KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Solving Systems of Polynomial Equations

Simon Telen

Supervisor:
Prof. dr. ir. M. Van Barel

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

September 2020

Solving Systems of Polynomial Equations

Simon TELEN

Examination committee:
Prof. dr. ir. Y. Willems, chair
Prof. dr. ir. M. Van Barel, supervisor
Prof. dr. N. Vannieuwenhoven
Prof. dr. W. Veys
Prof. dr. N. Budur
Prof. dr. ir. L. De Lathauwer
Prof. dr. D. A. Cox
(Amherst College)
Prof. dr. B. Mourrain
(INRIA)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Computer Science

September 2020

Uitgegeven in eigen beheer, Simon Telen, Celestijnenlaan 200A box 2402, B-3001 Leuven (Belgium)

Preface

This text is the result of four years of research which I conducted as a PhD student at the department of computer science of KU Leuven under the supervision of Prof. dr. ir. Marc Van Barel. It describes the insights I gained during this journey, which would have never been possible without the help and support of many other people.

I would like to thank all members of the examination committee for agreeing to be on my jury, for reading this thesis and for providing me with helpful feedback.

I consider myself very lucky to be under the supervision of a mentor who, on top of being an excellent researcher, is open for ideas and suggestions of his mentees and cares a lot about their general wellbeing. Marc, I cannot thank you enough for the opportunities you have given me to travel and meet other mathematicians, to pursue my own research interests even though they were sometimes outside both of our comfort zones, and for giving me the feeling that I can always contact you for advice or just a conversation. I could not have asked for better guidance.

During my PhD I have had the opportunity to visit several other research institutes. I am grateful to Bernard Mourrain for having me as a visitor at INRIA in Sophia Antipolis in September 2017 and for teaching me most of what I know about resultants and normal forms. I want to thank Tyler Jarvis and his students for inviting me to come to BYU in Provo in April 2018 and for our interesting discussions together with Alex Townsend on real root finding. I owe a big thank you to Jan Verschelde for two great weeks at UIC in Chicaco in August 2018, during which I learned a lot about homotopy continuation. My visit at MPI Leipzig in December 2018 was filled with fascinating mathematical discussions and encounters with great colleagues. I want to thank Bernd Sturmfels for the invitation and I am excited to start as a postdoc at the institute. I am grateful to Tomas Pajdla for having me over as a visitor at CIIRC in Prague in March 2019 and for introducing me to some fascinating applications of polynomials in computer vision. My longest research stay abroad was made possible by the MATH$^+$ Thematic Einstein Semester on Algebraic Geometry, Varieties, Polyhedra, Computation from September 2019 until January 2020 in Berlin. I am thankful to Peter Bürgisser for giving me the opportunity to participate. To all the people with whom I either shared an office (thanks, Oliver!), a research project (thanks, Alessandro, Elise, Matías, Marta, Roser, Sascha, Thomas, Tim!), a session in our intersection

theory reading seminar or a table at *Dave B's*: thank you for making this such a great experience. We also had the privilege to welcome some visitors to Leuven. I want to thank Milena Wrobel for the great week we spent in and around Leuven in August 2019 and for our many discussions about Cox rings and their applications. I'm also very thankful for Sascha Timme's visit in March 2020. I enjoyed our collaboration and our time in Leuven together with Katy a lot.

I would like to thank my colleagues at the department for our daily discussions. I am grateful to Nick Vannieuwenhoven for answering my questions, for giving me excellent advice on many occasions and for sharing several conference experiences together. Thank you Andrew, Daan, Dries, Luca, Marcus, Niel and Nina for being part of some of my best 'after work' memories. As the research topics of this thesis require some mathematical tools that are outside the expertise of our department, it was a great help for me to have a group of specialists working in the very next building. Thanks Alexander Lemmens, Filip Cools, Marcel Rubió, Naud Potemans, Nero Budur, Wim Veys and Wouter Castryck for answering several of my questions.
During the past four years I have also had the chance to talk to many mathematicians from other institutions whom I admire very much. I want to thank Carlos D'Andrea, Kim Batselier, Alessandra Bernardi, Laurent Busé, Lieven De Lathauwer, Alicia Dickenstein, Philippe Dreesen, Mateusz Michalek, Vanni Noferini, Gregory Smith, Ivan Soprunov, Frank Sottile, and Bernd Sturmfels for being so kind and available for all my questions. I would like to express my special thanks to David Cox for setting a great example as a mathematician, a communicator and a person.
Mathematical conferences are great occasions to learn and to brainstorm, but also to meet new people and to catch up with friends. Throughout the years, this job has taken me from open mic nights in Hong Kong and karaoke bars in Valencia to line dancing in Texas and rooftop bars in New York. I want to thank all of you who have made my conference experiences so memorable.
I also owe a big thank you to my office mates, who have made me look forward to going to work each day. Przemek, thank you for putting up with me and Andreas in our first years. Sahar, thank you for doing the same thing during the last couple of years and for always being happy and cheerful. Andreas, thank you for being my office mate and companion throughout the entire journey (including travels, construction works and pandemics) and for being patient with me and my computer illiteracy.

I want to thank my mom and dad for giving me the opportunity to study, work and travel and for being supportive of whatever I do. I am grateful to my sister Valérie for her interest, love and support, and to Ruben for being a great partner for her. I thank both my grandmothers for always being there for me and for being (embarrassingly) proud of what I do.
Finally, I want to thank my friends for keeping me sane and reminding me that there is life outside of mathematics. Thank you for the skiing trips, the concerts and festivals, the PÄÄL and Detlev weekend trips, the new years eve reunions, the good conversations .... It goes without saying that I am more proud of all of you than of any thesis or paper I will ever write. Special thanks to Brent, Bruno, Deborah, Dries, Pieter, Sebastiaan and Sophie.

Abstract

Systems of polynomial equations arise naturally from many problems in applied mathematics and engineering. Examples of such problems come from robotics, chemical engineering, computer vision, dynamical systems theory, signal processing and geometric modeling, among others. The numerical solution of systems of polynomial equations is considered a challenging problem in computational mathematics. Important classes of existing methods are *algebraic methods*, which solve the problem using eigenvalue computations, and *homotopy methods*, which track solution paths in a continuous deformation of the system. In this text, we propose new algorithms of both these types which address some of the most important (numerical) shortcomings of existing methods.

Classical examples of algebraic techniques use Gröbner bases, border bases or resultants. These methods take advantage of the fact that the solutions are encoded by the structure of an algebra that is naturally defined by the equations of the system. In order to do computations in this algebra, the algorithms choose a *representation* of it which is usually given by a set of monomials satisfying some conditions. In this thesis we show that these conditions are often too restrictive and may lead to *severe numerical instability* of the algorithms. This results in the fact that they are not feasible for finite precision arithmetic. We propose the framework of *truncated normal forms* to remedy this and develop new, robust and stabilized methods. The framework generalizes Gröbner and border bases as well as some resultant based algorithms. We present explicit constructions for square systems which show 'generic' behavior with respect to the Bézout root count in affine space or the Bernstein-Khovanskii-Kushnirenko root count in the algebraic torus. We show how the presented techniques can be used in a *homogeneous* context by introducing *homogeneous normal forms*, which offer an elegant way of dealing with solutions 'at infinity'. For instance, homogeneous normal forms can be used to solve systems which define finitely many solutions in projective space by working in its graded, homogeneous coordinate ring. We develop the necessary theory for generalizing this approach to the homogeneous coordinate ring (or *Cox ring*) of compact toric varieties. In this way we obtain an algorithm for solving systems on a compactification of the algebraic torus which takes the polyhedral structure of the equations into account. This approach is especially effective in the case where the system defines solutions on or near the boundary of the torus in its compactification, which typically causes difficulties for other solvers. Each of the proposed methods is tested extensively in numerical experiments and compared to existing implementations.

Homotopy methods are perhaps the most popular methods for the numerical solution of systems of polynomial equations. One of the reasons is that, in general, their computational complexity scales much better with the number of variables in the system than that of algebraic methods. However, the reliability of these methods depends strongly on some design choices in the algorithm. An important example is the choice of *step size* in the discretization of the solution paths. Choosing this too small leads to a large computational cost and prohibitively long computation times, while choosing it too large may lead to *path jumping*, which is a typical cause for *missing solutions* in the output of a homotopy algorithm. In this thesis, a new *adaptive step size path tracking algorithm* is proposed which is shown to be much less prone to path jumping than the state of the art software.

Beknopte samenvatting

Stelsels veeltermvergelijkingen duiken op in talrijke problemen in toegepaste wiskunde en ingenieurswetenschappen. Voorbeelden van zulke problemen kan men vinden in onder meer de robotica, chemische ingenieurstechnieken, computervisie, dynamische systeemtheorie, signaalverwerking en geometrische modellering. Het numeriek oplossen van een stelsel veeltermvergelijkingen wordt beschouwd als een uitdagend probleem in de computationele wiskunde. Belangrijke klassen van bestaande methodes zijn *algebraïsche methodes*, die het probleem oplossen via eigenwaardenberekeningen, en *homotopiemethodes*, die oplossingspaden volgen in een continue vervorming van het stelsel. In deze tekst stellen we nieuwe algoritmes voor van beide soorten die op verschillende vlakken beter presteren dan de bestaande methodes.

Klassieke voorbeelden van algebraïsche technieken maken gebruik van Gröbner-basissen, border-basissen of resultanten. Deze methodes zijn gebaseerd op het feit dat de oplossingen geëncodeerd zijn in de structuur van een algebra die op een natuurlijke manier door de vergelijkingen van het stelsel wordt gedefiniëerd. Om berekeningen te doen in deze algebra kiezen de algoritmes een *voorstelling* ervan die gebruikelijk bestaat uit een aantal monomen die aan zekere voorwaarden voldoen. In deze thesis tonen we aan dat deze voorwaarden vaak te strikt zijn en mogelijk leiden tot ernstige numerieke onstabiliteit van de algoritmes. Dit resulteert in het feit dat ze niet geschikt zijn voor berekeningen in eindige precisie. We stellen het raamwerk van *afgeknotte normaalvormen* (*truncated normal forms*, TNFs) voor om deze tekortkoming te verhelpen en ontwikkelen nieuwe, robuuste en gestabiliseerde methodes. Het raamwerk veralgemeent Gröbner- en border-basissen, alsook een aantal resultant-gebaseerde algoritmes. We stellen expliciete constructies voor om vierkante systemen op te lossen die 'generiek' gedrag vertonen, waarmee we bedoelen dat ze het verwachte aantal oplossingen hebben in de zin van Bézout of Bernstein-Khovanskii-Kushnirenko. We tonen aan hoe de voorgestelde technieken gebruikt kunnen worden in een *homogene* context door het definiëren van *homogene normaalvormen* (*homogeneous normal forms*, HNFs) die een elegante manier bieden om oplossingen 'op oneindig' af te handelen. Bijvoorbeeld, homogene normaalvormen kunnen gebruikt worden om stelsels op te lossen die eindig veel oplossingen definiëren in de projectieve ruimte door te werken in de homogene coördinaatring. We ontwikkelen de nodige theorie om deze aanpak te veralgemenen naar de homogene coördinaatring (of *Cox ring*) van een compacte

torische variëteit. Op deze manier bekomen we een algoritme voor het oplossen van veeltermstelsels in een compactificatie van de algebraïsche torus die rekening houdt met de polyhedrale structuur van de vergelijkingen. Deze aanpak is vooral effectief in het geval waarin het systeem oplossingen definiëert nabij de rand van de torus in zijn compactificatie, hetgeen typisch een probleem vormt voor andere methodes. Elk van de voorgestelde algoritmes wordt getest in numerieke experimenten en vergeleken met bestaande implementaties.

Homotopiemethodes zijn wellicht de meest populaire methodes voor het numeriek oplossen van een stelsel veeltermvergelijkingen. Één van de redenen daarvoor is dat de rekenkost veel beter schaalt met het aantal variableen in het stelsel dan voor algebraïsche methodes. Echter, de betrouwbaarheid van deze methodes hangt sterk af van een aantal ontwerpkeuzes in het algoritme. Een belangrijk voorbeeld is de keuze van de *stapgrootte* in de discretisatie van de oplossingspaden. Kiezen we deze te klein dan leidt dit tot lange rekentijden. Kiezen we deze te groot dan kan dit leiden tot *path jumping*, wat een typische oorzaak is voor *verloren* oplossingen in de output van een homotopie algoritme. In deze thesis ontwerpen we een *nieuw homotopie algoritme* dat gebruik maakt van een *adaptieve stapgrootte* en tonen we aan dat dit algoritme beduidend minder last heeft van path jumping dan state-of-the-art alternatieven.

List of Abbreviations

CPC convex polyhedral cone.

DCT discrete cosine transform.

GIT geometric invariant theory.

HNF homogeneous normal form.

IDCT inverse discrete cosine transform.

SVD singular value decomposition.

TNF truncated normal form.

List of Symbols

Sets

$\mathbb{C}$	The field of complex numbers
$\varnothing$	The empty set
$\mathbb{N} = \mathbb{Z}_{\geq 0}$	The nonnegative integers
$\mathbb{N}_{>0}$	The positive integers
$\mathbb{R}$	The field of real numbers
$\sqcup$	Disjoint union
$\subset$	Inclusion
$\subsetneq$	Strict inclusion
$\mathbb{Z}$	The ring of integers
$\mathbb{Z}_{<0}$	The negative integers

Rings and ideals

$\mathbb{C}[[t]]$	The ring of formal power series with coefficients in $\mathbb{C}$
$\mathbb{C}[x_1, \ldots, x_n]$	The polynomial ring with coefficients in $\mathbb{C}$ and variables $x_1, \ldots, x_n$
$\mathbb{C}[Y]$	Coordinate ring of Y
HF_I	Hilbert function of a homogeneous ideal I
HP_I	Hilbert polynomial of a homogeneous ideal I
$\langle \mathcal{P} \rangle$	Ideal generated by the elements in $\mathcal{P}$
$\mathfrak{B}$	Irrelevant ideal
$\mathscr{O}_{\mathbb{P}^n}(U)$	Ring of regular functions on an open subset $U \subset \mathbb{P}^n$
$\mathcal{P}$	A set of polynomials
$\mathrm{MaxSpec}(R)$	Maximal spectrum of the ring R
$\sqrt{I}$	The radical of I
I	An ideal

$I(Y)$	The vanishing ideal of the set $Y \subset \mathbb{C}^n$.
I^c	Contraction of I
I^e	Extension of I
$I_S(X)$	Homogeneous vanishing ideal of a projective variety X
$K(R)$	Field of fractions of an integral domain R
R	A commutative ring with 1
R_f	Localization of R at $f \in R$
S	Graded ring
S_d	Graded piece of degree d in the graded ring S
Varieties	
$(\mathbb{C}^*)^n$	The n-dimensional complex algebraic torus
$\mathbb{C}[X]$	Homogeneous coordinate ring of a projective variety X
$\mathbb{C}^n$	n-dimensional affine space
$\dim X$	Dimension of X as a quasi-projective variety
$\dim Y$	Dimension of Y as an affine variety
$\overline{Y}$	Zariski closure of Y
ϕ	A morphism of affine varieties
ϕ^*	The pullback of ϕ
$\mathbb{P}^n$	The n-dimensional complex projective space
$V_X(\mathcal{P})$	subvariety of X defined by the elements of $\mathcal{P}$
$V_X(f_1, \ldots, f_s)$	subvariety of X defined by $\mathcal{P} = \{f_1, \ldots, f_s\}$
$V_X(I)$	The subvariety of X defined by the ideal I
X	A quasi-projective variety
Y	An affine variety
Y_f	The affine variety $\mathrm{MaxSpec}(\mathbb{C}[Y]_f)$
Zero-dimensional ideals and varieties	
δ	Number of solutions
δ^+	Number of solutions, counting multiplicities
ev_{z_i}	Functional representing 'evaluation at the solution z_i'
$\mathcal{G}$	Gröbner basis of an ideal of a polynomial ring
$\mathcal{H}$	Border basis of a zero-dimensional ideal of a polynomial ring
$\mathrm{in}_\prec(f)$	Initial monomial of f w.r.t. $\prec$

μ_i	Multiplicity of a solution z_i
$\prec$	Monomial order on a polynomial ring
$\mathrm{Reg}(I)$	Regularity of a homogeneous ideal I
M_g	$\mathbb{C}$-linear map representing 'multiplication with g'

Normal forms and resultants

$\mathrm{New}_{\mathscr{A}_0,\ldots,\mathscr{A}_n}$	A Canny-Emiris toric resultant matrix
$\mathrm{Mac}_{d_0,\ldots,d_n}$	Macaulay resultant matrix
$\mathcal{N}$	Normal form
$\mathcal{N}_{\mathcal{G}}$	Normal form corresponding to the Gröbner basis $\mathcal{G}$
$\mathcal{N}_{\mathcal{H}}$	Normal form corresponding to the border basis $\mathcal{H}$
$\mathcal{N}_V$	Truncated normal form on V
$\mathcal{N}_{\alpha,\alpha_0}$	Homogeneous normal form for a regularity pair (α,α_0)
$\mathrm{Res}_{\mathscr{A}_0,\ldots,\mathscr{A}_n}$	Toric resultant
$\mathrm{Res}_{d_0,\ldots,d_n}$	Projective resultant
$\mathrm{res}_{f_1,\ldots,f_s}$	Resultant map defined by $f_1,\ldots,f_s$

Families of polynomial systems

$\mathcal{F}_R(d_1,\ldots,d_s)$	Family of total degree systems with degrees $d_1,\ldots,d_s$
$\mathcal{F}_S(d_1,\ldots,d_s)$	Family of total degree homogeneous systems with degrees $d_1,\ldots,d_s$
$\mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1,\ldots,\mathscr{A}_s)$	Family of systems supported in $\mathscr{A}_1,\ldots,\mathscr{A}_s$
$\mathcal{F}_{\mathbb{C}[M]}(P_1,\ldots,P_s)$	Polyhedral family with polytopes $P_1,\ldots,P_s$

Homotopy continuation

$[L/M]_x$	Padé approximant of $x(t)$ of type (L,M)
Γ	A continuous map defining a parameter path in $\mathbb{C}$
γ	A random complex constant
Π	Branched covering of $\mathbb{C}$ associated to a homotopy H
$\mathcal{S}$	Branch locus of a branched covering Π
H	Homotopy
J_H	Jacobian matrix of a polynomial map H

Toric varieties

$\mathbb{C}[\mathsf{S}]$	Algebra of the affine semigroup S
π	Almost geometric quotient in the Cox construction
S	An affine semigroup

Σ	Fan
$\sigma, \sigma^\vee$	Rational polyhedral cone and its dual
Σ_P	Normal fan of a polytope P
G	Reductive group of a toric variety
M	Character lattice of a torus
$M_\mathbb{R}$	Real vector space $M \otimes_\mathbb{Z} \mathbb{R}$ associated to a lattice M
N	Cocharacter lattice or one parameter subgroup lattice of a torus
S	Cox ring of a toric variety
$X_\mathscr{A}$	Projective toric variety defined by the exponents in $\mathscr{A}$
X_Σ	Toric variety associated to a fan Σ
X_P	Toric variety associated to a polytope P
$Y_\mathscr{A}$	Affine toric variety defined by the exponents in $\mathscr{A}$
Z	Base locus of a toric variety

Vector spaces

$\dim_\mathbb{C} V$	Dimension of V as a vector space
$\mathrm{span}_\mathbb{C}(\mathcal{W})$	The $\mathbb{C}$-linear span of the vectors in $\mathcal{W}$
V	A $\mathbb{C}$-vector space
$V^\vee$	Dual vector space of V

Other symbols

Δ_n	The n-dimensional elementary simplex
η_α	Homogenization of degree α
$\mathrm{id}_\mathcal{P}$	The identity map on a set $\mathcal{P}$.
$\mathrm{MV}(P_1, \ldots, P_n)$	Mixed volume of $P_1, \ldots, P_n$
$\simeq$	Isomorphism
$\mathrm{Supp}(f)$	Support of a Laurent polynomial f

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This text is about the mathematical problem of solving a system of polynomial equations, which is a fundamental problem in nonlinear algebra and algebraic geometry. Application areas of this problem include cryptography, signal processing, data science, chemical engineering, robotics and computer vision, to name a few.

With motivations coming mainly from pure mathematics, the research on algorithms for solving polynomial equations in the 19th and most of the 20th century focused on symbolic methods. This led to major advances in computer algebra with the development of powerful tools for testing theories, formulating conjectures and even proving theorems. Although very useful for such purposes, symbolic manipulation is often unfeasible for problems coming from applications. There are two main reasons for this. Firstly, the scale of such problems can be very large, requiring too much time for symbolic algorithms to terminate. Secondly, the input data of the problems (e.g. the coefficients of the polynomials) may come from measurements or previous numerical computations. The representation of these data as rational numbers requires the use of large integers, which rapidly leads to memory issues. These observations establish the need for robust numerical algorithms that produce reliable results in finite precision arithmetic. Somewhat surprisingly, the fields of numerical nonlinear algebra and numerical algebraic geometry have remained largely uncharted territory until the end of the 20th century. One possible explanation is that numerical analysts have rarely been exposed to commutative algebra or algebraic geometry in their undergraduate years. On top of that, the classical sources on these subjects often assume a background in algebra and topology that excludes numerical analysts and engineers from their reading audience. Books such as *Ideals, Varieties and Algorithms* and *Using Algebraic Geometry* by Cox, Little and O'Shea are game changers from this perspective. Among other things, the publication of such books has paved the way for today's growing community of applied and numerical algebraic geometers.

In this text, we have aimed to include background information on basic algebraic geometry, commutative algebra, numerical analysis and numerical linear algebra. We

assume basic knowledge of algebraic structures, linear algebra and floating point numbers. For the sake of readability, some of the preliminary material is moved to an appendix and references are provided where a full discussion would be too lengthy.
In the first section of this chapter, we state the problem of solving a system of polynomial equations in its simplest form and discuss some conventions used in this thesis. In Section 1.2 we present a selection of applications of the problem in some more detail. Section 1.3 gives an overview of some state of the art methods. In Section 1.4 we describe the goals of this thesis and our main contributions. Finally, in Section 1.5 we discuss the outline of the thesis.

1.1 Polynomial systems

Let $R = \mathbb{C}[x_1, \ldots, x_n]$ be the ring of n-variate polynomials with coefficients in $\mathbb{C}$. An element $f \in R$ defines a function $f : \mathbb{C}^n \to \mathbb{C}$. We will use the short notation $x = (x_1, \ldots, x_n) \in \mathbb{C}^n$ and when $n \leq 3$ we may use variable names such as x, y, z instead of x_1, x_2, x_3 to avoid subscripts. Given s elements $f_1, \ldots, f_s \in R$, we define the map $F : \mathbb{C}^n \to \mathbb{C}^s$ such that

$$F(x) = (f_1(x), \ldots, f_s(x)).$$

We will be interested in the inverse image of the origin in $\mathbb{C}^s$ under this map, i.e., in the fiber

$$F^{-1}(0) = \{x \in \mathbb{C}^n \mid F(x) = 0\}.$$

This set consists of all the points satisfying the relations

$$f_1(x) = \cdots = f_s(x) = 0.$$

Therefore, $F^{-1}(0)$ is called the *set of solutions* of the *system of polynomial equations* defined by $f_1, \ldots, f_s$. In this context, by *solving* the system of polynomial equations $f_1 = \cdots = f_s = 0$ we mean 'computing' $F^{-1}(0)$. Here we have to specify what we mean by 'computing' a set of points in $\mathbb{C}^n$. Some issues are:

1. The set $F^{-1}(0)$ may be infinite.
2. There may be no expression in radicals for the coordinates of the points in $F^{-1}(0)$, i.e. there is no algorithm that computes these coordinates in finite time.

Example 1.1.1. If $n = 2$ and $s = 1$, then $f(x, y) = 0$ defines infinitely many points in $\mathbb{C}^2$ unless f is a nonzero constant function. If $n = s = 1$, then there is no expression in radicals for the roots of a general quintic $a_5 x^5 + a_4 x^4 + \cdots + a_0 = 0$ by the famous Abel-Ruffini theorem. △

In this thesis, we will assume that $f_1, \ldots, f_s$ are such that the first situation does not occur. That is, we will assume that $F^{-1}(0)$ consists of isolated points, and this implies that $F^{-1}(0)$ is finite by Bézout's theorem 3.1.2. As we will see (Theorem 2.2.4), in order for this assumption to be satisfied we must have $s \geq n$. A system with finitely many solutions is called *zero-dimensional*, which refers to the dimension of $F^{-1}(0)$ as an affine algebraic variety. We will say more about dimension in Chapter 2 and use it as an intuitive concept for now. If the f_i are non-constant, the 'expected dimension' of $F^{-1}(0)$ is $n - s$, where negative dimensions (for $s > n$) indicate that $F^{-1}(0)$ is expected to be the empty set.

Example 1.1.2. If $f_i = a_{i0} + a_{i1}x_1 + \cdots + a_{in}x_n$ are affine functions, then $F^{-1}(0)$ is the affine space of solutions of a linear system of equations defined by an $s \times n$ matrix $A = (a_{ij})_{1 \leq i \leq s, 1 \leq j \leq n}$. The dimension of the solution space is $n - s$, except when the matrix A is not of full rank. $\triangle$

This means that systems given by n equations in n variables are expected to have finitely many isolated solutions. Systems for which $n = s$ are called *square systems*. They form an important class of polynomial systems and they will play an important role in this thesis.
The second issue listed above means that there is no hope for developing algorithms for computing *exactly* the coordinates of the solutions of any system of polynomial equations in finite time. However, the solutions can be approximated to arbitrary precision by using, for instance, Newton's method. Motivated by this, by 'computing' the solutions of $f_1 = \cdots = f_s = 0$ we mean computing *satisfactory numerical approximations* of the *coordinates* of the solutions in $\mathbb{C}^n$. A way of measuring the quality of an approximate solution is discussed in Appendix C.

In the formulation above, $\mathbb{C}^n$ is called the *solution space* of the system $f_1 = \cdots = f_s = 0$. Especially when dealing with systems in more than one variable ($n > 1$) it may be convenient to work with different solution spaces X, as we will do later on in this text. In the more general context, on which we will not elaborate until Section 3.2, F will be a section of a rank s algebraic *vector bundle* on X, and the set of solutions is the *zero locus* of F in X. One of the reasons for changing the solution space is that systems may define solutions 'at infinity', and for numerical stability reasons we may want to include 'infinity' in our solution space. This leads for instance to the projective solution space $X = \mathbb{P}^n$ (see Section 2.2) or other compact toric varieties (see Chapter 5). In all these cases, we will define coordinates on our solution space X, and by solving we mean computing *satisfactory numerical approximations* of the *coordinates* of the solutions in X.

Throughout this thesis, we will mostly work with polynomials, varieties and matrices over the complex numbers $\mathbb{C}$. This choice needs to be motivated, since many systems arising from applications have real coefficients and it is often only important to compute the real solutions. On top of that, the number of real solutions can be much smaller than the number of complex solutions. Real solutions of polynomial systems are studied in the field of *real algebraic geometry* [BCR13, Sot03]. Finding only the real solutions without computing all complex solutions first is a hard problem that is still largely open. One reason is the fact that $\mathbb{C}$ is *algebraically closed* and $\mathbb{R}$ is not.

In fact, $\mathbb{C}$ is the algebraic closure $\overline{\mathbb{R}}$ of $\mathbb{R}$, which means that $\mathbb{C}$ is the smallest of all fields K containing $\mathbb{R}$ such that every non-constant polynomial in $\mathbb{R}[x]$ has a solution in K. This implies that we can invoke Hilbert's Nullstellensatz (see Subsection 2.1.3), which is a celebrated result in algebraic geometry. It also leads to the fact that for certain families of polynomial systems and varieties, one can make statements about what happens *in general* or *generically*. Finally, working over the complex numbers is essential for the success of homotopy continuation methods (see Chapter 6) for solving polynomial systems. In conclusion, although many of the polynomial systems we are interested in have coefficients in $\mathbb{R}$, we will solve them over $\mathbb{C} = \overline{\mathbb{R}}$, and if we are only interested in real solutions, we will adopt the usual strategy of computing all complex solutions in $\mathbb{C}^n$ and taking the intersection with $\mathbb{R}^n$.

Example 1.1.3. Consider a general quadratic polynomial $f = ax^2 + bx + c \in \mathbb{R}[x]$ with $a \neq 0$. The polynomial f has two solutions in $\mathbb{R}$ when $b^2 - 4ac > 0$, one solution in $\mathbb{R}$ when $b^2 - 4ac = 0$ and no solutions in $\mathbb{R}$ when $b^2 - 4ac < 0$. A geometric way of thinking about this is the following. The discriminant surface $\{(a,b,c) \in \mathbb{R}^3 \mid b^2 - 4ac = 0\}$ partitions the parameter space $\{(a,b,c) \in \mathbb{R}^3 \mid a \neq 0\}$ into two compartments, each with a different real root count. A quadratic equation $f = ax^2 + bx + c \in \mathbb{C}[x]$ with $a \neq 0$ always has a solution in $\mathbb{C}$, and for general a, b, c there are two solutions in $\mathbb{C}$. If there is only one solution, then $b^2 - 4ac = 0$. A general cubic $f = ax^3 + bx^2 + cx + d \in \mathbb{R}[x], a \neq 0$ may have 1 or 3 solutions in $\mathbb{R}$. The discriminant is now given by $\Delta_f = b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd = 0$. If the coefficients are complex, there are 3 solutions except when $\Delta_f = 0$. $\triangle$

1.2 Applications

Systems of polynomial equations are at the heart of many problems in pure and applied mathematics. Some examples are computing all possible conformations of molecules in molecular biology [EM99a], the design of wavelet families in signal processing [Tel16, Section 1.2], analyzing feasible robot configurations in robotics [WS11], computing Nash equilibria in economics and game theory [Stu02, Chapter 6] (or [WS05, Chapter 9]), numerous applications in statistics [Sul18], curvature and bottleneck computation in topological data analysis [Bre20] and solving linear partial differential equations with constant coefficients [Stu02, Chapter 10]. The author learned about several of these applications and others in a course taught by David Cox at the 2018 CBMS conference on 'Applications of Polynomial Systems'. The course material has recently been published in [Cox20a]. In the remainder of this section we present a selection of other applications of polynomial systems in some more detail.

1.2.1 Polynomial optimization

Systems of polynomial equations often arise in applications in the form of a *polynomial optimization problem* [AL11], where the goal is to minimize a polynomial objective function $g(x_1,\ldots,x_k) \in \mathbb{R}[x_1,\ldots,x_k] \subset \mathbb{C}[x_1,\ldots,x_k]$ over a real algebraic set (the zero locus of a set of polynomials $h_1,\ldots,h_\ell \in \mathbb{R}[x_1,\ldots,x_k]$ in $\mathbb{R}^k$). That is, we consider the optimization problem

$$\begin{aligned} \min_{x\in\mathbb{R}} \qquad & g(x_1,\ldots,x_k), \\ \text{subject to} \qquad & h_1(x_1,\ldots,x_k) = \cdots = h_\ell(x_1,\ldots,x_k) = 0. \end{aligned} \tag{1.2.1}$$

This is an example where one is only interested in real solutions: minimizing over the complex numbers does not make much sense. Introducing new variables $\lambda_1,\ldots,\lambda_\ell$ we obtain the Lagrangian $L = g - \lambda_1 h_1 - \cdots - \lambda_\ell h_\ell$, whose partial derivatives give the optimality conditions

$$\frac{\partial L}{\partial x_1} = \cdots = \frac{\partial L}{\partial x_k} = h_1 = \cdots = h_\ell = 0. \tag{1.2.2}$$

This is a polynomial system with $n = s = k + \ell$. The real solutions are obtained by computing all the complex solutions and intersecting with $\mathbb{R}^k$. By the discussion in Section 1.1, the number of solutions is typically finite.

Example 1.2.1 (Euclidean distance degree)**.** Given a general point $y = (y_1,\ldots,y_k) \in \mathbb{R}^k$, we consider the (squared) Euclidean distance function $g(x_1,\ldots,x_k) = \|x-y\|_2^2 = (x_1-y_1)^2 + \cdots + (x_k-y_k)^2$. Let Y be the zero-locus of $h_1,\ldots,h_\ell \in \mathbb{R}[x_1,\ldots,x_k]$:

$$Y = \{x \in \mathbb{R}^k \mid h_1 = \cdots = h_\ell = 0\}.$$

Consider the optimization problem (1.2.1) given by these data. The solution y^* is the point on Y that's closest to y. The number of *complex* solutions of (1.2.2) is called the *Euclidean distance degree* of Y [DHO+16]. The authors of [DHO+16] point out that if y is a noisy sample from Y, then y^* is the maximum likelihood estimate for y under the assumption that the noise has a standard Gaussian distribution in $\mathbb{R}^n$. △

Example 1.2.2 (Computing critical points)**.** In many applications one is interested in finding the critical points of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, not necessarily polynomial, in a bounded domain $\Omega \subset \mathbb{R}^n$. These are the real solutions in Ω of

$$\frac{\partial f}{\partial x_1} = \cdots = \frac{\partial f}{\partial x_n} = 0. \tag{1.2.3}$$

A possible strategy for finding these points is approximating f by a polynomial function $\tilde{f}$ on Ω and computing the critical points of $\tilde{f}$ in Ω instead. An effective way of doing this approximation numerically is by the use of multivariate Chebyshev interpolants [Mas80, Tre17]. Replacing f in (1.2.3) by $\tilde{f}$ gives the optimality conditions for an unconstrained version of (1.2.1). This approach is used in [NNT15] (in combination with domain subdivision) for solving one of the SIAM 100-Digit Challenge problems [Tre02]. △

Example 1.2.3 (Parameter estimation for system identification)**.** System identification is an engineering discipline that aims at constructing models for dynamical systems from measured data [Lju86]. The general model for a discrete time, single-input single-output linear time-invariant system with input sequence $u : \mathbb{Z} \to \mathbb{R}$, output sequence $y : \mathbb{Z} \to \mathbb{R}$ and white noise sequence $e : \mathbb{Z} \to \mathbb{R}$ is

$$A(q)y(t) = \frac{B_1(q)}{B_2(q)}u(t) + \frac{C_1(q)}{C_2(q)}e(t).$$

Here $A, B_1, B_2, C_1, C_2 \in \mathbb{C}[q]$ are unknown polynomials in the backward shift operator q which acts on any sequence $s : \mathbb{Z} \to \mathbb{R}$ by $qs(t) = s(t-1)$. Let $d_A, d_{B_1}, d_{B_2}, d_{C_1}, d_{C_2}$ be the degrees of these polynomials, which depend on the choice of model. Clearing denominators gives

$$A(q)B_2(q)C_2(q)y(t) = B_1(q)C_2(q)u(t) + B_2(q)C_1(q)e(t). \tag{1.2.4}$$

Suppose we have measured $u(0), \ldots, u(N), y(0), \ldots, y(N)$. Then we can find algebraic relations among the coefficients of A, B_1, B_2, C_1, C_2 by writing (1.2.4) down for $t = d, d+1, \ldots, N$ where

$$d = \max(d_A + d_{B_2} + d_{C_2}, d_{B_1} + d_{C_2}, d_{B_2} + d_{C_1}).$$

The coefficients of these polynomials are then estimated by solving the polynomial optimization problem

$$\begin{aligned} \min_{\Theta \in \mathbb{R}^k} \quad & e(0)^2 + \ldots + e(N)^2 \\ \text{subject to} \quad & (1.2.4) \text{ is satisfied for } t = d, \ldots, N \end{aligned}$$

where Θ is the set of parameters consisting of $e(0), \ldots, e(N)$ and the unknown coefficients of A, B_1, B_2, C_1, C_2. The interested reader is referred to [Lju86, chapter 7] for a detailed treatment of parameter estimation in system identification, and to [Bat13, Subsection 1.1.1] for a worked out example. △

1.2.2 Chemical reaction networks

The equilibrium concentrations of the chemical species occuring in a *chemical reaction network* satisfy algebraic relations. Taking advantage of the algebraic structure of these networks has led to advances in the understanding of their dynamical behaviour. We refer the interested reader to [Dic16] and references therein. The network below involves 4 species A, B, C, D and models T cell signal transduction (see [Dic16]).

$$\begin{tikzcd}
 & A+B \arrow[dr, shift left, "\kappa_{12}"] & \\
D \arrow[ur, "\kappa_{31}"] & & C \arrow[ll, "\kappa_{23}"'] \arrow[ul, shift left, "\kappa_{21}"]
\end{tikzcd}$$

The parameters $\kappa_{12}, \kappa_{21}, \kappa_{31}, \kappa_{23} \in \mathbb{R}_{>0}$ are the reaction rate constants. Let x_A, x_B, x_C, x_D denote the time dependent concentrations of the species A, B, C, D respectively. The law of mass action gives the relations

$$
\begin{aligned}
f_A = \frac{dx_A}{dt} &= -\kappa_{12} x_A x_B + \kappa_{21} x_C + \kappa_{31} x_D, \\
f_B = \frac{dx_B}{dt} &= -\kappa_{12} x_A x_B + \kappa_{21} x_C + \kappa_{31} x_D, \\
f_C = \frac{dx_C}{dt} &= \kappa_{12} x_A x_B - \kappa_{21} x_C - \kappa_{23} x_C, \\
f_D = \frac{dx_D}{dt} &= \kappa_{23} x_C - \kappa_{31} x_D.
\end{aligned}
$$

The set $\{(x_A, x_B, x_C, x_D) \in (\mathbb{R}_{>0})^4 \mid f_A = f_B = f_C = f_D = 0\}$ is called the *steady state variety* of the chemical reaction network. By the structure of the equations, for given initial concentrations, the solution (x_A, x_B, x_C, x_D) cannot leave its *stoichiometric compatibility class*, which is an affine subspace of $(\mathbb{R}_{>0})^4$. Adding the affine equations of the stoichiometric compatibility class to the system, we get the set of all candidate steady states. We conclude by pointing out that there are remarkable connections with toric geometry [CDSS09] and geometric modeling [CGPS08].

