comm. Soc. Math. (2) **14** (1915), 227–228.
- [10] R. Berr and T. Wörmann, *Positive polynomials on compact sets*, Manuscripta Math. **104** (2001), no. 2, 135–143. MR 1821179
- [11] B. Bertrand and F. Bihan, *Euler characteristic of real nondegenerate tropical complete intersections*, Preprint, arXiv:math/0710.1222, 2007.

- [12] E. Bézout, *Théorie générale des équations algébriques*, Ph.-D. Pierres, 1779.
- [13] ———, *General theory of algebraic equations*, Princeton University Press, 2006, Translated from French original by Eric Feron.
- [14] G. Blekherman, P. A. Parrilo, and R. R. Thomas (eds.), *Semidefinite optimization and convex algebraic geometry*, MOS-SIAM Series on Optimization, vol. 13, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2013. MR 3075433
- [15] G. Blekherman, D. Plaumann, R. Sinn, and C. Vinzant, *Low-rank sum-of-squares representations on varieties of minimal degree*, To appear in International Mathematics Research Notices, 2016.
- [16] J. Bochnak, M. Coste, and M.-F. Roy, *Real Algebraic Geometry*, Ergebnisse der Mathematik und ihrer Grenzgebiete, vol. 36, Springer-Verlag, Berlin, 1998. MR 1659509 (2000a:14067)
- [17] C. Borcea, X. Goaoc, S. Lazard, and S. Petitjean, *Common tangents to spheres in \mathbb{R}^3* , Discrete Comput. Geom. **35** (2006), no. 2, 287–300. MR 2195056
- [18] J. Borcea and P. Brändén, *Applications of stable polynomials to mixed determinants: Johnson’s conjectures, unimodality, and symmetrized Fischer products*, Duke Math. J. **143** (2008), no. 2, 205–223. MR 2420507 (2009b:15015)
- [19] L. 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)
- [20] B. Buchberger, *Ein algorithmisches Kriterium für die Lösbarkeit eines algebraischen Gleichungssystems*, Aequationes Math. **4** (1970), 374–383.
- [21] ———, *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.
- [22] A. Bunse-Gerstner, R. Byers, and V. Mehrmann, *Numerical methods for simultaneous diagonalization*, SIAM J. Matrix Anal. Appl. **14** (1993), no. 4, 927–949. MR 1238912 (94h:65036)
- [23] M. D. Choi and T. Y. Lam, *Extremal positive semidefinite forms*, Math. Ann. **231** (1977/78), no. 1, 1–18. MR 0498384 (58 #16512)
- [24] G. E. Collins, *Quantifier elimination for real closed fields by cylindrical algebraic decomposition*, Automata theory and formal languages (Second GI Conf., Kaiserslautern, 1975), Springer, Berlin, 1975, pp. 134–183. Lecture Notes in Comput. Sci., Vol. 33. MR 0403962

- [25] D. Cox, J. Little, and D. O’Shea, *Ideals, varieties, and algorithms*, third ed., Undergraduate Texts in Mathematics, Springer, New York, 2007.
- [26] D. A. Cox, J. Little, and D. O’Shea, *Using algebraic geometry*, second ed., Graduate Texts in Mathematics, vol. 185, Springer, New York, 2005.
- [27] R. E. Curto and L. A. Fialkow, *Flat extensions of positive moment matrices: recursively generated relations*, Mem. Amer. Math. Soc. **136** (1998), no. 648, x+56. MR 1445490 (99d:47015)
- [28] ———, *The truncated complex K-moment problem*, Trans. Amer. Math. Soc. **352** (2000), no. 6, 2825–2855. MR 1661305 (2000j:47027)
- [29] E. 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)
- [30] R. Descartes, *La géometrie (discours de la méthode, third part)*, 1637, available via Project Gutenberg, <http://www.gutenberg.org/ebooks/26400>.
- [31] A. Dickenstein and I. 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)
- [32] P. Dietmaier, *The Stewart-Gough platform of general geometry can have 40 real postures*, Advances in robot kinematics: analysis and control (Salzburg, 1998), Kluwer Acad. Publ., Dordrecht, 1998, pp. 7–16. MR 1643130
- [33] J. Draisma, *A tropical approach to secant dimensions*, J. Pure Appl. Algebra **212** (2008), no. 2, 349–363. MR MR2357337 (2008j:14102)
- [34] M. Einsiedler, M. M. Kapranov, and D. Lind, *Non-archimedean amoebas and tropical varieties*, J. Reine Angew. Math. **601** (2006), 139–157.
- [35] D. Eisenbud, *Commutative algebra*, Graduate Texts in Mathematics, vol. 150, Springer-Verlag, New York, 1995, With a view toward algebraic geometry. MR 1322960 (97a:13001)
- [36] G. Ewald, *Combinatorial convexity and algebraic geometry*, Graduate Texts in Mathematics, vol. 168, Springer-Verlag, New York, 1996.
- [37] J.-C. Faugère, *FGb*, See <http://fgbtrs.lip6.fr/jcf/Software/FGb/index.html>.
- [38] 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.

