Invitation to Nonlinear Algebra

Mateusz Michałek
Bernd Sturmfels

Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, 04103 Leipzig, Germany
E-mail address: michalek@mis.mpg.de

Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, 04103 Leipzig, Germany
E-mail address: bernd@mis.mpg.de
2010 Mathematics Subject Classification. Primary:
Secondary

Key words and phrases. amsbook, AMS-\LaTeX
## Contents

**Preface** vii

**Chapter 1. Polynomial Rings**
- 1.1. Ideals 1
- 1.2. Gröbner Bases 5
- 1.3. Dimension and Degree 9
- Exercises 12

**Chapter 2. Varieties**
- 2.1. Affine varieties 15
- 2.2. Projective varieties 20
- 2.3. Geometry in Low Dimensions 22
- Exercises 28

**Chapter 3. Solving and Decomposing**
- 3.1. Zero-dimensional Ideals 31
- 3.2. Primary Decomposition 33
- 3.3. Linear PDE with Constant Coefficients 38
- Exercises 42

**Chapter 4. Mapping and Projecting**
- 4.1. Elimination 45
- 4.2. Implicitization 48
- 4.3. The Image of a Polynomial Map 51
- Exercises 52

**Chapter 5. Linear Spaces and Grassmannians**
- 5.1. Coordinates for Linear Spaces 55
- 5.2. Plücker Relations 56
- 5.3. Schubert Calculus 59
- Exercises 61

**Chapter 6. Nullstellensätze**
- 6.1. Certificates for Infeasibility 63
- 6.2. Hilbert’s Nullstellensatz 65
- 6.3. Let’s Get Real 66
- Exercises 69

**Chapter 7. Tropical Algebra**
- 7.1. Arithmetic 71
- 7.2. Linear Algebra 72
| 7.3. Towards Tropical Varieties | 77 |
| Ex exercises | 80 |
| Chapter 8. Toric Varieties | 83 |
| 8.1. Tori and Lattices | 83 |
| 8.2. Affine Toric Varieties | 84 |
| 8.3. Projective Toric Varieties | 87 |
| Exercises | 90 |
| Chapter 9. Tensors | 91 |
| 9.1. Eigenvectors | 91 |
| 9.2. Ranks of Tensors | 97 |
| 9.3. Matrix Multiplication | 100 |
| Exercises | 102 |
| Chapter 10. Representation Theory | 105 |
| 10.1. Irreducible Representations | 105 |
| 10.2. Schur-Weyl Duality | 112 |
| 10.3. Exploiting Symmetry | 113 |
| Exercises | 113 |
| Chapter 11. Invariant Theory | 115 |
| 11.1. Finite Groups | 115 |
| 11.2. Classical Invariant Theory | 118 |
| 11.3. Geometric Invariant Theory | 120 |
| Exercises | 123 |
| Chapter 12. Semidefinite Programming | 125 |
| 12.1. Spectrahedra | 125 |
| 12.2. Optimization and Duality | 128 |
| 12.3. Sums of Squares | 131 |
| Exercises | 134 |
| Chapter 13. Combinatorics | 135 |
| 13.1. Matroids | 135 |
| 13.2. Lattice Polytopes | 137 |
| 13.3. Generating Functions | 142 |
| Exercises | 146 |
| Bibliography | 147 |
| Index | 149 |
Preface

This book grew out of the lecture notes for a graduate course we taught during the Summer Semester 2018 at the Max-Planck Institute (MPI) for Mathematics in the Sciences in Leipzig, Germany. This was part of the Ringvorlesung series which is offered biannually by the International Max-Planck Research School (IMPRS). The aim of our course was to introduce the theme of Nonlinear Algebra, which is also the name of the research group that started at MPI Leipzig in early 2017.

Linear algebra is the foundation of much of mathematics, particularly in applied mathematics. Numerical linear algebra is the basis of most of scientific computing, and its importance for the sciences and engineering can hardly be overestimated.

This ubiquity of linear algebra masks the fairly recent growth of nonlinear models across the mathematical sciences. There has been a proliferation of methods that are based on systems of multivariate polynomial equations and inequalities. This is fueled by recent theoretical advances, efficient software, and an increased awareness of these tools. At the heart of this lies algebraic geometry, but there are links to numerous other branches, such as combinatorics, algebraic topology, commutative algebra, convex and discrete geometry, multilinear and tensor algebra, number theory, representation theory, and symbolic and numerical computation. Application areas include optimization, statistics, complexity theory, and more.

Nonlinear algebra is not simply a rebranding of algebraic geometry. It is a recognition that a focus on computation and applications, and the theoretical needs that this requires, results in a body of inquiry that is complementary to the existing curriculum. The term nonlinear algebra is intended to capture these trends, and to be more friendly to applied scientists. A special research semester with that title, held in the fall of 2018 at the Institute for Computational and Experimental Research in Mathematics (ICERM) at Brown University, explored the theoretical and computational challenges that have arisen, and it charted the course for the future. This book supports this effort by offering a warm welcome to nonlinear algebra.

Our presentation is structured into 13 chapters, one per week of the semester. Many of our chapters are rather ambitious in that they promise a first introduction to an area of mathematics that would normally be covered in a full-year course. But, what we offer is really just an invitation. Our readers are strongly encouraged to go further in their studies. We hope that students will find our presentation of use and that nonlinear algebra will encourage them to think critically and deeply, and to question the historic boundaries between “pure” and “applied” mathematics.

Mateusz Michalek and Bernd Sturmfels
CHAPTER 1

Polynomial Rings

“Algebra is but written geometry”, Sophie Germain

A natural next step after linear algebra is commutative algebra. In that subject area one studies fields, rings and ideals. In this chapter we introduce the relevant basics, with a focus on polynomials and Gröbner basis. We show how to use these for computing basic invariants of a polynomial ideal, like dimension or degree. The formalism we develop now will be applied to geometric situations in later chapters.

1.1. Ideals

Our basic algebraic structure is a field. The elements of the field serve as numbers, also called scalars. We add, subtract, multiply and divide them. It is customary to denote fields by the letter $K$, for the German word Körper. Our favorite field is the set $K = \mathbb{Q}$ of rational numbers. Another important field is the set $K = \mathbb{R}$ of real numbers. In practise, these two fields are very different. Numbers in $\mathbb{Q}$ can be manipulated by exact symbolic computation, whereas numbers in $\mathbb{R}$ are approximated by floating point representations and manipulated by numerical computation.

Other widely used fields are the complex numbers $\mathbb{C}$, the finite field $\mathbb{F}_q$ with $q$ elements, and the field $\mathbb{Q}_p$ of $p$-adic numbers. If $K$ is not algebraically closed then we write $\overline{K}$ for its algebraic closure. This is the smallest field in which every non-constant polynomial with coefficients in $K$ has a root. For instance, $\overline{\mathbb{Q}}, \overline{\mathbb{F}_q}$ and $\overline{\mathbb{Q}_p}$ are the algebraic closures of fields listed above. Another important example is the field of rational functions $\mathbb{Q}(t)$. Its algebraic closure $\overline{\mathbb{Q}(t)}$ is strictly contained in the field of Puiseux series $\mathbb{C}\{\{t\}\} = \mathbb{C}\{\langle t \rangle\}$, which is also algebraically closed.

In this section we study the ring of polynomials in $n$ variables $x_1, x_2, \ldots, x_n$ with coefficients in our field $K$. It is denoted $K[x] = K[x_1, x_2, \ldots, x_n]$. If the number $n$ is small then we typically use letters without indices to denote the variables. For instance, we write $K[x], K[x, y], \text{or } K[x, y, z]$ for the polynomial ring when $n \leq 3$.

Many of the constructions we present work for an arbitrary commutative ring $R$ with unit 1. For the most part, the reader may assume $R = K[x]$. But, of course, it would not hurt to peruse a standard text book on abstract algebra and look up the axioms of a ring and the formal definitions of commutative and unit. Important examples of rings are the integers $\mathbb{Z}$, the polynomial ring of the integers $\mathbb{Z}[x]$, or the quotient of a polynomial ring by an ideal. The latter will be discussed soon.

The polynomial ring $K[x]$ is an infinite-dimensional $K$-vector space. A distinguished basis of this vector space consists of the monomials $x^a = x_1^{a_1}x_2^{a_2}\cdots x_n^{a_n}$. There is one monomial for each nonnegative integer vector $a = (a_1, a_2, \ldots, a_n) \in \mathbb{N}^n$. It would not hurt to peruse a standard text book on abstract algebra and look up the axioms of a ring and the formal definitions of commutative and unit. Important examples of rings are the integers $\mathbb{Z}$, the polynomial ring of the integers $\mathbb{Z}[x]$, or the quotient of a polynomial ring by an ideal. The latter will be discussed soon.
Every polynomial \( f \in K[x] \) is written uniquely as a finite sum of monomials
\[
f = \sum_{a} c_{a}x^{a}.
\]
The degree of \( f \) is the maximum of the quantities \( |a| = a_1 + a_2 + \cdots + a_n \), where \( c_{a} \neq 0 \). Polynomials of degree 1, 2, 3, 4 are called linear, quadratic, cubic, quartic.

For example, the following is a cubic polynomial in \( n = 3 \) variables:
\[
(1.1) \quad f = \det \begin{pmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{pmatrix} = 2xyz - x^2 - y^2 - z^2 + 1.
\]
The zero set of this polynomial \( f \) is the surface in \( \mathbb{R}^3 \) that is shown in Figure 1. It consists of all points at which the rank of the 3 \( \times \) 3-matrix in (1.1) drops below 3.

We note that our cubic determinantal surface has four singular points, namely the points \((1,1,1), (1,-1,-1), (-1,1,-1), \) and \((-1,-1,1)\). These four points are the common zeros in \( \mathbb{R}^3 \) of the cubic \( f \) and its three partial derivatives of \( f \):
\[
\frac{\partial f}{\partial x} = 2yz - 2x, \quad \frac{\partial f}{\partial y} = 2xz - 2y, \quad \frac{\partial f}{\partial z} = 2xy - 2z.
\]
These are the points at which the rank of the 3 \( \times \) 3-matrix in (1.1) drops to 1.

**Figure 1.** A cubic surface with four singular points.

For example, the following is a cubic polynomial in \( n = 3 \) variables:

(1.1) \quad f = \det \begin{pmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{pmatrix} = 2xyz - x^2 - y^2 - z^2 + 1.

The zero set of this polynomial \( f \) is the surface in \( \mathbb{R}^3 \) that is shown in Figure 1. It consists of all points at which the rank of the 3 \( \times \) 3-matrix in (1.1) drops below 3.

We note that our cubic determinantal surface has four singular points, namely the points \((1,1,1), (1,-1,-1), (-1,1,-1), \) and \((-1,-1,1)\). These four points are the common zeros in \( \mathbb{R}^3 \) of the cubic \( f \) and its three partial derivatives of \( f \):
\[
\frac{\partial f}{\partial x} = 2yz - 2x, \quad \frac{\partial f}{\partial y} = 2xz - 2y, \quad \frac{\partial f}{\partial z} = 2xy - 2z.
\]
These are the points at which the rank of the 3 \( \times \) 3-matrix in (1.1) drops to 1.

**Definition 1.1.** An ideal in a ring \( R \) is a subset \( I \) of \( R \) such that
(a) if \( f \in R \) and \( g \in I \) then \( fg \in I \);
(b) if \( g, h \in I \) then \( g + h \in I \).

If \( R = K[x] \) then an ideal \( I \) is a subset of \( K[x] \) that is closed under taking linear combinations with polynomial coefficients. An alternative definition is as follows: A subset \( I \) of a ring \( R \) is an ideal if and only if there exists a ring homomorphism \( \phi : R \to S \) whose kernel \( \ker \phi = \phi^{-1}(0) \) is equal to \( I \). For instance, if \( R = \mathbb{Z} \) then the set \( I \) of even integers is an ideal. It is the kernel of the ring homomorphism \( \mathbb{Z} \to \mathbb{Z}/2\mathbb{Z} = \{0,1\} \) that takes an integer to either 0 or 1, depending on its parity.

Ideals in a ring play the same role as normal subgroups in a group. They are the subobjects that are used to define quotients. Consider the quotient of abelian
groups $R/I$. Its elements are the congruence classes $f + I$ modulo $I$. The ideal axioms (a) and (b) in Definition 1.1 ensure that the following identities hold:

$$(f + I) + (g + I) = (f + g) + I \quad \text{and} \quad (f + I)(g + I) = fg + I.$$ (1.2)

**Corollary 1.2.** If $I \subset R$ is an ideal then the quotient $R/I$ is a ring.

Given any subset $\mathcal{F}$ of $R$, we write $\langle \mathcal{F} \rangle$ for the smallest ideal containing $\mathcal{F}$. This is the ideal generated by $\mathcal{F}$. If $R = K[x]$ then the ideal $\langle F \rangle$ is the set of all polynomial linear combinations of finite subsets of $F$.

**Proposition 1.3.** If $I$ and $J$ are ideals in a ring $R$ then the following subsets of $R$ are ideals as well: the sum $I + J$, the intersection $I \cap J$, the product $IJ$, and the quotient $(I : J)$. The latter two subsets of $R$ are defined as follows:

$$IJ = \{fg : f \in I, g \in J\} \quad \text{and} \quad (I : J) = \{f \in R : fJ \subseteq I\}.$$

**Proof.** The product $IJ$ is an ideal by definition. For the others one checks that conditions (a) and (b) hold. We shall carry this out for the ideal quotient $(I : J)$. To show (a), suppose that $f \in R$ and $g \in (I : J)$. We have:

$$(fg)J = f(gJ) \subseteq fI \subseteq I.$$

For (b), suppose $f, g \in (I : J)$. We have:

$$(f + g)J \subseteq fJ + gJ \subseteq I + I = I.$$ 

This implies $g + h \in (I : J)$. We have shown that $(I : J)$ is an ideal. □

The **Euclidean algorithm** works in the polynomial ring $K[x]$ in one variable $x$ over a field $K$. This implies that $K[x]$ is a principal ideal domain (PID), i.e. every ideal $I$ is generated by one element. That generator can be uniquely factored into irreducible polynomials. The polynomial ring $K[x]$ in $n$ variables is a unique factorization domain (UFD). However, $K[x]$ is not a PID when $n \geq 2$.

**Example 1.4 (n = 1).** Consider the following two ideals in $\mathbb{Q}[x]$:

$$I = \langle x^3 + 6x^2 + 12x + 8 \rangle \quad \text{and} \quad J = \langle x^2 + x - 2 \rangle.$$

We wish to compute the four ideals in Proposition 1.3. For this, it helps to factor:

$$I = \langle (x + 2)^3 \rangle \quad \text{and} \quad J = \langle (x - 1)(x + 2) \rangle.$$

The four new ideals are

$$I \cap J = \langle (x - 1)(x + 2)^3 \rangle \quad IJ = \langle (x - 1)(x + 2)^4 \rangle$$

$$I + J = \langle x + 2 \rangle \quad I : J = \langle (x + 2)^2 \rangle.$$

We conclude that arithmetic in $\mathbb{Q}[x]$ is just like arithmetic in the ring of integers $\mathbb{Z}$.

A non-zero element $f$ in a ring $R$ is called

- a nilpotent if $f^m = 0$ for some positive integer $m$,
- a zero divisor if there exists $0 \neq g \in R$ such that $gf = 0$.

A ring $R$ is called an integral domain if it has no zero divisors.
We examine these properties for the quotient ring \( R/I \) where \( I \) is an ideal in \( R \). Properties of the ideal \( I \) correspond to properties of the ring \( R/I \), as follows:

<table>
<thead>
<tr>
<th>property</th>
<th>definition</th>
<th>the quotient ring ( R/I )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( I ) is maximal</td>
<td>( I ) is not contained in any proper ideal</td>
<td>is a field</td>
</tr>
<tr>
<td>( I ) is prime</td>
<td>( fg \in I \Rightarrow f \in I ) or ( g \in I )</td>
<td>is an integral domain</td>
</tr>
<tr>
<td>( I ) is radical</td>
<td>( (\exists s : f^s \in I) \Rightarrow f \in I )</td>
<td>has no nilpotent elements</td>
</tr>
<tr>
<td>( I ) is primary</td>
<td>( fg \in I ) and ( g \notin I \Rightarrow (\exists s : f^s \in I) )</td>
<td>every zero divisor is nilpotent</td>
</tr>
</tbody>
</table>

Example 1.5. The ideal \( I = \langle x^2 + 10x + 34, 3y - 2x - 13 \rangle \) is maximal in the polynomial ring \( R[x, y] \). The field \( R[x, y]/I \) is isomorphic to the complex numbers \( \mathbb{C} = \mathbb{R}[i]/(i^2 + 1) \). One isomorphism is gotten by sending \( i = \sqrt{-1} \) to \( \frac{1}{13}(x + 5y) \).

The square of that expression is \(-1 \) mod \( I \). The principal ideal \( J = \langle x^2 + 10x + 34 \rangle \) is prime. The quotient \( R[x, y]/J \) is an integral domain. It is isomorphic to \( \mathbb{C}[y] \).

Examples for the other two classes of ideals are given in the next proof.

Proposition 1.6. We have the following implications:

- \( I \) maximal \( \Rightarrow \) \( I \) prime \( \Rightarrow \) \( I \) radical,
- \( I \) prime \( \Rightarrow \) \( I \) primary.

None of these implications is reversible, however every radical primary ideal is prime. Every intersection of prime ideals is radical.

Proof. The first implication holds because there are no zero divisors in a field.

To see that prime implies radical, we take \( g = f^{s-1} \) and we use induction on \( s \). Prime implies primary is clear. To prove that every radical primary ideal is prime assume \( fg \in I \) and \( f \notin I \). Then, as \( I \) is primary, we have \( g^s \in I \) for some \( s \in \mathbb{N} \).

As \( I \) is radical, we now conclude that \( g \notin I \).

To see that no implication is reversible, we consider the following three ideals in the polynomial ring \( R[x, y] \) with \( n = 2 \):

- \( I = \langle x^2 \rangle \) is primary but not radical,
- \( I = \langle x(x - 1) \rangle \) is radical but not primary,
- \( I = \langle x \rangle \) is prime but not maximal.

The last statement holds because every intersection of radical ideals is radical. \( \square \)

We now revisit the singular surface in Figure [1] from the perspective of ideals.

Example 1.7 (\( n = 3 \)). We consider the ideal generated by the partial derivatives of the cubic \( f = 2xyz - x^2 - y^2 - z^2 + 1 \). This is the ideal

\[
I = \langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \rangle = \langle xy - z, xz - y, yz - x \rangle \subset \mathbb{R}[x, y, z].
\]

The given cubic \( f \) is not in this ideal because every polynomial in \( I \) has zero constant term. The ideal \( I \) is radical because we can write it as the intersection of five maximal ideals. Using a computer algebra system, we find that \( I \) equals

\[
\langle x, y, z \rangle \cap \langle x-1, y-1, z-1 \rangle \cap \langle x-1, y+1, z+1 \rangle \cap \langle x+1, y-1, z+1 \rangle \cap \langle x+1, y+1, z-1 \rangle.
\]

The cubic \( f \) lies in the last four maximal ideals. Their intersection is equal to \( I + \langle f \rangle \). The zero set of the radical ideal \( I + \langle f \rangle \) consists of the four singular points on the surface. The Chinese Remainder Theorem implies that the quotient ring is a product of fields. Namely, we have an isomorphism \( \mathbb{R}[x, y]/I \simeq \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \).
1.2. Gröbner Bases

Every ideal has many different generating sets, and there is no canonical notion of basis. For instance, the set \( \mathcal{F} = \{x^6 - 1, x^{10} - 1, x^{15} - 1\} \) minimally generates the ideal \((x - 1)\) in the polynomial ring \(\mathbb{Q}[x]\) in one variable. Of course, the singleton \(\{x - 1\}\) is a preferable generating set, given that every ideal in \(\mathbb{Q}[x]\) is principal for \(n = 1\). The Euclidean algorithm transforms the set \(\mathcal{F}\) into the set \(\{x - 1\}\).

A certificate for the fact that \(x - 1\) is in the ideal generated by \(\mathcal{F}\) is the identity

\[
x^5 \cdot (x^6 - 1) - (x^5 + x) \cdot (x^{10} - 1) + 1 \cdot (x^{15} - 1) = x - 1.
\]

Such certificates can be found with the Extended Euclidean Algorithm. Please google this. Finding certificates for ideal membership when \(n \geq 2\) is a harder problem. This topic comes up when we discuss Hilbert’s Nullstellensatz in Chapter 6.

Gaussian elimination is familiar from linear algebra. We here regard it as a process for manipulating ideals that are generated by linear polynomials. For example, the following two polynomial ideals in three variables are identical in \(\mathbb{Q}[x, y, z]\):

\[
\langle 2x + 3y + 5z + 7, 11x + 13y + 17z + 19, 23x + 29y + 31z + 37 \rangle = \langle 7x - 16, 7y + 12, 7z + 9 \rangle.
\]

Undergraduate linear algebra taught us how to transform the three generators on the left into the simpler ones on the right. This is the process of solving a system of linear equations. In our example there is a unique solution, namely \((\frac{16}{7}, -\frac{12}{7}, -\frac{9}{7})\).

We next introduce Gröbner bases. The framework of Gröbner bases offers practical methods for computing with ideals in a polynomial ring \(K[x]\) in \(n\) variables. Implementations of Gröbner bases are available in all major computer algebra systems, and we strongly encourage our readers to experiment with these tools.

Informally, we can think of Gröbner bases as a version of the Euclidean algorithm for polynomials in \(n \geq 2\) variables, or as a version of Gaussian elimination for polynomials of degree \(\geq 2\). Gröbner bases for ideals in \(K[x]\) are fundamental in non-linear algebra, just like Gaussian elimination for matrices is fundamental when one studies linear algebra. Nonlinear algebra is the next step after linear algebra.

We identify the set \(\mathbb{N}^n\) of non-negative integer vectors with the monomial basis of the polynomial ring \(K[x]\). The coordinatewise order on \(\mathbb{N}^n\) corresponds to divisibility of monomials, i.e. we have \(a \leq b\) in \(\mathbb{N}^n\) if and only if \(x^a\) divides \(x^b\).

**Theorem 1.8 (Dickson’s Lemma).** Any infinite subset of \(\mathbb{N}^n\) contains a pair \(\{a, b\}\) that satisfies \(a \leq b\).

**Proof.** We proceed by induction on \(n\). The statement is trivial for \(n = 1\). Any subset of cardinality at least two in \(\mathbb{N}\) contains a comparable pair. Suppose now that Dickson’s Lemma has been proved for \(n - 1\), and consider an infinite subset \(\mathcal{M}\) of \(\mathbb{N}^n\). For each \(i \in \mathbb{N}\) let \(\mathcal{M}_i\) denote the set of all vectors \(a \in \mathbb{N}^{n-1}\) such that \((a, i) \in \mathcal{M}\). If some \(\mathcal{M}_i\) is infinite then we are done by induction. Hence each \(\mathcal{M}_i\) is finite, and we have \(\mathcal{M}_i \neq \emptyset\) for infinitely many \(i\).

The infinite subset \(\bigcup_{i=0}^{\infty} \mathcal{M}_i\) of \(\mathbb{N}^{n-1}\) satisfies the assertion. This means that its subset of minimal elements with respect to the coordinatewise order is finite. Hence there exists an index \(j\) such that all minimal elements are contained in the finite set \(\bigcup_{i=0}^{j} \mathcal{M}_i\). Pick any element \((b, k) \in \mathcal{M}_k\) for \(k > j\). Since \(b\) is not minimal in \(\bigcup_{i=0}^{\infty} \mathcal{M}_i\), there exists an index \(i\) with \(i \leq j < k\) and an element \(a \in \mathcal{M}_i\) with \(a \leq b\). Then we have \((a, i) \leq (b, k)\) in \(\mathcal{M}\). \(\square\)
Corollary 1.9. For any nonempty set $\mathcal{M} \subset \mathbb{N}^n$, its subset of coordinatewise minimal elements is finite and nonempty.

Proof. The fact that it is nonempty follows by induction on $n$. The set is finite by Dickson’s Lemma. \qed

Definition 1.10. A total ordering $\prec$ of the set $\mathbb{N}^n$ is a monomial order if, for all $a, b, c \in \mathbb{N}^n$ we have

- $(0,0,\ldots,0) \preceq a$;
- $a \preceq b$ implies $a + c \preceq b + c$.

This gives a total order on all monomials in $K[x]$. Three standard examples are:

- the lexicographic ordering: we set $a \prec_{\text{lex}} b$ if the leftmost non-zero entry of $b - a$ is positive.
- the degree lexicographic order: we set $a \prec_{\text{deglex}} b$ if either $|a| < |b|$, or $|a| = |b|$ and the leftmost non-zero entry of $b - a$ is positive.
- the degree reverse lexicographic order: we set $a \prec_{\text{revlex}} b$ if either $|a| < |b|$, or $|a| = |b|$ and the rightmost non-zero entry of $b - a$ is negative.

All three orders satisfy $x_1 \succ x_2 \succ \cdots \succ x_n$, but they differ on monomials of higher degree. It is instructive to list the 10 quadratic monomials for $n = 4$ in each order.

Throughout this book we specify a monomial order by giving the name of the order and how the variables are sorted. For instance, we might say: “let $\prec$ denote the degree lexicographic order on $K[x,y,z]$ given by $y \prec z \prec x$”. Further choices of monomial orderings can be obtained by assigning positive weights to the variables. See [7] Exercise 11 in §2.4. We also note that any monomial order is a refinement of the partial, coordinatewise order on $\mathbb{N}^n$: if $x^a$ divides $x^b$ then $a \preceq b$.

Remark 1.11. Fix a monomial order $\prec$ and let $\mathcal{M}$ be any nonempty subset of $\mathbb{N}^n$. Then $\mathcal{M}$ has a unique minimal element with respect to $\prec$. To see this, apply Dickson’s Lemma as in Corollary 1.9. Our set $\mathcal{M}$ has a finite, nonempty subset of minimal elements with respect to the component-wise partial order on $\mathbb{N}^n$. This finite subset is linearly ordered by $\prec$, and we select its minimal element.

We now fix a monomial order $\prec$. Given any nonzero polynomial $f \in K[x]$, its initial monomial $\text{in}_\prec(f)$ is the $\prec$-largest monomial $x^a$ among those that appear in $f$ with non-zero coefficient.

For a comparison among our orders, let $n = 3$ with variable order $x \succ y \succ z$.

If $f = x^2 + xz^2 + y^3$ then $\text{in}_{\text{lex}}(f) = x^2$, $\text{in}_{\text{deglex}}(f) = xz^2$ and $\text{in}_{\text{revlex}}(f) = y^3$.

For any ideal $I \subset K[x]$, we define the initial ideal of $I$ with respect to $\prec$ as follows:

$$\text{in}_\prec(I) = \langle \text{in}_\prec(f) : f \in I \rangle.$$  

This is a monomial ideal, i.e. it is generated by a set of monomials. A priori, this generating set is infinite. However, it turns out that we can always choose a finite subset that suffices to generate this monomial ideal.

Proposition 1.12. Fix a monomial order $\prec$. Every ideal $I$ in the polynomial ring $K[x]$ has a finite subset $\mathcal{G}$ such that

$$\text{in}_\prec(I) = \langle \text{in}_\prec(f) : f \in \mathcal{G} \rangle.$$  

Such a finite subset $\mathcal{G}$ of $I$ is called a Gröbner basis for $I$ with respect to $\prec$. 


1.2. GRÖBNER BASES

Proof. Suppose no such finite set $G$ exists. Then we can create a list of infinitely many polynomials $f_1, f_2, f_3, \ldots$ in $I$ such that none of the initial monomials in $\prec(f_i)$ divides any other initial monomial in $\prec(f_j)$. This would be a contradiction to Dickson’s Lemma.

We next show that every Gröbner basis actually generates its ideal.

Theorem 1.13. If $G$ is a Gröbner basis for an ideal $I$ in $K[x]$ then $I = \langle G \rangle$.

Proof. Suppose that $G$ does not generate $I$. Among all polynomials $f$ in the set difference $I \setminus \langle G \rangle$, there exists an $f$ whose initial monomial $x^b = \text{in}_\prec(f)$ is minimal with respect to $\prec$. This follows from Remark 1.11. Since $x^b \in \text{in}_\prec(I)$, there exists an element $g \in G$ whose initial monomial divides $x^b$, say $x^b = x^c \cdot \text{in}_\prec(g)$. Now, $f - x^c g$ is a polynomial with strictly smaller leading monomial. It lies in $I$ but it does not lie in the ideal $\langle G \rangle$. This is a contradiction to the choice of $f$.

Corollary 1.14 (Hilbert’s Basis Theorem). Every ideal $I$ in $K[x]$ is finitely generated.

Proof. Fix any monomial order $\prec$. By Proposition 1.12, the ideal $I$ has a finite Gröbner basis $G$. By Theorem 1.13, the Gröbner basis $G$ generates $I$.

Gröbner bases are not unique. If $G$ is a Gröbner basis of an ideal $I$ with respect to a monomial order $\prec$ then so is every other finite subset of $I$ that contains $G$. In that sense, Gröbner bases differ from the bases we know from linear algebra. The issue of minimality and uniqueness is addressed next.

Definition 1.15. Fix $I$ and $\prec$. A Gröbner basis $G$ is reduced if the following two conditions hold:

(a) The leading coefficient of each polynomial $g \in G$ is 1.

(b) For any two distinct $g, h \in G$, no monomial in $g$ is a multiple of $\text{in}_\prec(h)$.

In what follows we fix an ideal $I \subset K[x]$ and a monomial ordering $\prec$.

Theorem 1.16. The ideal $I$ has a unique reduced Gröbner basis for $\prec$.

Proof idea. We refer to [7, §2.7, Theorem 5]. The idea is as follows. We start with any Gröbner basis $G$ and we turn it into a reduced Gröbner basis by applying the following steps. First we divide each $g \in G$ by its leading coefficient to make it monic, so that (a) holds. We remove all elements $g$ from $G$ whose initial monomial is not a minimal generator of $\text{in}_\prec(I)$. For any pair of polynomials with the same initial monomial we delete one of them. Next we apply the division algorithm [7, §2.3] to any trailing monomial until no more trailing monomial is divisible by any leading monomial. The resulting set is the reduced Gröbner basis.

Let $S_\prec(I)$ be the set of all monomials $x^b$ that are not in the initial ideal in $\prec(I)$. We call these $x^b$ the standard monomials of $I$ with respect to $\prec$.

Theorem 1.17. The set $S_\prec(I)$ of standard monomials is a basis for the $K$-vector space $K[x]/I$.

Proof. The image of $S_\prec(I)$ in $K[x]/I$ is linearly independent because every non-zero polynomial $f$ has at least one monomial, namely $\text{in}_\prec(f)$, that is not in $S$. We next prove that $S_\prec(I)$ spans $K[x]/I$. Suppose not. Then there exists a monomial $x^c$ which is not in the $K$-span of $S_\prec(I)$ modulo $I$. We may assume that
x^c is minimal with respect to the monomial order ≺. Since x^c is not in S≺(I), it lies in the initial ideal in≺(I). Hence there exists h ∈ I with in≺(h) = x^c. Each monomial in h other than x^c is smaller with respect to ≺, so it lies in the K-span of S≺(I) modulo I. Hence x^c has the same property. This is a contradiction. □

Software for Gröbner bases rests on Buchberger’s Algorithm [7] S 2.7]. This is implemented in all major computer algebra systems. It takes as its input a monomial order ≻ and a finite set F of polynomials in K[x]. The output of Buchberger’s Algorithm is the unique reduced Gröbner basis G for the ideal I = ⟨F⟩ with respect to ≻. In what follows we present some examples of input-output pairs (F, G) for n = 3. In each case we take the lexicographic monomial order with x ≻ y ≻ z.

Example 1.18. The input F ⊂ Q[x, y, z] is transformed into the reduced Gröbner basis G. The initial monomials are underlined in each case.

- For n = 1, the reduced Gröbner basis is just the greatest common divisor: F = \{x^3 - 6x^2 - 5x - 14, 3x^4 + 8x^2 + 11x + 10, 4x^4 + 4x^3 + 7x^2 - x - 2\}, G = \{(x^2 + x + 2)\}.

- For linear polynomials, running Buchberger’s algorithm amounts to Gaussian elimination: For F = \{2x + 3y + 5z + 7, 11x + 13y + 17z + 19, 23x + 29y + 31z + 37\}, the reduced Gröbner basis G = \{(x - \frac{16}{7}, y + \frac{12}{7}, z + \frac{2}{7})\}.

- Here is another ideal we saw earlier: F = \{xy - z, xz - y, yz - x\} yields G = \{(x - yz, y^2 - z^2, yz^2 - y, z^3 - z)\}. There are precisely five standard monomials: S≺(I) = \{1, y, z, yz, z^2\}. This is consistent with Example 1.7 where we saw that F has precisely five zeros in C^3.

- This input is a curve in the (y, z)-plane parametrized by two cubics in one variable x: F = \{y - x^3 + 4x, z - x^3 - x + 1\}. The Gröbner basis has the implicit equation of this curve as its second element: G = \{(\frac{1}{2}x^5 + \frac{1}{5}y + \frac{1}{5}, y^3 - 3y^2 z - 3yz^2 + 6yz + 28y - z^3 - 3z^2 + 97z + 99\}.

- Let z be the sum of x = √7 and y = √5. We encode this in the set F = \{x^3 - 7, y^4 - 5, z - x - y\}. The number z = √7 + √5 is algebraic of degree 12 over Q. Its minimal polynomial appears in G = \{(z^{12} - 28z^9 - 15z^6 + 294z^5 - 1680z^3 + 75z^4 - 1372z^3 - 7350z^2 - 2100z + 2276, \ldots\}.

- The elementary symmetric polynomials F = \{x + y + z, xy + xz + yz, xyz\} have the reduced Gröbner basis G = \{(x + y + z, y^2 + yz + z^2, z^3\}.

For each of the six ideals above, what is the reduced Gröbner basis for the degree lexicographic order? What are the possible initial monomial ideals?

In general, the choice of monomial order can make a huge difference in the complexity of the size of the reduced Gröbner basis, even for two input polynomials.

Example 1.19 (Intersecting two quartic surfaces in P^3). A random homogeneous polynomial of degree four in n = 4 variables has 35 monomials. Consider the ideal I generated by two such polynomials. If ≺ is the degree reverse lexicographic order then the reduced Gröbner basis G contains 5 elements of degree up to 7. If ≺ is the lexicographic order then G contains 150 elements of degree up to 73.
Naturally, one uses a computer to find the 150 elements in the aforementioned Gröbner basis. Many computer algebra systems offer an implementation of Buchberger’s algorithm for Gröbner bases. We reiterate that our readers are strongly encouraged to experiment with a computer algebra system while studying this book.

1.3. Dimension and Degree

The two most important invariants of an ideal $I$ in a polynomial ring $K[x]$ are its dimension and its degree. We shall define these invariants, starting with the case of monomial ideals. In this section we focus on combinatorial aspects. The geometric interpretation of dimension and degree will be discussed in Chapter 2.

Definition 1.20 (Hilbert function). Let $I \subset K[x]$ be a monomial ideal. The associated Hilbert function $h_I$ takes nonnegative integers to nonnegative integers. The value $h_I(q)$ is the number of monomials of degree $q$ that do not belong to $I$.

A convenient way to represent a function $\mathbb{N} \to \mathbb{N}$ is by its generating function, which is a formal power series with nonnegative integer coefficients. The generating function for the Hilbert function is known as the Hilbert series.

Definition 1.21 (Hilbert series). Let $I \subset K[x]$ be a monomial ideal. We introduce a formal variable $z$. The Hilbert series of $I$ is the generating function $HS_I(z) = \sum_{q=0}^{\infty} h_I(q) z^q$.

We begin with zero ideal $I = \{0\}$. Here we count all monomials in $K[x]$.

Example 1.22. The Hilbert series of the zero ideal is the rational function $HS_{\{0\}}(q) = \frac{1}{(1-z)^n} = \sum_{q=0}^{\infty} \binom{n+q-1}{n-1} z^q$. The number of monomials of degree $q$ in $n$ variables equals $h_{\{0\}}(q) = \binom{n+q-1}{n-1}$. Note that the Hilbert function $h_{\{0\}}(q)$ is a polynomial of degree $n-1$ in the unknown $q$.

We next consider the case of a principal ideal.

Example 1.23. Let $I = \langle x_1^{a_1} \cdots x_n^{a_n} \rangle$, where $\sum_{i=1}^{n} a_i = e$. We need to count all monomials of degree $q$ that are not divisible by the generator of $I$. Equivalently, we can count all monomials and then subtract those that are in $I$. This leads to

$$HS_I(z) = \frac{1 - z^e}{(1-z)^n} = \sum_{q=0}^{\infty} \left( \binom{n+q-1}{n-1} - \binom{n+q-e-1}{n-1} \right) z^q.$$ 

The second binomial coefficient is zero when $q < e$. For all $q \geq e$, the Hilbert function $h_I(q) = \binom{n+q-1}{n-1} - \binom{n+q-e-1}{n-1}$ is a polynomial in $q$ of degree $n-2$.

Our third example concerns monomial ideals with two minimal generators.

Example 1.24. Consider an ideal $I = \langle m_1, m_2 \rangle$ in $K[x]$, where $m_i$ is a monomial of degree $e_i$ for $i = 1, 2$. We can count the monomials in $I$ of degree $q$ by

1. computing the number of monomials divisible by $m_1$,
2. adding the number of monomials divisible by $m_2$,
3. subtracting the number of monomials divisible both by $m_1$ and $m_2$. 
This expression agrees with a polynomial in $q$. Case (3) concerns monomials that are divisible by the least common multiple $m_{12} = \text{lcm}(m_1, m_2)$. Let $e_{12}$ denote the degree of $m_{12}$. Then the Hilbert series equals

$$\text{HS}_I(z) = \frac{1 - z^{e_{12}}}{(1 - z)^n}.$$ 

This means that the Hilbert function is an alternating sum of binomial coefficients:

$$h_I(q) = \binom{n + q - 1}{n-1} - \binom{n + q - e_{12} - 1}{n-1} - \binom{n + q - e_{12} - 1}{n-1} + \binom{n + q - e_{12} - 1}{n-1}. $$

This expression agrees with a polynomial in $q$, provided $q \geq e_{12}$.

**Theorem 1.25.** The Hilbert series of a monomial ideal $I \subset K[x]$ has the form

$$\text{HS}_I(z) = \frac{\kappa_I(z)}{(1-z)^n},$$

where $\kappa_I(z)$ is polynomial with integer coefficients and $\kappa_I(0) = 1$. There exists a polynomial $\text{HP}$ in one unknown $q$ of degree $\leq n-1$, known as the Hilbert polynomial of the ideal $I$, such that $\text{HP}(q) = h_I(q)$ for all values of $q$ that are sufficiently large.

**Proof.** We prove this result by counting monomials using inclusion-exclusion, as hinted at in the three examples above. Let $m_1, m_2, \ldots, m_r$ be the monomials that minimally generate $I$. For any subset $\tau$ of the index set $\{1, 2, \ldots, r\}$, we write $m_\tau$ for the least common multiple of the set $\{m_i : i \in \tau\}$, and we set $e_\tau = \text{degree}(m_\tau)$. This includes the empty set $\tau = \emptyset$, for which $m_\emptyset = 1$ and $e_\emptyset = 0$. The desired numerator polynomial (1.3) can be written as an alternating sum of $2^r$ powers of $z$:

$$\kappa_I(z) = \sum_{\tau \subseteq \{1, 2, \ldots, r\}} (-1)^{|\tau|} \cdot z^{e_\tau}.$$ 

The cases $r = 0, 1, 2$ were seen above, and the general case is inclusion-exclusion. Note that $\kappa_I \in \mathbb{Z}[z]$ with $\kappa_I(0) = 1$. By regrouping the terms of the series (1.3),

$$h_I(q) = \sum_{\tau \subseteq \{1, 2, \ldots, r\}} (-1)^{|\tau|} \binom{n + q - e_\tau - 1}{n-1}. $$

This expression is a polynomial for $q \gg 0$. More precisely, the Hilbert function $h_I(q)$ coincides with the Hilbert polynomial $\text{HP}_I(q)$ for all $q$ that exceed $e_{\{1, 2, \ldots, r\}}$. This number is the degree of the least common multiple of all generators of $I$. □

**Remark 1.26.** The inclusion-exclusion principal carried out in the proof of Theorem 1.25 is a very powerful idea, but it also hints at possible simplifications. We wrote the numerator polynomial $\kappa_I(z)$ and the Hilbert polynomial $\text{HP}_I(q)$ as an alternating sum of $2^r$ terms. However, in most applications $r$ is much larger than $n$, and the vast majorities of terms will cancel each other. Doing the correct bookkeeping leads us the the topic of minimal free resolutions of monomial ideals. Identifying such resolutions is a main theme in combinatorial commutative algebra.

**Example 1.27.** Let $n = 2$ and consider the monomial ideal

$$I = \langle x \rangle \cap \langle y \rangle \cap \langle x, y \rangle^{r+1} = \langle x^r y, x^{r-1} y^2, x^{r-2} y^3, \ldots, x^2 y^{r-1}, xy^r \rangle.$$ 

Our formula for $\kappa_I$ involves $2^r$ terms. After cancellations, only $2r$ of them remain:

$$\kappa_I(z) = 1 - rz^{r+1} + (r-1)z^{r+2}.$$ 

The Hilbert polynomial is the constant $h_I(q) = 2$, and this is also the value of the Hilbert function $\text{HF}_I(q)$ for $q > r$. Note that we have $\text{HF}_I(q) = q + 1$ for $q \leq r$. 


1.3. Dimension and Degree

**Definition 1.28 (Dimension, Degree).** Let $I$ be a monomial ideal and write

$$\text{HP}_I(q) = \frac{g}{(d-1)!} q^{d-1} + \text{lower order terms in } q.$$ 

The **dimension** of $I$ is $d$ and the **degree** of $I$ is $g$. Here $g$ is a positive integer.

**Remark 1.29.** The fact that $g$ is a positive integer is a non-trivial piece of combinatorics. From the inclusion-exclusion formulas above, one can show that the numerator of the Hilbert series factors as $\kappa_I(z) = \lambda_I(z) \cdot (1 - z)^{n-d}$, where $\lambda_I(z)$ is also a polynomial with integer coefficients. The degree of $I$ equals $g = \lambda_I(1)$.

**Example 1.30.** Let $I$ be a principal ideal as in Example 1.23 generated by a monomial of degree $e > 0$. Then the dimension $I$ is $n - 1$ and the degree of $I$ is $e$.

**Example 1.31.** Let $n = 2m$ be even and consider the following monomial ideal.

$$I = \langle x_1 x_2, x_3 x_4, x_5 x_6, \ldots, x_{2m-3} x_{2m-2}, x_{2m-1} x_{2m} \rangle.$$ 

The dimension of $I$ equals $m$ and the degree of $I$ equals $2^m$. It is instructive to work out the Hilbert series and the Hilbert polynomial of $I$ for small values of $m$.

We now consider arbitrary ideals $I$ in $K[x]$. Let $\prec$ be any degree-compatible monomial order. This means that $|a| < |b|$ implies $a \prec b$ for all $a, b \in I$.

**Lemma 1.32.** The number of standard monomials of $I$ in a given degree $q$ is independent of the choice of monomial order $\prec$, provided $\prec$ is degree-compatible.

**Proof.** Let $K[x]_{\leq q}$ denote the vector space of polynomials of degree $\leq q$. We write $I_{\leq q} := I \cap K[x]_{\leq q}$ for the subspace of polynomials that lie in the ideal $I$. Also, consider the set of standard monomials of degree at most $q$:

$$S_{\prec}(I)_{\leq q} = S_{\prec}(I) \cap K[x]_{\leq q}$$

We claim that $S_{\prec}(I)_{\leq q}$ is a $K$-vector space basis for the quotient space $K[x]_{\leq q}/I_{\leq q}$. It is clearly linearly independent since no $K$-linear combination of $S_{\prec}(I)$ lies in $I$. But, given that $\prec$ is degree compatible, it also spans because taking the normal form of a polynomial modulo the Gröbner basis can only decrease the total degree. 

**Remark 1.33.** The function that associates to $q$ the dimension of the quotient space $\dim K[x]_{\leq q}/I_{\leq q}$ is known as the **affine Hilbert function**. We will see its relations to the Hilbert function, as in Definition 1.20 in Chapter 2, after introducing projective varieties and homogenization.

**Definition 1.34.** Given any ideal $I$ in a polynomial ring $K[x]$, we define its **Hilbert function** $h_I$ to be that of its initial ideal $\text{in}_\prec(I)$, where $\prec$ is any degree-compatible term order. For all $q \in \mathbb{N}$ we have

$$h_I(q) = h_{\text{in}_\prec(I)}(q) = \frac{|S_{\prec}(I)_{\leq q}| - |S_{\prec}(I)_{\leq q-1}|}{\dim(K[x]_{\leq q}/I_{\leq q}) - \dim(K[x]_{\leq q-1}/I_{\leq q-1})}.$$ 

This is the number of standard monomials whose degree is exactly $q$. This number is independent of $\prec$, thanks to Lemma 1.32. We also define the **Hilbert series** and the **Hilbert polynomial** to be that of any degree-compatible initial monomial ideal:

$$\text{HS}_I(z) = \text{HS}_{\text{in}_\prec(I)}(z) \quad \text{and} \quad \text{HP}_I(q) = \text{HP}_{\text{in}_\prec(I)}(q).$$

Finally, we define the dimension of $I$ as the dimension of $\text{in}_\prec(I)$, and similarly for the degree of $I$. All of these concepts are now well-defined, thanks to Lemma 1.32.
Let $X$ be a principal ideal generated by a polynomial $f$ of degree $e$ in $n$ variables. Then the dimension of $I$ is $n - 1$ and the degree of $I$ is $e$. This follows from Example 1.30 because the singleton $\{f\}$ is a Gröbner basis, and its initial monomial $\text{in}_<(f)$ has degree $e$ in any degree-compatible monomial order $\prec$.

What we have accomplished in this section is to give a purely combinatorial definition of dimension and degree of an ideal $I$. In Chapter 2, we shall see that that notion of dimension agrees with the intuitive one for the associated algebraic variety $V(I)$. Namely, a variety has dimension 0 if and only it consists of finitely many points. The number of these points is counted by the degree of the corresponding radical ideal. Likewise, the ideal of a curve has dimension 1, the ideal of a surface has dimension 2 etc. The degree is a measure for how curvy these shapes are. A prime ideal has degree 1 if and only if it is generated by linear polynomials.

Example 1.35. Let $I$ be a principal ideal generated by a polynomial $f$ of degree $e$ in $n$ variables. Then the dimension of $I$ is $n - 1$ and the degree of $I$ is $e$. This follows from Example 1.30 because the singleton $\{f\}$ is a Gröbner basis, and its initial monomial $\text{in}_<(f)$ has degree $e$ in any degree-compatible monomial order $\prec$.

What we have accomplished in this section is to give a purely combinatorial definition of dimension and degree of an ideal $I$. In Chapter 2, we shall see that that notion of dimension agrees with the intuitive one for the associated algebraic variety $V(I)$. Namely, a variety has dimension 0 if and only it consists of finitely many points. The number of these points is counted by the degree of the corresponding radical ideal. Likewise, the ideal of a curve has dimension 1, the ideal of a surface has dimension 2 etc. The degree is a measure for how curvy these shapes are. A prime ideal has degree 1 if and only if it is generated by linear polynomials.

Example 1.36. Fix the polynomial ring $K[x, y, z]$ and let $f = xyz - x^2 - y^2 - z^2 + 1$ as in (1.1). The ideal $\langle f \rangle$ has dimension 2 and degree 3. Let $I$ be the ideal of its partial derivatives, as in Example 1.7. Then $I$ has dimension 0 and degree 5. The ideal $I + \langle f \rangle$, whose zeros are the singular points, has dimension 0 and degree 4.

Exercises

1. Show that the polynomial $f = 5x^3 - 25x^2y + 25y^3 + 15xy - 50y^2 - 5x + 25y - 1$ is a product of three linear factors in $R[x, y]$. Draw the plane curve $\{f = 0\}$.

2. For $n = 2$, define a monomial ordering $\prec$ such that $(2, 3) \prec (4, 2) \prec (1, 4)$.

3. Let $n = 2$ and fix the ideals $I = \langle x, y^2 \rangle$ and $J = \langle x^2, y \rangle$. Compute $I + J$, $I \cap J$, $IJ$ and $I^3J^4$. How many generators does the ideal $I^{123}J^{234}$ have?

4. The radical $\sqrt{I}$ of an ideal $I$ is the smallest radical ideal containing $I$. Prove that the radical of a primary ideal is prime. For a polynomial ring, prove that
   - The radical of a principal ideal is principal.
   - The radical of a monomial ideal is a monomial ideal.

5. Show that the following inclusions always hold and they are strict in general:
   $$\sqrt{I} \sqrt{J} \subseteq \sqrt{IJ} \quad \text{and} \quad \sqrt{\text{in}_<(I)} \subseteq \sqrt{\text{in}_<(J)}.$$

6. Using Gröbner bases, find the minimal polynomials of $\sqrt{6} + \sqrt{3}$ and $\sqrt{6} - \sqrt{3}$.

7. Find the implicit equation of the curve $\{(x^5 - 6, x^7 - 8) : x \in \mathbb{R}\}$.

8. Investigate the ideal $I = \langle x^3 - yz, y^3 - xz, z^3 - xy \rangle$. Is it radical? If not, find $\sqrt{I}$. Regarding $I$ as a system of 3 equations, what are its solutions in $\mathbb{R}^3$?

9. For the ideals $I$ and $\sqrt{I}$ in the previous exercise, determine the Hilbert function, the Hilbert series, the Hilbert polynomial, the dimension, and the degree.

10. Find an ideal in $\mathbb{Q}[x, y]$ whose reduced Gröbner basis (for the lexicographic order) has precisely 5 elements and there are precisely 19 standard monomials.

11. Prove: An ideal in a polynomial ring $K[x]$ is principal if and only if its reduced Gröbner basis is a singleton.

12. Let $I$ be the ideal generated by the $n$ elementary symmetric polynomials in $x_1, \ldots, x_n$. Pick a monomial order and find the initial monomial ideal $\text{in}_<(I)$.

13. Let $X$ be a $2 \times 2$-matrix whose $n = 4$ entries are variables. Let $I_s$ be the ideal generated by the entries of the matrix power $X^s$ for $s = 2, 3, 4, \ldots$. Investigate these ideals. What can you say about the dimension and degree of $I_s$?
(14) A symmetric $3 \times 3$-matrix with unknown entries has seven principal minors: three of size $1 \times 1$, three of size $2 \times 2$, and one of size $3 \times 3$. Does there exist an algebraic relation among these minors? Hint: lexicographic Gröbner basis.

(15) Prove that if $\text{in}_<(I)$ is radical then $I$ is a radical ideal. Does the converse hold?

(16) Does the cubic surface in Figure 1 contain any straight line? Find all such lines.

(17) Characterize all maximal, prime, radical and primary ideals in the ring $R = \mathbb{Z}$.

(18) Let $I$ be the ideal generated by all $2 \times 2$ minors of a $2 \times n$ matrix filled with $2n$ distinct variables. What is the degree and dimension of $I$ for $n = 2, 3, 4$?

(19) Find a prime ideal of degree three and dimension one in $n$ variables for

- $n = 2$,
- $n = 3$,
- $n = 3$ with further assumption that $h_I(1) = 4$.

(20) Compute the dimension and degree of the ideal generated by two random degree four polynomials in $n = 4$ variables, as in Example 1.19.
CHAPTER 2

Varieties

“Geometry is but drawn algebra”, Sophie Germain

A variety is the set of solutions to a system of polynomial equations in several unknowns. These are the main objects of study in algebraic geometry. Varieties are the geometric counterparts to ideals, which live on the algebraic side. We distinguish between affine varieties and projective varieties. The former arise from arbitrary polynomials, while the latter represent the zero sets of systems of homogeneous polynomials. Geometers prefer projective varieties because of their nice properties, explained in some of the results we present, like Theorem 2.21. But, for starters, our readers are invited to peruse the pictures shown in this chapter.

2.1. Affine varieties

Algebraic varieties represent solutions of a system of polynomial equations. Fix a field $K$ and consider polynomials $f_1, \ldots, f_k$ in $K[x] = K[x_1, \ldots, x_n]$. The (algebraic) variety defined by these polynomials is the set of their common zeros:

$$V(f_1, \ldots, f_k) := \{ p = (p_1, \ldots, p_n) \in K^n : f_1(p) = \cdots = f_k(p) = 0 \}.$$ 

Different sets of polynomials can define the same variety. For instance, (2.1)

$$V(f_1, f_2) = V(f_1^2, f_2^2) = V(f_1, f_1 + f_2).$$ 

Instead of thinking about the polynomials themselves, we consider the ideal they generate, $I = \langle f_1, \ldots, f_k \rangle$, and we define $V(I) := V(f_1, \ldots, f_k)$. A subset of $K^n$ is a variety if it has the form $V(I)$ for some ideal $I \subset K[x]$. Given any ideal $I \subset K[x]$, by Hilbert’s Basis Theorem 1.14 we can always find a finite set of generators. By Exercise 1, the definition of $V(I)$ does not depend on the choice of generators of $I$.

Remark 2.1. Two distinct ideals may define the same variety. For instance, for two non-constant polynomials $f_1$ and $f_2$, the ideal $\langle f_1^2, f_2^2 \rangle$ is strictly contained in $\langle f_1, f_2 \rangle = \langle f_1, f_1 + f_2 \rangle$, but they define the same variety in (2.1). Chapter 6 on the Nullstellensätze deals with this issue for fields $K$ that are either algebraically closed, like the complex numbers $K = \mathbb{C}$, or real closed, like the reals $K = \mathbb{R}$.

Algebraic geometry is the study of the geometry of varieties. As in many branches of mathematics, one considers the basic, irreducible building blocks for the objects of study. A variety $V(I)$ is called irreducible if it cannot be written as a union of proper subvarieties in $K^n$. In symbols, $V(I)$ is irreducible if and only if $V(I) = V(J) \cup V(J') \implies V(I) = V(J) \lor V(I) = V(J')$ for any ideals $J$ and $J'$.

We can decompose any variety into irreducible varieties. The underlying algebraic technique is primary decomposition. This will be discussed in the next chapter.
Example 2.2. Consider the ideal \( I = \langle xy \rangle \subset \mathbb{R}[x,y] \). Its variety \( \mathcal{V}(I) = \mathcal{V}(x) \cup \mathcal{V}(y) \) is a union of two lines. Hence, this is a reducible variety. However, \( I \) is the intersection of two larger ideals \( \langle x \rangle \) and \( \langle y \rangle \). Their varieties \( \mathcal{V}(x) \) and \( \mathcal{V}(y) \) are irreducible. This follows from Proposition 2.3 because \( \langle x \rangle \) and \( \langle y \rangle \) are prime.

For any field \( K \), we can turn \( K^n \) into a topological space, using the Zariski topology, in which varieties are closed sets. In this setting, the definition of an irreducible variety coincides with the definition of an irreducible topological space. If \( K = \mathbb{R} \) or \( K = \mathbb{C} \) then we also have the classical Euclidean topology on \( K^n \). This is much finer than the Zariski topology because it has many more open sets.

Our aim is to relate geometric properties of \( \mathcal{V}(I) \) to algebraic properties of \( I \). Consider a maximal ideal \( m := \langle x_1 - p_1, \ldots, x_n - p_n \rangle \subset K[x] \). The point \( (p_1, \ldots, p_n) \) lies in \( \mathcal{V}(I) \) if and only if \( I \subset m \). Given any subset \( W \subset K^n \), we consider the set of all polynomials that vanish on \( W \). This set is an ideal, denoted

\[
\mathcal{I}(W) := \{ f \in K[x] : f(p) = 0 \text{ for all } p \in W \}.
\]

Note that the set \( W \) is a variety if and only if \( W = \mathcal{V}(\mathcal{I}(W)) \). Furthermore, given any two varieties \( V \) and \( W \) in \( K^n \), we have \( V \subseteq W \) if and only if \( \mathcal{I}(W) \subseteq \mathcal{I}(V) \).

Proposition 2.3. A variety \( W \subset K^n \) is irreducible if and only if its ideal \( \mathcal{I}(W) \) is prime.

Proof. Suppose \( \mathcal{I}(W) \) is prime and \( W = \mathcal{V}(J) \cup \mathcal{V}(J') \). If \( W \neq \mathcal{V}(J) \) then there exists \( f \in J \) and \( v \in W \) such that \( f(v) \neq 0 \). Therefore, \( f \notin \mathcal{I}(W) \). For any \( g \in J' \) we know that \( fg \) vanishes on \( \mathcal{V}(J) \) and \( \mathcal{V}(J') \), hence on \( W \). Thus \( fg \in \mathcal{I}(W) \). As \( \mathcal{I}(W) \) is prime, we have \( g \notin \mathcal{I}(W) \). We conclude that \( J' \subset \mathcal{I}(W) \). By Exercise 2, this implies \( W = \mathcal{V}(\mathcal{I}(W)) \subset \mathcal{V}(J') \).

For the converse, suppose that \( W \) is irreducible and \( fg \in \mathcal{I}(W) \). Hence

\[
W = W \cap \mathcal{V}(fg) = W \cap (\mathcal{V}(f) \cup \mathcal{V}(g)) = (W \cap \mathcal{V}(f)) \cup (W \cap \mathcal{V}(g)).
\]

Without loss of generality, \( W = W \cap \mathcal{V}(f) \). This means that \( W \subseteq \mathcal{V}(f) \) and hence \( f \in \mathcal{I}(W) \). This argument proves that \( \mathcal{I}(W) \) is a prime ideal.

Remark 2.4. Proposition 2.3 points to the interplay between geometry and number theory. Prime ideals in a polynomial ring correspond to irreducible varieties, while prime ideals in the ring of integers \( \mathbb{Z} \) correspond to prime numbers (or zero). Hence irreducible varieties are to varieties what primes are to all integers.

Many examples of varieties appearing from applications are given as (closures) of images of polynomial maps. Often we think about the domain \( K^n \) as the space of parameters and the codomain as the space of possible (observable) outcomes – cf. Exercise 9. We note that the (Zariski) closure of the image must be irreducible.

Example 2.5. Consider two independent discrete random variables \( X \) and \( Y \) each one with \( n \) states. Probability distribution of \( X \) (resp. \( Y \)) are points \( (p_1, \ldots, p_n) \) (resp. \( (q_1, \ldots, q_n) \)) in \( \mathbb{R}^n \) with nonnegative entries that sum to 1. The probability that \( X \) (resp. \( Y \)) is in state \( i \) equals \( p_i \) (resp. \( q_i \)). The joint random variable \( (X,Y) \) has \( n^2 \) states. Consider he map that associates to a distribution of \( X \) and a distribution of \( Y \) the joint distribution of \( (X,Y) \). This map extends to a polynomial map \( \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{n^2} \). Explicitly, that polynomial map is given by

\[
(p_1, \ldots, p_n, q_1, \ldots, q_n) \mapsto (p_1q_1, p_1q_2, \ldots, p_1q_n, p_2q_1, \ldots, p_nq_n).
\]
For statistics, it is important to incorporate the requirement \( \sum p_i = \sum q_i = 1 \), namely by restricting the domain. We write the resulting map explicitly for \( n = 3 \):

\[
(p_1, p_2, q_1, q_2) \mapsto (p_1 q_1, p_1 q_2, p_1 (1 - q_1 - q_2), p_2 q_1, p_2 q_2, p_2 (1 - q_1 - q_2), (1 - p_1 - p_2) q_1, (1 - p_1 - p_2) q_2, (1 - p_1 - p_2)(1 - q_1 - q_2)).
\]

In Exercise 9 we ask for the varieties and ideals given by the images of these maps.

Prime ideals play a central role in algebraic geometry. This motivates the following definition. We now take \( R \) to be any commutative ring with unity. One important example is the quotient \( R = K[x]/I \) of a polynomial ring by some ideal \( I \).