1.2.3 Tensor decomposition

Tensors, as a generalization of matrices, are represented in coordinates by multi-dimensional arrays. They have numerous applications in signal processing, chemistry and data mining, among others [KB09, Com02, CMDL$^+$15, SDLF$^+$17]. In these applications, a frequently encountered problem is to find a decomposition of a tensor into a sum of 'simple' tensors. For example, the tensor rank decomposition or *Canonical Polyadic Decomposition* (CPD) of a third order tensor $\mathcal{A} \in \mathbb{C}^l \otimes \mathbb{C}^m \otimes \mathbb{C}^n$ is

$$
\mathcal{A} = \sum_{i=1}^{r} x_i \otimes y_i \otimes z_i \tag{1.2.5}
$$

where r is the *rank* of $\mathcal{A}$ (it is the minimal number for which such a decomposition exists), $x_i \in \mathbb{C}^l, y_i \in \mathbb{C}^m, z_i \in \mathbb{C}^n$ and a term $x_i \otimes y_i \otimes z_i$ is called *a rank-one tensor*, or *elementary tensor* [DSL08]. Equivalently, in coordinates we can write (1.2.5) as

$$
\mathcal{A}_{jk\ell} = \sum_{i=1}^{r} x_{ij} y_{ik} z_{i\ell}, \quad 1 \leq j \leq l, 1 \leq k \leq m, 1 \leq \ell \leq n. \tag{1.2.6}
$$

Even when the rank r is known, it is considered a difficult problem to find the rank-one summands in (1.2.5). It is clear from (1.2.6) that the entries of the x_i, y_i, z_i are the solutions to a set of polynomial equations. Some variables can be eliminated by observing that $ax_i \otimes by_i \otimes cz_i = (abc)(x_i \otimes y_i \otimes z_i)$. That is, with an appropriate change of coordinates one can assume that $x_{i1} = y_{i1} = 1$, and the solution space has dimension $r(l+m+n-2)$. For some formats (l,m,n), there exists $r \in \mathbb{N}$ such that the resulting polynomial system is square. These are the formats (l,m,n) for which $lmn/(l+m+n-2) \in \mathbb{N}$. Such tensor formats are called *perfect* and homotopy methods of numerical algebraic geometry have proved very useful for investigating the identifiability and the generic number of possible decompositions [HOOS19].

In applications, the data in $\mathcal{A}$ are often contaminated by noise and there is no hope for having equality in (1.2.5) for low ranks r. One is usually interested in finding a rank r tensor that *approximates* $\mathcal{A}$. In the case where $r=1$ one computes the critical points of the algebraic function

$$\sum_{jk\ell} (\mathcal{A}_{jk\ell} - x_j y_k z_\ell)^2,$$

which is another example of polynomial optimization (Subsection 1.2.1).

In [KL18], homotopy continuation methods have been successfully applied for decomposing *unbalanced* tensors (in our example, these are tensors with $r < \max(l,m,n)$). The key ingredient is an alternative algebraic formulation for the decomposition problem using basic (multi-)linear algebra techniques.

Symmetric tensors $\mathcal{A}$ in $(\mathbb{C}^l)^{\otimes d}$ (i.e., tensors for which the coordinates are invariant under permutation of the indices) are homogeneous polynomials $f_{\mathcal{A}}$ of degree d in l variables (see Section 2.2). For $d=2$, this statement reduces to the standard observation that a matrix $A \in \mathbb{C}^{l\times l} = \mathbb{C}^l \otimes \mathbb{C}^l$ defines a quadratic form $f_A(u) = u^\top A u$ where $u = (u_1, \ldots, u_l)^\top$. The *symmetric tensor rank decomposition* of a symmetric tensor $\mathcal{A}$ is the decomposition of $\mathcal{A}$ into a minimal sum of *symmetric* elementary tensors. The number of summands is called the *symmetric rank*. This decomposition is given by the *Waring decomposition* of the corresponding homogeneous polynomial, which is its minimal decomposition into a sum of powers of linear forms. For instance

$$(\mathbb{C}^l)^{\otimes 3} \ni \mathcal{A} = \sum_{i=1}^{r} x_i \otimes x_i \otimes x_i \quad \sim \quad f_{\mathcal{A}}(u) = \sum_{i=1}^{r} l_i(u)^3,$$

where $l_i(u) = x_{i1}u_1 + \cdots + x_{il}u_l$. *Apolarity theory* relates the problem of finding the Waring decomposition to the theory of polynomial system solving [IK99, Chapters 1-2]. This was exploited in [BCMT10, BT20b] to design an algorithm for symmetric tensor decomposition which combines ideas from algebraic polynomial system solving methods and homotopy methods.

So far, we have discussed how polynomial system solving techniques can be applied to solve tensor decomposition problems. Going the other way around, in [VSDL17a, VSDL17b] the authors use tensor decomposition as the last step in their algorithm for solving systems of polynomial equations. The connection between the CPD of third order tensors and joint eigenvalue decomposition of commuting matrices, as discussed in [DL06], is exploited. Multiple roots of the polynomial system are handled using the *block term decomposition* and the algorithms can be used in particular for solving noisy, overdetermined systems.

1.2.4 Computer vision

An important problem in computer vision is that of estimating internal calibration parameters of a camera or camera displacement from point correspondences in a sequence of images [HZ03]. Every such point correspondence imposes an algebraic relation on the parameters that are to be estimated. For some minimal number of points, the number of solutions to the resulting system is finite. Problems that can be formulated in this way are called *minimal problems* [Kuk13].

Example 1.2.4 (relative pose problems)**.** Consider a moving, fully calibrated camera taking two pictures of the same object at different moments in time. In these pictures, there are certain points that correspond to one another. For instance, if the object is a cube, one of its vertices might appear in both pictures. A question one could ask is: 'What is the minimal number of point correspondences that we need to know such that there are only finitely many possible displacements of the camera that can realize these correspondences?' The answer to this question is five [Nis04]. If the focal length of the camera needs to be estimated as well (i.e. the camera is not fully calibrated), we need six point correspondences. △

Example 1.2.5 (the 8-point radial distortion problem)**.** The epipolar geometry and one parameter radial lens distortion of a camera can be estimated simultaneously from eight point correspondences [KP07]. This problem has several alternative formulations. See [Kuk13, Section 7.1] for a formulation as a polynomial system with 7 equations in 7 unknowns, and a different formulation as a system with 3 equations and 3 unknowns. In the first formulation with $n = s = 7$, there are 6 equations of degree 2 and one of degree 3. In the formulation with $n = s = 3$, two equations have degree 3 and one has degree 5. Geometric problems coming from applications can often be described by different polynomial models with solution spaces of different dimensions. Typically, as is the case in this example, the price one pays for reducing the number of variables is an increase of the degree of the equations and vice versa. We will say a bit more about the structure of the equations in the $n = s = 3$ formulation in Experiment 5.5.2. △

1.3 State of the art

In this section we give an overview of the available methods for solving systems of polynomial equations. We will elaborate more on methods related to those proposed in this thesis in later chapters. For more information, the reader can consult overview books such as [Stu02, WS05, EM07, CCC$^+$05]. Strategies for solving polynomial equations over the complex numbers can be roughly subdivided into two classes. One class of methods reduces the problem to a univariate root finding problem or an eigenvalue problem via algebraic manipulations of the input polynomials. We refer to such methods as *algebraic methods*. Other methods use a topological approach, where a polynomial system is continuously deformed into another one and numerical methods are used to track the paths of the isolated solutions. Such methods are referred to

as *homotopy methods*. We give an overview of algebraic and homotopy methods in Subsections 1.3.1 and 1.3.2 respectively.
We should mention that there is another popular class of methods, called *subdivision methods*, for finding solutions in bounded domains of $\mathbb{R}^n$ [MP09]. The approach uses a combination of domain reduction and domain subdivision for iterative refinement of the subregions where solutions may be located. We will not give any details here, since both the used techniques and the scope of these methods are fundamentally different from the ones in this thesis. We refer the interested reader to [MP09] and references therein.

1.3.1 Algebraic methods

We denote by $I \subset R$ the ideal generated by the polynomials $f_1, \ldots, f_s$ defining our system of polynomial equations. As explained in Section 3.1, the solutions of the polynomial system are encoded in the $\mathbb{C}$-algebra structure of the residue ring R/I. Algebraic methods for polynomial system solving deduce the algebraic structure of R/I by performing linear algebra operations on vector subspaces of I.

This approach finds its origins in 18th, 19th and early 20th century works on elimination theory and resultants by Bézout, Waring, Poisson, Sylvester, Cayley, Macaulay… [Béz79, War91, Poi02, Syl40, Cay64, Mac02, Mac94]. Matrices whose entries are coefficients of the polynomials $f_1, \ldots, f_s$ play a key role in these works, and they continue to do so in research on algebraic solving methods today. An explicit construction of such matrices was introduced for computing projective resultants, see e.g. [Mac02]. These matrices are also called *Macaulay resultant matrices* or, in the case of two homogeneous polynomials in two variables, *Sylvester resultant matrices*. See [CLO06, Chapter 3] for a detailed treatment. Analogous constructions have been described for computing *toric* or *sparse resultants* [EC93, PS93, D'A02, DS15]. These are among the main objects of study in sparse elimination theory and find their origins in the foundational work of Gel'fand, Kapranov and Zelevinsky [GKZ94]. Other types of matrix constructions come from *residual resultants* [Bus01] and *Bézoutians* [CCC$^+$05, Chapter1]. An overview of these matrix techniques can be found in [EM99b] and a nice summary of the history of elimination theory is given in [Cox20a, Chapter 1]. Although the original application of the theory of elimination and resultants was mainly in symbolic computing, the methods have been analyzed and used in a numerical context; see for instance [Tel16, JV05, BKM05]. We will say more about resultants in Section 3.4. In [Bat13, Dre13, DBDM12] (non-square) Macaulay-type matrices are used for root finding in a numerical linear algebra context. The authors have also developed algorithms that exploit the structure of these matrices (see, e.g., [BDDM14]) and show that their methods are useful in an overdetermined context where equations may be contaminated by noise. All of these tools can be used to reduce the problem of solving polynomial systems to a classical, generalized or polynomial eigenvalue problem.

Another well-established approach to describe the algebra R/I uses *Gröbner bases*. A Gröbner basis for I with respect to a certain *term order* is a finite set of generators for the ideal I satisfying some criteria (see Section 3.3). These criteria make the set of generators extremely useful for computations with and modulo the ideal. Gröbner bases were introduced in 1965 by Bruno Buchberger in his Ph. D. thesis [Buc06] entitled *An algorithm for finding the basis elements of the residue class ring of a zero dimensional polynomial ideal.* In this thesis he also presents what is now called the *Buchberger algorithm* for computing Gröbner bases. Many algorithms in computer algebra rely on (optimized versions of) this algorithm. A great introduction to the basics of Gröbner bases and the Buchberger algorithm can be found in [CLO13, Chapters 2-3] or [AL94, Chapter 1]. More advanced topics are discussed in [Stu96]. Great improvements on the efficiency of Gröbner basis computation have been made by using linear algebra tools. This has led to Faugère's F4 and F5 algorithms [Fau99, Fau02], which are considered the state of the art algorithms. The FGb library [Fau10] has an implementation of these algorithms and an interface to Maple [Map18]. The development of specialized Gröbner basis algorithms is an active area of research; see e.g. [BFT19] for Gröbner bases in a toric context.

Gröbner basis computations depend strongly on a choice of term order (see Section 3.3). *H-bases*, introduced by Macaulay [Mac94], are a different type of ideal bases which can be viewed as a 'coarser' version of Gröbner bases. The term order (which is always a total order on monomials) is replaced by a coarser order on monomials given by the total degree. Such bases have interesting properties and can be used, like Gröbner bases, for computing normal forms and to describe R/I [MS00].

Although Gröbner bases are indispensable symbolic tools for algebraic root finding, their use in a numerical context has remained limited. The reason is that Gröbner basis computations are *numerically unstable*. One of the causes is the fact that the set of *standard monomials* (these are the residue classes of monomials corresponding to a term order that form a basis for R/I, see Section 3.3) change discontinuously with the coefficients of the input polynomials $f_1, \ldots, f_s$ [Ste97, Mou99]. We will give an example in Subsection 3.3.2. To address this drawback of Gröbner bases, *border bases* have been introduced [AS88, MMM91, Möl93, Ste97, Mou99, KK05, KKR05, KK06, MT08]. A border basis for I is a finite set of generators of I satisfying criteria that are less strict than those imposed on Gröbner bases. For example, border bases do not necessarily correspond to a term order. For some finite dimensional vector subspace $B \subset R$, a border basis establishes the equality $R = B \oplus I$ identifying $R/I \simeq B$ as vector spaces. It is commonly required that B be *connected to 1* (see [Mou99]). If B is spanned by a set $\mathcal{B}$ of monomials of R, this restriction is sometimes made stronger by imposing that $\mathcal{B}$ be an *order ideal* (e.g. [KKR05]). Both restrictions are satisfied by the span of the standard monomials coming from a Gröbner basis computation. These generalizations lead to more robust numerical methods than Gröbner bases. The algorithms work with matrices that are usually smaller than resultant constructions because of their incremental nature [MT00]. However, these techniques do not offer a canonical choice for the representation of R/I that is optimized for numerical stability. This is mentioned as an open problem in [Mou07] and will be addressed in this thesis.

1.3.2 Homotopy methods

The strategy of homotopy continuation methods for solving systems of polynomial equations can be described (omitting many subtleties) as follows. Consider $F : \mathbb{C}^n \to \mathbb{C}^n$ as in Section 1.1 where we take $s = n$. Suppose $G : \mathbb{C}^n \to \mathbb{C}^n$ represents a different, square polynomial system whose solutions we know or can be easily computed. On top of that, assume that G has the same number of solutions as F. The next step is to construct a polynomial map

$$H : \mathbb{C}^n \times \mathbb{C} \to \mathbb{C}^n \quad \text{such that} \quad H(x, 0) = G(x) \quad \text{and} \quad H(x, 1) = F(x).$$

For instance

$$H(x, t) = (1 - t)G(x) + tF(x).$$

In this setup G is called the *start system* and F is called the *target system* of the *homotopy* H. As t goes from 0 to 1 along any continuous 1-real dimensional path in $\mathbb{C}$, the polynomial map G deforms continuously into F. If this path is 'nice', the solutions will describe smooth, continuous paths in $\mathbb{C}^n$ during this deformation, and the idea of homotopy continuation is to track these paths numerically. This is usually done by discretizing the path into small steps and applying a *predictor-corrector* scheme. An introduction to homotopy continuation can be found in [AG12, MS87, Li97, SVW01, SVW05, WS05].

Working over the complex numbers is crucial for the success of homotopy continuation methods (although recently, in [EdW19], the authors have made some progress in investigating what is possible over the reals). This means that these methods have an intrinsic numerical character. In fact, numerical path tracking is strongly related to numerically solving initial value problems given by ordinary differential equations [WS11, Part 2].

Constructing an appropriate start system G is an interesting problem on its own. One issue is that if G has too many solutions, some paths will diverge to infinity as t approaches 1. This leads to waste of computational efforts, which is of course undesirable. If all paths converge to a solution of F, the homotopy is called *optimal* [HSS98]. Optimal homotopy constructions exist for some important types of polynomial systems. Examples are *total degree homotopies* for square systems with the Bézout number of solutions [WS05, Subsection 8.4.1], *multihomogeneous homotopies* for square systems with the multihomogeneous Bézout number of solutions [Wam93] or *polyhedral homotopies* for square systems with the BKK number of solutions [HS95, VVC94]. We will say more about these solution counts in Sections 3.1 and 5.1.

Under the right assumptions on the path that is followed in the parameter space, the solution paths are smooth and do not cross each other along the way. However, if the system F has singular solutions, some paths may come together at $t = 1$. Also, if we were not able to construct an optimal homotopy, some paths may diverge to infinity. For dealing with this type of situations, so-called *end games* have been developed [MSW92a, MSW92b, HV98]. An alternative way of dealing with diverging paths is compactifying the solution space. It is common practice to track paths in projective and multiprojective spaces [WS05, Chapter 3].

An important reliability issue of these methods is the possibility of *path jumping*. This is the phenomenon where the numerical approximation of a point on one path jumps to another path along the way. This happens, for instance, when the predicted next point on the path is too far off and lands in the Newton basin of attraction of a different path. In order to avoid this problem, the steps taken in the discretization of the path should be small enough. On the other hand, taking the step size too small would result in a high computational cost. Motivated by this, *adaptive step size methods* have been developed that aim to choose the step size adaptively by detecting which regions of the path are easy/hard to track [SC87, KX94, GS04, Tim20]. In this thesis, we will propose a path tracking algorithm that proves to be more robust than the state of the art implementations with respect to path jumping.

Some state of the art implementations of the homotopy continuation method for solving systems of polynomial equations are Bertini [BSHW13], PHCpack [Ver99], HOM4PS [LLT08] and the recently developed Julia package HomotopyContinuation.jl [BT18]. We should also mention that *certified* path trackers have been developed [HS12, HLJ16, XBY18], which avoid path jumping and provably compute approximate solutions to the polynomial system F in the sense of Smale's α-theory [BCSS12]. However, these methods are computationally significantly more expensive and the certification assumes that the coefficients of the input systems are known *exactly*.

If one of the solutions of F is known, one could construct a homotopy $H(x,t)$ such that $H(x,0) = H(x,1) = F(x)$ by describing a closed loop in the parameter space. If this loop encircles some branchpoints, tracking the corresponding solution path will give us a new solution of the system. This is (again, omitting many details) the approach taken in *monodromy solvers* [DHJ$^+$19], which turn out to be very successful for generating start systems and starting solutions.

1.4 Research goals and contributions

Given a polynomial system $f_1 = \cdots = f_s = 0$ with solution space X defining finitely many solutions, our aim in this thesis is to develop *new algorithms* that work in *finite precision arithmetic* for finding *numerical approximations* of the *coordinates* of the solutions on X. In particular, with these algorithms we seek to address *numerical stability* and *robustness* issues of existing implementations. We develop the necessary theory for presenting the algorithms and perform numerical experiments to show their effectiveness in comparison with the state of the art. The numerical algorithms we present in this thesis are of two different types: some are algebraic solvers using normal forms and eigenvalue computations, others are homotopy algorithms.

Classical algebraic methods impose restrictions on the representation of the quotient algebra associated to a polynomial system which may lead to *ill-conditioned rewriting*

rules. More specifically, often *monomial* bases are used which either come from a *monomial ordering* or which satisfy some connectedness property (see Section 3.3). We develop the framework of *truncated normal forms* (TNFs) which allows more general, possibly non-monomial representations for the quotient algebra and leads to significant improvements in the stability of normal form algorithms. For example, an algorithm based on the classical Macaulay resultant construction fails at computing the 400 intersection points of two general degree 20 curves in the plane: the backward error is $O(1)$. With the TNF algorithm proposed in Subsection 4.3.2 we can compute all 28900 intersection points of two general degree 170 curves with a backward error no larger than 10^{-8} (see Subsection 4.3.3). The key feature of the algorithm that realizes this improvement is an *automatic choice of representation* for the quotient algebra with *good numerical properties* by applying standard tools from numerical linear algebra.
Truncated normal forms generalize both Gröbner and border bases. We develop the theory and propose *explicit constructions for square polynomial systems* which show 'generic' behavior with respect to their degrees or their monomial supports (Algorithms 4.1 and 5.3). These constructions are strongly related to Macaulay and toric resultant constructions. Just like in these constructions, exploiting the polyhedral structure of the system instead of only considering the degrees of the equations gives a significant reduction of the sizes of the matrices involved in our algorithms.
The systems encountered in applications are often 'non-generic': the number of isolated solutions may be much smaller than the expected number for a system with the same degree or support. Enlarging our solution space to projective space or a more general compact toric variety X, we can present constructions which allow isolated solutions 'at infinity'. The methods rely on a homogeneous interpretation of the theory of truncated normal forms. The 'normal forms' in this context work in the (multi-)graded homogeneous coordinate ring or *Cox ring* of X. We call them *homogeneous normal forms* and show how they lead to algorithms which can deal with solutions at or 'near' infinity (i.e. with large coordinates) in a robust way and which can help to understand the solution count in the torus for certain families of systems. For this, we prove a toric version of the classical eigenvalue-eigenvector theorems and prove new regularity results for homogeneous ideals in the Cox ring, defining finitely many points on X.

Perhaps the most important reliability issue for homotopy continuation methods is the possibility of *path jumping*, which happens when a numerical path tracker jumps 'too far off' the path that is currently being tracked, onto a different solution path. This is a typical way of how solutions are lost during the path tracking. To address this issue, we develop an adaptive stepsize algorithm that uses Padé approximants in the predictor to detect 'difficult' regions along the path. It detects where there is danger for path jumping and adjusts the discretization step of the path accordingly by using a new heuristic. The resulting algorithm can reliably solve challenging problems where other implementations fail.

1.5 Outline

To conclude this chapter, we give an overview of the contents of this thesis by summarizing the subject and goal of each of the next chapters.

In Chapter 2 we give an overview of some basic concepts from algebraic geometry and we fix our notation for varieties, rings and ideals. We have included examples which are instructive for later chapters. The goal of the chapter is to recall important concepts such as the correspondence between varieties and their coordinate rings, the definition of projective space and its standard affine open covering, homogeneous coordinate rings of projective varieties and the gluing construction, which play a prominent role in this thesis.

Chapter 3 consists of four sections, of which the first two recall some specific properties of *zero-dimensional* varieties in affine and projective space and the last two describe some classical methods for *computing* zero-dimensional varieties. The main goal of the first part of the chapter is to state two versions of the *eigenvalue-eigenvector theorem* for isolated root finding and to describe *generic properties* of systems of equations, introducing Bézout's theorem as an important example. The second part of the chapter focuses on *how these results are used* by Gröbner basis, border basis and resultant algorithms for solving equations. These methods have strong connections to the algorithms proposed in this thesis.

Chapter 4 introduces *truncated normal forms* (TNFs) and algorithms based on this framework for solving square polynomial systems. Different choices of representations for the quotient ring are discussed together with several adaptations and improvements of the proposed algorithms. The last section introduces *homogeneous normal forms* (HNFs) for solving square systems in projective space. Several numerical experiments illustrate the effectiveness of the proposed methods. The chapter is strongly based on the papers [TVB18, TMVB18, MTVB19].

In Chapter 5 we show how TNFs and HNFs can be used to solve more general families of polynomial systems. More specifically, we consider systems that are called *sparse* in the literature, referring to the fact that not all monomials up to a certain degree occur in the equations. Taking the polyhedral structure of the equations into account leads to smaller matrices than those of the constructions in Chapter 4. In order to use HNFs in this setting, we work in the Cox ring of a compact toric variety which is a natural solution space for our polyhedral system. We generalize the homogeneous version of the *eigenvalue-eigenvector theorem* to use it in this setting and answer some questions regarding the *regularity* of a homogeneous ideal in the Cox ring. The chapter is based on [TMVB18, Tel20, BT20a].

Chapter 6 is fairly independent of Chapters 3-5 since it deals with a different type of methods for solving polynomial systems. It discusses *homotopy continuation algorithms*. We recall the definition of Padé approximants, discuss some of their properties in the context of homotopy continuation and propose a new numerical path tracking algorithm. In several numerical experiments, this algorithm proves to be significantly more robust with respect to the issue of *path jumping* than existing implementations. This chapter is based on [TVBV19].

The text is supported by a total of five appendices which contain some supplementary material. Appendix A contains a summary of definitions and results from *commutative algebra* which are relevant to the text. Appendix B gives an overview of the used methods and concepts from *numerical linear algebra*. Appendix C motivates and defines the way in which we measure the *error* of computed approximate solutions to a system. Appendix D discusses objects and results from *polyhedral geometry*. Finally, Appendix E contains a crash course in basic *toric geometry*.

For the reader's convenience, we have summarized the most important dependencies between the different parts of the text in the table below.

	Section …	depends on …
Chapter 2	2.1	Appendix A
	2.2	Appendix A, Section 2.1
	2.3	Sections 2.1 and 2.2
Chapter 3	3.1	Appendix A, Section 2.1
	3.2	Appendix A, Section 2.2
	3.3	Section 3.1
	3.4	Section 3.2
Chapter 4	4.1	Sections 3.1 and 3.3
	4.2	Appendix A, Sections 3.1 and 3.3
	4.3	Appendices B and C, Sections 4.2 and 3.4
	4.4	Section 4.3
	4.5	Appendix C, Sections 3.2, 3.4 and 4.2
Chapter 5	5.1	Appendices A and D, Sections 2.1 and 3.1
	5.2	Sections 2.1, 3.4, 5.1
	5.3	Appendix C, Sections 5.1 and 5.2
	5.4	Appendix E, Sections 2.3 and 5.1
	5.5	Appendices C and E, Sections 3.2, 3.4, 5.1, 5.3
Chapter 6	6.1	Section 2.1
	6.2	Section 6.1
	6.3	Section 6.2
	6.4	Appendix B, Sections 6.3 and 6.2
	6.5	Appendix C, Section 6.4
Appendix E	E.1	Appendices A and D, Section 2.1
	E.2	Appendices A and D, Sections 2.2 and 2.3

Chapter 2

Basic algebraic geometry

Algebraic geometry is the study of geometric objects described by algebraic equations. These objects are called *algebraic varieties*. The goal of this chapter is to introduce some basic concepts from algebraic geometry on which the methods for system solving proposed in this thesis are built. We limit ourselves to the concepts that are instructive for the rest of the material in this thesis.

Many of the powerful results in modern algebraic geometry have been made possible by the rigorous algebraic foundations laid out by pioneers such as David Hilbert, Emmy Noether, Jean-Pierre Serre, Bartel Leendert van der Waerden, André Weil, Oscar Zariski and the high level of abstraction in the works of Alexander Grothendieck. However, it is this same level of abstraction that has given the subject the reputation of being rather unaccessible for outsiders. In order to appreciate the field to the fullest, it is crucial to start with the right book. Which book that is depends, of course, on the reader's background. An excellent introduction for readers with an engineering or applied mathematics background is [CLO13], and so is the follow-up book [CLO06]. Other gentle treatments can be found in [SKKT04, SR94]. The book of Hartshorne [Har77] is a standard, more advanced reference. Other advanced and complete treatments can be found in [Mum96, Eis13, Vak17, Cut18], and [Har13] is an excellent source of examples.

Just like differentiable manifolds locally look like open subsets of Euclidean space, algebraic varieties locally look like *affine varieties*. These can be viewed as the building blocks of algebraic varieties, and they are a natural starting point for this chapter. We will discuss affine varieties in Section 2.1. After that, we will introduce projective and quasi-projective varieties in Section 2.2. Finally, we briefly describe how affine varieties can be glued together to obtain more general, abstract varieties in Section 2.3. This gluing construction gives us a good way to think about toric varieties, which will play an important role in later chapters.

2.1 Affine varieties

Our starting point is the *n-dimensional complex affine space* $\mathbb{C}^n$. As a set, $\mathbb{C}^n$ consists of all n-tuples of complex numbers. Some authors write $\mathbb{A}^n$ for this space to emphasize that the origin $0 \in \mathbb{C}^n$ does not play a special role here, as it does when we think of $\mathbb{C}^n$ as a vector space over $\mathbb{C}$. We believe this will not be a source of confusion here and write $\mathbb{C}^n$ to avoid introducing too much notation. Let $R = \mathbb{C}[x_1, \ldots, x_n]$ be the ring of polynomial functions on $\mathbb{C}^n$. As stated in the introduction, if n is small $(n = 1, 2, 3)$ we will use variable names such as x, y, z to avoid subscripts.

2.1.1 Definition

We are interested in special subsets of $\mathbb{C}^n$, namely the zero sets of polynomials.

Definition 2.1.1 (affine variety)**.** An *affine variety* in $\mathbb{C}^n$ is a subset $Y \subset \mathbb{C}^n$ such that there is a set $\mathcal{P} \subset R$ of polynomials for which

$$Y = \{x \in \mathbb{C}^n \mid f(x) = 0, \forall f \in \mathcal{P}\}.$$

In this case, we denote $Y = V_{\mathbb{C}^n}(\mathcal{P})$ or, for short, $Y = V(\mathcal{P})$ when the ambient affine space is clear from the context. If $\mathcal{P} = \{f_1, \ldots, f_s\}$ is finite[1] we will write $V_{\mathbb{C}^n}(f_1, \ldots, f_s)$ for $V_{\mathbb{C}^n}(\{f_1, \ldots, f_s\})$.

Although we work over the complex numbers, for visualization purposes we often consider the real part $Y \cap \mathbb{R}^n$ of an affine variety, especially when $n = 2, 3$.

Example 2.1.1 (Plane curves)**.** Let $R = \mathbb{C}[x, y]$. *Algebraic plane curves* are affine varieties $Y = V_{\mathbb{C}^2}(\mathcal{P})$ where $\mathcal{P}$ is a singleton $\{f\}$, $f \in R \setminus \mathbb{C}$. A nice class of examples of algebraic curves is given by *Lissajous curves.* These are curves parametrized by $x = \sin(t), y = \sin(a_1 t + a_2)$ with $0 \leq a_2 \leq \pi/2$. Under the assumption that $a_1 \in \mathbb{Q}$, the curve is the zero set of a polynomial in $\mathbb{R}$ intersected with the box $[-1, 1]^2$. These curves have applications, for instance, in polynomial approximation and interpolation [BCDM+06]. An example is shown in Figure 2.1. △

Example 2.1.2 (Algebraic surfaces)**.** Let $R = \mathbb{C}[x, y, z]$. If $Y = V_{\mathbb{C}^3}(\mathcal{P})$ where $\mathcal{P}$ is a singleton $\{f\}, f \in R \setminus \mathbb{C}$, then Y is called an *algebraic surface.* As an example we consider the surface given by the equation

$$f = (x^2 - y^2)^2 - 2x^2 - 2y^2 - 16z^2 + 1 = 0.$$

Its real part is shown in Figure 2.2. This surface is obtained from projecting the *double pillow surface*, which lives in a 4-dimensional space, to a 3-dimensional space. The interested reader can find more information in [Sot17, Subsection 3.3]. It is clear from the figure that the surface contains one 'pillow' embracing the origin. The second pillow is in fact embracing a point 'at infinity', which we will make more concrete in Section 2.2. △

[1] In fact, by Hilbert's basis theorem (see Theorem A.1.1), $\mathcal{P}$ can always be assumed to be finite, since $V(\mathcal{P}) = V(\{f_1, \ldots, f_s\})$ for some $f_1, \ldots, f_s \in R$.

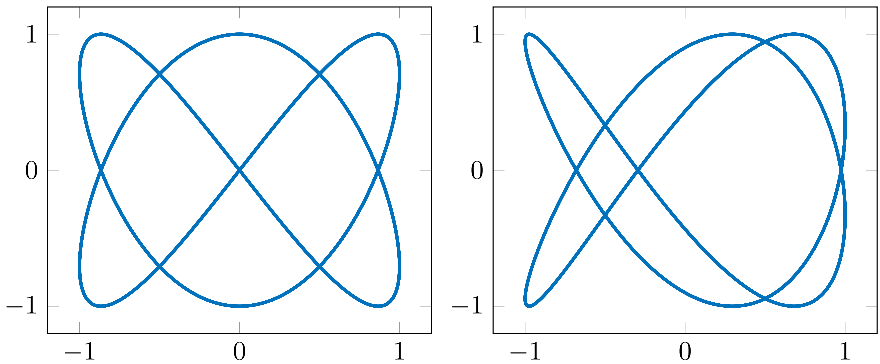


Figure 2.1: Lissajous curves with parameters $a_1 = 3/2, a_2 = 0$ (left) and $a_1 = 3/2, a_2 = \pi/7$ (right). The left curve is equal to the real part of $V(x^2(4x^2-3)^2 + 4y^2(y^2-1))$.

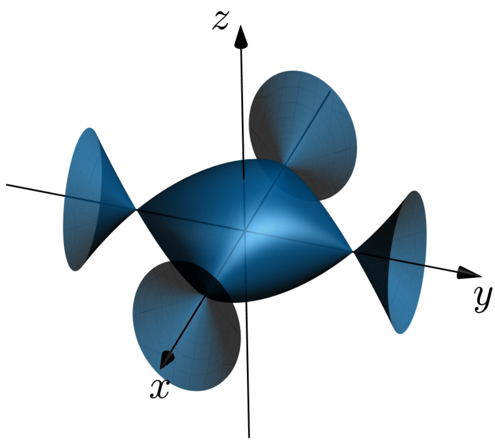


Figure 2.2: The double pillow.