- [39] W. Feller, *An introduction to probability theory and its applications. Vol. II*, Second edition, John Wiley & Sons, Inc., New York-London-Sydney, 1971. MR 0270403
- [40] G. Fischer and J. Piontkowski, *Ruled varieties*, Advanced Lectures in Mathematics, Friedr. Vieweg & Sohn, Braunschweig, 2001, An introduction to algebraic differential geometry. MR 1876644
- [41] L. Gårding, *Linear hyperbolic partial differential equations with constant coefficients*, Acta Math. **85** (1951), 1–62. MR 0041336
- [42] ———, *An inequality for hyperbolic polynomials*, J. Math. Mech. **8** (1959), 957–965.
- [43] D. R. Grayson and M. E. Stillman, *Macaulay2, a software system for research in algebraic geometry*, Available at <http://www.math.uiuc.edu/Macaulay2/>.
- [44] 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>.
- [45] W. Habicht, *Über die Zerlegung strikter definiter Formen in Quadrate.*, Comment. Math. Helv. **12** (1940), 317–322 (German).
- [46] D Handelman, *Representing polynomials by positive linear functions on compact convex polyhedra*, Pacific J. Math. **132** (1988), no. 1, 35–62. MR 929582 (90e:52005)
- [47] 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)
- [48] J. Harris, *Algebraic geometry*, Graduate Texts in Mathematics, vol. 133, Springer-Verlag, New York, 1992, A first course. MR 1182558 (93j:14001)
- [49] R. Hartshorne, *Algebraic geometry*, Springer-Verlag, New York, 1977, Graduate Texts in Mathematics, No. 52. MR 0463157 (57 #3116)
- [50] E. K. Haviland, *On the momentum problem for distribution functions in more than one dimension. II*, Amer. J. Math. **58** (1936), no. 1, 164–168. MR 1507139
- [51] K. Helmes and S. Röhl, *A geometrical characterization of multidimensional Hausdorff and dale polytopes with applications to exit time problems*, Technical Report ZIB 04-05, Konrad-Zuse-Zentrum für Informationstechnik, Berlin, 2004.
- [52] ———, *A geometrical characterization of multidimensional Hausdorff polytopes with applications to exit time problems*, Math. Oper. Res. **33** (2008), no. 2, 315–326. MR 2415995

- [53] B. Helton and V. Vinnikov, *Linear matrix inequality representation of sets*, Communications on Pure and Applied Mathematics **60** (2007), 654–674.
- [54] J. W. Helton and M. Putinar, *Positive polynomials in scalar and matrix variables, the spectral theorem, and optimization*, Operator theory, structured matrices, and dilations, Theta Ser. Adv. Math., vol. 7, Theta, Bucharest, 2007, pp. 229–306. MR 2389626
- [55] R. Hemmecke and P. Malkin, *Computing generating sets of lattice ideals*, math.CO/0508359.
- [56] D. Henrion and J.-B. Lasserre, *Detecting global optimality and extracting solutions in GloptiPoly*, Positive polynomials in control, Lect. Notes Control Inf. Sci., vol. 312, Springer, Berlin, 2005, pp. 293–310. MR 2123528
- [57] K. Hept and T. Theobald, *Tropical bases by regular projections*, Proc. Amer. Math. Soc. **137** (2009), no. 7, 2233–2241. MR MR2495256
- [58] C. Hermite, *Remarques sur le théorème de sturm*, C. R. Acad. Sci. Paris **36** (1853), 52–54.
- [59] D. Hilbert, *Herrmann minkowski*, Math. Annalen **68** (1910), 445–471.
- [60] H. Hironaka, *Resolution of singularities of an algebraic variety over a field of characteristic zero*, Ann. Math. **79** (1964), 109–326.
- [61] W. V. D. Hodge and D. Pedoe, *Methods of algebraic geometry. Vol. I*, Cambridge, at the University Press; New York, The Macmillan Company, 1947. MR 0028055
- [62] A. Holme, *A royal road to algebraic geometry*, Springer, Heidelberg, 2012. MR 2858123
- [63] K. 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)
- [64] S. Ilman and T. de Wolff, *Amoebas, nonnegative polynomials and sums of squares supported on circuits*, Res. Math. Sci. **3** (2016), Paper No. 9, 35. MR 3481195
- [65] N. V. Ilyushechkin, *The discriminant of the characteristic polynomial of a normal matrix*, Mat. Zametki **51** (1992), no. 3, 16–23, 143. MR 1172221
- [66] T. Jacobi and A. Prestel, *Distinguished representations of strictly positive polynomials*, J. Reine Angew. Math. **532** (2001), 223–235. MR 1817508 (2001m:14080)
- [67] A. Jensen, *Algorithmic aspects of gröbner fans and tropical varieties*, Ph.D. thesis, University of Aarhus, 2007.