**Definition 2.6.** The *spectrum* of the ring \( R \) is the set of proper prime ideals:

\[
\text{Spec}(R) := \{ p \subseteq R : p \text{ is a prime ideal} \}.
\]

The set \( \text{Spec}(R) \) is a topological space with the Zariski topology. Here the closed sets are the *varieties* \( \mathcal{V}(I) = \{ p \in \text{Spec} R : I \subseteq p \} \), where \( I \) is any ideal in \( R \).

The spectrum of the ring remembers a lot of information: all prime ideals and how they are related geometrically. In particular, \( \text{Spec}(K[x]) \) has points corresponding to *all* irreducible subvarieties of \( K^n \) - not only to usual points \( (p_1, \ldots, p_n) \in K^n \), which correspond to maximal ideals of the form \((x_1 - p_1, \ldots, x_n - p_n)\). In this manner, \( K^n \) is a subset of \( \text{Spec}(K[x]) \). Exercise 4 asks you to prove that the Zariski topology on \( K^n \) is the one induced one the Zariski topology on \( \text{Spec} K[x] \).

The polynomial ring \( K[x] \) consists of all (polynomial) functions on \( K^n \). We next introduce the ring \( K[W] \) of functions on a variety \( W \subset K^n \). Since we are interested in polynomial functions, there should be a surjection \( K[x] \rightarrow K[W] \).

Two functions on \( K^n \) coincide on the subvariety \( W \) if and only if their difference vanishes on \( W \). Thus the kernel of the above map equals \( I(W) \). We therefore define the ring of functions on \( W \) to be the quotient ring \( K[W] := K[x]/I(W) \). The advantage of this approach is that we may consider the ring \( K[W] \) as an object representing \( W \), without referring to any embedding. As before, we identify points \( p = (p_1, \ldots, p_n) \in W \) with maximal ideals \( (x_1 - p_1, \ldots, x_n - p_n) \subset K[W] \). The Zariski topologies on \( W \) and \( \text{Spec}(K[W]) \) are compatible, in the sense of Exercise 4.

**Example 2.7.** A parabola in \( \mathbb{R}^2 \) is defined by the equation \( y = x^2 \). The associated ideal equals \( I = (y - x^2) \). The ring of (polynomial) functions on the parabola equals \( \mathbb{R}[x, y]/I \). What are the Gröbner bases of \( I \) and what are the standard monomials? How about the dimension and the degree of \( I \) ?

Given two geometric objects \( X, Y \) and a map \( f : X \rightarrow Y \) between them, one may pull-back functions on \( Y \). Explicitly, given \( g : Y \rightarrow K \) we define the pull-back \( f^*(g) := g \circ f \). As we are dealing with algebraic varieties, we would like the pull-backs of polynomials to be polynomials. Hence, given an algebraic map \( f : W_1 \rightarrow W_2 \) between varieties, we would like the induced map \( f^* : K[W_2] \rightarrow K[W_1] \) to be a well-defined ring morphism. In Exercise 5 you will show that any ring morphism \( K[W_2] \rightarrow K[W_1] \) induces a map \( \text{Spec} K[W_2] \rightarrow \text{Spec} K[W_1] \). Hence, we may think about algebraic maps between varieties as morphisms among their rings of functions in the opposite direction. Using slightly more sophisticated language there is a contravariant functor, inducing an equivalence of categories of affine irreducible varieties (over \( K \)) and finitely generated integral \( K \)-algebras. We note that (algebraic) maps between varieties are continuous in the Zariski topology.
Example 2.8. As seen in Example 2.7, the ring of functions for the parabola is \( \mathbb{R}[x, y]/\langle y - x^2 \rangle \). Further, a line is simply represented by the polynomial \( \mathbb{R}[z] \) in one variable \( z \). The map that sends \( x \) to \( z \) and \( y \) to \( z^2 \) defines a ring isomorphism

\[
\mathbb{R}[x, y]/\langle y - x^2 \rangle \rightarrow \mathbb{R}[z].
\]

This corresponds to an isomorphism of varieties, between the line and the parabola:

\[
\mathbb{R} \ni z_0 \mapsto (z_0, z_0^2) \in \mathcal{V}(y - x^2).
\]

Indeed, notice that the function \( x \), that takes the first coordinate pulls back to the function \( z \) and the function \( y \) that takes the second coordinate pulls back to \( z^2 \).

Remark 2.9. Text books in algebraic geometry generally define affine algebraic varieties to be \( \text{Spec } R \), with its Zariski topology, for any (commutative, with unity) ring \( R \). Here \( R \) need not be a finitely generated algebra over a field \( K \). However, in this book all affine varieties come from zero sets of polynomials defined over \( K \).

We next note that the dependence on the field is crucial for many properties of ideals. Consider the squaring map \( K^1 \rightarrow K^1, \lambda \mapsto \lambda^2 \) from the affine line to itself.

- If \( K = \mathbb{C} \) then the squaring map is surjective.
- If \( K = \mathbb{R} \) then its image is the set of nonnegative real numbers. In both cases, the Zariski closure of the image is the whole line.
- If \( K = \mathbb{F}_p \) and \( p \neq 2 \), then the image is a proper subset of \( K^1 \). It coincides with its Zariski closure. What if we replace \( K \) by its algebraic closure?

From the perspective of ideals, it is instructive to consider \( I = \langle x^2 + 1 \rangle \) in \( K[x] \). The reader is asked to provide a description of \( \mathcal{V}(I) \) in Exercise 6.

Example 2.10. We consider the three ideals \( I_1 = \langle x^2 - y^2 \rangle \), \( I_2 = \langle x^2 - 2y^2 \rangle \) and \( I_3 = \langle x^2 + y^2 \rangle \) in \( K[x, y] \). The first one is not prime for any \( K \). The second one is not prime for \( K = \mathbb{R} \) or \( K = \mathbb{C} \). However, it is a prime ideal when \( K = \mathbb{Q} \). The last \( I_3 \) is not prime for \( K = \mathbb{C} \), but is a prime ideal for \( K = \mathbb{Q} \) or \( K = \mathbb{R} \).

We prove the last statement. Suppose \( fg \in I_3 \subset \mathbb{R}[x, y] \). This means that \( fg = (x^2 + y^2) \cdot h \), where \( f, g, h \in \mathbb{R}[x, y] \). By the Fundamental Theorem of Algebra, every homogeneous polynomial \( p \) in two variables has a unique (up to multiplication by constants) representation as a product of linear forms with complex coefficients \( p = \prod l_j \). If \( p \) has real coefficients, then the decomposition is stable under complex conjugation: for every \( j \), either \( l_j \) has real coefficients or \( \overline{l_j} \) must also appear in the decomposition. We have \( x^2 + y^2 = (x + iy)(x - iy) \). In the ring \( \mathbb{C}[x, y] \), without loss of generality, we may assume \( (x + iy)|f \). But then, by the above argument also \( (x - iy)|f \). Thus \( f = (x + iy)(x - iy) \prod l_i \in \mathcal{O}[x, y] \). However, \( \prod l_i \) is stable under conjugation, i.e. defines a real polynomial. Thus \( x^2 + y^2 \) divides \( f \) in \( \mathbb{R}[x, y] \).

The image of a variety under a polynomial map need not be closed if \( K = \mathbb{C} \). And, it need not be dense in its Zariski closure if \( K = \mathbb{R} \). This will be discussed in detail in Chapter 4 where the following definition will play an important role.

Definition 2.11. A subset \( A \subset \mathbb{R}^n \) is constructible if it can be described as a finite union of (set-theoretic) differences of varieties. Over the real numbers, a subset \( B \subset \mathbb{R}^n \) is semi-algebraic if it can be described as a set of solutions of a finite system of polynomial inequalities (both \( \geq \) and \( > \)) or a finite union of such.

Remark 2.12. Every constructible subset of \( \mathbb{R}^n \) is semi-algebraic, but the converse is not true. See below. The (set-theoretic) complement of a constructible set is constructible, and the complement of a semi-algebraic set is semi-algebraic.
Example 2.13. Take \( n = 2 \) and \( K = \mathbb{R} \). The singleton \( \mathcal{V}(x, y) = \{(0, 0)\} \) is constructible and hence so is \( \mathbb{R}^2 \setminus \{(0, 0)\} = \mathcal{V}(0) \setminus \mathcal{V}(x, y) \). The orthant \( \mathbb{R}^2_{>0} = \{(u, v) \in \mathbb{R}^2 : u \geq 0 \text{ and } v \geq 0\} \) is semi-algebraic but it is not constructible. (Why?) Its complement \( \mathcal{B} = \{(u, v) \in \mathbb{R}^2 : u < 0 \text{ or } v < 0\} \) is also semi-algebraic. Can you write \( \mathcal{B} \) as the set of solutions of a finite system of polynomial inequalities?

The two most important invariants of a variety \( V \) in \( K^n \) are its dimension and its degree. We define these as in Section 1.3, namely for the ideal \( \mathcal{I}(V) \subset K[x] \).

Example 2.14. Let \( V \) be a linear subspace of \( K^n \). The dimension of \( V \) as a variety equals its dimension as the linear subspace. The degree of \( V \) is one. Indeed, we may assume \( \mathcal{I}(V) = \langle x_1, \ldots, x_s \rangle \) where \( s = n - \dim(V) \). Using the isomorphism \( K[x]/\langle x_1, \ldots, x_s \rangle \simeq K[x_{s+1}, \ldots, x_n] \), the result follows from Example 1.22.

We note two important properties of dimension. If \( V_1 \subset V_2 \) then \( \dim(V_1) \leq \dim(V_2) \). If \( V_2 \) is irreducible, then the inequality is strict. The latter is not so easy to prove from the definition we gave. Readers are encouraged to consult a textbook in commutative algebra for alternative (but equivalent) definitions of dimension.

Here is a method for computing the dimension of a variety \( \mathcal{V}(I) \). It rests on the fact that \( \mathcal{V}(\text{in}_<(I)) \) is a union of coordinate subspaces, for any monomial order \( < \).

1. Compute a Gröbner basis of \( I \) and hence the initial monomial ideal \( \text{in}_<(I) \).
2. Suppose \( \text{in}_<(I) \) is generated by the monomials \( m_1, \ldots, m_k \). Find the smallest (with respect to cardinality) subset of variables \( S = \{x_{i_1}, \ldots, x_{i_d}\} \) such that every monomial generator \( m_j \) is divisible by some variable in \( S \).
3. The cardinality \( d \) of \( S \) is the common dimension of \( \mathcal{V}(\text{in}_<(I)) \) and \( \mathcal{V}(I) \).

The second most important invariant of a variety is the degree. We now provide its geometric interpretation. Suppose \( K \) is algebraically closed. A general subspace \( L \subset K^n \) of dimension equal to the codimension of \( V = \mathcal{V}(I) \subset K^n \) will intersect \( V \) only in finitely many, say \( d \), points. This is the degree of \( V \). Indeed, this follows inductively by the fact that a general linear form is not a zero-divisor in the ring of the variety. Hence, adding it to the ideal changes the Hilbert function in such a way that the dimension drops by one and the degree remains the same. In the zero dimensional (reduced) case the degree simply counts the number of points.

Some points on a variety are singular, like the four nodes of the cubic surface in Figure 1. Our aim is now to discuss singularities in general. We start with the case of a hypersurface, i.e. the variety \( \mathcal{V}(f) \) defined by one polynomial \( f \in K[x] \). A point \( p \in \mathcal{V}(f) \) is singular if all of the partial derivatives vanish, i.e. \( \frac{\partial f}{\partial x_i}(p) = 0 \) for \( i = 1, \ldots, n \). Thus the singular locus of \( f \) is the variety of the ideal \( \langle f, \frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \rangle \).

If this ideal has no zeros, we say that the hypersurface \( \mathcal{V}(f) \) is smooth.

Smoothness is a very important condition. It tells us that our variety can be locally arbitrary well approximated by a linear space - the tangent space. Let \( I = \langle f_1, f_2, \ldots, f_k \rangle \subset K[x] \) be a prime ideal, and suppose that its irreducible variety \( Y = \mathcal{V}(I) \) in \( K^n \) has dimension \( d \). A point \( p \in Y \) is singular if and only if the rank of the Jacobian matrix at the point \( p \) is smaller than the codimension \( Y \), i.e.

\[
\begin{pmatrix}
\frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\
\frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_2}{\partial x_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial f_k}{\partial x_1} & \cdots & \frac{\partial f_k}{\partial x_n}
\end{pmatrix}
(p) < n - d.
\]
A point that is not singular is called smooth. For a smooth point the inequality above turns into equality. The singular locus Sing(\(Y\)) is a variety in \(K^n\). Its ideal is the sum of the ideal \(I\) and the ideal generated by \((n - d) \times (n - d)\) minors of the Jacobian matrix. Note that the kernel of the Jacobian matrix evaluated at \(p\) is, by definition, the vector space parallel to the tangent space to \(V\) at \(p\). Hence, the definition of the smooth point assures that the tangent space and the variety are of the same dimension.

If a variety \(X \subset K^n\) is reducible and \(p\) belongs to more than one irreducible component of \(X\), we always say that \(p\) is singular in \(X\). If \(p\) belongs to a unique irreducible component \(Y\) then \(p\) is singular in \(X\) if and only if it is singular in \(Y\).

### 2.2. Projective varieties

The geometric objects we encountered so far are subsets of \(K^n\). We called them varieties, but more precisely we should refer to them as affine varieties. We now change our perspective with the aim of understanding projective algebraic varieties.

We start by recalling the construction of a projective space \(\mathbb{P}(V)\) over the vector space \(V\) of dimension \(n + 1\). Points of \(\mathbb{P}(V)\) correspond to lines in \(V\). Hence \([a_0 : \ldots : a_n] \in \mathbb{P}(V)\) represents a line going through the point \((a_0, \ldots, a_n) \in V\), where we assume that not all \(a_i\) are zero. Formally, \(\mathbb{P}(V)\) is the set of equivalence classes \([v]\), for \(v \in V \setminus \{0\}\), modulo the relation \(v_1 \sim v_2\) if and only if \(v_1 = \lambda v_2\) for some \(\lambda \in K^* = K \setminus \{0\}\). For the topological construction over \(\mathbb{R}\) or \(\mathbb{C}\), we note that each line through the origin in \(V\) intersects the unit sphere precisely in two points. Thus \(\mathbb{P}(V)\) may be regarded as a quotient of the sphere, identifying antipodal points. In particular, \(\mathbb{P}(V)\) is compact with respect to the classical topology.

On the subset \(S_i = \{a_i \neq 0\}\) of \(\mathbb{P}(V)\) we rescale and assume \(a_i = 1\). In this way we identify \(S_i\) with \(K^n\). As every point has some nonzero coordinate, the \(n + 1\) affine spaces \(S_i = K^n\) cover \(\mathbb{P}^n := \mathbb{P}(V)\). We obtain \(\mathbb{P}^n\) by glueing these charts.

As before, we are interested in functions on \(\mathbb{P}^n\). The first problem is that, for a polynomial \(f\), it does not make sense to evaluate \(f\) on \([a_0 : \ldots : a_n]\), as the result depends on the choice of the representative. It may even happen that \(f\) vanishes for some representatives, while it does not for others. Thus, from now on we focus on homogeneous polynomials, i.e. linear combinations of monomials of fixed degree. If \(f\) is a homogeneous polynomial of degree \(d\) in \(n + 1\) variables, then \(f(ta_0, \ldots, ta_n) = t^df(a_0, \ldots, a_n)\). In particular, \(f\) vanishes on some representative of \([a_0 : \ldots : a_n]\) if and only if it vanishes on any representative. Given homogeneous polynomials \(f_1, \ldots, f_k\), possibly of distinct degrees, we define the associated projective variety:

\[ V(f_1, \ldots, f_k) = \{[a_0 : \ldots : a_n] \in \mathbb{P}(V) : f_1(a_0, \ldots, a_n) = \cdots = f_k(a_0, \ldots, a_n) = 0\}. \]

An ideal \(I\) in \(K[x]\) is called homogeneous if it is generated by homogeneous polynomials \(f_1, \ldots, f_k\). In analogy to the affine case we define \(V(I) := V(f_1, \ldots, f_k) \subset \mathbb{P}^n\).

**Remark 2.15.** Homogeneous ideals contain (many) nonhomogeneous polynomials. For instance, \((x + y^2, y)\) is a homogeneous ideal. We refer to Exercise 11.

In theory, instead of considering a projective variety \(X \subset \mathbb{P}^n\) one can consider the affine cone \(\hat{X}\) over it, i.e. the variety defined by the same ideal, but in \(V = K^{n+1}\). The dimension and degree of a projective variety can be defined via its affine cone:

\[
\dim(X) := \dim(\hat{X}) - 1 \quad \text{and} \quad \deg(X) := \deg(\hat{X}).
\]
It is usually preferable to work with projective spaces and projective varieties. The geometry of varieties in \( \mathbb{P}^n \) is simpler than that in \( K^n \). For instance, parallel lines in \( K^2 \) do not intersect, but any two lines in \( \mathbb{P}^2 \) intersect. If \( X \) is any projective variety of degree \( \geq 2 \) then the affine cone \( \tilde{X} \) is always singular at the point \( 0 \in V \). However, this may be the only singular point of \( \tilde{X} \), in which case \( X \subset \mathbb{P}^n \) is smooth.

If \( Y \) is any variety in \( K^n \) then there is an associated projective variety \( \bar{Y} \) in \( \mathbb{P}^n \), called the projective closure of \( Y \). This can be defined by appealing to ideals. If \( I \subset K[x_1, \ldots, x_n] \) is the ideal of \( Y \) then the ideal \( \bar{I} \) of \( \bar{Y} \) lives in \( K[x_0, x_1, \ldots, x_n] \). It is generated by the following infinite set of homogeneous polynomials:

\[
\{ x_0^{\deg(g)} \cdot g\left(\frac{x_1}{x_0}, \ldots, \frac{x_n}{x_0}\right) : g \in I \}.
\]

The following is an algorithm for computing the ideal \( \bar{I} \) of the projective closure \( \bar{Y} \).

**Proposition 2.16.** Let \( I \) be an ideal in \( K[x_1, \ldots, x_n] \) and let \( \mathcal{G} \) be its reduced Gröbner basis with respect to a degree-compatible monomial ordering. Then \( \bar{I} \) is generated by the homogeneous polynomials in \( \{ 2.5 \} \) where \( g \) runs only over \( \mathcal{G} \).

**Proof.** Let \( f = f(x_0, x_1, \ldots, x_n) \) be any homogeneous polynomial in \( \bar{I} \). Suppose \( \mathcal{G} = \{ g_1, g_2, \ldots, g_s \} \). The dehomogenization \( f(1, x_1, \ldots, x_n) \) lies in \( I \) and hence its normal form modulo the Gröbner basis \( \mathcal{G} \) is zero. This gives a representation

\[
f(1, x_1, \ldots, x_n) = \sum_{i=1}^{s} h_i(x_1, \ldots, x_n)g_i(x_1, \ldots, x_n),
\]

where \( \deg(h_i g_i) \leq \deg(f) \) for all \( i \). By homogenizing the summands in this identity,

\[
f(x_0, x_1, \ldots, x_n) = \sum_{i=1}^{s} x_0^{\deg(f) - \deg(h_i g_i)} h_i\left(\frac{x_1}{x_0}, \ldots, \frac{x_n}{x_0}\right) \cdot g_i\left(\frac{x_1}{x_0}, \ldots, \frac{x_n}{x_0}\right).
\]

Hence \( f \) lies in the ideal generated by the set \( \{ 2.5 \} \) with \( I \) replaced by \( \mathcal{G} \). \( \square \)

**Corollary 2.17.** The dimension and degree of an affine variety \( Y \subset K^n \) are preserved when passing to its projective closure \( \bar{Y} \subset \mathbb{P}^n \):

\[
\dim(\bar{Y}) = \dim(Y) \quad \text{and} \quad \deg(\bar{Y}) = \deg(Y).
\]

**Proof.** The initial ideal of \( \bar{I} \) and the initial ideal of \( I \) have the same generators. These monomials in \( x_1, \ldots, x_n \) determine dimension and degree. \( \square \)

**Example 2.18.** Let \( I \) be the ideal generated by \( x_i - x_1^i \) for \( i = 2, 3, \ldots, n \). Then \( Y = \mathcal{V}(I) \) is a curve of degree \( n \) in \( K^n \). For the degree-lexicographic monomial order \( \prec \), the reduced Gröbner basis has \( \binom{n}{2} \) elements, and \( \text{in}_{\prec}(I) = \langle x_1, x_2, \ldots, x_{n-1} \rangle^2 \).

The ideal \( \bar{I} \) is minimally generated by the \( 2 \times 2 \)-minors of the \( 2 \times (n-1) \) matrix

\[
\begin{pmatrix}
x_0 & x_1 & x_2 & \cdots & x_{n-1} \\
x_1 & x_2 & x_3 & \cdots & x_n
\end{pmatrix}.
\]

The initial monomials of the \( \binom{n-1}{2} \) minors are the antidiagonal products. The projective variety \( \bar{Y} = \mathcal{V}(\bar{I}) \) is the rational normal curve of degree \( n \) in \( \mathbb{P}^n \).

We now return to our discussion of desirable properties of projective varieties.
Remark 2.19. If \( K = \mathbb{C} \) or \( K = \mathbb{R} \) then every projective variety is compact in the classical topology. Indeed, the projective space \( \mathbb{P}^n \) is compact, and every subvariety \( X \) is closed in the classical topology. Hence \( X \) is compact. If \( X \) is also smooth of dimension \( d \) then \( X \) is a compact real manifold, of dimension \( d \) if \( K = \mathbb{R} \) and of dimension \( 2d \) if \( K = \mathbb{C} \). Many interesting manifolds arise in this manner.

The following theorem discussed in greater detail in Chapter 4, Theorem 4.17 shows one aspect of a nice behaviour of projective varieties over \( \mathbb{C} \).

Theorem 2.20. Over an algebraically closed field, the image of a projective variety \( X \) under a polynomial map (that is defined on all of \( X \)) is Zariski closed.

Another nice property of projective varieties is their behavior under intersection.

Theorem 2.21. [34] 6.2 Theorem 6] Fix an algebraically closed field \( K \). Let \( X, Y \) be two projective varieties in the \( n \)-dimensional ambient space \( \mathbb{P}^n \), where \( d_1 = \dim(X) \) and \( d_2 = \dim(Y) \). Then their intersection \( X \cap Y \) has dimension at least \( d_1 + d_2 - n \). In particular, if \( d_1 + d_2 \geq n \) then \( X \cap Y \) is always non-empty.

The hypotheses are needed in this theorem. Consider the intersection of two surfaces in \( \mathbb{P}^3 \), where \( n = 3 \) and \( d_1 = d_2 = 2 \). The statement fails in affine space \( \mathbb{C}^3 \) where we can take two parallel planes. It also fails in \( \mathbb{P}^3 \) if the field is \( K = \mathbb{R} \).

Example 2.22. Consider the two surfaces \( X = \mathcal{V}(x_0^2 + x_1^2 - x_2^3 + x_3^2) \) and \( Y = \mathcal{V}(x_0^2 + x_1^2 + x_2^3 - x_3^2) \) in \( \mathbb{P}^3 \). Over \( \mathbb{C} \), their intersection is the union of four lines, so \( \dim(X \cap Y) = 1 = 2 + 2 - 3 \) as expected. However, over \( \mathbb{R} \), the intersection consists of two points, so \( \dim(X \cap Y) = 0 < 1 \), which would violate Theorem 2.21.

Many models in the sciences and engineering are given by homogeneous polynomial equations. Typically, these constraints arise from a construction familiar from linear algebra. Whenever one encounters such a model then it makes much sense to regard it as a projective variety. We close this section with two examples.

Example 2.23 (Nilpotent Matrices). An \( n \times n \)-matrix \( A \) is a point in a projective space \( \mathbb{P}^{n^2 - 1} \). The set of nilpotent matrices \( A \) is an irreducible projective variety \( X \subset \mathbb{P}^{n^2 - 1} \). We have \( \dim(X) = n^2 - n - 1 \) and \( \deg(X) = n! \). Indeed, \( X \) is a complete intersection. Its prime ideal \( \mathcal{I}(X) \) is generated by the coefficients of the characteristic polynomial of \( A \). For instance, if \( n = 2 \) then \( \mathcal{I}(X) = \langle \text{trace}(A), \det(A) \rangle \).

Example 2.24 (Kalman Varieties). In control theory, one is interested in the set of \( n \times n \)-matrices \( A \) that have an eigenvector in a given linear subspace of \( K^n \). This set is a projective variety in \( \mathbb{P}^{n^2 - 1} \). For instance, let \( n = 4 \) and consider \( 4 \times 4 \)-matrices that have an eigenvector with the last two coordinates zero. This Kalman variety has dimension 13 and degree 4 in \( \mathbb{P}^{15} \). It is defined by the \( 2 \times 2 \)-minors of

\[
\begin{pmatrix}
a_{31} & a_{41} & a_{11}a_{31} + a_{21}a_{32} + a_{31}a_{33} + a_{34}a_{41} & a_{11}a_{41} + a_{21}a_{42} + a_{31}a_{43} + a_{41}a_{44} \\
a_{32} & a_{42} & a_{12}a_{31} + a_{22}a_{32} + a_{32}a_{33} + a_{34}a_{42} & a_{12}a_{41} + a_{22}a_{42} + a_{32}a_{43} + a_{42}a_{44}
\end{pmatrix}.
\]

2.3. Geometry in Low Dimensions

Smooth projective varieties in low dimensions furnish interesting manifolds. Studying the geometry and topology of these manifolds leads to valuable insights that prove to be very useful also for understanding higher-dimensional scenarios.
2.3. GEOMETRY IN LOW DIMENSIONS

Figure 1. Schematic representation of \( P^2_\mathbb{R} \).

We work in projective spaces over the real numbers \( \mathbb{R} \) and over the complex numbers \( \mathbb{C} \). To distinguish these, we use the notations \( P^n_\mathbb{R} \) and \( P^n_\mathbb{C} \). We regard both of these as compact real manifolds, of dimension \( n \) and \( 2n \) respectively. Students of topology are encouraged to review the homology groups of these manifolds.

Let us start with \( n = 1 \). The real projective line \( P^1_\mathbb{R} \) is a circle. The complex projective line \( P^1_\mathbb{C} \) is a sphere, known as the Riemann sphere. Every subvariety of \( P^1_\mathbb{R} \) or \( P^1_\mathbb{C} \) is a finite collection of points, defined by a binary form \( f(x,y) \), i.e. a homogeneous polynomial in two variables. For instance, let \( f = x^{11}y - 11x^6y^6 - xy^{11} \). The variety \( X = V(f) \) has dimension 0 and degree 12 in \( P^1_\mathbb{C} \). These 12 points on the Riemann sphere are famous in the history of geometry and arithmetic. They serve as the vertices of the icosahedron in Felix Klein’s Lectures on the Icosahedron. Out of these 12 complex solutions four are real. The remaining eight come in four conjugate pairs.

We now move on to the \( n = 2 \) case. The real projective plane \( P^2_\mathbb{R} \) is a surface. However, it cannot be embedded homeomorphically in \( \mathbb{R}^3 \) (only in \( \mathbb{R}^4 \)), thus it is impossible to make a good picture. The simplest curve in the projective plane \( P^2_\mathbb{K} \) is a line \( L \), defined by one linear form in three variables. Of course, \( L \) is a projective line \( L \cong P^1_\mathbb{K} \), so the discussion in the previous paragraph applies to \( L \). The complement \( P^2_\mathbb{K} \setminus L \) is the affine plane \( \mathbb{K}^2 \). In particular, this complement is connected when \( K = \mathbb{R} \). The decomposition into \( L \) and \( \mathbb{K}^2 \) may be used to give a schematic picture of \( P^2_\mathbb{R} \). We identify \( \mathbb{R}^2 \) with the interior of the square. The boundary of the square should represent the line \( L \). However, we need to identify the opposite points of the boundary - this is often represented by putting directed arrows on the boundary as in Figure 1.

For any curve \( C \) in \( P^2_\mathbb{C} \), the complement \( P^2_\mathbb{C} \setminus C \) is connected because \( C \) is a surface in the 4-dimensional manifold \( P^2_\mathbb{C} \). By contrast, consider a smooth conic \( C \) in \( P^2_\mathbb{R} \). Then \( P^2_\mathbb{R} \setminus C \) has two connected components. One is a disk and the other is a Möbius strip (cf. Figure 2). The former is the inside of \( C \) and the latter is the outside of \( C \). A curve \( D \) in \( P^2_\mathbb{R} \) is called a pseudoline if \( P^2_\mathbb{R} \setminus D \) is connected and it is an oval otherwise. Every oval behaves like a conic in \( P^2_\mathbb{R} \): it has an inside and an outside.

**Theorem 2.25.** Let \( C \) be a smooth curve of degree \( d \) in the projective plane \( P^2_\mathbb{K} \). If \( K = \mathbb{C} \) then \( C \) is an orientable surface of genus \( g = \frac{(d-1)(d-2)}{2} \). If \( K = \mathbb{R} \) then \( C \) is a curve with at most \( g+1 \) connected components. If \( d \) is even then all components are ovals. If \( d \) is odd then one component is a pseudoline but all others are ovals.
Figure 2. Six topological ovals in $\mathbb{P}^2_{\mathbb{R}}$ with shaded interiors. It is possible to cut out the interior of the lower right oval from the square and glue together the remaining antipodal points on the boundary. This shows that indeed the complement of the interior of the oval is the Möbius strip.

Figure 3. Real elliptic curve $y^2 = x^3 - x$. The component on the left is an oval. The component on the right is a pseudoline. For a more complete picture see Figures 4 and 5.

Let us illustrate the above theorem for $d = 3$, i.e. $g = 1$. For example consider a cubic curve given in its Weierstrass form:

$$f(x, y, z) = zy^2 - x^3 - xz^2.$$  

We decompose the projective space $\mathbb{P}^2_{\mathbb{R}}$ into a line $L$ given by the equation $z = 0$ and its complement: the affine space $A = \mathbb{R}^2$. The cubic has two components: an oval and a pseudoline. We can see both of them by intersecting $C$ with $A$, as depicted in Figures 3 and 5. There is an additional point $P$ of the curve we do not see on
2.3. GEOMETRY IN LOW DIMENSIONS

Figure 4. The cone $zy^2 - x^3 - xz^2 = 0$ over an elliptic curve. The irreducible variety is depicted in yellow and blue according to two connected components in the real projective space. We intersect the cone with the grey plane given by $z = 1$. This corresponds to the affine chart, and the blue curve $C$ we obtain is exactly the same as in Figure 3.

How can we imagine the complex elliptic curve? This is not simple, as the correct picture just of the affine part would be in $\mathbb{C}^2 \simeq \mathbb{R}^4$. However, there exists a homeomorphism (but not a polynomial map!) of the complex curve $C$ with the real topological torus, i.e. the product of two circles $S^1 \times S^1$. It can be described as follows. We fix a point $p \in C$. For any point $q \in C$ consider a path $\gamma$ from $p$ to $q$ - this is always possible as over the complex numbers $C$ is connected. To a point $q$ we associate the complex number $\int_{\gamma} \frac{dx}{y}$. Identifying the complex plane with $\mathbb{R}^2$ we obtain a map $f : C \to \mathbb{R}^2$. It turns out that $f(q)$ depends on our choice of $\gamma$. Indeed, let us choose $p$ given by $z = 1, y = 0$ and $x = -1$. We may choose $q = p$ and $\gamma$ equal to the oval depicted in Figure 3. The integral $\int_{\gamma} \frac{dx}{y}$ will be a nonzero real number $\lambda$. Thus $f(p)$ may be equal to any integral multiple of $\lambda$. Further, on the curve $C$ there exists another loop $\gamma'$ giving rise to the integral $\int_{\gamma'} \frac{dx}{y}$ that is a complex number $\tau$. We may consider a lattice $M$, that is a subset of $\mathbb{C} \simeq \mathbb{R}^2$ given by all integral combinations $a\lambda + b\tau$ for $a, b \in \mathbb{Z}$. We know that $f(p)$ may be any point in $M$. Let $\pi : \mathbb{R}^2 \to \mathbb{R}^2/M$ be the natural projection. The map $\pi \circ f : C \to \mathbb{R}^2/M$ is now well-defined!

this picture, that belongs to the line $L$. It is given by $z = x = 0$ and $y = 1$. We may consider the surface in $\mathbb{R}^3$ that is the affine cone over our curve, as in Figure 4.
Figure 5. A cone $zy^2 - x^3 - xz^2 = 0$ over an elliptic curve, as in Figure 4. The grey sphere represents the projective space $\mathbb{P}^2_\mathbb{R}$, where we have to identify the antipodal points. The intersection of the surface with the sphere has three connected components. Two of them are identified, when we identify the antipodal points. These two components correspond to the oval – indeed cutting it out of the sphere, separates it into two pieces, even after identifying antipodal points. The other component corresponds to a pseudo-line. It does not separate a sphere after identifying the antipodal points. The points on the blue curve $C$ in Figure 4 correspond to pairs of antipodal points on the blue curve in this picture with one exception. This curve has one more pair of antipodal points - these are represented by a thickened blue point. Indeed, the line through that point is parallel to the grey plane in Figure 4. This point corresponds to the unique point of the projective curve that does not belong to the affine chart given by $z = 1$. It is precisely $z = 0, x = 0, y = 1$.

As $\mathbb{R}^2/M$ may be identified with the torus, we indeed obtain a homeomorphism $C \simeq \mathbb{R}^2/M$. The real part of the curve $C$ is mapped to two disjoint circles, as shown in Figure 6. Indeed, both the oval and the pseudoline are circles - they are only distinguished by their embedding in the real projective plane.

Remark 2.26. We contrast the topological torus mentioned here with the algebraic torus $(\mathbb{C}^\ast)^n$ playing a central role in Chapters 8 and 10. Indeed, a variety with a dense algebraic torus action will be called toric. The elliptic curve $C$ is the most basic example of a smooth, projective variety that is not toric.
2.3. GEOMETRY IN LOW DIMENSIONS

Figure 6. Two pictures showing a real torus, that is homeomorphic to the elliptic curve. The left picture presents the torus as $\mathbb{R}^2/\mathbb{Z}^2$. The right one is the familiar figure we know from topology. The two thickened circles, on both pictures, correspond to the real part of the curve.

Figure 7. Real part of a cuspidal curve

Remark 2.27. Elliptic curves make probably their first appearance in third century. Diophantus of Alexandria, in modern terms, asked for a (positive) rational point on a specific elliptic curve $y(6 - y) = x^3 - x$. As we argued above, an elliptic curve has a structure of a group (torus). The geometric interpretation of this was already well-known in 19th century. Since early 20th century, elliptic curves play a central role in (modern) number theory (studied mainly over fields of finite characteristic or rational numbers). By the end of the 20th century the group structure (over fields with finite characteristic!) started to be intensively used in applied cryptography.

Example 2.28. Let us consider a cuspidal curve defined by $x^3 - y^2$. Over $\mathbb{R}$ it is presented in Figure 7. How can we draw it over $\mathbb{C}$? If we identify $\mathbb{C}$ with $\mathbb{R}^2$ we obtain a surface in $\mathbb{R}^4$. Indeed:

$$(x_1 + ix_2)^3 - (y_1 + iy_2)^2 = 0 \Leftrightarrow x_1^3 - 3x_1x_2^2 = y_1^2 - y_2^2 \text{ and } 3x_1^2x_2 - x_2^3 = 2y_1y_2.$$
Hence, interpreted as a surface in $\mathbb{R}^4$ the variety is cut out by two polynomials. Although we may not make a picture in $\mathbb{R}^4$ we can project the given surface to $\mathbb{R}^3$. The result is the surface presented in Figure 7 together with the black line, that is the real part.

The surface seems more singular - this is the result of projection. The original surface in $\mathbb{R}^4$ has just one singular point. In Chapter 4 methods allowing the computation of projections of algebraic varieties are presented.

### Exercises

1. Prove that the definition of $\mathcal{V}(I)$ does not depend on the choice of the generators of $I$.
2. (a) Show that $J \subseteq I$ implies $\mathcal{V}(I) \subseteq \mathcal{V}(J)$.
   (b) Show that for any subsets $A, B \subseteq K^n$ if $A \subset B$ then $\mathcal{I}(B) \subseteq \mathcal{I}(A)$.
   (c) Give counterexamples to both opposite implications.
3. Prove that varieties (in $K^n$) satisfy the axioms of closed sets.
4. By identifying the point $(p_i) \in K^n$ with the prime ideal $\langle x_1 - p_1, \ldots, x_n - p_n \rangle$ consider $K^n$ as a subset of $\text{Spec} \ K[x]$. Show that the Zariski topology induced from $\text{Spec} \ K[x]$ to $K^n$ is the Zariski topology on $K^n$.
5. Show that a morphism of rings $f : R_1 \to R_2$ induces a map $f^* : \text{Spec} \ R_2 \to \text{Spec} \ R_1$, by proving that a pull-back of a prime ideal is prime. Show that the induced map is continuous with respect to the Zariski topology.
6. Describe the variety $\mathcal{V}(I)$ in the affine line $K^1$ for $I = \langle x^2 + 1 \rangle$ when $K = \mathbb{C}, \mathbb{R}, \mathbb{Q}$. Also, describe $\mathcal{V}(I) \subset \text{Spec}(K[x])$ for each of these three fields.
7. Realize the set of $n \times n$ nilpotent matrices as an affine variety. What is its dimension?
8. (a) Consider a polynomial $f \in K[x]$ (e.g. $f = x$). Let $D$ be the (open) set $D_f = \{p \in K^n : f(p) \neq 0\}$. Construct an affine variety $V$ and a polynomial map inducing a bijection $V \to D$.
   (b) Realize nondegenerate $n \times n$ matrices as an affine variety.
9. (a) Use (or not) your favorite computer algebra system to determine the ideal of the image of the map given by formula (2.3). What is the meaning of the lowest degree polynomial in this ideal?
   (b) Describe the ideal of the image of the map given by formula (2.2).
   (c) Generalize the previous point to more (independent) variables possibly with different (but finite) number of states.
10. Determine for which prime numbers $p$, the ideal $I_2 = \langle x^2 - 2y^2 \rangle \subset \mathbb{F}_p[x,y]$ is prime.
11. For a polynomial $f = \sum \alpha \in \mathbb{N}^n a_{\alpha} x^{\alpha}$ we call the degree $k$ part of $f$ the homogeneous polynomial $\sum_{\alpha:|\alpha|=k} a_{\alpha} x^{\alpha}$.
   (a) Provide an example of a homogeneous ideal generated by nonhomogeneous polynomials.
   (b) Prove that an ideal $I = \langle f_1, \ldots, f_j \rangle$ is homogeneous if and only if for any $f_i$ and any $k$ the degree $k$ part of $f_i$ belongs to $I$.
   (c) Propose an algorithm that, given a set of generators of $I \subset K[x]$, decides if $I$ is a homogeneous ideal.
12. (a) Let $I \subset K[x]$ be a monomial ideal. Prove that $\mathcal{V}(I)$ is a union of (some) vector subspaces of $K^n$ spanned by basis vectors.
(b) How do you characterize the sets of basis vectors that span a subspace belonging to \( \mathcal{V}(I) \)?

(13) Draw various pseudolines in \( \mathbb{P}^2_{\mathbb{R}} \), in analogy to Figure 2. Topologically, what is the complement of a pseudoline?

(14) Can you solve the problem of Diophantus of Alexandria in Remark 2.27?

Hint: Consider a tangent line to the elliptic curve at the point \((-1, 0)\).
CHAPTER 3

Solving and Decomposing

Solving systems of polynomial equations is a key task in nonlinear algebra. But, what does it mean to solve such a system? How should the solutions be presented? The answer to this question depends on the dimension of the variety of solutions. If the variety is zero-dimensional then it consists of finitely many points in $K^n$ and we aim to list each point explicitly. If $K = \mathbb{R}$ or $K = \mathbb{C}$ then this is usually done by displaying a floating point approximation to each of the $n$ coordinates of a solution.

If the solution variety has positive dimension then it has infinitely many points and we cannot list them all. In that case, the answer consists of a description of each irreducible component. Algebraically, this leads us to the topic of primary decomposition. If the given ideal is not radical then its constituents are primary ideals and we distinguish between minimal primes and embedded primes. To some readers, these objects may seem unnatural at first. However, they become quite natural in the setting of linear partial differential equations with constant coefficients.

3.1. Zero-dimensional Ideals

Let $K$ be a field and consider the polynomial ring $K[x]$ in one variable $x$. Every ideal in $K[x]$ is principal, so it has the form $I = \langle f \rangle$. The variety $V(I)$ consists of the zeros of $f$ and is zero-dimensional (unless $f = 0$). The polynomial $f$ has a unique factorization $f = \prod_{i=1}^{k} g_i^{a_i}$, where each $g_i$ is irreducible and the $a_i$ are positive integers. The set of solutions decomposes as $V(I) = V(g_1) \cup \cdots \cup V(g_k)$.

On the level of ideals we have the following decomposition as an intersection:

$$I = \langle g_1 \rangle^{a_1} \cap \cdots \cap \langle g_k \rangle^{a_k}.$$

This primary decomposition remembers the multiplicity $a_i$ of each irreducible factor $g_i$, so it contains more information than the irreducible decomposition of $V(I)$.

The precise nature of the decomposition depends on the field $K$. If $K$ is algebraically closed, such as $K = \mathbb{C}$, then each factor $g_i$ is a linear polynomial $g_i(x) = x - u_i$, where $u_1, \ldots, u_k$ are the zeros of $f$. If $K = \mathbb{R}$ then each $g_i$ is either linear or quadratic. If $K = \mathbb{Q}$ then $g_i$ can have arbitrarily high degree. In each case, the quotient ring $K[x]/\langle g_i \rangle$ is a field, namely it is an algebraic extension of $K$.

**Example 3.1.** Consider the polynomial $f = x^3 - 2x^2 + x - 2 \in \mathbb{R}[x]$. We have:

$$\langle f \rangle = \langle x - 2 \rangle \cap \langle x^2 + 1 \rangle.$$

Here the first ideal corresponds to the real solution 2, while the second to the pair of complex solutions $i$ and $-i$. Such factorizations are easy to compute in all computer algebra systems. What if we now replace the given polynomial by $g = x^3 - 2x^2 + x - 1$? How does the ideal $\langle g \rangle$ decompose in $\mathbb{R}[x]$? And in $\mathbb{C}[x]$?
Polynomials of degree \( m \) can have up to \( m \) zeros. The number \( m \) can be large. Often we are not interested in all the zeros, but only in specific ones. For instance, we might only be interested in solutions that are real and positive. This restriction is very important for many applications, e.g. in statistics and mathematical biology.

**Example 3.2.** Let \( I = (x^m - x - 1) \), where \( m \geq 2 \). The variety \( \mathcal{V}(I) \) has \( m \) complex points but only one of them is real and positive. Thus, \( \mathcal{V}(I) \cap \mathbb{R}_{>0} \) is a singleton. This follows from Descartes’ Rule of Signs, which states that the number of positive real solutions is bounded above by the number of sign alternations in the coefficient sequence. If \( m \) is even then there is also one negative solution.

In many applications one encounters polynomials whose coefficients depend on parameters. For instance, let \( \epsilon \) be an unknown that represents a very small positive real number. Let \( \mathbb{Q}(\epsilon) \) be the field of rational functions in that unknown and \( K = \overline{\mathbb{Q}(\epsilon)} \) its algebraic closure. Elements in \( K \) can be expressed by their Puiseux series expansion. This is analogous to the floating point expansion of numbers in \( \mathbb{R} \).

**Example 3.3.** The polynomial \( f = \epsilon^2 x^3 + x^2 + x - \epsilon \) is irreducible in \( \mathbb{Q}(\epsilon)[x] \). It factors into three linear factors \( f = (x - u_1)(x - u_2)(x - u_3) \) in \( K[x] \). We have

\[
\begin{align*}
  u_1 &= -\epsilon^2 + 1 + \epsilon^2 + 2\epsilon^4 + 3\epsilon^5 + 5\epsilon^6 + 10\epsilon^7 + \cdots \\
  u_2 &= -1 - \epsilon - 3\epsilon^3 + 3\epsilon^4 - 16\epsilon^5 + 32\epsilon^6 - 121\epsilon^7 + \cdots \\
  u_3 &= \epsilon - \epsilon^2 + 2\epsilon^3 - 5\epsilon^4 + 13\epsilon^5 - 37\epsilon^6 + 111\epsilon^7 + \cdots
\end{align*}
\]

Each of these three roots is an algebraic number over \( \mathbb{Q}(\epsilon) \) which we have written in its expansion as a Puiseux series. If we think of \( \epsilon \) as a very small positive quantity then we have \( u_1 \sim -\epsilon^2 \), \( u_2 \sim -\epsilon^0 \) and \( u_3 \sim \epsilon^1 \). The exponents \(-2, 0, 1\) tell us the asymptotic behavior. They are known as tropical solutions; cf. Chapter 7.

We have seen that solving a polynomial equation \( f = 0 \) amounts to decomposing the principal ideal \( I = \langle f \rangle \), i.e. presenting it as an intersection of simpler ideals. The situation is analogous for systems of polynomials in \( n \geq 2 \) variables, i.e. ideals \( I \subset K[x] \), where \( x = (x_1, \ldots, x_n) \). Suppose now that \( K \) is algebraically closed and assume that \( \mathcal{V}(I) \) is zero-dimensional. This means that the quotient ring \( K[x]/I \) is a finite-dimensional vector space over \( K \). A basis is given by the standard monomials for a given monomial order. The number of standard monomials is an upper bound for the cardinality of \( \mathcal{V}(I) \). Equality holds if and only if \( I \) is radical.

We will see in the next section that our zero-dimensional ideal \( I \) decomposes as

\[
I = \bigcap_{i=1}^{k} q_i
\]

where \( \text{rad}(q_i) \) is a prime ideal. Every prime ideal of dimension 0 in \( K[x] \) is a maximal ideal, so each \( \text{rad}(q_i) \) is a maximal ideal. Since \( K \) is algebraically closed, \( \mathcal{V}(q_i) \) is a point in \( K^n \). These points are precisely the solutions to our system.

**Example 3.4.** Let \( n = 2 \) and \( I = \langle xy, x^2 - x, y^2 - y \rangle \). We have:

\[
I = \langle x, y \rangle \cap \langle x - 1, y \rangle \cap \langle x, y - 1 \rangle.
\]

The variety of this radical ideal consists of three points: \( \mathcal{V}(I) = \{(0, 0), (1, 0), (0, 1)\} \).

Note however that when the given ideal is not radical we cannot express it as an intersection of maximal ideals. This should not be surprising: already in the case of one variable, if a root had a multiplicity we needed powers of linear forms.
Example 3.5. Let \( I = \langle xy, y^2 - y, x^2y - x^2 \rangle \). We have the decomposition
\[
I = \langle y - 1, x \rangle \cap \langle y, x^2 \rangle.
\]
The varieties of both ideals are points: \((0, 1)\) and \((0, 0)\) respectively. However, the second ideal remembers additional data, it is not simply equal to \(\langle x, y \rangle\), which indicates a 'multiplicity' of the solution at the point \((0, 0)\). We are now equipped with tools to measure this multiplicity! The degree of \(I\) equals three. The first ideal in the decomposition contributes with degree one, while the second with degree two.

We close this section with the example that was seen in Exercise 8 of Chapter [1]

Example 3.6. Fix the rationals \( K = \mathbb{Q} \) and \( I = \langle x^3 - yz, y^3 - xz, z^3 - xy \rangle \) in \( K[x, y, z] \). This ideal is the following irredundant intersection of 11 distinct ideals:
\[
I = \langle z - 1, y - 1, x - 1 \rangle \cap \langle z - 1, y + 1, x + 1 \rangle \cap \langle z - 1, x + y, y^2 + 1 \rangle
\cap \langle z + 1, y - 1, x + 1 \rangle \cap \langle z + 1, y + 1, x - 1 \rangle \cap \langle z + 1, x - y, y^2 + 1 \rangle
\cap \langle y - 1, x + z, z^2 + 1 \rangle \cap \langle y + 1, x - z, z^2 + 1 \rangle
\cap \langle y - z, x + 1, z^2 + 1 \rangle \cap \langle x - 1, y + z, z^2 + 1 \rangle.
\]
The first intersectand is the following primary ideal with radical \(\text{rad}(Q) = \langle x, y, z \rangle\):
\[
Q = \langle x^2y, x^2z, xy^2, xz^2, y^2z, yz^2, x^3 - yz, y^3 - xz, z^3 - xy \rangle.
\]
Each of the other 10 intersectands is a prime ideal. If we were to replace \( K \) by the complex numbers \( \mathbb{C} \) then six of the prime ideals decompose further:
\[
\langle x - 1, y - i, z + i \rangle \cap \langle x - 1, y + i, z - i \rangle
\]
We learn that \( V(I) \) consists of 17 complex points. Among these, only five are real.

3.2. Primary Decomposition

The idea of decomposing a mathematical object into simpler pieces is important. In this section we present a general theory of decomposing ideals. We shall express them as intersections of simpler ideals. Our point of departure is the following proposition. It shows how algebraic varieties may be decomposed.

Proposition 3.7. Any variety in \( K^n \) can be uniquely represented as a finite union of irreducible varieties (pairwise not contained in each other).

Proof. We start by proving the existence of such a decomposition. Any variety \( W \) is either irreducible or may be represented as a union \( W_1 \cup V_1 \). We may continue presenting \( W_1 \) as a union \( W_2 \cup V_2 \) etc. We obtain an ascending chain of ideals, \( I(W_1) \subseteq I(W_2) \subseteq \ldots \). This chain stabilizes by Hilbert Basis Theorem. Thus the decomposition procedure finishes with finitely many irreducible varieties.

Suppose we have two irreducible decompositions \( V_1 \cup \cdots \cup V_k = W_1 \cup \cdots \cup W_k \). As each \( W_{i_0} \) is irreducible and covered by \( \bigcup_j (V_j \cap W_{i_0}) \) we have \( W_{i_0} \subseteq V_{j_0} \) for some \( j_0 \). But similarly \( V_{j_0} \subseteq W_{i_0} \) for some \( i_1 \). As we cannot have \( W_{i_0} \subseteq W_{i_1} \) it follows that \( W_{i_0} = V_{j_0} \). Hence, for every component \( W_{i_0} \) there exists a (unique) component \( V_{j_0} \) equal to it. The uniqueness of the decomposition follows. \( \square \)

In what follows we present a vast generalization of the following two basic facts:

1. Every integer \( n > 1 \) can be uniquely decomposed as a product of powers of prime numbers:
\[
n = p_1^{a_1} \cdots p_k^{a_k}.
\]
(2) Any variety can be uniquely decomposed as a union of irreducible varieties - Proposition 3.7.

The algebraic notion that connects the first (number-theoretic) and the second (geometric) statement is that of an ideal. Indeed, any integer $n$ can be identified with the ideal $\langle n \rangle$ in the ring $\mathbb{Z}$. The elements of $\langle n \rangle$ are the integers that are divisible by $n$. The ideal $\langle n \rangle$ is prime if and only if $n$ is a prime number. We can restate fact (1) in terms of intersections of powers of prime ideals as follows:

(1') Every nonzero ideal $I \subset \mathbb{Z}$ has a unique decomposition as an intersection

$I = (I_1)^{a_1} \cap \cdots \cap (I_k)^{a_k},$

where the $I_i$ are prime ideals.

Likewise, over an algebraically closed field, we have an identification of varieties with radical ideals (cf. Chapter 6). This yields the following restatement of (2):

(2') Every radical ideal $I \subset \mathbb{C}[x]$ has a unique decomposition as an intersection of prime ideals, pairwise not contained in each other:

$I = p_1 \cap \cdots \cap p_k.$

From the above examples we see that our aim should be to decompose ideals $I$ in a ring $R$. Further, decomposition should mean that we present them as an intersection of other ideals. However, we still need to answer the following questions:

(1) What kind of ideals should be allowed in the intersection?
(2) What restrictions should be put on the ring $R$?
(3) Can we expect the decomposition to be unique?

We start with the first question. The number-theoretic example suggests all ideals might be intersections of powers of prime ideals. But this is not true.

Example 3.8. The ideal $I = \langle x^2, y \rangle$ is not an intersection of powers of prime ideals in $\mathbb{C}[x, y]$. Indeed, suppose $I = \bigcap p_i^{a_i}$. For all $i$, we have $p_i \supset I$. Hence $p_i = \langle x, y \rangle$: this is the only prime ideal containing $I$. But, $I$ is not a power of $\langle x, y \rangle$.

It turns out that the right class of ideals are primary ideals. Recall that $I$ is primary if and only if, for all $a, b$, if $ab \in I$ and $a \notin I$ then $b^n \in I$ for some $n$.

Now we pass to the second question: which rings $R$ to take? Clearly, $\mathbb{Z}$ and $\mathbb{C}[x]$ share a lot of nice properties. But there is a larger class of rings that works.

Definition 3.9. A ring $R$ is Noetherian if every ascending chain of ideals $I_1 \subseteq I_2 \subseteq I_3 \subseteq \ldots$ stabilizes, i.e. there exists $k$ such that $I_k = I_{k+1} = I_{k+2} = \ldots$.

Noetherian rings are named after one of the most famous German algebraists: Emmy Noether. A hint how important they are is given in Exercises 4 and 5. Note that $\mathbb{Z}$ and $K[x]$ are Noetherian rings because their ideals are finitely generated.

Before stating our main existence theorem let us introduce a technical definition.

Definition 3.10. An ideal $I$ is irreducible if and only if whenever $I = J_1 \cap J_2$ for some ideals $J_1, J_2$ then $I = J_1$ or $I = J_2$.

Theorem 3.11. Let $I$ be an ideal in a Noetherian ring $R$. Then there exist primary ideals $q_1, \ldots, q_k$ in $R$ such that:

$I = q_1 \cap \cdots \cap q_k.$
3.2. PRIMARY DECOMPOSITION

Proof. First we show that every ideal in \( R \) is a finite intersection of irreducible ideals. Suppose not, and let \( I \) be an ideal that cannot be presented in this way. In particular, it is not irreducible. Thus, \( I = J_1 \cap J_2 \) and each \( J_i \) strictly contains \( I \).

If \( J_1 \) and \( J_2 \) are finite intersections of irreducible ideals, then so is \( I \). Hence, we may assume \( J_1 \) cannot be presented in such a way. Let \( I_2 := J_1 \). We have \( I_1 \subsetneq I_2 \).

We repeat the construction starting with \( I_2 \) and get an ideal \( I_3 \) with \( I_1 \subsetneq I_2 \subsetneq I_3 \), where \( I_3 \) is not a finite intersection of irreducible ideals. Continuing, we get a chain of strictly ascending ideals. However, this is not possible in a Noetherian ring.

We next prove that every irreducible ideal \( q \) in \( R \) is primary. By replacing the ring \( R \) with \( R/q \), we may assume \( q = \{0\} \). Suppose that \( ab = 0 \) and \( a \neq 0 \). We have to prove that \( b \) is nilpotent. Consider the following ascending chain of ideals:

\[
\{x \in R : bx = 0\} =: \text{Ann}(b) \subseteq \text{Ann}(b^2) \subseteq \text{Ann}(b^3) \subseteq \ldots
\]

Since \( R \) is Noetherian, we have \( \text{Ann}(b^n) = \text{Ann}(b^{n+1}) \) for some \( n \). We claim that \( \langle a \rangle \cap \langle b^n \rangle = \{0\} \). Indeed, suppose \( \lambda a = \mu b^n \in \langle a \rangle \cap \langle b^n \rangle \) for some \( \lambda, \mu \in R \).

\[
0 = \lambda ab = \mu b^{n+1}.
\]

Hence, \( \mu \in \text{Ann}(b^{n+1}) = \text{Ann}(b^n) \). Thus, \( \mu b^n = 0 \). As \( \{0\} \) was assumed to be irreducible, and \( \langle a \rangle \supsetneq \{0\} \), we must have \( b^n = 0 \). This completes the proof. \( \square \)

We now pass to the third question, concerning uniqueness. We need not assume that \( R \) is Noetherian, as long as the ideal is an intersection of finitely many primary ideals: \( I = \bigcap_{i=1}^k q_i \). Here it is assumed that each \( q_i \) is necessary, i.e. \( \bigcap_{j \neq i} q_j \not\subset q_i \) for all \( 1 \leq i \leq k \). The next two lemmas suggest grouping the \( q_i \) by their radical.

**Lemma 3.12.** The radical of a primary ideal \( q \) is the unique smallest prime ideal containing it.

The proof is left as Exercise 6 for the reader. A primary ideal \( q \) with radical equal to \( p \) is called \( p \)-primary. An easy way to create a primary ideal is to take a power of a prime ideal generated by some of the variables in \( C \). Not true that the power of a prime ideal is always a primary ideal, even for \( C \).

**Example 3.13.** Let \( P \) be the ideal generated by the nine 2 \( \times \) 2 minors of a 3 \( \times \) 3 matrix \( X = (x_{ij}) \) of unknowns. This ideal is prime. The ideal \( P^2 \) is not primary. To see this, we verify (using Gröbner bases) that \( x_{ij} \cdot \text{det}(X) \) lies in \( P^2 \) for all \( 1 \leq i, j \leq 3 \). However, \( P^2 \) is generated by quartics and it contains no cubics.

We now focus on \( p \)-primary ideals for a fixed prime \( p \) in a Noetherian ring \( R \).

**Lemma 3.14.** If \( q_1, \ldots, q_k \) are \( p \)-primary ideals, then so is \( q_1 \cap \cdots \cap q_k \).

**Proof.** The following shows that the radical of \( I := \bigcap_{i=1}^k q_i \) equals \( p \):

\[
a \in \text{rad}(I) \iff \exists n : a^n \in I \iff \exists n \forall i : a^n \in q_i \iff \forall i : a \in \text{rad}(q_i) = p \iff a \in p.
\]

To see that \( I \) is primary, assume \( ab \in I \) and \( a \not\in I \). Then \( a \not\in q_{i_0} \) for some \( i_0 \). Since \( ab \in q_{i_0} \) and \( q_{i_0} \) is primary, \( b \in \text{rad}(q_{i_0}) = p = \text{rad}(I) \). Hence \( b^n \in I \) for some \( n \). \( \square \)

Lemma 3.14 suggests that, given a primary decomposition \( I = \bigcap_{i=1}^k q_i \), we should first group together \( q_i \)’s that have the same radical and replace them by
their intersection. Hence, we can always bring any presentation \( I = \bigcap_{i=1}^{k} q_i \) to the following form, which from now on will be called minimal primary decomposition:

\[
I = q_1 \cap q_2 \cap \cdots \cap q_k.
\] (3.1)

Here we assume that all \( q_i \)’s are primary ideals, that the \( q_i \)'s have pairwise distinct radicals, and that \( \bigcap_{j \neq i_0} q_j \not\subseteq q_{i_0} \) for all \( 1 \leq i_0 \leq k \).

To sum up, we have proved that:

1. in a Noetherian ring every ideal has a (finite) primary decomposition, and
2. (in any ring) any (finite) primary decomposition of an ideal can be changed to a minimal one (apply Lemma 3.14 and remove unnecessary ideals).

We next show that minimal primary decompositions may still be not unique.

**Example 3.15.** The following are two minimal primary decompositions:

\[
(x^2, xy) = (x) \cap (x, y)^2 = (x) \cap (x^2, y) \subset \mathbb{C}[x, y].
\] (3.2)

It turns out that, while the primary ideals \( q_i \) in the decomposition (3.1) are not unique, their radicals are. Recall the ideal quotient, \( I : a = \{ b \in R : ab \in I \} \).

**Theorem 3.16.** For any ideal \( I \) in a ring \( R \), the set of radicals \( \text{rad}(q_i) \) does not depend on the choice of a minimal primary decomposition (3.1). These radicals are precisely the prime ideals of the form \( \text{rad}(I : a) \) for some \( a \in R \). Further, if \( R \) is Noetherian, then these are also (exactly) prime ideals of the form \( I : a \) for \( a \in R \).