Example 2.1.3 (Space curves)**.** Let $\mathcal{P} = \{y - x^2, z - x^3\} \subset R = \mathbb{C}[x, y, z]$. The affine variety $V_{\mathbb{C}^3}(\mathcal{P})$ is the intersection of the algebraic surfaces $V_{\mathbb{C}^3}(y - x^2)$ and $V_{\mathbb{C}^3}(z - x^3)$. This is a standard example of an *algebraic space curve* (i.e., an algebraic curve in 3-space) called the *twisted cubic*. It is the image of the map $\phi : \mathbb{C} \to \mathbb{C}^3$ defined by $\phi(t) = (t, t^2, t^3)$. This is illustrated in Figure 2.3. $\triangle$

Example 2.1.4. Note that $\mathbb{C}^n = V(0)$ is itself an affine variety, and so is each point $p = (a_1, \ldots, a_n) \in \mathbb{C}^n$, as $p = V(x_1 - a_1, \ldots, x_n - a_n)$. Also the empty set $\varnothing$ is an

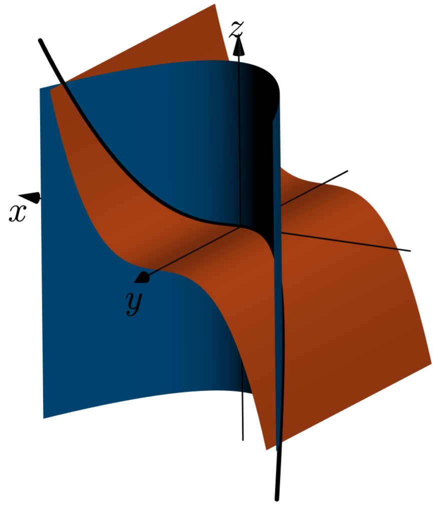


Figure 2.3: The twisted cubic.

affine variety, by $\varnothing = V(1)$. △

2.1.2 Affine varieties as topological spaces

Definition 2.1.1 defines affine varieties as sets. In this subsection we will define them as topological spaces (that is, we will specify which subsets are *closed* and which subsets are *open* in an affine variety Y). One way to do this is by considering the classical topology on $\mathbb{C}^n$ and the induced topology on affine varieties, which are among the closed subsets of $\mathbb{C}^n$ (by continuity of polynomial maps). However, in algebraic geometry we mostly work with a different topology on $\mathbb{C}^n$, called the *Zariski topology*.

Definition 2.1.2 (Zariski topology on $\mathbb{C}^n$)**.** The *Zariski topology* on $\mathbb{C}^n$ is the topology where the closed subsets are the affine varieties.

One can check that affine varieties satisfy the axioms for closed sets in a topology: both $\mathbb{C}^n$ and $\varnothing$ are closed by Example 2.1.4, intersections of affine varieties are affine varieties and finite unions of affine varieties are affine varieties.

Definition 2.1.3 (Zariski topology on an affine variety)**.** Let $Y \subset \mathbb{C}^n$ be an affine variety. The *Zariski topology* on Y is the subspace topology induced by the Zariski topology on $\mathbb{C}^n$.

This means that the closed subsets of Y are the intersections of Y with closed subsets of $\mathbb{C}^n$, which are affine varieties. Closed subsets of Y are also called *subvarieties* of

Y. The *Zariski closure* $\overline{Y}$ of a subset $Y \subset \mathbb{C}^n$ is the smallest Zariski closed subset containing Y.

Example 2.1.5. The only closed subsets of $\mathbb{C}$ are $\mathbb{C}, \varnothing$ and finite subsets. The set $\{(x, y) \in \mathbb{C}^2 \mid |x| \leq 1, |y| \leq 1\}$ is closed in the classical topology, but it is neither open nor closed in the Zariski topology. In fact, its Zariski closure is $\mathbb{C}^2$. The same is true for the set $\{(x, y) \in \mathbb{C}^2 \mid y = \exp(x)\}$. △

Definition 2.1.4 (Reducibility)**.** An affine variety Y is called *reducible* if it can be written as a union $Y = Y_1 \cup Y_2$ with Y_1 and Y_2 proper closed subsets. A variety that is not reducible is called *irreducible*.

2.1.3 The Nullstellensatz

It is a simple observation that $V(\mathcal{P}) = V(I)$, where $I = \langle \mathcal{P} \rangle = \{\sum_i g_i f_i \mid g_i \in R, f_i \in \mathcal{P}\}$ is the ideal generated by the elements in $\mathcal{P}$. For some basic properties and definitions related to ideals, we refer the reader to Appendix A. By Hilbert's basis theorem (see Theorem A.1.1) we can always find a finite set $\{f_1, \ldots, f_s\} \subset \mathcal{P} \subset R$ such that

$$I = \langle f_1, \ldots, f_s \rangle = \{g_1 f_1 + \cdots + g_s f_s \mid g_i \in R\}.$$

Given an ideal $I \subset R$, the operator $V(\cdot)$ gives an affine variety $Y \subset \mathbb{C}^n$. Going the other way around, one could start from a subset $Y \subset \mathbb{C}^n$ and define its *vanishing ideal*

$$I(Y) = \{f \in R \mid f(x) = 0, \forall x \in Y\} \subset R.$$

It is clear that $V(I(Y)) = \overline{Y}$ is the Zariski closure of Y in $\mathbb{C}^n$. In particular, if Y is an affine variety, then $V(I(Y)) = Y$. A natural question to ask is whether $I(V(I)) = I$? Although it is not hard to show that $I \subset I(V(I))$, a simple counterexample shows that the other inclusion does not hold in general.

Example 2.1.6. Let $I = \langle x^2 \rangle \subset \mathbb{C}[x]$. Then $V(I) = \{0\}$ and $I(V(I)) = \langle x \rangle \neq I$. △

Example 2.1.6 gives us an intuition about what can go wrong for the other inclusion. The ideal $\langle x^2 \rangle$ consists of all polynomials with a root of multiplicity at least 2 at the origin. The operator $V(\cdot)$ does not 'see' the multiplicity: for a polynomial to be in the ideal $I(V(I))$, it need only vanish at $x = 0$. A celebrated result by David Hilbert tells us that $I = I(V(I))$ for a subclass of ideals in R.

Theorem 2.1.1 (Hilbert's Nullstellensatz)**.** *Let $I \subset R = \mathbb{C}[x_1, \ldots, x_n]$ be an ideal and let $Y \subset \mathbb{C}^n$ be an affine variety. Then*

$$V(I(Y)) = Y \quad \text{and} \quad I(V(I)) = \sqrt{I},$$

where $\sqrt{I} = \{f \in R \mid f^m \in I \text{ for some } m \in \mathbb{N}\}$ *is the radical of I.*

Proof. Proofs can be found, for example, in [CLO13, Chapter 4], [Eis13, Chapter 4] or [Rei95, Chapter 5]. □

Theorem 2.1.1 establishes a nice interplay between algebra and geometry. More specifically, it tells us that there is a one-to-one correspondence between affine varieties in $\mathbb{C}^n$ and radical ideals of R.

2.1.4 Coordinate rings and morphisms

Our goal in this subsection is to establish a one-to-one correspondence between affine varieties and some special commutative rings with identity. As a first step, given an affine variety $Y \subset \mathbb{C}^n$, we want to understand the *polynomial functions* on Y. That is, we want to characterize the set $\mathbb{C}[Y]$ of functions $Y \to \mathbb{C}$ that are the restriction of a polynomial in R. It is clear that this set has a ring structure and there is a surjective ring homomorphism $R \to \mathbb{C}[Y]$ given by 'restriction to Y'. The elements of R that restrict to 0 on Y are exactly the elements in $I(Y)$. This gives a short exact sequence (see Subsection A.2.2)

$$0 \to I(Y) \to R \to \mathbb{C}[Y] \to 0. \tag{2.1.1}$$

By the first isomorphism theorem (Theorem A.2.2) we find that $\mathbb{C}[Y] = R/I(Y)$. The quotient ring $R/I(Y)$ is called the *coordinate ring* of Y. It is a finitely generated $\mathbb{C}$-algebra with no nilpotents[2] (by the fact that $I(Y)$ is radical), see Section A.1 for definitions.

Example 2.1.7 (Some trivial coordinate rings)**.** Note that $\mathbb{C}[\mathbb{C}^n] = R$ and $\mathbb{C}[\varnothing] = \{0\}$. △

Example 2.1.8 (Coordinate rings of points)**.** If $Y = \{p\}$ is a single point $p = (a_1, \ldots, a_n) \in \mathbb{C}^n$, then $I(Y) = \langle x_1 - a_1, \ldots, x_n - a_n \rangle$ is a maximal ideal of R. In fact, all maximal ideals of R are of this form [CLO13, Chapter 4, §5, Theorem 11]. In this case $\mathbb{C}[Y] = \mathbb{C}$ and the map $R \to \mathbb{C}[Y]$ in (2.1.1) sends f to $f(p)$. △

Example 2.1.9 (Irreducible varieties)**.** The geometric notion of an affine variety being *irreducible* (which means it cannot be written as the union of two strict subvarieties) corresponds to the equivalent algebraic notions of the ideal $I(Y)$ being prime and the ring $R/I(Y)$ being an integral domain [CLO13, Chapter 4, §5, Proposition 3]. △

Definition 2.1.5 (Morphisms of varieties)**.** Let $Y \subset \mathbb{C}^n$ and $Y' \subset \mathbb{C}^m$ be affine varieties. A *morphism* between Y and Y' is a map $\phi : Y \to Y'$ given by polynomials:

$$\phi(x) = (f_1(x), \ldots, f_m(x)), \quad f_i \in \mathbb{C}[x_1, \ldots, x_n].$$

Example 2.1.10 (Morphisms)**.** The parametrization $t \mapsto (t, t^2, t^3)$ of the twisted cubic in Example 2.1.3 is a morphism between $\mathbb{C}$ and $\mathbb{C}^3$, and between $\mathbb{C}$ and the twisted cubic. The coordinate ring of an affine variety Y is the ring of morphisms $Y \to \mathbb{C}$. △

[2] We say that a ring *has no nilpotents* or is *nilpotent-free* if its only nilpotent element is 0.

Note that the composition of two morphisms is again a morphism. A morphism $\phi : Y \to Y'$ gives a $\mathbb{C}$-algebra homomorphism $\phi^* : \mathbb{C}[Y'] \to \mathbb{C}[Y]$ by composing $f \in \mathbb{C}[Y']$ with ϕ: $\phi^*(f) = f \circ \phi$. The map ϕ^* is called the *pullback map* or simply the *pullback* of ϕ.

Definition 2.1.6 (Isomorphism)**.** A morphism $\phi : Y \to Y'$ is an *isomorphism* if the pullback $\phi^* : \mathbb{C}[Y'] \to \mathbb{C}[Y]$ is an isomorphism of $\mathbb{C}$-algebras. Two affine varieties $Y \subset \mathbb{C}^n$, $Y' \subset \mathbb{C}^m$ are called *isomorphic* if there exists an isomorphism $\phi : Y \to Y'$.

One can check that Y and Y' are isomorphic if and only if there exists morphisms $\phi : Y \to Y'$ and $\phi' : Y' \to Y$ with $\phi \circ \phi' = \mathrm{id}_{Y'}$ and $\phi' \circ \phi = \mathrm{id}_Y$ [CLO13, Chapter 5, §4, Theorem 9]. If Y and Y' are isomorphic, we write $Y \simeq Y'$ and sometimes, with a slight abuse of notation, $Y = Y'$.

Example 2.1.11. Let $Y \subset \mathbb{C}^3$ be the twisted cubic as in Example 2.1.3. The pullback of the map $\phi : \mathbb{C} \to Y$ given by $\phi(t) = (t, t^2, t^3)$ is the map ϕ^* that sends $f + \langle y - x^2, z - x^3 \rangle \in \mathbb{C}[x, y, z]/\langle y - x^2, z - x^3 \rangle$ to $f(t, t^2, t^3) \in \mathbb{C}[t]$. It is clearly surjective because $t = \phi^*(x + \langle y - x^2, z - x^3 \rangle)$. It is also injective because if $f(t, t^2, t^3) = 0$, then f vanishes at every point of Y, hence $f \in \langle y - x^2, z - x^3 \rangle$. It follows that Y is isomorphic to $\mathbb{C}$. △

Example 2.1.11 tells us that the twisted cubic in $\mathbb{C}^3$ and the affine line $\mathbb{C}$ are basically the same affine varieties, they are just embedded in a different ambient space. The intrinsic reason for this is that the algebras of polynomial functions on the twisted cubic and on $\mathbb{C}$ are the same. That is,

$$\mathbb{C}[x, y, z]/\langle y - x^2, z - x^3 \rangle \simeq \mathbb{C}[t].$$

The different embeddings come from a *choice of representation* of the $\mathbb{C}$-algebra $\mathbb{C}[t]$ as an image of a polynomial ring: it is the image of $\mathbb{C}[t]$ under the identity morphism but it is also the image of $\mathbb{C}[x, y, z]$ under the map $f \mapsto f(t, t^2, t^3)$ with kernel $\langle y - x^2, z - x^3 \rangle$. This hints at a more general procedure for associating an affine variety to a finitely generated $\mathbb{C}$-algebra A. We first represent A as the image of a polynomial ring: $R \to A \to 0$. Next, we consider the kernel of this map, which is an ideal $I \subset R$, to obtain the affine variety $Y = V(I)$. If A is nilpotent free, then I is radical and by the Nullstellensatz $I(V(I)) = I(Y) = I$. Therefore, $\mathbb{C}[Y] = R/I(Y) = R/I \simeq A$. The following theorem is a consequence of this.

Theorem 2.1.2. *There is a one-to-one correspondence between isomorphism classes of affine varieties and isomorphism classes of finitely generated, nilpotent free $\mathbb{C}$-algebras.*

We have the notation $Y \mapsto \mathbb{C}[Y]$ to make this correspondence explicit. To go in the other direction, we introduce the notation $A \mapsto \mathrm{MaxSpec}(A)$ which associates to a finitely generated, nilpotent free $\mathbb{C}$-algebra A an affine variety by the procedure presented above. The notation $\mathrm{MaxSpec}(A)$ is motivated by the fact that for an affine variety $Y \subset \mathbb{C}^n$, the points in Y are in one-to-one correspondence with maximal ideals in $R/I(Y)$. This was established in Example 2.1.8 in the case where $Y = \mathbb{C}^n$. The general case is described in [CLO13, Chapter 5, §4, Theorem 5].

Morphisms between varieties give homomorphisms between $\mathbb{C}$-algebras going in the opposite direction by considering the pullback morphism. Going the other way around, a $\mathbb{C}$-algebra homomorphism $\phi^* : A' \to A$ with A, A' finitely generated and nilpotent free gives a morphism $\phi : \mathrm{MaxSpec}(A) \to \mathrm{MaxSpec}(A')$ defined as follows. A point $p \in \mathrm{MaxSpec}(A)$ corresponds to a maximal ideal $I(p)$ of A. The inverse image $(\phi^*)^{-1}(I(p))$ is again a maximal ideal in A' (see, e.g., [SKKT04, Section 2.6]) and corresponds to a point $p' \in \mathrm{MaxSpec}(A')$. We set $\psi(p) = p'$. One can check that ψ is a morphism and that $\psi^* = \phi^*$. For readers familiar with category theory, we remark that this construction makes the correspondence in Theorem 2.1.2 functorial: the functor $Y \mapsto \mathbb{C}[Y]$ establishes a contravariant equivalence of categories between affine varieties and finitely generated nilpotent free $\mathbb{C}$-algebras [Har77, Chapter I, Corollary 3.8].

The machinery introduced in this chapter allows us to state a more general version of the Nullstellensatz which identifies subvarieties of an affine variety Y with radical ideals in its coordinate ring. For a subvariety $Y' \subset Y = \mathrm{MaxSpec}(A)$ and an ideal $I \subset A = \mathbb{C}[Y]$ we define the vanishing ideal of Y' and subvariety of I as

$$I_A(Y') = \{f \in A \mid f(p) = 0, \forall p \in Y'\}, \quad V_Y(I) = \{p \in Y \mid f(p) = 0, \forall f \in I\}$$

respectively. In the following theorem we recover Theorem 2.1.1 when $A = R$.

Theorem 2.1.3. *Let A be a finitely generated nilpotent free $\mathbb{C}$-algebra and let $Y =$ $\mathrm{MaxSpec}(A)$ be the corresponding affine variety. Let $I \subset A$ be an ideal and let $Y' \subset Y$ be a subvariety. Then*

$$V_Y(I_A(Y')) = Y' \quad \text{and} \quad I_A(V_Y(I)) = \sqrt{I},$$

where $\sqrt{I} = \{f \in A \mid f^m \in I \text{ for some } m \in \mathbb{N}\}$ is the radical of I.

Proof. See [CLO13, Chapter 5, §4, Theorem 5]. □

Example 2.1.12 (Localization at f)**.** Let A be a finitely generated nilpotent free $\mathbb{C}$-algebra and $Y = \mathrm{MaxSpec}(A)$. Let A_f be the localization of A at $f \in A, f \neq 0$ (see Subsection A.1.4). Note that A_f is finitely generated and nilpotent free. When A is an integral domain with field of fractions $K(A)$, then the canonical map $A \to A_f$ is injective and A_f is given by

$$A_f = \left\{ \frac{g}{f^\ell} \in K(A) \mid g \in A, \ell \in \mathbb{N} \right\},$$

see for instance [CLS11, Exercise 1.0.3]. We will now describe the corresponding affine variety $Y_f = \mathrm{MaxSpec}(A_f)$. The maximal ideals of A_f are the maximal ideals of A not containing f [AM69, Chapter 3]. Since points of Y_f are maximal ideals of A_f, the points of Y_f are the points $p \in Y$ such that $f(p) \neq 0$. This shows, somewhat surprisingly, that the open subset of Y consisting of the complement of $V_Y(f)$ can be given the structure of an affine variety. A standard example that clarifies this is the case where $Y = \mathbb{C}$ is the affine line and $f = t \in A = \mathbb{C}[t]$. Here $Y_f = \mathbb{C} \setminus \{0\} = \mathbb{C}^*$ and $A_f = \mathbb{C}[t]_t \simeq \mathbb{C}[x,y]/\langle xy-1\rangle$. This isomorphism of algebras is given explicitly by $\phi^* : \mathbb{C}[x,y]/\langle xy-1\rangle \to \mathbb{C}[t]_t$ defined as

$$\phi^*(f + \langle xy-1\rangle) \mapsto f(t, t^{-1}).$$

This corresponds to the morphism $\phi : \mathbb{C}^* \to V_{\mathbb{C}^2}(xy-1)$ given by $\phi(t) = (t, t^{-1})$. This morphism is illustrated in Figure 2.4.

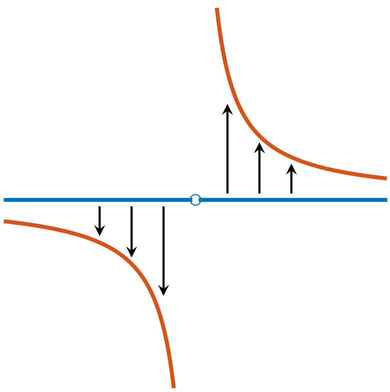


Figure 2.4: Illustration of the morphism $\phi : \mathbb{C}^* \to V_{\mathbb{C}^2}(xy-1)$ from Example 2.1.12.

The affine variety $\mathbb{C}^*$ is an example of an *algebraic torus*: the n-dimensional algebraic torus is the affine variety $(\mathbb{C} \setminus \{0\})^n = (\mathbb{C}^*)^n = \mathrm{MaxSpec}(\mathbb{C}[x_1, \ldots, x_n]_{x_1 \cdots x_n})$. Algebraic tori will play an important role in later chapters of this thesis. Subvarieties of algebraic tori are defined by elements of $\mathbb{C}[x_1, \ldots, x_n]_{x_1 \cdots x_n} = \mathbb{C}[x_1, x_1^{-1}, \ldots, x_n, x_n^{-1}]$ which are called *Laurent polynomials*. $\triangle$

2.1.5 Dimension

Although the geometric concept of dimension is very intuitive, formal definitions of dimension often are not. For completeness, we will include some formal, equivalent definitions of dimension in this subsection. The equivalence of these definitions establishes nicely the interplay between algebra and geometry. More elaborate treatments can be found in [CLO13, Chapter 9], [AM69, Chapter 11], [SR94, Chapter 1, Section 6], [Cut18, Chapter 2, Section 2.4], [Eis13, Chapter 2]. We should mention that, since we are working over the complex numbers, we always think of *complex dimension*. For instance, $\mathbb{C}$ has complex dimension one, but real dimension 2. Therefore, we will think of $\mathbb{C}$ as the *affine line* (a terminology that has been used a few times above) as opposed to the *complex plane*.

A first observation is that a reducible affine variety may have components of different dimension. For instance, the affine variety $Y = V_{\mathbb{C}^3}(xy, xz)$ is a union of the yz-plane where $x = 0$ and the x-axis defined by $y = z = 0$. We will define dimension for irreducible affine varieties and say that the dimension of an affine variety Y is the maximum among the dimensions of its irreducible components (which are always finite in number, see [Har77, Chapter I, Proposition 1.5]).

Definition 2.1.7 (Dimension of an irreducible affine variety)**.** Let $Y \subset \mathbb{C}^n$ be an irreducible affine variety. The *dimension* of Y, denoted $\dim Y$, is the length k of the longest possible chain of strict inclusions

$$Y_0 \subsetneq Y_1 \subsetneq \cdots \subsetneq Y_k = Y$$

where Y_i are irreducible subvarieties.

An affine variety is called *pure dimensional* if all its irreducible components have the same dimension. Pure dimensional affine varieties of dimension 1 are called (affine) *curves*, those of dimension 2 are called (affine) *surfaces* and those of dimension n are called (affine) *n-folds*. When embedded in an affine space $\mathbb{C}^n$ of dimension n, an affine variety Y has *codimension* $n - \dim Y$ and affine varieties of codimension 1 are called (affine) hypersurfaces. More generally, for a subvariety $Y' \subset Y$ we define $\operatorname{codim}_Y Y' = \dim Y - \dim Y'$.

Example 2.1.13. Consider the affine varieties

$$\begin{aligned}
Y_2 &= V_{\mathbb{C}^3}(x^2 + y^2 + z^2 - 1),\\
Y_1 &= V_{Y_2}(x^2 + y^2 - x + \langle x^2 + y^2 + z^2 - 1\rangle) = V_{\mathbb{C}^3}(x^2 + y^2 + z^2 - 1, x^2 + y^2 - x),\\
Y_0 &= V_{Y_1}(z - 1 + \langle x^2 + y^2 + z^2 - 1, x^2 + y^2 - x\rangle)\\
&= V_{\mathbb{C}^3}(x^2 + y^2 + z^2 - 1, x^2 + y^2 - x, z - 1).
\end{aligned}$$

This gives $Y_0 \subsetneq Y_1 \subsetneq Y_2$, which is a chain of maximal length as in Definition 2.1.7. This shows that the sphere has dimension 2 in $\mathbb{C}^3$. It also shows that $\dim Y_1 = 1$ and $\dim Y_0 = 0$. The (real part of the) curve Y_1 in this example is known as *Viviani's curve.* The situation is illustrated in Figure 2.5. △

The following theorem establishes the equivalence of the geometric (topological) Definition 2.1.7 with an algebraic definition of dimension. It shows, for instance, that the dimension is independent of the choice of embedding.

Theorem 2.1.4. *Let Y be an irreducible affine variety with coordinate ring $\mathbb{C}[Y]$. The following natural numbers are all equal to* $\dim Y$*:*

1. *the Krull dimension of $\mathbb{C}[Y]$ (see Subsection A.1.3),*
2. *the transcendence degree of the quotient field $\mathbb{C}(Y)$ of $\mathbb{C}[Y]$ over $\mathbb{C}$,*
3. *the maximal number of elements of $\mathbb{C}[Y]$ that are algebraically independent over $\mathbb{C}$,*

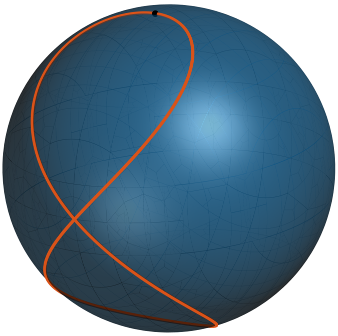


Figure 2.5: Illustration of the affine varieties Y_2 (blue surface), Y_1 (orange curve) and Y_0 (black point) from Example 2.1.13.

4. *the degree of the* affine Hilbert polynomial *as defined in [CLO13, Chapter 9, §3].*

Proof. The first statement follows directly from the fact that prime ideals of $\mathbb{C}[Y]$ are irreducible subvarieties of Y, and the correspondence is inclusion reversing [CLO13, Chapter 5, §4, Theorem 5]. For the equivalence between the first and the second definition see [Har77, Chapter 1, Section 1, Proposition 1.7 and Theorem 1.8A], [AM69, Chapter 11]. The equivalence of the second, third and fourth definition is established in [CLO13, Chapter 9, §3 and §5]. □

2.1.6 Affine schemes

For any ideal $I \subset R$, we can consider the affine variety $V(I)$. However, if I is not radical, some information is lost in making this association. There are many more ideals than affine varieties. Looking more closely, two ideals $I \neq I' \subset R$ with $V(I) = V(I')$ determine objects with different geometric behavior. Here are two examples.

Example 2.1.14. Consider the ideals $\langle f \rangle = \langle x^2(x-1) \rangle \subset \mathbb{C}[x]$ and $\langle g \rangle = \langle x(x-1)^2 \rangle \subset \mathbb{C}[x]$. It is clear that $V(f) = V(g) = \{0,1\} \subset \mathbb{C}$. However, f has the point $x = 0$ as a *double* root, since $f(0) = \frac{\partial f}{\partial x}(0) = 0$, whereas $\frac{\partial g}{\partial x}(0) = 1$. Slightly perturbing the polynomial f would result in a variety consisting of two points near $x = 0$ (although they may be far away from $x = 0$ relative to the 'size' of the perturbation) and a point near $x = 1$. On the other hand, slightly perturbing g would result in the opposite scenario. The situation is illustrated in Figure 2.6. △

Example 2.1.15. Consider the parametrized ideal $I(t) = \langle (x-t)(x+t) \rangle \subset \mathbb{C}[x]$. For $t \neq 0$, $V(I(t)) = \{t, -t\}$ consists of two points in $\mathbb{C}$ and $\mathbb{C}[x]/I(t)$ has no nilpotents. As $t \to 0$, the two points collide and $I(0) = \langle x^2 \rangle$ is the ideal from Example 2.1.6 and

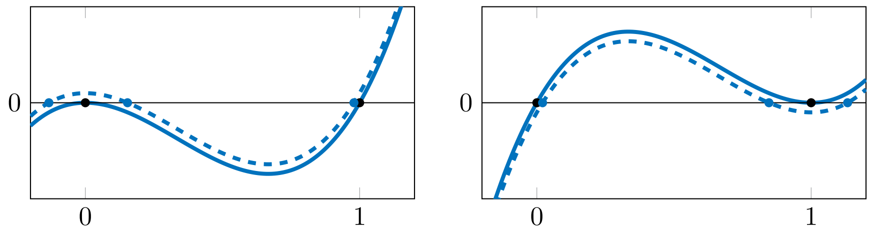


Figure 2.6: Illustration of $V(f)$ (left, black dots) and $V(g)$ (right, black dots) from Example 2.1.14 and the varieties (blue dots) corresponding to perturbed polynomials (dashed curves).

$x + I(0)$ is a nilpotent element of $\mathbb{C}[x]/I(0)$. This illustrates that finitely generated $\mathbb{C}$-algebras with nilpotent elements may arise as a limit of a sequence of finitely generated, nilpotent free $\mathbb{C}$-algebras. △

From Example 2.1.15 one can imagine more complicated situations such as points in higher dimensional affine spaces moving together resulting in *multiple points* or *fat points* (i.e. points with *multiplicity* > 1), curves moving together resulting in *multiple curves*, points moving into curves resulting in *embedded points* inside these curves, embedded curves in surfaces, and so on. In order to take these limiting situations into account, it is clear that we have to extend our correspondence between affine varieties and finitely generated, nilpotent free $\mathbb{C}$-algebras to larger classes of objects (i.e. larger categories). For instance, we want to allow nilpotent elements in our algebras. A powerful extension of this correspondence is given by the theory of *affine schemes*. Affine schemes form a category of geometric objects of which 'affine varieties' can be considered a subcategory. The equivalent category on the algebraic side consists of commutative rings with identity, containing the finitely generated, nilpotent free $\mathbb{C}$-algebras. The power and extent of this generalization can be seen from how small the subset of finitely generated, nilpotent free $\mathbb{C}$-algebras is in the commutative rings with identity.

The theory of schemes uses high levels of abstraction and defining them formally would require notions of sheaf theory, which would take us too far. Affine schemes will only make a modest appearance in this text: we will only consider finitely generated $\mathbb{C}$-algebras but we will sometimes allow nilpotents. Such schemes are sometimes called *affine* $\mathbb{C}$*-schemes*, and they are in one-to-one correspondence (up to isomorphism) with all rings of the form R/I where R is a polynomial ring over $\mathbb{C}$ and $I \subset R$ is any ideal of R. Among affine $\mathbb{C}$-schemes there are the affine varieties, whose algebras are nilpotent free. Affine schemes corresponding to nilpotent free rings are called *reduced*. We will also mostly be interested in *zero-dimensional affine* $\mathbb{C}$*-schemes*. Fortunately, these schemes have a very explicit and relatively simple description, which will be given in Subsection 3.1.3. For more information about schemes, we refer to [EH06] for a gentle introduction with many examples or [Har77, Chapter 2] for a denser treatment.

2.2 Projective varieties

The *projective n-space* $\mathbb{P}^n$ is defined as the set of all lines through the origin in $\mathbb{C}^{n+1}$. If $x_0, \ldots, x_n$ are coordinates on $\mathbb{C}^{n+1}$,

$$\mathbb{P}^n = (\mathbb{C}^{n+1} \setminus \{0\}) / \sim$$

where the quotient is by the equivalence relation

$$(x_0, \ldots, x_n) \sim (x_0', \ldots, x_n') \Leftrightarrow (x_0', \ldots, x_n') = (\lambda x_1, \ldots, \lambda x_n) \text{ for some } \lambda \in \mathbb{C}^*. \quad (2.2.1)$$

Points in $\mathbb{P}^n$ are denoted by $x = (x_0 : \cdots : x_n)$, where $(x_0 : \cdots : x_n) = (\lambda x_0 : \cdots : \lambda x_n)$ for $\lambda \in \mathbb{C}^*$. We will use the notation $S = \mathbb{C}[x_0, \ldots, x_n]$ for the coordinate ring of $\mathbb{C}^{n+1}$.

2.2.1 Definition

For a monomial $x^a = x_0^{a_0} \cdots x_n^{a_n} \in S$ with $a = (a_0, \ldots, a_n) \in \mathbb{N}^{n+1}$, we define its *degree* to be $\deg(x^a) = |a| = a_0 + \cdots + a_n$. We will consider S as a $\mathbb{Z}$-graded ring (see Subsection A.2.4). The $\mathbb{C}$-vector subspaces of the polynomial ring S spanned by the monomials of a fixed degree are called the *graded pieces* of S. They are denoted by

$$S_d = \bigoplus_{|a|=d} \mathbb{C} \cdot x^a, \quad d \in \mathbb{Z}_{\geq 0} \quad \text{and} \quad S_d = \{0\}, \quad d \in \mathbb{Z}_{<0},$$

where a ranges over $\mathbb{N}^{n+1}$. The decomposition

$$S = \bigoplus_{d \in \mathbb{N}} S_d$$

of S into its graded pieces coarsens the decomposition $S = \bigoplus_{a \in \mathbb{N}^{n+1}} \mathbb{C} \cdot x^a$ corresponding to the monomial basis. Note that for all $d, e \in \mathbb{N}$, $S_e \cdot S_d = \{fg \mid f \in S_e, g \in S_d\} \subset S_{d+e}$.

Definition 2.2.1 (Homogeneous polynomial)**.** A polynomial $f \in S$ is called *homogeneous* if it is contained in a graded piece of S, that is, if $f \in S_d$ for some $d \in \mathbb{Z}$. The *degree* of a nonzero homogeneous polynomial f, denoted $\deg(f)$, is d such that $f \in S_d$. The zero polynomial is homogeneous and its degree is $-\infty$ by convention.

Example 2.2.1. A homogeneous polynomial of degree 1 is called a *linear form*. A homogeneous polynomial of degree 2, 3, 4, 5, 6, ... is called a *quadratic, cubic, quartic, quintic, sextic, ... form*. Homogeneous polynomials in 2, 3, 4, 5, 6, ... variables are called *binary, ternary, quaternary, quinary, senary, ... forms*. For example, a general binary quintic form is given by

$$c_5 x_1^5 + c_4 x_1^4 x_0 + c_3 x_1^3 x_0^2 + c_2 x_1^2 x_0^3 + c_1 x_1 x_0^4 + c_0 x_0^5, \quad c_i \in \mathbb{C}.$$

Often the word ‘form’ is dropped: a *binary quintic* is a binary quintic form. $\triangle$

Just like affine varieties in $\mathbb{C}^n$ were defined as subsets of affine space given by polynomials in R, we will define projective varieties as subsets of $\mathbb{P}^n$ given by polynomials in S. In order to do so we investigate which polynomials have well defined zero sets on $\mathbb{P}^n$. As we saw in Section 2.1, elements of S are polynomial functions on $\mathbb{C}^{n+1}$. Note that for a homogeneous polynomial $f \in S_d$ we have $f(\lambda x) = \lambda^d f(x)$, $x \in \mathbb{C}^{n+1}, \lambda \in \mathbb{C}^*$. Therefore, for an element $f \in S_d$, the set

$$V_{\mathbb{P}^n}(f) = \{(x_0 : \cdots : x_n) \in \mathbb{P}^n \mid f(x_0, \ldots, x_n) = 0\}$$

is well defined. This leads to the following definition.

Definition 2.2.2 (Projective variety)**.** A *projective variety* is a subset $X \subset \mathbb{P}^n$ such that there is a subset $\mathcal{P} \subset S$ of homogeneous polynomials for which

$$X = \{(x_0 : \cdots : x_n) \in \mathbb{P}^n \mid f(x_0, \ldots, x_n) = 0, \forall f \in \mathcal{P}\}.$$

In this case, we denote $X = V_{\mathbb{P}^n}(\mathcal{P})$. If $\mathcal{P} = \{f_1, \ldots, f_s\}$ we will write $V_{\mathbb{P}^n}(f_1, \ldots, f_s) = V_{\mathbb{P}^n}(\{f_1, \ldots, f_s\})$.

Every polynomial $f \in S$ can be decomposed uniquely as

$$f = f_d + f_{d-1} + \cdots + f_0, \quad f_i \in S_i.$$

Therefore $f(\lambda x) = \lambda^d f_d(x) + \lambda^{d-1} f_{d-1}(x) + \cdots + f_0$. We conclude that a polynomial $f \in S$ gives a function

$$f : \mathbb{P}^n \to \mathbb{C} \quad \text{given by} \quad f((x_0 : \cdots : x_n)) = f(x_0, \ldots, x_n)$$

if and only if f is homogeneous and $\deg(f) = 0$. Indeed, homogeneous polynomials of degree $d > 0$ *do not* give functions on $\mathbb{P}^n$, but they do have well defined zero sets. A set of homogeneous elements $\mathcal{P} \subset S$ also defines an affine variety

$$V_{\mathbb{C}^{n+1}}(\mathcal{P}) = \{x \in \mathbb{C}^{n+1} \mid f(x) = 0, \forall f \in \mathcal{P}\}$$

which is called the *affine cone* over $V_{\mathbb{P}^n}(\mathcal{P})$.

Example 2.2.2. The projective space $\mathbb{P}^n$ itself and the empty set $\varnothing \subset \mathbb{P}^n$ are projective varieties. One can easily check that any finite union of projective varieties is again a projective variety, and so is any intersection of projective varieties. △

Example 2.2.3 (Linear subspaces)**.** The image under the quotient by (2.2.1) of a vector subspace of $\mathbb{C}^{n+1}$ is a projective variety, for which $\mathcal{P}$ consists of linear forms. △

2.2.2 Projective varieties as topological spaces

Just like affine varieties, projective varieties are topological spaces where closed sets are subvarieties.