- [68] A. N. Jensen, H. Markwig, and T. Markwig, *An algorithm for lifting points in a tropical variety*, Collect. Math. **59** (2008), no. 2, 129–165. MR MR2414142 (2009a:14077)
- [69] M. Joswig and T. Theobald, *Polyhedral and algebraic methods in computational geometry*, Universitext, Springer, London, 2013, Revised and updated translation of the 2008 German original. MR 2905853
- [70] S. Karlin and W. 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)
- [71] E. Katz, *A tropical toolkit*, Expo. Math. **27** (2009), no. 1, 1–36. MR MR2503041 (2010f:14069)
- [72] János Kollár, *Sharp effective nullstellensatz*, J. Amer. Math. Soc. **1** (1988), no. 4, 963–975.
- [73] J.-L. Krivine, *Anneaux préordonnés*, J. Analyse Math. **12** (1964), 307–326. MR MR0175937 (31 #213)
- [74] H. W. Kuhn, *Solvability and consistency for linear equations and inequalities*, Amer. Math. Monthly **63** (1956), 217–232. MR 0081538
- [75] J. 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)
- [76] ———, *Moments, positive polynomials and their applications*, Imperial College Press Optimization Series, vol. 1, Imperial College Press, London, 2010. MR 2589247
- [77] M. 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)
- [78] ———, *Semidefinite representations for finite varieties*, Math. Program. **109** (2007), no. 1, Ser. A, 1–26. MR 2291590 (2008g:90094)
- [79] ———, *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)
- [80] D. Lazard, *On the representations of rigid-body motions and its application to generalized platform manipulators.*, Computational Kinematics (P. Kovács J. Angeles, G. Hommel, ed.), Solid Mechanics and Its Applications, vol. volume 28, Kluwer, Dordrecht, 1993, pp. 175–182.

- [81] A.S. Lewis, P.A. Parrilo, and M.V. Ramana, *The Lax conjecture is true*, Proc. Amer. Math. Soc. **133** (2005), 2495–2499.
- [82] T. Y. Li, T. 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.
- [83] F. 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)
- [84] F.S. Macaulay, *Some properties of enumeration in the theory of modular systems*, Proc. London Math. Soc. **26** (1927), 531–555.
- [85] I. G. Macdonald, J. Pach, and T. Theobald, *Common tangents to four unit balls in \mathbb{R}^3* , Discrete Comput. Geom. **26** (2001), no. 1, 1–17. MR 1832726
- [86] Diane Maclagan and Bernd Sturmfels, *Introduction to tropical geometry*, Graduate Studies in Mathematics, vol. 161, American Mathematical Society, Providence, RI, 2015. MR 3287221
- [87] T. Markwig, *A field of generalised puiseux series for tropical geometry*, Rend. Semin. Mat. Torino **xx** (xx), xx.
- [88] M. Marshall, *Positive polynomials and sums of squares*, Mathematical Surveys and Monographs, vol. 146, American Mathematical Society, Providence, RI, 2008. MR 2383959 (2009a:13044)
- [89] 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
- [90] G. Megyesi, *Lines tangent to four unit spheres with coplanar centres*, Discrete Comput. Geom. **26** (2001), no. 4, 493–497. MR 1863804
- [91] E. Meissner, *Über positive Darstellungen von Polynomen*, Math. Ann. **70** (1911), no. 2, 223–235. MR 1511619
- [92] G. Mikhalkin, *Enumerative tropical algebraic geometry in \mathbb{R}^2* , J. Amer. Math. Soc. **18** (2005), no. 2, 313–377. MR MR2137980 (2006b:14097)
- [93] T. Netzer, D. Plaumann, and M. Schweighofer, *Exposed faces of semidefinitely representable sets*, SIAM J. Optim. **20** (2010), 1944–1955.