**Proof.** Fix a minimal primary decomposition \( I = \bigcap_{i=1}^{k} q_i \). Since intersection commutes with ideal quotients, we have \( I : a = \bigcap_{b=1}^{k} (q_i : a) = \bigcap_{a \not\subseteq q_j} (q_j : a) \). It also commutes with radicals, so \( \text{rad}(I : a) = \bigcap_{b=1}^{k} \text{rad}(q_i : a) = \bigcap_{a \not\subseteq q_j} \text{rad}(q_j : a) \).

We next argue that \( a \not\subseteq q_i \) implies \( \text{rad}(q_i : a) = \text{rad}(q_i) \). Suppose \( b \in \text{rad}(q_i : a) \), i.e. \( b^n a \in q_i \). As \( q_i \) is primary and \( a \not\subseteq q_i \), then \( (b^n)^m \in q_i \), i.e. \( b \in \text{rad}(q_i) \). Hence, \( \text{rad}(q_i : a) \subseteq \text{rad}(q_i) \) and the other inclusion is obvious. At this point we conclude that \( \text{rad}(I : a) \) equals the intersection of the prime ideals \( \text{rad}(q_j) \) satisfying \( a \not\subseteq q_j \).

By Exercise 8, if \( \text{rad}(I : a) \) is prime it has to equal \( \text{rad}(q_j) \) for some \( j \). Conversely, consider any \( \text{rad}(q_{i_0}) \). As the primary decomposition is minimal, there exists \( a \in \bigcap_{j \neq i_0} q_j \setminus q_{i_0} \). The conclusion above shows that \( \text{rad}(I : a) = \text{rad}(q_{i_0}) \).

It remains to prove the last assertion. Clearly, if \( I : a \) is prime, then it is equal to its radical. Thus, we have to consider a prime ideal \( \text{rad}(I : a) \) and show it equals \( I : a' \) for some \( a' \in I \). We already know that \( \text{rad}(I : a) = \text{rad}(q_{i_0}) \) for some \( i_0 \). By Exercise 9, \( \text{rad}(q_{i_0})^n \subseteq q_{i_0} \) for some \( n \). Hence, there exists \( n \) such that \( (\bigcap_{j \neq i_0} q_j) \cdot (\text{rad}(q_{i_0}))^n \subseteq I \). We fix the smallest \( n \) with this property. Then we pick \( a' \in \left( \bigcap_{j \neq i_0} q_j \right) \setminus (\text{rad}(q_{i_0}))^{n-1} \setminus I \).

(Here, if \( n = 1 \) then \( \text{rad}(q_{i_0})^{n-1} \) should be considered as the whole ring.) By definition, \( a' \cdot \text{rad}(q_{i_0}) \subseteq I \), and thus \( \text{rad}(q_{i_0}) \subseteq I : a' \). However, \( a' \in (\bigcap_{j \neq i_0} q_j) \setminus I \), thus \( a' \not\in q_{i_0} \). We have the inclusions \( \text{rad}(q_{i_0}) \subseteq I : a' \subseteq \text{rad}(I : a') = \text{rad}(q_{i_0}) \), which are in fact equalities. The last equation follows from the previous paragraph.

**Definition 3.17.** The associated primes of an ideal \( I \) are the radicals of the primary ideals appearing in a minimal primary decomposition. Equivalently, these are the prime ideals of the form \( \text{rad}(I : a) \) for some element \( a \) of the ring, or in case the ring is Noetherian, these are the prime ideals of the form \( I : a \).
Before going further, let us discuss the geometry behind the associated primes. If \( I = \bigcap_{i=1}^{k} q_i \) is a minimal primary decomposition then \( \text{rad}(I) = \bigcap_{i=1}^{k} \text{rad}(q_i) \). Thus, every component in the irreducible decomposition of the variety \( V(I) \) corresponds to one of the associated primes. However, the converse is not true.

**Example 3.18.** Let \( I = \langle x^2, xy \rangle \) as in Example 3.15. We have \( \text{rad}(I) = \langle x \rangle \), i.e. the variety \( V(I) \) is irreducible - a line in a plane. However, the minimal primary decompositions 3.2 reveal that \( I \) has two associated primes. The expected prime \( \langle x \rangle \) and the unexpected prime \( \langle x, y \rangle \) - a point on the line. Thus, the associated primes remember more information than just the variety associated to the ideal; there is a ‘hidden’ - embedded - point on that line distinguished by the ideal \( I \).

The formal replacement of varieties (corresponding to radical ideals) by arbitrary ideals allowed a tremendous advance of 20th century algebraic geometry. One is now able to work with ‘functions’ that are nonzero, but their square is zero, using basic, well-understood algebra. This advance should be compared to the introduction of complex numbers in 18th and 19th century, where (basically in the same way) instead of answering the question ‘does there exist a square root of \(-1\)’ one introduces imaginary numbers and shows how to use them in an efficient way. Still, we should not forget the classical geometry we started from. The line from Example 3.18 is of a different nature than the point, and these two should be distinguished.

**Definition 3.19.** For an ideal \( I \), let \( \text{Ass}(I) \) be the set of associated primes. The minimal (with respect to inclusion) elements of \( \text{Ass}(I) \) are called the minimal (or isolated) primes. The associated primes that are not minimal are called embedded.

An embedded prime \( p \) of an ideal \( I \) must contain a minimal prime \( p' \). This means that the variety \( V(p') \) contains the variety \( V(p) \). We do no see \( V(p) \) geometrically: it is embedded in \( V(p') \). Further the minimal primes correspond exactly to irreducible components of \( V(I) \). They are the irredundant intersectands in
\[
\text{rad}(I) = \bigcap_{i=1}^{k} \text{rad}(q_i).
\]
The lemma below gives one more explanation for the name for minimal primes.

**Lemma 3.20.** A prime ideal is a minimal (associated) prime of \( I \) if and only if it is a minimal element (with respect to inclusion) among the primes that contain \( I \).

**Proof.** It is enough to prove that every prime \( p \) containing \( I \) contains also a prime in \( \text{Ass}(I) \). Then \( p \) also contains a minimal prime. They are equal if \( p \) is minimal with respect to inclusion. Thus, suppose \( p \) contains \( I = \bigcap_{i=1}^{k} q_i \). By Exercise 8, we have \( p \supseteq q_{i_0} \) for some some \( i_0 \). Hence, \( p = \text{rad}(p) \supset \text{rad}(q_{i_0}) \). □

The geometry that led us to distinguish embedded and minimal associated primes shows us an idea how to get additional uniqueness properties about the primary decomposition. Indeed, in Example 3.15 it is the ideal corresponding to the embedded component that changes, while the minimal prime remains the same.

**Theorem 3.21.** Let \( I = \bigcap_{i=1}^{k} q_i \) be a minimal primary decomposition. The primary ideals \( q_i \) corresponding to minimal primes are uniquely determined by \( I \).

**Proof.** Let \( q_{i_0} \) be such that \( \text{rad}(q_{i_0}) \) is a minimal prime. We claim that
\[
q_{i_0} = \{ a : ab \in I \text{ for some } b \notin \text{rad}(q_{i_0}) \}.
\]
We already proved that the right hand side does not depend on the decomposition of $I$. Thus this equation suffices to establish the theorem. We show both inclusions.

Let $a \in q_{i_0}$. For every $i \neq i_0$ we have $q_i \not\subseteq \text{rad}(q_{i_0})$. Otherwise, $\text{rad}(q_i) \subseteq \text{rad}(q_{i_0})$, which would contradict the hypothesis that $\text{rad}(q_{i_0})$ is minimal. Hence, there exists $b_i \in q_i \setminus \text{rad}(q_{i_0})$. We define $b := \prod_{j \neq i_0} b_j$. As $\text{rad}(q_{i_0})$ is prime we have $b \not\in \text{rad}(q_{i_0})$. However, $ab \in q_j$ for $j \neq i_0$, as $b \in q_j$. Furthermore, $ab \in q_{i_0}$, as $a \in q_{i_0}$. This implies $ab \in I = \bigcap_{i=1}^{k} q_i$. This means that $a$ is in the right hand side.

Now we pick $a$ and $b \not\in \text{rad}(q_{i_0})$ such that $ab \in I$. In particular, $ab \in q_{i_0}$. If $a \not\in q_{i_0}$, we get a contradiction to the fact that $q_{i_0}$ is primary. This shows that the right hand side is contained in the left hand side, and the proof is complete. □

Primary decomposition for monomial ideals is easier than for general polynomial ideals. The associated primes are generated by subsets of the variables and the right hand side is contained in the left hand side, and the proof is complete.

Example 3.22. Let $n = 3$ and $I = \langle xy^2z^3, x^2yz^2, xy^3z^2, x^3yz^2, x^2y^3z, x^3y^2z \rangle$. This has seven associated prime ideals. A minimal primary decomposition equals

$I = \langle x \rangle \cap \langle y \rangle \cap \langle z \rangle \cap \langle x^2, y^2 \rangle \cap \langle x^2, z^2 \rangle \cap \langle y^2, z^2 \rangle \cap \langle x^3, y^3, z^3 \rangle$.

This example generalizes to $n \geq 4$, with $I$ generated by the $n!$ monomials $\prod_{i=1}^{n} x_i^{\pi_i}$, $\pi \in S_n$, and $\text{Ass}(I)$ consisting of $2^n - 1$ ideals generated by proper subsets of $\{x_1, \ldots, x_n\}$.

There are many algorithms and implementation for primary decomposition of polynomial ideals $I$. The output is the set $\text{Ass}(I)$ and primary ideals $q_1, \ldots, q_k$ satisfying (3.1). Traditionally, these are symbolic methods built upon Gröbner bases. In recent years, numerical tools for decomposing ideals and varieties have received much attention. Solving polynomial systems means running such software.

3.3. Linear PDE with Constant Coefficients

In this section we offer an alternative perspective on the problem of solving systems of polynomial equations. This is aimed at highlighting the role of embedded primes and primary ideals in a context of considerable practical importance.

Every polynomial with real or complex coefficients can be interpreted as a linear differential operator with constant coefficients. This operator is obtained by simply replacing $x_i$ by the differential operator $\frac{\partial}{\partial x_i}$. Every ideal $I$ in $\mathbb{R}[x_1, x_2, \ldots, x_n]$ can thus be interpreted as a system of linear partial differential equations (PDE) with constant coefficients. Suppose we are interested in the solutions to these PDE within some nice class of functions, like polynomial functions, real analytic functions $\mathbb{R}^n \to \mathbb{R}$, or complex holomorphic functions $\mathbb{C}^n \to \mathbb{C}$. Then the set of solutions to our PDE is a linear space over $\mathbb{R}$ or $\mathbb{C}$. We are interested in computing a basis for that solutions space. This computation rests on the primary decomposition of the ideal $I$. Both minimal primes and embedded primes will play a role, and all primary components will contribute to our basis for the solution space. But, first of all, let us start by interpreting the usual points of $V(I)$ in terms of PDE.

Lemma 3.23. A point $(a_1, \ldots, a_n) \in \mathbb{C}^n$ lies in $V(I)$ if and only if the function $\exp(a_1x_1 + \cdots + a_nx_n)$ is a solution of the partial differential equations given by $I$.

Proof. Let $f(x) = \exp(a_1x_1 + \cdots + a_nx_n)$. Then $\frac{\partial f}{\partial x_i} = a_i \cdot f$ for $i = 1, \ldots, n$. Let $g$ be any polynomial and $g\left(\frac{\partial}{\partial x_i}\right)$ the corresponding differential operator. By
induction on the degree of \( g \), and using degree one as our base case, we find that the application of the operator \( g(\frac{\partial}{\partial x}) \) to the function \( f(x) \) equals \( g(a_1, \ldots, a_n) \cdot f(x) \). \( \square \)

Lemma 3.23 embeds the classical solutions of a polynomial system into the solution space of the associated linear PDE. But, if the ideal is not radical then there are more solutions, governed by the primary decomposition. We shall explain this for the ideal in Example 3.6 and in Exercise 8 of the first chapter.

**Example 3.24.** Let \( n = 3 \) and \( I = \langle x^3 - yz, y^3 - xz, z^3 - xy \rangle \). The corresponding system of linear PDE asks for all functions \( f = f(x, y, z) \) that satisfy

\[
\frac{\partial^3 f}{\partial x^3} = \frac{\partial^2 f}{\partial y \partial z} \quad \text{and} \quad \frac{\partial^3 f}{\partial y^3} = \frac{\partial^2 f}{\partial x \partial z},
\]

To be more precise, we need to specify the class of functions \( f \) that are allowed. For instance, we might take all holomorphic functions \( f : \mathbb{C}^3 \to \mathbb{C} \). Or we might seek real analytic solutions \( f : \mathbb{R}^3 \to \mathbb{R} \), or, among these, all polynomial solutions.

The degree of our ideal \( I \) is 27 = 3 \times 3 \times 3, which comes from the degrees of the three generators of \( I \). The number 27 is also the dimension of the space of holomorphic solutions \( f \) to (3.3). A basis of that solution space is given by

\[
\begin{align*}
1, & \ x, \ y, \ z, \ x^2, \ y^2, \ z^2, \ x^3 + 6yz, \ y^3 + 6xz, \ z^3 + 6xy, \ x^4 + y^4 + z^4 + 24xyz, \\
& \exp(x - y - z), \ \exp(x + y + z), \ \exp(-x - y + z), \ \exp(-x + y - z), \\
& \exp(x - iy + iz), \ \exp(x + iy - iz), \ \exp(-x - iy - iz), \ \exp(-x + iy + iz), \\
& \exp(ix - y + iz), \ \exp(ix + y - iz), \ \exp(ix - iy + z), \ \exp(ix + iy - z), \\
& \exp(-ix - y - iz), \ \exp(-ix + y + iz), \ \exp(-ix - iy - z), \ \exp(-ix + iy + z).
\end{align*}
\]

The space of polynomial solutions has dimension 11 and is spanned by the first two rows. The space of real analytic solutions has dimension 15 and is spanned by the first two rows. All other basis functions are exponentials of linear forms that have \( i = \sqrt{-1} \) among its coefficients. The 16 basis solutions in the last four rows, along with the solution \( 1 = \exp(0x + 0y + 0z) \), are explained by Lemma 3.23. They are the exponentials corresponding to the 17 distinct points in \( V(I) \subset \mathbb{C}^3 \).

This basis in (3.4) was derived from the primary decomposition of our ideal:

\[
I = Q \cap \bigcap_{a = b = c \equiv 0 \mod 4} \langle x - i^a, y - i^b, z - i^c \rangle \quad \text{in} \ C[x, y, z].
\]

This is obtained by refining the decomposition over \( \mathbb{Q} \) shown in Example 3.6. The 16 ideals in the intersection on the right hand side are maximal and hence prime. They correspond to the 16 exponential solutions in (3.4). The ideal \( Q \) is primary to the maximal ideal \( \text{rad}(Q) = \langle x, y, z \rangle \). Since all associated primes are minimal, by Theorem 3.21, this primary ideal is uniquely determined from \( I \):

\[
Q = \langle x^2y, x^2z, xy^2, x^2z, y^2z, yz^2, x^3 - yz, y^3 - xz, z^3 - xy \rangle.
\]

This ideal has degree 11. It contributes the 11 polynomial solutions to our PDE.

Here is a general result explaining our observations from the previous example.

**Theorem 3.25.** Let \( I \) be a zero-dimensional ideal in \( \mathbb{C}[x_1, \ldots, x_n] \), interpreted as a system of linear PDE. The space of holomorphic solutions has dimension equal to the degree of \( I \). There exist non-zero polynomial solutions if and only if the maximal ideal \( M = \langle x_1, \ldots, x_n \rangle \) is associated to \( I \). In that case, the polynomial solutions are precisely the solutions of the \( M \)-primary component \( I : (I : M^\infty) \).
3. SOLVING AND DECOMPOSING

**Proof.** Fix a degree compatible monomial order and let in($I$) be the initial monomial ideal of $I$. The set $S$ of standard monomials is finite. It consists of monomials of degree $< D$ for some $D$ and these form a basis for the solution space of the partial differential equations associated with in($I$).

Now, for any $d \geq D$, we have the following isomorphism of vector spaces
\[ \mathbb{C}[x]/I \cong \mathbb{C}[x]_{\leq d}/I_{\leq d}. \]

The pairing between monomials in the $\partial/\partial x_i$ and the monomials in the $x_j$ defines a nondegenerate inner product on this vector space. Interpret each element of $I_{\leq d}$ as a differential operator, its solution space in $\mathbb{C}[x]_{\leq d}$ is the orthogonal complement with respect to that inner product. Our Gröbner basis for $I$ translates into a triangular vector space basis for $I_{\leq d}$. By solving that triangular system, we construct a unique solution basis whose elements have the form
\[ p_u(x_1, \ldots, x_n) = x^u + \text{higher order terms}, \quad \text{where } x^u \text{ runs over } S. \]

By increasing the value of $d$, we obtain a formal series (3.5) that solves our PDE. This formal series terminates as a polynomial if and only if it is annihilated by the operators $(\partial/\partial x_i)^d$ for $i = 1, 2, \ldots, n$. This is equivalent to saying that $I$ is $M$-primary. In that case, all of the above solutions $p_u$ are polynomials. The solution space is thus spanned by polynomials and its dimension equals $|S| = \text{degree}(I)$.

Suppose now that $I$ is primary. Since $I$ is zero-dimensional, its radical is the maximal ideal rad($I$) = $\langle x_1 - a_1, \ldots, x_n - a_n \rangle$, where $V(I) = \{(a_1, \ldots, a_n) \in \mathbb{C}^n\} \subset \mathbb{C}^n$. By translating $(a_1, \ldots, a_n)$ to the origin $(0, \ldots, 0)$, we can apply the analysis in the previous paragraph. From this and Lemma 3.23 we obtain degree($I$) many polynomials $p_u$, each with its lowest term a standard monomial $x^u \in S$, such that
\[ p_u(x_1, \ldots, x_n) \cdot \exp(a_1 x_1 + \cdots + a_n x_n) \]
solves the PDE given by $I$. These functions form a basis for the holomorphic solutions to $I$. None of them is a polynomial unless $(a_1, \ldots, a_n) = (0, \ldots, 0)$.

Next, let $I$ be an arbitrary zero-dimensional ideal. Its minimal primary decomposition (3.1) is unique, by Theorem 3.21. The solution space to $I$, regarded as a PDE, is spanned by the solution spaces of its primary components $q_1, q_2, \ldots, q_k$. For each of these, we constructed a basis of holomorphic functions (3.6). The union of these bases is a basis for the solution space of $I$, and its cardinality equals degree($I$).

Finally, suppose that $q_1$ is $M$-primary. Then $(I : M^\infty) = q_2 \cap \cdots \cap q_k$. Taking the ideal quotient of $I$ by $q_2 \cap \cdots \cap q_k$ is equal to $q_1$. This completes the proof. \[ \Box \]

In the preceding discussion, we studied the solutions to zero-dimensional polynomial systems in the guise of linear PDE with constant coefficients. We saw that the solution space of such an ideal $I$ is always a vector space of the same dimension, namely the degree of $I$, independently of whether the ideal $I$ is radical or not. In fact, this solution space, unlike the variety $V(I)$, varies gracefully under parameter changes. This underscores the utility with primary decompositions in the context of solving equations. We demonstrate this perspective in a simple example.

**Example 3.26** (n=2). Consider the ideal $I = \langle x^2 - \delta^2, y^2 - \epsilon^2 \rangle \subset \mathbb{R}[x, y]$, where $\delta, \epsilon$ are small real parameters, viewed as PDE. For $\delta, \epsilon \neq 0$, the solution space is spanned by the four functions $f_{ij} := \exp((-1)^i \delta x + (-1)^j \epsilon y)$, where $i, j \in \{0, 1\}$. However these four functions become linearly dependent when $\delta \epsilon = 0$. We therefore...
change the basis of our four-dimensional solution space as follows:

\[
\begin{align*}
g_{00} &= \frac{1}{4}(f_{00} + f_{01} + f_{10} + f_{11}) = 1 + \frac{\delta^2}{2} x^2 + \frac{\epsilon^2}{2} y^2 + \cdots \\
g_{01} &= \frac{1}{4}(f_{00} - f_{01} + f_{10} - f_{11}) = y + \frac{\delta^2}{2} x y + \frac{\epsilon^2}{2} y^3 + \cdots \\
g_{10} &= \frac{1}{4}(f_{00} + f_{01} - f_{10} - f_{11}) = x + \frac{\delta^2}{6} x^3 + \frac{\epsilon^2}{2} x y^2 + \cdots \\
g_{11} &= \frac{1}{4}(f_{00} - f_{01} - f_{10} + f_{11}) = x y + \frac{\delta^2}{6} x^3 y + \frac{\epsilon^2}{2} x y^3 + \cdots \\
\end{align*}
\]

This family remains linearly independent for all values of \( \delta \) and \( \epsilon \). In particular, for \( \delta = \epsilon = 0 \), we obtain the standard basis \( S = \{1, x, y, xy\} \) modulo the ideal \( \langle x^2, y^2 \rangle \).

We next explore what happens for polynomial ideals \( I \) that are not zero-dimensional. It is still true that the primary decomposition of \( I \) reveals the solution space of the associated PDE. The precise statement is an important result in analysis known as Ehrenpreis' Fundamental Principle or as Palamodov-Ehrenpreis Theorem. The details of this theorem are outside our scope. We will not state it here. Instead, we simply illustrate the role of primary decomposition in an example.

One observation is that embedded primes reveal spurious solutions spaces.

**Example 3.27.** Let \( n = 4 \) and consider the ideal

\[ J = \langle xw, xz + yw, yz \rangle. \]

Somewhat surprisingly, this ideal is not radical. Its radical is the monomial ideal

\[ \sqrt{J} = \langle x, y \rangle \cap \langle z, w \rangle = \langle xw, xz, yw, yz \rangle. \]

The given ideal \( J \) has three associated primes. The primes \( \langle x, y \rangle \) and \( \langle z, w \rangle \) are minimal primes, and the maximal ideal \( \langle x, y, z, w \rangle \) is an embedded prime. A minimal primary decomposition is given by

\[ J = \langle x, y \rangle \cap \langle z, w \rangle \cap (J + \langle x, y, z, w \rangle^3). \]

The third primary component is embedded. It is not unique. We can replace the third power of the maximal ideal by any higher power and get the same intersection.

As before, we interpret the generators of \( J \) as a system of linear PDE:

\[
\frac{\partial^2 f}{\partial x \partial w} = \frac{\partial^2 f}{\partial x \partial z} + \frac{\partial^2 f}{\partial y \partial w} = \frac{\partial^2 f}{\partial y \partial z} = 0.
\]

The linear space of solutions \( f(x, y, z, w) \) is infinite-dimensional. It is spanned by all functions of the form \( g(y, z) \) and \( h(x, y) \), together with the one special function \( xz - yw \). The former correspond to the two minimal primes, whereas the latter spurious solution arises from the embedded primary component.

Whenever one encounters a system of polynomial equations with special structure, and one is curious about the variety of solutions, it pays to explore the primary decomposition. This decomposition often reveals interesting structures, and it tells us how to break up the solutions into meaningful pieces. As an illustration consider the following question from the repertoire of linear algebra: *Let \( A, B, C \) be \( 2 \times 2 \)-matrices. How is it possible that the triple product \( ABC \) is the zero matrix?*

We approach this problem as follows. We set \( n = 12 \) and we consider the polynomial ring \( \mathbb{R}[a_{ij}, b_{ij}, c_{ij}] \) whose variables are the 12 entries of the matrices \( A, B, C \). Let \( K \) be the ideal of \( \mathbb{R}[a_{ij}, b_{ij}, c_{ij}] \) that is generated by the four entries of the matrix product \( ABC \). For example, one generator of \( K \) is the trilinear form

\[
a_{11}b_{11}c_{11} + a_{12}b_{21}c_{11} + a_{11}b_{12}c_{21} + a_{12}b_{22}c_{21}.
\]
As always, in the back of our minds, we think of this as a differential equation:

\[
\frac{\partial^3 f}{\partial a_{11}\partial b_{11}\partial c_{11}} + \frac{\partial^3 f}{\partial a_{12}\partial b_{21}\partial c_{12}} + \frac{\partial^3 f}{\partial a_{11}\partial b_{12}\partial c_{21}} + \frac{\partial^3 f}{\partial a_{12}\partial b_{22}\partial c_{21}} = 0.
\]

Which functions on matrix triples satisfy these four partial differential equations?

A computation reveals that \( K \) is radical. It is the intersection of six prime ideals. Three of them are ideals generated by the entries of \( A \) or \( B \) or \( C \) respectively. The next two primes are generated respectively by the 2 \( \times \) 2 minors of the matrices

\[
\begin{pmatrix}
a_{11} & a_{21} \\
a_{12} & a_{22}
\end{pmatrix}
\quad \text{and} \quad
\begin{pmatrix}
b_{11} & b_{21} & -c_{21} & -c_{22} \\
b_{12} & b_{22} & c_{11} & c_{12}
\end{pmatrix}.
\]

Finally, the last associated prime of \( K \) is the ideal \( K + \langle \det(A), \det(C) \rangle \). Thus \( \text{Ass}(K) \) consists of six primes, and all are minimal. One checks with a computer algebra system that our ideal \( K \) is the intersection of these six minimal primes.

Geometrically, we have studied the variety \( V(K) \) which is defined by four cubic equations and which lives in \( \mathbb{C}^{12} \). It is the union of six irreducible components. Three of them are linear spaces of dimension 8. The other three irreducible components have dimension 9 and they are not linear spaces. Their degrees are 4, 4 and 8 respectively. In terms of the original linear algebra question, the six irreducible components correspond to the following six scenarios for a triple of \( 2 \times 2 \)-matrices:

\[
\begin{align*}
\text{rank}(A) &= 0 \quad \text{or} \quad \text{rank}(B) = 0 \quad \text{or} \quad \text{rank}(C) = 0 \quad \text{or} \\
\text{rank}(A) &= \text{rank}(B) = 1 \quad \text{or} \quad \text{rank}(B) = \text{rank}(C) = 1 \quad \text{or} \quad \text{rank}(A) = \text{rank}(C) = 1.
\end{align*}
\]

Each of the six irreducible components is a rational variety, and admits a nice exponential parametrization. Using Lemma 3.23, we can then write down all exponential solutions to the four partial differential equations given by \( K \), like (3.7). The solutions contributed by the first irreducible component are the functions \( f(B, C) \). The solutions contributed by the last irreducible component have the form

\[
f(A, B, C) = \exp \left[ r_1 s_1 a_{11} + r_1 s_2 a_{12} + r_2 s_1 a_{21} + r_2 s_2 a_{22} + (t_{11} u_{2} - s_2 t_{12}) b_{11}
\right.
\]

\[
+ (s_2 t_{21} - t_{11} u_1) b_{12} + (s_1 t_{12} - t_{22} u_2) b_{21} + (t_{22} u_1 - s_1 t_{21}) b_{22}
\]

\[
+ u_1 v_1 c_{11} + u_1 v_2 c_{12} + u_2 v_1 c_{21} + u_2 v_2 c_{22} \],
\]

where \( r_1, s_j, t_{ij}, u_i, v_j \) are arbitrary complex constants. We invite the reader to verify that these functions satisfy the PDE. A conclusion, valid for this entire book: taking a fresh look at linear algebra offers a point of entry to nonlinear algebra.

### Exercises

1. Consider the ring \( \mathbb{C}[x, y]/(x^2, xy, y^2) \). Is \( \{0\} \) an irreducible ideal? Is it primary?
   a) Prove that the intersection of prime ideals is radical.
   b) Prove the opposite implication. Hint: Apply Kuratowski-Zorn lemma.

2. Prove that an ideal \( I \subseteq Z \) is a power of a prime ideal if and only if it is primary.

3. Prove that a ring is Noetherian if and only if every ideal is finitely generated.

4. a) Prove that if \( R \) is Noetherian, then so is \( R/I \) for any ideal \( I \).
   b) Prove Hilbert Basis Theorem: If \( R \) is Noetherian, then so is \( R[x] \).

5. Prove Lemma 3.12

6. Check that Example 3.15 provides two distinct minimal primary decompositions.

7. a) Prove that a prime ideal \( p \) cannot be equal to an intersection of (finitely many, more than one, incomparable) ideals.
b) More generally prove that if a prime ideal contains an intersection of finitely many ideals, then it contains one of them.

(8) Prove that in a Noetherian ring every ideal contains a power of its radical. Give a counterexample in case of a non-Noetherian ring.

(9) Find three polynomials in three unknowns, each having degree precisely five, whose variety in $\mathbb{C}^3$ consists of precisely 37 complex solutions.

(10) Determine all solutions $(x, y)$ of the two equations $x^2 + y = \epsilon$ and $y^2 + x = \epsilon$ over the algebraic closure of the field $\mathbb{Q}(\epsilon)$. Write explicit Puiseux series.

(11) Let $I$ be the ideal generated by the $2 \times 2$-subpermanents $x_iy_j + x_jx_i$ of a $2 \times 5$-matrix of unknowns. Determine a minimal primary decomposition of $I$, and interpret your result in terms of solving linear partial differential equations.
CHAPTER 4

Mapping and Projecting

A frequently encountered challenge is to compute the image of a polynomial map. Such an image need not be an algebraic variety. However, it can be closely approximated by its Zariski closure. The Zariski closure of the image is a variety, described by the polynomials that vanish on it. In this chapter we show how this variety can be found by eliminating variables. Gröbner bases and resultants serve as our primary tools. Further, we provide theorems that allow to understand the difference between the image and its closure. The answer we obtain depends heavily on the setting, where we work over the complex numbers \( \mathbb{C} \) or over the real numbers \( \mathbb{R} \), and whether the given polynomials are homogeneous or nonhomogeneous.

4.1. Elimination

In this section we introduce the main algebraic tool to compute the closure of the image of a map: elimination of variables. We show how to carry it out efficiently using Gröbner basis.

We fix an algebraically closed field \( K \) and the polynomial ring \( K[x] = K[x_1, \ldots, x_n] \).

Every ideal \( I \subseteq K[x] \) has an associated variety in \( n \)-dimensional affine space:

\[
V(I) = \{ p \in K^n : f(p) = 0 \text{ for all } f \in I \}.
\]

Consider the projection from \( K^n \) onto the subspace given by the first \( m \) coordinates:

\[
\pi : K^n \to K^m, (p_1, \ldots, p_m, p_{m+1}, \ldots, p_n) \mapsto (p_1, \ldots, p_m).
\]

If \( V \) is a variety in \( K^m \) then its image \( \pi(V) \) need not be a variety.

Example 4.1 \((n = 2, m = 1)\). The image of the hyperbola \( V = V(xy-1) \) under the projection \( K^2 \to K^1 \) from the plane to the \( x \)-axis equals \( \pi(V) = K^1 \setminus \{0\} \). This is not a variety in \( K^1 \). Note that the image will be closed if we first perform a change of coordinates, e.g. if we replace \( V \) by the hyperbola \( V' = V((x+y)(x-y)-1) \).

By definition, the Zariski closure \( \overline{\pi(V)} \) of the image \( \pi(V) \) is a variety in \( K^m \). It is the smallest variety containing \( \pi(V) \). We call the variety \( \pi(V) \) the closed image of \( V \) under the map \( \pi \). The following theorem characterizes its ideal.

Theorem 4.2. Let \( V = V(I) \) be the variety given by an ideal \( I \subseteq K[x] \). Then its closed image in \( K^m \) is the variety \( \overline{\pi(V)} = V(J) \) defined by the elimination ideal

\[
J = I \cap K[x_1, \ldots, x_m].
\]  

If the ideal \( I \) is radical or prime then the elimination ideal \( J \) has the same property.

Proof. If \( J \) is not a prime ideal then there exist polynomials \( f \) and \( g \) in \( K[x_1, \ldots, x_m] \) such that \( fg \in J \) but \( f, g \notin J \). The same polynomials show that \( I \) is not prime. Similarly if \( J \) is not radical then there exists \( f \in K[x_1, \ldots, x_m] \) and
r ≥ 2 such that f∗ ∈ J but f ∉ J. The same f shows that J is not radical. A similar reasoning implies that arbitrary ideals I ⊂ K[x] satisfy
\[
\text{Rad}(I) \cap K[x_1, \ldots, x_m] = \text{Rad}(I \cap K[x_1, \ldots, x_m]).
\]
Since passing to the radical does not change the variety of a given ideal, we may assume that I and J are radical ideals. We shall now make a forward reference and use the Nullstellensatz (Chapter [1]). A polynomial belongs to I if and only if it vanishes on \( V = \mathcal{V}(I) \). This holds, in particular, for polynomials \( f \) in \( K[x_1, \ldots, x_m] \). Such an \( f \) belongs to \( J \) if and only if it vanishes on \( \pi(V) \) if and only if it vanishes on \( \pi(V) \). The latter condition means that \( f \) lies in the radical ideal of \( \pi(V) \).

We conclude that the the radical ideal of the closed image \( \pi(V) \) is precisely the elimination ideal \( J \). For further details we refer to [7], §2.2, Theorem 3. □

Theorem 4.2 says that the algebraic operation of elimination corresponds to the geometric operation of projection. This holds in many settings, not just in algebraic geometry. For instance, Gaussian elimination in linear algebra corresponds to projection of linear subspaces, and Fourier-Motzkin elimination in convex geometry corresponds to projection of polyhedra. Alternatively, from the perspective of logic, we can think of our projection as quantifier elimination. We are eliminating the \( n - m \) existentially quantifiable variables from the statement \( \exists x_{m+1}, \ldots, x_n : x \in V \).

**Example 4.3 (Matrix Completion).** Fix \( n = 15 \) and let \( V \) be the irreducible variety of symmetric \( 5 \times 5 \)-matrices \( X = (x_{ij}) \) of rank \( \leq 2 \). Its prime ideal \( I = \mathcal{I}(V) \) is minimally generated by 50 homogeneous cubic polynomials, namely the 3 \( \times \) 3 minors of \( X \). These cubics form a Gröbner basis for the degree reverse lexicographic order. Now let \( m = 10 \) and order the variables so that the five diagonal entries \( x_{11}, x_{22}, x_{33}, x_{44}, x_{55} \) come last. Then the elimination ideal is principal:
\[
J = (x_{14}x_{15}x_{23}x_{25}x_{34} - x_{13}x_{15}x_{24}x_{25}x_{34} - x_{14}x_{15}x_{23}x_{24}x_{35} + x_{13}x_{14}x_{24}x_{25}x_{35} + x_{12}x_{15}x_{24}x_{34}x_{35} - x_{12}x_{14}x_{25}x_{34}x_{35} + x_{13}x_{15}x_{23}x_{24}x_{45} - x_{13}x_{14}x_{23}x_{25}x_{45} - x_{12}x_{15}x_{23}x_{34}x_{45} + x_{12}x_{13}x_{25}x_{34}x_{45} + x_{12}x_{14}x_{25}x_{35}x_{45} - x_{12}x_{13}x_{24}x_{35}x_{45} \).
\]
The ideal generator is known as the *pentad* in algebraic statistics [12], Example 4.2.8. The 15 terms correspond to the 15 maximal matchings in the complete graph \( K_5 \). The hypersurface \( \mathcal{V}(J) \) equals the image \( \pi(V) \) of the determinantal variety \( V \) under the projection onto the \( K^{10} \) given by the off-diagonal entries.

If the 10 off-diagonal entries of a symmetric \( 5 \times 5 \)-matrix are given then that matrix can be completed to a matrix of rank \( \leq 2 \) if and only if the pentad vanishes. This constraint appears in the statistical theory of *factor analysis* [12]. It represents a widely studied class of problems known as (low rank) matrix completion.

**Example 4.4.** The first four power sums in three variables are \( x^i + y^i + z^i \) for \( i = 1, 2, 3, 4 \). These four must be algebraically dependent since they involve only three variables. But, what is the algebraic relation satisfied by these power sums? We approach this question by setting \( n = 7, m = 4 \) and introducing the ideal
\[
I = (x + y + z - p_1, x^2 + y^2 + z^2 - p_2, x^3 + y^3 + z^3 - p_3, x^4 + y^4 + z^4 - p_4).
\]
This ideal lives in a polynomial ring in seven variables. We wish to eliminate the three original variables \( x, y, z \). Thus, we are asking for the elimination ideal
\[
J = I \cap K[p_1, p_2, p_3, p_4].
\]
This is a principal prime ideal. Its generator is a polynomial of degree four:

\[ J = \langle p_1^4 - 6p_1^2p_2 + 3p_2^2 + 8p_1p_3 - 6p_1 \rangle. \]

This is the desired relation. Please check by plugging in the power sums.

The computations in our two examples were carried out using Gröbner bases. Here is how this works. We first fix the lexicographic monomial order \( \prec \) on \( K[x] \) with \( x_1 \prec x_2 \prec \cdots \prec x_n \). We then compute the reduced Gröbner basis for the ideal generated by the given polynomials. And, finally, we select those polynomials from the output that use only the first \( m \) variables.

**Theorem 4.5.** If \( G \) is a lexicographic Gröbner basis for an ideal \( I \) in \( K[x] \) then its elimination ideal \( J \) in \([4.7]\) has the Gröbner basis \( G' = G \cap K[x_1, \ldots, x_m] \). If \( G \) is the reduced Gröbner basis of \( I \) then \( G' \) is the reduced Gröbner basis of \( J \).

**Proof.** Clearly, \( G' \) is contained in \( J = I \cap K[x_1, \ldots, x_m] \). Consider any nonzero polynomial \( f \in J \). The initial monomial \( \in_\prec(f) \) is divisible by \( \in_\prec(g) \) for some \( g \in G \). None of the variables \( x_{m+1}, \ldots, x_n \) appears in the monomial \( \in_\prec(g) \). Every trailing term of \( g \) is lexicographically smaller, so it cannot use any of the last \( n - m \) variables. Hence \( g \) lies in \( G' \). We have shown that some initial monomial from \( G' \) divides \( \in_\prec(f) \). Since \( f \) was chosen arbitrarily from \( J \setminus \{0\} \), this means that \( G' \) is a Gröbner basis for \( J \). If the given Gröbner basis \( G \) is reduced then \( G' \) also satisfies the two requirements for being a reduced Gröbner basis.

This result shows that the lexicographic Gröbner basis \( G \) solves the elimination problem simultaneously for all \( m \). Thus computing \( G \) means triangularizing a given system of polynomial equations. We saw in Chapter 4 that it can be quite costly to compute a lexicographic Gröbner basis. One therefore often uses different strategies to carry out the elimination process. But Theorem 4.5 represents the main idea that underlies these strategies. Lexicographic elimination is a key tool for solving systems of polynomial equations. The Gröbner basis in Theorem 4.5 triangularizes the given system. It is instructive to try this for some zero-dimensional varieties.

**Example 4.6.** Here is a simple question: can you find three real numbers \( x, y, z \) whose \( i \)-th power sum equals \( i \) for \( i = 1, 2, 3 \)? To answer this question, we compute the lexicographic Gröbner basis of the ideal that models the question:

\[ I = \langle x + y + z - 1, x^2 + y^2 + z^2 - 2, x^3 + y^3 + z^3 - 3 \rangle. \]

This Gröbner basis equals

\[ G = \{ 6z^3 - 6z^2 - 3z - 1, 2y^2 + 2yz - 2y + 2z^2 - 2z - 1, x + y + z - 1 \}. \]

Theorem 4.2 says that we can solve our equations by back-substitution. Indeed, the equations have six complex zeros. We first compute the three roots of the cubic in \( z \), we substitute them into the second equation and solve for \( y \), and then we set \( x = 1 - y - z \). The cubic has one real root and two complex conjugate roots:

\[ z \in \{ 1.4308, -0.21542 - 0.26471i, -0.21542 + 0.26471i \}. \]

By symmetry, the zeros of \( I \) are precisely the six points in \( \mathbb{C}^3 \) whose coordinates are permutations of the three complex numbers above. Hence the answer to our question is “no”. The variety \( V(I) \) has no real points.
Implicitization is a special instance of elimination. Here, the problem is to compute the image of a polynomial map between two affine spaces. This can be done by forming the graph of the map and then projecting onto the image coordinates. To be precise, we consider a map of the form

\[ f : K^m \to K^n, \quad p = (p_1, \ldots, p_m) \mapsto (f_1(p), \ldots, f_n(p)), \]

where \( f_1, \ldots, f_n \) are polynomials in \( K[z_1, \ldots, z_m] \). We write \( \text{image}(f) \) for the image of \( K^m \) under this map. This need not be a variety, as the following example shows:

**Example 4.7.** Let \( m = 2, n = 3 \) and consider the map given by \((z_1, z_1^2, z_1 z_2^2)\). The Zariski closure of the image is the surface \( V = \mathcal{V}(x_1 x_3 - x_2^2) \) in \( K^3 \). The point \((0, 0, 1)\) is in the surface but not in \( \text{image}(f) \). For \( K = \mathbb{C} \) we approximate \((0, 0, 1)\) by a sequence of points in the image, e.g. by taking \( z_1 = \epsilon^2 \) and \( z_2 = \epsilon^{-1} \) for \( \epsilon \to 0 \).

Recall that the closed image of the map \( f : K^m \to K^n \) is the Zariski closure of the set-theoretic image \( \text{image}(f) \). The closed image is denoted \( \overline{\text{image}(f)} \).

**Corollary 4.8.** Let \( I \) be the ideal in the polynomial ring \( K[x, z] \) in \( n + m \) variables which is generated by \( f_i(z_1, \ldots, z_m) - x_i \) for \( i = 1, 2, \ldots, n \). The closed image of \( f : K^m \to K^n \) is the variety defined the elimination ideal \( \overline{J} = I \cap K[x] \).

In symbols, \( \overline{\text{image}(f)} = \mathcal{V}(J) \).

**Proof.** The graph of \( f \) is Zariski closed in \( K^{n+m} \), and \( I \) is the ideal that defines it. The image of \( f \) is the projection of the graph onto \( K^n \). With this, the claim follows from Theorem 4.2.

**Example 4.9** (Plücker relations). What are the algebraic relations among the 2 \( \times \) 2-minors of a 2 \( \times \) 5-matrix? We answer this question by setting \( m = n = 10 \) and considering the map \( f : K^{10} \to K^{10} \) that takes a matrix

\[
\begin{pmatrix}
z_{11} & z_{12} & z_{13} & z_{14} & z_{15} \\
z_{21} & z_{22} & z_{23} & z_{24} & z_{25}
\end{pmatrix}
\]

to the vector \((x_{12}, x_{13}, \ldots, x_{45})\) where \( x_{ij} = z_{i1} z_{j2} - z_{i2} z_{j1} \) for \( 1 \leq i < j \leq 5 \). The graph of \( f \) is described by an ideal \( I \) in the polynomial ring \( K[x, z] \) in 20 variables. Note that \( I \) is generated by 10 polynomials. The desired elimination ideal equals

\[
I \cap K[x] = \{ x_{12} x_{34} + x_{13} x_{24} + x_{14} x_{23}, x_{12} x_{35} + x_{13} x_{25} + x_{15} x_{23}, \\
x_{12} x_{45} - x_{14} x_{25} + x_{15} x_{24}, x_{13} x_{45} - x_{14} x_{35} + x_{15} x_{34}, \\
x_{23} x_{45} - x_{24} x_{35} + x_{25} x_{34} \}.
\]

These five quadrics are the Plücker relations among the maximal minors. They will play a key role in our study of Grassmannians in the next chapter. The ten variables in \( K[x] \) can be written as the entries of a skew-symmetric 5 \( \times \) 5 matrix

\[
X = \begin{pmatrix}
0 & x_{12} & x_{13} & x_{14} & x_{15} \\
-x_{12} & 0 & x_{23} & x_{24} & x_{25} \\
-x_{13} & -x_{23} & 0 & x_{34} & x_{35} \\
-x_{14} & -x_{24} & -x_{34} & 0 & x_{45} \\
-x_{15} & -x_{25} & -x_{35} & -x_{45} & 0
\end{pmatrix}.
\]

The Plücker relations are the *pfaffians* of size 4 \( \times \) 4, that is, the square roots of the principal 4 \( \times \) 4 minors of \( X \). Thus \( \mathcal{V}(I \cap K[x]) \) is the variety of skew-symmetric 5 \( \times \) 5 matrices of rank \( \leq 2 \). We shall see that, as a projective variety in \( \mathbb{P}^9 \), this is the Grassmannian of lines in \( \mathbb{P}^4 \). Each such line is the image of the rank 2 matrix \( X \).
4.2. IMPLICITIZATION 49

Example 4.10 (Hyperdeterminant). Let \( X = (x_{ij}) \) be a tensor of format \( 2 \times 2 \times 2 \), where the \( n = 8 \) tensor entries are variables. The tensor represents an affine-trilinear polynomial in \( m = 3 \) variables:

\[
f = x_{000} + x_{100}z_1 + x_{010}z_2 + x_{001}z_3 + x_{110}z_1z_2 + x_{101}z_1z_3 + x_{011}z_2z_3 + x_{111}z_1z_2z_3.
\]

For any fixed \( X \), this polynomial defines a surface \( \mathcal{V}(f) \) in \( K^3 \). The surface is singular at the point \( z \) if and only if the pair \( (X, z) \in K^3 \) lies in the variety of

\[
I = \langle f, \frac{\partial f}{\partial z_1}, \frac{\partial f}{\partial z_2}, \frac{\partial f}{\partial z_3} \rangle.
\]

The elimination ideal \( I \cap K[x] \) is principal. We find that its generator equals

\[
\begin{align*}
x_{210}x_{001}^2 + x_{010}x_{201}^2 + x_{000}x_{110}^2 + x_{011}x_{100}^2 + 4x_{000}x_{100}x_{011}x_{110} & + 4x_{010}x_{100}x_{001}x_{111} \\
-2x_{000}x_{100}x_{011}x_{110} & - 2x_{010}x_{100}x_{001}x_{110} - 2x_{010}x_{100}x_{011}x_{101} & - 2x_{000}x_{110}x_{011}x_{110} & - 2x_{000}x_{110}x_{011}x_{111} & - 2x_{000}x_{100}x_{111}x_{101} & - 2x_{010}x_{110}x_{011}x_{111}.
\end{align*}
\]

This quartic is known as the \( 2 \times 2 \times 2 \) hyperdeterminant. It vanishes whenever the surface \( \mathcal{V}(f) \) fails to be smooth in \( K^3 \). Hyperdeterminants exist for larger tensors of many other formats. Their study is a fascinating topic in nonlinear algebra.

The most basic scenario in elimination arises when \( m \) variables are eliminated from a system of \( m + 1 \) polynomial equations. In this case one expects the result to be a single polynomial in the coefficients of that system. We saw this for \( m = 3 \) in Examples 4.4 and 4.10. The theory of resultants is custom-tailored to predict the elimnant in such scenarios. We set this up over the field \( \mathbb{Q} \) as follows.

Let \( i \in \{1, 2, \ldots, m+1\} \) and fix a general inhomogeneous polynomial \( f_i \) of degree \( d_i \) in \( z_1, \ldots, z_m \). This polynomial has \( \binom{d_i+m}{m} \) unknown coefficients \( x_{i,u} \), one for each monomial \( z^u \) of degree \( \leq d_i \). The total number of unknown coefficients equals \( n = \sum_{i=1}^{m+1} \binom{d_i+m}{m} \). We write \( \mathbb{Q}[x,z] \) for the resulting polynomial ring in \( n + m \) variables. Inside this ring we consider the ideal

\[
I = \langle f_1, f_2, \ldots, f_m, f_{m+1} \rangle \subset \mathbb{Q}[x,z].
\]

We are interested in the ideal in \( \mathbb{Q}[x] \) obtained by eliminating the \( m \) variables \( z_i \).

Theorem 4.11. The elimination ideal \( I \cap \mathbb{Q}[x] \) is principal. Its generator is an irreducible polynomial in the entries of the coefficient vector \( z \). This generator is denoted \( \text{Res}(f_1, \ldots, f_{m+1}) \) and called the resultant. The degree of the resultant in the coefficients of \( f_i \) equals \( d_1 \cdots d_{i-1}d_{i+1} \cdots d_{m+1} \) for \( i = 1, 2, \ldots, m+1 \).

Proof. We refer to [8, Chapter 3] for the proof. In that source, and many others, the \( f_i \) are taken to be homogeneous polynomials in \( m + 1 \) variables. We here prefer the inhomogeneous case, which allows for a simpler formulation as an elimination ideal. The two versions are equivalent. \( \square \)

Example 4.12 (Determinants). Let \( d_1 = d_2 = \cdots = d_{m+1} = 1 \). The \( m + 1 \) polynomials \( f_i \) are affine-linear. They can be expressed in the matrix-vector product

\[
\begin{pmatrix}
f_1 \\
f_2 \\
\vdots \\
f_m \\
f_{m+1}
\end{pmatrix} =
\begin{pmatrix}
x_{1,1} & x_{1,2} & \cdots & x_{1,m} & x_{1,m+1} \\
x_{2,1} & x_{2,2} & \cdots & x_{2,m} & x_{2,m+1} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
x_{m,1} & x_{m,2} & \cdots & x_{m,m} & x_{m,m+1} \\
x_{m+1,1} & x_{m+1,2} & \cdots & x_{m+1,m} & x_{m+1,m+1}
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
\vdots \\
x_m \\
1
\end{pmatrix}
\]
The resultant \( \det(f_1, \ldots, f_{m+1}) \) is the determinant of the coefficient matrix \((x_{i,j})\).
This is a homogeneous polynomial of degree \(m+1\) in \(n = (m+1)^2\) unknowns having \((m+1)!\) terms. As predicted, it has degree one in the coefficients of each \(f_i\).

**Example 4.13 (Eliminating one variable from two quadratic polynomials).** Let \(m = 1\) and \(d_1 = d_2 = 2\) and abbreviate \(z = z_1\). Then our system consists of two univariate polynomials of degree two, with six coefficients that are parameters:

\[
   f_1 = x_{11}z^2 + x_{12}z + x_{13} \quad \text{and} \quad f_2 = x_{21}z^2 + x_{22}z + x_{23}.
\]

The generator of the elimination ideal \(\langle f_1, f_2 \rangle \cap \mathbb{Q}[x] \) is the Sylvester resultant

\[
   \text{Res}(f_1, f_2) = \det \begin{pmatrix}
      x_{11} & x_{12} & x_{13} & 0 \\
      0 & x_{11} & x_{12} & x_{13} \\
      x_{21} & x_{22} & x_{23} & 0 \\
      0 & x_{21} & x_{22} & x_{23}
   \end{pmatrix}.
\]

This is a bi-homogeneous polynomial of bidegree \((d_1, d_2) = (2, 2)\). Its expansion has 7 terms. It vanishes precisely when the two quadrics have a common zero.

The formula \((4.2)\) generalizes to two polynomials in \(z\) of arbitrary degrees \(d_1, d_2\).

The following is the Sylvester matrix of format \((d_2+d_1) \times (d_2+d_1)\):

\[
   \text{Syl}_{d_1, d_2} = \begin{pmatrix}
      x_{11} & x_{12} & \cdots & x_{1:d_1+1} & 0 & \cdots & 0 & 0 \\
      0 & x_{11} & x_{12} & \cdots & x_{1:d_1+1} & 0 & \cdots & 0 \\
      \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
      0 & \cdots & 0 & x_{11} & x_{12} & \cdots & x_{1:d_1+1} & 0 \\
      0 & 0 & \cdots & 0 & x_{11} & x_{12} & \cdots & x_{1:d_1+1} \\
      x_{21} & x_{22} & \cdots & x_{2:d_2+1} & 0 & \cdots & 0 & 0 \\
      0 & x_{21} & x_{22} & \cdots & x_{2:d_2+1} & 0 & \cdots & 0 \\
      \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
      0 & \cdots & 0 & x_{21} & x_{22} & \cdots & x_{2:d_2+1} & 0 \\
      0 & 0 & \cdots & 0 & x_{21} & x_{22} & \cdots & x_{2:d_2+1}
   \end{pmatrix}
\]

For \(d_1 = d_2 = 2\) this is the \(4 \times 4\) matrix seen in \((4.2)\).

**Theorem 4.14.** The determinant of the Sylvester matrix \(\text{Syl}_{d_1, d_2}\) is equal to the resultant \(\text{Res}(f_1, f_2)\) of the two univariate polynomials

\[
   f_1(z) = x_{11}z^{d_1} + \cdots + x_{1:d_1}z + x_{1:d_1+1} \quad \text{and} \quad f_2(z) = x_{21}z^{d_2} + \cdots + x_{2:d_2}z + x_{2:d_2+1}.
\]

**Proof.** We first note that the determinant \(\det(\text{Syl}_{d_1, d_2})\) is a non-zero polynomial. We can see this by specializing \(f_1 = z^{d_1}\) and \(f_2 = 1\). Here the Sylvester matrix \(\text{Syl}_{d_1, d_2}\) specializes to the identity matrix, so its determinant is non-zero.

Let \(Z\) denote the column vector with entries \(z^{d_1+d_2-1}, z^{d_1+d_2-2}, \ldots, z^2, z, 1\), and let \(F\) denote the column vector with entries \(z^{d_2-1}f_1, \ldots, zf_1, f_1, z^{d_1-1}f_2, \ldots, zf_2, f_2\). Both vectors have length \(d_1 + d_2\). They are related by the Sylvester matrix:

\[
   \text{Syl}_{d_1, d_2} \cdot Z = F.
\]

Multiplying this equation on the left by the adjoint of the Sylvester matrix, we get

\[
   \det(\text{Syl}_{d_1, d_2}) \cdot Z = \text{adj}(\text{Syl}_{d_1, d_2}) \cdot F.
\]
The last coordinate of the column vector $Z$ equals 1. Hence the last coordinate in this equation shows that $\det(\text{Syl}_{d_1,d_2})$ is a polynomial linear combination of the entries of $F$, and hence it lies in the ideal $(f_1,f_2)$. The Sylvester determinant is a non-zero homogeneous polynomial of degree $d_1 + d_2$ that lies in the ideal $(f_1,f_2) \cap \mathbb{Q}[x]$. We know from Theorem 4.11 that this ideal is principal, and its generator $\text{Res}(f_1,f_2)$ also has degree $d_1 + d_2$. This implies that $\text{Res}(f_1,f_2)$ is equal to the Sylvester determinant $\det(\text{Syl}_{d_1,d_2})$, up to a non-zero multiplicative constant. □

Example 4.15. Let $f_1(z)$ and $f_2(z)$ be univariate polynomials of degree $d_1$ and $d_2$ in $\mathbb{Q}[z]$. This defines a map $f : \mathbb{C} \to \mathbb{C}^2$ whose closed image is an algebraic curve in the plane $\mathbb{C}^2$ with coordinates $x_1,x_2$. The implicit equation of this curve is the resultant $\text{Res}_z(x_1 - f(z),x_2 - g(z))$, taken with respect to the variable $z$.

If $m \geq 2$ then the resultant $\text{Res}(f_1,f_2,\ldots,f_{m+1})$ is more difficult to compute, and there does not always exists a formula as the determinant whose entries are linear expressions in the coefficients of $f_1,f_2,\ldots,f_{m+1}$. In some cases, however, such formulas are available in the literature. For instance, Sylvester already gave such a formula for $m = 2$ and $d_1 = d_2 = d_3$. A considerable body of information on such formulas can be found in the book by Gel’fand, Kapranov and Zelevinsky.

4.3. The Image of a Polynomial Map

We have discussed methods for computing the Zariski closure of the image of a polynomial map. Can we say something about the image itself? As we will see the answer is always ‘yes’, however the situation very much depends over which field we work and whether we are in the projective case or the affine case. In this section we discuss tools for computing such images. We begin by highlighting the difference between the real numbers and complex numbers with regard to this problem.

We start with a real, affine variety $X \subset \mathbb{R}^m$. We would like to understand the image of $X$ under a polynomial map:

$$f = (f_1,\ldots,f_n) : \mathbb{R}^m \to \mathbb{R}^n.$$

An easy example when $n = m = 1$, $X = \mathbb{R}$ and $f(x) = x^2$ shows that the Zariski closure of the image, $\mathbb{R}$, and the image itself, $\mathbb{R}_{\geq 0}$ can differ a lot. Is there a chance in general of describe the image using polynomials? The following theorem provides a positive answer.

**Theorem 4.16 (Tarski-Seidenberg).** Over the field of real numbers the image of a variety is a semi-algebraic set (recall Definition 2.17).

**Proof.** See [4] Section 1.4. □

Thus to provide a description of the image of a real variety we need two ingredients: the polynomials that it satisfies and the polynomials providing the inequalities. If we pass to an algebraically closed field, the situation is much simpler.

**Theorem 4.17 (Chevalley).** If $K$ is algebraically closed, then the image of a variety is a constructible set. In particular, if $K = \mathbb{C}$ the Euclidean closure and the Zariski closure coincide.

**Proof sketch.** The image is a projection of the graph of the map. One can apply Nullstellensatz - see Chapter [3] - to turn the problem into one from linear algebra. Resultants and their matrices play a central role. Details can be found in e.g. [44] 7.4.6–7.4.8. □
Suppose we want to check if a random (in a reasonable sense) point belongs to the image and we work over $\mathbb{C}$. It is enough to check polynomial equations, which can be obtained as described in previous two subsection. Obtaining the whole description of the image is slightly more complicated. We may proceed as follows:

- Compute the closed image $X_0$.
- Subtract from it a proper subvariety $X_1$.
- Add back $X_2$ - a proper subvariety of $X_1$, etc.

This procedure must finish in a finite number of steps, by Hilbert’s Basis Theorem. For an implementation we refer to [17].

The situation is nicest when we are in the projective case. The following theorem may be regarded as an algebraic analog of the fact that images of compact sets under continuous maps are compact.

**Theorem 4.18.** Let $X$ be a projective variety over an algebraically closed field. Then the image of $X$ is (Zariski) closed.

**Proof.** See [34, Section 5.2] \qed

This is a very powerful theorem. In particular it implies the following. Consider a map $f = (f_1, \ldots, f_n)$ given by homogeneous polynomials of the same degree. Assume that $V(f_1, \ldots, f_n) = 0$. Then the image of $f$ is Zariski closed - we may compute it using elimination. For an application we refer to Exercise 18. More generally we have the following theorem.

**Theorem 4.19.** We fix the field of complex numbers. Consider a map $f = (f_1, \ldots, f_n)$ given by homogeneous polynomials of the same degree in $m + 1$ variables. Suppose that the closed image of $f$ has affine dimension $d + 1$, i.e. projective dimension $d$. Assume that $\dim V(f_1, \ldots, f_n) = b + 1$. If $d + b < m$ then the image of $f$ is closed.

**Proof.** We may consider $f$ as a map from $\mathbb{P}^m \setminus V(f_1, \ldots, f_n)$. Let $\mathbb{P}^d \subset \mathbb{P}^m$ be a random subspace. With probability one it will be disjoint from $V(f_1, \ldots, f_n)$. Thus we may assume that $f$ is well-defined on $\mathbb{P}^d$. Further, also with probability one, the image of $\mathbb{P}^d$ and $\mathbb{P}^m$ by $f$ will coincide. As the former is closed by Theorem 4.18 so must be the latter. \qed

**Exercises**

1. Eliminate the variable $z$ from the equations $x^3y^3z^3 - x - y - z = 1$ and $x^5 + y^5 + z^5 = 2$.
2. Prove: If an ideal $I$ is prime then so are its elimination ideals, and same for radical. Give the examples when the converse does not hold. What is the geometric meaning of these statements?
3. Compute the determinants of the Sylvester matrices $Syl_{1,5}$, $Syl_{2,4}$ and $Syl_{3,3}$. Each of them is a polynomial of degree 6 in 8 unknowns. Which of them has the most terms?
4. A plane curve has the parametrization $z \mapsto (f(z), g(z))$ where $f$ and $g$ are polynomials of degree 10. At most how many terms do you expect the implicit equation to have?
5. Can you find an invertible $5 \times 5$-matrix that is skewsymmetric?
(6) You are given all entries of a skewsymmetric $5 \times 5$ matrix $X = (x_{ij})$ except for $x_{12}$ and $x_{45}$. Under which condition on the 8 visible entries can you complete with rank($X$) $\leq 2$?

(7) Let $\pi$ be the linear map from $\mathbb{C}^3$ to $\mathbb{C}^2$ given by the matrix $\begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}$. Given an algebraic curve $V$ in $\mathbb{C}^3$, explain how one can compute the plane curve $\pi(V) \subset \mathbb{C}^2$.