Definition 2.2.3 (Zariski topology on projective varieties)**.** The *Zariski topology* on $\mathbb{P}^n$ is the topology where the closed subsets are projective varieties. The Zariski topology on a projective variety $X \subset \mathbb{P}^n$ is the induced topology on X as a closed subset of $\mathbb{P}^n$.

By Example 2.2.2, projective varieties satisfy the axioms on closed sets. As in the affine case, a projective variety is called *reducible* if it can be written as a union of two proper closed subsets. If a projective variety is not reducible, it is called *irreducible*. We will also be interested in subsets of $\mathbb{P}^n$ that are *almost* projective varieties, but not quite.

Definition 2.2.4 (Quasi-projective variety)**.** A *quasi-projective variety* is an open subset of a projective variety with its induced subspace topology.

2.2.3 Projective Nullstellensatz

A natural question to ask is whether we also have a nice correspondence between radical ideals of S and projective varieties, as in the affine case (see Subsection 2.1.3). A first observation is that ideals of S corresponding to a projective variety X should have a special structure: their elements vanish on the affine cone over X in $\mathbb{C}^{n+1}$.

Definition 2.2.5 (Homogeneous ideal)**.** An ideal $I \subset S$ is called *homogeneous* if it can be generated by homogeneous elements of S. Equivalently, I is homogeneous if and only if for every element $f \in I$ with decomposition $f = f_d + \cdots + f_0$, $f_i \in S_i$, we have $f_i \in I, i = 0, \ldots, d$.

For a homogeneous ideal $I = \langle \mathcal{P} \rangle \subset S$ generated by a set $\mathcal{P}$ of homogeneous polynomials, we set $V_{\mathbb{P}^n}(I) = V_{\mathbb{P}^n}(\mathcal{P})$. Given a projective variety $X \subset \mathbb{P}^n$, we can associate an ideal to it by defining

$$I_S(X) = \{f \in S \mid f(x_0, \ldots, x_n) = 0, \forall (x_0 : \cdots : x_n) \in X\}.$$

Ideals arising in this way are homogeneous (see [CLO13, Chapter 8, §3, Proposition 4]). They are also radical since either $I_S(X) \subset S$ is the vanishing ideal of the affine cone over X or it is the ring S itself.[3] We conclude that radical homogeneous ideals define projective varieties, and projective varieties define radical homogeneous ideals. The question is whether this correspondence is one-to-one. The following observation shows that we should be careful.

Remark 2.2.1. The radical homogeneous ideal $\mathfrak{B} = \langle x_0, \ldots, x_n \rangle$ defines the affine variety $V_{\mathbb{C}^{n+1}}(\mathfrak{B}) = \{0\}$, but $V_{\mathbb{P}^n}(\mathfrak{B}) = \varnothing$. However, also $V_{\mathbb{P}^n}(S) = \varnothing$. $\triangle$

[3]Here's a proof. If X is empty, $I_S(X) = S$. Otherwise every $f = f_d + \cdots + f_0 \in I_S(X)$ is such that $f_i \in I_S(X), \forall I$. In particular $f_0 \in I_S(X)$ and since $X \neq \varnothing$ this implies $f_0 = 0$ and f vanishes at the origin in $\mathbb{C}^{n+1}$.

Theorem 2.2.1 (Projective Nullstellensatz)**.** *Let $I \subset S = \mathbb{C}[x_0, \ldots, x_n]$ be a homogeneous ideal and let $X = V_{\mathbb{P}^n}(I) \subset \mathbb{P}^n$. If $X \neq \varnothing$, we have*

$$V_{\mathbb{P}^n}(I_S(X)) = X \quad \text{and} \quad I_S(V_{\mathbb{P}^n}(I)) = \sqrt{I}.$$

Proof. The first statement follows from $V_{\mathbb{P}^n}(I_S(X)) = \overline{X} = X$ where $\overline{X}$ is the closure of X in $\mathbb{P}^n$ in its Zariski topology. The second statement follows from Theorem 2.1.1 and from the fact that $I_S(X)$ is the vanishing ideal of the affine cone over X (see above). □

Note that the ideal $\mathfrak{B} \subset S$ from Remark 2.2.1 is left out of the correspondence between radical homogeneous ideals and projective varieties in Theorem 2.2.1. Because this ideal has no corresponding closed subset, it is called the *irrelevant ideal* of S.

2.2.4 Homogeneous coordinate rings

For an affine variety $Y \subset \mathbb{C}^n$, we defined its coordinate ring as $\mathbb{C}[Y] = R/I_R(Y)$ where $R = \mathbb{C}[x_1, \ldots, x_n] = \mathbb{C}[\mathbb{C}^n]$. Similarly, for a projective variety X we define the *homogeneous coordinate ring* of X as $\mathbb{C}[X] = S/I_S(X)$. If $X \neq \varnothing$, $\mathbb{C}[X]$ is the ring of polynomial functions on the affine cone over X.

For any homogeneous ideal $I \subset S$, the grading on S induces a grading on I:

$$I = \bigoplus_{d \in \mathbb{Z}} I_d, \quad \text{where} \quad I_d = I \cap S_d.$$

The grading on S also induces a grading on the quotient ring S/I:

$$S/I = \bigoplus_{d \in \mathbb{Z}} (S/I)_d, \quad \text{where} \quad (S/I)_d = S_d/I_d.$$

Therefore the homogeneous coordinate ring $\mathbb{C}[X]$ of X has the natural structure of a graded ring.

Closed subsets of a projective variety X are given by homogeneous ideals of $\mathbb{C}[X]$: for $I = \langle f_1 + I_S(X), \ldots, f_s + I_S(X) \rangle \subset \mathbb{C}[X]$ we define

$$V_X(I) = \{(x_0 : \cdots : x_n) \in X \mid f_i(x_0, \ldots, x_n) = 0, i = 1, \ldots, s\}.$$

Conversely, a closed subset $X' \subset X$ gives a homogeneous ideal

$$I_{\mathbb{C}[X]}(X') = \{f + I_S(X) \in \mathbb{C}[X] \mid f(x_0, \ldots, x_n) = 0, \forall (x_0 : \cdots : x_n) \in X'\}.$$

2.2.5 Affine coverings

In the introduction to this chapter we claimed that varieties locally look like affine varieties. We will make this precise for projective varieties in this subsection. We define the Zariski open subsets

$$U_i = \{(x_0 : \cdots : x_n) \in \mathbb{P}^n \mid x_i \neq 0\}, i = 0, \ldots, n$$

of $\mathbb{P}^n$. These correspond to the Zariski open subsets

$$U_i' = \{x \in \mathbb{C}^{n+1} \mid x_i \neq 0\}$$

of $\mathbb{C}^{n+1}$ via $U_i = U_i'/\sim$. As we saw in Example 2.1.12, U_i' is an affine variety with coordinate ring $\mathbb{C}[U_i'] = S_{x_i}$ (the localization of S at x_i). The grading on S induces a grading on S_{x_i}, such that if a nonzero element of S_{x_i} is represented by f/x_i^ℓ, its degree is $\deg(f) - \ell$. The rational functions in S_{x_i} that give well defined functions on U_i are those of the form f/x_i^ℓ with $\deg(f) = \ell$. Indeed, if $\deg(f) = \ell$ then

$$\frac{f}{x_i^\ell}(\lambda x) = \frac{\lambda^\ell f(x)}{\lambda^\ell x_i^\ell} = \frac{f}{x_i^\ell}(x).$$

These are the elements of degree zero, denoted by $(S_{x_i})_0 = \mathbb{C}[U_i']_0$. Note that

$$\mathbb{C}[U_i']_0 = \left\{ \frac{f}{x_i^\ell} \mid f \in S_\ell, \ell \in \mathbb{N} \right\} = \mathbb{C}\left[\frac{x_0}{x_i}, \ldots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \ldots, \frac{x_n}{x_i} \right].$$

By the results of Subsection 2.1.4, the inclusion of finitely generated, nilpotent free $\mathbb{C}$-algebras $\mathbb{C}[U_i']_0 \to \mathbb{C}[U_i']$ gives a morphism $U_i' \to \mathbb{C}^n$ given by

$$(x_0, \ldots, x_n) \mapsto \left(\frac{x_0}{x_i}, \ldots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \ldots, \frac{x_n}{x_i} \right).$$

This morphism factors through U_i: $U_i' \to U_i \to \mathbb{C}^n$ and $U_i \to \mathbb{C}^n$ is clearly bijective. The following theorem tells us that it also identifies U_i and $\mathbb{C}^n$ as topological spaces.

Theorem 2.2.2. *The map $\phi_i : U_i \to \mathbb{C}^n$ given by*

$$(x_0 : \ldots : x_n) \mapsto \left(\frac{x_0}{x_i}, \ldots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \ldots, \frac{x_n}{x_i} \right) \tag{2.2.2}$$

is a homeomorphism of topological spaces with respect to the Zariski topology on both U_i and $\mathbb{C}^n$.

Proof. We need to show that closed subsets of U_i are identified with closed subsets of $\mathbb{C}^n$ under ϕ_i. We identify $\mathbb{C}^n$ with MaxSpec $\mathbb{C}[y_0, \ldots, y_{i-1}, y_{i+1}, y_n]$. Let $X_i \subset U_i$ be a closed subset with closure $X = \overline{X_i}$ in $\mathbb{P}^n$. The projective variety X gives a homogeneous ideal $I = I_S(X) = \langle f_1, \ldots, f_s \rangle \subset S$ with f_j homogeneous $j = 1, \ldots, s$. We set

$$\hat{f}_{ij} = f_j(y_1, \ldots, y_{i-1}, 1, y_{i+1}, \ldots, y_n). \tag{2.2.3}$$

and find that $\phi_i(X_i) = V_{\mathbb{C}^n}(\hat{f}_{i1}, \ldots, \hat{f}_{is})$. Conversely, for a closed subset $Y = V_{\mathbb{C}^n}(\hat{f}_{i1}, \ldots, \hat{f}_{is}) \subset \mathbb{C}^n$ let d_j be the smallest integer such that

$$f_j = x_i^{d_j} \hat{f}_{ij} \left(\frac{x_0}{x_i}, \ldots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \ldots, \frac{x_n}{x_i} \right) \tag{2.2.4}$$

is a homogeneous polynomial. We have that $\phi_i^{-1}(Y) = V_{\mathbb{P}^n}(f_1, \ldots, f_s) \cap U_i$. □

Theorem 2.2.2 shows that the affine variety $\mathbb{C}^n$ can be identified as a topological space with an open subset of $\mathbb{P}^n$. This makes $\mathbb{C}^n$ into a quasi-projective variety. Also, an affine variety $Y \subset \mathbb{C}^n$ corresponds to a closed subset $X_i \subset U_i$, which is open in its closure $X = \overline{X_i}$ in $\mathbb{P}^n$. Therefore, any affine variety is a quasi-projective variety.

The theorem also shows that $\mathbb{P}^n = \bigcup_{i=0}^n U_i$ writes $\mathbb{P}^n$ as a union of affine spaces. Each of the U_i is Zariski open in $\mathbb{P}^n$ and every $x \in \mathbb{P}^n$ belongs to at least one of the U_i. We say that $\{U_0, \ldots, U_n\}$ is an *affine open covering* of $\mathbb{P}^n$. The U_i are called the *affine charts* of $\mathbb{P}^n$. More generally, any projective variety $X \subset \mathbb{P}^n$ can be written as $X = \bigcup_{i=0}^n (X \cap U_i)$. As $X \cap U_i$ is closed in U_i, Theorem 2.2.2 shows that it can be identified with an affine variety $Y_i \subset \mathbb{C}^n$. We say that $\{X \cap U_0, \ldots, X \cap U_n\}$ is an *affine open covering* of X ($X \cap U_i$ is closed in U_i, but open in X). The Y_i are called the *affine charts* of X. It is slightly less straightforward that any quasi-projective variety has an affine open covering.

Theorem 2.2.3. *Any quasi-projective variety $X \subset \mathbb{P}^n$ can be written as $X = \bigcup_{i=1}^s Y_i$ where $Y_1, \ldots, Y_s$ are isomorphic to affine varieties. The set $\{Y_1, \ldots, Y_s\}$ is called an* affine open covering *of X.*

Proof. First, we write $X = \bigcup_{i=0}^n X \cap U_i$, which writes X as a union of open subsets of affine varieties. By Hilbert's basis theorem, every open subset U of an affine variety Y can be written as a finite union $U = Y_{f_1} \cup \ldots \cup Y_{f_{s'}}$ for some $f_1, \ldots, f_{s'} \in \mathbb{C}[Y]$ where

$$Y_{f_i} = \{x \in Y \mid f_i(x) \neq 0\}.$$

By Example 2.1.12, each Y_{f_i} is affine, which proves the theorem. □

Example 2.2.4 (The projective line)**.** The *projective line* $\mathbb{P}^1$ is covered by two copies of $\mathbb{C}$:

$$U_0 = \{(x_0 : x_1) \in \mathbb{P}^n \mid x_0 \neq 0\}, \quad U_1 = \{(x_0 : x_1) \in \mathbb{P}^n \mid x_1 \neq 0\}.$$

Note that $\mathbb{P}^1 \setminus U_0 = \{(0 : 1)\}$. We can send $\mathbb{C}$ into $\mathbb{P}^1$ by identifying it with U_0. This gives the map $\phi : t \mapsto (1 : t)$. Note that the point $(0 : 1) = \lim_{t\to\infty} \phi(t)$. For this reason, if $\mathbb{C}$ is identified with U_0, the point $(0 : 1) \in \mathbb{P}^n$ is called *the point at infinity* and with a slight abuse of notation we write $\mathbb{P}^1$ as the disjoint union $\mathbb{P}^1 = \mathbb{C} \sqcup \{\infty\}$. If we choose to identify $\mathbb{C}$ with $U_1 \subset \mathbb{P}^1$, the point $(1 : 0)$ is the point at infinity. △

Example 2.2.5 (Affine stratification of $\mathbb{P}^n$)**.** The construction in Example 2.2.4 generalizes to higher dimensions. If we choose to identify $\mathbb{C}^n$ with U_0, the complement $H_0 = \mathbb{P}^n \setminus U_0$ is called the *hyperplane at infinity*. This is the closed subspace

$$H_0 = V_{\mathbb{P}^n}(x_0) = \{(0 : x_1 : \cdots : x_n) \in \mathbb{P}^n \mid (x_1 : \cdots : x_n) \in \mathbb{P}^{n-1}\} = \mathbb{P}^{n-1}.$$

This shows that $\mathbb{P}^n$ can be written as the disjoint union

$$\mathbb{P}^n = \mathbb{C}^n \sqcup H_0 = \mathbb{C}^n \sqcup \mathbb{P}^{n-1} = \mathbb{C}^n \sqcup \mathbb{C}^{n-1} \sqcup \mathbb{C}^{n-2} \sqcup \cdots \sqcup \mathbb{C} \sqcup \{\infty\},$$

where $\mathbb{P}^1 = \mathbb{C} \sqcup \{\infty\}$ as in Example 2.2.4. This is called an *affine stratification of* $\mathbb{P}^n$. △

Example 2.2.6. Consider the homogeneous polynomial $f = xy - z^2 \in S_2$ with $S = \mathbb{C}[x, y, z]$. We consider the projective variety $X = V_{\mathbb{P}^2}(f)$. In the affine chart $U_x = \{(x : y : z) \in \mathbb{P}^2 \mid x \neq 0\}$, $X \cap U_x = Y_x$ has equation $y - z^2 = 0$ and looks like a parabola. On the other hand, $Y_z \simeq X \cap U_z$ has equation $xy - 1 = 0$, which is a hyperbola. A picture of (the real part of) these affine charts can be obtained by cutting the affine cone over X with the planes with equation $x = 1$ and $z = 1$ respectively. This is illustrated in Figure 2.7. We note that in $\mathbb{P}^2$, hyperbolas and parabolas look exactly the same, and they all look like an ellipse. The reason is that any ternary quadric corresponds to a symmetric 3×3 matrix, and any full rank 3×3 matrix is similar to any other full rank symmetric 3×3 matrix. Since full rank symmetric 3×3 matrices are exactly the ellipses/parabolas/hyperbolas in $\mathbb{P}^2$, they are all equal up to a change of coordinates. Rank two symmetric 3×3 matrices correspond to the union of two different lines (i.e. 2 copies of $\mathbb{P}^1$, e.g. $V_{\mathbb{P}^2}(xy)$) in $\mathbb{P}^2$, and the rank one case corresponds to a line with multiplicity 2 (e.g. $V_{\mathbb{P}^2}(x^2)$). See [Eis13, Exercise 1.15]. △

Remark 2.2.2. Note that for any nonzero polynomial $h = c_0x_0 + \ldots + c_nx_n \in S_1$, $U_h = \mathbb{P}^n \setminus V_{\mathbb{P}^n}(h)$ is an affine space. To see this, we can either consider a transformation of coordinates such that $x_i \leftarrow c_0x_0 + \ldots + c_nx_n$ or consider the map $U_h \to \mathbb{C}^{n+1}$

$$(x_0 : \cdots : x_n) \mapsto \left(\frac{x_0}{h(x_0, \ldots, x_n)}, \ldots, \frac{x_n}{h(x_0, \ldots, x_n)} \right)$$

which identifies U_h with $V_{\mathbb{C}^{n+1}}(h - 1) \simeq \mathbb{C}^n$ and proceed as in the proof of Theorem 2.2.2. △

The maps (2.2.3) and (2.2.4) establish an isomorphism of vector spaces

$$\eta_d : R_{\leq d} = \left\{ \sum_a c_a y^a \in R \mid \max_{c_a \neq 0} |a| \leq d \right\} \to S_d,$$

where $R = \mathbb{C}[y_1, \ldots, y_n]$ is the polynomial ring in n variables, $|a| = a_1 + \ldots + a_n$ and $S = \mathbb{C}[x_0, \ldots, x_n]$. The map η_d is defined by sending $\hat{f}_{ij} \in R_{\leq d}$ to $f_j \in S_d$ as in (2.2.4), but with d_j replaced by d. This map is called *homogenization of degree* d, and its inverse η_d^{-1} is called *dehomogenization*.

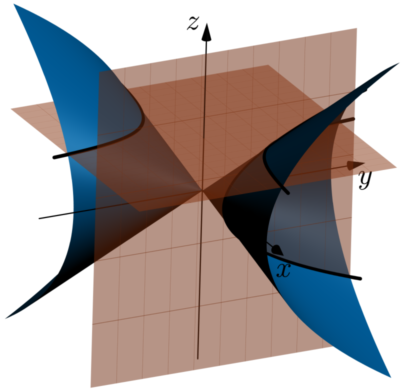


Figure 2.7: Two affine charts of $X = V_{\mathbb{P}^2}(xy - z^2)$ as in Example 2.2.6.

Example 2.2.7 (Projective closure of an affine variety)**.** It is sometimes useful to think of an affine variety as an affine chart of a projective variety. Let $Y \subset \mathbb{C}^n$ be an affine variety. We identify Y with the closed subset of $U_0 \subset \mathbb{P}^n$ given by $X_0 = \phi_0^{-1}(Y)$, where ϕ_0 is the map from Theorem 2.2.2. We define the *projective closure* of Y to be the Zariski closure $X = \overline{X_0}$ in $\mathbb{P}^n$. Given equations for $Y \subset \mathbb{C}^n$, we would like to know homogeneous equations for X. Suppose $Y = V_{\mathbb{C}^n}(\hat{f}_1, \ldots, \hat{f}_s)$ and let $d_i \in \mathbb{N}$ be the smallest number such that $\hat{f}_i \in R_{\leq d_i}$. A first guess would be that $X = V_{\mathbb{P}^n}(f_1, \ldots, f_s)$ where $f_i = \eta_{d_i}(\hat{f}_i)$. This is not true in general. It does work if $Y = V_{\mathbb{C}^n}(f)$ is an affine variety defined by only one equation. For instance, the projective closure of $Y = V_{\mathbb{C}^2}(y - z^2)$ is $X = V_{\mathbb{P}^2}(xy - z^2) = Y \sqcup \{(0 : 1 : 0)\}$ (with homogeneous coordinates $(x : y : z)$ on $\mathbb{P}^2$), see [SKKT04, Section 3.3]. An example where this doesn't work is the twisted cubic (see Example 2.1.3). This is the affine variety $Y = V_{\mathbb{C}^3}(y - x^2, z - x^3)$. Using homogeneous coordinates $(x : y : z : w)$ on $\mathbb{P}^3$ and thinking of Y as a subset of U_w, the projective variety $X = V_{\mathbb{P}^3}(wy - x^2, w^2 z - x^3)$ is a union of the closure of the twisted cubic and the line $\{(0 : y : z : 0) \mid (y : z) \in \mathbb{P}^1\} \simeq \mathbb{P}^1$. As the twisted cubic is irreducible in $\mathbb{C}^3$, so should its projective closure be in $\mathbb{P}^3$. The reason for this 'extra' component is that this is not a good representation of the vanishing ideal of the twisted cubic for the purpose of taking its projective closure. For more information, the reader can consult [CLO13, Chapter 8, §4]. $\triangle$

2.2.6 Regular functions and morphisms

In Subsection 2.1.4 we defined rings of polynomial functions on affine varieties and morphisms between affine varieties. Since affine varieties are quasi-projective varieties, we are now looking at a strictly larger class of objects. In this subsection, our goal is to define the *ring of regular functions* of quasi-projective varieties, and morphisms between them. The most important results of this subsection for the purpose of this thesis are the rings of regular functions in Example 2.2.8 and the fact that there is a notion of (iso-)morphisms which generalizes (iso-)morphisms in the affine setting.

We have established earlier that the only polynomial functions on $\mathbb{P}^n$ are the constants. However, if we consider open subsets and allow rational functions that are well defined on these subsets, we get much larger rings of functions. Just like elements of $(S_{x_i})_0$ give well-defined functions on U_i, rational functions of the form

$$\frac{f}{g}, \quad f, g \in S_\ell \text{ for some } \ell$$

give well defined functions on $\mathbb{P}^n \setminus V_{\mathbb{P}^n}(g)$. The proof of Theorem 2.2.2 shows that considering functions in $(S_{x_i})_0$ on U_i agrees with considering the polynomial functions on the affine variety $\mathbb{C}^n$ as we did in the previous section. The following definition associates the ring $(S_{x_i})_0$ to U_i as its *ring of regular functions*.

Definition 2.2.6 (Regular functions)**.** Let $X \subset \mathbb{P}^n$ be a quasi-projective variety and let $U \subset X$ be an open subset. A function $\phi = U \to \mathbb{C}$ is called *regular at* $x \in U$ if

$$\phi(p) = \frac{f}{g}(p), \quad \text{with } f, g \in S_\ell \text{ for some } \ell,$$

for all p in an open subset $U' \subset U$ containing x and such that $V_{\mathbb{P}^n}(g) \cap U' = \varnothing$. If ϕ is regular at all $x \in U$, we say that ϕ is *regular on* U. The ring of all regular functions on U is denoted by $\mathscr{O}_X(U)$.

Note that an open subset U of a quasi-projective variety X is again a quasi-projective variety and $\mathscr{O}_U(U) = \mathscr{O}_X(U)$. If it is not important that we think of U as a subset of X we will use the short notation $\mathscr{O}(U)$.

Remark 2.2.3. Definition 2.2.6 is quite technical. It is important that it has the following consequences.

1. A regular function ϕ on an open subset $U \subset X$ gives a regular function ϕ' on a smaller open subset $U' \subset U$ by *restricting* ϕ to U'.

2. Suppose an open subset $U \subset X$ is covered by open subsets $\{U_i'\}_{i \in \mathscr{I}}$ for some index set $\mathscr{I}$ (i.e. $U = \bigcup_{i \in \mathscr{I}} U_i'$). If a regular function $\phi : U \to \mathbb{C}$ restricts to 0 on U_i', for all $i \in \mathscr{I}$, then $\phi = 0$.

3. If ϕ_i' is a regular function on U_i', for all $i \in \mathscr{I}$, such that $\phi'_{i|U_i' \cap U_j'} = \phi'_{j|U_i' \cap U_j'}$ for all $i, j \in \mathscr{I}$, then $\{\phi_i'\}_{i \in \mathscr{I}}$ 'glue together' to a regular function ϕ on U (given by $\phi(x) = \phi_i'(x)$ when $x \in U_i'$). Indeed: at any point $x \in U$, choose $i \in \mathscr{I}$ such that $x \in U_i'$. Since ϕ_i' is regular, it looks like a rational function on an open subset of U_i' containing x, which is open in U. Since the $\phi_i'(x)$ agree on overlaps, the value of $\phi(x)$ is independent of the choice of i.

△

Example 2.2.8. If $Y \subset \mathbb{C}^n$ is affine, we have $\mathscr{O}_Y(Y) = \mathbb{C}[Y]$. If f is a nonzero element of $\mathbb{C}[Y]$, consider the open set

$$Y_f = \{x \in Y \mid f(x) \neq 0\}.$$

Then we have $\mathscr{O}_Y(Y_f) = \mathbb{C}[Y]_f = \mathbb{C}[Y_f]$ and we can think of the canonical map $\mathbb{C}[Y] \to \mathbb{C}[Y]_f$ as the *restriction* of a function on Y to the open subset Y_f. For any nonempty projective variety $X \subset \mathbb{P}^n$, $\mathscr{O}_X(X) = \mathbb{C}$. For $f \in \mathbb{C}[X]$, the quasi-projective variety

$$X_f = \{(x_0 : \cdots : x_n) \mid f(x) \neq 0\}$$

has ring of regular functions $\mathscr{O}_X(X_f) = (\mathbb{C}[X]_f)_0$. Restriction from X to X_f is given by the inclusion $\mathbb{C} \to (\mathbb{C}[X]_f)_0$. △

In the affine case, we defined morphisms $Y \to Y'$ between affine varieties as maps that pull back to $\mathbb{C}$-algebra homomorphisms $\mathbb{C}[Y'] \to \mathbb{C}[Y]$. This definition is valid for morphisms between open subsets of the form Y_f, since these are again affine (see Example 2.1.12). We extend this definition to general open subsets of affine varieties first.

Definition 2.2.7. Let $U \subset Y, U' \subset Y'$ be open subsets of affine varieties Y, Y'. A function $\Phi : U \to U'$ is a *morphism* if the composition of any regular function $\phi' : U' \to \mathbb{C}$ with Φ is a regular function $\phi = \phi' \circ \Phi : U \to \mathbb{C}$. Equivalently, Φ is a morphism if $\phi' \mapsto \phi' \circ \Phi$ is a map of rings $\Phi^* : \mathscr{O}_{Y'}(U') \to \mathscr{O}_Y(U)$.

Remark 2.2.4. The map sending a function ϕ' to a composition $\phi' \circ \Phi$ is always a $\mathbb{C}$-algebra homomorphism: $(c\phi') \mapsto (c\phi') \circ \Phi = c(\phi' \circ \Phi), c \in \mathbb{C}$. △

Definition 2.2.8. Let $X \subset \mathbb{P}^n, X' \subset \mathbb{P}^m$ be quasi-projective varieties. Let $\{Y_1, \dots, Y_s\}$ and $\{Y'_1, \dots, Y'_{s'}\}$ be affine open coverings of X and X' respectively. A function $\Phi : X \to X'$ is a *morphism* if for all i, j,

$$\Phi_{Y_i \cap \Phi^{-1}(Y'_j)} : Y_i \cap \Phi^{-1}(Y'_j) \to Y'_j$$

is a morphism as defined in Definition 2.2.7.

Two quasi-projective varieties X, X' are *isomorphic* if there exist morphisms $\Phi : X \to X'$ and $\Psi : X' \to X$ such that $\Phi \circ \Psi = \mathrm{id}_{X'}$ and $\Psi \circ \Phi = \mathrm{id}_X$.

Example 2.2.9. The homeomorphism ϕ_i in Theorem 2.2.2 is an isomorphism of quasi-projective varieties, since $\phi_i^*(f) \in (S_{x_i})_0 = \mathscr{O}(U_i) = (\mathbb{C}[\mathbb{P}^n]_{x_i})_0$ for all $f \in \mathscr{O}(\mathbb{C}^n) = \mathbb{C}[\mathbb{C}^n] = \mathbb{C}[y_1, \dots, y_n]$. △

Example 2.2.10. A composition of morphisms is a morphism and the identity map $\mathrm{id}_X : X \to X$ is an isomorphism. Every inclusion $U \subset U'$ of open subsets of X is a morphism which gives a *restriction* map $\mathscr{O}_X(U') \to \mathscr{O}_X(U)$, and if $U = U'$ this is the identity map $\mathrm{id}_{\mathscr{O}_X(U)}$. In the language of category theory, $\mathscr{O}_X$ is a contravariant functor from 'open subsets U of X with inclusion maps' to 'rings $\mathscr{O}_X(U)$ with restriction maps'. This, together with the observations in Remark 2.2.3, makes $\mathscr{O}_X$ into a *sheaf of rings* on X, called the *structure sheaf* of X. Going more into detail would take us to far. We refer the reader to [EH06, Section I.1.3], [Har77, Chapter 2] or [Ser55]. △

2.2.7 Dimension and degree

In this subsection we introduce the concepts of dimension and degree for a projective variety. For the dimension, we could use a topological definition such as Definition 2.1.7. Instead (but equivalently), we will use the definition of dimension for affine varieties.

Definition 2.2.9 (Dimension of a quasi-projective variety)**.** The *dimension* of a quasi-projective variety X, denoted $\dim X$, with affine open covering $\{Y_1, \ldots, Y_s\}$ is $\max_i \dim Y_i$ (as affine varieties). The *codimension* of a quasi-projective variety $X \subset \mathbb{P}^n$ is $\operatorname{codim} X = n - \dim X$.

Theorem 2.2.4. *Let $X, X' \subset \mathbb{P}^n$ be irreducible projective varieties of dimension k, ℓ respectively. Then every irreducible component of the projective variety $X \cap X' \subset \mathbb{P}^n$ has dimension at least $k + \ell - n$. In particular, if $k + \ell \geq n$ then $X \cap X' \neq \varnothing$.*

Proof. See [Har77, Chapter 1, Theorem 7.2]. □

Example 2.2.11. Two lines in the projective plane $\mathbb{P}^2$ always meet, which corresponds to the intuition that parallel lines in $\mathbb{C}^2$ meet 'at infinity'. △

The *degree* of a projective variety tells us 'how far' the variety is from being linear (i.e. given by linear equations). A first definition is very intuitive but hard to make rigorous.

Definition 2.2.10 (Degree of a projective variety)**.** Let $X \subset \mathbb{P}^n$ be a projective variety such that all irreducible components of X have dimension k. The *degree* of X, denoted $\deg X$, is the number of intersection points of X with a 'general' linear subvariety of $\mathbb{P}^n$ of codimension k.

A *linear subvariety* or *linear subspace* of $\mathbb{P}^n$ is a projective subvariety defined by linear equations (i.e. elements of S_1). The problem with Definition 2.2.10 is that it is rather complicated to make the word 'general' precise. We will mention an algebraic definition of degree below, but Definition 2.2.10 will often be more useful for our purposes as it is more intuitive. The reader should think of a 'general' linear subvariety as one defined by linear equations with *random* complex coefficients (e.g. with real and imaginary part drawn from a normal distribution).

Example 2.2.12. If $f \in S \setminus \{0\}, \deg(f) = d$, then $V_{\mathbb{P}^n}(f)$ is called a *hyperplane* if $d = 1$ and a *hypersurface of degree* d for general d. For $n = 2$, a hypersurface is called a curve. A curve of degree 2, 3, 4, ... is called a plane conic, cubic, quartic, ... For $n = 3$, a hypersurface is called a surface. A surface of degree 2, 3, 4, ... is called a quadratic, cubic, quartic, ... surface. △

An algebraic definition of dimension and degree for projective varieties is provided by an important tool called the *Hilbert function.* It is defined as follows.

Definition 2.2.11 (Hilbert function)**.** Let $I \subset S$ be a homogeneous ideal of S. The *Hilbert function* of I is

$$\mathrm{HF}_I : \mathbb{Z} \to \mathbb{N} \quad \text{given by} \quad \mathrm{HF}_I(d) = \dim_{\mathbb{C}}(S/I)_d.$$

The Hilbert function of a projective variety X is $\mathrm{HF}_X = \mathrm{HF}_{I_S(X)}$, i.e. $\mathrm{HF}_X(d) = \dim_{\mathbb{C}} \mathbb{C}[X]_d$.

The Hilbert function can be defined for any graded S-module, but considering modules of the form S/I for some homogeneous ideal $I \subset S$ suffices for us. Remarkably, the Hilbert function HF_X of a projective variety carries a lot of geometric information.

Theorem 2.2.5 (Hilbert-Serre)**.** *Let $I \subset S$ be a homogeneous ideal and let $X = V_{\mathbb{P}^n}(I)$. There exists a unique polynomial $\mathrm{HP}_I \in \mathbb{Q}[t]$ such that for some $\ell \in \mathbb{N}$, $\mathrm{HF}_I(d) = \mathrm{HP}_I(d)$ for all $d \geq \ell$. Moreover, the degree of $\mathrm{HP}_I(t)$ is $\dim X$ and if $I = I_S(X)$, the degree of X is defined as the leading coefficient of $\mathrm{HP}_I(t)$, multiplied with $(\dim X)!$. That is,*

$$\mathrm{HP}_{I_S(X)}(t) = \frac{\deg X}{(\dim X)!} t^{\dim X} + \textit{lower order terms.}$$

If all irreducible components of X have the same dimension, this definition of degree agrees with Definition 2.2.10.

Proof. See [CLO06, Chapter 6, §4, Proposition 4.7] for the existence of HP_I, [Har77, Chapter 1, Theorem 7.5] for the statement about $\dim X$ and [Cut18, Theorem 16.9] for the equivalence of the definitions for $\deg X$. □

The polynomial HP_I in Theorem 2.2.5 is called the *Hilbert polynomial* of I, and the Hilbert polynomial of a projective variety $X \subset \mathbb{P}^n$ is defined as $\mathrm{HP}_X = \mathrm{HP}_{I_S(X)}$. The theorem implies by the projective Nullstellensatz that $\deg \mathrm{HP}_I = \deg \mathrm{HP}_{\sqrt{I}}$.