- [94] J. Nie, P. A. Parrilo, and B. Sturmfels, *Semidefinite representation of the k -ellipse*, Algorithms in algebraic geometry (A. Sommese A. Dickenstein, F.-O. Schreyer, ed.), The IMA Volumes in Mathematics and its Applications, vol. 146, Springer, New York, 2008, pp. 117–132.
- [95] P. 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)
- [96] P. A. Parrilo and B. 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)
- [97] P.A. Parrilo, *An explicit construction of distinguished representations of polynomials nonnegative over finite sets*, ETH Zürich, IfA Technical Report AUT02-02, 2002.
- [98] S. Payne, *Fibers of tropicalization*, Math. Z. **262** (2009), no. 2, 301–311. MR 2504879
- [99] ———, *Erratum to: Fibers of tropicalization*, Math. Z. **272** (2012), no. 3-4, 1403–1406. MR 2995174
- [100] P. Pedersen, M.-F. Roy, 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)
- [101] R. Pemantle, *Hyperbolicity and stable polynomials in combinatorics and probability*, Current developments in mathematics, 2011, Int. Press, Somerville, MA, 2012, pp. 57–123. MR 3098077
- [102] D. Perrin, *Algebraic geometry*, Universitext, Springer-Verlag London Ltd., London, 2008, An introduction, Translated from the 1995 French original by Catriiona Maclean. MR 2372337 (2008k:14001)
- [103] H. Poincaré, *Sur les équations algébriques.*, C. R. XCVII, (1884), 1418–1419 (French).
- [104] G. Pólya, *Über positive darstellung von polynomen*, Vierteljahrsschrift der Naturforschenden Gesellschaft in Zürich **73** (1928), 141–145, Reprinted in: Collected Papers, Volume 2, 309–313, MIT Press, Cambridge.
- [105] 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)

- [106] L. Porkolab and L. Khachiyan, *On the complexity of semidefinite programs*, J. Global Optim. **10** (1997), no. 4, 351–365. MR 1457182
- [107] H. Pottmann and J. Wallner, *Computational line geometry*, Mathematics and Visualization, Springer-Verlag, Berlin, 2001. MR 1849803
- [108] V. Powers and B. Reznick, *Polynomials that are positive on an interval*, Trans. Amer. Math. Soc. **352** (2000), no. 10, 4677–4692. MR 1707203 (2001b:12002)
- [109] V. Powers, B. Reznick, C. Scheiderer, and F. Sottile, *A new approach to hilbert’s theorem on ternary quartics*, Comptes rendus. Mathématique **339** (2004), no. 9, 617–620.
- [110] A. Prestel and C. N. Delzell, *Positive polynomials*, Springer Monographs in Mathematics, Springer-Verlag, Berlin, 2001, From Hilbert’s 17th problem to real algebra. MR 1829790 (2002k:13044)
- [111] Q. I. Rahman and G. Schmeisser, *Analytic theory of polynomials*, London Math. Society Monographs, vol. 26, Clarendon Press, Oxford, 2002.
- [112] M. Ramana and A. J. Goldman, *Some geometric results in semidefinite programming*, Journal of Global Optimization **7** (1995), 33–50.
- [113] B. Reznick, *Extremal PSD forms with few terms*, Duke Math. J. **45** (1978), no. 2, 363–374. MR 0480338 (58 #511)
- [114] 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.
- [115] M. Riesz, *Sur le problème des moments. III.*, Ark. Mat. Astron. Fys. **17** (1923), no. 16, 52 (French).
- [116] R. M. Robinson, *Some definite polynomials which are not sums of squares of real polynomials*, Selected questions of algebra and logic (collection dedicated to the memory of A. I. Mal’cev) (Russian), Izdat. “Nauka” Sibirsk. Otdel., Novosibirsk, 1973, pp. 264–282. MR 0337878
- [117] R. T. Rockafellar, *Convex analysis*, Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, N.J., 1970. MR 0274683 (43 #445)
- [118] Felice Ronga and Thierry Vust, *Stewart platforms without computer?*, Real analytic and algebraic geometry (Trento, 1992), de Gruyter, Berlin, 1995, pp. 197–212. MR 1320320
- [119] C. Scheiderer, *Spectrahedral shadows*, SIAM J. Appl. Algebra Geom. **2** (2018), no. 1, 26–44. MR 3755652