(8) Consider the Fermat curve $V = \mathcal{V}(x^3 + y^3 + z^3)$ in the projective plane $\mathbb{P}^2$. Compute the ideal in 6 variables whose variety is the image of $V$ under the Veronese map

$$\mathbb{P}^2 \to \mathbb{P}^5, (x : y : z) \mapsto (x^2 : xy : xz : y^2 : yz : z^2).$$

(9) Determine the prime ideal of relations among the $3 \times 3$-minors of a $3 \times 6$-matrix.

(10) Let $V_1$ and $V_2$ be curves in $\mathbb{C}^3$ and $V_1 + V_2$ their pointwise sum. The Zariski closure $\overline{V_1 + V_2}$ is an algebraic variety in $\mathbb{C}^3$. Explain how one can compute its ideal $I(V_1 + V_2)$.

(11) Compute the hyperdeterminant of a $2 \times 2 \times 3$ tensor whose 12 entries are unknowns.

(12) Apply the method in Example 4.15 to compute the implicit equation of the plane curve that has the parametrization $z \mapsto (2z^3 + 3z^2 + 5z + 7, 11z^3 + 13z^2 + 17z + 19)$.

(13) Let $m = 2$, $d_1 = 1$, $d_2 = d_3 = 2$. The total number of coefficients is $n = 15 = 3+6+6$. Compute the resultant $\text{Res}(f_1, f_2, f_3)$ explicitly, as a polynomial in all 15 unknowns.

(14) Which constraints hold for off-diagonal entries of a rank one $3 \times 3$-matrix?

(15) Which constraints hold for off-diagonal entries of a nilpotent $3 \times 3$-matrix?

(16) Which constraints hold for off-diagonal entries of an orthogonal $3 \times 3$-matrix?

(17) Let $m = 2$ and $d_1 = d_2 = d_3 = 2$. Then $\text{Res}(f_1, f_2, f_3)$ is the resultant of three quadrics in the plane. This is a polynomial in 18 = 6+6+6 variables of degree 12 = 4+4+4. How many terms does it have? Find an explicit matrix formula for $\text{Res}(f_1, f_2, f_3)$.

(18) Let $V$ be the complex vector space of homogeneous polynomials in $n$ variables of degree $d$. The $d$-th powers of linear forms form a subset of $V$. Is it Zariski closed for any $n$ and $d$? What happens if we change the field to the real numbers?

(19) Consider a degree two polynomial $ax^2 + bx + c$, where $a, b, c$ are unknown. When does it have a double root and how is this related to resultants? What happens for higher degree polynomials?
CHAPTER 5

Linear Spaces and Grassmannians

In previous chapters we saw the construction of projective space, and we argued that in many applications projective varieties are preferable to affine varieties. Points in a projective space correspond to lines through the origin in the underlying vector space. In this chapter we replace lines with higher-dimensional linear subspaces. The role of the projective space is now played by a Grassmannian. This is a smooth projective variety whose points correspond to linear subspaces of a fixed dimension. For instance, the Grassmannian of lines in projective 3-space is a variety of dimension four. Its subvarieties represent families of lines. Counting lines that satisfy a certain property (e.g. lying on a cubic surface) leads us to enumerative algebraic geometry, a subject in which Grassmannians play a fundamental role.

5.1. Coordinates for Linear Spaces

Let $V$ be a vector space of dimension $n$ over a field $K$. In Chapter 2 we constructed the projective space $\mathbb{P}(V)$. Its points are the 1-dimensional subspaces of $V$. We note that $\mathbb{P}(V)$ is the key example of a compact algebraic variety when $K = \mathbb{C}$. Our aim is to generalize this construction from lines to subspaces of arbitrary dimension $k$. We will construct a projective variety $G(k, V)$ whose points correspond bijectively to $k$-dimensional subspaces of $V$. This variety is called the Grassmannian, after the 19th century mathematician Hermann Grassmann. If $V = K^n$ then we use the notations $\mathbb{P}^{n-1}$ for $\mathbb{P}(V)$ and $G(k, n)$ for $G(k, V)$.

We start with an explicit construction in coordinates, by fixing a basis $e_1, \ldots, e_n$ of $V$. Consider any $k$ linearly independent vectors $v_1, \ldots, v_k \in V$. We represent them in a form of a $k \times n$ matrix $M_W$ of rank $k$. To these vectors, or equivalently to a full rank matrix, we associate the linear subspace $\langle v_1, \ldots, v_k \rangle$ in $V$. This association is surjective, but not injective, as we may replace the $v_i$’s by linear combinations. In other words, the group $GL(k)$ acts on the set of $k \times n$ matrices by left multiplication, and this does not change the linear span of the rows.

We know some polynomial functions that do not change (up to scaling) under taking linear combinations of the rows: these are the $k \times k$ minors of the $k \times n$ matrix. Suppose that $W$ is a $k$-dimensional subspace of $V$. Pick any basis and express $W$ as the row space of a $k \times n$-matrix. We then write $i(W)$ for the vector of all $k \times k$-minor of that matrix, up to scale. This construction defines a map

$$i : \{k\text{-dimensional subspaces of } V\} \to \mathbb{P}(K(k)).$$

The map $i$ is well-defined since $i(W)$ does not depend on the basis chosen for $W$.

**Lemma 5.1.** The map $i$ is injective.

**Proof.** Consider two $k$-dimensional subspaces $W_1, W_2 \subseteq V$. Assume $i(W_1) = i(W_2)$. As the matrices $M_{W_1}$ and $M_{W_2}$ representing respectively $W_1$ and $W_2$ are
of full rank, without loss of generality we may assume that the minor given by
first $k$ columns is nonzero. By performing linear operations on the rows of both
matrices, we transform $M_{W_1}$ to a matrix $\tilde{M}_{W_1}$, whose left-most $k \times k$ submatrix
is the identity. We observe that any entry of $\tilde{M}_{W_1}$ not in the first $k$ columns, is equal
to some maximal minor or its negation. Thus, if $i(W_1) = i(W_2)$ the two matrices
$\tilde{M}_{W_1}, M_{W_2}$ must be equal. This implies $W_1 = W_2$. \hfill \square

The image of $i$ is known as the Grassmannian $G(k,n)$ and the inclusion in
$\mathbb{P}(K^{(k)}_n)$ as the Plücker embedding. For readers familiar with the exterior power of
a vector space, here is a more invariant way to describe the Grassmannian:

$$G(k,n) = \{ [v_1 \wedge \cdots \wedge v_k] \in \mathbb{P}(\bigwedge^k V) : v_1, \ldots, v_k \in V \text{ are linearly independent} \}.$$ 

Indeed, first we may identify $\mathbb{P}(K^{(k)}_n)$ with $\mathbb{P}(\bigwedge^k V)$ by fixing a basis of $V$ and an
induced basis of $\bigwedge^k V$. Expanding $v_1 \wedge \cdots \wedge v_k$ in a basis we indeed obtain the
minors of $[v_1, \ldots, v_k]$. The group $GL(V)$ acts naturally on $V$, taking subspaces to
subspaces. This induces an action on $\mathbb{P}(\bigwedge^k V)$, that restricts to the Grassmannian.
Precisely $g \in GL(V)$ transforms $v_1 \wedge \cdots \wedge v_k$ to $g(v_1) \wedge \cdots \wedge g(v_k)$. We note
that the action is transitive: for any $p_1, p_2 \in G(k,V)$ there exists a (non-unique)
automorphism $g \in GL(V)$ such that $g(p_1) = p_2$. This holds because any set of $k$
linearly independent vectors may be transformed by an invertible linear map to any
other such set. Hence, $G(k,V)$ is an orbit under the action of $GL(V)$ on $\mathbb{P}(\bigwedge^k V)$.
In fact, $G(k,V)$ is the unique closed orbit in this space.

Projective algebraic varieties that are orbits of linear algebraic groups are called
homogeneous, the Grassmannians being prominent examples. Homogeneous vari-
eties are always smooth. Indeed, any algebraic variety always contains a smooth
point and an action of a group must take a smooth point to a smooth point - a
version of this statement is given in Exercise 2.

5.2. Plücker Relations

Our aim is now to demonstrate that the Grassmannian is a projective variety.
Equivalently, we need to express the fact that $(\ell \choose k)$ numbers are minors of a matrix,
by vanishing of (homogeneous) polynomials.

**Theorem 5.2.** The Grassmannian $G(k,n) \subset \mathbb{P}(K^{(k)}_n)$ is Zariski closed and
irreducible.

**Proof.** Lemma 5.1 gives us an idea how to proceed. Namely, first let us
assume that the matrix $M_{W_1}$ representing $W$ is of the form:

$$
\begin{bmatrix}
1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & 1
\end{bmatrix}
\begin{bmatrix}
A
\end{bmatrix},
$$

where $A$ is a $k \times (n-k)$ matrix. Each maximal minor of $M_{W_1}$ is now, up to sign,
a minor of $A$ of some size. Further, by Laplace expansion, a $q \times q$ minor of $A$ for
$q > 1$ may be expressed, as a quadratic polynomial, in terms of smaller minors. This
exactly provides us a collection of $\sum_{q=2}^{\min(k,n-k)} \binom{n-k}{q} \binom{n-k}{q}$ inhomogeneous quadratic
equations in the entries of the $k \times (n-k)$-matrix $A$. These quadratic equations define
the part of image of our map $i$ that lies in the affine open set $K^{(k)}_{n-1} \subseteq \mathbb{P}(K^{(k)}_{n})$ given by the nonvanishing of the first coordinate.

If $i(W)$ has its first coordinate zero then some other coordinate will be nonzero. In other words, the matrix $M_W$ must have some invertible $k \times k$ submatrix. If we multiply $M_W$ on the left by the inverse of that matrix then we obtain a matrix that looks like (5.1) but with its columns permuted. The same construction as before gives us a system of $\sum_{q=2}^{\min(k,n-k)} \binom{k}{q} \binom{n-k}{q}$ inhomogeneous quadratic equations in the $k(n - k)$ entries of the new matrix $A$.

Each of the quadratic equations in $k(n - k)$ variables obtained above can be written as a homogeneous quadratic equation in the $\binom{n}{k}$ coordinates on $\mathbb{P}(K^{(k)}_{n})$. Namely, a minor of $A$ is replaced by the corresponding maximal minor of $M_W$, and then the quadric is homogenized by the special minor that corresponds to the identity matrix in (5.1). The collection of all these homogeneous quadratic equations gives a full polynomial description of $G(k,n)$.

The Grassmannian $G(k,n)$ is an irreducible subvariety of $\mathbb{P}(K^{(k)}_{n})$ because it is the image of a polynomial map $i$, namely the image of the space $K^{k\times n}$ of all $k \times n$ matrices under taking all maximal minors.

\[ \square \]

We have proved that the set $G(k,n)$ is cut out by quadratic equations. In fact, with slightly more effort one can show that $I(G(k,n))$ may be generated by quadratic polynomials, known as Plücker relations [27, Chapter 3]. Further below we will discuss the Plücker relations for $k = 2$. From a more algebraic perspective the equations vanishing on $G(k,n)$ are exactly the polynomial relations among maximal minors. We point out that finding polynomial equations among nonmaximal minors of a fixed size is an open problem in commutative algebra.

Another fact that follows from the proof, is that the intersection $G(k,n) \cap K^{(k)}_{n-1}$ of the Grassmannian with the open affine is the affine space $K^{k\times(n-k)}$.

\[ \text{Corollary 5.3. The dimension of the Grassmannian } G(k,n) \text{ equals } k(n-k). \]

\[ \text{Remark 5.4. The Grassmannian } G(k,n) \text{ parametrizes } k \text{-dimensional vector subspaces of an } n \text{-dimensional vector space, or equivalently } k - 1 \text{ dimensional projective subspaces of an } n - 1 \text{ dimensional projective space.} \]

\[ \text{Example 5.5. We consider } G(2,4). \text{ The Grassmannian is the image of a map:} \]
\[
\begin{bmatrix}
a & b & c & d \\
e & f & g & h \\
\end{bmatrix}
\rightarrow (af - be : ag - ce : ah - de : bg - cf : bh - df : ch - dg) \in \mathbb{P}^5.
\]

Alternatively, fixing a basis $(v_1, v_2, v_3, v_4)$ of $V$ we may write:
\[
(av_1 + bv_2 + cv_3 + dv_4) \land (ev_1 + fv_2 + gv_3 + hv_4) =
(af - be)v_1 \land v_2 + (ag - ce)v_1 \land v_3 + (ah - de)v_1 \land v_4
+(bg - cf)v_2 \land v_3 + (bh - df)v_2 \land v_4 + (ch - dg)v_3 \land v_4.
\]

This Grassmannian has dimension 4, i.e. it is a hypersurface in $\mathbb{P}^5$. We write the coordinates on $\mathbb{P}^5$ as $(p_{12} : p_{13} : p_{14} : p_{23} : p_{24} : p_{34})$. The indices refer to the minors of a $2 \times 4$-matrix. Following the proof of Theorem 5.2 we look at matrices (5.1). They take the form
\[
\begin{bmatrix}
1 & 0 & c & d \\
0 & 1 & g & h \\
\end{bmatrix}.
\]
The expansion of the rightmost $2 \times 2$-minor yields the inhomogeneous quadratic equation $p_{41} = ch - dg = (-p_{23})p_{14} - (-p_{24})p_{13}$. We homogenize this equation with the extra variable $p_{12}$. We conclude that $G(2, 4)$ is the hypersurface in $\mathbb{P}^5$ defined by the Plücker quadric

$$p_{23}p_{14} - p_{13}p_{24} + p_{12}p_{34}. \tag{5.2}$$

We now discuss the homogeneous prime ideal $I(\Gr(k, n))$ of the Grassmannian $\Gr(k, n)$. A complete answer is known, in terms of certain quadratic relations that form a Gröbner basis. These are known as straightening relations. For a derivation and explanation see e.g. [38, Chapter 3]. These relations are implemented in the computer algebra system Macaulay2, where one obtains generators for $I(\Gr(k, n))$ with the convenient command Grassmannian.

We here present the answer in the special case $k = 2$. The corresponding Grassmannian $\Gr(2, n)$ is the space of lines in $\mathbb{P}^{n-1}$. It is convenient to write the \binom{n}{2} Plücker coordinates as the entries of a skew-symmetric $n \times n$-matrix $\mathbf{P} = (p_{ij})$. We are interested in the principal submatrices of $\mathbf{P}$ having size $4 \times 4$. One such submatrix is given by taking the first four rows and first four columns. The determinant of that matrix is the square of the Plücker quadric (5.2). One refers to the square root of the determinant of a skew-symmetric matrix of even order as its pfaffian. Thus the $4 \times 4$ pfaffians of our matrix $\mathbf{P}$ are the \binom{n}{4} quadrics

$$p_{il}p_{jk} - p_{ik}p_{jl} + p_{ij}p_{kl} \quad \text{for} \quad 1 \leq i < j < k < l \leq n. \tag{5.3}$$

**Theorem 5.6.** The \binom{n}{4} quadrics listed in (5.3) form the reduced Gröbner basis of the Plücker ideal $I(\Gr(2, n))$, for any monomial ordering on the polynomial ring in the \binom{n}{2} variables $p_{ij}$ that selects the underlined leading terms.

**Proof.** The argument in the proof of Theorem 5.2 shows that the quadrics (5.3) cut out $\Gr(2, n)$ as a subset in $\mathbb{P}^{\binom{n}{2}-1}$. In other words, our Grassmannian is given as the set of skew-symmetric $n \times n$-matrices whose $4 \times 4$ pfaffians vanish. These are skew-symmetric matrices of rank 2.

By Hilbert’s Nullstellensatz, we now know that the radical of $I(\Gr(2, n))$ is generated by (5.3). We need to argue that this ideal is radical. However, this follows from the assertion that (5.3) form a Gröbner basis. Indeed, the leading monomials $p_{il}p_{jk}$ are square-free, so they generate a radical monomial ideal. However, if the initial ideal $\text{in}(J)$ of some ideal $J$ is radical then also $J$ itself is radical. So, all we need to do is to verify the Gröbner basis property for our quadrics. That Gröbner basis is then automatically a reduced Gröbner basis because none of the two trailing terms in (5.3) is a multiple of some other leading term.

To verify the Gröbner basis property, we argue as follows. For $n = 4$, it is trivial because there is only one generator. For $n = 5, 6, 7$, this is a direct computation, e.g. using Macaulay2. One checks that the S-polynomial of any two quadrics in (5.3) reduces to zero. Suppose that $n \geq 8$ and consider two Plücker quadrics. These involve at most 8 distinct indices. If the number of distinct indices is 7 or less then we are done by the aforementioned computation, which verified the claim for $n \leq 7$. Hence we may assume that all eight indices occurring in the two Plücker quadrics are distinct. In that case, the two underlined leading monomials are relatively prime. Here, Buchberger’s Second Criterion applies, and we can conclude that the S-polynomial automatically reduces to zero. In conclusion, all S-polynomials formed by pairs from (5.3) reduce to zero. This completes the proof. \qed
5.3. Schubert Calculus

In this section we show how Grassmannians can help us answer enumerative problems, i.e., questions of type: how many lines or planes in space satisfy some properties. This branch is known as enumerative geometry and the answers we provide are based on Schubert calculus. In other words, we describe an intersection theory for subvarieties of a Grassmannian.

We start by describing special subvarieties of Grassmannians, on the example of \( G(2, 4) \). Let us fix a complete flag in \( \mathbb{P}(K^4) \), i.e. \( f_0 = \mathbb{P}^0 \subset f_1 = \mathbb{P}^1 \subset f_2 = \mathbb{P}^2 \subset \mathbb{P}^3 \). Our aim is to group families of projective lines, i.e. subvarieties of \( G(2, 4) \), according to how they intersect the flag. Clearly the flag distinguishes a point in \( G(2, 4) \), namely \( X_0 := f_1 \in G(2, 4) \). There is also a distinguished one dimensional variety \( X_1 \): lines \( l \) such that \( f_0 \in l \subset f_2 \). The most interesting is the case of two dimensional subvarieties. There are two types of those:

1. \( X_2 \) consisting of lines \( l \) such that \( f_0 \in l \) and\( l \subset f_2 \).
2. \( X_{2'} \) consisting of lines \( l \) such that \( l \subset f_2 \).

There is also one three dimensional variety \( X_3 \), consisting of all lines that intersect the given line \( f_1 \). The varieties \( X_1, X_2, X_{2'} \) and \( X_3 \) described are called Schubert subvarieties. In Exercise 1 you will generalize their construction to arbitrary Grassmannians.

We now describe the geometry of above varieties explicitly in \( \mathbb{P}^5 \), as in Example 5.5. We assume \( f_i = \langle v_0, \ldots, v_i \rangle \), for \( i = 0, 1, 2 \). The point \( f_1 \in G(2, 4) \) is given by vanishing of all the coordinates \( p_{ij} \) apart from \( p_{12} \). We have \( f_0 \in l \subset f_2 \) if and only if \( l \) admits the following matrix representation:

\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & f & g & 0
\end{bmatrix}
\]

In particular, the subvariety \( X_1 \) is given by vanishing of all \( p_{ij} \) apart from \( p_{12} \) and \( p_{13} \) and, as those two minors can be arbitrary, it is a line \( \mathbb{P}^1 \subset G(2, 4) \subset \mathbb{P}^5 \). Similarly, \( X_2 \) is a \( \mathbb{P}^2 \) with coordinates \( p_{12}, p_{13}, p_{14} \) and \( X_{2'} \) is a different \( \mathbb{P}^2 \) with coordinates \( p_{12}, p_{13}, p_{23} \). Recall from Example 5.5 that \( G(2, 4) \) is a four dimensional quadric in \( \mathbb{P}^5 \).

In general, for a nondegenerate \( 2k \) dimensional quadric \( Q \subset \mathbb{P}^{2k+1} \) and any \( L = \mathbb{P}^{k-1} \subset Q \) there exist exactly two \( k \)-dimensional subspaces that contain \( L \) and are contained in \( Q \). These are \( X_2 \) and \( X_{2'} \) for \( L = X_1 \). For \( k = 1 \) this is a classical fact of projective geometry. A quadratic surface in \( \mathbb{P}^3 \) is isomorphic to \( \mathbb{P}^1 \times \mathbb{P}^1 \). Through any point \( p = L \) there pass precisely two lines contained in the quadric. In Figure the blue quadric contains the red point \( p \) and the two lines are green. This three dimensional picture may be seen in fact as a cut of \( \mathbb{P}^5 \) with a codimension two projective space \( H \) defined by \( p_{12} = p_{34} = 0 \). The red point \( p \) equals \( H \cap X_1 \). The quadric is simply \( H \cap G(2, 4) \) and the two green lines correspond to \( X_2 \cap H \) and \( X_{2'} \cap H \).

For readers familiar with algebraic topology, we remark that Schubert varieties represent cohomology classes in the Grassmannian when \( K = \mathbb{C} \). Schubert calculus is the study of intersection of these classes. When intersecting Schubert varieties as cohomology classes one should consider them coming from different flags. As we have seen in our construction \( X_2 \cap X_{2'} = X_1 \). However, if we consider all lines in \( \mathbb{P}^3 \) going through a fixed point and all lines contained in an unrelated \( \mathbb{P}^2 \subset \mathbb{P}^3 \), we...
see that both sets of lines do not have common elements. Hence, the intersection of the two classes is empty. On the other hand, if we consider lines going through two points, or contained in two planes, in each case we obtain exactly one line. The selfintersections of $X_2$ and $X_2'$ (as cohomology classes) give one (point).

The discussion above can be summarized by the following multiplicative relations that hold in the cohomology ring of $G(2, 4)$. Recall that multiplication represents intersection:

\[
[X_3][X_3] = [X_2] + [X_2'],\quad [X_2][X_2] = [X_2'][X_2'] = [X_3][X_1] = 1,\quad [X_2][X_2'] = 0.
\]

With this we can now answer questions like the following: How many lines pass through four general lines in $\mathbb{P}^3$? The set of such lines is a finite set in $G(2, 4)$. It is the intersection of four hypersurfaces, all of the form $X_3$. Since intersections are represented by multiplication in the cohomology ring, the following computation shows that the answer is two:

\[
[X_3]^4 = ([X_3][X_3])^2 = ([X_2] + [X_2'])^2 = [X_2]^2 + 2[X_2][X_2'] + [X_2']^2 = 1 + 2 \cdot 0 + 1 = 2.
\]

Here is another question that can be answered by Schubert calculus: How many lines are simultaneously tangent to four general quadratic surfaces in $\mathbb{P}^3$? Suppose a given quadric $Q$ in $\mathbb{P}^3$ is represented as a symmetric $4 \times 4$-matrix. Let $\wedge^2 Q$ be the symmetric $6 \times 6$ matrix with entries given by the $2 \times 2$ minors of the matrix representing $Q$. The condition for a line to be tangent to $Q$ is expressed by the vanishing of the quadratic form $P(\wedge^2 Q)P^T$ in the Plücker coordinates $P = (p_{12}, p_{13}, \ldots, p_{34})$ of that line. Therefore, the cohomology class of lines tangent to $Q$ is $2[X_3]$. Therefore, the number of lines tangent to four given general quadrics in $\mathbb{P}^3$ equals

\[
(2[X_3])^4 = 16[X_3]^4 = 16 \cdot 2 = 32.
\]

In order to actually compute the 32 lines over $\mathbb{C}$, given four concrete quadrics in $\mathbb{P}^3$, we would need to carry out some serious Gröbner basis computations.

Remark 5.7. Grassmannians are named after Hermann Grassmann. However, it was Julius Plücker who first realized that lines in 3-space may be studied as a four-dimensional object [31]. The (earlier) discoveries of Grassmann were much more
fundamental: he was the one to realize that algebraic setting of geometry allows to consider objects not in three-dimensional space, but in any dimension.

Exercises

(1) Fix a complete flag in $\mathbb{P}^n$. Construct a bijection between:
- subvarieties of $G(k,n)$, that can be defined as $l \in G(k,n)$ that intersect each element of the flag in at most the given dimension, and
- Young diagrams contained in a $k \times (n-k)$ rectangle.

The codimension (or dimension - depending on the construction you choose) of the subvariety in $G(k,n)$ equals the number of boxes in the corresponding Young diagram.

(2) Let $G$ be a subgroup of $GL(V)$ and let $X \subset V$ be a variety, such that the action of $G$ on $V$ restricts to $X$. Prove that if $x$ is a smooth point of $X$ and $g \in G$, then $gx$ is also a smooth point. Hint: Consider the action of $G$ on the polynomial ring.

(3) Consider $G(2,4) \times G(2,4) \subset \mathbb{P}^5 \times \mathbb{P}^5$. Describe the locus of pairs of lines $(l_1, l_2) \in G(2,4) \times G(2,4)$ such that $l_1$ intersects $l_2$ in $\mathbb{P}^3$. Hint: Present both lines as $2 \times 4$ matrices. Note that two lines in $\mathbb{P}^3$ intersect if and only if they do not span the whole ambient space. Apply Laplace expansion of the determinant.

(4) For a variety $X \subset \mathbb{P}^n$, one considers a subset of $G(k+1,n+1)$ of $P_k \subset X$. This is known as the Fano variety of $k$ dimensional subspaces of $X$. Fix a nondegenerate quadric $Q \subset \mathbb{P}^3$. Describe the Fano variety of lines in it. Hint: One may solve this exercise either theoretically or using algebra software. Also Figure 1 gives a hint about the answer.

(5) How many real lines in 3-space can be simultaneously tangent to four given spheres?

(6) The two lines incident to four given real lines in $\mathbb{P}^3$ can be either real or complex. In the latter case they form a complex conjugate pair. Write down a polynomial in the $24 = 4 \cdot 6$ Plücker coordinates of four given lines whose sign distinguishes the two cases.

(7) How many lines in $\mathbb{P}^3$ are simultaneously incident to two given lines and tangent to two given quadratic surfaces?

(8) Consider the set of all lines in $\mathbb{P}^3$ that are tangent to the cubic Fermat surface $\{x_1^3 + x_2^3 + x_3^3 + x_4^3 = 0\}$. This set is an irreducible hypersurface in the Grassmannian $Gr(2, 4)$. Compute a polynomial in $p_{12}, p_{13}, \ldots, p_{34}$ that defines this hypersurface.

(9) Write down a minimal generating set for the ideal of the Grassmannian $Gr(3, 6)$.

(10) Prove that the determinant of a skew-symmetric $n \times n$-matrix is zero if $n$ is odd, and it is the square of a polynomial when $n$ is even.

(11) Examine the monomial ideal that is generated by the underlined initial monomials in (5.3). Express this ideal as the intersection of prime ideals. How many primes occur?

(12) Fix six general planes $\mathbb{P}^2$ in $\mathbb{P}^4$. How many lines in $\mathbb{P}^4$ intersect all six planes?

(13) Let $n = 2k$ and suppose that the $n \times n$-matrix $A = (a_{ij})$ in (5.1) is symmetric, i.e. its entries satisfy the equations $a_{ij} = a_{ji}$. Write these equations in terms of the $\binom{2n}{n}$ Plücker coordinates. The resulting subvariety of $Gr(n, 2n)$ is the Lagrangian Grassmannian.
CHAPTER 6

Nullstellensätze

The German noun Nullstellensatz refers to a theorem that characterizes the existence of a zero (= Nullstelle) for a system of polynomials. The classical version, due to Hilbert, works over algebraically closed fields. It says that the nonexistence of zeros is equivalent to the existence of a partition of unity for the given polynomials. A more general version furnishes a bijection between varieties and radical ideals. In this chapter we also discuss the analogous results over the field of real numbers. Here the main results are the real Nullstellensatz and the Positivstellensatz. These furnish criteria for polynomial equations and inequalities to have no real solutions. This leads us to real radical ideals and their characterization via sums of squares.

6.1. Certificates for Infeasibility

In Chapter 3 we have discussed how to find and represent a solution of a system of polynomial equations, but what if such a solution does not exist? In this section we present methods to prove that a given system has no solution.

Throughout this section, we fix an algebraically closed field $K$, and we write $K[x] = K[x_1, \ldots, x_n]$ for its polynomial ring in $n$ variables. For an ideal $I \subset K[x]$ we denote the associated variety in $K^n$ by $V(I)$. We begin with the following weak version of the Nullstellensatz. This result appears as Theorem 1 in [7, §4.1].

Theorem 6.1. If $I$ is a proper ideal in $K[x]$ then its variety $V(I)$ in $K^n$ is non-empty.

Proof. We use induction on $n$, following [7, §4.1]. For $n = 1$ our statement holds because every non-constant polynomial in one variable has a zero in the algebraically closed field $K$.

Let now $n \geq 2$. For $a \in K$, we write $I_{x_n = a}$ for the ideal in $K[x_1, \ldots, x_{n-1}]$ that is obtained by setting $x_n = a$ in each element of $I$. One easily checks that this is indeed an ideal. We claim that there exists a scalar $a \in K$ such that $1 \notin I_{x_n = a}$. By induction, there is a point $(a_1, \ldots, a_{n-1})$ in $V(I_{x_n = a})$. This implies that $(a_1, \ldots, a_{n-1}, a)$ is a point in the variety $V(I)$.

To prove the claim, we distinguish two cases. First suppose $I \cap K[x_n] \neq \{0\}$. Since $1 \notin I$, the principal ideal $I \cap K[x_n]$ is generated by a nonconstant polynomial

$$f(x_n) = \prod_{i=1}^r (x_n - b_i)^{m_i}.$$ 

Suppose that $1 \in I_{x_n = b_i}$ for $i = 1, 2, \ldots, r$. If this is not the case then we are done. Hence there exist $B_1, \ldots, B_r \in I$ such that $B_i(x_1, \ldots, x_{n-1}, b_i) = 1$ for all $i$. Note that $B_i$ is congruent to 1 modulo $(x_i - b_i)$ in $K[x]$. This implies that the product $\prod_{i=1}^r B_i^{m_i}$ is congruent to 1 modulo $(f)$. Since $f \in I$, we conclude that $1 \in I$.
Next suppose \( I \cap K[x_n] = \{0\} \). Let \( \{g_1, \ldots, g_r\} \) be a Gröbner basis for \( I \) with respect to the lexicographic order with \( x_1 > \cdots > x_n \). Write \( g_i = c_i(x_n)x^{\alpha_i} + \text{lower order terms} \), where \( x^{\alpha_i} \) is a monomial in \( x_1, \ldots, x_{n-1} \). Since \( K \) is infinite, we can choose \( a \in K \) such that \( c_i(a) \neq 0 \) for all \( i \). The polynomials \( \overline{g}_i = g_i(x_1, \ldots, x_{n-1}, a) \) form a Gröbner basis for \( I_{x_n=a} \), for the lexicographic monomial order, with leading monomials \( x^{\alpha_i} \) for \( i = 1, \ldots, r \). None of these monomials is 1, since \( I \cap K[x_n] = \{0\} \). This implies that 1 is not in the ideal \( I_{x_n=a} \).

Theorem 6.1 gives a certificate for the non-existence of solutions to polynomial equations, namely the partition of unity promised in the opening paragraph above.

**Corollary 6.2.** A collection of polynomials \( f_1, \ldots, f_r \in K[x] \) either has a common zero in \( K^n \) or there exists a certificate \( g_1f_1 + \cdots + g_rf_r = 1 \) with polynomial multipliers \( g_1, \ldots, g_r \in K[x] \).

**Proof.** Let \( I = \langle f_1, \ldots, f_r \rangle \). The either \( \mathcal{V}(I) \neq \emptyset \) or \( \mathcal{V}(I) = \emptyset \). In the latter case, \( 1 \in I \).

**Example 6.3.** Let \( n = 2 \) and consider the following three polynomials:

\[
\begin{align*}
    f_1 &= (x+y-1)(x+y-2), & f_2 &= (x-y+3)(x+2y-5), & f_3 &= (2x-y)(3x+y-4).
\end{align*}
\]

These do not have a common zero. This is proved by the Nullstellensatz certificate

\[
\begin{align*}
    g_1f_1 + g_2f_2 + g_3f_3 &= 1,
    \end{align*}
\]

where

\[
\begin{align*}
    g_1 &= \frac{895}{591}x^2 - \frac{6263}{591}x - \frac{2617}{591}y + \frac{4327}{591}, \\
    g_2 &= \frac{5191}{3760}x^2 + \frac{945}{3760}xy - \frac{3027}{3760}x - \frac{15125}{3760}y + \frac{7560}{3760}, \\
    g_3 &= -\frac{179}{725}x^2 - \frac{716}{725}xy + \frac{1453}{725}x - \frac{716}{725}y + \frac{1377}{725}.
\end{align*}
\]

The reader is invited to verify the identity (6.1), or to find other multipliers \( g_1, g_2, g_3 \).

There are two possible methods for computing the multipliers \( (g_1, \ldots, g_r) \) in Corollary 6.2. The first is to use the Extended Buchberger Algorithm. This is analogous to the Extended Euclidean Algorithm for integers or polynomials in one variable. For instance, given a collection of relatively prime integers, this writes 1 as a \( \mathbb{Z} \)-linear combination of these integers.

In the Extended Buchberger Algorithm one keeps track of the polynomial multipliers that are used to generate new S-polynomials from current basis polynomials. In the end, each element in the final Gröbner basis is written explicitly as a polynomial linear combination of the input polynomials. If \( \mathcal{V}(I) = \emptyset \) then that final Gröbner basis is the singleton \( \{1\} \).

The second method for computing Nullstellensatz certificates is to use degree bounds plus linear algebra. Let \( d \) be any integer that exceeds the degree of each \( f_i \). Let \( g_i \) be a polynomial of degree \( d - \deg(f_i) \) with coefficients that are unknowns, for \( i = 1,2,\ldots, r \). The desired identity \( \sum_{i=1}^rg_if_i = 1 \) translates into a system of linear equations in all of these unknowns. We solve this system. If a solution is found then this gives a certificate. If not then there is no certificate in degree \( d \), and we try a higher degree.

The two methods, in complete generality, can be very complicated to carry out in practice. Both: the computation of Gröbner basis and degree bounds on the polynomials \( g_i \) are not polynomial - a vast majority of mathematicians believe there is no polynomial algorithm to decide if a given polynomial system has a complex solution. The situation is even more complicated over \( \mathbb{R} \) or \( \mathbb{Z} \). In the latter case,
it is known there exists no algorithm at all – irrespective of the complexity – to
decide if a system has an integral solution. This was Hilbert’s 10-th problem.

6.2. Hilbert’s Nullstellensatz

Hilbert’s Nullstellensatz also offers a characterization of the set of all polyno-
mials that vanish on a given variety. This classical result from 1890 works over any
algebraically closed field \( K \), such as the complex numbers \( K = \mathbb{C} \). In this section
we present this theorem and some of its ramifications.

Recall that the \textit{radical} of an ideal \( I \) in \( K[x] \) is the (possibly larger) ideal
\[ \sqrt{I} = \{ f \in K[x] : f^m \in I \text{ for some } m \in \mathbb{N} \}. \]
This is a radical ideal, hence it is an intersection of prime ideals.

\textbf{Example 6.4.} Let \( n = 4 \) and consider the ideal \( I = \langle x_1x_3, x_1x_4+x_2x_3, x_2x_4 \rangle \).
This is not radical. To see this, note that the monomial \( f = x_1x_4 \) is not in \( I \) but
\( f^2 \) is in \( I \). The radical of \( I \) equals
\[ \sqrt{I} = \langle x_1x_3, x_1x_4, x_2x_3, x_2x_4 \rangle = \langle x_1, x_2 \rangle \cap \langle x_3, x_4 \rangle. \]

How many associated primes does the ideal \( I \) have? Do Gröbner bases of \( I \) give
any hints? We refer to Example 3.27 for an explicit primary decomposition of \( I \).

We now show that \( \sqrt{I} \) comprises all polynomials that vanish on \( \mathcal{V}(I) \).

\textbf{Theorem 6.5 (Hilbert’s Nullstellensatz).} For any ideal \( I \) in the polynomial
ring \( K[x] \) over an algebraically closed field \( K \), we have
\begin{equation}
\mathcal{I}(\mathcal{V}(I)) = \sqrt{I}. \tag{6.2}
\end{equation}

\textbf{Proof.} The radical \( \sqrt{I} \) is contained in \( \mathcal{I}(\mathcal{V}(I)) \), because \( f^m(a) = 0 \) implies
\( f(a) = 0 \) for all \( a \). We must show the left hand side is a subset of the right hand
side in (6.2). Let \( I = \langle f_1, \ldots, f_r \rangle \) and suppose that \( f \) is a polynomial that vanishes
on \( \mathcal{V}(I) \). Let \( y \) be a new variable and consider the ideal \( J = \langle f_1, \ldots, f_r, yf - 1 \rangle \)
in polynomial ring \( K[x,y] = K[x_1, \ldots, x_n, y] \). The variety \( \mathcal{V}(J) \) is empty because
\( f = 0 \) on every zero of \( f_1, \ldots, f_r \) and \( f \neq 0 \) on every zero of \( yf - 1 \). By Theorem
6.1 there exist multipliers \( g_1, \ldots, g_r, h \) in \( K[x,y] \) such that
\[ \sum_{i=1}^r g_i(x,y) \cdot f_i(x) + h(x,y) \cdot (yf(x) - 1) = 1. \]

We now substitute \( y = 1/f(x) \) into this identity. This yields the following identify
do rational functions in \( n \) variables:
\[ \sum_{i=1}^r g_i(x, \frac{1}{f(x)}) \cdot f_i(x) = 1. \]

The common denominator equals \( f(x)^m \) for some \( m \in \mathbb{N} \). Multiplying both sides
with this common denominator, we obtain a polynomial identity of the form
\[ \sum_{i=1}^r p_i(x) \cdot f_i(x) = f(x)^m. \]

This shows that \( f^m \) lies in \( I \), and hence \( f \) lies in the radical \( \sqrt{I} \). \( \square \)
EXAMPLE 6.6. Which polynomial functions vanish on all nilpotent $3 \times 3$-matrices? We set $n = 9$ and take $I$ to be ideal generated by the entries of $X^3$, where $X = (x_{ij})$ is a $3 \times 3$-matrix with variables as entries. These are nine homogeneous cubic polynomials in nine unknowns $x_{ij}$. The radical of $I$ is generated by the coefficients of the characteristic polynomial of $X$:

$$
\sqrt{I} = \left\langle x_{11} + x_{22} + x_{33}, x_{11}x_{22} + x_{11}x_{33} - x_{12}x_{21} - x_{13}x_{31} + x_{22}x_{33} - x_{23}x_{32}, \det(X) \right\rangle
$$

This reflects the familiar fact that a square matrix is nilpotent if and only if it has no eigenvalues other than zero. Theorem 6.5 implies that every polynomial that vanishes on nilpotent $3 \times 3$-matrices is a polynomial linear combination of the three generators above.

The Nullstellensatz implies a one-to-one correspondence between ideals and radical ideals.

COROLLARY 6.7. The map $V \mapsto \mathcal{I}(V)$ defines a bijection between varieties in $K^n$ and radical ideals in $K[x]$. The inverse map that takes radical ideals to varieties is given by $I \mapsto \mathcal{V}(I)$.

PROOF. The Nullstellensatz tells us that $V = \mathcal{V}(%I(V))$ and $I = \mathcal{I}(\mathcal{V}(I))$. This shows that both maps are one-to-one and onto, and that they are the inverses of each other. \Box

COROLLARY 6.8. The map $V \mapsto \mathcal{I}(V)$ defines a bijection between irreducible varieties in $K^n$ and prime ideals in $K[x]$. As before, the inverse map is given by $I \mapsto \mathcal{V}(I)$.

PROOF. A variety $V$ is irreducible if and only if its associated radical ideal $\mathcal{I}(V)$ is prime. \Box

6.3. Let’s Get Real

Does there exist an analog of Hilbert’s Nullstellensatz over an ordered field, such as the real numbers $K = \mathbb{R}$? We shall see that the answer is affirmative. In this section we discuss the real Nullstellensatz and the Positivstellensatz. These concern systems of polynomial equations and inequalities over the real numbers. They generalize Linear Programming Duality, for systems of linear equations and linear inequalities over $\mathbb{R}$. Moreover, as we shall see in Chapter 12, the Positivstellensatz plays an important role in Nonlinear Optimization.

None the results of the previous section are valid when $K = \mathbb{R}$ is the field of real numbers. To see this, let $n = 2$ and consider varieties in the real plane $\mathbb{R}^2$. For Theorem 6.1 we take $I = \langle x^2 + y^2 + 1 \rangle$. This is proper ideal in $\mathbb{R}[x, y]$ but $\mathcal{V}_\mathbb{R}(I) = \emptyset$. For Theorem 6.5 we take $I = \langle x^2 + y^2 \rangle$. This is a radical ideal, but

$$
\mathcal{I}(\mathcal{V}_\mathbb{R}(I)) = \langle x, y \rangle \quad \text{strictly contains} \quad \sqrt{I} = I.
$$

This raises the following questions about ideals $I$ in $\mathbb{R}[x]$ and their varieties in $\mathbb{R}^n$:

- Is there an algebraic rule for certifying that the real variety $\mathcal{V}_\mathbb{R}(I)$ is empty?
- Is there an algebraic rule for computing $\mathcal{I}(\mathcal{V}_\mathbb{R}(I))$ from generators of $I$?

The answer to these questions is given by the real Nullstellensatz. Our point of departure for this result is the observation that any polynomial in $\mathbb{R}[x]$ that is a sum of squares must be nonnegative, i.e. the inequality $f(u) \geq 0$ holds for all
u ∈ ℝ^n. A natural question is whether the converse holds: can every nonnegative polynomial be written as a sum of squares?

Hilbert showed in 1893 that the answer is negative if one asks for squares of polynomials. However, the answer is positive if one allows squares of rational functions. This was the 17th problem in Hilbert’s famous list from the International Congress of Mathematicians in 1900. It was solved by Emil Artin in 1927.

**Theorem 6.9 (Artin’s Theorem).** Let f be a polynomial in ℝ[x] that is nonnegative on ℝ^n. Then there exist polynomials p_1, p_2, ..., p_r, q_1, q_2, ..., q_r ∈ ℝ[x] such that

\[
\left( \frac{p_1}{q_1} \right)^2 + \left( \frac{p_2}{q_2} \right)^2 + \cdots + \left( \frac{p_r}{q_r} \right)^2.
\]

**Example 6.10 (Motzkin Polynomial).** Let n = 2 and consider the polynomial

\[ M(x, y) = x^4y^2 + x^2y^4 + 1 - 3x^2y^2 \]

Distributing the three terms of the factor \[x^2 + y^2 + 1\], we see that the right hand side is a sum of four squares of rational functions. This shows that the Motzkin polynomial \[M(x, y)\] is nonnegative. However, \[M(x, y)\] is not a sum of squares in \([x, y]\). Suppose it were. Then there exist \(\alpha, \beta, \gamma, \delta \in \mathbb{R}\) such that

\[ M(x, y) = \sum_i (\alpha_ix^2y + \beta_ixy^2 + \gamma_ixy + \delta_i)^2, \]

The coefficient \(-3\) of \(x^2y^2\) in \(M(x, y)\) would then be equal to \(\sum_i \gamma_i^2 \geq 0\).

We shall view Artin’s Theorem 6.9 as a special case of the following more general statement, which is the real number analogue to the weak form of the Nullstellensatz, seen in Theorem 6.1.

**Theorem 6.11.** Let I be an ideal in \(\mathbb{R}[x]\) whose real variety \(\mathcal{V}_R(I)\) is empty. Then \(-1\) is a sum of squares of polynomials modulo I, i.e. there exist \(p_1, p_2, \ldots, p_r \in \mathbb{R}[x]\) such that

\[
(6.3) \quad 1 + p_1^2 + p_2^2 + \cdots + p_r^2 \in I.
\]

For the proof of Theorem 6.11 see the book by Murray Marshall [28 §2.3].

**Derivation of Theorem 6.9 from Theorem 6.11** Let y be a new variable and consider the \(g = f(x)y^2 + 1\) in \(\mathbb{R}[x, y]\). Since \(f\) is nonnegative, the real variety \(\mathcal{V}_R(g)\) is empty in \(\mathbb{R}^{n+1}\). Theorem 6.11 says that there exists a polynomial identity of the form

\[
(6.4) \quad 1 + p_1(x, y)^2 + p_2(x, y)^2 + \cdots + p_r(x, y)^2 + h(x, y)g(x, y) = 0.
\]

We substitute \(y = \pm \frac{1}{\sqrt{-f(x)}}\) into (6.4), which makes the last term cancel in both substitutions. Thereafter we multiply the two resulting expressions. The result no longer contains any radicals. We obtain an identity

\[
1 + \frac{1}{(-f(x))^d} \left( g_1(x)^2 + g_2(x)^2 + \cdots + g_r(x)^2 \right) = 0,
\]

where \(g_1, g_2, \ldots, g_r\) are polynomials, and \(d\) is a positive integer, necessarily odd. We subtract the constant 1 on both sides of this identity, and we multiply by \(-f(x)\) to obtain a representation of \(f(x)\) as a sum of squares of rational functions. This gives Artin’s Theorem. □
We next come to the Positivstellensätze, which concerns systems that have both equations and inequalities. To motivate this, we briefly review the corresponding statements for linear polynomials. This is known as Farkas’ Lemma, and it is at the heart of Linear Programming Duality. Informally, Farkas’ Lemma states that a system of linear equations and inequalities either has a solution in $\mathbb{R}^n$, or it has a dual solution which certifies that the original system has no solution. The precise statement can be stated in many equivalent versions. Here is one of them, selected to make the extension to higher-degree polynomials more transparent.

Let $f_1, \ldots, f_r, g_1, \ldots, g_s$ be polynomials of degree 1 in $\mathbb{R}[x]$, and consider the system
\begin{equation}
(6.5) \quad f_1(u) = 0, \ldots, f_r(u) = 0, \quad g_1(u) \geq 0, \ldots, g_s(u) \geq 0.
\end{equation}
In the dual problem, we seek real numbers $a_1, \ldots, a_r, b_1, \ldots, b_s \in \mathbb{R}$ such that
\begin{equation}
(6.6) \quad a_1 \cdot f_1 + \cdots + a_r \cdot f_r + b_1^2 \cdot g_1 + \cdots + b_s^2 \cdot f_s = -1 \quad \text{in} \ \mathbb{R}[x],
\end{equation}
It is clear that at most one of these two systems can have a solution. Indeed, since $b_1^2, \ldots, b_s^2$ are nonnegative, the left hand side of (6.6) is nonnegative for every vector $x$ that solves (6.5).

**Theorem 6.12 (Farkas’ Lemma).** Given any choice of linear polynomials $f_1, \ldots, f_r$ and $g_1, \ldots, g_s$ in $\mathbb{R}[x]$, exactly one of the following two statements is true:

(P) There exists a point $u \in \mathbb{R}^n$ such that (6.5) holds.

(D) There exist real numbers $a_1, \ldots, a_r, b_1, \ldots, b_s \in \mathbb{R}$ such that (6.6) holds.

Consider the system (6.5) where the $f_i$ and $g_j$ are allowed to be arbitrary polynomials. In the dual problem, we seek polynomials $a_i$ and $b_{j\nu}$ in $\mathbb{R}[x]$ such that
\begin{equation}
(6.7) \quad a_1 \cdot f_1 + \cdots + a_r \cdot f_r + \sum_{\nu \in \{0,1\}^s} (\sum_j b_{j\nu})^2 \cdot g_1^{\nu_1} \cdots g_s^{\nu_s} = -1.
\end{equation}
In the double sum on the right, we see linear combinations of squarefree monomials in $g_1, \ldots, g_s$ whose coefficients are sums of squares. The set of polynomials that admit such a representation is the quadratic module generated by $g_1, \ldots, g_s$. Quadratic modules associated with inequality constraints are fundamental in the study of semi-algebraic sets $\mathbb{R}^n$. Qua-

**Theorem 6.13 (Positivstellensatz).** Given any choice of polynomials $f_1, \ldots, f_r$ and $g_1, \ldots, g_s$ in $\mathbb{R}[x]$, exactly one of the following two statements is true:

(P) There exists a point $u \in \mathbb{R}^n$ such that (6.5) holds.

(D) There exist polynomials $a_i$ and $b_{j\nu}$ in $\mathbb{R}[x]$ such that (6.7) holds.

**Proof.** See [28] §2.3. $\square$

The dual solution (D) in Theorem 6.13 is similar in nature to that in Farkas’ Lemma. The one extra complication is that we now need products of the $g_i$. The result be rephrased in words as follows: if a system of polynomial equations and inequalities is infeasible then $-1$ lies in the sum of the ideal of the equations and the quadratic module of the inequalities. There is a more general version of the Positivstellensatz which also incorporates strict inequalities $h_1 > 0, \ldots, h_t > 0$. This is stated in [40] Theorem 7.5 and it is also proved in [28] §2.3.

The radical $\sqrt{I}$ of a polynomial ideal $I$ was the main player in the strong form of Hilbert’s Nullstellensatz (Theorem 6.5). It offers an algebraic representation for
polynomials that vanish on a given complex variety. We now come to the analogous result over the real numbers.

Given any ideal \( I \) in \( \mathbb{R}[x] \), we define its \textit{real radical} to be the following set
\[
\sqrt{\mathbb{R}}I = \{ f \in \mathbb{R}[x] : f^{2m} + g_1^2 + \cdots + g_s^2 \in I \text{ for some } m \in \mathbb{N} \text{ and } g_1, \ldots, g_s \in \mathbb{R}[x] \}.
\]
One checks that this is also an ideal in \( \mathbb{R}[x] \). We have the following analogue to Theorem 6.5.

**Theorem 6.14 (Real Nullstellensatz).** For any ideal in the polynomial ring \( \mathbb{R}[x] \), we have
\[
\sqrt{\mathbb{R}}I = \sqrt{I}.
\]

**Proof.** The argument is similar to that in the proof of Theorem 6.5. Again, it is clear that \( \sqrt{\mathbb{R}}I \) is contained in \( I(\mathbb{R}^k(I)) \). We need to show the reverse inclusion. Suppose that \( f \) vanishes on the real variety of \( I = (f_1, \ldots, f_r) \subset \mathbb{R}[x] \). We introduce a new variable \( y \) and consider ideal \( J = (f_1, \ldots, f_r, yf - 1) \) in \( \mathbb{R}[x, y] \). It satisfies \( \mathbb{R}^k(J) = \emptyset \). By Theorem 6.11 there exists an identity of the form (6.3) for the ideal \( J \). Substituting \( y = 1/f(x) \) into that identity and clearing denominators, we find that some even power of \( f \) plus a sum of squares lies in \( I \). This means that the polynomial \( f \) is in the real radical \( \sqrt{I} \).

**Example 6.15.** Fix the principal ideal generated by the Motzkin polynomial
\[
I = (M(x, y)) = (x^4y^2 + x^2y^4 + 1 - 3x^2y^2).
\]
Building on Example 6.14, we wish to compute the real radical \( \sqrt{\mathbb{R}}I \). It must contain the numerators of the four summands in the sum of squares representation of \( M \). This leads us to consider the ideal
\[
J = (xy(x^2 + y^2 - 2), x^2 - y^2).
\]
This ideal is not radical. Its radical is found to equal the Jacobian ideal of the Motzkin polynomial:
\[
\sqrt{J} = (M, \partial M/\partial x, \partial M/\partial y).
\]
This radical ideal is precisely the real radical we were looking for:
\[
\sqrt{\mathbb{R}}I = \sqrt{J} = (x, y) \cap (x - 1, y - 1) \cap (x - 1, y + 1) \cap (x + 1, y - 1) \cap (x + 1, y + 1).
\]
This means that the real variety \( \mathbb{R}^k(M) \) defined by the Motzkin polynomial consists of the five points \((1, 1), (1, -1), (-1, 1), (-1, -1)\) and \((0, 0)\) in \( \mathbb{R}^2 \). Since \( M \) is nonnegative, these zeros are singular points of the complex curve \( \mathbb{V}(M) \subset \mathbb{C}^2 \).

**Exercises**

1. Find univariate polynomials \( g_1, g_2, g_3, g_4 \) in \( \mathbb{Q}[x] \) such that
\[
g_1(x - 2)(x - 3)(x - 4) + g_2(x - 1)(x - 3)(x - 4) + g_3(x - 1)(x - 2)(x - 4) + g_4(x - 1)(x - 2)(x - 3) = 1.
\]
2. Prove that an ideal \( I \) in \( \mathbb{C}[x] \) contains a monomial if and only if all points in \( \mathbb{V}(I) \) have at least one zero coordinate. Describe an algorithm for testing whether this holds.
3. Let \( M \) be an ideal generated by monomials in \( K[x] \). How to compute the radical \( \sqrt{M} \)?
(4) For \( n = 4 \) let \( I \) be the ideal generated by the two cubics \( x_1^2 x_2 - x_4^2 x_1 \) and \( x_1 x_2^3 - x_3^4 \). Describe the projective variety \( \mathcal{V}(I) \) in \( \mathbb{P}^3 \). Find the radical ideal \( \sqrt{I} \). How many minimal generators does \( \sqrt{I} \) have and what are their degrees?

(5) Let \( V \) be the variety of orthogonal Hankel matrices of format \( 4 \times 4 \). This lives in \( \mathbb{R}^7 \). Describe the ideal \( \mathcal{I}(V) \). What are the irreducible components of \( V \)?

(6) For \( n = 3 \) let \( I \) be the ideal generated by the two quartics \( x_1^4 - x_1^2 x_2^2 \) and \( x_1^2 - x_3^4 \) in \( \mathbb{R}[x_1, x_2, x_3] \). Determine the radical \( \sqrt{I} \) and the real radical \( \sqrt[\mathbb{R}]{I} \). Write each of these two radical ideals as an intersection of prime ideals.

(7) Let \( f_1, \ldots, f_r \) and \( f \) be polynomials in \( \mathbb{Q}[x] \). Explain how Gröbner bases can be used to test whether \( f \) lies in the radical of the ideal \( I = \langle f_1, \ldots, f_r \rangle \).

(8) The circle defined by \( f = x^2 + y^2 - 4 \) does not intersect the hyperbola defined by \( g = xy - 10 \) in the real plane \( \mathbb{R}^2 \). Find a real Nullstellensatz certificate for this fact, i.e. write \(-1\) as a sum of squares modulo the ideal \( \langle f, g \rangle \) in \( \mathbb{R}[x, y] \).

(9) For any positive integer \( d \), exhibit a polynomial \( f \) and an ideal \( I \) in \( K[x] \) such that \( f^d \not\in I \) but \( f^{d+1} \in I \). How small can the degrees of the generators of \( I \) be?

(10) Let \( I \) be the ideal in \( \mathbb{R}[x, y, z] \) generated by the Robinson polynomial

\[
x^6 + y^6 + z^6 + 3x^2 y^2 z^2 - x^4 y^2 - y^4 z^2 - x^2 y^4 - x^4 z^2 - y^4 z^2 - y^2 z^4.
\]

Determine the real radical \( \sqrt[\mathbb{R}]{I} \) and the real variety \( \mathcal{V}_{\mathbb{R}}(I) \) in the plane \( \mathbb{P}^2_{\mathbb{R}} \).

(11) Show that Theorem 6.14 implies Theorem 6.9.

(12) What is the Effective Nullstellensatz?

(13) Find the radical and the real radical of the ideal \( I = \langle x^7 - y^7, x^8 - z^8 \rangle \) in \( \mathbb{R}[x, y, z] \). Explain the difference between these two radical ideals.
Tropical Algebra

The operations of addition and multiplication are familiar from primary school. We here redefine them. While tropical arithmetic may at first seem unnatural to the reader, we justified it with several applications, e.g. in the design of dynamic programming algorithms. The primary focus of this chapter is on tropical linear algebra. The point is that the piecewise linear structures of tropical mathematics offer yet another transition point between linear and nonlinear algebra. On the fully nonlinear side lies tropical algebraic geometry. We offer a first glimpse at this subject with a brief discussion of tropical varieties and their geometric properties.

7.1. Arithmetic

The tropical semiring \((\mathbb{R} \cup \{\infty\}, \oplus, \odot)\) consists of the real numbers \(\mathbb{R}\), together with an extra element \(\infty\) called infinity. The arithmetic operations of addition and multiplication are

\[
x \oplus y := \min(x, y) \quad \text{and} \quad x \odot y := x + y.
\]

The tropical sum of two numbers is their minimum, and the tropical product of two numbers is their usual sum. Here are some examples of how to do arithmetic in the tropical world:

\[
3 \oplus 7 = 3 \quad \text{and} \quad 3 \odot 7 = 10.
\]

Tropical addition and tropical multiplication are both commutative:

\[
x \oplus y = y \oplus x \quad \text{and} \quad x \odot y = y \odot x.
\]

These two arithmetic operations are also associative, and the times operator \(\odot\) takes precedence when plus \(\oplus\) and times \(\odot\) occur in the same expression. The distributive law holds:

\[
x \odot (y \oplus z) = x \odot y \oplus x \odot z.
\]

Here is a numerical example to show distributivity:

\[
3 \odot (7 \oplus 11) = 3 \odot 7 = 10,
\]

\[
3 \odot 7 \oplus 3 \odot 11 = 10 \odot 14 = 10.
\]

Both arithmetic operations have an identity element. Infinity is the identity element for addition and zero is the identity element for multiplication:

\[
x \oplus \infty = x \quad \text{and} \quad x \odot 0 = x.
\]

We also note the following identities involving the two identity elements:

\[
x \odot \infty = \infty \quad \text{and} \quad x \oplus 0 = \begin{cases} 0 & \text{if } x \geq 0, \\ x & \text{if } x < 0. \end{cases}
\]
Tropical Algebra

There is no subtraction in tropical arithmetic. There is no real number \( x \) that we can call “17 minus 8” because the equation \( 8 \oplus x = 17 \) has no solution \( x \). Tropical division is defined to be classical subtraction, so \((\mathbb{R} \cup \{ \infty \}, \oplus, \odot)\) satisfies all ring axioms except for the existence of an additive inverse. Such algebraic structures are called semirings, whence the name tropical semiring. It is essential to remember that “0” is the multiplicative identity element. For instance, all coefficients in the Binomial Theorem that “0” is the multiplicative identity element. For instance, all coefficients in the Binomial Theorem

\[
(x \oplus y)^3 = (x \oplus y) \odot (x \oplus y) \odot (x \oplus y) = 0 \circ x^3 \oplus 0 \circ x^2 y \oplus 0 \circ xy^2 \oplus 0 \circ y^3.
\]

Of course, the zero coefficients can here be dropped.

\[
(x \oplus y)^3 = x^3 \oplus x^2 y \oplus xy^2 \oplus y^3 = x^3 \oplus y^3.
\]

These identifies holds for all real numbers \( x, y \in \mathbb{R} \).

7.2. Linear Algebra

The familiar algebra of vectors and matrices make sense over the tropical semiring. For instance, the tropical scalar product in \( \mathbb{R}^3 \) of a row vector with a column vector is the scalar

\[
(u_1, u_2, u_3) \odot (v_1, v_2, v_3)^T = u_1 \odot v_1 \oplus u_2 \odot v_2 \oplus u_3 \odot v_3 = \min\{u_1 + v_1, u_2 + v_2, u_3 + v_3\}.
\]

Here is the product of a column vector and a row vector of length three:

\[
(u_1, u_2, u_3)^T \odot (v_1, v_2, v_3) = \begin{pmatrix}
  u_1 \odot v_1 \\
  u_2 \odot v_1 \\
  u_3 \odot v_1
\end{pmatrix} \odot \begin{pmatrix}
  u_1 \odot v_2 \\
  u_2 \odot v_2 \\
  u_3 \odot v_2
\end{pmatrix} = \begin{pmatrix}
  u_1 + v_1 & u_1 + v_2 & u_1 + v_3 \\
  u_2 + v_1 & u_2 + v_2 & u_2 + v_3 \\
  u_3 + v_1 & u_3 + v_2 & u_3 + v_3
\end{pmatrix}.
\]

Any matrix which can be expressed as such a product has tropical rank one.

Given a \( d \times n \)-matrix \( A \), we might be interested in computing its image \( \{ A \odot x : x \in \mathbb{R}^n \} \), and in solving the linear systems \( A \odot x = b \) for various right hand sides \( b \). For an introduction to tropical linear systems and their applications we recommend the books on Max-linear Systems by Butković [5] and Essentials of Tropical Combinatorics by Joswig [20].