Remark 2.2.5. In the notation of Theorem 2.2.5, if $I \subsetneq I_S(X)$, the leading coefficient of HP_I encodes the degree of the *projective scheme* associated to I. This takes into account, for instance, that certain irreducible components of $V_{\mathbb{P}^n}(I)$ may occur with arbitrary multiplicities. For more information, see [EH06, Chapter 3]. △

Example 2.2.13 (The Hilbert function of $\mathbb{P}^n$)**.** The Hilbert function of the projective space $\mathbb{P}^n$ is given by

$$\mathrm{HF}_{\mathbb{P}^n}(d) = \dim_{\mathbb{C}}(S_d) = \binom{n+d}{n} \quad \text{where } \binom{\ell}{k} = \begin{cases} \frac{\ell!}{k!(\ell-k)!} & \ell \geq k \\ 0 & \text{otherwise} \end{cases}.$$

In this case $\mathrm{HF}_{\mathbb{P}^n}(d) = \mathrm{HP}_{\mathbb{P}^n}(d)$ for $d \geq 0$. △

Example 2.2.14 (The Hilbert function of a hypersurface)**.** Let $X = V_{\mathbb{P}^n}(f)$ for $f \in S_{d_f}$ homogeneous and of degree d_f. Assume moreover that f is square-free, which means that $I_S(V_{\mathbb{P}^n}(f)) = \langle f \rangle$. For $I = \langle f \rangle$, we have

$$\dim_{\mathbb{C}} I_d = \dim_{\mathbb{C}}\{gf \mid g \in S_{d-d_f}\} = \begin{cases} \mathrm{HF}_{\mathbb{P}^n}(d-d_f) & d \geq d_f \\ 0 & \text{otherwise} \end{cases}.$$

Since $\mathrm{HF}_X(d) = \dim_{\mathbb{C}}(S/I)_d = \dim_{\mathbb{C}} S_d - \dim_{\mathbb{C}} I_d$ we get

$$\mathrm{HF}_X(d) = \begin{cases} \mathrm{HF}_{\mathbb{P}^n}(d) & d < d_f \\ \mathrm{HF}_{\mathbb{P}^n}(d) - \mathrm{HF}_{\mathbb{P}^n}(d-d_f) & d \geq d_f \end{cases}$$

and the Hilbert polynomial HP_X agrees with the Hilbert function for $d \geq d_f$. It is given by

$$\mathrm{HP}_X(d) = \binom{n+d}{n} - \binom{n+d-d_f}{n} = \frac{d_f}{(n-1)!} d^{n-1} + \dots.$$

△

2.3 Abstract varieties

In the previous section we have started by defining the projective n-space $\mathbb{P}^n$ and showed that it is covered by affine open subsets which overlap on Zariski open subsets. In this section, we will go the other way around and define a topological space by 'gluing together' affine varieties. This construction will give us a good way of thinking about toric varieties, which will play an important role in later chapters.

Consider a set $\{Y_i\}_{i \in \mathscr{T}}$ of affine varieties for some index set $\mathscr{T}$. Suppose that for all pairs $i, j \in \mathscr{T}$, we have isomorphic Zariski open subsets $Y_{ij} \subset Y_i$, $Y_{ji} \subset Y_j$. Let $\{\phi_{ij}\}_{i,j \in \mathscr{T}}$ be isomorphisms such that for all $i, j, k \in \mathscr{T}$,

1. $\phi_{ij} : Y_{ij} \to Y_{ji}$ and $\phi_{ji} : Y_{ji} \to Y_{ij}$ satisfy $\phi_{ij} \circ \phi_{ji} = \mathrm{id}_{Y_{ji}}, \phi_{ji} \circ \phi_{ij} = \mathrm{id}_{Y_{ij}}$,
2. $\phi_{ij}(Y_{ij} \cap Y_{ik}) = Y_{ji} \cap Y_{jk}$,
3. $\phi_{ik} = \phi_{jk} \circ \phi_{ij}$ on $Y_{ik} \cap Y_{ij}$.

The disjoint union $\bigsqcup_{i\in\mathscr{I}} Y_i$ is the set

$$\hat{X} = \bigsqcup_{i\in\mathscr{I}} Y_i = \{(x, Y_i) \mid i \in \mathscr{I}, x \in Y_i\}.$$

It is a topological space with the disjoint union topology, which is such that the open subsets of $\hat{X}$ are disjoint unions of open subsets in the Y_i. We define an equivalence relation $\sim$ on $\hat{X}$ by setting $(x, Y_i) \sim (y, Y_j)$ if $x \in Y_{ij}$, $y \in Y_{ji}$ and $\phi_{ij}(x) = y$. The first condition on the ϕ_{ij} makes $\sim$ reflexive and symmetric, the second and third conditions make it transitive. We consider the quotient space $X = \hat{X}/\sim$ with its quotient topology, in which

$$U_i = \{[(x, Y_i)] \mid x \in Y_i\} \subset X$$

are open subsets isomorphic to Y_i (here we denoted $[\cdot]$ for an equivalence class in the quotient). The topological space X is called the *gluing* of the affine varieties in $\{Y_i\}_{i\in\mathscr{I}}$ and $\{Y_i\}_{i\in\mathscr{I}}, \{\phi_{ij}\}_{i,j\in\mathscr{I}}$ are called the *gluing data.*

Example 2.3.1 (Gluing of $\mathbb{P}^1$)**.** The projective line $\mathbb{P}^1$ is covered by $\mathbb{P}^1 = U_x \cup U_y$ where

$$U_x = \{(x : y) \in \mathbb{P}^1 \mid x \neq 0\}, \quad U_y = \{(x : y) \in \mathbb{P}^1 \mid y \neq 0\}.$$

Consider the isomorphisms

$$h_x : U_x \to \mathbb{C}_t \quad \text{and} \quad h_y : U_y \to \mathbb{C}_u,$$

where $\mathbb{C}_t$ is $\mathbb{C}$ with coordinate t and analogously for u, given by $h_x(x : y) = y/x$ and $h_y(x : y) = x/y$ (these are the maps ϕ_i in Theorem 2.2.2). For a point $(x : y) \in U_x \cap U_y$, we have $h_x(x : y) = h_y(x : y)^{-1}$. Let

$$\mathbb{C}_{tu} = \mathbb{C}_t^* = \mathbb{C}_t \setminus \{0\}, \quad \mathbb{C}_{ut} = \mathbb{C}_u^* = \mathbb{C}_u \setminus \{0\}$$

and $\phi_{tu} : \mathbb{C}_{tu} \to \mathbb{C}_{ut}$ given by $\phi_{tu}(t) = t^{-1}$, $\phi_{ut} = \phi_{tu}^{-1}$. This gives the following commutative diagram.

$$\begin{tikzcd}
U_x \cap U_y \arrow[r, "h_x"] \arrow[d, "h_y"'] & \mathbb{C}_{tu} \arrow[dl, shift left, "\phi_{tu}"] \\
\mathbb{C}_{ut} \arrow[ur, shift left, "\phi_{ut}"] &
\end{tikzcd}$$

The projective line $\mathbb{P}^1$ is a gluing of two copies of $\mathbb{C}$ with gluing data $\{\mathbb{C}_t, \mathbb{C}_u\}$ and $\{\phi_{tu}, \phi_{ut}\}$. The two affine lines $\mathbb{C}_t$ and $\mathbb{C}_u$ are glued together along the open subsets $\mathbb{C}_t^*$ and $\mathbb{C}_u^*$ to get the open subset $U_x \cap U_y \subset \mathbb{P}^1$. The missing points $\mathbb{P}^1 \setminus (U_x \cap U_y) = \{(1 : 0), (0 : 1)\}$ correspond to the origin in $\mathbb{C}_t$ and $\mathbb{C}_u$. If we consider $\mathbb{P}^1$ as the projective closure of $\mathbb{C}_t$, the *point at infinity* (see Example 2.2.4) corresponds to the origin in $\mathbb{C}_u$. This gluing construction is illustrated in Figure 2.8. $\triangle$

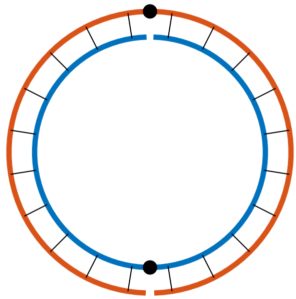


Figure 2.8: Illustration of the construction of $\mathbb{P}^1$ as the gluing of two affine lines. The affine lines are represented as circles with a missing point ('at infinity'). The origin in each line is indicated with a black dot and the gluing isomorphism is illustrated by black line segments.

Example 2.3.2 (Gluing of $\mathbb{P}^2$)**.** One can repeat Example 2.3.1 for higher dimensional projective spaces. For $\mathbb{P}^2$, we consider the isomorphisms

$$h_x : U_x \to \mathbb{C}^2_t, \ h_y : U_y \to \mathbb{C}^2_u, \text{ and } h_z : U_z \to \mathbb{C}^2_v$$

where $\mathbb{C}^2_t$ is the affine plane with coordinates t_1, t_2 (analogously for u, v) and

$$h_x(x : y : z) = (y/x, z/x), \ h_y(x : y : z) = (x/y, z/y), \ h_z(x : y : z) = (x/z, y/z).$$

The gluing morphisms $\phi_{tv} = \phi_{vt}^{-1}$ come from identifying the images of points in $U_x \cap U_z$ under h_x and h_z, e.g. on $\mathbb{C}^2_{tv} = \mathbb{C}^2_t \setminus V(t_2)$

$$\phi_{tv}(t_1, t_2) = (t_2^{-1}, t_1 t_2^{-1}) \quad \text{comes from} \quad \left(\frac{x}{z}, \frac{y}{z}\right) = \left(\left(\frac{z}{x}\right)^{-1}, \left(\frac{y}{x}\right)\left(\frac{z}{x}\right)^{-1}\right).$$

The morphism sends the parabola Y_x from Example 2.2.6 (more precisely, its intersection with $\mathbb{C}^2_{tv}$) to the hyperbola Y_z (intersected with $\mathbb{C}^2_{vt}$). △

All quasi-projective varieties can be obtained via the gluing construction. From now on, we will use the word *variety* for any topological space that is obtained from a gluing of affine varieties as described in this section. Using Definitions 2.2.6, 2.2.7 and 2.2.8 it is straightforward to define regular functions on open subsets of varieties and morphisms between varieties. Dimension can also be defined locally. An analogous construction can be used for gluing affine schemes together to obtain general schemes [EH06, Section I.2.4]. As mentioned before, an important application in the context of this thesis is the gluing of a complete toric variety from a set of affine toric varieties. In this case, the gluing data has a particularly nice description in terms of a polytope (or in its normal fan). This construction generalizes Examples 2.3.1 and 2.3.2 and is described in Appendix E.

Chapter 3

Zero-dimensional varieties

In this chapter we discuss *zero-dimensional* subvarieties of $\mathbb{C}^n$ and $\mathbb{P}^n$. These are varieties consisting of finitely many points. Understanding their coordinate rings allows us to compute coordinates for these points via eigenvalue computations. In the affine case, this is a result called the *classical eigenvalue, eigenvector theorem*. Together with a description of the *multiplicity* (or *scheme*) *structure* of zero-dimensional algebras and an affine version of Bézout's theorem, this is the subject of Section 3.1. In Section 3.2, after introducing the necessary theory on Hilbert functions and Bézout's theorem, we formulate a projective version of the eigenvalue, eigenvector theorem and we discuss the effects of *homogenizing* a given set of affine equations. Among the methods for polynomial system solving that exploit these results are Gröbner and border basis techniques and Macaulay resultants. Since these approaches are strongly related to the framework of *truncated normal forms*, introduced in the next chapter, we will give an overview in Sections 3.3 and 3.4.

We use the following notation for some basic concepts from linear algebra. For a finite dimensional $\mathbb{C}$-vector space W, we write $W^\vee = \mathrm{Hom}_\mathbb{C}(W, \mathbb{C})$ for the dual vector space. For a vector space endomorphism $\phi : W \to W$, a *right eigenpair* is a tuple $(\lambda, w) \in \mathbb{C} \times (W \setminus \{0\})$ satisfying $\phi(w) = \lambda w$. Similarly, a *left eigenpair* is a tuple $(v, \lambda) \in (W^\vee \setminus \{0\}) \times \mathbb{C}$ such that $v \circ \phi = \lambda v$. The $\mathbb{C}$-linear span of a subset $\mathcal{W} \subset W$ is denoted by $\mathrm{span}_\mathbb{C}(\mathcal{W}) \subset W$.

3.1 Points in affine space

Throughout this section, let $R = \mathbb{C}[x_1, \ldots, x_n]$ be the n-variate polynomial ring over $\mathbb{C}$ and for $f_1, \ldots, f_s \in R$ let $I = \langle f_1, \ldots, f_s \rangle \subset R$ be an ideal. We assume that the affine variety defined by I consists of finitely many points:

$$V(I) = V_{\mathbb{C}^n}(I) = \{z_1, \ldots, z_\delta\} \subset \mathbb{C}^n.$$

Ideals of R satisfying this assumption are called *zero-dimensional*, which reflects the dimension of $V(I)$ as an algebraic variety and, equivalently, the Krull dimension of R/I (see Subsection 2.1.5). We remark, for the reader who is familiar with commutative algebra, that by [AM69, Theorem 8.5] these are exactly the ideals of R for which R/I is *Artin*.

3.1.1 The eigenvalue, eigenvector theorem

In this subsection we will make the extra assumption that $I = \sqrt{I}$ is a *radical ideal*. This is equivalent to the assumption that R/I is *nilpotent free* or *reduced*. We will discuss the more general case in Subsection 3.1.3. By the Nullstellensatz (Theorem 2.1.1), the assumption $I = \sqrt{I}$ implies

$$I = I(V(I)) = \{f \in R \mid f(z_i) = 0, i = 1, \ldots, \delta\}.$$

This makes it particularly easy to describe the quotient ring R/I. The following lemma will be helpful.

Lemma 3.1.1. *For a collection of $\delta < \infty$ points $\{z_1, \ldots, z_\delta\} \subset \mathbb{C}^n$, there is a linear form $h = h_1x_1 + \cdots + h_nx_n \in R$ such that $h(z_i) \neq h(z_j), i \neq j$.*

Proof. If $\delta = 1$, there is nothing to prove. For $\delta > 1$, the condition that $h(z_i) = h(z_j), i \neq j$ is a (nonzero) linear condition on the coefficients $h_1, \ldots, h_n$. Let

$$C = \binom{\delta}{2} = \frac{\delta!}{2(\delta - 2)!}.$$

In total, this gives at most C pairwise linearly independent conditions, which means that the points $(h_1, \ldots, h_n) \in \mathbb{C}^n$ for which h *does not* satisfy the desired property are on the union of at most C hyperplanes through the origin in $\mathbb{C}^n$. □

The proof of Lemma 3.1.1 shows that *almost all* linear forms $h = h_1x_1 + \cdots + h_nx_n \in R$ satisfy $h(z_i) \neq h(z_j), i \neq j$. We say that a *generic* linear form has this property. We will say more about the notion of genericity in Subsection 3.1.2.

Definition 3.1.1 (Evaluation map). Let $I = \sqrt{I}$ be a zero-dimensional ideal with $V(I) = \{z_1, \ldots, z_\delta\}$. For $i = 1, \ldots, \delta$, we define $\mathrm{ev}_{z_i} \in (R/I)^\vee$ by $\mathrm{ev}_{z_i}(f + I) = f(z_i)$. Furthermore, we define the *evaluation map* $\psi : R/I \to \mathbb{C}^\delta$ by $\psi = (\mathrm{ev}_{z_1}, \ldots, \mathrm{ev}_{z_\delta})$, that is

$$\psi(f + I) = (f(z_1), \ldots, f(z_\delta)).$$

Note that the map $\psi : R/I \to \mathbb{C}^\delta$ is well-defined: if $f, g \in R$ are such that $f - g \in I$, then $f(z_i) = g(z_i), i = 1, \ldots, \delta$. Moreover, the map ψ is $\mathbb{C}$-linear. Lemma 3.1.1 allows us to construct polynomials whose residue classes map to the standard basis vectors of $\mathbb{C}^\delta$ under ψ.

Lemma 3.1.2. *Consider the evaluation map from Definition 3.1.1. There exist polynomials* $\ell_1, \ldots, \ell_\delta \in R$ *satisfying*

$$\ell_i(z_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}. \tag{3.1.1}$$

Polynomials satisfying (3.1.1) *are called* Lagrange polynomials *for* $\{z_1, \ldots, z_\delta\}$.

Proof. Let h be as in Lemma 3.1.1 and set

$$\ell_i = \frac{\prod_{i \neq j}(h(x) - h(z_j))}{\prod_{i \neq j}(h(z_i) - h(z_j))}.$$

□

Proposition 3.1.1. *For a zero-dimensional ideal* $I = \sqrt{I}$*, an element* $f + I \in R/I$ *is completely determined by the values* $f(z_1), \ldots, f(z_\delta)$*. In particular, the evaluation map* $\psi : R/I \to \mathbb{C}^\delta$ *is an isomorphism of* $\mathbb{C}$*-vector spaces.*

Proof. Since $I = \sqrt{I}$, the map ψ is injective: $\psi(f + I) = 0$ implies $f \in I$. To show that it is also surjective, let $V(I) = \{z_1, \ldots, z_\delta\}$ and let $\ell_1, \ldots, \ell_\delta \in R$ be a set of Lagrange polynomials of $V(I)$ (these exist by Lemma 3.1.2). Then surjectivity follows from $\psi(\ell_i + I) = e_i$, where $e_i = (0, \ldots, 1, \ldots, 0)$ (1 in the i-th position) is the i-th standard basis vector of $\mathbb{C}^\delta$. □

Proposition 3.1.1 establishes the fact that, under the assumptions of this subsection, R/I has dimension δ as a $\mathbb{C}$-vector space (we write $\dim_\mathbb{C} R/I = \delta$, whereas $\dim R/I = 0$ denotes the Krull dimension) and the evaluation map gives us one way to define coordinates on R/I. It also shows that $\{\ell_1 + I, \ldots, \ell_\delta + I\}$ is a $\mathbb{C}$-basis for R/I with dual basis $\{\mathrm{ev}_{z_1}, \ldots, \mathrm{ev}_{z_\delta}\}$ for $(R/I)^\vee$. The next step is to understand the structure of R/I as an R-module in terms of linear algebra operations.

Definition 3.1.2 (Multiplication map)**.** For any $g \in R$ we define the *multiplication map* representing *multiplication with* g as the $\mathbb{C}$-linear map

$$M_g : R/I \to R/I \quad \text{with} \quad M_g(f + I) = fg + I.$$

Note that the multiplication maps define the structure of R/I as an R-module, in the sense that scalar multiplication is given by $R \times R/I \to R/I$ with $(g, f+I) \mapsto M_g(f+I)$. Since M_g is a $\mathbb{C}$-linear endomorphism on a finite dimensional vector space, it can be represented by a matrix once we fix coordinates. With the very special choice of coordinates discussed above, these matrices are diagonal. This leads immediately to a proof of the main theorem of this subsection.

Theorem 3.1.1 (Eigenvalue, eigenvector theorem)**.** *Let* $I = \sqrt{I}$ *be a zero-dimensional ideal of* R *with* $V(I) = \{z_1, \ldots, z_\delta\}$*. The multiplication maps* $M_g : R/I \to R/I$ *are pairwise commuting and have left and right eigenpairs*

$$(\mathrm{ev}_{z_i}, g(z_i)), \quad (g(z_i), \ell_i + I), \quad i = 1, \ldots, \delta.$$

Proof. The fact that $M_{g_1} \circ M_{g_2} = M_{g_2} \circ M_{g_1}$ for any $g_1, g_2 \in R$ follows directly from Definition 3.1.2. The statement about the eigenpairs follows from the fact that ψ is a vector space isomorphism and the diagram

$$\begin{tikzcd} R/I \arrow[r, "M_g"] \arrow[d, "\psi"] & R/I \arrow[d, "\psi"] \\ \mathbb{C}^\delta \arrow[r, "\Delta"] & \mathbb{C}^\delta \end{tikzcd}$$

commutes, where Δ is the linear map corresponding to the diagonal matrix $\mathrm{diag}(g(z_1), \ldots, g(z_\delta))$. □

Remark 3.1.1. The name of Ludwig Stickelberger is often attached to this theorem. See [Cox20b] for a discussion on why, and for an overview of the theorem's origins. △

Example 3.1.1 (Companion matrices for $n = 1$)**.** Let $f = c_0 + c_1 x + \cdots + c_\delta x^\delta \in \mathbb{C}[x]$ with $c_\delta \neq 0$ and $I = \langle f \rangle \subset \mathbb{C}[x]$. Moreover, suppose that $I = \sqrt{I}$ such that f has δ distinct roots $V(f) = \{z_1, \ldots, z_\delta\}$. The algebra $\mathbb{C}[x]/I$ has dimension δ as a $\mathbb{C}$-vector space and the Lagrange polynomials

$$\ell_i = \frac{\prod_{i \neq j}(x - z_j)}{\prod_{i \neq j}(z_i - z_j)}, \quad i = 1, \ldots, \delta$$

give the $\mathbb{C}$-basis $\{\ell_1 + I, \ldots, \ell_\delta + I\}$ for $\mathbb{C}[x]/I$. However, in order to compute the ℓ_i, we need to know the roots. An alternative basis for $\mathbb{C}[x]/I$ is given by $\{1 + I, x + I, \ldots, x^{\delta-1} + I\}$. It is easy to check that these monomials are indeed $\mathbb{C}$-linearly independent modulo I. Let us construct the matrix representation of $M_x : \mathbb{C}[x]/I \to \mathbb{C}[x]/I$ in this basis. By $M_x(x^a + I) = x^{a+1} + I$ and $x^\delta + I = -c_\delta^{-1}(c_0 + c_1 x + \cdots + c_{\delta-1}x^{\delta-1}) + I$, we get that

$$M_x = \begin{bmatrix} & & & & -c_0/c_\delta \\ 1 & & & & -c_1/c_\delta \\ & 1 & & & -c_2/c_\delta \\ & & \ddots & & \vdots \\ & & & 1 & -c_{\delta-1}/c_\delta \end{bmatrix},$$

where $e_i \in \mathbb{C}^\delta$ is identified with $x^{i-1} + I$. This is the so-called *Frobenius companion matrix* of f, whose eigenvalues are well-known to be the roots of f. This observation is at the heart of many numerical algorithms for univariate root finding, such as [AMVW15]. The roots $z_1, \ldots, z_\delta$ are indeed the values $g(z_1), \ldots, g(z_\delta)$ for $g = x$, and Theorem 3.1.1 also characterizes the left and right eigenvectors of this matrix. △

With a slight abuse of notation, where there is no confusion possible we let M_g denote both the linear map $M_g : R/I \to R/I$ and its matrix representation in some basis. Theorem 3.1.1 tells us that a matrix representation M_g has *eigenvalue decomposition* (see Appendix B)

$$D M_g D^{-1} = \mathrm{diag}(g(z_1), \ldots, g(z_\delta)),$$

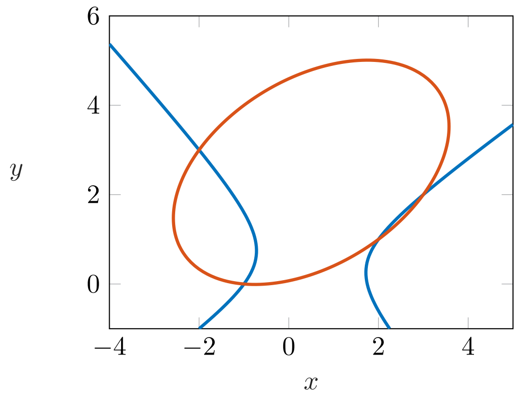


Figure 3.1: Picture in $\mathbb{R}^2$ of the algebraic curves $V(f_1)$ (in blue) and $V(f_2)$ (in orange) from Example 3.1.2.

where the rows of D represent the linear functionals ev_{z_i} and $\mathrm{diag}(g(z_1), \ldots, g(z_\delta))$ is a $\delta \times \delta$ diagonal matrix with the values $g(z_i)$ on its diagonal. Note that the matrix D does not depend on g. Indeed, $\{M_g \mid g \in R\}$ is a commuting family of matrices which share eigenvectors. This naturally leads to the following pseudo-algorithm for computing coordinates of $z_1, \ldots, z_\delta$.

1. For some basis of R/I, compute the matrices $M_{x_1}, \ldots, M_{x_n}$.
2. Diagonalize them simultaneously (compute $DM_{x_i}D^{-1} = \mathrm{diag}(z_{1i}, \ldots, z_{ni})$, $i = 1, \ldots, n$) and read off the coordinates from the diagonal.

Among the classical methods for performing step 1 are Gröbner basis, border basis or resultant techniques, as we will discuss in Sections 3.3 and 3.4. Section 4.2 is devoted to developing the framework of *truncated normal forms*, which generalizes the above mentioned approaches and is highly flexible for taking *numerical stability* into account. In this thesis, we leave step 2 mostly to a 'numerical linear algebra blackbox' which uses the standard techniques for computing (joint) eigenvalue decompositions. We will say a little more about this in Section 4.3.

Example 3.1.2 (Intersecting two conics in the plane)**.** This is an example taken from [TMVB18]. Let $R = \mathbb{C}[x, y]$ and consider the ideal $I = \langle f_1, f_2 \rangle$ with

$$f_1 = 7 + 3x - 6y - 4x^2 + 2xy + 5y^2,$$
$$f_2 = -1 - 3x + 14y - 2x^2 + 2xy - 3y^2.$$

As illustrated in Figure 3.1, the two curves $V(f_1)$ and $V(f_2)$ meet in four real points $z_1 = (-2, 3), z_2 = (3, 2), z_3 = (2, 1), z_4 = (-1, 0)$ and these are the only points in $V(I) \subset \mathbb{C}^2$. A $\mathbb{C}$-basis for R/I is $\mathcal{B} = \{x + I, y + I, x^2 + I, xy + I\}$ and one can check the identities

$$x^3 + I = -2x + 12y - 3x^2 + 6xy + I$$
$$x^2y + I = \frac{-15}{4}x + \frac{33}{2}y - \frac{15}{4}x^2 + 5xy + I$$

in R/I. Using the basis $\mathcal{B}$ (with its elements ordered as above) we obtain the matrix representation

$$M_x = \begin{bmatrix} 0 & 0 & -2 & -15/4 \\ 0 & 0 & 12 & 33/2 \\ 1 & 0 & -3 & -15/4 \\ 0 & 1 & 6 & 5 \end{bmatrix}.$$

This matrix has right eigenvector $(-3/8, 5/4, -3/8, 1/2)^\top$ corresponding to the eigenvalue 3, which is x evaluated at z_2. This represents the Lagrange polynomial

$$\ell_2 = \frac{-3}{8}x + \frac{5}{4}y - \frac{3}{8}x^2 + \frac{1}{2}xy.$$

△

3.1.2 Genericity and Bézout's theorem

Throughout this thesis we will work with polynomial systems on which we make certain *genericity assumptions*. More specifically, we usually assume that the polynomial system belongs to some *family* of polynomial systems, and it has the properties of a *general* or *generic member* of the family. We have already encountered some examples of genericity assumptions. In Lemma 3.1.2 we considered a linear polynomial $h = h_1x_1 + \cdots + h_nx_n$ from the *family of all linear polynomials* satisfying the condition of Lemma 3.1.1. The proof of Lemma 3.1.1 showed that *almost all* members of the family satisfy this condition. In our definition of degree for a projective variety (Definition 2.2.10) we considered 'general linear subvarieties of codimension k'. These correspond to general members of the family of polynomial systems given by k linear equations.

Definition 3.1.3 (Families and genericity)**.** Let R be a polynomial ring over $\mathbb{C}$ and let $W_1, \ldots, W_s \subset R$ be finite dimensional $\mathbb{C}$-vector subspaces of R. For some $p \in \mathbb{N}$, let

$$\phi : \mathbb{C}^p \to W_1 \times \cdots \times W_s$$

be a morphism ($W_1 \times \cdots \times W_s$ is thought of as an affine variety). We think of an element in $\operatorname{im} \phi$ as a polynomial system given by $f_1 = \cdots = f_s = 0$ where $(f_1, \ldots, f_s) = \phi(a)$ for some $a \in \mathbb{C}^p$. We say that the image of ϕ is a *family of polynomial systems parametrized by* $\mathbb{C}^p$. A property of a polynomial system is said to hold for a *generic* or *general* member of the family $\operatorname{im} \phi$ if there is a nonzero polynomial $f \in \mathbb{C}[\mathbb{C}^p]$ such that the property holds for all $\phi(a)$ with $a \in \mathbb{C}^p \setminus V_{\mathbb{C}^p}(f)$.

We note that if property A and property B hold for a generic member of a family, then so does property 'A *and* B' (the intersection of two nonempty open subsets of $\mathbb{C}^p$ is again open and nonempty). Working over the complex numbers allows us to think of many of the properties of polynomial systems we are interested in as *generic properties*. An important example is the *number of solutions* of the system. Here is an example for $n = 1$.

Example 3.1.3. Consider the family of polynomials given by $\phi : \mathbb{C}^3 \to \mathbb{C}[x]_{\leq 2}$ given by

$$\phi(a, b, c) = ax^2 + bx + c.$$

Generically, a member of this family has two solutions in $\mathbb{C}$. Indeed, $\phi(a, b, c)$ has two solutions unless $f(a, b, c) = a(b^2 - 4ac) = 0$. It is also true that a general member of this family has two solutions in $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$. This happens whenever $ac(b^2 - 4ac) \neq 0$. △

To give examples for larger n, we need to specify which family of systems we want to consider. A first example of a family of multivariate polynomial systems is the family of so-called *total degree* systems. As in Subsection 2.2.5, let

$$R_{\leq d} = \left\{ \sum_a c_a x^a \in R \mid \max_{c_a \neq 0} |a| \leq d \right\}.$$

Definition 3.1.4 (Total degree systems)**.** For an ordered tuple $(d_1, \ldots, d_s) \in \mathbb{N}^s$, the *family of total degree polynomial systems of degree* $(d_1, \ldots, d_s)$ is the image of

$$\phi : \mathbb{C}^{p_1} \times \cdots \times \mathbb{C}^{p_s} \to R_{\leq d_1} \times \cdots \times R_{\leq d_s}, \quad \text{where } p_i = \binom{n + d_i}{n}$$

and $\phi((c_{1,a})_{|a| \leq d_1}, \ldots, (c_{s,a})_{|a| \leq d_s}) = \left(\sum_{|a| \leq d_1} c_{1,a} x^a, \ldots, \sum_{|a| \leq d_s} c_{s,a} x^a \right)$. Here $|a| \leq d_i$ means that a runs over all tuples $a = (a_1, \ldots, a_n) \in \mathbb{N}^n$ satisfying $|a| = a_1 + \cdots + a_n \leq d_i$. We will denote this family by

$$\mathcal{F}_R(d_1, \ldots, d_s) = \operatorname{im} \phi = R_{\leq d_1} \times \cdots \times R_{\leq d_s}.$$

When $n = s$, the family $\mathcal{F}_R(d_1, \ldots, d_n)$ is called a *family of square total degree systems.* An important property that holds for general members $(f_1, \ldots, f_n) \in \mathcal{F}_R(d_1, \ldots, d_n)$ is given by *Bézout's theorem in* $\mathbb{C}^n$.

Theorem 3.1.2 (Bézout's theorem in $\mathbb{C}^n$)**.** *For* any *member* $(f_1, \ldots, f_n) \in \mathcal{F}_R(d_1, \ldots, d_n)$ *we have that the number of isolated points in* $V(f_1, \ldots, f_n)$*, counted with multiplicities (see Subsection 3.1.3), is bounded by* $\prod_{i=1}^n d_i$*. For a* general *member* $(f_1, \ldots, f_n) \in \mathcal{F}_R(d_1, \ldots, d_n)$ *we have that*

1. *the affine variety* $V(f_1, \ldots, f_n) \subset \mathbb{C}^n$ *consists of finitely many points,*
2. *the ideal* $\langle f_1, \ldots, f_n \rangle$ *is radical,*

3. the number of points in $V(f_1,\dots,f_n)$, counting multiplicities, is $\prod_{i=1}^n d_i$.

Note that the theorem implies that a general member of $\mathcal{F}_R(d_1,\dots,d_n)$ has $\prod_{i=1}^n d_i$ isolated solutions, all these solutions have multiplicity one and there are no positive dimensional components. We omit the proof of this theorem for now and we will say more about this result in the projective setting in Section 3.2. The theory of resultants will allow us to describe exactly when the generic properties of Theorem 3.1.2 fail to hold.

Remark 3.1.2. When $s < n$ and $d_i > 0, i = 1,\dots,s$, we have that for a general member $(f_1,\dots,f_s) \in \mathcal{F}_R(d_1,\dots,d_s)$, $\dim V_{\mathbb{C}^n}(f_1,\dots,f_s) = n - s$. When $s > n$, a general member has no solutions: $V_{\mathbb{C}^n}(f_1,\dots,f_s) = \varnothing$, which implies by the Nullstellensatz that $\langle f_1,\dots,f_s\rangle = R$. △

Example 3.1.4. The system in Example 3.1.2 is a general member of $\mathcal{F}_R(2,2)$, in the sense that all three generic properties of Theorem 3.1.2 are satisfied. △

3.1.3 Multiplicity

In this subsection, our aim is to generalize the results from Subsection 3.1.1 to the case where I is zero-dimensional, but not necessarily radical. An example for $n = 1$ gives us an idea of what to expect.

Example 3.1.5. Let $R = \mathbb{C}[x]$ and $I = \langle f\rangle$ where $f = x^2(x-1)$. Note that $I \subsetneq \sqrt{I}$, since $g = x(x-1) \notin I$ but $g^2 \in I$. The variety $V(f)$ consists of $\delta = 2$ points $\{0,1\}$. However, the dimension $\dim_{\mathbb{C}} R/I = 3$: the residue classes $1 + I, x + I, x^2 + I$ are $\mathbb{C}$-linearly independent in R/I and they generate R/I over $\mathbb{C}$. The reason for this discrepancy is that the point 0 in this example should be counted *twice*. That is, the point 0 has *multiplicity* 2 as a root of f. One way to see this is by decomposing R/I into smaller rings, each of which 'contributes' one root to $V(I)$. Observe that $I = \langle x^2\rangle \cap \langle x-1\rangle$ and $\langle x^2\rangle$ and $\langle x-1\rangle$ are *coprime ideals* since $x^2 - (x-1)(x+1) = 1$. By the *Chinese remainder theorem* (Theorem A.1.3) the map

$$R/I \to R/\langle x^2\rangle \times R/\langle x-1\rangle \quad \text{given by} \quad f + I \mapsto (f + \langle x^2\rangle, f + \langle x-1\rangle)$$

is an isomorphism. This shows that

$$\dim_{\mathbb{C}} R/I = \dim_{\mathbb{C}} R/\langle x^2\rangle + \dim_{\mathbb{C}} R/\langle x-1\rangle = 2 + 1,$$

where the root 0 contributes the term 2 in this sum. △

For general n, if $V(I) = \{z_1,\dots,z_\delta\}$ the Nullstellensatz tells us that

$$\sqrt{I} = \mathfrak{p}_1 \cap \cdots \cap \mathfrak{p}_\delta \tag{3.1.2}$$

where $\mathfrak{p}_i$ is the maximal ideal for which $V(\mathfrak{p}_i) = z_i$. Since $\mathfrak{p}_i + \mathfrak{p}_j = R$ for $i \neq j$, we can apply the Chinese remainder theorem to write

$$R/\sqrt{I} \simeq R/\mathfrak{p}_1 \times \cdots \times R/\mathfrak{p}_\delta \simeq \mathbb{C} \times \cdots \times \mathbb{C} \simeq \mathbb{C}^\delta.$$

The decomposition (3.1.2) of $\sqrt{I}$ into prime (in this case, maximal) ideals corresponds to the decomposition of $V(I)$ into irreducible varieties. The generalization of this operation for arbitrary ideals is given by the *primary decomposition* (see Theorem A.1.2). In our case, the primary decomposition writes I as an intersection

$$I = Q_1 \cap \cdots \cap Q_\delta \tag{3.1.3}$$

where the Q_i are primary ideals such that $\sqrt{Q_i} = \mathfrak{p}_i, i = 1, \ldots, \delta$. We say that Q_i is $\mathfrak{p}_i$*-primary.* Since $V(Q_i + Q_j) = \varnothing, i \neq j$, we have that the primary ideals $Q_1, \ldots, Q_s$ are pairwise coprime. By the Chinese remainder theorem this gives

$$R/I \simeq R/Q_1 \times \cdots \times R/Q_\delta. \tag{3.1.4}$$

We are now ready to define the *multiplicity* of the points in $V(I)$, generalizing the observations of Example 3.1.5.