- [120] K. 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)
- [121] ———, *The K-moment problem for compact semi-algebraic sets*, Math. Ann. **289** (1991), no. 2, 203–206. MR 1092173 (92b:44011)
- [122] R. Schneider, *Convex bodies: The Brunn-Minkowski theory*, Encyclopedia of Mathematics and its Applications, vol. 44, Cambridge University Press, Cambridge, 1993. MR MR1216521 (94d:52007)
- [123] A. Schrijver, *Theory of linear and integer programming*, Wiley-Interscience Series in Discrete Mathematics, John Wiley & Sons Ltd., Chichester, 1986. MR 874114 (88m:90090)
- [124] M. Schweighofer, *An algorithmic approach to Schmüdgen’s Positivstellensatz*, J. Pure Appl. Algebra **166** (2002), no. 3, 307–319. MR 1870623 (2002j:14063)
- [125] ———, *Optimization of polynomials on compact semialgebraic sets*, SIAM J. Optim. **15** (2005), no. 3, 805–825 (electronic). MR 2142861 (2006d:90136)
- [126] I. 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)
- [127] J.A. Shohat and J.D. Tamarkin, *The Problem of Moments*, American Mathematical Society Mathematical surveys, vol. I, American Mathematical Society, New York, 1943. MR 0008438
- [128] N.Z. Shor, *Quadratic optimization problems*, Izv. Akad. Nauk SSSR Tekhn. Kibernet. (1987), no. 1, 128–139, 222. MR 939596
- [129] K. E. Smith, L. Kahanpää, P. Kekäläinen, and W. Traves, *An invitation to algebraic geometry*, Universitext, Springer-Verlag, New York, 2000. MR 1788561 (2001k:14002)
- [130] 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)
- [131] F. Sottile, *Real solutions to equations from geometry*, University Lecture Series, vol. 57, American Mathematical Society, Providence, RI, 2011. MR 2830310
- [132] F. Sottile and T. Theobald, *Lines tangent to $2n - 2$ spheres in \mathbb{R}^n* , Trans. Amer. Math. Soc. **354** (2002), no. 12, 4815–4829. MR 1926838

- [133] 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)
- [134] D. Speyer and B. Sturmfels, *The tropical Grassmannian*, Adv. Geom. **4** (2004), no. 3, 389–411. MR MR2071813 (2005d:14089)
- [135] G. Stengle, *A Nullstellensatz and a Positivstellensatz in semialgebraic geometry*, Math. Ann. **207** (1974), 87–97. MR MR0332747 (48 \#11073)
- [136] R. H. Stockbridge, *The problem of moments on polytopes and other bounded regions*, J. Math. Anal. Appl. **285** (2003), no. 2, 356–375. MR 2005126
- [137] J. C. F. Sturm, *Mémoire sur la résolution des équations numériques*, Bull. Sci. Féruccac **11** (1829), 419–425.
- [138] 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.
- [139] B. Sturmfels and J. Tevelev, *Elimination theory for tropical varieties*, Math. Res. Lett. **15** (2008), no. 3, 543–562. MR MR2407231 (2009f:14124)
- [140] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev. **38** (1996), no. 1, 49–95. MR 1379041 (96m:90005)
- [141] O. 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)
- [142] D.G. Wagner, *Multivariate stable polynomials: theory and applications*, Bull. Amer. Math. Soc. **48** (2011), no. 1, 53–84. MR 2738906 (2012d:32006)
- [143] R. J. Walker, *Algebraic curves*, Dover Publications Inc., New York, 1962. MR MR0144897 (26 \#2438)
- [144] G. M. Ziegler, *Lectures on polytopes*, Springer-Verlag, New York, 1995.