For a first application of tropical algebra, consider the problem of finding shortest paths in a weighted directed graph. We fix a directed graph \( G \) with \( n \) nodes labeled \( 1, 2, \ldots, n \). Every directed edge \((i, j)\) in \( G \) has an associated length \( d_{ij} \) which is a non-negative real number. If \((i, j)\) is not an edge of \( G \) then we set \( d_{ij} = +\infty \). We represent \( G \) by its \( n \times n \) adjacency matrix \( D_G = (d_{ij}) \) with zeros on the diagonal and whose off-diagonal entries are the edge lengths \( d_{ij} \). The matrix \( D_G \) need not be symmetric; we allow \( d_{ij} \neq d_{ji} \) for some \( i, j \). However, if \( G \) is an undirected graph, then we represent \( G \) as a directed graph with two directed edges \((i, j)\) and \((j, i)\) for each undirected edge \(\{i, j\}\). In that special case, \( D_G \) is a symmetric matrix, where \( d_{ij} = d_{ji} \) is the distance between node \( i \) and node \( j \).

Consider the \( n \times n \)-matrix with entries in \( \mathbb{R}_{\geq 0} \cup \{ \infty \} \) that results from tropically multiplying the given adjacency matrix \( D_G \) with itself \( n - 1 \) times:

\[
D_G^{\odot(n-1)} = D_G \odot D_G \odot \cdots \odot D_G.
\]
Proposition 7.1. Let $G$ be a weighted directed graph on $n$ nodes with adjacency
determinant means solving the classical assignment problem of combinatorial
and column $j$ equals the length
of any path from node $i$ to node $j$ in the graph $G$.

Proof. Let $d_{ij}^{(r)}$ denote the minimum length of any path from node $i$ to node
which uses at most $r$ edges in $G$. We have $d_{ij}^{(1)} = d_{ij}$ for any two nodes $i$ and $j$. Since the edge weights $d_{ij}$ were assumed to be non-negative, a shortest path from node $i$ to node $j$ visits each node of $G$ at most once. In particular, any shortest path in the directed graph $G$ uses at most $n-1$ directed edges. Hence the length of a shortest path from $i$ to $j$ equals $d_{ij}^{(n-1)}$.

For $r \geq 2$ we have a recursive formula for the length of a shortest path:

$$d_{ij}^{(r)} = \min \{ d_{ik}^{(r-1)} + d_{kj} : k = 1, 2, \ldots, n \}.$$  

Using tropical arithmetic, this formula can be rewritten as follows:

$$d_{ij}^{(r)} = d_{i1}^{(r-1)} \circ d_{1j} + d_{i2}^{(r-1)} \circ d_{2j} + \cdots + d_{in}^{(r-1)} \circ d_{nj},$$

$$= (d_{i1}^{(r-1)}, d_{i2}^{(r-1)}, \ldots, d_{in}^{(r-1)}) \circ (d_{1j}, d_{2j}, \ldots, d_{nj})^T.$$  

From this it follows, by induction on $r$, that $d_{ij}^{(r)}$ equals the entry in row $i$ and
column $j$ of the $n \times n$ matrix $D_G^{\otimes r}$. Indeed, the right hand side of the recursive
formula is the tropical product of row $i$ of $D_G^{\otimes (r-1)}$ and column $j$ of $D_G$, which is
the $(i, j)$ entry of $D_G^{\otimes r}$. In particular, $d_{ij}^{(n-1)}$ is the entry in row $i$ and column $j$ of
$D_G^{\otimes (n-1)}$. This proves the claim. □

The above algorithm is an instance of what is known as Dynamic Programming
in Computer Science. For us, running that algorithm means performing the matrix
multiplication

$$D_G^{\otimes r} = D_G^{\otimes (r-1)} \circ D_G \quad \text{for } r = 2, \ldots, n-1.$$  

We next consider the notion of the tropical determinant. Fix an $n \times n$ matrix
$X = (x_{ij})$. As there is no negation in tropical arithmetic, we define this determinant
as the tropical sum over the tropical diagonal products obtained by taking all $n!$
permutations $\pi$ of $\{1, 2, \ldots, n\}$:

$$\text{tropdet}(X) := \bigoplus_{\pi \in S_n} x_{1\pi(1)} \circ x_{2\pi(2)} \circ \cdots \circ x_{n\pi(n)}.$$  

Here $S_n$ is the symmetric group of permutations of $\{1, 2, \ldots, n\}$. Evaluating the
tropical determinant means solving the classical assignment problem of combinatorial
optimization. Imagine a company that has $n$ jobs and $n$ workers, and each job
needs to be assigned to exactly one of the workers. Let $x_{ij}$ be the cost of assigning
job $i$ to worker $j$. The company wishes to find the cheapest assignment $\pi \in S_n$.
The optimal total cost equals

$$\min \{ x_{1\pi(1)} + x_{2\pi(2)} + \cdots + x_{n\pi(n)} : \pi \in S_n \}.$$  

This minimum is precisely the tropical determinant (7.4) of the matrix $X = (x_{ij})$:

Proposition 7.2. The tropical determinant solves the assignment problem.
In the assignment problem we seek the minimum over \( n! \) quantities. This appears to require exponentially many operations. However, there is a polynomial-time algorithm. It was developed by Harold Kuhn in 1955 who called it the Hungarian Assignment Method. This algorithm maintains a price for each job and a partial assignment of workers and jobs. At each iteration, an unassigned worker is chosen and a shortest augmenting path from this person to the set of jobs is chosen. The total number of arithmetic operations is \( O(n^3) \).

In classical arithmetic, the complexity of evaluating determinants and permanents differs greatly. The determinant of an \( n \times n \) matrix is a hard problem. Leslie Valiant proved that computing permanents is \#P-complete. In tropical arithmetic, computing the permanent is easier, thanks to the Hungarian Assignment Method. We can think of the Hungarian Method as a certain tropicalization of Gaussian Elimination.

Eigenvalues and eigenvectors of square matrices are a central topic in linear algebra. Let us now see their counterparts in tropical linear algebra. We fix an \( n \times n \)-matrix \( A = (a_{ij}) \) whose entries \( a_{ij} \) are in \( \mathbb{R} = \mathbb{R} \cup \{ \infty \} \). An eigenvalue of \( A \) is a real number \( \lambda \) such that

\begin{equation}
A \odot \mathbf{v} = \lambda \odot \mathbf{v} \quad \text{for some } \mathbf{v} \in \mathbb{R}^n.
\end{equation}

We say that \( \mathbf{v} \) is an eigenvector of the tropical matrix \( A \). The arithmetic operations in (7.6) are tropical. For instance, for \( n = 2 \), the left hand side of (7.6) equals

\[
\begin{pmatrix}
a_{11} & a_{12} \\
a_{21} & a_{22}
\end{pmatrix}
\odot
\begin{pmatrix}
v_1 \\
v_2
\end{pmatrix}
= \begin{pmatrix}
\min\{a_{11} + v_1, a_{12} + v_2\} \\
\min\{a_{21} + v_1, a_{22} + v_2\}
\end{pmatrix}.
\]

The right hand side of (7.6) is equal to

\[
\lambda \odot \begin{pmatrix}
v_1 \\
v_2
\end{pmatrix} = \begin{pmatrix}
\lambda \odot v_1 \\
\lambda \odot v_2
\end{pmatrix} = \begin{pmatrix}
\lambda + v_1 \\
\lambda + v_2
\end{pmatrix}.
\]

Let \( G(A) \) denote the directed graph with adjacency matrix \( A \). Its nodes are labeled by \( [n] = \{1, 2, \ldots, n\} \). There is an edge from node \( i \) to node \( j \) if and only if \( a_{ij} < \infty \), and the edge has length \( a_{ij} \). The normalized length of a directed path \( i_0, i_1, \ldots, i_k \) in \( G(A) \) is the sum (in classical arithmetic) of the lengths of the edges divided by the length \( k \) of the path. Thus the normalized length is \( (a_{i_0 i_1} + a_{i_1 i_2} + \cdots + a_{i_{k-1} i_k})/k \). If \( i_k = i_0 \) then the path is a directed cycle, and this quantity is the normalized length of the cycle. Recall that a directed graph is strongly connected if there is a directed path from any node to any other node.

**Theorem 7.3.** Let \( A \) be an \( n \times n \)-matrix such that \( G(A) \) is strongly connected. Then \( A \) has precisely one eigenvalue \( \lambda(A) \). It equals the minimum normalized length of a directed cycle.

**Proof.** Let \( \lambda = \lambda(A) \) be the minimum of the normalized length over all directed cycles in \( G(A) \). We first prove that \( \lambda(A) \) is the only possibility for an eigenvalue. Suppose that \( \mathbf{z} \in \mathbb{R}^n \) is any eigenvector of \( A \), and let \( \gamma \) be the corresponding eigenvalue. For any cycle \( (i_1, i_2, \ldots, i_k, i_1) \) in \( G(A) \) we have

\[
a_{i_1 i_2} + z_{i_2} \geq \gamma + z_{i_1}, \quad a_{i_2 i_3} + z_{i_3} \geq \gamma + z_{i_2},
\]
\[
a_{i_3 i_4} + z_{i_4} \geq \gamma + z_{i_3}, \ldots, \quad a_{i_{k-1} i_k} + z_{i_k} \geq \gamma + z_{i_{k-1}}.
\]
Adding the left-hand sides and the right-hand sides, we find that the normalized length of the cycle is greater than or equal to $\gamma$. In particular, we have $\lambda(A) \geq \gamma$. For the reverse inequality, start with any index $i_1$. Since $z$ is an eigenvector with eigenvalue $\gamma$, there exists $i_2$ such that $a_{i_1 i_2} + z_{i_2} = \gamma + z_{i_1}$. Likewise, there exists $i_3$ such that $a_{i_2 i_3} + z_{i_3} = \gamma + z_{i_2}$. We continue in this manner until we reach an index $i_l$ which was already in the sequence, say, $i_k = i_l$ for $k < l$. By adding the equations along this cycle, we find that
\[
(a_{i_k i_{k+1}} + z_{i_{k+1}}) + (a_{i_{k+1} i_{k+2}} + z_{i_{k+2}}) + \cdots + (a_{i_{l-1} i_l} + z_{i_l}) = (\gamma + z_{i_k}) + (\gamma + z_{i_{k+1}}) + \cdots + (\gamma + z_{i_{l-1}}).
\]
We conclude that the normalized length of the cycle $(i_k, i_{k+1}, \ldots, i_l = i_k)$ in $G(A)$ is equal to $\gamma$. In particular, $\gamma \geq \lambda(A)$. This proves that $\gamma = \lambda(A)$.

It remains to prove the existence of an eigenvector. Let $B$ be the matrix obtained from $A$ by (classically) subtracting $\lambda(A)$ from every entry in $A$. All cycles in $G(B)$ have non-negative length, and there exists a cycle of length zero. Using tropical matrix operations we define
\[
B^+ = B \oplus B^2 \oplus B^3 \oplus \cdots \oplus B^n.
\]
This matrix is known as the Kleene plus of the matrix $B$. The entry $B^+_{ij}$ in row $i$ and column $j$ of $B^+$ is the length of a shortest path from node $i$ to node $j$ in the weighted directed graph $G(B)$. Since this graph is strongly connected, we have $B^+_{ij} < \infty$ for all $i$ and $j$.

Fix any node $j$ that lies on a zero length cycle of $G(B)$. Let $x = B^+_{ij}$ denote the $j$th column vector of the matrix $B^+$. We have $x_j = B^+_{ij} = 0$, as there is a path from $j$ to itself of length zero, and there are no negative weight cycles. This implies $B^+ \odot x \leq B^+ = x$. Next note that $(B \odot x)_i = \min_l (B_{il} + x_i) = \min_l (B_{il} + B^+_{ij}) \geq B^+_{ij} = x_i$, since lengths of shortest paths obey the triangle inequality. In vector notation this states $B \odot x \geq x$. Since tropical linear maps preserve coordinatewise inequalities among vectors, we have $B^2 \odot x \geq B \odot x$, and $B^3 \odot x \geq B^2 \odot x$, etc. Therefore, $B^+ \odot x = B \odot x \oplus B^2 \odot x \oplus \cdots \oplus B^n \odot x = B \odot x$. This yields $x \leq B \odot x = B^+ \odot x \leq x$. This means that $B \odot x = x$, so $x$ is an eigenvector of $B$ with eigenvalue $0$. We conclude that $x$ is an eigenvector with eigenvalue $\lambda$ of our matrix $A$:
\[
A \odot x = (\lambda \odot B) \odot x = \lambda \odot (B \odot x) = \lambda \odot x.
\]
This completes the proof of Theorem 7.3.

The eigenvalue $\lambda$ of a tropical $n \times n$-matrix can be computed efficiently. Given a matrix $A = (a_{ij})$, one sets up the following linear program with $n + 1$ decision variables $v_1, \ldots, v_n, \lambda$:
\[
\text{(7.7) Maximize } \gamma \text{ subject to } a_{ij} + v_j \geq \gamma + v_i \text{ for all } 1 \leq i, j \leq n.
\]

Proposition 7.4. The unique eigenvalue $\lambda(A)$ of the given $n \times n$-matrix $A = (a_{ij})$ coincides with the optimal value $\gamma^*$ of the linear program (7.7).

Proof. See [26 Proposition 5.1.2].

We next determine the eigenspace of the matrix $A$, which is the set
\[
\text{Eig}(A) = \{ x \in \mathbb{R}^n : A \odot x = \lambda(A) \odot x \}.
\]
The set \( \text{Eig}(A) \) is closed under tropical scalar multiplication: if \( x \in \text{Eig}(A) \) and \( c \in \mathbb{R} \) then \( c \odot x \) is also in \( \text{Eig}(A) \). We can thus identify \( \text{Eig}(A) \) with its image in the quotient space \( \mathbb{R}^n / \mathbb{R}^1 \simeq \mathbb{R}^{n-1} \). Here \( 1 = (1, 1, \ldots, 1) \). This space is called the tropical projective torus; cf. [20, Section 1.4]. We saw that every eigenvector of the matrix \( A \) is also an eigenvector of the matrix \( B = (-\lambda(A)) \odot A \) and vice versa. Hence the eigenspace \( \text{Eig}(A) \) is equal to \( \text{Eig}(B) = \{ x \in \mathbb{R}^n : B \odot x = x \} \).

**Theorem 7.5.** Let \( B^+_0 \) be the submatrix of the Kleene plus \( B^+ \) given by the columns whose diagonal entry \( B^+_{jj} \) is zero. The image of this matrix (with respect to tropical multiplication of vectors on the right) is equal to the desired eigenspace:

\[
\text{Eig}(A) = \text{Eig}(B) = \text{Image}(B^+_0).
\]

**Proof.** See [26, Theorem 5.1.3]. \( \square \)

**Example 7.6.** We demonstrate the computation of eigenvalues and eigenvectors for \( n = 3 \). In our first example, the minimal normalized cycle lengths are attained by the loops:

\[
A = \begin{pmatrix}
3 & 4 & 4 \\
4 & 3 & 4 \\
4 & 4 & 3
\end{pmatrix} \Rightarrow \lambda(A) = 3 \Rightarrow B = B^+ = B^+_0 = \begin{pmatrix}
0 & 1 & 1 \\
1 & 0 & 1 \\
1 & 1 & 0
\end{pmatrix}.
\]

The eigenspace is the tropical linear span in \( \mathbb{R}^3 \) of the column vectors of \( B \). Its image in \( \mathbb{R}^3 / \mathbb{R}^1 \) is the hexagon with vertices \( (0,1,1) \), \( (0,0,1) \), \( (1,0,1) \), \( (1,0,0) \), \( (1,1,0) \) and \( (0,1,0) \). In our second example, the shortest normalized cycle is the loop between nodes 1 and 2:

\[
A = \begin{pmatrix}
3 & 1 & 4 \\
1 & 3 & 2 \\
4 & 4 & 3
\end{pmatrix} \Rightarrow \lambda(A) = 1 \Rightarrow B = \begin{pmatrix}
2 & 0 & 3 \\
0 & 2 & 1 \\
3 & 3 & 2
\end{pmatrix} \Rightarrow B^+ = \begin{pmatrix}
0 & 0 & 1 \\
0 & 0 & 1 \\
3 & 3 & 2
\end{pmatrix}.
\]

The eigenspace of \( A \) is the tropical linear space spanned by the first column of \( B^+ \):

\[
\text{Eig}(A) = \text{Eig}(B) = \{ c \odot (0,0,3)^T : c \in \mathbb{R} \} = \{ (c, c+3)^T : c \in \mathbb{R} \}.
\]

So, here \( \text{Eig}(A) \) is just a single point in the tropical projective 2-torus \( \mathbb{R}^3 / \mathbb{R}^1 \). \( \diamond \)

We computed the eigenspace of a square matrix as the image of another matrix. This motivates the study of images of tropical linear maps \( \mathbb{R}^m \to \mathbb{R}^n \). Such images are not tropical linear spaces. They are known as tropical polytopes. Indeed, one defines tropical convexity in \( \mathbb{R}^n / \mathbb{R}^1 \) by taking tropical linear combinations. Tropical convexity is a rich and beautiful theory with many applications. For textbook introductions see [20] Chapter 5 and [26] §5.2.

We give a brief illustration in the case \( m = n = 3 \). The image of a \( 3 \times 3 \)-matrix \( X \) is the set of all tropical linear combinations of three vectors in \( \mathbb{R}^3 \). We represent this set by its image in the plane \( \mathbb{R}^3 / \mathbb{R}^1 \). That image is a tropical triangle, because it is the tropical convex hull of three points in the plane. It is possible that this triangle degenerates because three points are tropically collinear in \( \mathbb{R}^3 / \mathbb{R}^1 \). This happens when the minimum in the tropical determinant (7.4) is attained twice. In that case, the matrix \( X \) is called tropically singular.
Example 7.7. Let $T = \text{image}(A)$ be the tropical triangle defined by either of the matrices

$$A = \begin{pmatrix} 0 & 0 & 2 \\ 0 & 3 & 1 \\ 1 & 0 & 0 \end{pmatrix} \quad \text{or} \quad A' = \begin{pmatrix} -1 & 0 & 2 \\ -1 & 3 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$ 

Each point in the quotient $\mathbb{R}^3 / \mathbb{R}1$ can be represented uniquely by a vector $(u, v, 0)^T$ with last coordinate zero. The tropical triangle $T$ consists of the segment between $(-1, -1, 0)^T$ and $(0, 0, 0)^T$, the segment between $(0, 0, 3)^T$ and $(0, 1, 0)^T$, the segment between $(2, 1, 0)^T$ and $(1, 1, 0)^T$, and the classical triangle with vertices $(0, 0, 0)^T$, $(0, 1, 0)^T$ and $(1, 1, 0)^T$.

There are five distinct combinatorial types of tropical triangles in the plane. Similarly, there are 35 types of tropical quadrilaterals in the plane. They are shown in [26] Figure 5.2.4.

7.3. Towards Tropical Varieties

Up to this point, this section has explored the tropical counterparts of concepts from linear algebra. In what follows we move on to nonlinear algebra, and we discuss the tropical counterparts of algebraic varieties. This will also show how the tropical semiring arises naturally from the familiar arithmetic operations over a field $K$.

We fix an algebraically closed field with a valuation, for instance the field $\mathbb{C}$. Here two examples of scalars in $\mathbb{K}$ and their valuations:

$$c = \frac{1}{t^2 + 2t^3 + t^5} = t^{-2} - 2t^{-1} + 4 - 9t + 20t^2 - 44t^3 + 97t^4 - 214t^5 + \cdots$$

has $\text{val}(c) = -2$, while the following scalar has $\text{val}(c') = \frac{2}{7}$:

$$c' = t^{2/7} \sqrt{1 - t^{2/7}} = t^{2/7} - \frac{1}{2} t^{20/21} - \frac{1}{8} t^{34/21} - \frac{1}{16} t^{16/7} - \frac{5}{128} t^{62/21} - \cdots$$

Every polynomial of degree $d$ in $K[x]$ has $d$ distinct roots, counting multiplicities.

Example 7.8 (Puiseux series). Every cubic polynomial in $K[x]$ has three roots, easily found using computer algebra. For instance, the three roots of $f(x) = t x^3 - x^2 + 3tx - 2t^5$ are

$$t^{-1} - 3t - 9t^3 - 54t^5 + 2t^6 - 405t^7 + 18t^8 - 3402t^9 + 180t^{10} - 30618t^{11} + \cdots$$

(7.8) $3t + 9t^3 - \frac{2}{7} t^4 + 54t^5 - 2t^6 + \frac{10931}{2} t^7 - 18t^8 + 3402t^9 - \frac{43756}{21} t^{10} + 30618t^{11} + \cdots$

The valuations of the three roots are $-1, 1$ and $4$. These valuations characterize the asymptotic behavior of the roots when the parameter $t$ is a real number very close to zero.

Consider any polynomial in $n$ variables over the Puiseux series field $K$:

$$f = c_1 x^{a_1} + c_2 x^{a_2} + \cdots + c_n x^{a_n}.$$
The tropicalization of $f$ is the following expression in tropical arithmetic:
\[
\text{trop}(f) = \text{val}(c_1) \odot x^{a_{11}} \odot \text{val}(c_2) \odot x^{a_{21}} \odot \cdots \odot \text{val}(c_s) \odot x^{a_{s1}}.
\]
To evaluate this tropical polynomial at a point $u = (u_1, \ldots, u_n)$, we take the minimum of the $s$ expressions
\[
\text{val}(c_i) \odot u_i^{a_{ii}} = \text{val}(c_i) \odot u_1^{a_{1i}} \odot \cdots \odot u_n^{a_{ni}} = c_i + a_{i1}u_1 + \cdots + a_{in}u_n,
\]
where the index $i$ runs over $\{1, \ldots, s\}$. If this minimum is attained at least twice then we say that $u$ is a tropical zero of $\text{trop}(f)$.

**Proposition 7.9.** If $z = (z_1, \ldots, z_n) \in K^n$ is a zero of a polynomial $f$ in $K[x]$ then its coordinatewise valuation $\text{val}(z) = (\text{val}(z_1), \ldots, \text{val}(z_n)) \in \mathbb{Q}^n$ is a tropical zero of $\text{trop}(f)$.

**Proof.** Note that the valuation of the Puiseux series $c_i z_i^n$ equals $\text{val}(c_i) \odot u_i^{a_{ii}}$. The sum of these $r$ Puiseux series is zero in $K$, so the terms of lowest valuation must cancel. This implies that the minimum valuation is attained by two or more of the expressions $\text{val}(c_i) \odot u_i^{a_{ii}}$. By definition, this means that the vector $u \in \mathbb{Q}^n$ is a tropical zero of $\text{trop}(f)$. \qed

A celebrated result due to Kapranov states that the converse holds as well. Namely, if $f \in K[x]$ and $u \in \mathbb{Q}^n$ is a tropical zero of $\text{trop}(f)$ then there exists a point $z \in K^n$ such that $f(z) = 0$ and $\text{val}(z) = u$. We refer to [26 Theorem 3.1.3] the proof and further details.

**Example 7.10 ($n = 1$).** If $f$ is the cubic polynomial in Example 7.8 then its tropicalization is
\[
\text{trop}(f) = 1 \odot x^{3} \odot 0 \odot x^{2} + 1 \odot x + 5.
\]
The tropical zeros are the rational numbers $x$ such that the minimum of $1+3x$, $0+2x$, $1+x$ and $5$ is attained twice. There are three solutions: $x = -1$, $x = 1$ and $x = 4$. Each of these is the valuation of an element in $K$ that is a zero of $f$. These solutions are listed in (7.8).

The extra element $+\infty$ arises naturally from the arithmetic in a field $K$ with valuation because $\text{val}(0) = \infty$. Sometimes it is preferable to restrict tropical algebra to $\mathbb{R}$, or to $\mathbb{Q}$, thus excluding $+\infty$. This is accomplished by disallowing zero coordinates among the solutions of a polynomial equation. To be precise, we set $K^* = K \setminus \{0\}$ and we introduce the algebraic torus $(K^*)^n$. The ring of polynomial functions on $(K^*)^n$ is the Laurent polynomial ring
\[
K[x^\pm] := K[x_1^{\pm 1}, x_2^{\pm 1}, \ldots, x_n^{\pm 1}].
\]
Its elements are polynomials as in (7.9) but we now allow negative integers among the coordinates of the exponent vectors $a_i$. For every $u \in \mathbb{R}^n$, the initial form $\text{in}_u(f)$ is the subsum of terms $\tau_j x^{a_j}$ in (7.9) for which $\text{val}(c_j) \odot u^{a_j}$ is minimal. Here $\tau_j$ is the term of lowest order in the Puiseux series $c_j$. For instance, if $c$ is the scalar in the middle line in (7.8) then $\tau = 3t$.

**Lemma 7.11.** For any Laurent polynomial $f \in K[x^\pm]$ and any point $u \in \mathbb{R}^n$, the following three conditions are equivalent:
- The initial form $\text{in}_u(f)$ is a unit in $K[x^\pm]$.
- The initial form $\text{in}_u(f)$ is a not monomial.
- The point $u$ is a tropical zero of $\text{trop}(f)$.
7.3. TOWARDS TROPICAL VARIETIES

Fix any ideal $I$ in $K[x^k]$ and let $\mathcal{V}(I)$ be its variety in the algebraic torus $(K^*)^n$. We define the tropical variety of $I$ to be the following subset of $\mathbb{R}^n$:

$$\text{trop}(\mathcal{V}(I)) = \{ u \in \mathbb{R}^n : u \text{ is a tropical zero of trop}(f) \text{ for all } f \in I \}.$$

We also refer to this set as the tropicalization of the variety $\mathcal{V}(I)$.

The study of tropical varieties is the subject of tropical algebraic geometry. Two important results are the Fundamental Theorem ([26 Theorem 3.2.3]) and the Structure Theorem ([26 Theorem 3.3.5]). The former extends the theorem of Kapranov mentioned above. It states that the set of rational points in trop($\mathcal{V}(I)$) is the image of the classical variety $\mathcal{V}(I) \subset (K^*)^n$ under the coordinatewise valuation map. The latter states that trop($\mathcal{V}(I)$) is a balanced polyhedral complex, whose dimension agrees that the dimension of $\mathcal{V}(I)$. Numerous concrete examples of such polyhedral complexes are found in the books [20] and [26].

**Example 7.12.** Fix $n = 9$ and let $x = (x_{ij})$ be a $3 \times 3$-matrix whose entries are unknowns. Let $I$ be the ideal generated by the nine $2 \times 2$-minors of $x$. Then $\mathcal{V}(I)$ is the 5-dimensional variety of $3 \times 3$-matrices of rank 1 in $(K^*)^{3 \times 3}$. The tropical variety $\text{trop}(\mathcal{V}(I))$ is the set of $3 \times 3$-matrices in $[7.1]$, that is, the matrices of tropical rank one. This is the linear subspace of dimension 5 in $\mathbb{R}^{3 \times 3}$ defined by the tropical $2 \times 2$-determinants $u_{ij} \odot u_{kl} \oplus u_{ik} \odot u_{kj}$. Of course, this minimum is attained twice if and only if $u_{ij} + u_{kl} - u_{ik} - u_{kj} = 0$. Every matrix $u = (u_{ij})$ that satisfies these linear equations, and has its entries in $\mathbb{Q}$, arises as the valuation $u = \text{val}(u)$ of a rank one matrix $z = (z_{ij})$ with entries in $K^*$.

The situation becomes more interesting when we pass from rank 1 to rank 2. Let $J$ be the principal ideal generated by the determinant of $x$. The $\mathcal{V}(J)$ is a hypersurface of degree three in $(K^*)^{3 \times 3}$. The tropical hypersurface trop($\mathcal{V}(J)$) is defined by the tropical determinant

$$\text{tropdet}(u) = u_{11} \odot u_{22} \odot u_{33} \oplus u_{11} \odot u_{23} \odot u_{32} \oplus \cdots \oplus u_{13} \odot u_{22} \odot u_{31}.$$  

Thus trop($\mathcal{V}(J)$) is set of all $3 \times 3$-matrices $u = (u_{ij})$ such that this minimum is attained twice. For such a matrix, there is more than one optimal assignment of the three workers to the three jobs in (7.4). The set trop($\mathcal{V}(J)$) is a polyhedral fan of dimension 8. It is a cone with apex trop($\mathcal{V}(J)$) $\simeq \mathbb{R}^5$ over the 2-dimensional polyhedral complex shown in Figure 4.

The six triangles represent matrices $u$ where the minimum in (7.10) is attained by two permutations in $S_3$ that have the same sign. The nine squares on the right in Figure 4 are glued to form a torus. These represent matrices $u'$ where the minimum in (7.10) is attained by two permutations in $S_3$ that have opposite signs. Concrete examples for the two cases are

$$u = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad u' = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$  

Here are classical matrices of rank 1 that map to $u$ and $u'$ under tropicalization:

$$z = \begin{pmatrix} t + 1 & -1 + t & 2t \\ t & 1 & 1 + t \\ 1 & t & 1 + t \end{pmatrix} \quad \text{and} \quad z' = \begin{pmatrix} 1 & 2 & t \\ 2 & 4 & 5 \\ 3t & 6t & 7 \end{pmatrix}.$$  

The supports of the matrices $u = \text{trop}(z)$ match the labels of the corresponding 2-cells in Figure 4. The matrix $u$ has support 13, 21, 32, which labels the bottom
triangle on the left. The matrix \( u \) has support 13, 23, 31, 32, which labels the middle left square on the right.

We close with a remark on lifting Proposition 7.1 from tropical algebra to algebra over the field \( K \). Given a directed graph \( G \) with rational edge weights \( d_{ij} \), we now define a new adjacency matrix \( A_G \). The entry of \( A_G \) in row \( i \) and column \( j \) equals \( t^{ij} \) if \( (i, j) \) is an edge of \( G \), and 0 otherwise. By construction, the valuation of the matrix \( A_G \) is the earlier adjacency matrix \( D_G \). Moreover, the matrix in (7.2) is the valuation of the classical matrix power of \( A_G \):

\[
D_G^{\odot(n-1)} = (\text{val}(A_G))^{\odot(n-1)} = \text{val}(A_G^{n-1}).
\]

Indeed, the \((i, j)\) entry of \( A_G^{n-1} \) is the generating function for all paths. To be precise, it is the Puiseux polynomial \( \sum_{\ell} c_{\ell} t^\ell \), where \( c_{\ell} \) is the number of paths from \( i \) to \( j \) having length \( \ell \).

**Exercises**

1. Let \( u, v, w \) be real numbers and let \( x, y, z \) be variables. What are the coefficients in the expansion of the expression \((u \odot x \oplus v \odot y \oplus w \odot z)^{\odot n}\) in tropical arithmetic?
2. Prove that the tropical multiplication of square matrices is an associative operation.
3. Draw the graph of the function \( \mathbb{R} \to \mathbb{R} \), \( x \mapsto 1 \oplus 2 \odot x \oplus 3 \odot x^\odot 2 \oplus 6 \odot x^\odot 3 \oplus 10 \odot x^\odot 4 \). What are the tropical zeros of this tropical polynomial?
4. How would you define the tropical characteristic polynomial of a square matrix? Compute your characteristic polynomial for the 3 \( \times \) 3-matrices in Example 7.6.
5. Draw the graph of the function \( \mathbb{R}^2 \to \mathbb{R} \), \( (x, y) \mapsto 1 \oplus 2 \odot x \oplus 3 \odot y \oplus 6 \odot xy \oplus 10 \odot xy^\odot 2 \). What are the tropical zeros of this tropical polynomial?
(6) Let $G$ be the directed graph on $n$ nodes with edge weights $d_{ij} = i \cdot j$ for $i, j \in \{1, 2, \ldots, n\}$. Compute the tropical powers $D_G^{\circ i}$ of the matrix $D_G$ for $i = 1, 2, \ldots, n-1$. What are their tropical ranks? Interpret the entries of these matrices in terms of paths.

(7) Take the graph $G$ from above with $n = 5$. Compute the powers $A_G^i$ of the matrix $A_G$ for $i < n$. What are their ranks? Interpret the entries in terms of paths. Verify equation (7.11).

(8) Take the graph $G$ from above with $n = 3$. Find the eigenvalues and eigenspaces of $A_G$. Find the tropical eigenvalue and the tropical eigenspace of $D_G$. Do you see a relationship?

(9) Take the graph $G$ from above with $n = 10$. Compute the determinant of $A_G$ and the tropical determinant of $D_G$. Do you see a relationship? Can you generalize to arbitrary $n$?

(10) Take the graph $G$ from above. The matrix $D_G$ defines a tropical linear map from $\mathbb{R}^n$ to itself. Determine the image of this map for $n = 2, 3, 4$. Draw pictures in $\mathbb{R}^n / \mathbb{R}1 \simeq \mathbb{R}^{n-1}$.

(11) Consider the quartic polynomial $f(x) = t + t^2x + t^3x^2 + t^6x^3 + t^{10}x^4$ in $K[x]$. Identify its four roots. Write the first 10 terms of these Puiseux series. What are their valuations?

(12) Let $J$ be the ideal generated by the determinant of a symmetric $3 \times 3$-matrix. This lives in a Laurent polynomial ring with six variables. Determine the tropical hypersurface $\text{trop}(\mathcal{V}(J))$. Write a discussion analogous to Example 7.12. Draw the analog to Figure 1.

(13) Analyze the complexity of the algorithm described in Proposition 7.1. How would you improve the computation of $D_G^{\circ (n-1)}$? What happens if some weights of the edges of $G$ are negative? What happens if the graph contains cycles of negative total weight? How would you detect if such a cycle exists?
Toric varieties are arguably the simplest and most accessible varieties. They often appear in applications, both within mathematics and across the sciences. A toric variety is an irreducible variety that is parametrized by a vector of monomials. The relations among these monomials are binomials, i.e. polynomials with only two terms. Thus, an irreducible variety is toric if and only if its prime ideal is generated by binomials. Monomials and binomials correspond to points in an integer lattice, and we think of these as the vertices of a lattice polytope. Toric varieties appear prominently in optimization and statistics, thanks to the purely combinatorial description given above. This description also makes them a perfect “model organism” for algebraic geometers. They use toric varieties to test conjectures, teach geometric concepts, and compute invariants. For instance, the dimension and degree of a toric variety are the dimension and volume of the associated lattice polytope.

8.1. Tori and Lattices

In this section we introduce the algebraic torus and describe its relations to lattices. The set $(\mathbb{K}^\times)^n = \text{Spec} \mathbb{K}\left[ x_1^{\pm 1}, \ldots, x_n^{\pm 1} \right]$ is known as the algebraic torus. It is an algebraic group with the action given by coordinatewise multiplication. If $n = 2$ and $\mathbb{K} = \mathbb{C}$ then the algebraic torus $(\mathbb{C}^\times)^2 \simeq (\mathbb{R}_+ \times S^1)^2$ coincides with the usual topological torus $S^1 \times S^1$ up to multiplication with the contractible factor $\mathbb{R}_+^2$.

**Definition 8.1 (Character of a torus).** A character of a torus $T = (\mathbb{K}^\times)^n$ is an algebraic map $T \to \mathbb{K}^\times$ that is also a group morphism.

In Exercise 1 the reader is asked to prove that all characters are given by Laurent monomials. The characters of $T$ are hence the elements of $\mathbb{Z}^n$. Characters of a torus $T$ can be identified with $\mathbb{Z}^n$ not only as a set but also as a group $(\mathbb{Z}^n, +)$ with the action given by:

$$(\chi_1 + \chi_2)(t) := \chi_1(t)\chi_2(t).$$

A group isomorphic to $\mathbb{Z}^n$ is called a lattice. The lattice of characters of $T$ will be denoted by $\tilde{M}_T$ or simply $\tilde{M}$. As a subgroup of a free abelian group is free, a finite set of characters generates a sublattice $\tilde{M} \subset M$.

**Proposition 8.2.** Fix characters $a_1, \ldots, a_N \in M_T$ generating a sublattice $\tilde{M}$. The image of $T$ in $(\mathbb{K}^\times)^N$ by the map $x \to (x^{a_1}, \ldots, x^{a_N})$ is also a torus $\tilde{T}$ with the character lattice equal to $\tilde{M}$.

**Proof.** The monomial map $f : T \to (\mathbb{K}^\times)^N$ induces the ring homomorphism

$$f^* : K[y_1^{\pm 1}, \ldots, y_N^{\pm 1}] \to K[x_1^{\pm 1}, \ldots, x_n^{\pm 1}], \ y_i \mapsto x^{a_i}.$$ 

The spectrum of the image of the ring map $f^*$ is the image $\tilde{T}$ we are interested in. Note that $\text{im} f^*$ equals the group algebra $K[\tilde{M}]$. By definition, this is the vector
space over $K$ with basis given by elements of $\tilde{M}$ and multiplication induced from addition in $M_T$. The lattice $\tilde{M}$ is isomorphic to the group $\mathbb{Z}^d$ for some $d \in \mathbb{N}$. We have $\tilde{T} = \text{Spec } K[\tilde{M}] = (K^*)^d$. The associated toric variety in $K^N$ is the Zariski closure of $\tilde{T}$. □

8.2. Affine Toric Varieties

We start directly with a definition of a toric variety. Recall that nonnegative integer vectors $b = (b_1, \ldots, b_n)$ are identified with monomials $x^b := x_1^{b_1} x_2^{b_2} \cdots x_n^{b_n}$.

In the following definition we allow $b \in \mathbb{Z}^n$ to have negative entries. This means that $x^b$ is a Laurent monomial, i.e. a monomial with possibly negative exponents.

**Definition 8.3 (Affine toric variety).** An affine toric variety is the closed image of a monomial map $(K^*)^n \rightarrow K^N$, $x \mapsto (x^{a_1}, x^{a_2}, \ldots, x^{a_N})$, where $a_i \in \mathbb{Z}^n$ and $K^* = K \setminus \{0\}$.

To specify a toric variety, we need to specify $N$ characters of a torus, equivalently $N$ integer points in $\mathbb{Z}^n$. Toric geometry relates the geometric properties of a toric variety $X$ with combinatorics of a finite set of lattice points defining $X$.

**Example 8.4.**

(1) Any affine space is a toric variety.
(2) The cuspidal cubic curve $x^3 - y^2$ is a toric variety. It is the image of the monomial map $z \mapsto (z^2, z^3)$.

**Proposition 8.5.** Let us fix characters $a_1, \ldots, a_N \in \mathbb{Z}^n = M$, generating a sublattice $\tilde{M}$. The dimension of the associated toric variety equals the rank of $\tilde{M}$.

**Proof.** By Proposition we know that the image of $(\mathbb{C}^*)^n$ is a torus of dimension equal to the rank of $\tilde{M}$. As the Zariski closure does not change the dimension, the toric variety is also of the same dimension. □

We see that we may equivalently define toric varieties as closures of a subtorus of the torus $(K^*)^N \subset K^N$. Further, in analogy to the proof of Proposition 8.2 we see that the toric variety equals Spec $K[S]$, where $S$ is the monoid in $M_T$ generated by the distinguished characters, i.e. the smallest set containing 0, the chosen characters and closed under addition.

**Example 8.6.**

(1) The cuspidal curve defined by the equation $x^3 - y^2$ equals Spec $K[z^2, z^3]$. Here, the associated monoid equals $\{0, 2, 3, 4, \ldots\}$.
(2) The affine line is the closure of the image of the map $K^* \ni x \rightarrow x \in K$.

Here the character lattice is $M = \mathbb{Z}$, the distinguished character corresponds to $1 \in M$ and the monoid equals $\{0, 1, 2, \ldots\}$.

There is a fundamental difference between the example of the cuspidal curve and affine line. When we look at the monoid for the cuspidal curve, there is a 'hole' in it: the character corresponding to 1.
8.2. AFFINE TORIC VARIETIES

Definition 8.7. A submonoid $S$ in a lattice $M$ is called saturated if and only if for any $x \in M$ and $k \in \mathbb{Z}_+$ the following implication holds:

$$kx \in S \Rightarrow x \in S.$$ 

Affine toric varieties for which $S$ is saturated (in the lattice $\tilde{M}$ that it generates) are called normal. For the algebraic definition of normal varieties we refer to [1, Chapter 5]. Nonnormal varieties are always singular and for curves the two notions coincide. Hence, Example 8.6 shows one nonnormal (equivalently singular) curve - as seen in Figure 1 - and one normal (equivalently smooth) curve.

Further, we can find the generators of the ideal of the variety $X$ from the characters that define it. In general, given a variety defined as a Zariski closure of the image of a map, finding the defining equations is a hard problem, known as implicitization. We discussed this in Chapter 4. The implicitization problem greatly simplifies when the variety is toric. Recall that a binomial is a polynomial that is a difference of two monomials.

Lemma 8.8. Let $X$ be the toric variety defined by $a_1, \ldots, a_N \in \mathbb{Z}^n$. Then:

1. any relation $\sum_i b_i a_i = \sum_j c_j a_j$, with positive integral coefficients $b_i, c_j \in \mathbb{Z}_+$ provides a binomial $\prod y_i^{b_i} - \prod y_j^{c_j}$ in the prime ideal $I_X$ of $X$;
2. every binomial in the ideal $I_X$ is of the form described in point 1;
3. the ideal $I_X$ is generated by binomials.

Sketch of the proof: Properties 1 and 2 follow from the fact that a polynomial vanishes on the toric variety $X$ if and only if we obtain zero after substituting $y_i$ by $x^{a_i}$. However, such a substitution turns monomials (in variables $y$) to monomials (in variables $x$). The fact that the monomials in $x$ cancel is precisely encoded by the integral relations in point 1. Property 3 follows similarly, by induction on the support of a polynomial in the ideal of $X$. □

Example 8.9. Let $n = 3, N = 7$ and take $a_1, \ldots, a_7$ to be the column vectors of the matrix

$$A = \begin{pmatrix} 2 & 2 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 2 & 2 & 1 \\ 0 & 1 & 2 & 2 & 1 & 0 & 1 \end{pmatrix}$$
The associated toric variety $X$ is a threefold in $K^7$. Its prime ideal $I_X$ equals
\[
\langle y_1y_3 - y_2y_7, y_1y_4 - y_2^2, y_1y_5 - y_6y_7, y_2y_4 - y_3y_7, y_2y_5 - y_2^2, \\
y_2y_6 - y_1y_7, y_3y_5 - y_4y_7, y_3y_6 - y_2^2, y_4y_6 - y_2^2 \rangle
\]
Since the nine binomial generators of $I_X$ are homogeneous, the variety $X$ is a cone in $K^7$. It can thus also be regarded as projective variety in $\mathbb{P}^6$. That projective toric variety is a smooth surface of degree six. The reader is invited to check this.

**Theorem 8.10.** A prime ideal generated by binomials defines a toric variety.

**Proof.** This follows from the fact that binomials may be translated to Laurent monomials on $(K^*)^N$, where they have to define a torus. For details see [9, Proposition 1.1.11].

**Definition 8.11.** A convex polyhedral cone in a real vector space $V$ is a subset of elements of the form $\lambda_1v_1 + \cdots + \lambda_kv_k$ for some fixed integer $k$, $v_1, \ldots, v_k \in V$ and $\lambda_1, \ldots, \lambda_k \in \mathbb{R}_{\geq 0}$. If we identify $V$ with $\mathbb{R}^d$ we call a convex polyhedral cone rational if all the $v_i$’s can be chosen as rational vectors.

We will refer to rational convex polyhedral cones simply as cones.

In Exercise 4 the reader is asked to show that a finitely generated, saturated monoid in a lattice $\mathbb{Z}^n$ is the same as the set of integral points in a cone in the corresponding $\mathbb{R}^n$.

**Definition 8.12.** A face of a cone $C \subset V$ is a subset $F \subset C$ defined by:
\[
F = \{c \in C : f(c) = 0\},
\]
where $f$ is such a linear function $f \in V^*$ that for any $p \in C$ we have $f(c) \geq 0$. If $\dim C = \dim V = \dim F + 1$, then $f$ is uniquely determined, up to scalar. In such a case $F$ is called a facet and the hyperplane defined by $f$ is called a supporting hyperplane of $C$. We point out that if $f = 0$ then we obtain $F = C$. Further, any face of a cone is also a cone.

**Example 8.13.** Let $C$ be the positive quadrant in $\mathbb{R}^2$. It has one 2-dimensional face - the whole cone, two 1-dimensional facets and one 0-dimensional face $\{0\} \subset C$.

By Proposition 8.2 a toric variety $X$ is a closure of the torus
\[
T = \{t \in X : \text{all coordinates of } t \text{ are nonzero} \} \subset X \subset \mathbb{C}^N.
\]
As $T$ is a group that acts both on itself and $\mathbb{C}^n$, it must also act on $X = \overline{T}$. Our next aim is to provide a combinatorial and geometric description of the orbits of this action.

Let us make the following assumptions. The toric variety $X$ is defined by characters $\lambda$ that generate a saturated monoid. Let $C$ be the corresponding cone and $T \subset X$ the torus dense in $X$.

**Theorem 8.14.** Using the above notation, the $T$-orbits in $X$ are in bijection with the faces of the cone $C$. The orbit corresponding to a face $F$ consists exactly of those points $x \in X$ that have a nonzero coordinate corresponding to a character $\lambda \in A$ if and only if $\lambda \in F$. Further, the closure of the orbit corresponding to $F$ is the toric variety $\text{Spec } \mathbb{C}[F \cap A]$, where formally $F \cap A$ represents the monoid generated by $F \cap A$. In particular, the dimension of $F$ equals the dimension of the orbit. Moreover, an orbit corresponding to face $F_1$ belongs to the closure of the orbit corresponding to face $F_2$ if and only if $F_1 \subset F_2$. 

Example 8.15. Consider the toric variety associated to characters $(1, 0, 0)$, $(1, 0, 1)$, $(1, 1, 0)$, $(1, 1, 1)$. It is the affine cone over the quadric $xt - yz$. The four 2-dimensional facets of the cone correspond to four two dimensional tori contained in it. For example the face generated by $(1, 0, 0), (1, 0, 1)$ corresponds to the set of points of the type $(\ast, \ast, 0, 0)$, where $\ast$ are nonzero. The four 1-dimensional faces correspond to coordinate axis (minus $\{0\}$). Note that the intersection of faces in the cone and intersection of the corresponding closures of orbits agree.

As we can see, the geometry of $X$ is read off from the cone $C$ representing it.

8.3. Projective Toric Varieties

Let us now pass to projective toric varieties. We stress they relations to polytopes and present the construction of the moment map.

The definition of a projective toric variety is completely analogous to 8.3.

Definition 8.16 (Projective toric variety). A projective toric variety is the closed image of a monomial map

$$(K^*)^n \to \mathbb{P}(K^N), \ x \mapsto [x^{a_1} : x^{a_2} : \cdots : x^{a_N}],$$

where $a_i \in \mathbb{Z}^n$ and $K^* = K \setminus \{0\}$.

Example 8.17.

- Any projective space is a toric variety;
- More generally, Veronese reembeddings and Segre products of projective spaces are toric varieties.

We note that given a set $A$ of $N$ monomials, we obtain the same projective variety if we multiply every monomial by a new variable $x_0$. In many aspects such a description is better, as it also parameterizes the affine cone over the projective variety. Thus, instead of working in lattice $\mathbb{Z}^n$, we will be working in the lattice $\mathbb{Z}^{n+1} = \mathbb{Z} \times \mathbb{Z}^n$ assuming that the defining set of characters/monomials $A$ is contained in $\{1\} \times \mathbb{Z}^n$.

Definition 8.18. A polytope in a vector space $V$ is a convex hull of a finite set of vectors. A polytope is called a lattice polytope if it is a convex hull of points of a lattice $M \subset V$.

If we want $A$ to generate a saturated monoid a necessary condition is that $A = \text{conv}(A) \cap \tilde{M}$ where $\tilde{M}$ is the lattice generated by $A$. In other words, $A$ is the set of integral points of an integral polytope. However, in general this is not enough.

Definition 8.19. A lattice polytope $P$ (in a lattice $M$) is called normal if and only if for any integer $k$ and any point $p \in kP \cap M$ there exist $p_1, \ldots, p_k \in P \cap M$ such that $p = \sum_{i=1}^k p_i$.

Exercise 5 gives an example of a non-normal polytope. In Exercise 6 the reader is asked to prove that a lattice polytope $P$ is normal if and only if $(\{1\} \times P) \cap M$ generates a saturated monoid. The orbit cone correspondence from Theorem 8.14 in a trivial way generalizes to projective toric varieties and polytopes. As we will see below there is another way to explain why the geometry of the polytope coincides with the geometry of the toric variety. Our aim is to define a map, called the moment map, that takes the toric variety $X$ onto the associated polytope $P$. 
Let \( X \subset \mathbb{P}(\mathbb{C}^N) \) be a toric variety defined by a set of characters \( A \subset \mathbb{Z}^n \). In particular, \(|A| = N\) and the coordinates of \( \mathbb{C}^N \) correspond to elements of \( A \). For a point \( y \in \mathbb{C}^N \) and \( a \in A \) we denote by \( a(y) \in \mathbb{C} \) the coordinate of \( y \) corresponding to \( a \). In other words \( y = (y_a)_{a \in A} \).

**Definition 8.20.** The algebraic moment map \( \mu_A : X \to \mathbb{R}^n \) is defined by:

\[
\mu_A(x) = \frac{\sum_{a \in A} |a(x)| a}{\sum_{a \in A} |a(x)|}.
\]

Here, as \( x \in \mathbb{P}(\mathbb{C}^N) \), the value \( a(x) \) is defined only up to a scalar. However, as \( \mu_A(x) \) is a fraction it does not depend on the choice of the scalar.

The numerator in the above definition is a nonnegative combination of integral points \( A \) defining \( X \). The denominator assures that \( \mu_A(x) \in \text{conv}(A) \). Consider a torus fixed point \( x_0 \in X \). By Theorem 8.14 it must have all coordinates equal to zero, apart from one, corresponding to a vertex \( a_0 \in A \) of \( \text{conv} A \). In particular, \( \mu_A(x_0) = a_0 \). Our aim is to present a vast generalization of the above fact, which explains why the geometry of \( X \) is related to the geometry of \( \text{conv}(A) \).

**Definition 8.21.** For a set of characters \( A \), we define the nonnegative part of the associated toric variety \( X \subset \mathbb{C}^{|A|} \) as \( X_{\geq 0} := X \cap \{ (\mathbb{R}_{\geq 0})^{\text{dim}} \} \). Similarly, the positive part of the toric variety is the semialgebraic set \( X > 0 := X \cap \{ (\mathbb{R}_{> 0})^{\text{dim}} \} \).

Let \( T \to \mathbb{C}^{|A|} \) be the map defining \( X \). By Proposition 8.2, \( T \) maps surjectively to a torus \( \hat{T} \) that can be identified with those points of \( X \) that have all coordinates nonzero. Further, \( T_{> 0} \) maps surjectively to \( X_{> 0} = \hat{T}_{> 0} \). This is especially useful in statistics, where our defining map can be interpreted as a statistical model and coordinates as probabilities - cf. Chapter 2, Example 2.5 and many more examples in [39], Chapter 5, Chapter 14]. More generally, we have a map \( X \to X_{\geq 0} \) given by \( r : (x_1, \ldots, x_{|A|}) \to (|x_1|, \ldots, |x_{|A|}|) \). Note that \( \hat{T} = (\mathbb{C}^*)^d \) contains a topological torus \( S_d \), by taking points with coordinates of module one. Further, \( S_d \) is a subgroup of \( \hat{T} \) that acts transitively on each fiber of \( r \). Thus \( \hat{T} \) may be regarded as a quotient map \( X \to X/(S_d) = X_{\geq 0} \). Hence \( r : X \to X_{\geq 0} \) has fibers that are real tori, with dimension equal to the dimension of the orbit of \( \hat{T} \) they belong to. A formal statement and a proof can be found for example in [9], Proposition 12.2.3]. We have now related the geometry of \( X \) with the geometry of \( X_{\geq 0} \). We note that we can make the same definitions when \( X \) is projective, where a point is positive if and only if it has a positive representative.

**Theorem 8.22.** Let \( A \) be the set of lattice points in a lattice polytope \( P \subset \mathbb{R}^n \) and let \( X \) be the associated toric variety. The moment map: \( \mu_A : X_{\geq 0} \to \mathbb{R}^n \) is a homeomorphism onto \( P \).

The proof, along with other interesting facts, is found in [37], Theorem 8.4].

**Example 8.23.** Let us continue the statistically motivated Example 2.5 from Chapter 2 in the case \( n = 2 \). We obtain the Segre embedding:

\[
\mathbb{P}^1 \times \mathbb{P}^1 \to \mathbb{P}^3,
\]

where our toric variety is represented as a unit square and is defined as a quadric \( xt - yz \). If we consider the affine set \( \mathbb{R}^3 \subset \mathbb{P}^3 \) defined by \( x + y + z + t = 1 \), then the moment map, restricted to \( X_{\geq 0} \), becomes simply a linear projection \( \mu : \mathbb{R}^3 \to \mathbb{R}^2 \):

\[
[x : y : z : t] \to x(0, 0) + y(1, 0) + z(0, 1) + t(1, 1) = (y + t, z + t).
\]
Below we present the picture of $X_{\geq 0}$ in coordinates $y, z, t$. The red line shows the direction of the projection of the moment map:

If we rotate the picture so that the red line becomes (nearly) a point, we see that the projection is indeed a square:

**Example 8.24.** Suppose we throw two (possibly biased) coins 1024 times (each time two coins at once) and observe:

- 128 times both heads,
- 128 times the first coin gives heads, the second tails,
- 384 times both tails,
- 384 times the first coin gives tails, the second heads.

This can be translated to:

- $1/8$ times $(0, 0)$,
- $1/8$ times $(0, 1)$,
- $3/8$ times $(1, 1)$,
- $3/8$ times $(1, 0)$.

The data can be represented as a point in the square:

$$p := \frac{1}{8}(0, 0) + \frac{1}{8}(0, 1) + \frac{3}{8}(1, 1) + \frac{3}{8}(1, 0) = (3/4, 1/2).$$

The unique preimage of the point $p$ by the moment map, as in Example 8.23, has coordinates $(y, z, t) = (1/8, 3/8, 3/8)$ and translates to $(x, y, z, t) = (1/8, 1/8, 3/8, 3/8)$. 
Looking at the associated monomial map,
\[(a, b, c, d) \mapsto (ac, ad, bc, bd),\]
under the assumption \(a + b = c + d = 1\), we obtain \(a = 1/4, b = 3/4, c = d = 1/2\).
This is the correct estimate of the probability distribution for the coins: the first coin is biased (with probability of tails 3/4 and heads 1/4) and the second is fair.

The previous example is a very special case of a general theorem in algebraic statistics. Toric varieties correspond to discrete statistical models. The inverse image of any point \(p\) in the polytope, by the moment map, is known as the Birch point or Maximum Likelihood Estimator. For further reading on toric models see [30, Section 1.2.2]. For a more theoretical application of toric varieties notice that:

- \(\mathbb{P}^n\) has a representation as a(n \(n\)-dimensional) simplex;
- the product \(\mathbb{P}^{a_1} \times \cdots \times \mathbb{P}^{a_n}\) has a representation as a product of simplices;
- more generally, a product of projective toric varieties represented by polytopes is the toric variety represented by the product of the polytopes.

As an application of toric geometry we see that a product of projective spaces is a projective variety, naturally embedded in another projective space. Hence, if we are given (possibly nontoric) projective algebraic varieties \(X \subset \mathbb{P}^a, Y \subset \mathbb{P}^b\), we see that the product \(X \times Y \subset \mathbb{P}^a \times \mathbb{P}^b\) is also a projective variety. Notice however, that the natural embedding is not in \(\mathbb{P}^{a+b}\), as one could expect from the affine case. In fact \(\mathbb{P}^a \times \mathbb{P}^b \subset \mathbb{P}^{ab+a+b}\).

**Exercises**

1. Prove that every character \((\mathbb{C}^*)^n \rightarrow \mathbb{C}^*\) is given by \(x \rightarrow x^a\) for some \(a \in \mathbb{Z}^n\).
2. Prove that every polynomial in the ideal of an affine toric variety is a linear combination of binomials - cf. point 3. in Lemma [8.8]
3. Describe the ideals of the Segre product \(\mathbb{P}^{a_1} \times \cdots \times \mathbb{P}^{a_n}\) and of the (arbitrary) Veronese reembeddings of \(\mathbb{P}^a\).
4. Prove that for a fixed lattice \(\mathbb{Z}^d \subset \mathbb{R}^d\) there is a natural bijection between (convex, rational, polyhedral) cones (in \(\mathbb{R}^d\)) and finitely generated saturated monoids (in \(\mathbb{Z}^d\)).
5. Prove that the convex hull of points \((0, 0, 0), (0, 1, 0), (1, 0, 0), (1, 1, 2), (1, 1, 3)\) is not a normal polytope (in the lattice \(\mathbb{Z}^3\)).
6. Prove that a lattice polytope \(P\) is normal if and only if \((\{1\} \times P) \cap M\) generates a saturated monoid.
7. (a) The \(f\)-vector \((f_0, \ldots, f_m) \in Z^{m+1}\) for an \(m\)-dimensional polytope \(P\) is a sequence of positive integers, where \(f_i\) equals the number of \(i\) dimensional faces of \(P\). Compute the number of points of a projective toric variety \(X\) defined by a lattice polytope \(P\) over a finite field, in terms of the \(f\)-vector.
   (b) * Assuming \(X\) is smooth, use the Weil conjectures (now proved due to work of Grothendieck and Deligne), to give a formula for Betti numbers of \(X\), again in terms of the \(f\)-vector.
8. Prove that for any lattice polytope \(P\) in \(\mathbb{R}^d\), the polytope \((d-1)P\) is normal.
9. Prove the following theorem due to Mumford in the case of toric varieties; Let \(X\) be a projective toric variety. For \(r\) large enough the \(r\)-th Veronese reembedding \(v_r(X)\) of \(X\) is defined by quadratic equations.
Tensors

Tensors are ubiquitous in many different branches of modern mathematics. They are higher dimensional analogs of matrices. Just as matrices are basic objects in linear algebra, tensors are fundamental for nonlinear algebra. The only reason they appear so late in this book, is that before they were appearing in disguise: homogeneous polynomials are special tensors. As we will see basic properties of matrices like rank or the set of eigenvectors can be defined also for tensors. However, their behavior is far more interesting. We will present applications of tensors, focusing on a central algorithmic problem: how fast can one multiply two matrices? As always, linear algebra is our entry point to nonlinear algebra. But, also, the new nonlinear tools will be applied to answer fundamental questions in linear algebra.

9.1. Eigenvectors

In this section we show how the concepts of eigenvectors, rank and singular values an be extended from the case of matrices to tensors.

Let us start by reviewing some basics of linear algebra, beginning with the study of symmetric matrices. Symmetric \( n \times n \) matrices are important in statistics where they encode the covariance structure of a joint distribution of \( n \) random variables. In an algebraic setting, symmetric matrices are important because they uniquely represent quadratic forms.

For instance, consider the following quadratic form in three variables \( x, y \) and \( z \):

\[
Q = 2x^2 + 7y^2 + 23z^2 + 6xy + 10xz + 22yz.
\]

This quadratic form is represented uniquely by a symmetric \( 3 \times 3 \)-matrix, as follows:

\[
Q = \begin{pmatrix} 2 & 3 & 5 \\ 3 & 7 & 11 \\ 5 & 11 & 23 \end{pmatrix}
\begin{pmatrix} x \\ y \\ z \end{pmatrix}
\]

The gradient of the quadratic form \( Q \) is a vector of linear forms. It defines a linear map from \( \mathbb{R}^3 \) to itself. Up to multiplication by 2, this is the map one associates with a square matrix:

\[
\nabla Q = \begin{pmatrix} \frac{\partial Q}{\partial x} \\ \frac{\partial Q}{\partial y} \\ \frac{\partial Q}{\partial z} \end{pmatrix} = 2 \cdot \begin{pmatrix} 2 & 3 & 5 \\ 3 & 7 & 11 \\ 5 & 11 & 23 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}
\]

We call \( v \in \mathbb{R}^n \) an eigenvector of \( Q \) if \( v \) is mapped to a scalar multiple of \( v \) by the gradient map:

\[
(\nabla Q)(v) = \lambda \cdot v \quad \text{for some } \lambda \in \mathbb{R}
\]

Just like in the earlier chapters, it is convenient to replace the affine space \( \mathbb{R}^n \) with the projective space \( \mathbb{P}^{n-1} \). Two nonzero vectors are identified if they are parallel.
From $Q$ we obtain an induced self-map on projective space:

$$
\nabla Q : \mathbb{P}^{n-1} \to \mathbb{P}^{n-1}.
$$

This is a rational map. It is undefined at points where the gradient vanishes. We conclude our discussion with the following remark concerning the map in (9.3).