Definition 3.1.5. Let $I \subset R$ be a zero-dimensional ideal with $V(I) = \{z_1, \ldots, z_\delta\} \subset \mathbb{C}^n$. Let $\mathfrak{p}_i = I(\{z_i\}), i = 1, \ldots, \delta$ be the corresponding maximal ideals of R and consider the primary decomposition $I = Q_1 \cap \cdots \cap Q_\delta$ such that Q_i is $\mathfrak{p}_i$-primary. For each i, the *multiplicity* μ_i of the point z_i as a solution of I is given by

$$\mu_i = \dim_{\mathbb{C}} R/Q_i.$$

We denote $\delta^+ = \mu_1 + \cdots + \mu_\delta = \dim_{\mathbb{C}} R/I$. Recall that in the case where $I = \sqrt{I}$, $\mu_i = 1, i = 1, \ldots, \delta$ and

$$f \in I \quad \Longleftrightarrow \quad \mathrm{ev}_{z_i}(f + I) = f(z_i) = 0,\ i = 1, \ldots, \delta.$$

In words, to check whether $f \in I$, it is enough to check whether f vanishes at all points of $V(I)$. In the case where $n = 1$ and I is not necessarily radical, the multiplicities of $z_1, \ldots, z_\delta$ impose vanishing conditions on the derivatives of f in order for f to be in the ideal:

$$f \in I \quad \Longleftrightarrow \quad \frac{d^\ell f}{dx^\ell}(z_i) = 0,\ \ell = 0, \ldots, \mu_i - 1,\ i = 1, \ldots, \delta.$$

This generalizes nicely for general n: the decomposition (3.1.4) of the algebra R/I gives a way of writing the condition $f \in I$ in terms of the vanishing of some *differential operators.* We now describe how this works.

For an n-tuple $a = (a_1, \ldots, a_n) \in \mathbb{N}^n$ we define the $\mathbb{C}$-linear map $\partial_a : R \to R$ given by

$$\partial_a(f) = \frac{1}{a_1! \cdots a_n!} \frac{\partial^{a_1 + \cdots + a_n} f}{\partial x_1^{a_1} \cdots \partial x_n^{a_n}}.$$

These differential operators generate the $\mathbb{C}$-vector space

$$\mathscr{D} = \left\{ \sum_{a \in \mathbb{N}^n} c_a \partial_a \mid \text{finitely many } c_a \text{ are nonzero} \right\}.$$

For each $a \in \mathbb{N}^n$, we also define the *antidifferentiation operator* $s_a : \mathscr{D} \to \mathscr{D}$ by

$$s_a \left(\sum_b c_b \partial_b \right) = \sum_{b-a \geq 0} c_b \partial_{b-a},$$

where the sum on the right hand side ranges over all $b = (b_1, \ldots, b_n) \in \mathbb{N}^n$ such that $b_i - a_i \geq 0, i = 1, \ldots, n$. These operators allow for a very simple formulation of Leibniz' rule, which says that for $\partial \in \mathscr{D}$,

$$\partial(fg) = \sum_{b \in \mathbb{N}^n} \partial_b(g)(s_b(\partial))(f). \tag{3.1.5}$$

Definition 3.1.6. A $\mathbb{C}$-vector subspace $D \subset \mathscr{D}$ is *closed* if $\dim_{\mathbb{C}}(D) < \infty$ and for each $\partial \in D$ and each $a \in \mathbb{N}^n$, $s_a(\partial) \in D$.

Note that if $D \subset \mathscr{D}$ is closed, then $\partial_0 = \mathrm{id}_R \in D$ (here id_R is our notation for the identity map $f \mapsto f$ on R). The motivation for defining closed subsets of $\mathscr{D}$ in this way is the fact that they 'annihilate' zero-dimensional primary ideals of R.

Theorem 3.1.3. *Let $z = (z_1, \ldots, z_n) \in \mathbb{C}^n$. There is a one-to-one correspondence between $\langle x - z_1, \ldots, x - z_n \rangle$-primary ideals Q of R and closed subspaces D of $\mathscr{D}$. Explicitly, the correspondence is given by*

$$Q \mapsto \{\partial \in \mathscr{D} \mid \partial(f)(z) = 0, \textit{ for all } f \in Q\}$$

and

$$D \mapsto \{f \in R \mid \partial(f)(z) = 0, \textit{ for all } \partial \in D\}.$$

Moreover, we have that $\dim_{\mathbb{C}} D = \dim_{\mathbb{C}} R/Q$.

Proof. See [MMM93, Theorem 2.6]. □

It follows from Theorem 3.1.3 that the ideals Q_i from (3.1.4) give closed subspaces

$$D_i = \{\partial \in \mathscr{D} \mid \partial(f)(z_i) = 0, \text{ for all } f \in Q_i\}.$$

Note that any $\partial \in D_i$ gives a well-defined functional

$$\mathrm{ev}_{z_i} \circ \partial : R/I \to \mathbb{C} \quad \text{with} \quad (\mathrm{ev}_{z_i} \circ \partial)(f + I) = \partial(f)(z_i).$$

This follows from the fact that D_i can be identified with $(R/Q_i)^\vee \subset (R/I)^\vee$ via

$$\partial \mapsto (f + Q_i \mapsto \partial(f)(z_i)).$$

In particular, the theorem also implies that $\dim_{\mathbb{C}} D_i = \mu_i$. For a differential operator $\partial = \sum_a c_a \partial_a \in \mathscr{D}$ we define $\operatorname{ord}(\partial) = \max_{c_a \neq 0} |a|$. We denote by $(D_i)_{\leq d} = \{\partial \in D_i \mid \operatorname{ord}(\partial) \leq d\}$ the subspace of differential operators in D_i of order bounded by d. For giving explicit descriptions of the eigenstructure of multiplication maps (defined below), it is convenient to work with a special type of basis for the spaces D_i (see [MS95, Section 5]).

Definition 3.1.7. An ordered tuple $(\partial_{i1}, \ldots, \partial_{i\mu_i})$ with $\partial_{ij} \in D_i$ is called a *consistently ordered basis* for D_i if for every $d \geq 0$ there is j_d such that $\{\partial_{i1}, \ldots, \partial_{ij_d}\}$ is a $\mathbb{C}$-vector space basis for $(D_i)_{\leq d}$.

Note that a consistently ordered basis always exists for any closed subspace D, its first differential operator is always ∂_0 and it is a $\mathbb{C}$-vector space basis for D.

Lemma 3.1.3. *For $i = 1, \ldots, \delta$, let $(\partial_{i1}, \ldots, \partial_{i\mu_i})$ be a consistently ordered basis for D_i. The linear map $R/I \to \mathbb{C}^{\delta^+}$ given by*

$$f + I \mapsto ((\mathrm{ev}_{z_1} \circ \partial_{11})(f), \ldots, (\mathrm{ev}_{z_1} \circ \partial_{1\mu_1})(f), \ldots, (\mathrm{ev}_{z_\delta} \circ \partial_{\delta 1})(f), \ldots, (\mathrm{ev}_{z_\delta} \circ \partial_{\delta \mu_\delta})(f))$$

is an isomorphism of vector spaces.

Proof. The map is injective because $f \in I \Leftrightarrow f \in Q_1 \cap \cdots \cap Q_\delta$, which is equivalent to $(\mathrm{ev}_{z_i} \circ \partial)(f) = 0, \forall \partial \in D_i, i = 1, \ldots, \delta$. The lemma follows since $\dim_{\mathbb{C}} R/I = \delta^+$. □

Note that if $I = \sqrt{I}$, the map from Lemma 3.1.3 is the map ψ from Proposition 3.1.1. As in Lemma 3.1.3, for $i = 1, \ldots, \delta$, let $(\partial_{i1}, \ldots, \partial_{i\mu_i})$ be a consistently ordered basis for D_i. Note that by Leibniz' rule, for all $f + I \in R/I$ we have

$$\begin{aligned} ((\mathrm{ev}_{z_i} \circ \partial_{ij}) \circ M_g)(f + I) &= \mathrm{ev}_{z_i}(\partial_{ij}(fg) + I) \\ &= \mathrm{ev}_{z_i}\left(\sum_{b \in \mathbb{N}^n} \partial_b(g) s_b(\partial_{ij})(f) + I\right) \qquad (3.1.6) \\ &= \sum_{b \in \mathbb{N}^n} \partial_b(g)(z_i) \cdot (\mathrm{ev}_{z_i} \circ s_b(\partial_{ij}))(f + I). \qquad (3.1.7) \end{aligned}$$

In particular, for $\partial_{i1} = \partial_0 = \mathrm{id}_R$ we get

$$\mathrm{ev}_{z_i} \circ M_g = g(z_i)\, \mathrm{ev}_{z_i},$$

which shows that the evaluation functionals ev_{z_i} are (left) eigenvectors of M_g with eigenvalues $g(z_i)$. In general, by the property of being closed, $s_b(\partial_{ij})$ can be written as a $\mathbb{C}$-linear combination of $\partial_{i1}, \ldots, \partial_{i\mu_i}$. For $b \neq 0$, by the property of being consistently ordered and $\operatorname{ord}(s_b(\partial)) < \operatorname{ord}(\partial)$, $s_b(\partial_{ij})$ can be written as a $\mathbb{C}$-linear combination of $\partial_{i1}, \ldots, \partial_{i,j-1}$ (in fact, we only need the differentials of order strictly lower than $\operatorname{ord}(\partial_{ij})$). Therefore, we can write

$$\sum_{b \in \mathbb{N}^n} \partial_b(g)(z_i) \cdot (\mathrm{ev}_{z_i} \circ s_b(\partial_{ij})) = g(z_i)(\mathrm{ev}_{z_i} \circ \partial_{ij}) + \sum_{k=1}^{j-1} c_{ij}^{(k)} (\mathrm{ev}_{z_i} \circ \partial_{ik}).$$

Then, in matrix notation, (3.1.6) becomes

$$D_i \circ M_g = \begin{bmatrix} \mathrm{ev}_{z_i} \circ \partial_{i1} \\ \mathrm{ev}_{z_i} \circ \partial_{i2} \\ \vdots \\ \mathrm{ev}_{z_i} \circ \partial_{i\mu_i} \end{bmatrix} \circ M_g = \begin{bmatrix} g(z_i) & & & \\ c_{i2}^{(1)} & g(z_i) & & \\ \vdots & & \ddots & \\ c_{i\mu_i}^{(1)} & c_{i\mu_i}^{(2)} & \dots & g(z_i) \end{bmatrix} \begin{bmatrix} \mathrm{ev}_{z_i} \circ \partial_{i1} \\ \mathrm{ev}_{z_i} \circ \partial_{i2} \\ \vdots \\ \mathrm{ev}_{z_i} \circ \partial_{i\mu_i} \end{bmatrix} = L_i \circ D_i. \quad (3.1.8)$$

Here the notation D_i is (ab-)used for the linear map represented by a consistently ordered basis for D_i composed with ev_{z_i}. Putting the equations (3.1.8) together for $i = 1, \dots, \delta$ we get

$$\underbrace{\begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_\delta \end{bmatrix}}_{D} \circ M_g = \underbrace{\begin{bmatrix} L_1 & & & \\ & L_2 & & \\ & & \ddots & \\ & & & L_\delta \end{bmatrix}}_{L} \circ \underbrace{\begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_\delta \end{bmatrix}}_{D}. \quad (3.1.9)$$

By observing that the map D in (3.1.9) is exactly the map from Lemma 3.1.3, we get that any matrix representation of M_g is similar to the lower triangular matrix L, whose diagonal is

$$\underbrace{g(z_1), \dots, g(z_1)}_{\mu_1 \text{ times}}, \dots, \underbrace{g(z_\delta), \dots, g(z_\delta)}_{\mu_\delta \text{ times}}.$$

The following theorem follows easily.

Theorem 3.1.4. *For any matrix representation of the multiplication map $M_g : R/I \to R/I$, we have that*

$$\det(\lambda \, \mathrm{id}_{\mathbb{C}^{\delta^+}} - M_g) = \prod_{i=1}^{\delta} (\lambda - g(z_i))^{\mu_i}.$$

Remark 3.1.3. Describing the multiplicity structure by means of differential operators has the advantage that it gives a very explicit description of the invariant subspaces of the multiplication operators. An alternative way of decomposing the algebra R/I into subalgebras coming from the different points in $V(I)$ is via localization. This is the approach taken in, for instance, [CLO06, Chapter 4, §2]. The key idea is to establish an isomorphism

$$R/I \to R_{\mathfrak{p}_1}/IR_{\mathfrak{p}_1} \times \dots \times R_{\mathfrak{p}_\delta}/IR_{\mathfrak{p}_\delta},$$

where $R_{\mathfrak{p}_i}$ is the localization of R at the maximal ideal $\mathfrak{p}_i = I(\{z_i\})$ (see Subsection A.1.4). The equivalence of the approaches follows from the exact sequence

$$0 \to Q_i \to R \to R_{\mathfrak{p}_i}/IR_{\mathfrak{p}_i} \to 0,$$

from which $R/Q_i \simeq R_{\mathfrak{p}_i}/IR_{\mathfrak{p}_i}$. This is discussed in [CLO06, Chapter 4, §2, Exercise 11]. △

Remark 3.1.4. A solution $z_j \in V(I) = V(f_1, \dots, f_s)$ has multiplicity $\mu_j > 1$ if and only if there is a differential operator $\partial = \sum_{i=1}^n c_i \partial_{e_i} \in \mathscr{D}$ with $\mathrm{ord}(\partial) = 1$ such that $\partial \in D_j$. This is equivalent to the condition that $\partial(f_i)(z_j) = 0$ for $i = 1, \dots, s$, which means that the Jacobian

$$J(z_i) = \left(\frac{\partial f_k}{\partial x_\ell}(z_i) \right)_{1 \le k \le s, 1 \le \ell \le n}$$

has the vector $c = (c_1, \dots, c_n)^\top$ in its kernel: $J(z_i)c = 0$. In particular, if $n = s$, the root z_i has multiplicity $\mu_i > 1$ if and only if $\det J(z_i) = 0$. $\triangle$

Given an isolated point $z_i \in V(I)$, there is a numerical linear algebra based algorithm for computing a basis of D_i [DZ05]. A description of this algorithm is outside the scope of this thesis.

Example 3.1.6. Consider the ideal $I = \langle f_1, f_2 \rangle \subset R = \mathbb{C}[x, y]$ generated by

$$f_1 = x + \frac{1}{3}y^2 - x^2, \quad f_2 = \frac{-1}{3}x + \frac{1}{3}x^2.$$

The variety $V(I) = \{z_1, z_2\}$ consists of the two points $z_1 = (1, 0), z_2 = (0, 0)$. One can easily check that

$$(\mathrm{ev}_{z_j} \circ \partial_{(0,1)})(f_i) = \frac{\partial f_i}{\partial y}(z_j) = 0, \quad i = 1, 2, \ j = 1, 2.$$

It follows that V_1, V_2 have at least dimension two, and by Bézout's theorem (Theorem 3.1.2), the sum of these dimensions is at most 4. We conclude that $\{\partial_{(0,0)}, \partial_{(0,1)}\} \subset \mathscr{D}$ is a basis for D_1 as well as for D_2. In the algebra R/I we have the equalities

$$y^2 + I = 0 + I, \quad x^2 + I = x + I,$$

and $\mathcal{B} = \{1 + I, y + I, xy + I, x^2 + I\}$ is a $\mathbb{C}$-basis for R/I. Using the basis $\mathcal{B}$ with its elements in this order we find that 'multiplication with y' is given by

$$M_y = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The matrix D from (3.1.9) is given by

$$D = \begin{bmatrix} D_1 \\ D_2 \end{bmatrix} = \begin{bmatrix} \mathrm{ev}_{z_1} \circ \partial_{(0,0)} \\ \mathrm{ev}_{z_1} \circ \partial_{(0,1)} \\ \mathrm{ev}_{z_2} \circ \partial_{(0,0)} \\ \mathrm{ev}_{z_2} \circ \partial_{(0,1)} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Note that D is indeed invertible (Lemma 3.1.3). For $j = 1, 2$ and any $g \in R$ we have

$$(\mathrm{ev}_{z_j} \circ \partial_{(0,0)}) \circ M_g(f + I) = g(z_j)(\mathrm{ev}_{z_j} \circ \partial_{(0,0)})(f + I),$$

$$(\mathrm{ev}_{z_j} \circ \partial_{(0,1)}) \circ M_g(f + I) = g(z_j)(\mathrm{ev}_{z_j} \circ \partial_{(0,1)})(f + I) + \frac{\partial g}{\partial y}(z_j)(\mathrm{ev}_{z_j} \circ \partial_{(0,0)})(f + I).$$

In matrix notation, this gives $DM_g = LD$ where

$$L = \begin{bmatrix} g(z_1) & & & \\ \frac{\partial g}{\partial y}(z_1) & g(z_1) & & \\ & & g(z_2) & \\ & & \frac{\partial g}{\partial y}(z_2) & g(z_2) \end{bmatrix}.$$

In particular, for $g = y$ this gives

$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

△

3.2 Points in projective space

In this section, we work in the $\mathbb{Z}$-graded ring $S = \mathbb{C}[x_0, \ldots, x_n]$ and consider zero-dimensional homogeneous ideals of S (see Section 2.2). These are the homogeneous ideals $I \subset S$ such that $V_{\mathbb{P}^n}(I) = \{\zeta_1, \ldots, \zeta_\delta\}$ consists of finitely many points. Each of the points $\zeta_i \in V_{\mathbb{P}^n}(I)$ can be represented by a set of *homogeneous coordinates* $z_i = (z_{i0}, \ldots, z_{in}) \in \mathbb{C}^{n+1} \setminus \{0\}$ such that $\zeta_i = (z_{i0} : \ldots : z_{in})$ and $z_i \in V_{\mathbb{C}^{n+1}}(I)$. Our motivation for studying zero-dimensional homogeneous ideals is twofold. Firstly, the solutions of some problems coming from applications have a natural interpretation as points in $\mathbb{P}^n$. Think for instance about the case where solutions are elements in the kernel of some matrix, eigenvectors of a (nonlinear) eigenvalue problem [GT17] or conics in $\mathbb{P}^2$ [BST19]. Secondly, it is sometimes beneficial to *reinterpret* equations on $\mathbb{C}^n$ as equations on $\mathbb{P}^n$ via a process called *homogenization*. After describing some basic properties of zero-dimensional homogeneous ideals and formulating a projective eigenvalue, eigenvector theorem in Subsections 3.2.1 and 3.2.2, we will discuss homogenization in Subsection 3.2.3.

3.2.1 The Hilbert function and Bézout's theorem

Let $I = \langle f_1, \ldots, f_s \rangle \subset S$ be a zero-dimensional homogeneous ideal with $V_{\mathbb{P}^n}(I) = \{\zeta_1, \ldots, \zeta_\delta\}$ and such that $d_i = \deg(f_i), i = 1, \ldots, s$. Our goal in this subsection is to say something more about the *expected value* of δ in this setting. In the language of Subsection 3.1.2, we want to understand the number of solutions of a general member of the following family of homogeneous polynomial systems.

Definition 3.2.1 (Homogeneous systems)**.** For an ordered tuple $(d_1, \dots, d_s) \in \mathbb{N}^s$, the *family of homogeneous polynomial systems of degree* $(d_1, \dots, d_s)$ is the image of

$$\phi : \mathbb{C}^{p_1} \times \cdots \times \mathbb{C}^{p_s} \to S_{d_1} \times \cdots \times S_{d_s}, \quad \text{where } p_i = \binom{n + d_i}{n}$$

and $\phi((c_{1,a})_{|a|=d_1}, \dots, (c_{s,a})_{|a|=d_s}) = \left(\sum_{|a|=d_1} c_{1,a}x^a, \dots, \sum_{|a|=d_s} c_{s,a}x^a\right)$. Here $|a| = d_i$ means that a runs over all tuples $a = (a_0, a_1, \dots, a_n) \in \mathbb{N}^{n+1}$ satisfying $|a| = a_0 + a_1 + \cdots + a_n = d_i$. We will denote this family by

$$\mathcal{F}_S(d_1, \dots, d_s) = \operatorname{im} \phi = S_{d_1} \times \cdots \times S_{d_s}.$$

The most interesting scenario happens when $n = s$, which is the case covered by *Bézout's theorem in projective space*. The tool we will use for understanding this theorem is the *Hilbert function*, see Subsection 2.2.7.

First, we define the concept of *multiplicity* for a point in $V_{\mathbb{P}^n}(I)$. We do this by restricting the equations to an affine chart. As in Section 2.2, let

$$U_i = \{(x_0 : \cdots : x_n) \in \mathbb{P}^n \mid x_i \neq 0\} \simeq \mathbb{C}^n.$$

A first observation is that for $i = 0, \dots, n$, the ideal I gives an ideal

$$\mathscr{I}(U_i) \subset \mathscr{O}_{\mathbb{P}^n}(U_i) = \mathbb{C}\left[\frac{x_0}{x_i}, \dots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \dots, \frac{x_n}{x_i}\right] = \mathbb{C}[y_0, \dots, y_{i-1}, y_{i+1}, \dots, y_n]$$

by *dehomogenization*. Here's how this works. For $j = 1, \dots, s$ let

$$\hat{f}_{ij} = \eta_{d_j}^{-1}(f_j) = f_j(y_0, \dots, y_{i-1}, 1, y_{i+1}, \dots, y_n),$$

where $\eta_{d_j} : \mathscr{O}_{\mathbb{P}^n}(U_i)_{\leq d_j} \to S_{d_j}$ is the homogenization isomorphism (see Subsection 2.2.4). We define $\mathscr{I}(U_i) = \langle \hat{f}_{i1}, \dots, \hat{f}_{is} \rangle$. Note that the polynomials f_j do *not* define functions on $\mathbb{P}^n$, but the functions $\hat{f}_{ij}$ *do* define functions on U_i and on the overlaps $U_i \cap U_k, k \neq i$, the functions $\hat{f}_{ij}$ and $\hat{f}_{kj}$ agree on where they are zero.[1] Indeed, for $x \in U_i \cap U_k$ we have

$$\hat{f}_{ij}(x) = \left(\frac{x_k}{x_i}\right)^{d_j} \hat{f}_{ik}(x),$$

where it should be clear that $\hat{f}_{ij}(x) = \hat{f}_{ij}(x_0/x_i, \dots, x_{i-1}/x_i, x_{i+1}/x_i, \dots, x_n/x_i)$, and the analogous notation is used for $\hat{f}_{ik}$.

The points $\zeta_j \in V_{\mathbb{P}^n}(I)$ can be assigned a multiplicity as in the affine case (see Subsection 3.1.3). The multiplicity of a point is defined locally, so for some affine chart $U_i \subset \mathbb{P}^n$ containing ζ_j, we can define the multiplicity μ_j of ζ_j as the multiplicity of

[1] For the reader who is familiar with vector bundles, we are describing f_j as a global section of the line bundle with sheaf of sections $\mathscr{O}_{\mathbb{P}^n}(d_j)$ on $\mathbb{P}^n$ with transition functions $(x_k/x_i)^{d_j}$. The tuple $(f_1, \dots, f_s)$ can be seen as a global section of the rank s algebraic vector bundle with sheaf of sections $\mathscr{O}_{\mathbb{P}^n}(d_1) \oplus \cdots \oplus \mathscr{O}_{\mathbb{P}^n}(d_s)$.

this point as a solution of $\mathscr{I}(U_i)$. This is independent of the choice of U_i containing ζ_j. We do not go into detail here.
Another concept which we have to introduce before talking about Hilbert functions is that of *saturation* with respect to the irrelevant ideal. Recall from Section 2.2 that the irrelevant ideal $\mathfrak{B} = \langle x_0, \ldots, x_n \rangle$ plays a special role in our graded ring: it is a proper ideal whose projective variety is the empty set. Here's an example of the kind of issues that this causes, similar to Remark 2.2.1 but for a nonempty projective variety.

Example 3.2.1. Let $S = \mathbb{C}[x_0, x_1]$ and consider $I = \langle x_0x_1, x_1^2 \rangle$ with $V_{\mathbb{P}^1}(I) = \{(1:0)\}$. Dehomogenizing this to the chart U_0 where $x_0 \neq 0$, we get the ideal $\mathscr{I}(U_0) = \langle y_1, y_1^2 \rangle = \langle y_1 \rangle \subset \mathbb{C}[y_1]$, which shows that the point $(1:0)$ has multiplicity 1. Therefore, the geometric object associated to I is *exactly the same* as the one associated to $\langle x_1 \rangle \subset S$, which is a strictly larger ideal of S. Note that $\mathscr{I}(U_1) = \mathscr{O}_{\mathbb{P}^1}(U_1)$, which reflects the fact that there are no points in $V_{\mathbb{P}^1}(I) \cap U_1$. △

The reason for the ambiguity in Example 3.2.1 is that the *affine scheme* defined by $\langle x_0x_1, x_1^2 \rangle$ in $\mathbb{C}^2$ consists of the line $x_1 = 0$ with an 'extra', 'distinguished', or *embedded point* at the origin. Think for instance of $\langle x_0x_1, x_1^2 \rangle$ as the limit of $\langle (x_0 - t)x_1, x_1^2 \rangle$ for $t \to 0$. This embedded point is no longer visible when moving to projective space. A remedy for this is provided by 'dividing the ideal $\mathfrak{B}$ out'. This is a process called *saturation.*

Definition 3.2.2 (Saturation)**.** For a homogeneous ideal $I \subset S$, the *saturation* of I (with respect to $\mathfrak{B}$) is the homogeneous ideal

$$(I : \mathfrak{B}^\infty) = \{f \in S \mid \text{for all } b \in \mathfrak{B},\ b^\ell f \in I \text{ for some } \ell \in \mathbb{N}\} \subset S.$$

If $I = (I : \mathfrak{B}^\infty)$, we say that I is ($\mathfrak{B}$-)*saturated.*

For any homogeneous ideal $I \subset S$, there is some $\ell \in \mathbb{N}$ such that the saturation of I equals the ideal quotient

$$(I : \mathfrak{B}^\infty) = (I : \mathfrak{B}^\ell)$$

of I by the ideal $\mathfrak{B}^\ell = \langle b_1 \cdots b_\ell \mid b_i \in \mathfrak{B}, i = 1, \ldots, \ell \rangle = \langle S_\ell \rangle$ (see [CLO13, Chapter 4, §4, Proposition 9]). The fact that the ideals I and $(I : \mathfrak{B}^\infty)$ carry the same geometric information is reflected in their behavior for high degrees.

Proposition 3.2.1. *Let $I \subset S$ be a homogeneous ideal. For some $\ell \in \mathbb{N}$, we have that*

$$I_d = (I : \mathfrak{B}^\infty)_d, \quad d \geq \ell.$$

Proof. The inclusion $I \subset (I : \mathfrak{B}^\infty)$ is clear (in all degrees). For the opposite inclusion, let $\hat{\ell}$ be such that $(I : \mathfrak{B}^\infty) = (I : \mathfrak{B}^{\hat{\ell}})$. Since S is Noetherian, $(I : \mathfrak{B}^\infty) = \langle g_1, \ldots, g_{s'} \rangle$ is finitely generated, where we can take g_i homogeneous of degree d_i. Take $\ell \in \mathbb{N}$ such that $\ell = \max_{i=1,\ldots,s'} \hat{\ell} + d_i$. Then

$$(I : \mathfrak{B}^\infty)_\ell = \{h_1 g_1 + \cdots + h_{s'} g_{s'} \mid h_i \in S_{\ell - d_i}\}$$

and since $\ell - d_i \geq \hat{\ell}, \ i = 1, \ldots, s'$, we have for each $f = h_1 g_1 + \cdots + h_{s'} g_{s'} \in (I : \mathfrak{B}^\infty)_\ell$ that $f \in I_\ell$, since $h_i \in \mathfrak{B}^{\hat{\ell}}$. □

Recall that in the affine setting, a zero-dimensional ideal is radical if and only if all the points in its variety have multiplicity 1. In the projective setting, we have to take the irrelevant ideal $\mathfrak{B}$ into account.

Proposition 3.2.2. *Let $I = \langle f_1, \ldots, f_s \rangle$ be zero-dimensional. If $V_{\mathbb{P}^n}(I) = \{\zeta_1, \ldots, \zeta_\delta\}$ with multiplicities $\mu_i = 1, i = 1, \ldots, \delta$, then $(I : \mathfrak{B}^\infty) = I_S(V_{\mathbb{P}^n}(I)) = \sqrt{(I : \mathfrak{B}^\infty)}$.*

Proof. Let $g \in (I : \mathfrak{B}^\infty)$. Without loss of generality, we may assume that g is homogeneous. By definition, we know that for some $\ell \in \mathbb{N}$ and for $i = 0, \ldots, n$, $x_i^\ell g \in I$. For all $\zeta_j \in V_{\mathbb{P}^n}(I)$, pick i such that $\zeta_j \in U_i$. Now $x_i^\ell g = h_1 f_1 + \cdots + h_s f_s$ vanishes at ζ_j, but x_i^ℓ does not. We conclude that $g(\zeta_j) = 0$, and hence $g \in I_S(V_{\mathbb{P}^n}(I))$. To prove the opposite inclusion, take $g \in I_S(V_{\mathbb{P}^n}(I))$ homogeneous. For $i = 0, \ldots, n$, let $\hat{g}_i = g(x_0/x_i, \ldots, x_{i-1}/x_i, 1, x_{i+1}/x_i, \ldots, x_n/x_i)$ be the dehomogenization. For each $\zeta_j \in U_i$, since all multiplicities are one we have

$$\hat{g}_i(\zeta_j) = 0 \quad \Rightarrow \quad \hat{g}_i \in \mathscr{I}(U_i) = \langle \hat{f}_{i1}, \ldots, \hat{f}_{is} \rangle.$$

It follows that for some $\hat{h}_i, i = 1, \ldots, s$ we can write

$$\hat{g}_i = \hat{h}_1 \hat{f}_{i1} + \cdots + \hat{h}_s \hat{f}_{is}. \tag{3.2.1}$$

There exists $\ell \in \mathbb{N}$ such that multiplying both sides of the equation (3.2.1) with x_i^ℓ clears the denominators and $\ell_i \geq \max(\deg(g), \deg(f_1), \ldots, \deg(f_s))$. Since $g = x_i^{\deg(g)} \hat{g}_i$ and $f_{ij} = x_i^{\deg(f_j)} \hat{f}_{ij}$ we find that $x_i^{\ell_i - \deg(g)} g \in I$. It follows that for $\ell = \max_{i=0,\ldots,n} \ell_i - \deg(g)$, $x_i^\ell g \in I, i = 0, \ldots, n$, which implies $g \in (I : \mathfrak{B}^\infty)$. □

The following theorem shows that for a zero-dimensional homogeneous ideal $I \subset S$, the Hilbert function HF_I stabilizes for high degrees, and it reveals the number of points in $V_{\mathbb{P}^n}$, counted with multiplicity.

Theorem 3.2.1. *Let $I \subset S$ be a $\mathfrak{B}$-saturated, zero-dimensional homogeneous ideal. Denote $V_{\mathbb{P}^n}(I) = \{\zeta_1, \ldots, \zeta_\delta\}$ where ζ_i has multiplicity μ_i and $\delta^+ = \mu_1 + \cdots + \mu_\delta$. For some $\ell \in \mathbb{N}$, the Hilbert function HF_I satisfies*

$$\mathrm{HF}_I(d) = \dim_{\mathbb{C}}(S/I)_d = \delta^+, \quad d \geq \ell.$$

Moreover, $\mathrm{HF}_I(d), d = 0, 1, 2, \ldots$ is a non-decreasing sequence.

Proof. See [EH06, Proposition III-59]. The fact that $\mathrm{HF}_I(d)$ is constant for large enough d follows from Theorem 2.2.5. □

d	0	1	2	3	4	$\cdots$
$\mathrm{HF}_I(d)$	1	2	1	1	1	$\cdots$
$\mathrm{HF}_{(I:\mathfrak{B}^\infty)}(d)$	1	1	1	1	1	$\cdots$

Table 3.1: Hilbert function of the ideals from Example 3.2.2.

Example 3.2.2. The ideal $I = \langle x_0x_1, x_1^2\rangle \subset S$ from Example 3.2.1 is not saturated: its saturation is $(I : \mathfrak{B}^\infty) = \langle x_1\rangle$. Some values of the Hilbert functions of these ideals are shown in Table 3.1. The table illustrates that HF_I stabilizes for $d \geq 2$, and $\mathrm{HF}_{(I:\mathfrak{B}^\infty)}$ stabilizes for $d \geq 0$. By Proposition 3.2.1, the Hilbert functions must agree for large enough degrees. This happens for $d = 2$ in this example: $I_2 = (I : \mathfrak{B}^\infty)_2$ is the $\mathbb{C}$-vector space spanned by x_0x_1 and x_1^2. △

An important and fascinating consequence of Theorem 3.2.1 is that if $I = \langle f_1, \ldots, f_n\rangle$ (note that $s = n$) is zero-dimensional, the number of points in $V(I)$ (counting multiplicities) *only depends on the degrees* $d_1, \ldots, d_n$ of the generators. In other words, it only depends on the family $\mathcal{F}_S(d_1, \ldots, d_n)$.

Theorem 3.2.2 (Bézout's theorem in $\mathbb{P}^n$)**.** *Let $(f_1, \ldots, f_n) \in \mathcal{F}_S(d_1, \ldots, d_n)$ be such that $I = \langle f_1, \ldots, f_n\rangle \subset S$ is zero-dimensional and $d_i > 0, i = 1, \ldots, n$. Denote $V_{\mathbb{P}^n}(I) = \{\zeta_1, \ldots, \zeta_\delta\}$ where ζ_i has multiplicity μ_i and $\delta^+ = \mu_1 + \cdots + \mu_\delta$. We have that $\delta^+ = \prod_{i=1}^n d_i$. Moreover, both the property that I is zero-dimensional and the property that $\mu_i = 1, i = 1, \ldots, \delta$ hold for general members of $\mathcal{F}_S(d_1, \ldots, d_n)$.*

Proof. The proof of this theorem will be our first application of the Koszul complex (see Subsection A.2.5). Since S is Cohen-Macaulay and $\operatorname{codim}_{\mathbb{P}^n} V_{\mathbb{P}^n}(I) = n$ is the number of homogeneous equations, $f_1, \ldots, f_n$ is a regular sequence in S, see [Ben19, Proposition 2.7.13] or the discussion in [EH06, page 144]. As a consequence (Theorem A.2.6), the augmented Koszul complex

$$\hat{\mathcal{K}}(f_1, \ldots, f_n): \quad 0 \longrightarrow K_n \xrightarrow{\phi_n} K_{n-1} \xrightarrow{\phi_{n-1}} \cdots \xrightarrow{\phi_2} K_1 \xrightarrow{\phi_1} S \longrightarrow S/I \longrightarrow 0 \quad (3.2.2)$$

where

$$K_\ell = \bigoplus_{1 \leq i_1 \leq \cdots \leq i_\ell \leq n} S(-d_{i_1} - \cdots - d_{i_\ell})$$

is exact. Also, all homomorphisms ϕ_ℓ are graded of degree 0. Restricting the sequence (3.2.2) to the degree d part and applying Theorem A.2.3 we find that

$$\mathrm{HF}_I(d) = \dim_{\mathbb{C}}(S/I)_d = \dim_{\mathbb{C}} S_d + \sum_{\ell=1}^n (-1)^\ell \dim_{\mathbb{C}}(K_\ell)_d.$$

In this formula, the dimensions of S_d and $(K_\ell)_d$ are easy to compute: these are all twisted free graded S-modules. One can work out the combinatorics (see [EH06, page

144-145]) to obtain

$$\mathrm{HF}_I(d) = \prod_{i=1}^{n} d_i, \quad d \geq d_1 + \cdots + d_n - n. \tag{3.2.3}$$

This proves the first statement. The proof of the rest of the theorem uses resultants (among other things). This is covered in [CLO06, Chapter 3, §5, Exercise 6]. □

Remark 3.2.1. There are versions of Bézout's theorem for positive dimensional solution sets. See for instance [EH06, Theorem III-71] or [Har77, Chapter I, Theorem 7.7]. △

3.2.2 Projective eigenvalue, eigenvector theorem

In this subsection, we will assume for simplicity that $I = \langle f_1, \ldots, f_s \rangle \subset S$ is a zero-dimensional ideal with $V_{\mathbb{P}^n}(I) = \{\zeta_1, \ldots, \zeta_\delta\}$ where each of the ζ_i has multiplicity one. This implies that the saturation $(I : \mathfrak{B}^\infty)$ is radical (Proposition 3.2.2). All results can be generalized to the case with arbitrary multiplicities. We would like to mimic the approach taken in Subsection 3.1.1 to construct matrices representing 'multiplication with a function' whose eigenvalues are the evaluations of that function at the points of $V_{\mathbb{P}^n}(I)$. Since the only regular functions on $\mathbb{P}^n$ are the constants, we will allow rational functions defined on $V_{\mathbb{P}^n}(I)$. A first thing to generalize is the evaluation map from Definition 3.1.1.