**Remark 9.1.** The eigenvectors of $Q$ are the fixed points $v$ in $\mathbb{P}^{n-1}$ of its gradient map $\nabla Q$.

A real $n \times n$-matrix usually has $n$ independent eigenvectors, over the complex numbers. When the matrix is symmetric, its eigenvectors have real coordinates and are orthogonal. For a rectangular matrix, one considers pairs of singular vectors, one on the left and one on the right. The number of these pairs is equal to the smaller of the two matrix dimensions.

Eigenvalues are familiar from linear algebra, where they are taught in concert with eigenvalues and singular values. Linear algebra is the foundation of applied mathematics and scientific computing. Specifically, the concept of eigenvectors and numerical algorithms for computing them, became a key technology during the 20th century.

Eigenvectors and singular vectors are familiar from linear algebra, where they are taught in concert with eigenvalues and singular values. Linear algebra is the foundation of applied mathematics and scientific computing. Specifically, the concept of eigenvectors and numerical algorithms for computing them, became a key technology during the 20th century.

Singular vectors are associated to rectangular matrices. We review their definition through the lens of Remark 9.1. Each rectangular matrix represents a bilinear form, e.g.

$$
B = 2ux + 3uy + 5uz + 3vx + 7vy + 11vz = (u \ v) \begin{pmatrix} 2 & 3 & 5 \\ 3 & 7 & 11 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}.
$$

The gradient of the bilinear form defines an endomorphism of the direct sum of the row space and the column space. This fuses left multiplication and right multiplication by our matrix into a single map. In the example, the gradient is the following vector of linear forms

$$
\nabla B = \left( \frac{\partial B}{\partial u}, \frac{\partial B}{\partial v}, \frac{\partial B}{\partial x}, \frac{\partial B}{\partial y} \right).
$$

The associated linear map $\nabla B : \mathbb{R}^3 \oplus \mathbb{R}^2 \to \mathbb{R}^2 \oplus \mathbb{R}^3$ takes $((x, y, z), (u, v))$ to this vector.

More generally, let $B$ be an $m \times n$-matrix over $\mathbb{R}$. We consider the equations

$$
Bx = \lambda y \quad \text{and} \quad B^t y = \lambda x,
$$

where $\lambda$ is a scalar, $x$ is a vector in $\mathbb{R}^n$, and $y$ is a vector in $\mathbb{R}^m$. These are our unknowns. Given a solution to (9.6), $x$ is an eigenvector of $B^t B$, $y$ is an eigenvector of $B B^t$, and $\lambda^2$ is a common eigenvalue of these two symmetric matrices. Its square root $\lambda \geq 0$ is a singular value of $B$. Associated to $\lambda$ are the right singular vector $x$ and the left singular vector $y$. In analogy to Remark 9.1, the process of solving (9.6) has the following dynamical interpretation:

**Remark 9.2.** The singular vector pairs $(x, y)$ of a rectangular matrix are the fixed points of the gradient map, taken on a product of projective spaces, of the associated bilinear form:

$$
\nabla B : \mathbb{P}^{m-1} \times \mathbb{P}^{n-1} \to \mathbb{P}^{m-1} \times \mathbb{P}^{n-1}
$$

$$(x, y) \mapsto \left( \frac{\partial B}{\partial x_1}, \ldots, \frac{\partial B}{\partial x_n}, \frac{\partial B}{\partial y_1}, \ldots, \frac{\partial B}{\partial y_m} \right).$$
We summarize our review of some linear algebra concepts in the following points:

- Symmetric matrices $Q$ are important since they represent quadratic forms.
- Rectangular matrices $B$ are important since they represent bilinear forms.
- Their gradients define the linear maps one usually identifies with $Q$ and $B$.
- Fixed points of these maps are called *eigenvectors* and *singular vectors*.
- These fixed points are computed via orthogonal decompositions:
  
  $$Q = O \cdot \text{diag} \cdot O^t \quad \text{and} \quad B = O_1 \cdot \text{diag} \cdot O_2.$$ 

  Here $O$, $O_1$ and $O_2$ are orthogonal matrices. The formulas above are known as the *spectral decomposition* and the *singular value decomposition*.

In the age of Big Data, the role of matrices is increasingly played by *tensors*, that is, multidimensional arrays of numbers. Principal component analysis tells us that eigenvectors of covariance matrices $Q = BB^t$ point to directions in which the data $B$ is most spread. One hopes to identify similar features in higher-dimensional data. This has encouraged engineers and scientists to spice up their linear algebra tool box with a pinch of algebraic geometry.

The spectral theory of tensors is the theme of the following discussion. This theory was pioneered around 2005 by Lek-Heng Lim and Liqun Qi. Our aim is to generalize familiar notions, such as rank, eigenvectors and singular vectors, from matrices to tensors. Specifically, we address the following questions. The answers are provided in Examples 9.7 and 9.12.

**Question 9.3.** How many eigenvectors does a $3 \times 3 \times 3$-tensor have?

**Question 9.4.** How many triples of singular vectors does a $3 \times 3 \times 3$-tensor have?

A *tensor* is a $d$-dimensional array $T = (t_{i_1i_2 \cdots i_d})$. The set of all tensors of format $n_1 \times n_2 \times \cdots \times n_d$ form a vector space of dimension $n_1 n_2 \cdots n_d$ over the ground field $K$. For $d = 1, 2$ we get vectors and matrices. A tensor has *rank 1* if it is the outer product of $d$ vectors, written $T = u \otimes v \otimes \cdots \otimes w$, or, in coordinates,

$$t_{i_1i_2 \cdots i_d} = u_{i_1} v_{i_2} \cdots w_{i_d}.$$ 

The problem of *tensor decomposition* concerns expressing $T$ as a sum of rank 1 tensors, using as few summands as possible. That minimal number of summands needed is the *rank* of $T$.

An $n \times n \times \cdots \times n$-tensor $T = (t_{i_1i_2 \cdots i_d})$ is *symmetric* if it is unchanged under permuting the indices. The space $\text{Sym}_d(\mathbb{R}^n)$ of such symmetric tensors has dimension $(n+d-1)$. It is identified with the space of homogeneous polynomials of degree $d$ in $n$ variables, written as

$$T = \sum_{i_1, \ldots, i_d = 1}^n t_{i_1i_2 \cdots i_d} x_{i_1} x_{i_2} \cdots x_{i_d}.$$ 

**Example 9.5.** A tensor $T$ of format $3 \times 3 \times 3$ has 27 entries. If $T$ is symmetric then it has ten distinct entries, one for each coefficient of the associated cubic polynomial in three variables. This polynomial defines a cubic curve in the projective plane $\mathbb{P}^2$, as indicated in Figure 1.
Symmetric tensor decomposition writes a polynomial as a sum of powers of linear forms:

\[
T = \sum_{j=1}^{r} \lambda_j v_j^{\otimes d} = \sum_{j=1}^{r} \lambda_j (v_{1j}x_1 + v_{2j}x_2 + \cdots + v_{nj}x_n)^d.
\]

The gradient of \(T\) defines a map \(\nabla T : \mathbb{R}^n \to \mathbb{R}^n\). A vector \(v \in \mathbb{R}^n\) is an eigenvector of \(T\) if

\[
(\nabla T)(v) = \lambda \cdot v
\]

for some \(\lambda \in \mathbb{R}\).

Eigenvectors of tensors arise naturally in optimization. Consider the problem of maximizing a polynomial function \(T\) over the unit sphere in \(\mathbb{R}^n\). If \(\lambda\) denotes a Lagrange multiplier, then one sees that the eigenvectors of \(T\) are the critical points of this optimization problem.

Algebraic geometers find it convenient to replace the unit sphere in \(\mathbb{R}^n\) by the projective space \(\mathbb{P}^{n-1}\). The gradient map is then a rational map from this projective space to itself:

\[
\nabla T : \mathbb{P}^{n-1} \to \mathbb{P}^{n-1}.
\]

The eigenvectors of \(T\) are fixed points (\(\lambda \neq 0\)) and base points (\(\lambda = 0\)) of \(\nabla T\). Thus the spectral theory of tensors is closely related to the study of dynamical systems on \(\mathbb{P}^{n-1}\). In the matrix case (\(d = 2\)), the linear map \(\nabla T\) is the gradient of the quadratic form

\[
T = \sum_{i=1}^{n} \sum_{j=1}^{n} t_{ij}x_i x_j.
\]

By the Spectral Theorem, \(T\) has a real decomposition \((9.7)\) with \(d = 2\). Here \(r\) is the rank, the \(\lambda_j\) are the eigenvalues of \(T\), and the eigenvectors \(v_j = (v_{1j}, v_{2j}, \ldots, v_{nj})\) are orthonormal. We can compute this by power iteration, namely, by applying \(\nabla T\) until a fixed point is reached.

For \(d \geq 3\), one can still use the power iteration to compute eigenvectors of \(T\). However, the eigenvectors are usually not the vectors \(v_i\) in the low rank decomposition \((9.7)\). One exception arises when the symmetric tensor is odec, or orthogonally decomposable. This means that \(T\) has the form \((9.7)\), where \(r = n\) and \(\{v_1, v_2, \ldots, v_r\}\) is an orthogonal basis of \(\mathbb{R}^n\). These basis vectors are the attractors of the dynamical system \(\nabla T\), provided \(\lambda_j > 0\).
Theorem 9.6. The number of complex eigenvectors of a general tensor $T \in \text{Sym}_d(\mathbb{R}^n)$ is

$$\frac{(d-1)^n - 1}{d-2} = \sum_{i=0}^{n-1} (d-1)^i.$$ 

Example 9.7. Let $n = d = 3$. The Fermat cubic $T = x^3 + y^3 + z^3$ is an odeco tensor. Its gradient map squares each coordinate: $\nabla T : \mathbb{P}^2 \to \mathbb{P}^2$, $(x : y : z) \mapsto (x^2 : y^2 : z^2)$. This dynamical system has seven fixed points, of which only the first three are attractors:

$$(1 : 0 : 0), (0 : 1 : 0), (0 : 0 : 1), (1 : 1 : 0), (1 : 0 : 1), (0 : 1 : 1), (1 : 1 : 1).$$

We conclude that $T$ has seven eigenvectors. The same property holds for $3\times3\times3$-tensors in general.

Figure 2. The polynomial $T = xyz(x + y + z)$ represents a symmetric $3\times3\times3$ tensor.

It is known that all eigenvectors can be real for suitable tensors. This was proved in 2017 by Khazhasov [21] using the theory of harmonic polynomials. For $n = 3$, this can be seen by the following simple argument, found earlier by Abo, Seigal and Sturmfels [2]. Let $T$ be a product of linear forms in three unknowns, defining $d$ lines in $\mathbb{P}^2$, then the $\binom{d}{2}$ vertices of the line arrangement are base points of $\nabla T$, and each of the $\binom{d}{2} + 1$ regions contain one fixed point. This accounts for all $1 + (d-1) + (d-1)^2$ eigenvectors, which are therefore real.

Example 9.8. Let $d = 4$ and fix the product of linear forms $T = xyz(x + y + z)$. Its curve in $\mathbb{P}^2$ is an arrangement of four lines, as shown in Figure 2. This quartic represents a symmetric $3\times3\times3\times3$ tensor. All $13 = 6 + 7$ eigenvectors of this tensor are real. The 6 vertices of the arrangement are the base points of $\nabla T$. Each of the 7 regions contains one fixed point.

For special tensors $T$, two of the eigenvectors in Theorem 9.6 may coincide. This corresponds to vanishing of the eigendiscriminant, which is a big polynomial in the $t_{i_1 i_2 \cdots i_d}$. In the matrix case ($d = 2$), it is the discriminant of the characteristic polynomial of an $n \times n$-matrix. For $3\times3\times3$ tensors, the eigendiscriminant is a polynomial of degree 24 in 27 unknowns. In general we have the following result:

Theorem 9.9 (Abo-Seigal-Sturmfels [2]). The eigendiscriminant is an irreducible homogeneous polynomial of degree $n(n-1)(d-1)^{n-1}$ in the coefficients $t_{i_1 i_2 \cdots i_d}$ of the tensor $T$. 

Singular value decomposition is a central notion in linear algebra and its applications. Consider a rectangular matrix \( T = (t_{ij}) \) of format \( n_1 \times n_2 \). The **singular values** \( \sigma \) of \( T \) satisfy
\[
Tu = \sigma v \quad \text{and} \quad T^*v = \sigma u,
\]
where \( u \) and \( v \) are the corresponding **singular vectors**. Just like with eigenvectors, we can associate to this a dynamical system. Namely, we interpret the matrix as a bilinear form
\[
T = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} t_{ij} x_i y_j.
\]
The gradient of \( T \) defines a rational self-map of a product of two projective spaces:
\[
\nabla T : \mathbb{P}^{n_1-1} \times \mathbb{P}^{n_2-1} \to \mathbb{P}^{n_1-1} \times \mathbb{P}^{n_2-1}
\]
\[
(u, v) \mapsto (Tv, Tu)
\]
The *fixed points* of this map are the pairs of singular vectors of \( T \).

Consider now an arbitrary \( d \)-dimensional tensor \( T \) in \( \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \). It corresponds to a multilinear form. The **singular vector tuples** of \( T \) are the fixed points of the gradient map
\[
\nabla T : \mathbb{P}^{n_1-1} \times \cdots \times \mathbb{P}^{n_d-1} \to \mathbb{P}^{n_1-1} \times \cdots \times \mathbb{P}^{n_d-1}.
\]

**Example 9.10.** The trilinear form \( T = x_1 y_1 z_1 + x_2 y_2 z_2 \) gives a \( 2 \times 2 \times 2 \) tensor. Its map \( \nabla T \) is
\[
\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \quad \to \quad \mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1,
\]
\[
((x_1 : x_2), (y_1 : y_2), (z_1 : z_2)) \quad \mapsto \quad ((y_1 z_1 : y_2 z_2), (x_1 z_1 : x_2 z_2), (x_1 y_1 : x_2 y_2)).
\]

This map has no base points, but it has six fixed points, namely \( (1:0), (1:0), (1:0)) \), \( (0:1), (0:1), (0:1)), (1:1), (1:1), (1:1)), (1:1), (1:1), (1:1)), (1:1), (1:1), (1:1)), and \( (1:1), (1:1), (1:1)) \). These are the triples of singular vectors of the given \( 2 \times 2 \times 2 \) tensor \( T \).

Here is an explicit formula for the expected number of singular vector tuples.

**Theorem 9.11** (Friedland and Ottaviani). For a general \( n_1 \times n_2 \times \cdots \times n_d \)-tensor \( T \), the number of singular vector tuples (over \( \mathbb{C} \)) is the coefficient of the monomial \( z_1^{n_1-1} \cdots z_d^{n_d-1} \) in the polynomial
\[
\prod_{i=1}^d \frac{z_i^{n_i}}{z_i - \hat{z}_i} \quad \text{where} \quad \hat{z}_i = z_1 + \cdots + z_{i-1} + z_{i+1} + \cdots + z_d.
\]

We conclude our excursion into the spectral theory of tensors by answering Question 2.

**Example 9.12.** Let \( d = 3 \) and \( n_1 = n_2 = n_3 = 3 \). The generating function in Theorem 9.11 equals
\[
(z_1^2 + z_1 z_3 + z_3^2)(z_2^2 + z_2 z_3 + z_3^2)(z_3^2 + z_3 z_1 + z_1^2) = \cdots + 37z_1^2 z_2 z_3 + \cdots
\]
This means that a general \( 3 \times 3 \times 3 \)-tensor has exactly 37 triples of singular vectors. Likewise, a general \( 3 \times 3 \times 3 \times 3 \)-tensor, as illustrated in Figure 2, has 997 quadruples of singular vectors.
9.2. Ranks of Tensors

In this section we describe in detail the behavior of tensor rank. There are several ways to define the rank of a matrix $M \in K^a \times K^b$. It is:

1. the smallest integer $r$ such that all $(r+1) \times (r+1)$ minors vanish,
2. the dimension of the image of the induced linear map $K^a \rightarrow K^b$,
3. the dimension of the image of the induced linear map $K^b \rightarrow K^a$,
4. the smallest integer $r$, such that there exist vectors $v_1, \ldots, v_r \in K^a$, $w_1, \ldots, w_r \in K^b$ for which:

$$M_{ij} = \sum_{k=1}^{r} (v_k)_i (w_k)_j.$$ 

The first point implies that matrices of rank at most $r$ form a variety. The last point implies that a matrix of rank $r$ is a sum of $r$ matrices of rank one. This is also true for symmetric matrices: a symmetric matrix of rank $r$ is a sum of $r$ symmetric rank one matrices. Another fact is that a real matrix of rank $r$ has also rank $r$, when regarded as a complex matrix. This seems obvious, but a priori, it is not clear why there is no shorter complex decomposition into rank one matrices. Our aim is to find analogous statements for arbitrary tensors.

We recall the definition of rank one tensor: it is the outer product of $d$ vectors, written $T = u \otimes v \otimes \cdots \otimes w$, i.e.

$$t_{i_1 i_2 \cdots i_d} = u_{i_1} v_{i_2} \cdots w_{i_d}.$$ 

Tensors of rank (at most) one form an algebraic variety. It is the affine cone over the Segre product $P^{n_1} \times \cdots \times P^{n_d}$. In fact, from Chapter 2 and 8 we know the equations of this variety! These are binomial quadrics that can be identified with $2 \times 2$ minors. In other words, a tensor $T \in V_1 \otimes \cdots \otimes V_d \simeq K^{n_1 \times \cdots \times n_d}$ has rank one if and only if all the induced linear maps/matrices, known as flattenings:

$$K^{\prod_{i \in I} d_i} = \bigotimes_{i \in I} V_i^* \rightarrow \bigotimes_{i \in [n] \setminus I} V_i = K^{\prod_{i \in [n] \setminus I} d_i}$$

have rank one, for any subset $I \subset [n]$.

**Example 9.13.** A tensor $T = (t_{ijk}) \in V_1 \otimes V_2 \otimes V_3$ induces a linear map:

$$V_1^* \rightarrow V_2 \otimes V_3,$$

given by:

$$e_i^* \rightarrow (t_{ijk})_{j,k} = \sum_{j,k} t_{ijk} f_j \otimes g_k,$$

where $(e_i), (f_j), (g_k)$ are respectively bases of $V_1, V_2, V_3$.

The conclusion is that rank one tensors behave in a very nice way. What is surprising, arbitrary tensors exhibit very strange properties. Recall that the rank of a tensor $T$ is the minimal $r$ such that $T$ is the sum of rank one tensors.

**Definition 9.14.** The following $2 \times 2 \times$ tensor is known in quantum physics as the $W$-state:

$$W = e_0 \otimes e_0 \otimes e_1 + e_0 \otimes e_1 \otimes e_0 + e_1 \otimes e_0 \otimes e_0.$$ 

It plays an important role in quantum information theory. As we will see below it may be defined as a tangent vector to the Segre product $P^1 \times P^1 \times P^1$. 

Clearly $W$ has rank at most three. In fact, $\text{rk} W = 3$, as the reader is asked to prove in Exercise 9. However, there exist rank two tensors arbitrary near $W$. For any $\epsilon \neq 0$ we have:

$$
\frac{1}{\epsilon} ((e_0 + \epsilon e_1) \otimes (e_0 + \epsilon e_1) \otimes (e_0 + \epsilon e_1) - e_0 \otimes e_0 \otimes e_0) = \\
W + \epsilon (e_1 \otimes e_1 \otimes e_0 + e_1 \otimes e_0 \otimes e_1 + e_0 \otimes e_1 \otimes e_1) + \epsilon^2 e_1 \otimes e_1 \otimes e_1.
$$

In particular, we have

$$
\lim_{\epsilon \to 0} \frac{1}{\epsilon} ((e_0 + \epsilon e_1) \otimes (e_0 + \epsilon e_1) \otimes (e_0 + \epsilon e_1) - e_0 \otimes e_0 \otimes e_0) = W.
$$

We conclude that the $W$-state is a tensor of rank three, but it can be approximated with arbitrary precision by a sequence of tensors of rank two.

**Definition 9.15.** The border rank $\text{brk} T$ of the tensor $T$ is the smallest $r$ such that there exist tensors of rank $r$ in any neighbourhood of $T$.

We note that the notion of border rank requires a topology on the space of tensors. The geometric locus of tensors of border rank at most $r$ is the closure of the locus of tensors of rank at most $r$. Over complex numbers, by Chevalley’s theorem [4,17] it does not matter if we take Zariski or Euclidean topology: the closures coincide. However, over the real numbers the situation is different.

**Example 9.16.** Consider $\mathbb{R}^2 \otimes \mathbb{R}^2 \otimes \mathbb{R}^2$. In Exercise 10 the reader is asked to prove that the Zariski closure of tensors of rank two is the whole space. However, the Euclidean closure of the locus of rank two tensors is a proper semialgebraic subset. Passing to the projective setting $X = \mathbb{P} \mathbb{R} \otimes \mathbb{P} \mathbb{R} \otimes \mathbb{P} \mathbb{R} \subset \mathbb{P}^7 \mathbb{R}$, the union of the tangent spaces to $X$, known as the tangential variety, is a hypersurface in $\mathbb{P}^7 \mathbb{R}$. The sign of the defining equation of the tangential variety determines if the tensor has (real) rank two or three.

Explicitly, given a tensor $T$ we obtain the associated map:

$$
\mathbb{R}^2 \to \mathbb{R}^2 \otimes \mathbb{R}^2.
$$

For a general tensor $T$, the image of this map is a two dimensional linear space $S$ of $2 \times 2$ matrices. If $T$ has rank two, i.e. $T = v_1 \otimes v_1 \otimes w_1 + v_2 \otimes v_2 \otimes w_2$ then $S$ must contain two rank one matrices: $v_1 \otimes w_1$ and $v_2 \otimes w_2$. Hence, we ask if $S$ intersects the locus of rank one matrices (at least) two points. In projective setting, rank one matrices are defined by the quadric, i.e. the determinant, and coincide with the Segre surface that is the image:

$$
\mathbb{P}^1 \times \mathbb{P}^1 \to \mathbb{P}^3.
$$

The line $\mathbb{P}(S)$ must intersect this surface over the field of complex numbers, however, does not have to over the field of real numbers.

Consider the tensor $T := e_1 \otimes f_1 \otimes g_1 - e_1 \otimes f_2 \otimes g_2 - e_2 \otimes f_1 \otimes g_2 - e_2 \otimes f_2 \otimes g_1$. The associated space $S$ consists of the matrices

$$
\begin{pmatrix}
a & b \\
b & -a
\end{pmatrix}.
$$

This space does not contain any real rank one matrix. The argument remains correct in a (Euclidean) neighbourhood of $T$. On the other hand we obtain two complex rank one matrices, which give rise to the decomposition:

$$
T = \frac{1}{2} ((e_1 + ie_2)^\otimes 3 + (e_1 - ie_2)^\otimes 3).$$
To conclude, contrary to the case of matrices or rank one tensors:

- tensors of rank at most \( r \) may not form a closed set,
- a real tensor may have different (smaller) rank, when regarded as a complex tensor,
- real tensors of bounded real border rank form semialgebraic sets.

We have described rank one tensors as the Segre product of projective spaces. It is natural to ask for a geometric description of tensors of rank at most \( r \).

**Definition 9.17 (Secant Variety).** Let \( X \) be a projective (resp. affine) algebraic variety. For a set \( S \) let \( \langle S \rangle \) be the smallest projective (resp. affine) subspace containing \( S \). The \( k \)-th secant variety of \( X \) is the closure of all \( k \)-secant planes:

\[
\sigma_k(X) := \bigcup_{p_1, \ldots, p_k \in X} \langle p_1, \ldots, p_k \rangle.
\]

In particular,

\[
X = \sigma_1(X) \subset \sigma_2(X) \subset \cdots \subset \sigma_{\dim(X)}(X) = \langle X \rangle.
\]

In fact, the containments is strict until \( \sigma_r = \langle X \rangle \). If \( X \) is the Segre product, then \( \bigcup_{p_1, \ldots, p_k \in X} \langle p_1, \ldots, p_k \rangle \) is the locus of tensors of rank at most \( r \). Hence, \( \sigma_r(X) \) coincides with the locus of tensors of border rank at most \( r \). It is a major open problem to describe the ideal of \( \sigma_r(\mathbb{P}^{a_1} \times \cdots \times \mathbb{P}^{a_n}) \). This would provide an algebraic test for a tensor to have border rank \( r \). Let us describe the simplest equations. A tensor \( T \in V_1 \otimes \cdots \otimes V_n \) and a subset \( I \subset \{1, \ldots, n\} \), induce the flattening map:

\[
\bigotimes_{i \in I} V_i^* \to \bigotimes_{i \in \{1, \ldots, n\} \setminus I} V_i.
\]

The rank of the flattening map is at most \( \text{rk} \ T \). It follows that the size \( r + 1 \) minors of the flattening matrix provide (some) equations of \( \sigma_r(\mathbb{P}(V_1) \times \cdots \times \mathbb{P}(V_n)) \). These are of degree \( r + 1 \). The are no polynomials of strictly smaller degree in this ideal.

An analogous notion of symmetric rank or Waring rank can be defined for symmetric tensors. A symmetric tensor \( T \) has symmetric/Waring rank one if the following equivalent conditions hold:

1. \( \text{rk} \ T = 1 \),
2. \( T = v \otimes \cdots \otimes v \) for some vector \( v \),
3. \( T \) represented as a polynomial is a power of a linear form.

The symmetric/Waring rank of a symmetric tensor \( T \) is the smallest \( r \) such that \( T \) is a linear combination of \( r \) rank one symmetric tensors.

**Remark 9.18.** We do not write that a tensor is a sum of symmetric tensors, as over real numbers, this may be not possible. For example, when \( T \) is represented by an even degree polynomial, then sum of even powers of real linear forms always is a polynomial that is nonnegative.

We also have a concept of symmetric/Waring border rank of a symmetric tensor \( T \); it is the smallest integer \( r \) such that \( T \) can be approximated by symmetric tensors of symmetric/Waring rank \( r \). Our previous discussion shows that \( W \)-state has symmetric rank three and symmetric border rank two. Clearly, for any symmetric tensor its symmetric rank (resp. symmetric border rank) is at least equal to its rank (resp. border rank).
Conjecture 9.19 (Comon’s Conjecture). For any symmetric tensor its symmetric rank equals its symmetric border rank.

The conjecture was confirmed in many specific cases, however recently a counterexample was presented by Yaroslav Shitov \[35\]. Unfortunately, the example he gave is far too complicated to be presented in this book. The border rank analogue of Comon’s conjecture remains open.

Just as rank one tensors correspond to Segre products \(\mathbb{P}^a \times \cdots \times \mathbb{P}^a\), symmetric rank one tensors correspond to Veronese reembeddings \(v_d(\mathbb{P}^n) \subset \mathbb{P}^{\binom{d+n}{n}-1}\). Here, we interpret \(\mathbb{P}^n\) as a space of linear forms in \(n + 1\) variables and \(\mathbb{P}^{\binom{d+n}{n}-1}\) as the space of degree \(d\) homogeneous polynomials. The map \(v_d\) sends a linear form \(l\) to its \(d\)-th power \(l^d\). This is the same Veronese map as discussed in Chapter 8, up to scaling the coordinates. Further, the locus of symmetric tensors of Waring rank at most \(r\) is precisely \(\sigma_r(v_d(\mathbb{P}^n))\).

Remark 9.20. For each integer \(k\) there exists a minimal number \(r\), such that for any \(n\) there exist positive integers \(a_1, \ldots, a_r\) such that

\[
n = \sum_{i=1}^{r} a_i^k.
\]

Waring’s original problem is, to determine \(r\) as a function of \(k\).

The problem for polynomials that we are facing is to represent a homogeneous polynomial of degree \(d\) as a linear combination of powers of linear forms. Thus, by analogy, the minimal number of linear forms that is needed is called the Waring rank. By seminal work of Alexander and Hirschowitz we know Waring ranks of general polynomials (of any degree in any number of variables). In other words we know the maximal border rank that a homogeneous polynomial may have. For usual (nonsymmetric) tensors, the problem of determining maximal border rank (or rank) in general remains open.

Although, it is easy to prove that general tensors have high rank and border rank it is extremely hard to find explicit examples. Here, we do not want to dive into precise definition of ‘explicit’, let us just say one seeks a tensor with not too big integer entries. In particular, it is not know how to provide examples of tensors \(T \in \mathbb{C}^n \otimes \mathbb{C}^n \otimes \mathbb{C}^n\) of either:

- rank greater than \(3n\),
- border rank greater than \(2n\).

Still, by Exercise 12, a general tensor in this space has border rank quadratic in \(n\).

9.3. Matrix Multiplication

In this section we show that the operation of multiplying two matrices may be identified with a tensor. On this example we explain how tensors may be regarded as computational problems, tensor decompositions as algorithms and tensor rank as a complexity measure.

Determining the rank of a tensor is one of central problems in non-linear algebra. In general one cannot hope for an efficient solution, as it is NP-hard [18]. However, special cases are of particular interest. The most well-known and important one is the matrix multiplication tensor. Let \(\text{Mat}_{a,b} \simeq \mathbb{C}^{ab}\) be the space of complex \(a \times b\) matrices. Matrix multiplication is a bilinear map: \(\text{Mat}_{a,b} \times \text{Mat}_{b,c} \to \text{Mat}_{a,c}\).
9.3. MATRIX MULTIPLICATION

Hence, it is a tensor \( M_{a,b,c} \in \text{Mat}_{a,b}^* \otimes \text{Mat}_{b,c}^* \otimes \text{Mat}_{a,c} \). To simplify notation let us define \( M_n := M_{a,n,n} \). Explicitly in coordinates we have:

\[
M_{a,b,c} = \sum_{i=1,j=1,k=1}^{a,b,c} e_{ij} \otimes f_{jk} \otimes g_{ik},
\]

where \( e_{ij}, f_{jk}, g_{ik} \) are respectively basis of the spaces \( \text{Mat}_{a,b}^*, \text{Mat}_{b,c}^* \) and \( \text{Mat}_{a,c} \).

Another presentation is provided in Exercise 13.

Let us note that the representation of the tensor \( M_{a,b,c} \) given above may be regarded as an algorithm to compute the product of the two matrices as follows:

Take the sum over \( i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, c \) of the following partial results: put the product of the \( (i, j) \)-th entry of the first matrix with the \( (j, k) \)-th entry of the second matrix in the \( (i, k) \)-th entry of the result matrix.

This is the familiar classical algorithm to multiply matrices that we know. It performs \( abc \) additions and \( abc \) multiplications. Thus for \( a = b = c = n \) its complexity is \( O(n^3) \). We note that the number of multiplications is exactly equal to the number of rank one tensors appearing in the decomposition.

What if we present \( M_{a,b,c} \) with a different decomposition? Could it be that the number of multiplications we perform is smaller than \( abc \) or equivalently is the rank of \( M_{a,b,c} \) smaller than \( abc \)? Half a century ago Strassen set up on the quest to prove that this is not possible. He quickly realized that the case of arbitrary \( a, b, c \) is extremely hard and focused on the first nontrivial case \( a = b = c = 2 \). Here he found out a most surprising answer:

\[
M_2 = (e_{11} + e_{22}) \otimes (f_{11} + f_{22}) \otimes (g_{11} + g_{22}) + (e_{21} + e_{22}) \otimes f_{11} \otimes (g_{21} + g_{22}) +
\]
\[
e_{11} \otimes (f_{12} - f_{22}) \otimes (g_{12} + g_{22}) + e_{22} \otimes (f_{21} - f_{11}) \otimes (g_{11} + g_{21}) +
\]
\[
(e_{11} + e_{12}) \otimes f_{22} \otimes (g_{12} - g_{11}) + (e_{21} - e_{11}) \otimes (f_{11} + f_{12}) \otimes g_{22} +
\]
\[
(e_{12} - e_{22}) \otimes (f_{21} + f_{22}) \otimes g_{11}.
\]

Thus the rank of \( M_2 \) is at most seven! In fact, it is known that the rank and border rank of \( M_2 \) equal exactly seven. The latter is a highly nontrivial statement and we are not aware of any accessible proof.

Why could such a decomposition be interesting? First let us note that it can be still regarded as an algorithm to multiply \( 2 \times 2 \) matrices that adds seven partial results. We describe only first two, as we are sure the reader can reconstruct the other five:

1. add \((1,1)\) entry to the \((2,2)\) entry of the first matrix and multiply by the sum of \((1,1)\) and \((2,2)\) entries of the second matrix. Put the result to the \((1,1)\) and \((2,2)\) entry of the first partial result.

2. add \((2,1)\) entry of the first matrix to the \((2,2)\) entry and multiply by the \((1,1)\) entry of the second matrix. Put the result to the \((2,1)\) and \((2,2)\) entry of the second partial result.

To compute each partial result we only need to perform one multiplication. Although we improved on the number of multiplications we have to do, we needed to increase the number of additions (and subtractions) to twenty one. Why should this be exciting? The reason is that multiplication of \( 2 \times 2 \) matrices is not our final aim. We would like to multiply fast very large matrices. Consider two \( 512 \times 512 \) matrices. How to multiply them? We may regard our matrices as \( 2 \times 2 \) matrices with entries that are \( 256 \times 256 \) matrices and apply Strassen’s algorithm! We will
have to add a lot of 256 × 256 matrices, but we only need to perform seven multiplications of such matrices. Further, these multiplications may be done recursively applying the same algorithm, reducing to multiplication of 128 × 128 matrices, etc. Anyone who tried multiplying or adding very large matrices knows that it beneficial to trade multiplication even for many additions. This is in fact a theorem: the complexity of the (optimal) algorithm to multiply matrices is governed by the rank of $M_n$. Precisely both values are measured by the very important constant

$$\omega := \inf \{ \tau : \text{the complexity of multiplication of two } n \times n \text{ matrices is } O(n^\tau) \}$$

$$= \inf \{ \tau : \text{rank of } M_n = O(n^\tau) \}.$$ 

The naive algorithm proves $\omega \leq 3$, however Strassen’s algorithm described above gives $\omega \leq \log_2 7$. As matrices are of size $n^2$ clearly $\omega \geq 2$. The central, astounding conjecture is:

**Conjecture 9.21.** The constant $\omega$ is equal to two.

The conjecture would imply that it is not much harder to multiply very large matrices then to add them (or even output the result)! At this point we note that our story is very much on the side of applied mathematics: Strassen’s algorithm is implemented and used in practice to multiply very large matrices, especially if they are not of any prescribed type.

A careful reader for sure now has an idea how to proceed with a proof of Conjecture 9.21. As Strassen looked at $2 \times 2$ matrices, we should focus on larger, say $3 \times 3$ matrices. The disappointing fact is that despite many attempts, no one knows either rank or border rank of $M_3$! For the current best estimates we refer to [25, 23, 24, 36]. We note that for each fixed $n$ deciding if rank (resp. border rank) of $M_n$ equals $r$ is reduced to deciding if $M_n$ belongs to the image (resp. closed image) of a particular polynomial map. Thus, methods of Chapter 4 apply. However, as tensor spaces are very large such computations are impossible to be carried out in practice, even for $n = 3$. What one can use instead is representation theory, as described in Chapter 10. The optimal estimates for $\omega$ are beyond the scope of this book. Currently we know $2 \leq \omega < 2.38$. What is fascinating, the upper bounds are based on border rank and nonconstructive methods: one proves the existence of an algorithm without explicitly providing it.

In general, we lack methods to show that a tensor has high rank or border rank, as already mentioned at the end of Section 9.2. To prove that $\omega > 2$ we would need to show that rank of $M_n \in \mathbb{C}^{n^2} \otimes \mathbb{C}^{n^2} \otimes \mathbb{C}^{n^2}$ grows superlinearly with the dimension $n^2$ of the space of matrices, while currently we cannot prove that any (explicit) given tensor has rank greater than $3n^2$.

**Exercises**

1. Fix the quadratic form $Q$ in (9.1). Compute the maxima and minima of $Q$ on the unit 2-sphere. Find all fixed points of the map $\nabla Q : \mathbb{P}^2 \to \mathbb{P}^2$. How are they related?

2. Compute all fixed points of the map $\nabla B : \mathbb{P}^2 \times \mathbb{P}^1 \to \mathbb{P}^2 \times \mathbb{P}^1$ given by $B$ in (9.4).

3. Consider the $3 \times 3 \times 2 \times 2$ tensor defined by the multilinear form $T = x_1 y_1 z_1 w_1 + x_2 y_2 z_2 w_2$? Determine all quadruples of singular vectors of $T$. 

(4) For \(d = 2, 3, 4\), pick random symmetric tensors of formats \(d \times d \times d\) and \(d \times d \times d \times d\). Compute all eigenvectors of your tensors.


(6) Write down an explicit \(3 \times 3 \times 3 \times 3\) tensor with precisely 13 real eigenvectors.

(7) What is the number of singular vector tuples of your tensors in Problem 4?

(8) Compute the eigendiscriminants for tensors of format \(2 \times 2\) and \(2 \times 2 \times 2 \times 2\). Write them explicitly as homogeneous polynomials in these entries of an unknown tensor.

(9) By showing that a particular system of polynomial equations has no solutions, prove that \(\text{rank } W = 3\).

(10) Prove that the Zariski closure of tensors of rank two in \(\mathbb{R}^2 \otimes \mathbb{R}^2 \otimes \mathbb{R}^2\) is the whole space (e.g. by computing the dimension of the locus of such tensors).

(11) Find the equation of the tangential variety to \(\mathbb{P}^1 \otimes \mathbb{P}^1 \otimes \mathbb{P}^1 \subset \mathbb{P}^7\).

(12) Prove that in \(\mathbb{C}^n \otimes \mathbb{C}^n \otimes \mathbb{C}^n\):

(a) there exists a tensor of border rank at least \(\frac{1}{3} n^2\),

(b) every tensor has rank at most \(n^2\).

(13) One may identify linear maps from \(V_1\) to \(V_2\) with tensors in \(V_1^* \otimes V_2\). As matrix multiplication corresponds to composition of linear maps we may regard it as a map:

\[
(V_1^* \otimes V_2) \times (V_2^* \otimes V_3) \to (V_1^* \otimes V_3).
\]

In particular, \(M_{\dim V_1, \dim V_2, \dim V_3}\) belongs to the space \((V_1^* \otimes V_2)^* \otimes (V_2^* \otimes V_3)^* \otimes (V_1^*)^* \otimes (V_2^*)^* \otimes (V_3^*)^*\).

(a) How to interpret \(M_{\dim V_1, \dim V_2, \dim V_3}\) as an element of the last space? Do not refer to the basis, but only linear maps \(V_i \to V_i\).

Hint: Note that there exists a distinguished element in \(V_1^* \otimes V_1\) corresponding to the identity map.

(b) Provide a natural isomorphism \(\text{Mat}_{a,b}^* \simeq \text{Mat}_{b,a}\).

(c) The ambient space of \(M_{a,b,c}\) may also be identified with \(((V_1^* \otimes V_2) \otimes (V_2^* \otimes V_3) \otimes (V_3^* \otimes V_1))^*\), i.e. a trilinear map:

\[
\text{Mat}_{a,b} \times \text{Mat}_{b,c} \times \text{Mat}_{c,a} \to \mathbb{C}.
\]

Describe this map without referring to coordinates.
CHAPTER 10

Representation Theory

Symmetry is the key to many applications and computations. While this is true across the mathematical sciences, it is especially pertinent in nonlinear algebra. In its most basic form, symmetry is expressed via the action of a group acting linearly on a vector space. The study of such actions is the subject of representation theory. For instance, the symmetric group on \( n \) letters acts on \((n - 1)\)-dimensional space by the rotations and reflections that fix a regular \((n - 1)\)-simplex. The map that takes each group element to its associated \( n \times n \) matrix is the representation of the group. The matrix representations of the groups we study here can be simultaneously block-diagonalized. The blocks are irreducible representations. Identifying these blocks is tantamount for exploiting symmetry in explicit computations. Our objective in this chapter is to give a first introduction to representation theory.

10.1. Irreducible Representations

The most important groups we study in this chapter are:

- \( \text{GL}(V) = \text{GL}(\dim V) \) - the group of linear isomorphisms of a finite-dimensional vector space \( V \). This group has the structure of an algebraic variety, given by Exercise 8 in Chapter 2.
- \( \text{SL}(V) = \text{SL}(\dim V) \) - the group of volume- and orientation-preserving linear automorphisms of \( V \), with the structure of an algebraic variety given by the equation \( \det A = 1 \);
- \( S_n \) - the group of permutations of a set with \( n \) elements; this is an algebraic variety consisting of \( n! \) distinct points in \( \text{GL}(n) \), namely the \( n \times n \) permutation matrices.

The groups that we consider have two structures: of an abstract group and of an algebraic variety. We note that basic group operations, like inverse or group action, are in fact morphisms of algebraic varieties. We call such groups algebraic. In what follows, we restrict our attention to algebraic groups and morphisms between them that are both group morphisms and morphisms of algebraic varieties. We work over an algebraically closed field \( K \) of characteristic zero.

In general, the following strategy to study an object can be very powerful:

- consider all maps from (resp. to) this object into (resp. from) another basic object.

This general approach could be seen as motivation to study homotopy, homology or the theory of embeddings. For groups, we obtain the following central definition.

**Definition 10.1.** A representation of a group \( G \) is a morphism \( G \to \text{GL}(V) \).

Given a representation \( \rho : G \to \text{GL}(V) \), every element of \( g \) induces a linear map \( \rho(g) : V \to V \). It is useful to think about a representation as a map \( G \times V \to V \).
with the notation
\[ g v := \rho(g)(v) \in V. \]
Here, we have the natural compatibilities
\[ (g_1 g_2) v = g_1 (g_2 v) \quad \text{and} \quad g(\lambda v_1 + v_2) = \lambda g v_1 + g v_2, \]
where \( \lambda \in K \), \( v, v_1, v_2 \in V \) and \( g, g_1, g_2 \in G \). We say that the group \( G \) acts on the vector space \( V \). If the action follows from the context then we call \( V \) a representation of \( G \).

**Example 10.2.** The groups \( \text{GL}(n) \) and \( \text{SL}(n) \) act (by linear change of coordinates) on the space \( V = K[x_1, \ldots, x_n]_k \simeq K^{(n+k-1)/k} \) of homogeneous polynomials of degree \( k \) in \( n \) variables. Using the monomial basis on \( V \), the representation \( \rho \) maps a small matrix, of size \( n \times n \), to a large matrix, with rows and columns indexed by monomials of degree \( k \). The entries in that large matrix are homogeneous polynomials of degree \( k \) in the entries of the small matrix. We recommend working this out for \( n = k = 2 \). This representation \( \rho \) of \( \text{GL}(n) \) plays an important role in classical Invariant Theory, the topic to be studied in the next Chapter 11.

The representations of a fixed group \( G \) are the objects of a category. In this category, a morphism \( f \) between representations \( \rho_1 : G \to \text{GL}(V_1) \) and \( \rho_2 : G \to \text{GL}(V_2) \) is a linear map \( f : V_1 \to V_2 \) that is compatible with the group action:
\[ f(\rho_1(g)(v)) = \rho_2(g)(f(v)) \quad \text{for all} \quad g \in G \quad \text{and} \quad v \in V_1. \]
This can also be written as \( f(g v) = g f(v) \). The category of representations of a group \( G \) is an abelian category. This means in particular that kernels and cokernels exist - cf. Exercise 3.

Our first aim is to describe the basic building blocks of representations.

**Definition 10.3.** A subrepresentation of a representation \( V \) of a group \( G \) is a linear subspace \( W \subset V \) such that the action of \( G \) restricts to \( W \), i.e.
\[ g w \in W \quad \text{for all} \quad w \in W \quad \text{and} \quad g \in G. \]
Equivalently, a subrepresentation is an injective map in the category of all representations of \( G \).

For any representation \( V \), the subspaces 0 and \( V \) are always subrepresentations.

**Definition 10.4.** A representation \( V \) is called irreducible if and only if 0 and \( V \) are its only subrepresentations. We next show that there are no nonzero morphisms between nonisomorphic irreducible representations.

**Lemma 10.5 (Schur’s Lemma).** Let \( V_1 \) and \( V_2 \) be irreducible representations of a group \( G \). If \( f : V_1 \to V_2 \) is a morphism of representations then either \( f \) is an isomorphism or \( f = 0 \). Further, any two isomorphisms between \( V_1 \) and \( V_2 \) differ by a scalar multiple.

**Proof.** Both the kernel \( \ker f \) and the image \( \text{im} f \) are representations. As \( V_1 \) is irreducible, either \( \ker f = V_1 \) or \( f \) is injective. In the latter case, \( \text{im} f \simeq V_1 \) is a nontrivial subrepresentation of \( V_2 \), hence \( f \) is also surjective, i.e. it is a linear isomorphism. The inverse of \( f \), as a linear map, is also the inverse as morphism of representations.
For the last part, consider two isomorphisms \( f_1 \) and \( f_2 \). We may assume that \( f_1 \) is the identity on \( V_1 \). If \( v \) be the eigenvector of \( f_2 \) with eigenvalue \( \lambda \in K \) then
\[
f_2(v) = \lambda v = \lambda f_1(v).
\]
Consider the morphism of representations \( f := f_2 - \lambda f_1 \). Clearly, \( v \in \ker f \). Hence, by the first part, \( f_2 - \lambda f_1 \) is the zero map, and hence \( f_2 = \lambda f_1 \). □

**Theorem 10.6 (Maschke’s theorem).** Let \( V \) be a finite-dimensional representation of a finite group \( G \). There exists a direct sum decomposition
\[
V = \bigoplus V_i,
\]
where each \( V_i \) is an irreducible representation of \( G \).

**Proof.** By induction on the dimension, it is enough to prove the following statement: if \( W \) is a subrepresentation of \( V \), then there exists a subrepresentation \( W' \) such that \( V = W \oplus W' \).

Let \( \pi : V \to W \) be any (surjective) projection. Let \( \tilde{\pi} : V \to W \) be defined by:
\[
\tilde{\pi}(v) = \frac{1}{|G|} \sum_{g \in G} \rho(g)|_W \circ \pi \circ \rho(g)^{-1}.
\]
We note that \( \tilde{\pi} \) is a morphism of representations and \( V = W \oplus \ker \tilde{\pi} \). □

**Remark 10.7.** The existence of decomposition into irreducible components holds not only for finite groups. It also holds for \( \text{GL}(n) \) and \( \text{SL}(n) \). One possible proof is similar to the one above and is known as the unitarian trick. It was introduced by Hurwitz and generalized by Weyl. A representation that allows such a decomposition is called semi-simple or completely reducible. If all representations of \( G \) have this property then the group \( G \) is called reductive.

**Example 10.8.** The group \( G = (\mathbb{C}, +) \) not reductive. Indeed, let us consider the following representation:
\[
G \ni a \to \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \in \text{GL}(2).
\]
Clearly, the one-dimensional subspace of \( \mathbb{C}^2 \) spanned by the second basis vector is invariant under the group action. However, it does not allow a complement - cf. Exercise 13.

The decomposition of into irreducible representations in Maschke’s Theorem is not unique. The following example makes this clear.

**Example 10.9.** Any group \( G \) acts on any vector space \( V \) trivially by \( gv = v \). Any subspace of \( V \) is a subrepresentation. The irreducible subrepresentations are the 1-dimensional subspaces of \( V \). Hence, any decomposition into 1-dimensional subspaces \( V = \bigoplus_{i=1}^{\dim V} K^1 \) is a decomposition into irreducible representations, but there is no distinguished one.

As we will see, the reason for nonuniqueness, is the fact that distinct \( V_i \)’s appearing in the decomposition may be isomorphic. Let us group the isomorphic \( V_i \)’s together obtaining:
\[
(10.1) \quad V = \bigoplus V_j^{\times a_j},
\]
where \( V_{j_0} \cong V_{j_1} \) if and only if \( j_0 = j_1 \). The subrepresentations \( V_j^{x_{a_j}} \) are called isotypic components. The number \( a_j \) is the multiplicity of the irreducible representation \( V_j \) in \( V \).

**Corollary 10.10 (to Schur’s Lemma).** The isotypic components and multiplicities of a semi-simple representation \( V \) are well defined, i.e. do not depend on the choice of the decomposition into irreducible representations.

**Proof.** Consider two decompositions:
\[
V = \bigoplus_j V_j^{x_{a_j}} = \bigoplus_k V_k^{x_{b_k}}.
\]
Allowing \( a_j, b_k \) to be equal to zero, we may assume that all irreducible representations occur and that the indexing in both sums \( \bigoplus \) is the same. First we prove that for a given irreducible representation \( V_i \) we have \( a_i = b_i \). The restriction of identity gives us an injective map:
\[
m : V_i^{x_{a_i}} \to \bigoplus_k V_k^{x_{b_k}}.
\]
By Schur’s Lemma, the composition of \( m \) with the projection
\[
\pi_s : \bigoplus_k V_k^{x_{b_k}} \to V_s^{x_{b_s}}
\]
equals zero, unless \( s = i \). Hence, \( \text{im} \ m \subset V_i^{x_{b_i}} \). In particular, by dimension count, \( a_i \leq b_i \). Analogously \( b_i \leq a_i \), i.e. the multiplicities do not depend on the decomposition. Further, the composition \( \pi_s \circ m \) is an isomorphism if \( s = i \) and is zero if \( s \neq i \). It follows that \( \text{im} \ m = V_i^{x_{b_i}} \). Thus, the identity maps isotypic components to (the same) isotypic components. \( \square \)

Our next aim is to understand the irreducible representations of a given group \( G \). The following definition provides us with the most important tool.

**Definition 10.11 (Character).** Let \( \rho : G \to GL(V) \) be a representation of \( G \). The character \( \chi_{V} = \chi_{\rho} \) of \( \rho \) is the function \( G \to K \) obtained by composing \( \rho \) with the trace function \( \text{Tr} \):
\[
\chi_{\rho}(g) = \text{Tr}(\rho(g)).
\]

Properties of the trace of a square matrix imply the following about characters:
- If \( V = \bigoplus V_i \) then \( \chi_V = \sum \chi_{V_i} \).
- If \( g_1 \) and \( g_2 \) are conjugate elements of \( G \), then \( \chi(g_1) = \chi(g_2) \) for any character \( \chi \).
- If \( V_1, V_2 \) are representations with characters \( \chi_1, \chi_2 \) then their tensor product \( V_1 \otimes V_2 \) is also a representation, and its character is the product \( \chi_1 \chi_2 \).
- We have \( \chi_V(e) = \dim V \), where \( e \in G \) is the neutral element.

For a finite group \( G \), we fix the following scalar product on the complex vector space \( C^G \) of all functions from \( G \) to \( C \):
\[
\langle \chi_1, \chi_2 \rangle := \frac{1}{|G|} \sum_{g \in G} \chi_1(g) \overline{\chi_2(g)}.
\]
It turns out that characters of all irreducible representations of \( G \) are orthonormal with respect to this scalar product. For details we refer to Serre’s book [33], Chapter 2. In particular, the characters are linearly independent elements in \( C^G \). Hence,
we can find the multiplicities $a_j$ in the isotypic decomposition $V = \bigoplus_j V_j^{a_j}$ by decomposing the character:

$$\chi_V = \sum_j a_j \chi_j.$$ 

For any finite group $G$ there are finitely many irreducible representations - the sum of squares of their dimensions equals the order of the group [33, Chapter 2.5, Corollary 2]. A class function if a function $G \to K$ that is constant on conjugacy classes. Characters in fact form a basis of the space of class functions. Often (all) characters are represented in a table, which makes the decomposition very easy, if we know the character of a representation.

**Example 10.12.** We present the character table for the symmetric group $S_3$ on three letters:

<table>
<thead>
<tr>
<th>Character</th>
<th>Trivial representation</th>
<th>Sign repr.</th>
<th>2-dimensional repr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 identity</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2 cycles $(ijk)$</td>
<td>1</td>
<td>1</td>
<td>−1</td>
</tr>
<tr>
<td>3 transpositions $(ij)$</td>
<td>1</td>
<td>−1</td>
<td>0</td>
</tr>
</tbody>
</table>

There are three conjugacy classes, hence there are three irreducible representations.

The first is the trivial representation $gv = v$, the second is the sign representation $gv = (\text{sgn } g)v$, and the third is the two-dimensional representation, given by the symmetries of a regular triangle. Each column in the table represents a function $S_3 \to \mathbb{C}$. Make sure to check these functions are orthonormal with respect to the inner product [10.2]. In fact, one builds the character table of a finite group by exploiting the orthonormality of the columns. In this manner, one obtains the $5 \times 5$ character table for $S_4$ and the $7 \times 7$ character table for $S_5$.

These ideas generalize to $GL(n)$ and $SL(n)$. However, we cannot represent their characters by tables. However, we can represent each character $\chi$ by its values on the Zariski dense subset of diagonalizable matrices. Hence, we fix a torus $T = (K^*)^n \subset GL(n)$ and restrict the character to $T$. As $\chi$ is constant on conjugacy class and any diagonalizable matrix is conjugate to an element of $T$, the function $\chi|_T$ characterizes $\chi$. Therefore, given any representation $W$ of $GL(n)$, we restrict the group and regard $W$ as a representation of $T$. By Exercise 1 and Corollary [10.10] we know that, as a representation of $T$, the space $W$ decomposes:

$$(10.3) W = \bigoplus_{b \in \mathbb{Z}^n} W_b^{a_b},$$

where $t = (t_1, \ldots, t_n)$ takes $w$ to $t^b w$ for $w \in W_b$. The isotypic components $W_b^{a_b}$ for the $T$-action are called weight spaces. The characters $b$ of $T$ for which $a_b \neq 0$ are called weights.

**Remark 10.13.** Let $T$ be the torus of diagonal matrices $t = \text{diag}(t_1, \ldots, t_n)$ in $GL(n)$. If $\chi$ is a character of $GL(n)$ then its restriction to $T$ is the function $\chi|_T : T \to K, t \mapsto \text{Tr}(\rho(t))$. Here $\text{Tr}$ denotes the trace of a (large) square matrix. The restricted character $\chi|_T$ equals

$$\chi|_T(t) = \sum_{b \in \mathbb{Z}^n} a_b t^b.$$ 

This Laurent polynomial in $t_1, \ldots, t_n$ is invariant under permuting its $n$ unknowns.
Example 10.14. Following Example [10.2] we consider the action of $GL(n)$ on homogeneous polynomials of degree $k$. Let $\chi$ be its character. Then $\chi|_T$ is the complete symmetric polynomial of degree $k$, i.e., $\chi|_T(t)$ is the sum of all monomials $t^a$ where $a \in \mathbb{N}^n$ and $|a| = k$.

Example 10.15. The group $GL(n)$ acts naturally on the $k$th exterior power $V = \wedge^k K^n$. Write $\rho$ for this representation and $\chi$ for its character. We identify $V$ with $K^{(k)}$ by fixing the standard basis $\{e_{i_1} \wedge \cdots \wedge e_{i_k} : 1 \leq i_1 < \cdots < i_k \leq n\}$. The image $\rho(g)$ of an $n \times n$-matrix $g = (g_{ij})$ is the $k$th compound matrix or $k$th exterior power, whose entries are the (suitably signed) $k \times k$ minors of $g$. We note that the determinant of $\rho(g)$ equals $\det(g)^{\binom{k}{2}}$. The restricted character $\chi|_T(t)$ is the $k$th elementary symmetric polynomial in $t_1, \ldots, t_n$.

For a concrete example, let $k = 2$. Then $\rho(g)$ is the $\binom{n}{2} \times \binom{n}{2}$ matrix whose rows and columns are labeled by ordered pairs from $\{1, 2, \ldots, n\}$, and whose entry in row $(i < j)$ and column $(k < l)$ equals $g_{ik}g_{jl} - g_{il}g_{jk}$. We have $\det(\rho(g)) = \det(g)^{n-1}$ and $\chi|_T(t) = \sum_{i<j} t_it_j$. For $k = 1$ we have $\rho(g) = g$, so $\chi|_T(t) = t_1 + t_2 + \cdots + t_n$. Finally, for $k = n$, we get the one dimensional representation where $\rho(g)$ is the $1 \times 1$-matrix with entry $\det(g)$, so we have $\chi|_T(t) = t_1t_2 \cdots t_n$. The latter gives the trivial representation when restricted to $SL(n)$.

Let $\rho$ be any representation of $GL(n)$. We fix the lexicographic order on the set of weights $b$ that occur in $\rho$. Of particular importance is the highest weight. The corresponding eigenvectors $w \in W_b$ in [10.3] are called highest weight vectors. They span the highest weight space. In Example [10.14] the highest weight is $(d, 0, \ldots, 0) \in \mathbb{Z}^n$, and a highest weight vector is the monomial $x_1^d$. In Example 1, the highest weight is $(1, \ldots, 1, 0, \ldots, 0)$, and a highest weight vector is $e_1 \wedge \cdots \wedge e_k$. In both cases, the highest weight space is 1-dimensional.

Example 10.16 (Adjoint representation). The space $V = K^{n \times n}$ of $n \times n$ matrices $M$ forms a representation of $GL(n)$ under the action by conjugation, where $\rho(g)(M) := gMg^{-1}$. This is the adjoint representation. The weights, known as roots in this case, are $t_i/t_j$ with highest weight $(1, 0, \ldots, 0, -1)$. If we restrict it to $SL(V)$ we have $t_n^{-1} = \prod_{i=1}^{n-1} t_i$ and the highest weight becomes $(2, 1, \ldots, 1) \in \mathbb{Z}^{n-1}$. Again, the highest weight space is 1-dimensional.

The following result provides a characterization of irreducible representations.

Proposition 10.17. Every irreducible representation of $SL(V)$ is determined (up to isomorphism) by its highest weight, and the highest weight space is 1-dimensional. A weight $(a_1, \ldots, a_{n-1}) \in \mathbb{Z}^{n-1}$ is the highest weight for some irreducible representation if and only if $a_1 \geq a_2 \geq \cdots \geq a_{n-1} \geq 0$.

Proof. For the proof we refer to [15] Chapter 15.

Here is a combinatorial tool for building irreducible representations from highest weights:

Definition 10.18. A Young diagram with $k$-rows is a nonincreasing sequence of $k$ positive integers. It is usually presented in the following graphical form, e.g. for a sequence $(2, 1, 1)$:

```
  •
  •
```

This diagram represents the Young diagram with two rows of single elements and one row of two elements.
This particular Young diagram encodes the adjoint representation of $\text{SL}(4)$.

Proposition 10.17 tells us that the irreducible representations of $\text{SL}(n)$ are in bijection with the Young diagrams with at most $n - 1$ rows. Representations of $\text{GL}(n)$ are not very different: first, every irreducible representation $V$ of $\text{GL}(n)$ it is also an irreducible representation of $\text{SL}(n)$, so it has a corresponding Young diagram $\lambda$. However, different representations of $\text{GL}(n)$ give the same representation of $\text{SL}(n)$ if they differ a power of the determinant. Precisely, consider a representation $\rho : \text{SL}(n) \rightarrow \text{GL}(V)$ with associated Young diagram $\lambda$. We have the following representations of $\text{GL}(n)$ for any $a \in \mathbb{Z}$:

$$\rho_a(g) := (\det g)^a \cdot \rho\left(\frac{1}{\sqrt{\det g}} \cdot g\right).$$

Here, the argument of $\rho$ is in $\text{SL}(n)$. The 1-dimensional representation $g \rightarrow \det(g)$ of $\text{GL}(n)$ corresponds to a Young diagram with one column and $n$ rows. Thus for $a \geq 0$ the representation $\rho_a$ corresponds to Young diagram $\lambda$ extended by $a$ columns of height $n$. The representation of $\text{GL}(U)$ corresponding to a Young diagram $\lambda$ is denoted by $S^\lambda(U)$.

Given a Young diagram $\lambda$, we write $\chi_\lambda$ for character of the irreducible representation $S^\lambda(U)$. This is a symmetric polynomial in $t = (t_1, \ldots, t_n)$, known as the Schur polynomial of $\lambda$. Schur polynomials include the complete symmetric polynomials in $t$, given by

$$\chi_{(1,\ldots,1)} = \det(t_1, \ldots, t_n).$$

Proposition 10.19. The Schur polynomial for $\lambda$ is the following ratio of $n \times n$ determinants:

$$\chi_\lambda(t) = \frac{\det(t_i^{\lambda_j+n-j})_{1 \leq i,j \leq n}}{\det(t_i^{n-j})_{1 \leq i,j \leq n}}.$$

We can find the decomposition (10.1) of a representation $V$ into irreducibles by writing the character $\chi_V$ as linear combination of Schur functions $\chi_\lambda$ with nonnegative integer coefficients $a_\lambda$. These coefficients are the multiplicities. This expression is unique because the Schur polynomials form a $\mathbb{Z}$-linear basis for the ring of symmetric polynomials in $n$ variables.

Example 10.20. Let $n = 3$. The Schur polynomial for $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ is the ternary form

$$\chi_\lambda(t) = \frac{1}{(t_1-t_2)(t_1-t_3)(t_2-t_3)} \cdot \det\left(\begin{array}{ccc}
t_1^{\lambda_1+2} & t_1^{\lambda_2+1} & t_1^{\lambda_3} \\
t_2^{\lambda_1+2} & t_2^{\lambda_2+1} & t_2^{\lambda_3} \\
t_3^{\lambda_1+2} & t_3^{\lambda_2+1} & t_3^{\lambda_3}
\end{array}\right).$$

From this, we compute the three Schur polynomials of degree $|\lambda| = 3$ as follows:

$$\chi_{(3,0,0)} = t_1^3 + t_1^2 t_2 + t_1 t_2^2 + t_2^3 + t_1^2 t_3 + t_1 t_2 t_3 + t_2^2 t_3 + t_1 t_3^2 + t_2 t_3^2 + t_3^3$$

$$\chi_{(2,1,0)} = (t_1 + t_2)(t_1 + t_3)(t_2 + t_3)$$

$$\chi_{(1,1,1)} = t_1 t_2 t_3$$

The action of $\text{GL}(3)$ on $U = K^3$ induces an action on the 27-dimensional space $U^\otimes 3$ of $3 \times 3 \times 3$-tensors. As characters are multiplicative under tensor product, its character equals

$$\chi_{U^\otimes 3} = (t_1 + t_2 + t_3)^3 = \chi_{(3,0,0)} + 2 \cdot \chi_{(2,1,0)} + \chi_{(1,1,1)}.$$
112 10. REPRESENTATION THEORY

From this decomposition into Schur polynomials, we conclude the irreducible de-
composition

\[(10.4) \quad U \otimes^3 = S^{(3)}(U) \oplus (S^{(21)}(U) \oplus S^{(21)}(U)) \oplus S^{(111)}(U).\]

The first summand is the symmetric tensors, the last summand is the antisymmetric
tensors, and the middle summand consists of two copies of the adjoint representation
(Example [10.16]).

The irreducible representations \(S^\lambda(U)\) of \(\text{SL}(U)\) come together with nice alge-
braic varieties. The group \(\text{SL}(U)\) acts also on the projective space \(\mathbb{P}(S^\lambda(U))\). The
letter action has unique closed orbit, namely the orbit of the highest weight vector.
Particular examples are:

1. The orbit of \([e_1 \cdots e_1] \in \mathbb{P}(S^k(U))\). This is the \(k\)-th Veronese embedding
   of \(\mathbb{P}(U)\).

2. The orbit of \([e_1 \wedge \cdots \wedge e_k] \in \mathbb{P}(\wedge^k(U))\) is the Grassmannian \(G(k, U)\) in its
   Plücker embedding. Here \(\lambda = (1, \ldots, 1)\) as in Example [10.15].

This result provides us with a unified approach to homogeneous varieties. It
could be also used to build some of the representations. Fix a Young diagram \(\lambda\) and
let \(k\lambda\) be a Young diagram where each row is scaled by \(k\). Given the homogeneous
variety \(X\) in \(\mathbb{P}(S^\lambda(U))\) we can take the \(k\)-th Veronese map \(v_k(X)\) of this projective
space and the linear span of \(v_k(X)\) is \(S^{k\lambda}(V)\). A special case of this construction
is point 1 above where \(X = \mathbb{P}(U)\).

10.2. Schur-Weyl Duality

In this section we present a beautiful connection between finite groups - \(S_n\) and
Lie groups - \(\text{SL}(n)\) or \(\text{GL}(n)\). This is the Schur-Weyl duality. We refer readers
interested in the topic to [15], Chapter 4].

Before stating it let us go back to irreducible representations of \(S_n\). Their
characters form a basis of class functions. Hence the number of irreducible rep-
resentations equals the number of conjugacy classes. Each conjugacy class can be
encoded by lengths of cycles in a decomposition of a permutation into cycles. These
can be further represented by a Young diagram with \(n\) boxes: the first row repre-
sents the length of the longest cycle, the last of the shortest. Thus, the number
of irreducible representations of \(S_n\) equals the number of Young diagrams with \(n\)
boxes. We shall exhibit a natural bijection between Young diagrams with \(n\) boxes
and irreducible representations of \(S_n\). Before, we see how to construct it, let us
assume that to each such Young diagram \(\lambda\) we can associate a representation \(S_\lambda\) of
\(S_n\).

Fix a vector space \(U\) and consider the \(n\)-fold tensor product \(U^\otimes n\). There are
two groups acting on it: \(\text{GL}(U)\) - on each factor - and \(S_n\) - by permuting factors.
Schur-Weyl duality provides a simultaneous decomposition of the space of tensors
with respect to both groups.

**Theorem 10.21** (Schur-Weyl duality). Let \(U\) be a vector space of dimension
at least \(n\). Then

\[(10.5) \quad U^\otimes n = \sum_{|\lambda|=n} S_\lambda \otimes S^\lambda(U),\]

where the sum is over all Young diagrams with precisely \(n\) boxes.
When \( n = 2 \) we obtain \( U^\otimes 2 = S^2(U) \oplus \wedge^2 U \), as there are only two irreducible representations of \( S_2 \), both 1-dimensional. This recovers the fact every \( n \times n \) matrix is uniquely the sum of a symmetric matrix and a skew-symmetric matrix. The \( S_2 \) action on the matrix space \( U^\otimes \) is transposition, which acts trivially on \( S^2(U) \) and changes the sign on \( \wedge^2 U \).

The case \( n = 3 \) is the first interesting one. The three irreducible representations \( S^\lambda \) of \( S_3 \) in Example 10.12 correspond to the three outer summands in (10.4). Note that \( \dim(S^\lambda) = 2 \) for \( \lambda = (2, 1) \). The middle summand in (10.4) is the 16-dimensional space \( S_{(21)} \otimes S_{(21)}(U) \).

By Schur-Weyl duality, the multiplicity of \( S^\lambda(U) \) in \( U^\otimes n \) equals the dimension of \( S^\lambda \). This provides us with a method for defining \( S^\lambda \). Consider the decomposition of \( U^\otimes n \) as an \( \text{GL}(U) \) representation, into isotypic components. Here \( a_\lambda \) can be found using Schur functions:

\[
U^\otimes n = \bigoplus \lambda (S^\lambda(U))^{a_\lambda}.
\]

For each isotypic component \( (S^\lambda(U))^{a_\lambda} \) consider the highest weight space, i.e. eigenvectors of the torus action with weight \( \lambda \). The permutation group \( S_n \) acts on the highest weight space. This representation of \( S_n \) is irreducible, and we find that it is precisely \( S^\lambda \).

Coming back to the example of matrices \( (n = 2) \), the highest weight vectors are as follows:

- The highest weight vector \( e_1 e_1 = e_1 \otimes e_1 \) of \( S^2(U) \) is invariant with respect to transposition, i.e. it provides the trivial representation of the two-element group \( S_2 \).
- The highest weight vector \( e_1 \wedge e_2 = \frac{1}{2}(e_1 \otimes e_2 - e_2 \otimes e_1) \) of \( \wedge^2(U) \) changes sign when transposed, i.e. it provides the sign representation of the two-element group \( S_2 \).

Example 10.22 \((n = 3)\). Let \( \lambda = (2, 1) \). The isotypic component \( (S_{(21)}(U))^2 \) in the middle of (10.4) has a 2-dimensional subspace \( S^\lambda \) of highest weight vectors. One possible basis of this space consists of the tensors \( e_{11} + e_{21} - 2e_{121} \) and \( e_{12} + e_{21} - 2e_{112} \), where \( e_{ijk} := e_i \otimes e_j \otimes e_k \).

Remark 10.23. We stress the fact that we worked under the assumption that the field is algebraically closed and of characteristic zero, which makes representation theory much better behaved. Representation theory in finite characteristic is considerably more complicated.

10.3. Exploiting Symmetry

In this section we will show how representation theory can be used for describing polynomial ideals and their varieties in the presence of a large symmetry group.

Exercises

1. (a) Prove that, over an algebraically closed field, every irreducible representation of an abelian group is 1-dimensional.

(b) Explain the correspondence between the characters of a torus \( T = (\mathbb{C}^*)^n \), as defined in Chapter 8, and the irreducible representations of \( T \).
(2) Derive the character table of the symmetric group $S_4$. Hint:

$$1^2 + 1^2 + 2^2 + 3^2 + 3^2 = 24.$$ 

What is the geometric meaning of the 3-dimensional irreducible representations?

(3) Let $f : V_1 \to V_2$ be a morphism between two representations of a group $G$.

- Prove that the kernel, image and cokernel of $f$ are also representations.
- Prove that morphisms of two representations are closed under taking scalar multiples and sums, i.e. they form a vector space.

(4) Derive the character table of the symmetric group $S_5$. Hint:

$$1^2 + 1^2 + 2^2 + 4^2 + 4^2 + 5^2 + 5^2 + 6^2 = 120.$$ 

Can you write matrices $ho(g)$ for the 6-dimensional irreducible representation?

(5) Let $V_1$ and $V_2$ be two representations of a group $G$.

(a) Prove that linear morphisms $\text{Hom}(V_1, V_2)$ have also a structure of a representation. How can you characterize morphisms of representations inside $\text{Hom}(V_1, V_2)$?

(b) In terms of multiplicities of isotypic components of $V_1$ and $V_2$, what is the dimension of the space of morphisms among these two representations?

(c) Conclude that the multiplicity of an irreducible representation $W$ in $V_1$ equals the dimension of morphisms of representations $W \to V_1$ (or equivalently of $V_1 \to W$).

(6) Let $V$ be a representation of $\text{GL}(n)$. Its character $\chi_V$ is a Laurent polynomial in $t_1, \ldots, t_n$. Show that the vector spaces $S^2(V)$ and $\Lambda^2 V$ are also representations of $\text{GL}(V)$, and compute the characters $\chi_{S^2(V)}$ and $\chi_{\Lambda^2 V}$ in terms of $\chi_V$.

(7) Describe the 2-dimensional irreducible representation from Example 10.12 explicitly, by assigning a $2 \times 2$ matrix to each of the six permutations of $\{1, 2, 3\}$.

(8) Consider the representation $\rho$ of $\text{GL}(3)$ action on $\Lambda^3 K^6$. What is the highest weight? What is the associated Young diagram? Find the entries of the $20 \times 20$ matrix $\rho(g)$.

(9) Is every $2 \times 2 \times 2$ tensor the sum of a symmetric and a skew-symmetric tensor?

(10) If $U = K^n$, what is the dimension of $S^2(U)$? Give a formula in terms of $n$.

(11) What is the dimensions of the vector space $S^3(S^3(K^n))$? Find a weight basis. Write down the character of this representation of $\text{GL}(3)$. Can you decompose it into Schur polynomials?

(12) What are the orbits of the adjoint representation? Are they closed? What is the dimension of a general orbit? What is the vanishing ideal such an orbit, e.g. for $n = 3$?

(13) Show that the representation $\mathbb{C}^2$ in Example 10.8 is not a sum of irreducible representations.
CHAPTER 11

Invariant Theory

What is geometry? An answer to this question was proposed by Felix Klein’s Erlanger Programm. According to Klein, a quantity is geometric if it is invariant under the action of an underlying of transformations. Thus, in short, geometry is invariant theory. For example, Euclidean geometry is the study of quantities, expressed in the coordinates of points, that are invariant under the Euclidean group. From the modern point of view, invariant theory can be seen as a branch of representation theory. However, that view does not do justice to the tremendous utility of invariant theory for dealing with geometric objects. In particular, in algebraic geometry, invariants are used to construct quotients of algebraic varieties modulo groups that act on them. This results in a concise description of orbit spaces. The study of such spaces is called Geometric Invariant Theory. Our aim in this chapter is to give a first introduction to this theory, starting with actions by finite groups.

11.1. Finite Groups

We fix the polynomial ring $K[x] = K[x_1, \ldots, x_n]$ over a field $K$ of characteristic zero. The group $\text{GL}(n, K)$ of invertible $n \times n$ matrices acts on $K^n$. This induces an action by $G$ on the ring of polynomial functions on $K^n$. Namely, if $\sigma = (\sigma_{ij})$ is a matrix in $\text{GL}(n, K)$ and $f$ is a polynomial in $K[x]$ then $\sigma f$ is the polynomial that is obtained from $f$ by replacing the variable $x_i$ by the linear form $\sum_{j=1}^n \sigma_{ij} x_j$ for $i = 1, \ldots, n$.

Let $G$ be a subgroup of $\text{GL}(n, K)$. A polynomial $f \in K[x]$ is an invariant of the group $G$ if $\sigma f = f$ for all $\sigma \in G$. We write $K[x]^G$ for the set of all such invariants. This set is a subring because the sum of two invariants is again an invariant, and same for the product.

In this chapter we discuss two scenarios. In this section we consider finite groups $G$, and in the next one we consider representations of nice infinite groups like $\text{SL}(d, K)$ and $\text{SO}(d, K)$. Such groups are called reductive. A celebrated theorem of Hilbert shows that the invariant ring is finitely generated in this case. After two initial examples, we begin by proving this for finite groups $G$.

**Example 11.1.** Let $G$ be the group of $n \times n$ permutation matrices. The invariant ring $K[x]^G$ consists of all polynomials $f$ that are invariant under permuting the coordinates, i.e.

$$f(x_{\pi_1}, x_{\pi_2}, \ldots, x_{\pi_n}) = f(x_1, x_2, \ldots, x_n) \text{ for all permutations } \pi \text{ of } \{1, 2, \ldots, n\}.$$ 

The invariant ring $K[x]^G$ is generated by the $n$ elementary symmetric polynomials $E_1, \ldots, E_n$. These are the coefficients of the following auxiliary polynomial in one
variable $z$:

\[(11.1) \quad (z + x_1)(z + x_2) \cdots (z + x_n) = z^n + \sum_{i=1}^{n} E_i(x) z^{n-i}.\]

We also set $E_0 = 1$. Alternatively, $K[x]^G$ can also be generated by the power sums $P_j(x) = x_1^j + x_2^j + \cdots + x_n^j$ for $j = 1, 2, \ldots, n$.

The formulas that connect the $E_i$ and the $P_j$ are known as *Newton’s Identities*:

\[(11.2) \quad kE_k = \sum_{i=1}^{k} (-1)^{i-1} E_{k-i} P_i \quad \text{and} \quad P_k = (-1)^{k-1} kE_k + \sum_{i=1}^{k-1} (-1)^{k-i-1} E_{k-i} P_i \quad \text{for} \quad 1 \leq k \leq n.\]

Invariants are polynomial functions that are constant along $G$-orbits on $K^n$. They offer an algebraic view on the space of orbits. Namely, we think of the spectrum of $K[x]^G$ as a quotient space $K^n//G$, whose points are these orbits. This interpretation is only informal, as the details are very subtle. Making it all precise is the aim of *Geometric Invariant Theory*.

**Example 11.2.** For $n = 2$, consider the following representation of the cyclic group of order $4$:

\[(11.3) \quad G = \{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \} \].

These are the rotational symmetries of the square. Its invariant ring is generated by $I_1 = x_1^2 + x_2^2, \quad I_2 = x_1^2 x_2^2, \quad I_3 = x_1^4 x_2 - x_1 x_3^3$.

These three invariants are algebraically dependent. Using their relation we can write

\[(11.4) \quad K[x_1, x_2]^G = K[I_1, I_2, I_3] \cong K[y_1, y_2, y_3]/(y_1 y_2 - 4y_2^2 - y_3^2).\]

The spectrum of the ring (11.4) is the cubic surface in $K^3$ defined by the equation $y_1 y_2 = 4y_2^2 + y_3^2$. The points on this surface are in one-to-one correspondence with the $G$-orbits on $K^2$.

In what follows, let $G$ be a finite subgroup of $\text{GL}(n, K)$. One can create invariants by averaging polynomials. The *Reynolds operator*, denoted by a star, is

\[(11.5) \quad * : K[x] \to K[x]^G, \quad f \mapsto f^* := \frac{1}{|G|} \sum_{\sigma \in G} \sigma f.\]

Each of the following properties of the Reynolds operator is easily verified:

**Lemma 11.3.** The Reynolds operators $*$ has the following three properties:

(a) * The map $*$ is $K$-linear, i.e. $(\lambda f + \nu g)^* = \lambda f^* + \nu g^*$ for all $f, g \in K[x]$ and $\lambda, \nu \in K$.

(b) * The map $*$ restricts to the identity on $K[x]^G$, i.e. $I^* = I$ for all invariant polynomials $I$.

(c) * The map $*$ is a $K[x]^G$-module homomorphism, i.e. $(fI)^* = f^* I$ for all $f \in K[x]$ and $I \in K[x]^G$.

The following result from 1890 marks the beginning of Commutative Algebra.

**Theorem 11.4 (Hilbert’s Finiteness Theorem).** The invariant ring $K[x]^G$ of any finite matrix group $G \subset \text{GL}(n, K)$ is finitely generated as a $K$-algebra.
We present the proof under the hypothesis that \( K \) has characteristic zero. However, the result holds for every field \( K \). For a proof see \[11\]. This is known as modular invariant theory.

**Proof.** Let \( \mathcal{I}_G = \langle K[x]^G \rangle \) be the ideal in \( K[x] \) that is generated by all homogeneous invariants of positive degree. By Lemma \[11.3\] (a), every invariant is a \( K \)-linear combination of symmetrized monomials \((x^a)^*\). These homogeneous invariants are the images of monomials under the Reynolds operator. Thus \( \mathcal{I}_G \) is generated by the set \( \{(x^a)^*: a \in \mathbb{N}^n\}\). By Hilbert’s Basis Theorem, the ideal \( \mathcal{I}_G \) is finitely generated, so that a finite subset of \( a \) in \( \mathbb{N}^n \) suffices. In conclusion, there exist invariants \( I_1, I_2, \ldots, I_m \) such that \( \mathcal{I}_G = \langle I_1, I_2, \ldots, I_m \rangle \).

We claim that these \( m \) invariants generate the invariant ring \( K[x]^G \) as a \( K \)-algebra. Suppose the contrary, and let \( I \) be a homogeneous element of minimal degree in \( K[x]^G/K[I_1, I_2, \ldots, I_m] \). Since \( I \in \mathcal{I}_G \), we have \( I = \sum_{j=1}^{m} f_j I_j \) for some homogeneous polynomials \( f_j \in K[x] \) whose degrees are all strictly less than \( \deg(I) \).

Applying the Reynolds operator on both sides of the equation \( I = \sum_{j=1}^{m} f_j I_j \), we obtain

\[
I = I^* = \left( \sum_{j=1}^{m} f_j^* I_j \right)^* = \sum_{j=1}^{m} f_j^* I_j.
\]

Here we are using the properties (b) and (c) in Lemma \[11.3\]. The new coefficients \( f_j^* \) are homogeneous invariants whose degrees are less than \( \deg(I) \). From the minimality assumption on the degree of \( I \), we get \( f_j^* \in K[I_1, \ldots, I_m] \) for \( j = 1, \ldots, m \). This implies \( I \in K[I_1, \ldots, I_m] \), which is a contradiction to our assumption. This completes the proof of Theorem \[11.4\].

**Theorem 11.5 (Noether’s Degree Bound).** If \( G \) is finite and \( \text{char}(K) = 0 \) then the invariant ring \( K[x]^G \) is generated by homogeneous invariants of degree \( \leq |G| \).

**Proof.** Let \( u = (u_1, \ldots, u_n) \) be new variables. For any \( d \in \mathbb{N} \), we consider the expression

\[
S_d(u, x) = \left[ (u_1 x_1 + \cdots + u_n x_n)^d \right]^*. 
\]

This is a polynomial in \( u \) whose coefficients are polynomials in \( x \). Up to a multiplicative constant, they are the invariants \((x^a)^*\) where \( |a| = d \). All polynomials in \( u \) are fixed under *.

Consider the \( |G| \) expressions \( u_1(x^1) + \cdots + u_n(x^n) \), one for each group element \( \sigma \in G \). The polynomial \( S_d(u, x) \) is the \( d \)th power sum of these expressions. The power sums for \( d > |G| \) are polynomials in the first \( |G| \) power sums. Such a representation is derived from Newton’s Identities \[11.2\]. It implies that all \( u \)-coefficients of \( S_d(u, x) \) for \( d > |G| \) are polynomial functions in the \( u \)-coefficients of \( S_d(u, x) \) for \( d \leq |G| \). Hence all invariants \((x^a)^*\) with \( |a| > |G| \) are polynomial functions (over \( K \)) in the invariants \((x^b)^*\) with \( |b| \leq |G| \). This proves the claim.

We note that Example \[11.2\] attains Noether’s degree bound. The cyclic group in that example has order 4, and the invariant ring requires a generator of degree 4.

Our next theorem is a useful tool for constructing the invariant ring. It says that we can count invariants by averaging the reciprocal characteristic polynomials of the group elements.
Theorem 11.6 (Molien). The Hilbert series of the invariant ring $K[x]^G$ equals

\begin{equation}
\sum_{d=0}^{\infty} \dim_{K}(K[x]_d^G) z^d = \frac{1}{|G|} \sum_{g \in G} \frac{1}{\det(id - z g)}.
\end{equation}

The coefficient of $z^d$ in this formal generating function is the number of linearly independent invariants of degree $d$.

Proof. See [38] Theorem 2.2.1. \hfill \square

Example 11.7. Consider the cyclic group $G = \mathbb{Z}_4$ in Example 11.2. For the four matrices $g$ in Example 11.2 the quadratic polynomials $\det(id - z g)$ are $(1-z)^2$, $1+z^2$, $(1+z)^2$ and $1+z^2$. Adding up their reciprocals and dividing by $|G| = 4$, we see that the Hilbert series of $K[x]^G$ is

\begin{equation}
\frac{1 + z^4}{(1-z^2)(1-z^4)} = 1 + z^2 + 3z^4 + 3z^6 + 5z^8 + \cdots.
\end{equation}

This agrees with the Hilbert series of the ring on the right in (11.4), where $\deg(y_1) = 2$ and $\deg(y_2) = \deg(y_3) = 4$. Indeed, the principal ideal exhibits a Noether normalization. We see that the ring is a free module with basis $\{1, y_3\}$ over $K[y_1, y_2]$. This explains the numerator and denominator on the left of (11.7), and it proves that $I_1, I_2, I_3$ do indeed generate $K[x]^G$.

11.2. Classical Invariant Theory

Hilbert’s Finiteness Theorem also holds for an infinite group $G \subset GL(n, K)$ that has a Reynolds operator $\ast$ satisfying the properties (a), (b) and (c) in Lemma 11.3. Such $G$ are known as reductive groups.

Corollary 11.8. Fix a reductive group $G$ of $n \times n$-matrices. If $\{g_1, g_2, \ldots, g_m\}$ is a set of homogeneous polynomials that generates the ideal $I_G$ then its image $\langle g_1^\ast, g_2^\ast, \ldots, g_m^\ast \rangle$ under the Reynolds operator generates the invariant ring $K[x]^G$ as a $K$-algebra.

Proof. Let $M = \langle x_1, \ldots, x_n \rangle$ be the homogeneous maximal ideal in $K[x]$, and consider the finite-dimensional vector space $I_G/MI_G$. It has a basis of invariants since $I_G$ is generated by invariants. This means that the Reynolds operator acts as the identity on $I_G/MI_G$. The images of $g_1, g_2, \ldots, g_m$ also span $I_G/MI_G$ as a vector space, and hence so do the invariants $g_1^\ast, g_2^\ast, \ldots, g_m^\ast$. By Nakayama’s Lemma, we find that $g_1^\ast, g_2^\ast, \ldots, g_m^\ast$ generate the ideal $I_G$. As in the proof of Theorem 11.4 we conclude that $g_1^\ast, g_2^\ast, \ldots, g_m^\ast$ generate the $K$-algebra $K[x]^G$. \hfill \square

Classical invariant theory was primarily concerned with the case when $G$ is a representation of the group $SL(d, K)$ of $d \times d$-matrices with determinant 1. Here $d$ is an integer that is usually much smaller than $n$ and $K$ is a field of characteristic zero. This means that $G$ is the image of a group homomorphism $SL(d, K) \to GL(n, K)$. It is known that $SL(d, K)$ is a reductive group, i.e. there also exists an averaging operator $\ast : K[x] \to K[x]^G$ which has the same formal properties as the averaging operator of a finite group, stated in Lemma 11.3.

That Reynolds operator $\ast$ can be realized either by integration or by differentiating. In the first realization, one replaces the sum in (11.5) by an integral. Namely, one takes $K = \mathbb{C}$ and one integrates over the compact subgroup $SU(d, \mathbb{C})$ with respect to Haar measure. The same kind of integral also works in Theorem 11.6.
If $G = \text{SL}(d, \mathbb{C})$ then one can compute the Hilbert series of the invariant ring by averaging reciprocal characteristic polynomials.

An alternative to integrating with respect to Haar measure on $\text{SU}(d, \mathbb{C})$ is a certain differential operator known as Cayley’s $\Omega$-process. This process, which is explained in [38, Section 4.3], can also be used to transform arbitrary polynomials into invariants.

A third method for computing invariants is plain old linear algebra. Indeed, suppose we fix an integer $d \in \mathbb{N}$ and we seek a basis for the space $K[x]^G$ of homogeneous invariants of degree $d$. We then pick a general polynomial $f$ of degree $d$ with unknown coefficients, and we examine the equations $\sigma f = f$ for $\sigma \in G$. Each of these translates into a linear system of equations in the unknown coefficients of $f$. By taking enough matrices $\sigma$, we obtain a linear system of equations whose solutions are precisely the invariants of degree $d$. In the case when $G$ is a connected Lie group, like $\text{SL}(d, \mathbb{C})$, one can replace the condition $\sigma f = f$ by requiring that $f$ is annihilated by the associated Lie algebra. Setting up these linear equations and solving them is usually quite efficient on small examples. See [38, Section 4.5].

In what follows we take the matrix group to be an $n$-dimensional polynomial representation of $G = \text{SL}(d, K)$ for some $d, n \in \mathbb{N}$. Each of these is a direct sum of irreducible representations, one for each integer partition, as seen in Chapter 10.

**Example 11.9.** Let $U = (K^d)^m$ be the space of $d \times m$-matrices. Thus $U$ is the direct sum of $m$ copies of the defining representation of $G$. The group $G$ acts on $U$ by matrix multiplication on the left. This induces an action on the ring $K[U]$ of polynomials in the entries of a $d \times m$ matrix of variables. If $m < d$ then this action has no non-constant invariants. If $m \geq d$ then the $\binom{m}{d}$ maximal minors of the $d \times m$ matrix are invariants. This invariance holds because the determinant of the product of two $d \times d$-matrices is the product of the determinants. It is known that the invariant ring $K[U]^G$ is generated by these $\binom{m}{d}$ determinants. This result is the First Fundamental Theorem of Invariant Theory; cf [38, Section 3.2].

Note that we already encountered the ring $K[U]^G$ in Chapter 5. It is the coordinate ring of the Grassmannian of $d$-dimensional subspaces in $K^m$. Thus, $K[U]^G$ is isomorphic to a polynomial ring in $\binom{m}{d}$ variables, modulo the ideal of quadratic Plücker relations.

Arguably, the most important irreducible representations of the group $G = \text{SL}(d, K)$ are the $p$-th symmetric powers of the defining representation $K^d$, where $p \in \mathbb{N}$. We denote such a symmetric power by $V = K[u_1, \ldots, u_d]^p = \text{Sym}_p(K^d)$. Its elements are homogeneous polynomials of degree $p$ in $d$ variables. The $G$-module $V$ has dimension $n = \binom{p+d-1}{p}$. The monomials form a basis. The action of $G$ on $V$ is simply by linear change of coordinates.

**Example 11.10 (d=2, p=3).** Fix the space $V = \text{Sym}_3(K^2)$ of binary cubics

\[ f(u_1, u_2) = x_1u_1^3 + x_2u_1^2u_2 + x_3u_1u_2^2 + x_4u_2^3. \]

The coefficients $x_i$ are the coordinates on $V \simeq K^4$. The way we set things up, the group $\text{SL}(2, K)$ acts on this space by left multiplication, in its guise as the group $G$ of $4 \times 4$-matrices of the form

\[ \phi(\sigma) = \begin{pmatrix} \sigma_{11}^3 & \sigma_{11}^2\sigma_{12} & \sigma_{11}\sigma_{12}^2 & \sigma_{12}^3 \\ 3\sigma_{11}\sigma_{21} & \sigma_{11}^2\sigma_{22} + 2\sigma_{11}\sigma_{12}\sigma_{21} & \sigma_{12}\sigma_{21} + 2\sigma_{11}\sigma_{12}\sigma_{22} & 3\sigma_{12}^2\sigma_{22} \\ 3\sigma_{11}\sigma_{21}^2 & \sigma_{11}\sigma_{21}\sigma_{22} + 2\sigma_{11}\sigma_{12}\sigma_{21} & \sigma_{11}\sigma_{22}^2 + 2\sigma_{12}\sigma_{21}\sigma_{22} & 3\sigma_{12}\sigma_{22}^2 \\ \sigma_{21}^3 & \sigma_{21}\sigma_{22}^2 & \sigma_{21}\sigma_{22}\sigma_{22} & \sigma_{22}^3 \end{pmatrix}. \]
The group $G$ corresponding invariant ring is generated by two invariants
\begin{equation}
\Delta = 27x_1^2x_2^2 - 18x_1x_2x_3x_4 + 4x_1^2x_3^2 + 4x_2^2x_4 - x_2^2x_3^2.
\end{equation}
It turns out that the discriminant generates the invariant ring, i.e. $K[x]^G = K[\Delta]$.

Invariants of binary forms ($d = 2$) are a well-studied subject in invariant theory. Complete lists of generators for the invariant ring are known up to degree $p$. For $p = 2$, there is also only the discriminant $\Delta = 4x_2 - x_1x_3$. For $p = 4$, we have two generating invariants of degree 2 and 3 respectively. For $p = 10$, the invariant ring has 104 minimal generators.

### 11.3. Geometric Invariant Theory

According to Felix Klein, invariant theory plays a fundamental role for geometry. Namely, a polynomial in the coordinates of a space is invariant under the group of interest if and only if that polynomial expresses a geometric property. For instance, consider the space $V$ of binary cubics $f$ in Example 11.10. The hypersurface defined by $f$ in $\mathbb{P}^4$ consists of three points. The vanishing of the invariant $\Delta$ means that these three points are not all distinct.

In geometric invariant theory, one considers the variety $\mathcal{V}(\mathcal{I}_G)$ defined by all homogeneous invariants of positive degree. This variety is known as the nullcone.

Its points are known as unstable points. For a finite group $G$, the nullcone consists just of the origin, $\mathcal{V}(\mathcal{I}_G) = \{0\}$. For $G = \text{SL}(d, K)$ the situation is more interesting, and the geometry of the nullcone is very important for understanding the invariant ring $K[x]^G$. Corollary 11.8 says, more or less, that computing $K[x]^G$ is equivalent to finding polynomial equations that define the nullcone.

**Example 11.11** ($d = p = 3$). Consider the 10-dimensional space $V = \text{Sym}_3(K^3)$ of ternary cubics

\[ f(u) = x_1u_1^3 + x_2u_2^3 + x_3u_3^3 + x_4u_1^2u_2 + x_5u_1^2u_3 + x_6u_2^2u_3 + x_7u_1^2u_3 + x_8u_2^2u_1 + x_9u_3^2u_1 + x_{10}u_1u_2u_3. \]

The group $G = \text{SL}(3, K)$ acts on $V$ by linear change of coordinates. The corresponding invariant ring is generated by two invariants $I_4$ and $I_6$ of degrees 4 and 6 respectively. In symbols, $K[x]^G = K[I_4, I_6]$. The degree 4 invariant is the following sum of 25 monomials:

\[ I_4 = x_0^4 - 8x_0^2x_4x_9 - 8x_0^2x_5x_7 - 8x_0^2x_6x_8 - 216x_0x_1x_2x_3 + 24x_0x_1x_7x_9 + 24x_0x_2x_5x_8 + 24x_0x_3x_4x_6 + 24x_5x_6x_9 + 144x_1x_2x_8x_9 + 144x_1x_3x_6x_7 - 48x_1x_6x_9 - 48x_1x_7x_8 + 144x_2x_3x_4x_5 - 48x_2x_4x_9 - 48x_2x_5x_7 - 48x_3x_5x_9 + 16x_3^2x_5^2 - 16x_4x_5x_7 - 16x_4x_6x_8 + 16x_5^2x_7^2. \]

The degree 6 invariant is also unique up to scaling. It is a sum of 103 monomials:

\[ I_6 = x_0^6 - 12x_0^4x_4x_9 - 12x_0^4x_5x_7 - 12x_0^4x_6x_8 + 540x_0^3x_1x_2x_3 + \cdots + 96x_0^2x_2^2x_7^2 - 64x_0^2x_6x_8. \]

The invariant $I_4$ is the Aronhold invariant. This plays an important role in the theory of tensor decomposition. Indeed, we can regard $f$ as a symmetric $3 \times 3 \times 3$-tensor. A random tensor $f$ has rank 4. The Aronhold invariant $f$ vanishes for those tensors of rank $\leq 3$. In other words, $I_4 = 0$ holds if and only if $f$ is a sum of...
three cubes of linear forms, or can be approximated by a sequence of such. See the discussion of ranks of tensors two weeks ago.

On the geometric side, we identify \( f \) with the cubic curve \( V(f) \) it defines in the projective plane \( \mathbb{P}^2 \). To a number theorist, this is an \textit{elliptic curve}. An important invariant of this curve is the \textit{discriminant} \( \Delta \). This invariant has degree 12 and its explicit formula equals

\[
\Delta = I_4^3 - I_6^2.
\]

This expression vanishes if and only if the curve \( V(f) \) has a singular point. Typically, this singularity is a \textit{node}. In the special case when both \( I_4 \) and \( I_6 \) vanish, that singular point is a \textit{cusp}. Thus, for ternary cubics, the nullcone \( \mathcal{N}(G) \) is given by plane cubics that have a cusp. The moduli space of elliptic curves is parametrized by the \textit{j-invariant} plane cubics that have a cusp. The moduli space of elliptic curves is parametrized by the \( j \)-\textit{invariant}.

We now present a general-purpose algorithm, due to Harm Derksen, for computing the invariant ring of a reductive algebraic group \( G \) that acts polynomially on a vector space \( V = \mathbb{C}^n \). The group \( G \) can be represented as an algebraic variety inside \( \text{GL}(n, \mathbb{K}) \), that is, by polynomial equations in the entries of an unknown \( n \times n \)-matrix. This works for both finite groups and for polynomial representations of \( \text{SL}(d, \mathbb{K}) \), such as the ones discussed about. As before, we use the notation \( \sigma \mapsto \phi(\sigma) \) to write the representation of \( G \) on \( \mathbb{C}^n \) explicitly.

The product \( G \times V \times V \) is an algebraic variety, with coordinates \((\sigma, x, y)\). Inside its coordinate ring \( K[\sigma, x, y] \), let \( \mathcal{J}_G \) be the ideal generated by the \( n \) entries of the vector \( y - \phi(\sigma)x \). This ideal is radical, and it is prime when \( G \) is a connected group like \( \text{SL}(d, \mathbb{K}) \). Its variety describes the action of the group. The elimination ideal \( \mathcal{J}_G \cap K[x, y] \) is also radical (resp. prime). Its variety consists of pairs of points in \( V \) that lie in the same \( G \)-orbit.

**Theorem 11.12 (Derksen’s Algorithm).** The ideal \( \mathcal{I}_G \) of the nullcone is the image in \( K[x] \) of the elimination ideal \( \mathcal{J}_G \cap K[x, y] \) under the substitution \( y = 0 \). From any finite list of ideal generators of \( \mathcal{I}_G \), algebra generators for the invariant ring \( K[x]^G \) are found via Corollary 11.8.

**Proof.** Let \( I \) be any homogeneous invariant of positive degree. Then \( I(x) \equiv I(\phi(\sigma)x) \equiv I(y) \) modulo the ideal \( \mathcal{J}_G \) that defines the group action. Therefore, \( I(x) - I(y) \) lies in the elimination ideal \( \mathcal{J}_G \cap K[x, y] \), and we find \( I(x) \) in the ideal that is obtained by substituting \( y = 0 \). This proves that \( \mathcal{I}_G \) is contained in the ideal that is computed by Derksen’s Algorithm. For the converse direction, we refer to the argument given in the proof of [10] Theorem 3.1.

**Example 11.13 (p=d=2).** Consider the 3-dimensional space \( V = \text{Sym}_2(K^2) \) of binary quadrics

\[
f(u_1, u_2) = x_1 u_1^2 + x_2 u_1 u_2 + x_3 u_2^2.
\]

The coordinate ring of the variety \( \text{SL}(2, K) \times V \times V \) is the polynomial ring

\[
K[\sigma, x, y] = K[\sigma_{11}, \sigma_{12}, \sigma_{21}, \sigma_{22}, x_1, x_2, x_3, y_1, y_2, y_3]
\]

modulo the principal ideal \( \langle \sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21} - 1 \rangle \). Note that this ring has 10 generators. The ideal that encodes our action equals

\[
\mathcal{J}_G = \langle \sigma_{11}^2 x_1 + \sigma_{11}\sigma_{21}x_2 + \sigma_{21}^2 x_1 - y_1, \sigma_{12}^2 x_1 + \sigma_{12}\sigma_{22}x_2 \sigma_{22}^2 x_3 - y_3, \sigma_{21}^2 x_2 + (\sigma_{12}\sigma_{21}) x_2 + 2\sigma_{21}\sigma_{22} x_3 - y_2 \rangle
\]
Elimination of the four variables for the group elements yields the principal ideal

\[ \mathcal{J}_G \cap K[x,y] = \langle 4x_1x_3 - x_2^2 - 4y_1y_3 + y_2^2 \rangle. \]

We now set \( y_1 = y_2 = y_3 = 0 \). The result is the familiar discriminant \( \Delta = 4x_1x_3 - x_2^2 \). In this manner, Derksen’s Algorithm finds the invariant ring for binary quadrics \( \bar{K}[x]G = K[\Delta] \).

In Example 11.10, we determined the invariant ring for \( \text{SL}(2,K) \) acting on \( 2 \times 2 \times 2 \) tensors that are symmetric. In what follows, we extend this computation to non-symmetric tensors. Thus, we present case study in invariant theory for \( d = 2 \) and \( n = 8 \). We identify \( K^8 \) with the space \( (K^3)^{\otimes 3} \) of \( 2 \times 2 \times 2 \)-tensors. The corresponding polynomial ring is denoted by

\[ K[x] = K[x_{111}, x_{112}, x_{121}, x_{122}, x_{211}, x_{212}, x_{221}, x_{222}]. \]

The group \( G = \text{SL}(2,K) \) acts on \( K^2 \) by matrix-vector multiplication. This action extends naturally to the triple tensor product of \( K^2 \). Explicitly, if \( \sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \) is a \( 2 \times 2 \)-matrix in \( G \) then \( \sigma \) acts by performing the following substitution in each polynomial on \( K[x] \):

\[ (11.12) \quad x_{ijk} \mapsto \sum_{r=1}^{2} \sum_{s=1}^{2} \sum_{t=1}^{2} x_{rst} \sigma_{ri} \sigma_{sj} \sigma_{tk}. \]

Here are two nice polynomials that are invariant under this action:

**Example 11.14.** Up to scaling, there is a unique polynomial of degree 2 that is invariant under \( G = \text{SL}(2,K) \). That invariant is the following quadric, which we call the hexagon invariant:

\[ \text{Hex}(x) = x_{112}x_{122} - x_{122}x_{121} + x_{121}x_{221} - x_{221}x_{211} + x_{211}x_{212} - x_{212}x_{112}. \]

Another nice invariant is homogeneous of degree four. This is the hyperdeterminant

\[ \text{Det}(x) = \frac{x_{112}x_{122} - x_{122}x_{121} + x_{121}x_{221} - x_{221}x_{211} + x_{211}x_{212} - x_{212}x_{112}}{x_{211}x_{212}x_{221}x_{222} + x_{121}x_{122}x_{121}x_{222} + x_{111}x_{212}x_{221}x_{121} - 2x_{122}x_{211}x_{112}x_{212} - 2x_{112}x_{211}x_{112}x_{212} - 2x_{111}x_{211}x_{212}x_{222} - 2x_{111}x_{212}x_{212}x_{222}.} \]

One checks by computation that the substitution \( (11.12) \) maps the hexagon invariant \( \text{Hex}(x) \) to itself times the third power of \( \text{det}(\sigma) = \sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21} \). Similarly, the hyperdeterminant \( \text{Det}(x) \) transforms to itself times \( \text{det}(\sigma)^6 \). Hence both are invariant when \( \text{det}(\sigma) = 1 \).

Invariants can be used to test whether two tensors lie in the same orbit. Here is a concrete example. We write our \( 2 \times 2 \times 2 \) tensors as vectors in \( \mathbb{R}^8 \) as follows:

\[ c = (c_{111}, c_{112}, c_{121}, c_{122}, c_{211}, c_{212}, c_{221}, c_{222}). \]

The following two tensors appear in the theory of signatures of paths. It is of interest to know whether their \( G \)-orbits agree up to scaling:

\[ c_{\text{axis}} = (\frac{1}{6}, \frac{1}{2}, 0, \frac{1}{2}, 0, 0, \frac{1}{6}) \quad \text{and} \quad c_{\text{mono}} = (\frac{1}{6}, \frac{1}{2}, \frac{1}{3}, \frac{1}{6}, \frac{4}{15}, \frac{1}{12}, \frac{1}{10}, \frac{1}{6}). \]

The two polynomials in Example \( 11.14 \) are relative invariants of the \( \text{GL}(2) \) action on the tensor space \( \mathbb{R}^8 \). The following rational function is an absolute invariant.
It is homogeneous of degree zero, so it represents an invariant rational function on the projective space $\mathbb{P}^7$:

$$\frac{\text{Hex}(x)^2}{\text{Det}(x)}.$$ \hfill (11.13)

We find that the invariant \[(11.13)\] evaluates to 81 on $c_{\text{axis}}$, and it evaluates to 45 on $c_{\text{mono}}$. Hence the orbit closures of our two special core tensors of format $2 \times 2 \times 2$ are disjoint in $\mathbb{P}^7$.

We now come to determination of the full ring of invariants for the $G$-action on the space $K^8$ of $2 \times 2 \times 2$ tensors. Using Derksen’s Algorithm, we derive:

**Theorem 11.15.** *The invariant ring $K[x]^\text{SL(2)}$ of $2 \times 2 \times 2$ tensors has Krull dimension five. It is minimally generated by 13 invariants, namely the hexagon invariant of degree two, eight invariants of degree four (including the hyperdeterminant), and four invariants of degree six.*

In addition to the hyperdeterminant, there are three additional invariants of degree four that deserve special attention. Each has 17 terms when expanded. One of these invariants is

$$x_{111}x_{222} - x_{121}x_{212}^2 + x_{121}x_{222}x_{112} + x_{111}x_{212}x_{121}^2 + x_{111}x_{222}x_{211}^2$$ \hfill (11.14)

The other two invariants in this family are obtained by permuting indices.

**Corollary 11.16.** *The three quartics in (11.14) together with Hex and Det form an algebraically independent system of five primary invariants. All other invariants in $K[x]^\text{SL(2)}$ are integral over the polynomial subring generated by these five. The five primary invariants cut out the null cone $\mathcal{V}(K[x]^\text{SL(2)})$, which is a variety of dimension four and degree 12 in $\mathbb{P}^7$.*

It is instructive to restrict the 13 generating invariants in Theorem 11.15 to the 4-dimensional subspace $\text{Sym}_3(K^2)$ of symmetric $2 \times 2 \times 2$ tensors, seen in Example 11.10. We do this by setting

$$x_{111} = x_1, \quad x_{112} = x_{121} = x_{211} = \frac{1}{3}x_2, \quad x_{122} = x_{212} = x_{221} = \frac{1}{3}x_3, \quad x_{222} = x_4.$$  

The resulting symmetric tensors correspond to binary cubics (11.8). The hyperdeterminant and five other generators of degree four specialize to the discriminant $\Delta$ of the binary cubic. The other eight generators of $K[x]^\text{SL(2)}$, including the hexagon invariant, specialize to zero. In this manner, the invariant ring in Theorem 11.15 maps onto the invariant ring of binary cubics.

**Exercises**

1. Let $G$ be the symmetry group of the square $[-1, 1]^2$ in the plane $\mathbb{R}^2$. This is an order 8 subgroup in $\text{GL}(2, \mathbb{R})$. List all eight matrices. Determine the invariant ring $\mathbb{R}[x_1, x_2]^G$.

2. Let $G$ be the symmetry group of the regular 3-cube, as a subgroup of $\text{GL}(3, \mathbb{R})$. How many matrices are in $G$, and what are their characteristic polynomials? Determine the Molien series (11.7) of this group. What does it tell you about the invariant ring?
(3) Fix \( n = 5 \). Let \( \psi(j) \) denote the number of monomials in the expansion of the power sum \( P_j \) in terms of the elementary symmetric functions \( E_1, E_2, E_3, E_4, E_5 \). Compute \( \psi(j) \) for some small values, say \( j \leq 20 \). Guess a formula for \( \psi(j) \). Can you prove it?

(4) Show that Noether’s Degree Bound is always tight for finite cyclic groups.

(5) Find a subgroup of \( \text{GL}(4, K) \) that has order 15. Compute the invariant ring.

(6) Let \( T \) be the group of \( 3 \times 3 \) diagonal matrices with determinant 1, acting on the space \( V = \text{Sym}_3(K^3) \) of ternary cubics. This group is the torus \( T \cong (K^*)^2 \). Determine the invariant ring \( K[V]^T \). Do you see any relationship to the invariants in Example 11.11?

(7) Let \( G = A_n \) be the alternating group of order \( n! / 2 \). Its elements are the even permutation matrices. Determine the invariant ring \( K[x]^G \).

(8) List all 103 monomials of the invariant \( I_6 \) of ternary cubics in Example 11.11. Give an explicit formula, in terms of \( x_1, x_2, \ldots, x_9, x_0 \), for the discriminant and the \( j \)-invariant.

(9) Consider the action of \( \text{SL}(3, K) \) on the space \( \text{Sym}_2(K^3) \cong K^6 \) of symmetric \( 3 \times 3 \)-matrices. The entries of the \( 6 \times 6 \) matrix \( \phi(\sigma) \) are quadratic forms in \( \sigma_{11}, \sigma_{12}, \ldots, \sigma_{33} \). Write this matrix explicitly, similarly to (11.9). What is the invariant ring?

(10) Using Derksen’s Algorithm, determine the invariant ring for binary quartics \((d = 2, p = 4)\). How many minimal generators does this ring have?

(11) The rotation group \( \text{SO}(2, \mathbb{R}) \) acts by left multiplication on the space of \( 2 \times 2 \)-matrices. Determine the invariant ring.

(12) Is the invariant ring of every matrix group \( G \subset \text{GL}(n, K) \) finitely generated?
CHAPTER 12

Semidefinite Programming

The transition from linear algebra to nonlinear algebra has a natural counterpart in convex optimization, namely the passage from linear programming to semidefinite programming. This transition is the topic of this chapter. Linear programming concerns the solution of linear systems of inequalities, and the optimization of linear functions subject to linear constraints. The feasible region is a convex polyhedron, and the optimal solutions form a face of that polyhedron. In semidefinite programming we work in the space of symmetric $n \times n$-matrices. The inequality constraints now stipulate that some linear combination of matrices be positive semidefinite. The feasible region given by such constraints is a closed convex set, known as a spectrahedron. We again wish to optimize a linear function. The condition for a polynomial to be a sum of squares as a semidefinite program. This furnishes a connection to the real Nullstellensatz (Chapter 6), thereby establishing semidefinite programming as a key tool for computing in real algebraic geometry.

12.1. Spectrahedra

In this chapter we work over the field $\mathbb{R}$ of real numbers. The Spectral Theorem in Linear Algebra states that all eigenvalues of a symmetric matrix $A \in \text{Sym}_2(\mathbb{R}^n)$ are real. Moreover, there is an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of $A$. We say that the matrix $A$ is positive definite if it satisfies the following conditions.

1. All $n$ eigenvalues of $A$ are positive real numbers.
2. All $2^n$ principal minors of $A$ are positive real numbers.
3. Every non-zero column vector $u \in \mathbb{R}^n$ satisfies $u^T A u > 0$.

Here, by a principal minor we mean the determinant of any square submatrix of $A$ whose set of column indices agree with its set of row indices. For the empty set, we get the $0 \times 0$ minor of $A$, which equals 1. Next there are the $n$ diagonal entries of $A$, which are the $1 \times 1$ principal minors, and finally the determinant of $A$, which is the unique $n \times n$ principal minor. Each of the three conditions (1), (2) and (3) behaves as expected when we pass to the closure. This is not obvious because the closure of an open semialgebraic set $\{ f > 0 \}$, where $f \in \mathbb{R}[x]$, is generally smaller than the corresponding closed semialgebraic set $\{ f \geq 0 \}$.

**Example 12.1.** Let $f = x^3 + x^2 y + x y^2 + y^3 - x^2 - y^2$. The set $\{ f > 0 \}$ is the open halfplane above the line $x + y = 1$ in $\mathbb{R}^2$. The closure of the set $\{ f > 0 \}$ is the corresponding closed halfplane. It is properly contained in $\{ f \geq 0 \}$ which also contains the origin $(0,0)$.

Luckily, no such thing happens with condition (2) for positive definite matrices.
Theorem 12.2. For a symmetric $n \times n$ matrix $A$, the following three conditions are equivalent:

(1') All $n$ eigenvalues of $A$ are nonnegative real numbers.
(2') All $2^n$ principal minors of $A$ are nonnegative real numbers.
(3') Every non-zero column vector $u \in \mathbb{R}^n$ satisfies $u^T A u \geq 0$.

If this holds then $A$ is called positive semidefinite. The semialgebraic set $\text{PSD}_n$ of positive semidefinite $n \times n$ matrices is a full-dimensional closed convex cone in $\text{Sym}_2(\mathbb{R}^n)$.

We use the notation $X \succeq 0$ to express that a symmetric matrix $X$ is positive semidefinite. A spectrahedron $S$ is the intersection of the cone $\text{PSD}_n$ with an affine-linear subspace $\mathcal{L}$ of the ambient space $\text{Sym}_2(\mathbb{R}^n)$. Hence, spectrahedra are closed convex semialgebraic sets.

A subspace $\mathcal{L}$ of symmetric matrices is either given parametrically, or as the solution set to an inhomogeneous system of linear equations. In the equational representation, we write

\[(12.1) \quad \mathcal{L} = \{ X \in \text{Sym}_2(\mathbb{R}^n) : \langle A_1, X \rangle = b_1, \langle A_2, X \rangle = b_2, \ldots, \langle A_s, X \rangle = b_s \}.
\]

Here $A_1, A_2, \ldots, A_s \in \text{Sym}_2(\mathbb{R}^n)$ and $b_1, b_2, \ldots, b_s \in \mathbb{R}$ are fixed. We employ the standard inner product in the space of square matrices, which is given by the trace of the matrix product:

\[(12.2) \quad \langle A, X \rangle := \text{trace}(AX) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_{ij}.
\]

The associated spectrahedron $S = \mathcal{L} \cap \text{PSD}_n$ consists of all positive semidefinite matrices that lie in the subspace $\mathcal{L}$. If the subspace is given by a parametric representation, say

\[(12.3) \quad \mathcal{L} = \{ A_0 + x_1 A_1 + \cdots + x_s A_s : (x_1, \ldots, x_s) \in \mathbb{R}^s \},
\]

then it is customary to identify the spectrahedron with its preimage in $\mathbb{R}^s$. Hence,

\[(12.4) \quad S = \{ (x_1, \ldots, x_s) \in \mathbb{R}^s : A_0 + x_1 A_1 + \cdots + x_s A_s \succeq 0 \}.
\]

Proposition 12.3. Every convex polyhedron is a spectrahedron. Convex polyhedra are precisely the spectrahedra that arise when the subspace $\mathcal{L}$ consists only of diagonal $n \times n$ matrices.

Proof. Suppose that the matrices $A_0, A_1, \ldots, A_s$ are diagonal matrices. Then \[(12.4)\] is the solution set in $\mathbb{R}^s$ of a system of $n$ inhomogeneous linear inequalities. Such a set is a convex polyhedron. Every convex polyhedron in $\mathbb{R}^s$ has such a representation. We simply write its defining linear inequalities as the diagonal entries of the matrix $A_0 + x_1 A_1 + \cdots + x_s A_s$.

The formula $S = \mathcal{L} \cap \text{PSD}_n$ with $\mathcal{L}$ as in \[(12.1)\] corresponds to the standard representation of a convex polyhedron, as the set of non-negative points in an affine-linear space. Here the equations in \[(12.1)\] include those that require the off-diagonal entries of all matrices to be zero:

\[\langle X, E_{ij} \rangle = x_{ij} = 0 \quad \text{for} \; i \neq j.
\]

In the other inequalities, the matrices $A_i$ are diagonal and the $b_i$ are typically nonzero. \qed
Example 12.4. Let $\mathcal{L}$ be the space of symmetric $3 \times 3$ matrices whose three diagonal entries are all equal to $1$. This is an affine-linear subspace of dimension $s = 3$ in $\text{Sym}_2(\mathbb{R}^3) \simeq \mathbb{R}^6$. The spectrahedron $\mathcal{S} = \mathcal{L} \cap \text{SDP}_3$ is the yellow convex body seen in Chapter 1, Figure 1. To draw this spectrahedron in $\mathbb{R}^3$, one uses the representation (12.4), namely

$$\mathcal{S} = \left\{ (x, y, z) \in \mathbb{R}^3 : \begin{pmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{pmatrix} \succeq 0 \right\}.$$ 

The boundary of $\mathcal{S}$ consists of all points $(x, y, z)$ where the matrix has determinant zero and its nonzero eigenvalues are positive. The determinant is a polynomial of degree three in $x, y, z$, so the boundary lies in cubic surface in $\mathbb{R}^3$. This cubic surface also contains points where the three eigenvalues are positive, zero and negative. Such points are drawn in red in our picture from Chapter 1. They lie in the Zariski closure of the yellow boundary points.

We next slice our 3-dimensional spectrahedron to get a picture in the plane.

Example 12.5. Suppose that $\mathcal{L} \subset \text{Sym}_2(\mathbb{R}^3)$ is a general plane that intersects the cone $\text{PSD}_3$. The spectrahedron $\mathcal{S}$ is a planar convex body whose boundary is a smooth cubic curve, drawn in red in Figure 1. On that boundary, the $3 \times 3$ determinant vanishes and the other two eigenvalues are positive. For points $(x, y) \in \mathbb{R}^2 \setminus \mathcal{S}$, the matrix has at least one negative eigenvalue. The black curve lie in the Zariski closure of the red curve. It separates points in $\mathbb{R}^2 \setminus \mathcal{S}$ whose remaining two eigenvalues are positive from those with two negative eigenvalues.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{A plane curve of degree three (left) and its dual curve of degree six (right). The red part on the left bounds a spectrahedron while that on the right bounds its convex dual.}
\end{figure}

To be explicit, suppose that our planar cubic spectrahedron is defined as follows:

$$\mathcal{S} = \left\{ (x, y) \in \mathbb{R}^3 : \begin{pmatrix} 1 & x & x + y \\ x & 1 & y \\ x + y & y & 1 \end{pmatrix} \succeq 0 \right\}. \tag{12.5}$$

The cubic curve is the locus where the $3 \times 3$ matrix is singular. Its determinant is

$$f = 2x^2 y + 2xy^2 - 2x^2 - 2xy - 2y^2 + 1. \tag{12.6}$$
The curve \( \{ f = 0 \} \) has four connected components in \( \mathbb{R}^2 \), one in red and three in black, as shown in Figure 1 (left). The boundary of the cubic spectrahedron \( S \) is the convex part of the curve that is shown in red.

The picture on the right in Figure 1 shows the dual curve. This lives in the dual plane whose points \((u,v)\) represent the lines \( \ell = \{(x,y) : ux + vy = 1\} \) in \( \mathbb{R}^2 \). The points in the dual curve correspond to lines \( \ell \) that are tangent to the original curve. The dual curve has degree six, and its equation is computed by the following ideal computation in \( \mathbb{R}[x,y,u,v] \):

\[
\langle f(x,y), u \cdot x + v \cdot y - 1, \partial f/\partial x \cdot v - \partial f/\partial y \cdot u \rangle \cap \mathbb{R}[u,v] = \langle 8u^6 - 24u^5v + 21u^4v^2 - 2u^3v^3 + 21u^2v^4 - 24uv^5 + 8v^6 - 24u^5 + 60u^4v - 24u^3v^2 - 24u^2v^3 + 60uv^4 - 24v^5 + 12u^4 - 24u^3v + 36u^2v^2 - 24uv^3 + 12v^4 + 24u^3 - 36u^2v - 36uv^2 + 24v^3 - 24u^2 + 24uv - 24v^2 + 4 \rangle.
\]

The points in the sextic correspond to lines that are tangent at black points of the cubic, and similarly for the red points. Moreover, the convex set enclosed by the red sextic on the right in Figure 1 is dual, in the sense of convexity, to the spectahedron on the left.

The polynomials in (12.6) and (12.7) have degree three and six respectively, confirming what was asserted in the caption to Figure 1. A random line \( L \) will meet the curve in three (left) or six (right) complex points. Consider the point on the other side that is dual to \( L \). There are three (right) or six (left) complex lines through that point that are tangent to the curve.

### 12.2. Optimization and Duality

We now finally come to semidefinite programming (SDP). This refers to the problem of maximizing or minimizing a linear function over a spectrahedron. Linear programming is the special case when the spectrahedron consists of diagonal matrices. If the spectrahedron is given in its standard form representation (12.1), then we get the SDP in its primal form:

\[
\text{Minimize } \langle C, X \rangle \text{ subject to } \langle A_1, X \rangle = b_1, \langle A_2, X \rangle = b_2, \ldots, \langle A_s, X \rangle = b_s \text{ and } X \succeq 0.
\]

Here \( C = (c_{ij}) \) is a matrix that represents the cost function. Every convex optimization problem has a dual problem. On first glance, it is not so easy to relate that duality to those for plane curves in Figure 1. The semidefinite problem dual to (12.8) takes the following form

\[
\text{Maximize } b^T x = \sum_{i=1}^s b_i x_i \text{ subject to } C - x_1 A_1 - x_2 A_2 - \cdots - x_s A_s \succeq 0.
\]

In this formulation, the spectrahedron of feasible points lives in \( \mathbb{R}^s \), similarly to (12.4). We refer to either formulation (12.8) or (12.9) as a semidefinite program, also abbreviated SDP. Here the term “program” is simply an old-fashioned way of saying “optimization problem”. The relationship between the primal and the dual SDP is given by the following theorem:

**Theorem 12.6 (Weak Duality).** If \( x \) is any feasible solution to (12.9) and \( X \) is any feasible solution to (12.8) then \( b^T x \leq \langle C, X \rangle \). If the equality \( b^T x = \langle C, X \rangle \) holds then both \( x \) and \( X \) are optimal.
The term feasible means only that the point \( x \) resp. \( X \) satisfies the equations and inequalities that are required in (12.8) resp. (12.9). The point is optimal if it is feasible and it solves the program, i.e. it attains the minimum resp. maximum value for that optimization problem.