Definition 3.2.3 (Homogeneous evaluation maps)**.** For $d \in \mathbb{N}$ and $h \in S_d$ such that $h(\zeta_i) \neq 0, i = 1, \ldots, \delta$, we define $\mathrm{ev}_{\zeta_i} \in (S/I)_d^\vee, i = 1, \ldots, \delta$ by $\mathrm{ev}_{\zeta_i}(f + I_d) = \frac{f}{h}(\zeta_i)$. Furthermore, we define the *homogeneous evaluation map* $\psi_d : (S/I)_d \to \mathbb{C}^\delta$ by $\psi_d = (\mathrm{ev}_{\zeta_1}, \ldots, \mathrm{ev}_{\zeta_\delta})$. That is,

$$\psi_d(f + I_d) = \left(\frac{f}{h}(\zeta_1), \cdots, \frac{f}{h}(\zeta_\delta) \right).$$

The maps ψ_d are well-defined because f and h are homogeneous of the same degree and h does not vanish at any of the points ζ_i. Note that for each d it is possible to find $h \in S_d$ satisfying the condition of Definition 3.2.3. In fact, a *general* member of $\mathcal{F}_S(d)$ satisfies the condition, for all $d \in \mathbb{N}$. A crucial property of the evaluation map from Definition 3.1.1 is that it can be used to define coordinates on the (affine) coordinate ring of a set of points in $\mathbb{C}^n$. The same happens in the homogeneous case for large enough degrees. We characterize what 'large enough' means first.

Definition 3.2.4 (Regularity)**.** The *regularity* $\mathrm{Reg}(I)$ of I is defined as

$$\mathrm{Reg}(I) = \{d \in \mathbb{Z} \mid \mathrm{HF}_I(d) = \delta \text{ and } I_d = (I : \mathfrak{B}^\infty)_d\}.$$

By the results from Subsection 3.2.1, we know that there is $\ell \in \mathbb{N}$ such that $d \in \mathrm{Reg}(I)$ for all $d \geq \ell$. For the case we are most interested in, the regularity has an easy description.

Theorem 3.2.3 (Regularity for square systems)**.** *If* $I = \langle f_1, \ldots, f_n \rangle$ *with* $f_i \in S_{d_i}$, $d_i > 0$, $i = 1, \ldots, n$ *is zero-dimensional, then* $\mathrm{Reg}(I) \supset \{d \in \mathbb{Z} \mid d \geq d_1 + \cdots + d_n - n\}$.

Proof. The fact that $\mathrm{HF}_I(d) = \delta$ for $d \geq d_1 + \ldots + d_n - n$ follows from the proof of Theorem 3.2.2. The condition that $I_d = (I : \mathfrak{B}^\infty)_d$ turns out to be satisfied for all d in this case. See Theorem 5.5.10. ☐

Proposition 3.2.3. *If* $I \subset S$ *is zero-dimensional such that all points in* $V_{\mathbb{P}^n}(I)$ *have multiplicity 1, then for all* $d \in \mathrm{Reg}(I)$ *the evaluation map* $\psi_d : (S/I)_d \to \mathbb{C}^\delta$ *from Definition 3.2.3 is an isomorphism of* $\mathbb{C}$*-vector spaces.*

Proof. It follows from $d \in \mathrm{Reg}(I)$ that $\dim_{\mathbb{C}}(S/I)_d = \delta$. Moreover, $d \in \mathrm{Reg}(I)$ also implies that ψ_d is injective, since $f(\zeta_i) = 0, i = 1, \ldots, \delta$ means $f \in (\sqrt{(I : \mathfrak{B}^\infty)})_d = (I : \mathfrak{B}^\infty)_d = I_d$. ☐

It is now clear what the generalization of the Lagrange polynomials in Subsection 3.1.1 should be.

Definition 3.2.5 (Homogeneous Lagrange polynomials)**.** For $d \in \mathrm{Reg}(I)$ and $j = 1, \ldots, \delta$, let $\ell_j \in S_d$ be any representative of the class $\psi_d^{-1}(e_j) \in (S/I)_d$. That is, any homogeneous polynomial satisfying

$$\ell_j(z_j) = h(z_j), \quad \ell_j(z_i) = 0, i \neq j$$

for any set of homogeneous coordinates z_j of ζ_j, where $h \in S_d$ is used to define the evaluation map ψ_d (Definition 3.2.3).

Note that the elements $\mathrm{ev}_{\zeta_i}, i = 1, \ldots, \delta$ from Definition 3.2.3 form the dual basis of $(S/I)_d^\vee$ with respect to the homogeneous Lagrange polynomials. The next step is to define multiplication maps for homogeneous polynomials.

Definition 3.2.6 (Homogeneous multiplication map)**.** Fix $d, d_0 \in \mathbb{N}$. For any $g \in S_{d_0}$ we define the *multiplication map* representing *multiplication with* g as the $\mathbb{C}$-linear map

$$M_g : (S/I)_d \to (S/I)_{d+d_0} \quad \text{with} \quad M_g(f + I_d) = fg + I_{d+d_0}.$$

The following lemma will be used to state the main result of this subsection.

Lemma 3.2.1. *Let* $d, d_0 \in \mathbb{N}$ *be such that* $d, d + d_0 \in \mathrm{Reg}(I)$. *For any* $h_0 \in S_{d_0}$ *such that* $h_0(\zeta_i) \neq 0, i = 1, \ldots, \delta$ *we have that the multiplication map* $M_{h_0} : (S/I)_d \to (S/I)_{d+d_0}$ *is an isomorphism of vector spaces.*

Proof. Let $h \in S_d$ such that $h(\zeta_i) \neq 0, i = 1, \ldots, \delta$ and use h to define ψ_d. Since hh_0 does not vanish at any of the ζ_i, we can use it to define ψ_{d+d_0}. The lemma follows from $\psi_{d+d_0} \circ M_{h_0} = \mathrm{diag}(h_0(\zeta_1), \ldots, h_0(\zeta_\delta)) \circ \psi_d$ and Proposition 3.2.3. ☐

Theorem 3.2.4 (Projective eigenvalue, eigenvector theorem)**.** *Let $d, d_0 \in \mathbb{N}$ be such that $d, d + d_0 \in \operatorname{Reg}(I)$ and take $h_0 \in S_{d_0}$ as in Lemma 3.2.1. Then for any $g \in S_{d_0}$, $M_{g/h_0} = M_{h_0}^{-1} \circ M_g : (S/I)_d \to (S/I)_d$ has eigenpairs*

$$\left(\frac{g}{h_0}(\zeta_j), \ell_j + I_d\right), \qquad \left(\mathrm{ev}_{\zeta_j}, \frac{g}{h_0}(\zeta_j)\right), \quad j = 1, \dots, \delta,$$

where the $\ell_j + I_d$ are cosets of homogeneous Lagrange polynomials of degree d and the ev_{ζ_j} form the dual basis of $(S/I)_d^\vee$.

Proof. The map M_{h_0} is an isomorphism by Lemma 3.2.1. We define ψ_d, ψ_{d+d_0} as in Definition 3.2.3 with $h \in S_d$, $hh_0 \in S_{d+d_0}$ respectively. A straightforward computation shows that $\psi_{d+d_0} \circ M_{h_0}(\ell_j + I_d) = e_j$. Analogously, we have $\psi_{d+d_0} \circ M_g(\ell_j + I_d) = \frac{g}{h_0}(\zeta_j)e_j$. It follows that

$$M_{g/h_0}(\ell_j + I_d) = \frac{g}{h_0}(\zeta_j)(\ell_j + I_d),$$

which proves the statement about the right eigenpairs, since the $\ell_j + I_d$ are linearly independent. For the statement about the left eigenpairs, note that for any $f \in S_d$

$$\mathrm{ev}_{\zeta_j} \circ M_{g/h_0}(f + I_d) = \mathrm{ev}_{\zeta_j} \circ M_{h_0}^{-1}(gf + I_{d+d_0})$$

and since M_{h_0} is an isomorphism, there is $\tilde{f} \in S_d$ such that $gf - h_0\tilde{f} \in I_{d+d_0}$. Therefore, for each $\zeta_j \in V_{\mathbb{P}^n}(I)$ we have

$$\frac{gf - h_0\tilde{f}}{h_0 h}(\zeta_j) = 0 \Rightarrow \frac{\tilde{f}}{h}(\zeta_j) = \frac{g}{h_0}(\zeta_j)\frac{f}{h}(\zeta_j)$$

and thus, since $M_{h_0}^{-1}(gf + I_{d+d_0}) = \tilde{f} + I_d$, we have

$$\mathrm{ev}_{\zeta_j} \circ M_{g/h_0}(f + I_d) = \mathrm{ev}_{\zeta_j}(\tilde{f} + I_d) = \frac{g}{h_0}(\zeta_j)\,\mathrm{ev}_{\zeta_j}(f + I_d).$$

The ev_{ζ_j} are linearly independent, so this concludes the proof. □

As in the affine case, this suggests the following pseudo-algorithm for computing homogeneous coordinates of $\zeta_1, \dots, \zeta_\delta$.

1. For $d, d+1 \in \operatorname{Reg}(I)$ and for some basis of $(S/I)_d$, pick a generic linear form $h_0 \in S_1$ and compute matrix representations of $M_{x_0/h_0}, \dots, M_{x_n/h_0}$.
2. Diagonalize these matrices simultaneously, i.e. compute

$$DM_{x_i/h_0}D^{-1} = \operatorname{diag}\left(\frac{x_i}{h_0}(\zeta_1), \dots, \frac{x_i}{h_0}(\zeta_\delta)\right), \ i = 0, \dots, n,$$

   and read off the homogeneous coordinates from the diagonal.

3.2.3 Homogenization

In Subsection 3.2.1 we have discussed how a zero-dimensional homogeneous ideal $I \subset S$ gives ideals $\mathscr{I}(U_i) \subset \mathscr{O}(U_i) = \mathbb{C}[\mathbb{C}^n]$ defining points in an affine chart of $\mathbb{P}^n$ by *dehomogenizing* the generators. This is used to obtain local information such as the multiplicities of the points defined by I. In this subsection we will study the way of obtaining a homogeneous ideal $I \subset S = \mathbb{C}[x_0, \ldots, x_n]$ by *homogenizing* the generators of a zero-dimensional ideal in $R = \mathbb{C}[\mathbb{C}^n] = \mathbb{C}[y_1, \ldots, y_n]$. Recall that *homogenization* of degree d is defined as

$$\eta_d : R_{\leq d} \to S_d \quad \text{with} \quad \eta_d(\hat{f}(y_1, \ldots, y_n)) = x_0^d \hat{f}\left(\frac{x_1}{x_0}, \ldots, \frac{x_n}{x_0}\right).$$

Let $J = \langle \hat{f}_1, \ldots, \hat{f}_s \rangle \subset R$ and define d_i as the smallest integer such that $\hat{f}_i \in R_{\leq d_i}$. We consider the homogeneous ideal $I \subset S$ obtained as

$$I = \langle f_1, \ldots, f_s \rangle \subset S, \quad \text{with} \quad f_i = \eta_{d_i}(\hat{f}_i), \; i = 1, \ldots, s.$$

With the notation of Subsection 3.2.1, it is clear that $J = \mathscr{I}(U_0)$. If J is zero-dimensional, it is clear that $V_{\mathbb{P}^n}(I) \cap U_0 = V_{\mathbb{C}^n}(J)$ and the isolated points in $V_{\mathbb{P}^n}(I) \cap U_0$ have the same multiplicity as the corresponding points in $V_{\mathbb{C}^n}(J)$ (for the reader who knows about schemes: J and I define the same zero-dimensional subscheme of $U_0 \simeq \mathbb{C}^n$). For the rest of this subsection, we will consider the case where $s = n$.

A first observation is that *generically* nothing happens when going from J to I, in the sense that the only points in $V_{\mathbb{P}^n}(I)$ are the ones corresponding to $V_{\mathbb{C}^n}(J)$. To be more precise, let $(\hat{f}_1, \ldots, \hat{f}_s) \in \mathcal{F}_R(d_1, \ldots, d_n)$ be a general member in the sense that $V_{\mathbb{C}^n}(J)$ consists of $d_1 \cdots d_n$ points with multiplicity 1 (Theorem 3.1.2). Homogenization establishes an isomorphism between $\mathcal{F}_R(d_1, \ldots, d_n)$ and $\mathcal{F}_S(d_1, \ldots, d_n)$. By Theorem 3.2.2 our general member $(\hat{f}_1, \ldots, \hat{f}_s) \in \mathcal{F}_R(d_1, \ldots, d_n)$ homogenizes to a general member $(f_1, \ldots, f_s) \in \mathcal{F}_S(d_1, \ldots, d_n)$ in the sense that $V_{\mathbb{P}^n}(I)$ consists of $d_1 \cdots d_n$ isolated points with multiplicity 1. It is clear that these points are in one-to-one correspondence. Homogenization can sometimes be useful to understand the case where $(\hat{f}_1, \ldots, \hat{f}_s) \in \mathcal{F}_R(d_1, \ldots, d_n)$ does *not* behave like a general member (in terms of the Bézout root count), but the homogenization $(f_1, \ldots, f_s) \in \mathcal{F}_S(d_1, \ldots, d_n)$ does.

Example 3.2.3. Consider the ideal $J = \langle \hat{f}_1, \hat{f}_2 \rangle \subset R = \mathbb{C}[y_1, y_2]$ given by

$$\hat{f}_1 = y_1^2 - 3y_1y_2 + 2y_2^2 + 1, \quad \hat{f}_2 = y_1^2 - y_2^2 - 3y_2 + 1.$$

The solutions (y_1, y_2) in $\mathbb{C}^2$ are $(\sqrt{-1}, 0), (-\sqrt{-1}, 0)$ and $(3, 2)$. Note that this is one less than expected: the Bézout root count is $d_1 d_2 = 4$. To see where this 'missing' solution has gone, we homogenize to obtain

$$f_1 = x_1^2 - 3x_1x_2 + 2x_2^2 + x_0^2, \quad f_2 = x_1^2 - x_2^2 - 3x_0x_2 + x_0^2.$$

The solutions $(x_0 : x_1 : x_2)$ in $\mathbb{P}^2$ are $(1 : \sqrt{-1} : 0), (1; -\sqrt{-1} : 0), (1 : 3 : 2)$ and $(0 : 1 : 1)$. The first three in this list correspond to the affine solutions, and the fourth one lies in the line defined by $x_0 = 0$, which is the complement of U_0 in $\mathbb{P}^2$. In this setting, this is the *line at infinity*, and the system of equations $\hat{f}_1 = \hat{f}_2 = 0$ is said to have a *solution at infinity.* Note that $(f_1, f_2) \in \mathcal{F}_S(2, 2)$ is a generic member, in the sense of Bézout's theorem. We remark that from a numerical point of view, it makes sense to compute such 'excess solutions' as well, rather than ignoring them. Indeed, the slightest perturbation of the coefficients of $\hat{f}_1, \hat{f}_2$ will move the solution $(0 : 1 : 1) \in \mathbb{P}^2$ into U_0, causing $\hat{f}_1 = \hat{f}_2 = 0$ to have four solutions in $\mathbb{C}^2$, one of which has 'large' coordinates. △

Another reason one might want to use $\mathbb{P}^n$ as a solution space instead of $\mathbb{C}^n$ is that we can compute representatives $z_1, \ldots, z_\delta$ of the solutions $\zeta_1, \ldots, \zeta_\delta$ of I in *any* affine subspace of $\mathbb{P}^n$. More precisely, the solutions of J correspond to points in $U_0 \subset \mathbb{P}^n$, which in turn correspond to lines through the origin of $\mathbb{C}^{n+1}$ that hit the hyperplane $V_{\mathbb{C}^{n+1}}(x_0 - 1)$. This hyperplane is identified with $\mathbb{C}^n$: the coordinates $(y_1, \ldots, y_n) \in \mathbb{C}^n$ of the affine solutions are the $x_1, \ldots, x_n$ coordinates of the intersection of these lines with $V_{\mathbb{C}^{n+1}}(x_0 - 1)$. Instead of choosing the hyperplane $V_{\mathbb{C}^{n+1}}(x_0 - 1)$, we could pick a different linear form $h_0 \in S_1$ and identify $\mathbb{C}^n$ with $V_{\mathbb{C}^{n+1}}(h_0 - 1)$ via the map from Remark 2.2.2. This may be advantageous if the coordinates for $x_0 = 1$ of a solution are very large (solutions 'near' infinity). In this case we can compute the coordinates for $h_0 = 1$ with h_0 chosen randomly (such that there is no reason to expect that the coordinates will be large) and afterwards we simply scale them to have $x_0 = 1$. More concretely, solutions on or near infinity cause numerical issues for computing the multiplication matrices M_{y_i} from Subsection 3.1.1, which are actually the matrices M_{x_i/x_0} from Subsection 3.2.2. Choosing a random element h_0 can help us get rid of this issue completely. We will say more about this in Section 4.5.

As we have noted in Example 2.2.7, homogenizing the generators of J may enlarge the variety by adding components contained in $\mathbb{P}^n \setminus U_0$. This is also what happened in Example 3.2.3. The fact that an extra point was added after homogenizing in Example 3.2.3 was due to the equations $\hat{f}_1, \hat{f}_2$ being *non-generic* in a sense. Indeed, the 4 solutions of a general member of $\mathcal{F}_R(2, 2)$ all lie in $\mathbb{C}^2$. Sometimes, however, extra points in $\mathbb{P}^n \setminus U_0$ are introduced *as an artifact of homogenization*, possibly even destroying the zero-dimensionality. This is illustrated by the following example.

Example 3.2.4. Let $R = \mathbb{C}[y_1, y_2, y_3]$ and consider the equations

$$\begin{aligned}
\hat{f}_1 &= a_1 + a_2y_1 + a_3y_2 + a_4y_3 + a_5y_1y_2 + a_6y_1y_3 + a_7y_2y_3 + a_8y_1y_2y_3, \\
\hat{f}_2 &= b_1 + b_2y_1 + b_3y_2 + b_4y_3 + b_5y_1y_2 + b_6y_1y_3 + b_7y_2y_3 + b_8y_1y_2y_3, \\
\hat{f}_3 &= c_1 + c_2y_1 + c_3y_2 + c_4y_3 + c_5y_1y_2 + c_6y_1y_3 + c_7y_2y_3 + c_8y_1y_2y_3.
\end{aligned}$$

Homogenizing these equations and setting $x_0 = 0$ we obtain

$$\begin{aligned}
f_1(0, x_1, x_2, x_3) &= a_8x_1x_2x_3, \\
f_2(0, x_1, x_2, x_3) &= b_8x_1x_2x_3, \\
f_3(0, x_1, x_2, x_3) &= c_8x_1x_2x_3.
\end{aligned}$$

This shows that *for any choice of the parameters* a_i, b_i, c_i, $V_{\mathbb{P}^n}(I)$ contains the three lines $\{(0:0:x_2:x_3)\}, \{(0:x_1:0:x_3)\}, \{(0:x_1:x_2:0)\}$ (each of which is isomorphic to $\mathbb{P}^1$). $\triangle$

Example 3.2.4 is an illustration of how homogenization has some undesirable properties for systems coming from a *subfamily* $\mathcal{F}' \subset \mathcal{F}_R(d_1, \ldots, d_n)$ which is such that generic elements of the subfamily do not behave like generic elements of $\mathcal{F}_R(d_1, \ldots, d_n)$. We argue that in this kind of situations, $\mathbb{P}^n$ is not the right solution space to consider. This raises the question '*which one is?*'. For an important class of subfamilies $\mathcal{F}' \subset \mathcal{F}_R(d_1, \ldots, d_n)$, containing the family considered in Example 3.2.4, the answer is a *compact toric variety* which is naturally associated to $\mathcal{F}'$. This is the subject of Chapter 5. For now, we will work with the isomorphic families $\mathcal{F}_R(d_1, \ldots, d_n)$ and $\mathcal{F}_S(d_1, \ldots, d_s)$ and solution spaces $\mathbb{C}^n$ or $\mathbb{P}^n$.

3.3 Gröbner and border bases

To use the results of the previous subsections for solving polynomial systems we need algorithmic tools for doing computations modulo an ideal I. The theory of *Gröbner bases* provides us with such a tool. Gröbner bases have led to great advances in computational algebraic geometry and computer algebra and give rise to a good example of what is called a *normal form* with respect to an ideal. This is a concept that plays an important role in this thesis. *Border bases* generalize Gröbner bases in several ways. In particular, they remove some of the restrictions that Gröbner bases impose on the basis of the quotient ring R/I in which we can work. Our aim is to present the main ideas. For references that cover Gröbner and border bases in more detail, see Subsection 1.3.1. Throughout this subsection we work with zero-dimensional ideals $I \subset R = \mathbb{C}[x_1, \ldots, x_n]$. In the context of Gröbner bases it is more common to work over fields that are more fit for symbolic computation, such as $\mathbb{Q}$ or finite fields. We stick to the complex numbers for the sake of consistency. The reader can safely replace $\mathbb{C}$ in this section with their favorite field.

3.3.1 Gröbner bases

The discussion on Gröbner bases included here is partly inspired by some lectures by Frank Sottile on *Algorithmic Algebraic Geometry*, attended by the author at FU Berlin in the fall semester of 2019.

In the case where $n = 1$, all ideals in $R = \mathbb{C}[x]$ are principal. If $f = c_0 + c_1 x + \cdots + c_d x^d$ with $c_d \neq 0$ and $I = \langle f \rangle$, a canonical choice of basis for R/I is $\mathcal{B} = \{1 + I, x + I, \ldots, x^{d-1} + I\}$. A well known way of expanding the residue class of any polynomial $g \in R$ in this basis is given by the *Euclidean division algorithm.* This algorithm writes g as

$$g = qf + r,$$

where $r, q \in R$ and the degree of r is smaller than d. It follows easily that $g+I = r+I$ and the coefficients of r (in the monomial basis) give the expansion of $g+I$ in terms of $\mathcal{B}$. One can think of the Euclidean division as a way of using f to *rewrite* g modulo I using 'smaller' monomials. Here smaller is with respect to the total order

$$1 < x < x^2 < x^3 < \cdots$$

on the monoid of monomials in R, or equivalently, with respect to the canonical total order on the natural numbers $\mathbb{N}$. A first step to generalize this to the multivariate case is to define what we mean by 'small' monomials. For $n > 1$, there is no canonical total ordering on the monomials in $\mathbb{R}^n$.

Definition 3.3.1 (Monomial order)**.** A *monomial order* is a total order '$\prec$' on the monomials of R such that for any $a, b, c \in \mathbb{N}^n$

1. $1 \preceq x^a$ for any $a \in \mathbb{N}^n$,
2. $x^a \prec x^b$ implies $x^{a+c} \prec x^{b+c}$.

Example 3.3.1 (Monomial orders)**.** Some important examples of monomial orders are

1. the *lexicographic* order, where $x^a \succ_{\text{lex}} x^b$ if the first nonzero entry of $a - b$ is positive,
2. the *degree lexicographic order*, where $x^a \succ_{\text{deglex}} x^b$ if $|a| > |b|$ or $|a| = |b|$ and $x^a \succ_{\text{lex}} x^b$,
3. the *degree reverse lexicographic order*, where $x^a \succ_{\text{drl}} x^b$ if $|a| > |b|$ or $|a| = |b|$ and the last nonzero entry of $a - b$ is negative.

For example, in $R = \mathbb{C}[x_1, x_2]$, $x_1 \succ_{\text{lex}} x_2^2$, yet $x_1 \prec_{\text{deglex}} x_2^2$. In $R = \mathbb{C}[x_1, x_2, x_3]$ we have

$$x_1^3 x_2 x_3^3 \succ_{\text{lex}} x_1 x_2^4 x_3^2, \quad x_1^3 x_2 x_3^3 \succ_{\text{deglex}} x_1 x_2^4 x_3^2 \quad \text{and} \quad x_1^3 x_2 x_3^3 \prec_{\text{drl}} x_1 x_2^4 x_3^2.$$

$\triangle$

In what follows, if we do not specify the monomial order we will assume that some monomial order '$\prec$' is fixed.

Definition 3.3.2 (Initial monomial)**.** For a polynomial $f = \sum_{a \in \mathbb{N}^n} c_a x^a \in R$ we define the *initial monomial* of f as

$$\text{in}_\prec(f) = x^a \quad \text{where } x^a \text{ is the maximal element w.r.t. } \prec \text{ such that } c_a \neq 0.$$

Theorem 3.3.1 (Multivariate division algorithm)**.** *There exists an algorithm which takes as an input the polynomials $g, f_1, \ldots, f_s \in R$ and a monomial order '$\prec$' and gives as an output a set of polynomials $q_1, \ldots, q_s, r \in R$ satisfying*

1. $g = q_1 f_1 + \cdots + q_s f_s + r$,
2. $\mathrm{in}_\prec(g) \succeq \mathrm{in}_\prec(r)$,
3. $\mathrm{in}_\prec(g) \succeq \mathrm{in}_\prec(q_i f_i),\ i = 1, \ldots, s$,
4. *no term of r is divisible by any of the initial monomials* $\mathrm{in}_\prec(f_i), i = 1, \ldots, s$.

Proof. The algorithm is a straightforward generalization of the Euclidean division algorithm for $n = 1$. It is given explicitly in the proof of Theorem 3 in [CLO13, Chapter 2, §3]. □

It is clear that if $I = \langle f_1, \ldots, f_s \rangle$ and the algorithm of Theorem 3.3.1 allows us to write $g = q_1 f_1 + \cdots + q_s f_s + r$, then $g + I = r + I$ in R/I. Unfortunately, in general this does not give a unique way of representing g modulo I. The output depends on the choice of generators $f_1, \ldots, f_s$ of I and on the way they are ordered. The following is Example 5 in [CLO13, Chapter 2, §3]. It shows that the conditions imposed on the output of the multivariate division algorithm do not guarantee that r is unique.

Example 3.3.2. Let $R = \mathbb{C}[x, y]$ with lexicographic monomial order where $x \succ y$. For $g = xy^2 - x$, $f_1 = xy - 1$, $f_2 = y^2 - 1$, the polynomials

$$q_1 = y, \quad q_2 = 0, \quad r = -x + y$$

satisfy the conditions of Theorem 3.3.1, and so do the polynomials

$$q_1' = 0, \quad q_2' = x, \quad r' = 0.$$

In fact, (q_1, q_2, r) is the output of the algorithm in [CLO13, Chapter 2, §3], whereas (q_1', q_2', r') is the output when the order of f_1, f_2 is changed. △

This 'imperfection' of the multivariate division algorithm can be removed by imposing some conditions on $f_1, \ldots, f_s$ such that r is unique under the conditions of Theorem 3.3.1. Such 'special' sets of generators for I are called *Gröbner bases.*

Definition 3.3.3 (Gröbner basis)**.** A finite subset $\mathcal{G} \subset I$ is called a *Gröbner basis* for I with respect to '$\prec$' if the *initial ideal*

$$\mathrm{in}_\prec(I) = \langle x^a \mid x^a = \mathrm{in}_\prec(g) \text{ for some } g \in I \rangle$$

satisfies $\mathrm{in}_\prec(I) = \langle \mathrm{in}_\prec(f) \mid f \in \mathcal{G} \rangle$.

It is a direct consequence of Dickson's lemma [CLO13, Chapter 2, §4, Theorem 5] that every ideal in R has a finite Gröbner basis. The terminology 'Gröbner *basis*' is justified by the fact that any Gröbner basis of an ideal I is a basis for the ideal, i.e. the elements of a Gröbner basis generate the ideal [CLO13, Chapter 2, §5, Corollary 6].

Proposition 3.3.1. *If $\{f_1, \dots, f_s\}$ is a Gröbner basis for $I = \langle f_1, \dots, f_s \rangle$, then $g \in I$ if and only if the polynomial r returned by the multivariate division algorithm is the zero polynomial. Moreover, for each $g \in R$ there is a unique polynomial $r \in R$ satisfying $r + I = g + I$ and condition 4 of Theorem 3.3.1.*

Proof. It is clear that if $r = 0$, $g \in I$. Conversely, if $r \neq 0$, then by the fourth condition of Theorem 3.3.1 no term of r lies in $\mathrm{in}_{\prec}(I)$. It follows that $r \notin I$, which implies $g \notin I$ since $g = q_1 f_1 + \cdots + q_s f_s + r$. To prove the second statement, suppose that

$$g = q_1 f_1 + \cdots + q_s f_s + r = q_1' f_1 + \cdots + q_s' f_s + r'.$$

Then $r - r' \in I$. If $r = r'$, we're done. If $r \neq r'$, we arrive at a contradiction because none of the terms in $r - r'$ are in $\mathrm{in}_{\prec}(I)$. □

The unique polynomial r returned by the multivariate division algorithm for a polynomial $g \in R$ and a Gröbner basis $\mathcal{G} \subset R$ of an ideal I is called the *remainder upon division of g by $\mathcal{G}$*. We denote $r = \mathcal{N}_{\mathcal{G}}(g)$. The set of monomials

$$\mathcal{B}_{\prec} = \{x^a \mid x^a \notin \mathrm{in}_{\prec}(I)\}$$

is called the *set of standard monomials* of I with respect to $\prec$. Their $\mathbb{C}$-linear span is denoted by

$$B_{\prec} = \mathrm{span}_{\mathbb{C}}(\mathcal{B}_{\prec}) = \left\{ \sum_{x^a \in \mathcal{B}_{\prec}} c_a x^a \mid \text{finitely many } c_a \text{ are nonzero} \right\} \subset R.$$

It follows from Proposition 3.3.1 that the map $\mathcal{N}_{\mathcal{G}} : R \to B_{\prec}$ is $\mathbb{C}$-linear and $\mathcal{N}_{\mathcal{G}}(b) = b$ for all $b \in B_{\prec}$.

Theorem 3.3.2. *Let $\mathcal{G} = \{f_1, \dots, f_s\}$ be a Gröbner basis for I. We have the short exact sequence of $\mathbb{C}$-vector spaces*

$$0 \longrightarrow I \longrightarrow R \xrightarrow{\mathcal{N}_{\mathcal{G}}} B_{\prec} \longrightarrow 0.$$

Proof. The fact that $\ker \mathcal{N}_{\mathcal{G}} = I$ follows immediately from Proposition 3.3.1. Surjectivity of $\mathcal{N}_{\mathcal{G}} : R \to B_{\prec}$ follows from $B_{\prec} \subset R$ and $\mathcal{N}_{\mathcal{G}}(b) = b$ for $b \in B_{\prec}$. □

Corollary 3.3.1. *If $I \subset R$ is a zero-dimensional ideal with $V_{\mathbb{C}^n}(I) = \{z_1, \dots, z_\delta\}$ such that z_i has multiplicity μ_i and $\delta^+ = \mu_1 + \cdots + \mu_\delta$, then for any monomial order '$\prec$', the set of standard monomials $\mathcal{B}_{\prec}$ consists of δ^+ monomials whose residue classes in R/I form a $\mathbb{C}$-basis of R/I.*

Remark 3.3.1. A Gröbner basis $\mathcal{G} = \{f_1, \dots, f_s\}$ is called *reduced* if for $i = 1, \dots, s$, the coefficient standing with the monomial $\mathrm{in}_{\prec}(f_i)$ equals 1 and no monomial occurring in f_i can be divided by any of the leading terms of the other elements of $\mathcal{G}$ (i.e. all monomials of f_i are not contained in $\langle \mathrm{in}_{\prec}(f_j) \mid j \neq i \rangle$. Reduced Gröbner bases have the nice property that every ideal $I \subset R$ has a *unique* reduced Gröbner basis for any monomial ordering [CLO13, Chapter 2, §7, Theorem 5]. △

Remark 3.3.2. The remainder upon division r of a polynomial g by a Gröbner basis $\mathcal{G} = \{f_1, \dots, f_s\}$ can be defined as the result of the multivariate division algorithm because of the uniqueness property in Proposition 3.3.1. However, the polynomials $q_1, \dots, q_s$ satisfying the conditions of Theorem 3.3.1 are not unique (for instance, replace q_i by $q_i + f_j$ and q_j by $q_j - f_i$). However, the polynomial $h = q_1 f_1 + \cdots + q_s f_s = g - r$ can be defined from any output of the multivariate division algorithm and is again unique. The map $g \mapsto h + r$ makes the isomorphism $R \simeq I \oplus B_\prec$ explicit. $\triangle$

Remark 3.3.3. Gröbner bases, along with an algorithm for computing them, were introduced by Bruno Buchberger. In his Ph. D. thesis [Buc06], the focus was on the zero-dimensional case. The general theory was developed in [Buc70]. Many improvements to the original algorithm have been made to reduce the complexity and memory usage. We have listed some references in Subsection 1.3.1. A more complete overview is given in [CLO13, Chapter 2, §10]. The development of specialized Gröbner basis methods is ongoing research. See, for instance, the Ph. D. thesis of Zuzana Kukelova [Kuk13] for Gröbner basis methods in computer vision, and the Ph. D. thesis of Matías Bender [Ben19] for specialized algorithms dealing with sparse polynomials. $\triangle$

Example 3.3.3. As an illustration, we compute Gröbner bases for the ideal of Example 3.1.2 using the computer algebra software Macaulay2 [GS] for two different monomial orderings. Using the (default) degree reverse lexicographic order, we obtain

$$\mathcal{G} = \{6\underline{xy} - y^2 - 3x + 22y + 5, 3\underline{x^2} + 4y^2 + 3x - 10y + 4, 98\underline{y^3} - 363y^2 - 189x + 888y + 107\},$$

where we have underlined the initial monomials. Here $\mathrm{in}_{\prec_{\mathrm{drl}}}(I) = \langle xy, x^2, y^3 \rangle$ and $\mathcal{B}_{\prec_{\mathrm{drl}}} = \{1, y, y^2, x\}$. For a lexicographic order with $y \succ_{\mathrm{lex}} x$ we obtain

$$\mathcal{G} = \{49\underline{x^4} + 374x^3 + 913x^2 + 840x + 1260, 906\underline{y} - 196x^3 - 859x^2 - 747x - 1272\}.$$

Here $\mathrm{in}_{\prec_{\mathrm{lex}}}(I) = \langle x^4, y \rangle$ and $\mathcal{B}_{\prec_{\mathrm{lex}}} = \{1, x, x^2, x^3\}$. We note that these computations happened in exact arithmetic: if the ideal can be generated by polynomials with coefficients in a field K, then it is a direct consequence of Buchberger's algorithm that the ideal has a Gröbner basis with coefficients in K (here $K = \mathbb{Q}$, for instance). Figure 3.2 shows how the partitioning of the monomials of $\mathbb{C}[x, y]$ into $\mathcal{B}$ and the monomials in $\mathrm{in}_\prec(I)$ leads to a typical *staircase pattern*, which depends on the monomial order. In this type of figures, we identify $a \in \mathbb{N}^2$ with the monomial $x^{a_1} y^{a_2}$. $\triangle$

What is essential for us is that a map $\mathcal{N}_\mathcal{G}$ having the property of Theorem 3.3.2 allows us to compute the multiplication maps from Subsection 3.1.1. Indeed, multiplication with g in the basis $\mathcal{B}_\prec = \{x^{a_1}, \dots, x^{a_\delta}\}$ looks like

$$M_g = \begin{array}{c} \\ x^{a_1} \\ \vdots \\ x^{a_\delta} \end{array} \begin{array}{c} \begin{array}{ccc} x^{a_1} & \cdots & x^{a_\delta} \end{array} \\ \begin{bmatrix} | & & | \\ \mathcal{N}_\mathcal{G}(gx^{a_1}) & \cdots & \mathcal{N}_\mathcal{G}(gx^{a_\delta}) \\ | & & | \end{bmatrix} \end{array}$$

where the columns are the expansions of $\{\mathcal{N}_\mathcal{G}(gx^a) \mid x^a \in \mathcal{B}_\prec\}$ in the basis $\mathcal{B}_\prec$. A map satisfying the property of Theorem 3.3.2 is what we will define to be a *normal form*. We will see another example in the next subsection.