**Proof.** The inner product of two positive semidefinite matrices is a non-negative real number:

\[
0 \leq \langle C - \sum_{i=1}^{s} x_i A_i, X \rangle = \langle C, X \rangle - \sum_{i=1}^{s} x_i \cdot \langle A_i, X \rangle = \langle C, X \rangle - b^T x.
\]

This shows that the optimal value of the minimization problem (12.8) is an upper bound for the optimal value of the maximization problem (12.9). If the equality is attained by a pair \((X, x)\) of feasible solutions then \( X \) must be optimal for (12.8) and \( x \) must be optimal for (12.9).

There is also Strong Duality Theorem which states that, under suitable hypotheses, the duality gap \( \langle C, X \rangle - b^T x \) must attain the value zero for some feasible pair \((X, x)\). These hypotheses are always satisfied for diagonal matrices, and we recover the Duality Theorem for Linear Programming as a special case. Interior point methods for Linear Programming are numerical algorithms that start at an interior point of the feasible polyhedron and create a path from that point towards an optimal vertex. The same class of algorithms works for Semidefinite Programming. These run in polynomial time and are well-behaved in practice.

Semidefinite Programming has a much larger expressive power than Linear Programming. Many more problems can be phrased as an SDP. We illustrate this with a simple example.

**Example 12.7 (The largest eigenvalue).** Let \( A \) be a real symmetric \( n \times n \) matrix, and consider the problem of computing its largest eigenvalue \( \lambda_{\text{max}}(A) \). We would like to solve this without having to write down the characteristic polynomial and extract its roots. Let \( C = \text{Id} \) be the identity matrix and consider the SDP problems (12.8) and (12.9) with \( s = 1 \) and \( b = 1 \). They are

- Minimize \( \text{trace}(X) \) subject to \( \langle A, X \rangle = 1 \).
- Maximize \( x \) subject to \( \text{Id} - xA \succeq 0 \).

If \( x^* \) is the common optimal value of these two problems then \( \lambda_{\text{max}}(A) = 1/x^* \).

The inner product \( \langle A, X \rangle = \text{trace}(A \cdot X) \) of two positive semidefinite matrices \( A \) and \( X \) can only be zero when their matrix product \( A \cdot X \) is zero. We record this for our situation:

**Lemma 12.8.** If the expression in (12.10) is zero then \( (C - \sum_{i=1}^{s} x_i A_i) \cdot X \) is the zero matrix.

This lemma allows us to state the following algebraic reformulation of SDP:

**Corollary 12.9.** Consider the following system of \( s \) linear equations and \( \binom{n+1}{2} \) bilinear equations in the \( \binom{n+1}{2} + s \) unknown coordinates of the pair \((X, x)\):

\[
\langle A_1, X \rangle = b_1, \ldots, \langle A_s, X \rangle = b_s \quad \text{and} \quad (C - \sum_{i=1}^{s} x_i A_i) \cdot X = 0.
\]

If \( X \succeq 0 \) and \( C - \sum_{i=1}^{s} x_i A_i \succeq 0 \) then \( X \) is optimal for (12.8) and \( x \) is optimal for (12.9).
The equations \[ (12.11) \] are known as the Karush-Kuhn-Tucker (KKT) equations. These play a major role when one explores semidefinite programming from an algebraic perspective. In particular, they allow us to study the nature of the optimal solution as function of the data. A key feature of the KKT system is that the two optimal matrices have complementary ranks. This follows from the complementary slackness condition on the right of \[ (12.11) \]:

\[
\text{rank}(C - \sum_{i=1}^{s} x_i A_i) + \text{rank}(X) \leq n.
\]

In particular, if \( X \) is known to be nonzero then the determinant of \( C - \sum_{i=1}^{s} x_i A_i \) vanishes. For instance, for the eigenvalue problem in Example 12.7, we have \((\text{Id} - x A) \cdot X = 0\) and \( \langle A, X \rangle = 1 \). This implies \( \det(\text{Id} - x A) = 0 \), so \( 1/x \) is a root of the characteristic polynomial.

**Example 12.10.** Consider the problem of maximizing a linear function \( \ell(x, y) = ux + vy \) over the spectrahedron \( S \) in \[ (12.5) \]. This is the primal SDP \[ (12.8) \] with \( s = 2 \) and \( b = (u, v) \) and

\[
A_1 = -\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad A_2 = -\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.
\]

The KKT system \[ (12.11) \] consists of eight equations in eight unknowns, with two parameters:

\[
2x_{12} + 2x_{13} + u = 2x_{13} + 2x_{23} + v = 0 \quad \text{and} \quad
\begin{pmatrix} 1 & x & x + y \\ x & 1 & y \\ x + y & y & 1 \end{pmatrix} \cdot \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{12} & x_{22} & x_{23} \\ x_{13} & x_{23} & x_{33} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.
\]

By eliminating the variables \( x_{ij} \) we obtain an ideal \( I \) in \( \mathbb{Q}[u, v, x, y] \) that characterizes the optimal solution \((x^*, y^*)\) to our SDP as an algebraic function of \((u, v)\). Let \( \ell^* \) now be a new unknown, and consider the elimination ideal \((I + \langle ux + vy - \ell^* \rangle) \cap \mathbb{Q}[u, v, \ell^*] \). Its generator is a ternary sextic in \( u, v, \ell^* \). This is precisely the homogenization of the dual sextic in \[ (12.7) \]. It expresses the optimal value \( \ell^* \) as an algebraic function of degree six in the cost \((u, v)\).

This relationship between the dual hypersurface and the optimal value function generalizes to arbitrary polynomial optimization problems, including semidefinite programs. This is the content of [3, Theorem 5.23]. We refer to the book [3], and especially Chapter 5, for further reading on spectrahedra, semidefinite programming, and the relevant duality theory.

A fundamental task in Convex Algebraic Geometry [3] is the computation of the convex hull of a given algebraic variety or semialgebraic set. Recall that the convex hull of a set is the small convex set containing the given set. Spectrahedra or their linear projections, known as spectrahedral shadows, can be used for this task. This matters for optimization since minimizing a linear function over a set is equivalent to minimizing over its convex hull.
Example 12.11 (Toeplitz Spectrahedron). Consider the convex body
\[
\{(x, y, z) \in \mathbb{R}^3 : \begin{bmatrix} 1 & x & y & z \\ x & 1 & x & y \\ y & x & 1 & x \\ z & y & x & 1 \end{bmatrix} \succeq 0 \}.
\]
The determinant of the given Toeplitz matrix of size \(4 \times 4\) factors as
\[(x^2 + 2xy + y^2 -xz -x-z-1)(x^2 - 2xy + y^2 -xz + x+z -1).
\]
The Toeplitz spectrahedron (12.12) is the convex hull of the cosine moment curve
\[\{(\cos(\theta), \cos(2\theta), \cos(3\theta)) : \theta \in [0, \pi]\}.
\]
The curve and its convex hull are shown on the left in Figure 2. The two endpoints, \((x, y, z) = (1, 1, 1)\) and \((x, y, z) = (-1, 1, -1)\), correspond to rank 1 matrices. All other points on the curve have rank 2. To construct the Toeplitz spectrahedron geometrically, we form the cone from each endpoint over the cosine curve, and we intersect these two quadratic cones. The two cones intersect along this curve and the line through the endpoints of the cosine curve.

Shown on the right in Figure 2 is the convex body dual to the Toeplitz spectrahedron. It is the set of trigonometric polynomials \(1 + a_1 \cos(\theta) + a_2 \cos(2\theta) + a_3 \cos(3\theta)\) that are nonnegative on \([0, \pi]\). This convex body is not a spectrahedron because it has a non-exposed edge (cf. [3 Exercise 6.13]).

12.3. Sums of Squares

Semidefinite programming can be used to model and solve arbitrary polynomial optimization problems. The key to this is the representation of nonnegative polynomials in terms of sums of squares, or, more generally, the Real Nullstellensatz (cf. Chapter 6). We explain this for the simplest scenario, namely the problem of unconstrained polynomial optimization.

Let \(f(x_1, \ldots, x_n)\) be a polynomial of even degree \(2p\), and suppose that \(f\) attains a minimal real value \(f^*\) on \(\mathbb{R}^n\). Our goal is to compute \(f^*\) and a point \(u^* \in \mathbb{R}^n\) such that \(f(u^*) = f^*\). Minimizing a function is equivalent to finding the best possible
lower bound \( \lambda \) for that function. Our goal is therefore equivalent to solving the following optimization problem:

\[
(12.13) \quad \text{Maximize } \lambda \text{ such that } f(x) - \lambda \geq 0 \text{ for all } x \in \mathbb{R}^n.
\]

This is a difficult problem. Instead, we consider the following relaxation:

\[
(12.14) \quad \text{Maximize } \lambda \text{ such that } f(x) - \lambda \text{ is a sum of squares in } \mathbb{R}[x].
\]

Here relaxation means that we enlarged the set of feasible solutions. Indeed, every sum of squares is nonnegative, but not every nonnegative polynomial is a sum of squares of polynomials. For instance, the Motzkin polynomial \( x^4y^2 + x^2y^4 + 1 - 3x^2y^2 \) is nonnegative but it is not a sum of squares of polynomials. For that reason, the optimal value of \((12.14)\) is always a lower bound for the optimal value of \((12.13)\), but the two values can be different in some cases. However, here is the good news:

**Proposition 12.12.** The optimization problem \((12.14)\) is a semidefinite program.

**Proof.** Let \( x^{[p]} \) be the column vector whose entries are all monomials in \( x_1, \ldots, x_n \) of degree \( \leq p \). Thus \( x^{[p]} \) has length \( \binom{n+p}{n} \). Let \( G = (g_{ij}) \) be a symmetric \( \binom{n+p}{n} \times \binom{n+p}{n} \) matrix with unknown entries. Then \( (x^{[p]})^T \cdot G \cdot x^{[p]} \) is a polynomial of degree \( d = 2p \) in \( x_1, \ldots, x_n \). We set

\[
(12.15) \quad f(x) - \lambda = (x^{[p]})^T \cdot G \cdot x^{[p]}.
\]

By collecting coefficients of the \( x \)-monomials, this gives a system of \( \binom{2p+n}{n} \) linear equations in the unknowns \( g_{ij} \) and \( \lambda \). The number of unknowns is \( \binom{(n+p+1)}{2} + 1 \).

Suppose the linear system \((12.15)\) has a solution \((G, \lambda)\) such that \( G \) is positive semidefinite. Then we can write \( G = H^T H \) where \( H \) is a real matrix with \( r \) rows and \( \binom{p+n}{n} \) columns. (This is known as a Cholesky factorization of \( H \).) The polynomial in \((12.15)\) then equals

\[
(12.16) \quad f(x) - \lambda = (Hx^{[p]})^T \cdot (Hx^{[p]}).
\]

This is the scalar product of a vector of length \( r \) with itself. Hence \( f(x) - \lambda \) is a sum of squares. Conversely, every representation of \( f(x) - \lambda \) as a sum of squares of polynomials uses polynomials of degree \( \leq p \), and it can hence be written in the form as in \((12.16)\).

Our argument shows that the optimization problem \((12.14)\) is equivalent to

\[
(12.17) \quad \text{Maximize } \lambda \text{ subject to } (G, \lambda) \text{ satisfying the linear equations } (12.16) \text{ and } G \succeq 0.
\]

This is a semidefinite programming problem, and so the proof is complete. \( \square \)

If \( n = 1 \) or \( d = 2 \) or \((n = 2 \text{ and } d = 4)\) then every nonnegative polynomial is a sum of squares. In those special cases, problems \((12.13)\) and \((12.17)\) are equivalent.

**Example 12.13** \((n = 1, p = 2, d = 4)\). Suppose we seek to find the minimum of the degree 4 polynomial \( f(x) = 3x^4 + 4x^3 - 12x^2 \). Of course, we know how to do this using Calculus. However, we here present the SDP approach. The linear equations \((12.15)\) have a one-dimensional space of solutions. Introducing a parameter \( \mu \) for that line, the solutions can be written as

\[
(12.18) \quad f(x) - \lambda = \begin{pmatrix} x^2 & x \\ x & 1 \end{pmatrix} \begin{pmatrix} 3 & 2 & \mu - 6 \\ 2 & -2\mu & 0 \\ \mu - 6 & 0 & -\lambda \end{pmatrix} \begin{pmatrix} x^2 \\ x \\ 1 \end{pmatrix}.
\]
Consider the set of all pairs $(\lambda, \mu)$ such that the $3 \times 3$ matrix in (12.18) is positive semidefinite. This set is a cubic spectrahedron in the plane $\mathbb{R}^2$, just like that shown on the left in (1). We seek to maximize $\lambda$ over all points in that cubic spectrahedron. The optimal point equals $(\lambda^*, \mu^*) = (-32, -2)$. Substituting this into the matrix in (12.18) we obtain a positive definite matrix of rank 2. This can be factored as $G = H^T H$, where $H$ has format $2 \times 3$. The resulting representation (12.16) as a sum of two squares equals

$$f(x) - \lambda^* = f(x) + 32 = \left((\sqrt{3}x - \frac{4}{\sqrt{3}}) \cdot (x + 2)\right)^2 + \frac{8}{3}(x + 2)^2.$$ 

The right hand side is nonnegative for all $x$. It takes the value 0 only when $x = -2$.

Any polynomial optimization problem can be translated into a relaxation that is a semidefinite programming problem. If we are minimizing $f(x)$ subject to some polynomial constraints, then we seek a certificate for $f(x) - \lambda < 0$ to have no solution. This certificate is promised by the Real Nullstellensatz or Positivstellensatz. If we fix a degree bound then the existence of a certificate translates into a semidefinite program, and so does the additional requirement for $\lambda$ to be minimal. This relaxation may or may not give the correct solution for some fixed degree bound. However, if one increases the degree bound then the SDP formulation is more likely to succeed, albeit at the expense of having to solve a much larger problem. This is a powerful and widely used approach to polynomial optimization, known as SOS programming. The term Lasserre hierarchy refers to varying the degree bounds.

Every spectrahedron $S = L \cap \text{PSD}_n$ has a special point in its relative interior. This point, defined as the unique matrix in $S$ whose determinant is maximal, is known as analytic center. Finding the analytic center of $S$ is a convex optimization problem, since the function $X \mapsto \log \det(X)$ is strictly convex on the cone of positive definite matrices $X$. The analytic center is important for semidefinite programming because it serves as the starting point for interior point methods. Indeed, the central path of an SDP starts at the analytic center and runs to the optimal face. It is computed by a sequence of numerical approximations.

**Example 12.14.** The determinant function takes on all values between 0 and 1 on the spectrahedron $S$ in (12.5). The value 1 is attained only by the identity matrix, for $(x, y) = (0, 0)$. This point is therefore the analytic center of $S$.

We close by relating spectrahedra and their analytic centers to statistics. Every positive definite $n \times n$ matrix $\Sigma = (\sigma_{ij})$ is the covariance matrix of a multivariate normal distribution. Its inverse $\Sigma^{-1}$ is the concentration matrix of that distribution.

A Gaussian graphical model is specified by requiring that some off-diagonal entries of $\Sigma^{-1}$ are zero. These entries correspond to the non-edges of the graph. Maximum likelihood estimation for this graphical model translates into a matrix completion problem. Suppose that $S$ is the sample covariance matrix of a given data set. We regard $S$ as a partial matrix, with visible entries only on the diagonal and on the edges of the graph. One considers the set of all completions of the non-edge entries that make the matrix $S$ positive definite. The set of all these completions is a spectrahedron. Maximum likelihood estimate for the data $S$ in the graphical model amounts to maximizing the logarithm of the determinant. We hence seek to compute the analytic center of the spectrahedron of all completions.
Example 12.15 (Positive definite matrix completion). Suppose that the eight entries \( \sigma_{ij} \) in the following symmetric \( 4 \times 4 \)-matrix are visible, but the entries \( x \) and \( y \) are unknown:

\[
\Sigma = \begin{pmatrix}
\sigma_{11} & \sigma_{12} & x & \sigma_{14} \\
\sigma_{12} & \sigma_{22} & \sigma_{23} & y \\
x & \sigma_{23} & \sigma_{33} & \sigma_{34} \\
\sigma_{14} & y & \sigma_{34} & \sigma_{44}
\end{pmatrix}.
\]

This corresponds to the graphical model of the four-cycle \( 12,23,34,41 \). Given visible entries \( \sigma_{ij} \), we consider the set of pairs \((x, y)\) that make \( \Sigma \) positive definite. This is the interior of a planar spectrahedron \( S_\sigma \), bounded by a quartic curve. The MLE is the analytic center of \( S_\sigma \).

One is also interested in conditions on the \( \sigma_{ij} \) such that \( \text{int}(S_\sigma) \) is non-empty. When can we find \((x, y)\) that make \( \Sigma \) positive definite? A necessary condition is that the diagonal entries \( \sigma_{ii} \) and the four visible principal \( 2 \times 2 \)-minors are positive:

\[
\sigma_{11} \sigma_{22} > \sigma_{12}^2, \quad \sigma_{22} \sigma_{33} > \sigma_{23}^2, \quad \sigma_{33} \sigma_{44} > \sigma_{34}^2, \quad \sigma_{11} \sigma_{44} > \sigma_{14}^2.
\]

But this is not sufficient. The answer is a cone that is bounded by the hypersurface

\[
\sigma_{11}^2 - \sigma_{12}^2 + 2 \sigma_{22}^2 \sigma_{12}^2 - \sigma_{23}^2 + \sigma_{33}^2 - \sigma_{11} \sigma_{33} \sigma_{44}^2 - \sigma_{12} \sigma_{23}^2 - 2 \sigma_{11} \sigma_{12} \sigma_{23}^2 + \sigma_{22}^2 \sigma_{12}^2 - 2 \sigma_{11} \sigma_{22} \sigma_{33} \sigma_{44} - \sigma_{14}^2 - \sigma_{23}^2 - 2 \sigma_{34}^2 + \sigma_{11} \sigma_{33} \sigma_{44}^2 - \sigma_{12} \sigma_{23}^2 + \sigma_{22}^2 \sigma_{12}^2 - \sigma_{11} \sigma_{12} \sigma_{23}^2 - \sigma_{22}^2 \sigma_{12}^2 - 2 \sigma_{11} \sigma_{22} \sigma_{33} \sigma_{44} - \sigma_{14}^2 - \sigma_{23}^2 - 2 \sigma_{34}^2 + \sigma_{11} \sigma_{33} \sigma_{44}^2 - \sigma_{12} \sigma_{23}^2 + \sigma_{22}^2 \sigma_{12}^2 - \sigma_{11} \sigma_{12} \sigma_{23}^2 - \sigma_{22}^2 \sigma_{12}^2 - 2 \sigma_{11} \sigma_{22} \sigma_{33} \sigma_{44} - \sigma_{14}^2 - \sigma_{23}^2.
\]

This polynomial of degree eight is found by eliminating \( x \) and \( y \) from the determinant and its partial derivatives with respect to \( x \) and \( y \), after saturating by the ideal of \( 3 \times 3 \)-minors. For more details on this example see to [41], Theorem 4.8.

Exercises

1. Prove Theorem 12.2.
2. Show that a real symmetric matrix \( G \) is positive semidefinite if and only if it admits a Cholesky factorization \( G = H^TH \) over the real numbers, with \( H \) upper triangular.
3. What is the largest eigenvalue of any of the \( 3 \times 3 \) matrices in the set \( S \) in (12.5)?
4. Maximize and minimize the linear function \( 13x + 17y + 23z \) over the spectrahedron \( S \) in Example 12.4. Use SDP software if you can.
5. Maximize and minimize the linear function \( 13x + 17y + 23z \) over the Toeplitz spectrahedron in Example 12.11. Use SDP software if you can.
6. Write the dual SDP and solve the KKT system for the previous two problems.
7. Determine the convex body dual to the spectrahedron \( S \) in Example 12.4.
8. Consider the problem of minimizing the univariate polynomial \( x^6 + 5x^5 + 3x^4 + x \). Express this problem as a semidefinite program.
9. In the partial matrix (12.19) set \( \sigma_{11} = \sigma_{22} = \sigma_{33} = \sigma_{44} = 5, \sigma_{12} = \sigma_{23} = \sigma_{34} = 1 \) and \( \sigma_{14} = 2 \). Compute the spectrahedron \( S_\sigma \), draw a picture, and find the analytic center.
10. Find numerical values for the eight entries \( \sigma_{ij} \) in (12.19) that satisfy (12.20) but \( \text{int}(S_\sigma) = \emptyset \).
CHAPTER 13

Combinatorics

“Anyone who has worked with matroids has come away with the conviction that matroids are one of the richest and most useful mathematical ideas of our days.”, Gian-Carlo Rota

It has been said that combinatorics is the “nanotechnology of mathematics”. This technology interacts in many fruitful ways with algebra and geometry, for instance in the interplay between convex polytopes and toric varieties. This chapter offers a pinch of combinatorics in a vast sea of algebra. We present topics that are important for nonlinear algebra. The first such topic is matroid theory. Matroids encode independence, just like groups encode symmetry. The theory of matroids has many connections to toric geometry and we will present a few of them. One of our main aims is to highlight connections between Grassmannians (Chapter 5), toric varieties (Chapter 8) and matroids. In all topics the lattice polytopes will play a prominent role. We will finish by presenting a snapshot of generating functions. Their role as Hilbert series brings us back to the two key invariants of a variety: dimension and degree. We would like to stress the fact that our emphasis is not on combinatorics itself, but rather on its interactions with algebra and geometry.

13.1. Matroids

In this section we give an introduction to the theory of matroids. The name matroids suggests that these should be regarded as generalizations of matrices. Indeed, as we will soon see every matrix defines a matroid. We fix a finite set $E$, which we will refer to as the ground set of a matroid. We would like to distinguish a family of subsets of $E$ that we could call independent. Thus a matroid $M$ will be a family $\mathcal{I} \subset 2^E$ of subsets $E$ that we refer to as independent sets. These are of course assumed to satisfy certain axioms.

A first observation is that whenever we have an independent set $I \subset E$, it is reasonable to assume that every subset of $I$ is also independent. We obtain the first axiom of a matroid for the family $\mathcal{I}$:

1. If $I \in \mathcal{I}$ and $J \subset I$, then $J \in \mathcal{I}$.

What we defined so far is a very important object in mathematics: simplicial complex. Another observation is that we would like $\mathcal{I}$ to be nonempty, or equivalently we want the empty set to be independent:

2. We have $\emptyset \in \mathcal{I}$.

It turns out that to obtain a matroid we need just one more axiom. To motivate it we make a following observation. Whenever we have finite linearly independent subsets $I, J \subset V$ of a vector space $V$, if $|I| < |J|$, then we may extend $I$ by an
element of \( j \in J \), in such a way that \( I \cup \{j\} \) is still linearly independent. This simple observation is precisely what we need to get the last axiom for the family \( \mathcal{J} \):

3. If \( I, J \in \mathcal{J} \) and \(|I| < |J|\) then there exists \( j \in J \) such that \( I \cup \{j\} \in \mathcal{J} \).

**Definition 13.1.** A matroid is a family of subsets \( \mathcal{J} \) satisfying Axioms 1, 2, 3.

Exercise 1 asks to prove that the following structures are matroids:

**Example 13.2.**

- (Representable/Realizable matroid) Let \( V \) be a vector space over an arbitrary field \( F \). Let \( E \subset V \) be a nonempty, finite subset. We define \( \mathcal{J} \) to be the family of subsets of \( E \) that are linearly independent. We say that the matroid is representable over \( F \).
- (Graphic matroid) Let \( G \) be a graph with edge set \( E \). Let \( \mathcal{J} \) be the family of those subsets of \( E \) that do not contain a cycle. Equivalently \( \mathcal{J} \) is the family of forests in \( G \).
- (Algebraic matroid) Let \( F \subset K \) be an arbitrary field extension. Let \( E \) be a finite subset of \( K \). Let \( \mathcal{J} \) be the family of subsets of \( E \) that are algebraically independent over \( F \).
- (Uniform matroid) Let \( E \) be a finite set and \( k \leq |E| \). Let \( \mathcal{J} \) be the family of subsets of cardinality at most \( k \). This matroid is denoted by \( U_{k,E} \) or \( U_{k,|E|} \).

Matroids are known for having many equivalent definitions, depending on the point of view on the matroid. For example, due to the first axiom to determine a matroid we do not have to know all independent sets, just those that are inclusion maximal. By analogy to linear algebra, the inclusion maximal independent sets are called basis. It turns out - as the reader is asked to prove in Exercise 2 - that a nonempty family \( \mathcal{B} \subset 2^E \) of subsets of \( E \) is a family of basis of some matroid if and only if the following axiom is satisfied:

- For all \( B_1, B_2 \in \mathcal{B} \), \( b_2 \in B_2 \setminus B_1 \) there exists \( b_1 \in B_1 \setminus B_2 \) such that \( (B_1 \setminus \{b_1\}) \cup \{b_2\} \in \mathcal{B} \).

The seemingly weak axiom on \( \mathcal{B} \) in fact implies the following two statements:

- For all \( B_1, B_2 \in \mathcal{B} \), \( b_2 \in B_2 \setminus B_1 \) there exists \( b_1 \in B_1 \setminus B_2 \) such that both \( (B_1 \setminus \{b_1\}) \cup \{b_2\}, (B_2 \setminus \{b_2\}) \cup \{b_1\} \in \mathcal{B} \).
- For all \( B_1, B_2 \in \mathcal{B} \) and any subset \( A_2 \subset B_2 \setminus B_1 \) there exists a subset \( A_1 \subset B_1 \setminus B_2 \) such that both \((B_1 \setminus A_1) \cup A_2, (B_2 \setminus A_2) \cup A_1 \in \mathcal{B} \).

The first point is known as the symmetric exchange property and the second one as multiple symmetric exchange property. The facts that both exchange properties hold is nontrivial - we refer the reader to the proofs in \([6, 43]\). We will soon see the algebraic meaning of the exchange properties.

Exercise 3 states that all basis of a matroid have the same cardinality. The cardinality of a basis is known as the rank of a matroid. More generally for a matorid on a ground set \( E \) we may define the rank of any subset of \( A \subset E \).

**Definition 13.3.** For a matroid on a ground set \( E \) and independent sets \( \mathcal{J} \subset 2^E \) we define the rank function:

\[
    r : 2^E \ni A \rightarrow \max_{I \in \mathcal{J}} \{|I \cap A|\} \in \mathbb{Z}.
\]

Equivalently, the rank of a set is the cardinality of a largest independent set contained in it.
We note that for a representable matroid the rank is simply the dimension of the vector subspace spanned by the given vectors. Clearly, for any matroid the rank function \( r \) satisfies the following:

- \( 0 \leq r(A) \) for all \( A \subseteq E \) and \( r(\emptyset) = 0 \).
- \( r(A) \leq r(A \cup \{b\}) \leq r(A) + 1 \) for all \( A \subseteq E, x \in E \).

Further, the rank function has one more property known as submodularity:

- for all \( A, B \subseteq E \) we have \( r(A \cup B) + r(A \cap B) \leq r(A) + r(B) \).

In Exercise 7 the reader is asked to prove that any function \( r : 2^E \to \mathbb{Z} \) satisfying the three axioms above is a rank function of a matroid. The independent sets can be reconstructed as those \( I \subseteq E \) for which \( r(I) = |I| \). This gives us another possible definition of a matroid.

### 13.2. Lattice Polytopes

In this section we discuss the interplay between toric geometry and matroids. It is not possible to even state all of the interesting results; we refer to [13, 19, 14, 29].

We start by recalling the definition of a lattice polytope.

**Definition 13.4 (Lattice polytope).** Let \( \mathbb{R}^n \) be a real vector space. A polytope \( P \) is the convex hull of a finite set of points \( p_1, \ldots, p_k \in \mathbb{R}^n \):

\[
P := \{ x \in \mathbb{R}^n : x = \sum_{i=1}^{k} \lambda_i p_i \text{ for some real } \lambda_1, \ldots, \lambda_k \geq 0, \sum_{i=1}^{k} \lambda_i = 1 \}.
\]

We say that \( P \) is a lattice polytope if we may find \( p_1, \ldots, p_k \in \mathbb{Z}^n \subset \mathbb{R}^n \). For each polytope \( P \) there is an inclusion minimal set of \( p_i \)'s of which it is a convex hull. We call these \( p_i \)'s the vertices of \( P \).

To pass from a combinatorial object, like a matroid, to a polytope, we apply the following 'standard' construction. Consider a vector space \( \mathbb{R}^{|E|} \) with basis elements \( b_e \) corresponding to the elements \( e \in E \). Any subset \( A \subseteq E \) can be identified with a point \( p_A := \sum_{e \in A} b_e \in \mathbb{R}^{|E|} \). In this way a family of subsets may be identified with a set of points.

**Definition 13.5 (Matroid basis polytope).** Let \( M \) be a matroid on the ground set \( E \) and basis set \( \mathcal{B} \). We use the notation introduced above. We define the matroid basis polytope \( P_M \subset \mathbb{R}^{|E|} \) as the convex hull of the points \( p_B := \sum_{e \in B} b_e \in \mathbb{R}^{|E|} \), where we take all \( B \in \mathcal{B} \).

Clearly \( P_M \) is a lattice polytope, hence we may consider the toric variety associated to it. Precisely it is the image of the map given by monomials, in variables corresponding to elements of \( E \), that are products of elements in a basis.

**Example 13.6.** Consider the rank two uniform matroid on the set \( E = \{1, 2, 3\} \). Its set of bases is

\[
\mathcal{B} = \{ \{1, 2\}, \{1, 3\}, \{2, 3\} \}.
\]

We consider \( \mathbb{R}^3 \). The three bases correspond, in the given order, to the three points

\[
(1, 1, 0), (1, 0, 1), (0, 1, 1) \in \mathbb{R}^3.
\]

Hence, the matroid basis polytope is a triangle. The polynomial map is:

\[
(\mathbb{C}^*)^3 \ni (x_1, x_2, x_3) \mapsto (x_1x_2, x_1x_3, x_2x_3) \in \mathbb{P}^2.
\]

The closure of the image is the whole \( \mathbb{P}^2 \), which is the associated toric variety.
The combinatorial statement equivalent to the proposition presented below was proved by White [42].

**Proposition 13.7.** A matroid basis polytope is normal in the lattice it spans.

In order to present the proof we state one of the most useful theorems about matroids.

**Theorem 13.8 (The matroid union theorem).** Let $M_1, \ldots, M_k$ be matroids on the same ground set $E$ with respective families of independent sets $I_1, \ldots, I_k$ and rank functions $r_1, \ldots, r_k$. Let

$$\mathcal{I} := \{I \subseteq E : I = \bigcup_{i=1}^k I_i \text{ for } I_i \in I_i\}.$$  

Then $\mathcal{I}$ is also a family of independent sets for a matroid, known as the union of $M_1, \ldots, M_k$. Further, the rank of any set $A \subseteq E$ for the union matroid is given by:

$$r(A) = \min_{B \subseteq A} \{|A \setminus B| + \sum_{i=1}^k r_i(B)\}.$$  

For the proof we refer to [29, 12.3.1]. As a corollary of the matroid union theorem we obtain the following theorem due to Edmonds.

**Theorem 13.9.** Let $M$ be a matroid on a ground set $E$ with rank function $r$. $E$ can be partitioned into $k$ independent sets if and only if $|A| \leq k \cdot r(A)$ for all subsets $A \subseteq E$.

**Proof.** The implication $\Rightarrow$ is straightforward.

For the other implication consider the union $U$ of $M$ with itself $k$ times. We apply the matroid union theorem to compute the rank of $E$:

$$r_U(E) = \min\{|E| - |B| + k \cdot r_M(B)|.$$

Clearly by assumption $|E| - |B| + k \cdot r_M(B) \geq |E|$ and equality holds for $B = \emptyset$. Hence, $r_U(E) = |E|$. This means that $E$ is an independent set in $U$, and hence by definition it is a union of $k$ independent sets of $M$. □

**Definition 13.10.** Let $M$ be a matroid on a ground set $E$ with the family of independent sets $\mathcal{I}$. Let $E' \subseteq E$. The *restriction* of $M$ to $E'$ is a matroid where $A \subseteq E'$ is independent if and only if $A \in \mathcal{I}$.

**Proof of Proposition 13.7.** Let $M$ be a matroid on the ground set $\{1, \ldots, n\}$. Let $p \in kP_M$. We know that $p = \sum_{B \in \mathbb{B}} \lambda_B p_B$ with $\sum \lambda_B = k$ and $0 \leq \lambda_B \in \mathbb{Q}$. After clearing the denominators we have:

$$dp = \sum_{B} \lambda'_B p_B,$$

where $\sum \lambda_B = dk$ and $0 \leq \lambda_B \in \mathbb{Z}$.

By restricting the matroid $M$ we may assume that all coordinates of $p = (p_1, \ldots, p_n)$ are nonzero.

We define two matroids. The first matroid $N$ is on the ground set $E_N := \{(i, j) : 1 \leq i \leq n, 1 \leq j \leq p_i\}$. In other words, we replace a point $i$ in the original matroid by $p_i$ equivalent points. A subset $\{(i_1, j_1), \ldots, (i_s, j_s)\} \subseteq E_N$ is independent if only if:

- all $i_s$’s are distinct,
\[ \{i_1, \ldots, i_s\} \text{ is an independent set in } M. \]

We note that a basis of \(N\) maps naturally to a basis of \(M\). Also the rank function for \(N\) is the same as the one for \(M\) if we forget the second coordinates. The point \(p\) has a decomposition as a sum of \(k\) points corresponding to basis of \(M\) if and only if the matroid \(N\) is covered by \(k\) basis (i.e. the ground set is a union of \(k\) basis). Hence, by Theorem 13.9 our aim is to prove the following statement:

For any \(A \subset E_N\) we have \(|A| \leq kr_N(A)\).

The second matroid \(N'\) is on the ground set \(E_{N'} := \{(i, j, l) : 1 \leq i \leq n, 1 \leq j \leq p_i, 1 \leq l \leq d\}\). In other words we replace any point of \(N\) by \(d\) equivalent points. A subset \(\{(i_1, j_1, l_1), \ldots, (i_s, j_s, l_s)\} \subset E_N\) is independent if only if:

- all \(i_s\)’s are distinct,
- \(\{i_1, \ldots, i_s\}\) is an independent set in \(M\).

We have a natural projection \(\pi : E_{N'} \to E_N\) given by forgetting the last coordinate. We note that \(r_{N'}(\pi^{-1}(A)) = r_N(A)\). As the point \(dp\) is decomposable we know that the matroid \(N'\) can be covered by \(kd\) basis. Hence, for any \(B \subset E_{N'}\) we have: \(|B| \leq dk \cdot r_{N'}(B)\). Applying this to \(\pi^{-1}(A)\) we obtain:

\[ k|A| = |\pi^{-1}(A)| \leq dk \cdot r_{N'}(\pi^{-1}(A)) = dk \cdot r_N(A). \]

This is equivalent to the statement we wanted to prove! \(\square\)

Our next aim is to relate matroids with the geometry of special subvarieties of Grassmannians. We recall that one of the possible definitions of a Grassmannian \(G(k, n)\) is an orbit of \([e_1 \wedge \cdots \wedge e_k] \in \mathbb{P}(\bigwedge^k \mathbb{C}^n)\) under the action of the group of \(n \times n\) invertible matrices \(GL(n)\). While the Grassmannian is an orbit of the big group \(GL(n)\), we may ask how smaller groups act on \(G(k, n)\). In particular, consider the torus \(T := (\mathbb{C}^\times)^n\) of diagonal matrices. This torus acts on \(\mathbb{P}(\bigwedge^k \mathbb{C}^n)\) and on \(G(k, n)\). However, in general \(G(k, n)\) is not an orbit of \(T\) or even a closure of an orbit of \(T\). Indeed, we already know that \(G(k, n)\) has dimension \(k(n - k)\) which may be much larger than \(n\). Let us fix a point \(p \in G(k, n)\). The questions that motivate us are:

- What is the \(T\)-orbit of \(p\)?
- What is the closure of this orbit?
- How can we describe this variety?

A beautiful answer was provided by Gelfand, Goresky, MacPherson and Serganova 16. The point \(p = [v_1 \wedge \cdots \wedge v_k] \in G(k, n)\) represents a \(k\)-dimensional subspace \(V = \langle v_1, \ldots, v_k \rangle\) in \(\mathbb{C}^n\). We may present the vectors \(v_1, \ldots, v_k\) as a \(k \times n\) matrix \(N_p\). From Chapter 4 we know that the coordinates of \(p \in \mathbb{P}(\bigwedge^k \mathbb{C}^n)\), are given by maximal minors of \(N_p\). How does a point \(t = (t_1, \ldots, t_n) \in T\) act on \(p\)? In general, \(t\) acts on the coordinate indexed by \(e_{i_1} \wedge \cdots \wedge e_{i_k}\) rescaling it by \(t_{i_1} \cdots t_{i_k}\). Hence, the orbit of \(p\) is the image of the map:

\[ T \ni (t_1, \ldots, t_n) \mapsto (t_{i_1} \cdots t_{i_k} \det((N_p)_{i_1, \ldots, i_k}))_{1 \leq i_1 < \cdots < i_k \leq n} \in \mathbb{P}(\bigwedge^k \mathbb{C}^n), \]

where \((N_p)_{i_1, \ldots, i_k}\) denotes the \(k \times k\) submatrix of \(N_p\) with the chosen columns indexed by \(i_1, \ldots, i_k\).

**Example 13.11.** Consider the two dimensional subspace of the four dimensional space spanned by the rows of the following matrix:

\[
\begin{bmatrix}
1 & 1 & 1 & 1 \\
1 & 2 & 3 & 4
\end{bmatrix}.
\]
In the coordinates of the Grassmannian we have the associated point:
\[(e_1+e_2+e_3+e_4)\wedge(e_1+2e_2+3e_3+4e_4) = e_1\wedge e_2 + 2e_1\wedge e_3 + 3e_1\wedge e_4 + e_2\wedge e_3 + 2e_2\wedge e_4 + e_3\wedge e_4.\]
The orbit in the coordinates above is parameterized as follows:
\[(t_1, t_2, t_3, t_4) \rightarrow (t_1t_2, 2t_1t_3, 3t_1t_4, 2t_2t_4, t_3t_4).\]

This is almost a monomial map! Indeed, the only thing that changes are the constants given by minors of the matrix $N_p$. However, these constants do not depend on $t \in T$ and hence we may define an automorphism of $\mathbb{P}(\wedge^k \mathbb{C}^n)$ that turns the orbit to an image of a monomial map, by simply rescaling the coordinates.

At this point one could have a false impression that the orbit is isomorphic to the image of a monomial map defined by all squarefree monomials of degree $k$. This is not the case, as some of the monomials may not appear at all. This happens if the corresponding minor was equal to zero - then we cannot rescale it.

**Example 13.12.** (1) First we continue Example 13.11. The polytope associated to the toric variety has the following vertices:
\[(1, 1, 0, 0), (1, 0, 1, 0), (1, 0, 0, 1), (0, 1, 1, 0), (0, 1, 0, 1), (0, 0, 1, 1).\]
This is the hypersimplex $\Delta_{2,4}$. The associated projective toric variety is three dimensional. It represents the closure of the $T$-orbit of a general point in the Grassmannian $G(2, 4)$. The associated matroid is the uniform rank two matroid on four elements.

(2) Let us now consider a different point of $G(2, 4)$ given by the rows of the following matrix:
\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
1 & 2 & 3 & 4
\end{bmatrix}.
\]
The orbit is parameterized as follows:
\[(t_1, t_2, t_3, t_4) \rightarrow (2t_1t_2, 3t_1t_3, 4t_1t_4, 0, 0, 0).\]
The polytope representing the toric variety has the following vertices:
\[(1, 1, 0, 0), (1, 0, 1, 0), (1, 0, 0, 1).\]
It is isomorphic to a two dimensional simplex, hence the closure of the orbit is a $\mathbb{P}^2$, as can be seen directly from the parameterization.

Which monomials are thus left? Exactly those for which the corresponding minor of $N_p$ was not zero.

Let us consider the representable matroid $M_p$ of $n$ points in $\mathbb{C}^k$, defined by the columns of the matrix $N_p$. Clearly a set of points is a basis of $M_p$ if and only if the corresponding minor of $N_p$ is nonzero. We have proved the following proposition.

**Proposition 13.13.** The closure of the $T$-orbit of any point $p = [v_1 \wedge \cdots \wedge v_k]$ in a Grassmannian $G(k, n)$ is the toric variety represented by the matroid base polytope, for the representable matroid defined by columns of the $k \times n$ matrix $N_p$ with $i$-th row equal to $v_i$.

The results of Chapter 8 combined with Proposition 13.7 show the following.

**Proposition 13.14.** Any torus orbit closure in any Grassmannian is projectively normal.
We now turn to the interpretation of basis exchange properties in terms of algebraic geometry. Consider a matroid with basis polytope $P$. We recall that:

- the ideal of the associated toric variety is generated by binomials,
- every binomial in the ideal corresponds to an integral relation among lattice points of $P$.

How do these statements specialize in the case of matroids? A lattice point of $P$ is the characteristic function of a basis. A sum of lattice points is the sum of these characteristic functions. This corresponds to taking a sum of basis as multisets.

**Example 13.15.** Consider the rank two uniform matroid on four elements $\{p_1, p_2, p_3, p_4\}$. An integral relation among the vertices of the matroid polytope is:

$$(1, 1, 0, 0) + (0, 0, 1, 1) = (1, 0, 1, 0) + (0, 1, 0, 1).$$

As a sum of basis elements this corresponds to:

$$\{p_1, p_2\} \cup \{p_3, p_4\} = \{p_1, p_3\} \cup \{p_2, p_4\}.$$

It is a degree two binomial in the ideal of the associated toric variety.

Hence, we say that two multisets of basis are *compatible* if their union (as multisets) is the same. Equivalently, every element of the base set belongs to the same number of basis in the first and second multiset of basis. Thus, the binomials in the ideal of the toric variety represented by matroid base polytope are in bijection with pairs of compatible multisets of basis.

What are the quadrics in such an ideal? Equivalently, when $\{B_1, B_2\}$ is equivalent to $\{B_3, B_4\}$? This is if and only if $B_1 \cup B_2 = B_3 \cup B_4$. In other words, this is if and only if we change:

- $B_1$ by subtracting from it a set $A_1 \subset B_1 \setminus B_2$ and adding to it $A_2 \subset B_2 \setminus B_1$ and
- $B_2$ by adding to it $A_1$ and subtracting $A_2$.

We see that quadrics in the ideal correspond to multiple symmetric exchanges. It follows that symmetric basis exchanges form a distinguished set of quadrics in the ideal. The following four conjectures are due to White.

**Conjecture 13.16.**

- **Representable case:** The ideal of any torus orbit closure in any Grassmannian is:
  
  (1) generated by quadrics,
  (2) generated by quadrics corresponding to symmetric basis exchanges.

- **General case:** For any matroid $M$ any two finite multisets of basis $(B_i), (B_j)$ such that $\bigcup B_i = \bigcup B_j$ can be transformed to one another in a finite number of such steps that:
  
  (1) we replace two basis $B, B'$ in one multiset, by two basis $\tilde{B}, \tilde{B}'$ obtained by multiple symmetric exchange (i.e. $B \cup B' = \tilde{B} \cup \tilde{B}'$),
  (2) we replace two basis $B, B'$ in one multiset, by two basis $\tilde{B}, \tilde{B}'$ obtained by a symmetric exchange (i.e. $B = \tilde{B} \cup \{b_1\} \setminus \{b_2\}$ and $B' = \tilde{B}' \cup \{b_2\} \setminus \{b_1\}$).

One can show that the general case implies the representable case.
13.3. Generating Functions

In this section we introduce multivariate generating functions that are given by rational functions. The key example is multigraded Hilbert series. We discuss methods for computing them, and we explore connections to regular triangulations. In particular, we discuss the Ehrhart series.

We start with the familiar example of the polynomial ring $K[x]$. From Chapter 8 we know we may interpret it as a semigroup algebra $K[Z_{\geq 0}]$. How to interpret the Hilbert function $h(q)$ introduced in Chapter 1? As the monomials of degree $d$ form a basis of $K[Z_{\geq 0}]$, the $h(d)$ equals the number of lattice points in the semigroup that belong to the hyperplane $H_d$ defined by $\sum x_i = d$. Let $\Delta$ be the standard simplex, i.e. the convex hull of basis vectors. We note that the Hilbert function counts the number of lattice points in dilations of $\Delta$, precisely $h(d) = |d\Delta \cap Z^d|$.

Further, for the Hilbert series we obtain

$$HS(z) = \sum_{q=0}^{\infty} |d\Delta \cap Z^d| z^q = \frac{1}{(1-z)^d},$$

as in Example 1.22. Our next aim is to refine our counting. So far we have treated all monomials of the same degree on an equal footing. What happens if we try to remember each monomial, not only its degree?

**Definition 13.17.** Let $C \subset \mathbb{R}^n$ be a rational pointed polyhedral cone. We define the associated **multigraded Hilbert series** as a formal power series:

$$MHS_C(x) = \sum_{c \in C \cap \mathbb{Z}^n} x^c.$$

If $C \subset \mathbb{R}_{\geq 0}^n$ then we may reconstruct the Hilbert series of $\mathbb{C}[C]$ from the multigraded Hilbert series, by setting $x_1 = \cdots = x_n = z$. However, the multigraded Hilbert series remembers much more information: all lattice points of the cone. Our next aim is to represent MHS as a rational function, just as we did with the Hilbert series. A cone generated by linearly independent vectors is called **simplicial**.

**Lemma 13.18.** Let $C$ be a simplicial cone with ray generators $c_1, \ldots, c_d \in \mathbb{Z}^n$. Then

$$MHS_C(x) = \frac{\kappa_C(x)}{\prod_{i=1}^d (1-x^{c_i})},$$

where $\kappa_C(x)$ is a Laurent polynomial with non-negative coefficients.

**Proof.** Consider the following half-open parallelepiped:

$$P := \{ x \in C : x = \sum_{i=1}^d \lambda_i c_i, 0 \leq \lambda_1, \ldots, \lambda_d < 1 \}.$$

As $c_i$ are linearly independent and generate $C$ as a cone, every lattice point $c \in C$ has a unique representation $c = p + \sum_{i=1}^d s_i c_i$, where $p \in P$ is a lattice point and $s_i$ are non-negative integers. We obtain:

$$MHS_C(x) = \sum_{c \in C \cap \mathbb{Z}^n} x^c = \sum_{p \in P \cap \mathbb{Z}^n} x^p \left( \prod_{i=1}^d \left( \sum_{s_i=0}^{\infty} x^{s_i c_i} \right) \right) =$$
\[
\sum_{p \in P \cap \mathbb{Z}^n} x^p \left( \prod_{i=1}^d \frac{1}{1 - x^{c_i}} \right) = \frac{\sum_{p \in P' \cap \mathbb{Z}^n} x^p}{\prod_{i=1}^d (1 - x^{c_i})}.
\]

Remark 13.19. We could freely manipulate with infinite summations in Lemma 13.18 as the assumption that \(C\) is pointed assures that the series is absolutely convergent in some neighborhood.

Proposition 13.20. Let \(C\) be a pointed, rational polyhedral cone with ray generators \(c_1, \ldots, c_d \in \mathbb{Z}^n\). Then
\[
\text{MHS}_C(x) = \frac{\kappa_C(x)}{\prod_{i=1}^d (1 - x^{c_i})},
\]
where \(\kappa_C(x)\) is a Laurent polynomial with integral coefficients.

Proof. We may triangulate the cone \(C\), i.e. present it as a union of simplicial cones which rays are rays of \(C\) and which intersect only in lower dimensional simplicial cones. This may be done e.g. by induction on the number of rays.

Lemma 13.18 together with induction on dimension of \(C\) and inclusion-exclusion allow us to conclude. \(\square\)

Remark 13.21. We note that the proofs of Proposition 13.20 and Lemma 13.18 give us an algorithm to compute the multigraded Hilbert series, as well as combinatorial interpretation of the numerator.

The case when \(C\) is a cone over a lattice polytope is particularly nice. Let \(P\) be a lattice polytope in \(\mathbb{R}^n\). We may regard it as a polytope \(P \times \{1\} \subset \mathbb{R}^n \times \mathbb{R}\). Let \(C \subset \mathbb{R}^{n+1}\) be the cone over \(P\), i.e. the smallest cone that contains \(P \times \{1\}\).

Proposition 13.22. Let \(C\) and \(P\) be as defined above. The Hilbert function for \(C\) with respect to the grading induced by the last variable is given by:
\[
h(q) = |qP \cap \mathbb{Z}^{n+1}|.
\]
It counts the number of lattice points in dilations of \(P\). It is known as the Ehrhart polynomial of \(P\) and indeed it coincides with a polynomial for \(q \in \mathbb{Z}_{\geq 0}\).

Proof. The only nontrivial statement is that \(h(q)\) is equal to a polynomial for all positive integers \(q\). By induction on \(\dim P\) and by triangulating \(P\) it is enough to prove the statement when \(P\) is a simplex with vertices \(v_1, \ldots, v_d\). Let \(f(q) := \frac{d+q-1}{d} \binom{d+q-1}{d-1}\) for \(q \geq 0\) and \(f(q) = 0\) for \(q < 0\). As in the proof of Lemma 13.18 we have:
\[
h(q) = \sum_{i=0}^{d-1} a_i f(q - i),
\]
where \(a_i\) is the number of lattice points with the last coordinate \(i\) in the set
\[
\{ x : x = \sum_{i=1}^d \lambda_i(v_i, 1), 0 \leq \lambda_1, \ldots, \lambda_d < 1 \}.
\]
We only have to consider \(a_i\) for \(i < d\), as there are no lattice points in this parallelepiped with last coordinate greater or equal to \(d\).
We note that $f$ is not a polynomial if we consider the negative arguments. The punchline is that the polynomial $q(q) := (d + q - 1)(d + q - 2) \cdots q/(d - 1)!$ equals $f$ also for negative, integral $q$, as long as $q \geq -d + 1$. Hence, for $q \in \mathbb{Z}_{\geq 0}$ we have:

$$h(q) = \sum_{i=0}^{d-1} a_ig(q - i).$$

Given a lattice polytope $P$ with $N$ lattice points, we may associate to it a projective toric variety in $\mathbb{P}^{N-1}$ as in Chapter 8. Hence, we obtain a binomial ideal $I_P \subset K[x] = K[x_1, \ldots, x_N]$. Let us fix a term order $\prec$ which induces the initial ideal in$_\prec(I_P)$. The latter is a monomial ideal. Thus the radical rad in$_\prec(I_P)$ has the following property:

- if a product $m$ of distinct variables does not belong to rad in$_\prec(I_P)$, then neither does any monomial that divides $m$.

This may be restated as:

- subsets $S$ of variables such that $\prod_{x \in S} x \notin$ rad in$_\prec(I_P)$ form a simplicial complex.

Our aim is to obtain a nice, geometric description of this simplicial complex. The main idea is that the variables in the ring are in bijection with lattice points of $P$. Let $\Delta$ be a subdivision of $P$ into polytopes $P_i$ that are convex hulls of sets of points $S$, such that the product of variables corresponding to $S$ is not in rad in$_\prec(I_P)$.

**Example 13.23.** Let $P$ be the square conv$((0, 0, 1), (0, 1, 1), (1, 0, 1), (1, 1, 1))$. The associated projective variety is a surface in $\mathbb{P}^3$ defined by $x_0x_3 - x_1x_2$. We fix a term order for which $x_0x_3$ is the leading term. The triangulation $\Delta$ of $P$ contains two triangles: conv$((0, 0, 1), (1, 1, 1), (1, 0, 1))$ and conv$((0, 1, 1), (1, 0, 1), (1, 1, 1))$. The minimal nonface is the pair of vertices $(0, 0, 1), (1, 1, 1)$ corresponding to the unique generator of the initial ideal.

If we change the term order so that $x_1x_2$ becomes the leading term we obtain a different triangulation of $P$, given by the other diagonal.

The following proposition relates Gröbner bases to triangulations.

**Proposition 13.24.** Using the notation introduced above $\Delta$ is a triangulation of $P$. The minimal non-faces of $\Delta$ correspond to (radicals of) generators of $\text{in}_\prec(I_P)$.

**Definition 13.25.** The triangulations of the form $\Delta$, induced by any term order, are called regular. There exist triangulations that are not regular.

The story that we are telling has in fact three sides: combinatorial, algebraic and geometric. From the combinatorial point of view, as described above, we are triangulating a lattice polytope $P$ into simplices. The algebraic part is the finest: we degenerate a toric, i.e. binomial, prime ideal $I_P$ to a monomial ideal that shares with $I_P$ all the most important invariants, like dimension and degree.

Let us now describe the geometry in this picture. Here we degenerate the variety $V(I_P)$ to $V(\text{in}_\prec(I_P))$. One of the problems is that $\text{in}_\prec(I_P)$ maybe not radical, thus we may lose some information, however let us ignore this for a moment.

What is $V(\text{in}_\prec(I_P))$? As $V(\text{in}_\prec(I_P)) = V(\text{rad}(\text{in}_\prec(I_P)))$ the question is what is the set of solutions of a squarefree monomial ideal. Let us state the solution to Exercise 12 in Chapter 2.
13.3. GENERATING FUNCTIONS

- the variety \( V(\text{rad}(\text{in}_{\prec}(I_P))) \) is the union of (coordinate) vector subspaces. Each subspace is spanned by basis vectors \((e_i)_{i \in S} \) such that \( \prod_{i \in S} x_i \notin \text{rad}(\text{in}_{\prec}(I_P)) \).

In particular: the simplices in the induced triangulation of \( P \) correspond naturally to components of \( V(\text{in}_{\prec}(I_P)) \) - as the triangulation breaks the polytope into simple pieces, our variety breaks into simple components.

We note that the idea of computing the dimension and degree of an ideal, by passing to the initial ideal is equivalent to the idea of computing the Hilbert series of a cone, by subdividing it into simplicial cones.

We know that the dimension of the (projective) toric variety associated to a polytope \( P \) equals the dimension of \( P \). What about the degree?

**Proposition 13.26.** Let \( P \) be a \( d \)-dimensional lattice polytope, whose lattice points generate the lattice \( \mathbb{Z}^d \). The degree of the ideal \( I_P \) equals the Euclidean volume of \( P \) times \( d! \).

**Sketch of the proof.**

1. The degree times the factorial of the dimension is the leading coefficient of the Ehrhart polynomial \( h \). Thus it is enough to show that for any \( \epsilon > 0 \) there exists a constant \( C \) such that:

\[
\frac{\text{vol } P - \epsilon}{d!} q^d - C \leq h_P(q) \leq \frac{\text{vol } P + \epsilon}{d!} q^d + C,
\]

for any positive integer \( q \), where \( h_P(q) \) is the number of lattice points in \( qP \).

2. It is easy to prove inequality (13.1) for rational polytopes that are products of intervals.

3. Point (2) implies point (1), by covering \( P \) with small products of intervals, according to the definition of the Lebesgue measure.

**Example 13.27.** Let \( P \) be a \( d \)-dimensional simplex, given as the convex hull of 0 and \( d \) basis vectors. The Ehrhart polynomial is given by \( h(q) = \binom{d+q}{q} = \frac{1}{d!} q^d + \text{lower order terms} \). Indeed \( \text{vol } P = \frac{1}{d!} \) and \( \dim P = d \).

The usual Euclidean volume multiplied by \( d! \) is known as the normalized volume. The simplex that is the convex hull of 0 and all standard basis vectors has normalized volume equal to one. Every lattice polytope has normalized volume that is a positive integer, equal to the degree of the associated variety, if one works in the correct lattice.

How is the triangulation compatible with the degree computation? Clearly the volume of the polytope is equal to the sum of volumes of (maximal) simplices in its triangulation.

**Theorem 13.28.** Let \( P \) be a \( d \)-dimensional lattice polytope whose lattice points generate the lattice \( \mathbb{Z}^d \). Let \( I_P \) be the associated toric ideal. Let \( \prec \) be a term order and \( \Delta \) the associated triangulation of \( P \).

1. The minimal primes of \( \text{in}_{\prec}(I_P) \) are in bijection with the maximal simplices in the triangulation \( \Delta \).

2. The (unique) primary ideal corresponding to a minimal prime of \( \text{in}_{\prec}(I_P) \) has degree equal to the normalized volume of the associated simplex in \( \Delta \).
Exercises

1. Show that Example 13.3 presents matroids.

2. (a) Fix a family of independent sets $\mathcal{I}$ for a matroid $M$. Prove that the inclusion maximal elements in $\mathcal{I}$ satisfy the axiom for the basis of a matroid.
   
   (b) Fix a nonempty set $\mathcal{B} \subset 2^E$ satisfying the axiom for basis of a matroid.
   
   Prove that $\mathcal{I} := \{I \subset E : \exists B \in \mathcal{B} : I \subset B\}$ satisfies the axioms for the independent sets.

3. Prove that all basis in a matroid have the same cardinality.

4. Prove that the points $p_B$ in Definition 13.5 are vertices of the polytope $P_M$.
   
   Prove that these are the only lattice points of $P_M$.

5. In this exercise we examine matroid duality.
   
   (a) Let $\mathcal{B} \subset 2^E$ be a set of basis of a matroid $M$. Let $\mathcal{B}^* := \{B \subset E : E \setminus B \in \mathcal{B}\}$. Prove that $\mathcal{B}^*$ is a set of basis of a matroid $M^*$. The matroid $M^*$ is known as the dual matroid (of $M$).

   (b) Prove that a dual of a representable matroid is representable.

6. Prove that for any matroid the rank function is submodular.

7. Prove that any function $2^E \to \mathbb{Z}$ satisfying the three axioms of the rank function is indeed a rank function of some matroid.

8. How many distinct torus orbit closures are there in $G(2, 4)$? How many up to isomorphism (of algebraic varieties)?

Bibliography

Index

affine cone, 20
associated prime, 36
Chevalley’s theorem, 22
constructible set, 18
cusp, 27
cuspidal curve, 27
degree
of a polynomial, 2
of a variety, 19
of an ideal, 11
Descartes’ Rule of Signs, 32
Dickson’s Lemma, 5
dimension, 11
of a variety, 19
Ehrhart polynomial, 143
eigenspace, 75
eigenvalue, 75
embedded prime, 37
Gröbner basis, 6
reduced, 7
Hilbert function, 11
affine, 11
for monomial ideal, 9
Hilbert polynomial, 10
Hilbert series, 11
for monomial ideal, 9
Hilbert’s Basis Theorem, 7
homogeneous ideal, 20
homogeneous polynomial, 20
icosahedron, 23
ideal, 2
$ p $-primary, 35
homogeneous, 20
irreducible, 34
maximal, 4
monomial, 6
primary, 4, 34
prime, 4
radical, 4
zero-dimensional, 31
ideal quotient, 3
initial monomial, 6
integral domain, 3
isolated prime, 37
Jacobian matrix, 19
Kalman Varieties, 22
matroid, 135
representable, 136
minimal prime, 37
monomial
standard, 7
nilpotent, 3
normalized volume, 145
order
degree lexicographic, 6
degree reverse lexicographic, 6
lexicographic, 6
monomial, 6
oval, 23
PDE, 38
PID, 3
primary decomposition, 35
minimal, 36
projective closure, 21
projective space, 20
Puiseux series, 32
rational normal curve, 21
ring
Noetherian, 34
semi-algebraic set, 18
singular
point, 19
singular locus, 20
of a hypersurface, 19
smooth
point, 20
spectrum
  of a ring, 17
standard monomial, 7

triangulation
  regular, 144

variety, 15
  irreducible, 15, 16
  projective, 20

Zariski topology, 16
zero divisor, 3