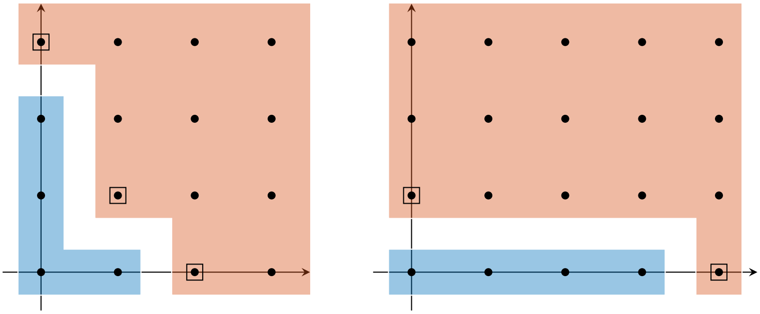


Figure 3.2: Illustration of the staircase patterns of a $\prec_{\mathrm{drl}}$ (left) and a $\prec_{\mathrm{lex}}$ (right) Gröbner basis for the ideal of Example 3.3.3. The initial terms in the Gröbner basis (i.e. the generators of $\mathrm{in}_\prec(I)$ are indicated with small boxes.

3.3.2 Border bases

The staircase patterns arising from Gröbner bases depend on the choice of monomial order, but they also depend on the ideal. This is natural in the sense that the subsets of monomials of R whose images in R/I can be used as a basis for R/I depends on I. However, the dependence of $\mathcal{B}_\prec$ on the ideal has some specific features that are artifacts of working with a monomial order '$\prec$' and can have bad consequences for the behavior of Gröbner bases in a numerical context. Here's an example that illustrates this.

Example 3.3.4. Let $R = \mathbb{C}[x, y]$ and consider the degree reverse lexicographic monomial order '$\prec_{\mathrm{drl}}$' with $y \prec_{\mathrm{drl}} x$. We consider the ideal $I = \langle f_1, f_2 \rangle$ from Example 3.1.6 with

$$f_1 = x + \frac{1}{3}y^2 - x^2, \quad \text{and} \quad f_2 = \frac{-1}{3}x + \frac{1}{3}x^2.$$

The resulting reduced Gröbner basis is $\mathcal{G} = \{\underline{x^2} - x, \underline{y^2}\}$ and $\mathcal{B}_{\prec_{\mathrm{drl}}} = \{1, x, y, xy\}$. If we perturb the polynomials f_1 and f_2 slightly to obtain $I' = \langle f_1', f_2' \rangle$ with $f_1' = f_1 - 10^{-7}xy$, $f_2' = f_2 + 10^{-7}xy$, the new reduced Gröbner basis becomes

$$\mathcal{G}' = \left\{ \underline{xy} + \frac{10^7}{6}y^2, \underline{x^2} - \frac{1}{2}y^2 - x, \underline{y^3} + \frac{30000000}{49999999999991}y^2 \right\}$$

with set of standard monomials $\mathcal{B}'_{\prec_{\mathrm{drl}}} = \{1, x, y, y^2\}$. In order to obtain the first two elements of $\mathcal{G}'$, we can use the equations f_1', f_2' to write that (modulo I')

$$\begin{bmatrix} -10^{-7} & -1 \\ 10^{-7} & 1/3 \end{bmatrix} \begin{bmatrix} xy \\ x^2 \end{bmatrix} = - \begin{bmatrix} 0 & 1 & 0 & 1/3 \\ 0 & -1/3 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & x & y & y^2 \end{bmatrix}^\top,$$

from which we get

$$\begin{bmatrix} xy \\ x^2 \end{bmatrix} = - \begin{bmatrix} -10^{-7} & -1 \\ 10^{-7} & 1/3 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 1 & 0 & 1/3 \\ 0 & -1/3 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & x & y & y^2 \end{bmatrix}^\top = \begin{bmatrix} -10^7/6y^2 \\ 1/2y^2 + x \end{bmatrix}.$$

The reader who is familiar with numerical analysis notices that this computation is not very suitable for finite precision arithmetic: we are inverting an ill-conditioned matrix (see Appendix B). Indeed, performing this computation in double precision arithmetic with the help of the following Matlab [MAT17] commands

```
A = [-1e-7 -1;1e-7 1/3];  B = [0 1 0 1/3;0 -1/3 0 0];
```

we get a relative forward error

```
>> norm(-A\B - [0 0 0 -1e7/6; 0 1 0 1/2])/norm(A)
```

of size `4.4177e-10`, which is roughly 10^6 times larger than our working precision! It is interesting to see what the analogous computation looks like when we stick to our set of standard monomials $\{1, x, y, xy\}$ from before. We now get

$$\begin{bmatrix} -1 & 1/3 \\ 1/3 & 0 \end{bmatrix} \begin{bmatrix} x^2 \\ y^2 \end{bmatrix} = - \begin{bmatrix} 0 & 1 & 0 & -10^{-7} \\ 0 & -1/3 & 0 & 10^{-7} \end{bmatrix} \begin{bmatrix} 1 & x & y & xy \end{bmatrix}^\top,$$

which leads to $x^2 - x + 3 \cdot 10^{-7} xy \in I'$ and $y^2 + 6 \cdot 10^{-7} xy \in I'$. The coefficient matrix is now perfectly well conditioned and the set of polynomials $\mathcal{H} = \{x^2 - x + 3 \cdot 10^{-7} xy, y^2 + 6 \cdot 10^{-7} xy\}$ can be computed up to machine precision. Note that the polynomials in $\mathcal{H}$ are slightly perturbed versions of the polynomials in $\mathcal{G}$. They are a basis for the ideal I' as they are just an invertible linear combination of f_1' and f_2'. Although not a Gröbner basis, the set $\mathcal{H}$ can be used to rewrite any polynomial $g \in R$ as a $\mathbb{C}$-linear combination of the monomials in $\mathcal{B}$ modulo the ideal (as we will see). Even though the slightly perturbed polynomials f_1', f_2' lead to a slightly perturbed set of polynomials $\mathcal{H}$ that allow us to compute modulo I' in the basis $\{1 + I', x + I', y + I', xy + I'\}$ of R/I', the Gröbner basis $\mathcal{G}'$ and its corresponding set of standard monomials change completely. Moreover, we are forced to solve a *nearly degenerate* system of linear equations in order to compute $\mathcal{G}'$. The reason for this is that the monomial order '$\prec_{\text{drl}}$' *really* prefers y^2 over xy as a candidate for the set of standard monomials. By adding the monomial xy to the equations, xy 'replaces' y^2 in the initial ideal. This causes an artificial discontinuity in the set of standard monomials picked by a Gröbner basis. Note that the condition number of the coefficient matrix in this example governs the magnitude of the coefficients in the reduced Gröbner basis. Also, the size 10^{-7} of the perturbation can be taken smaller: the situation can be made arbitrarily bad. △

Similar examples of the bad behavior of Gröbner bases in a numerical context can be found, for instance, in the introductions of [Ste97, Mou99]. *Border bases* have been developed to remedy this type of behavior. For instance, the set $\mathcal{H}$ of Example 3.3.4 is part of a border basis. The idea of the multivariate division algorithm is to use the elements $f_1, \ldots, f_s$ to *reduce* a polynomial g, where 'reduce' means 'lower' its initial monomial with respect to the chosen monomial order. A reduced Gröbner basis $\mathcal{G} = \{f_1, \ldots, f_s\}$ is such that

$$f_i = \text{in}_\prec(f_i) - \sum_{x^a \in \mathcal{B}} c_a x^a \tag{3.3.1}$$

gives an explicit way of rewriting $\mathrm{in}_\prec(f_i)$ as $\sum_{x^a\in\mathcal{B}} c_a x^a$ modulo the ideal $\langle f_1,\dots,f_s\rangle$. The initial monomial of any polynomial g that is not in $B_\prec$ divides one of the $\mathrm{in}_\prec(f_i)$. This means that there is an appropriate term cx^a such that $g - cx^a f_i$ lies 'closer' to $B_\prec$ than g does, in the sense that $\mathrm{in}_\prec(g - cx^a f_i) \prec \mathrm{in}_\prec(g)$. Border bases give a way of 'reducing' any polynomial g modulo I *without* the use of a monomial order. More precisely, for a $\mathbb{C}$-vector subspace $B \subset R$ satisfying some properties, a B-border basis for a zero-dimensional ideal I is a basis $\mathcal{H}$ of I that induces a map $\mathcal{N}_\mathcal{H} : R \to B$ such that $g - \mathcal{N}_\mathcal{H}(g) \in I$ and $g \mapsto (g - \mathcal{N}_\mathcal{H}(g), \mathcal{N}_\mathcal{H}(g))$ gives an isomorphism $R \simeq I \oplus B$. In particular, a Gröbner basis $\mathcal{G}$ gives a border basis with $\mathcal{N}_\mathcal{H} = \mathcal{N}_\mathcal{G}$. We will now fill in the gaps in this definition. First of all, let us specify which conditions the subspace B should satisfy. Two different definitions are commonly used in the literature, and we will give them both.

Definition 3.3.4 (Order ideal)**.** A nonempty subset $\mathcal{B}$ of monomials in R is called an *order ideal* or a *closed subset* if for each $x^b \in \mathcal{B}$ and $x^{b'}$ such that $x^{b'}$ divides x^b, we have $x^{b'} \in \mathcal{B}$.

Note that every order ideal contains 1. For instance, the references [MMM91, Ste97, KKR05, KK05] work with B-border bases where B is the $\mathbb{C}$-linear span of an order ideal.

Definition 3.3.5 (Connected to 1)**.** A $\mathbb{C}$-vector subspace $B \subset R$ is *connected to 1* if for every $b \in B$ there exist $b_1,\dots,b_n \in B$ such that

$$b = \sum_{i=1}^{n} x_i b_i.$$

Every connected to 1 subspace $B \subset R$ contains 1. Moreover, the $\mathbb{C}$-span of every order ideal is connected to 1. An example of a set of monomials that is an order ideal and one that is not, but its span is still connected to 1, are shown in Figure 3.3. The connected to 1 property is the restriction on B for the B-border bases discussed in [Mou99, MT05, LLM$^+$13]. Since subspaces that are connected to 1 contain the subspaces coming from an order ideal, we will work with this assumption in the remainder of this subsection. Next, in order to specify what we mean by 'reducing' a polynomial g with respect to B, we need a way of determining *how far* g is from being in B. To that end, following the approach in [Mou99], for any subspace $B \subset R$ we define

$$B^+ = B + x_1 \cdot B + \cdots + x_n \cdot B$$

where $x_i \cdot B = \{x_i b \mid b \in B\} \subset R$, and we let $B^{[d]}$ be the result of applying the operator $(\cdot)^+$ d times to B. We set $B^{[0]} = B$ by convention and we define $B^{[\star]} = \bigcup_{d=0}^{\infty} B^{[d]}$.

Definition 3.3.6 (B-index)**.** For a polynomial $g \in R$ and a subspace $B \subset R$, we define the *B-index* $\mathrm{ind}_B(g)$ of g as the smallest $d \in \mathbb{N}$ such that $g \in B^{[d]}$. If such a d does not exist, we set $\mathrm{ind}_B(g) = -\infty$.

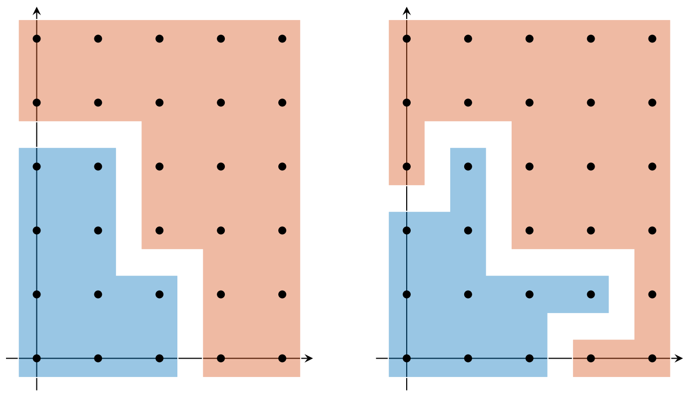


Figure 3.3: Illustration of an order ideal (left) and the 'connected to 1' property (left and right).

Note that if $1 \in B$, every $g \in R$ has a finite B-index and $B^{[\star]} = R$. Also, if $L \subset R$ is spanned by $\mathcal{H} = \{f_1, \ldots, f_s\}$ over $\mathbb{C}$, then $L^{[\star]} = \langle \mathcal{H} \rangle = \langle f_1, \ldots, f_s \rangle$.

Lemma 3.3.1. *If $1 \in B$ and L is such that $B^+ = B + L$, then every element g with $\operatorname{ind}_B(g) = d$ can be written as $g = h + r$ where $h \in L^{[d-1]}$ and $r \in B$.*

Proof. The proof is by induction on d [Mou99, Lemma 2.3]. □

The process of writing $g = h+r$ in Lemma 3.3.1 is called *B-reduction of g along L*. This is to border basis algorithms what the multivariate division algorithm is to Gröbner bases. Here B plays the role of $B_\prec$ and L plays the role of the $\mathbb{C}$-linear span of the generators of the ideal I. With the right assumptions on B and L we will have that the B-reduction along L is *canonical*, i.e. for each $d \in \mathbb{N}$ and each $g \in R$ with $\operatorname{ind}_B(g) = d$ there is a unique way of writing $g = h + r$ with $h \in L^{[d-1]}, b \in B$. Equivalently, B-reduction along L defines a map $\mathcal{N}_\mathcal{H} : R \to B$ where $\mathcal{N}_\mathcal{H}(g) = (g - \mathcal{N}_\mathcal{H}(g), \mathcal{N}_\mathcal{H}(g))$ is an isomorphism $R \simeq \langle \mathcal{H} \rangle \oplus B$ (here $\mathcal{H}$ is a $\mathbb{C}$-basis for L).

Definition 3.3.7 (Border basis)**.** Let $I \subset R$ be a zero-dimensional ideal. A *border basis* of I is a pair $(B, \mathcal{H})$ where

1. $B \subset R$ such that $\dim_\mathbb{C} B = \dim_\mathbb{C} R/I$ and B is connected to 1,
2. $L = I \cap B^+$ is supplementary to B in B^+: $B^+ = B \oplus L$,
3. $\mathcal{H}$ is a $\mathbb{C}$-basis for L.

We say that $\mathcal{H}$ is a *B-border basis* of I.

As we will see in Section 4.2, for a border basis $(B, \mathcal{H})$ of I we have that the B-reduction $\mathcal{N}_{\mathcal{H}} : R \to B$ along $L = I \cap B^+$ is canonical and $\langle L \rangle = \langle \mathcal{H} \rangle = I$, so $\mathcal{H}$ is indeed an ideal basis of I. In [Mou99] an algorithm is described for computing a border basis of I, based on *Mourrain's criterion* for normal form algorithms [Mou99, Theorem 3.1].

Definition 3.3.7 is mostly based on the results from [Mou99], even though in this article the terminology *border basis* is not used. To justify this definition, we remark the following. Definition 3.3.7 defines a border basis as *any* $\mathbb{C}$-basis $\mathcal{H}$ for $L = I \cap B^+$. However, for every subspace $\partial B \subset B^+$ such that $B^+ = B \oplus \partial B$ and for every choice of $\mathbb{C}$-basis $\partial\mathcal{B}$ for ∂B there is a canonical choice for $\mathcal{H}$. This choice of $\mathcal{H}$ leads to the definition of '$\mathcal{B}$-border basis' in [KK05, KKR05, KK06, Ste97] if $\mathcal{B}$ is a $\mathbb{C}$-basis for B which is an order ideal and that of a 'border basis for $\mathcal{B}$' in [MT08] if $\mathcal{B}$ consists of monomials and $B = \operatorname{span}_{\mathbb{C}}(\mathcal{B})$ is connected to 1. For a border basis $(B, \mathcal{H})$ we say that $\mathcal{H}$ is a *reduced B-border basis* with respect to a basis $\partial\mathcal{B} = \{g_1, \ldots, g_s\}$ of ∂B if $\mathcal{H} = \{f_1, \ldots, f_s\}$ with

$$f_i = g_i - \mathcal{N}_{\mathcal{H}}(g_i), \ i = 1, \ldots, s.$$

Note that $\{f_1, \ldots, f_s\}$ give an explicit way of rewriting the 'border' ∂B of B modulo the ideal.

Example 3.3.5. Let $\mathcal{G} = \{f_1, \ldots, f_s\}$ be a reduced Gröbner basis for I with respect to a monomial order '$\prec$'. The border $\partial B_\prec$ contains the initial monomials $\mathrm{in}_\prec(f_i)$. Let $\partial\mathcal{B}_\prec = \{x^a \mid x^a \in B^+ \text{ but } x^a \notin B\}$. Then $\partial\mathcal{B}_\prec$ is a basis for $\partial B_\prec$, the set

$$\mathcal{H} = \{x^a - \mathcal{N}_{\mathcal{G}}(x^a) \mid x^a \in \partial\mathcal{B}_\prec\}$$

contains $\mathcal{G}$ and is a reduced $B_\prec$-border basis with respect to $\partial\mathcal{B}_\prec$. △

Example 3.3.6. Let $B \subset R = \mathbb{C}[x, y]$ be the $\mathbb{C}$-span of $\{1, x, y, xy\}$ and consider the basis $\partial\mathcal{B} = \{x^2, y^2, x^2y, xy^2\}$ of $\partial B \simeq B^+/B$. A reduced B-border basis with respect to $\partial\mathcal{B}$ for the perturbed ideal I' from Example 3.3.4 is given by

$$\begin{aligned} \mathcal{H}' = \{&x^2 - x + 3 \cdot 10^{-7} xy, \ y^2 + 6 \cdot 10^{-7} xy, \\ &x^2y - \frac{1}{1 - 18 \cdot 10^{-14}} xy, \ xy^2 + \frac{6 \cdot 10^{-7}}{1 - 18 \cdot 10^{-14}} xy\}. \end{aligned}$$

This is a slightly perturbed version of the B-border basis

$$\mathcal{H} = \{x^2 - x, \ y^2, \ x^2y - xy, \ xy^2\}$$

of I from the same example. Note that the reduced Gröbner basis $\mathcal{G}$ is contained in $\mathcal{H}$ and the B-border basis varies continuously in a 'neighborhood' of I. △

Just like for Gröbner bases, the fact that the map $\mathcal{N}_{\mathcal{H}}$ identifies B with R/I allows us to compute multiplication with g in R/I as

$$M_g = \begin{array}{c c} & \begin{array}{ccc} b_1 & \cdots & b_\delta \end{array} \\ \begin{array}{c} b_1 \\ \vdots \\ b_\delta \end{array} & \left[\begin{array}{ccc} | & & | \\ \mathcal{N}_{\mathcal{H}}(gb_1) & \cdots & \mathcal{N}_{\mathcal{H}}(gb_\delta) \\ | & & | \end{array} \right] \end{array}$$

in a $\mathbb{C}$-basis $\mathcal{B} = \{b_1, \dots, b_\delta\}$ for B. The columns are the expansions of $\{\mathcal{N}_\mathcal{H}(gb) \mid b \in \mathcal{B}\}$ in this basis.

3.4 Resultants and Macaulay matrices

In this section, we discuss a different algebraic technique for computing points defined by zero-dimensional ideals, based on *resultants*. More specifically, we consider *projective resultants* and postpone the discussion on (more general) *toric resultants* to Chapter 5. As the name suggests, the natural solution space for studying these resultants is the projective space. Throughout this section, we work with (homogeneous) polynomials in $S = \mathbb{C}[x_0, \dots, x_n] = \mathbb{C}[\mathbb{P}^n]$. The main results and their proofs can be found in [Jou91, GKZ94, Mac02] and [CLO06, Chapter 3] contains an accessible treatment with a view towards computations. First, we state the definition and some properties of resultants. This will allow us to describe very explicitly when a member of the square family $\mathcal{F}_S(d_1, \dots, d_n)$ is 'generic' with respect to some properties. That is, we will give equations for the variety of members that are *not*. Next, in Subsection 3.4.2 we will describe a construction due to Macaulay to compute the resultant and a way of constructing (homogeneous) multiplication maps using resultants.

3.4.1 Definition and properties

We consider the family of homogeneous polynomial systems $\mathcal{F}_S(d_0, \dots, d_n) \simeq S_{d_0} \times \cdots \times S_{d_n}$ given by $n+1$ homogeneous equations $f_0 = \cdots = f_n = 0$ over $\mathbb{P}^n$, with $f_i \in S_{d_i}$. Note that this is not a square family: we are considering $n+1$ equations on an n-dimensional solution space. Recall that $\mathcal{F}_S(d_0, \dots, d_n)$ is isomorphic to the affine space $\mathbb{C}^p = \mathbb{C}^{p_0} \times \cdots \times \mathbb{C}^{p_n}$ where $p_i = \binom{n+d_i}{n}$ via

$$\phi((c_{0,a})_{|a|=d_0}, \dots, (c_{n,a})_{|a|=d_n}) = \left(\sum_{|a|=d_0} c_{0,a} x^a, \dots, \sum_{|a|=d_n} c_{n,a} x^a \right).$$

Here $|a| = d_i$ means that a runs over all tuples $a = (a_0, a_1, \dots, a_n) \in \mathbb{N}^{n+1}$ satisfying $|a| = a_0 + \cdots + a_n = d_i$. Let us denote

$$A = \mathbb{C}[\mathbb{C}^p] = \mathbb{C}[(c_{0,a})_{|a|=d_0}, \dots, (c_{n,a})_{|a|=d_n}]$$

for the ring of polynomials whose variables represent the coefficients of a member of $\mathcal{F}_S(d_0, \dots, d_n)$. A property is said to hold for a *generic* member of $\mathcal{F}_S(d_0, \dots, d_n)$ if there is some polynomial $g \in A$ such that the property holds for $\phi(\mathbb{C}^p \setminus V_{\mathbb{C}^p}(g))$. Resultants are a powerful tool for finding such a polynomial g for many interesting properties of polynomial systems.

Definition 3.4.1 (Resultant)**.** A *resultant* of the family $\mathcal{F}_S(d_0, \dots, d_n)$ with $d_i \geq 1, i = 0, \dots, n$ is a polynomial $\mathrm{Res}_{d_0,\dots,d_n} \in A$ such that $\mathrm{Res}_{d_0,\dots,d_n}(a) = 0$ if and only if $\phi(a)$ represents a homogeneous system which has a solution in $\mathbb{P}^n$ and $\mathrm{Res}_{d_0,\dots,d_n}(a) = 1$ for the point $a \in \mathbb{C}^p$ with $a = \phi^{-1}(x_0^{d_0}, \dots, x_n^{d_n})$.

Note that the second condition on the polynomial $\mathrm{Res}_{d_0,\dots,d_n} \in A$ is just a scaling condition. We will use the notation $\mathrm{Res}_{d_0,\dots,d_n}(a) = \mathrm{Res}_{d_0,\dots,d_n}(f_0, \dots, f_n) = \mathrm{Res}(f_0, \dots, f_n)$ for $a = \phi^{-1}(f_0, \dots, f_n)$. The following theorem tells us that Definition 3.4.1 makes sense and it gives a selection of some of the interesting properties of the resultant.

Theorem 3.4.1. *For any tuple* $(d_0, \dots, d_n) \in \mathbb{N}_{>0}^{n+1}$ *a resultant* $\mathrm{Res} = \mathrm{Res}_{d_0,\dots,d_n}$ *exists and it is unique. Moreover, it has the following properties:*

1. Res *has coefficients in* $\mathbb{Z}$*,*
2. Res *is an irreducible polynomial,*
3. *each term of* Res *has degree* $d_0 \cdots d_{i-1} d_{i+1} \cdots d_n$ *in the variables* $(c_{i,a})_{|a|=d_i}$.

Proof. All of these statements and more are discussed in [CLO06, Chapter 3, §2 and 3] with proofs or full references. □

Example 3.4.1 (Sylvester resultant)**.** Let $S = \mathbb{C}[x, y]$ and consider two general homogeneous polynomials

$$f_0 = a_0 y^{d_0} + a_1 x y^{d_0 - 1} + \cdots + a_{d_0} x^{d_0}, \; f_1 = b_0 y^{d_1} + b_1 x y^{d_1 - 1} + \cdots + b_{d_1} x^{d_1}.$$

In this example $A = \mathbb{C}[a_0, \dots, a_{d_0}, b_0, \dots, b_{d_1}]$. It is a classical result that f_0 and f_1 have a common root in $\mathbb{P}^1$ if and only if the determinant of the $(d_0 + d_1) \times (d_0 + d_1)$ matrix

$$\mathrm{Syl}(f_0, f_1) = \begin{array}{c} \\ y^{d_0+d_1-1} \\ xy^{d_0+d_1-2} \\ \vdots \\ x^{d_0}y^{d_1-1} \\ x^{d_0+1}y^{d_1-2} \\ \vdots \\ x^{d_0+d_1-1} \end{array} \begin{array}{c} \begin{array}{ccccccc} y^{d_1-1} & xy^{d_1-2} & \cdots & x^{d_1-1} & y^{d_0-1} & \cdots & x^{d_0-1} \end{array} \\ \begin{bmatrix} a_0 & & & & b_0 & & \\ a_1 & a_0 & & & b_1 & \ddots & \\ \vdots & a_1 & \ddots & & \vdots & \ddots & b_0 \\ a_{d_0} & \vdots & \ddots & a_0 & \vdots & & b_1 \\ & a_{d_0} & & a_1 & b_{d_1} & & \vdots \\ & & \ddots & \vdots & & \ddots & \vdots \\ & & & a_{d_0} & & & b_{d_1} \end{bmatrix} \end{array} \tag{3.4.1}$$

with coefficients a_i appearing in the first d_1 columns and b_i in the last d_0 columns, is zero (see [CLO06, Chapter 3, §1] and [CLO13, Chapter 3, §6] for the affine version).

The indexing of the rows and columns by monomials comes from the interpretation of $\mathrm{Syl}(f_0, f_1)$ as the matrix representation of the linear map

$$S_{d_1-1} \times S_{d_0-1} \to S_{d_0+d_1-1} \quad \text{given by} \quad (q_0, q_1) \mapsto q_0 f_0 + q_1 f_1$$

in monomial bases for $S_{d_1-1} \times S_{d_0-1}$ and $S_{d_0+d_1-1}$ (e.g. for S_{d_1-1} the basis $\{y^{d_1-1}, xy^{d_1-2}, \ldots, x^{d_1-1}\}$ is used). We set $\mathrm{Res}_{d_0,d_1} = \det(\mathrm{Syl}(f_0, f_1))$ and one can trivially check that Res satisfies the scaling condition $\mathrm{Res}(y^{d_0}, x^{d_1}) = 1$ (we let x play the role of x_1 and y the role of x_0 in Definition 3.4.1). △

Example 3.4.2 (The determinant of a square matrix)**.** The resultant $\mathrm{Res}_{1,1,\ldots,1}$ is the determinant of the matrix $(c_{i,e_j})_{0 \le i,j \le n}$ where e_j is the exponent vector corresponding to x_j. △

Remark 3.4.1. To gain some more insight in property 3 of Theorem 3.4.1, suppose that we let the coefficients of the polynomials $f_1, \ldots, f_n$ take on generic values $(c^*_{i,a})_{|a|=d_i}, i = 1, \ldots, n$. We investigate the condition on the coefficients $(c_{0,a})_{|a|=d_0}$ of f_0 such that $f_0 = f_1 = \ldots = f_n$ has a solution in $\mathbb{P}^n$. The condition that $f_0(\zeta) = 0$ for some $\zeta \in \mathbb{P}^n$ imposes a linear condition on the $(c_{0,a})_{|a|=d_0}$. Hence, for each of the common zeros $\zeta \in V_{\mathbb{P}^n}(f_1, \ldots, f_n)$ we get a linear condition $l_\zeta \in \mathbb{C}[(c_{0,a})_{|a|=d_0}]$. Then we have that $V_{\mathbb{P}^n}(f_0) \cap V_{\mathbb{P}^n}(f_1, \ldots, f_n)$ is nonempty if and only if $\prod_{\zeta \in V_{\mathbb{P}^n}(f_1,\ldots,f_n)} l_\zeta = 0$. By Bézout's theorem 3.2.2 this is a homogeneous polynomial of degree $d_1 \cdots d_n$. △

To conclude this subsection, we state some genericity conditions which we have used in previous subsections in terms of resultants.

- In Subsection 3.2.3 we stated that for a general member $(\hat{f}_1, \ldots, \hat{f}_n) \in \mathcal{F}_R(d_1, \ldots, d_n)$ the homogenization $(f_1, \ldots, f_n) = (\eta_{d_1}(\hat{f}_1), \ldots, \eta_{d_n}(\hat{f}_n))$ does not 'add' anything to the variety defined by $\hat{f}_1 = \cdots = \hat{f}_n = 0$, in the sense that $V_{\mathbb{P}^n}(f_1, \ldots, f_n)$ is generically contained in U_0. This is justified by the fact that $V_{\mathbb{P}^n}(f_1, \ldots, f_n)$ contains a point outside of U_0 if and only if

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

  has a common solution in the hyperplane 'at infinity'. Note that the $f_i(0, x_1, \ldots, x_n)$ are homogeneous of degree d_i in $x_1, \ldots, x_n$ and they have a common solution in $\mathbb{P}^{n-1}$ if and only if

  $$\mathrm{Res}_{d_1,\ldots,d_n}(f_1(0, x_1, \ldots, x_n), \ldots, f_n(0, x_1, \ldots, x_n)) = 0.$$

  This imposes a polynomial condition on the coefficients of $\hat{f}_1, \ldots, \hat{f}_n$ standing with the monomials of degree $d_1, \ldots, d_n$ respectively.

- A homogeneous version of the Jacobian condition of Remark 3.1.4 for a root $\zeta \in V_{\mathbb{P}^n}(f_1, \ldots, f_n)$ to have multiplicity > 1 is the following. For any set of homogeneous coordinates $z \in \mathbb{C}^{n+1}$ of ζ the gradient vectors

  $$\nabla f_i = \left(\frac{\partial f_i}{\partial x_0}(z), \ldots, \frac{\partial f_i}{\partial x_n}(z) \right) \in \mathbb{C}^{n+1}$$

must be linearly dependent. This gives $2n+1$ homogeneous equations

$$f_1 = \cdots = f_n = 0, \quad y_1 \nabla f_1 + \cdots + y_n \nabla f_n = 0$$

in the $2n+1$ variables $x_0, \ldots, x_n, y_1, \ldots, y_n$. These equations have more structure: they are homogeneous in the two sets of variables $\{x_0, \ldots, x_n\}$ and $\{y_1, \ldots, y_n\}$ separately. The meaningful solutions correspond to points in the product of projective spaces $\mathbb{P}^n \times \mathbb{P}^{n-1}$. The existence of such solutions corresponds to the vanishing of a *multihomogeneous resultant*. We omit the details and refer to [CLO06, Chapter 3, §5, Exercise 6]).

- Theorem 3.2.2 asserts that generic members of $\mathcal{F}_S(d_1, \ldots, d_n)$ have a zero-dimensional solution set. The condition for $V_{\mathbb{P}^n}(f_1, \ldots, f_n)$ to be positive dimensional is the following. For any hyperplane given by $f_0 = 0$, $f_0 \in S_1$ there is a nonempty intersection $V_{\mathbb{P}^n}(f_0) \cap V_{\mathbb{P}^n}(f_1, \ldots, f_n)$. This only happens for coefficients $(c^*_{i,a})_{|a|=d_i}, i = 1, \ldots, n$ that make the resultant $\text{Res}_{1,d_1,\ldots,d_n}$ identically equal to zero. This is equivalent to the vanishing of the coefficients of a degree $d_1 \cdots d_n$ polynomial in $c_{0,e_0}, \ldots, c_{0,e_n}$ where e_i is the exponent vector corresponding to x_i and each of these coefficients is a polynomial in the $(c^*_{i,a})_{|a|=d_i}, i = 1, \ldots, n$. In particular, the subvariety of $\mathcal{F}_S(d_1, \ldots, d_n)$ corresponding to systems with a positive dimensional solution set is contained in the variety of systems whose solution set intersects $V_{\mathbb{P}^n}(x_0)$. These are exactly the systems with solutions at infinity, whose variety we described above.

3.4.2 Macaulay matrices

There are several ways of using resultants for solving a system of polynomial equations numerically. One approach is via *u-resultants* which recover the coordinates of the points in $V_{\mathbb{P}^n}(I)$ via a generalized eigenvalue problem (see e.g. [JV05]). Another approach uses *hidden variable resultants* to eliminate variables from the equations. This leads to a *polynomial eigenvalue problem* which can be solved via, for instance, linearization or numerical contour integration techniques [GT17]. The hidden variable resultant approach has been studied quite extensively in the context of numerical computation, using different resultant constructions (Sylvester/Macaulay type as well as Bézoutian resultant constructions). The technique turns out to be quite effective, especially in the case where $n = 2$ [BKM05, SVBDL14, NNT15, Tel16]. We should mention that, even though in practice they usually give satisfying results, the fact that these methods 'project some variables away' makes them inherently numerically unstable. A proof and examples of worst-case scenarios are given in [NT16].

We will limit ourselves to the description of a way to obtain multiplication matrices from an important resultant construction of Macaulay. This is the resultant-based approach for solving equations that is most directly related to the methods proposed in this thesis. The Macaulay construction is a generalization of Sylvester's matrix (3.4.1) for the resultant of two homogeneous equations on $\mathbb{P}^1$. Our goal is to construct a matrix which we will call $\mathrm{Mac}_{d_0,\ldots,d_n}$ whose entries are coefficients of $f_0,\ldots,f_n$ (that is, variables of A), such that its determinant $\det \mathrm{Mac}_{d_0,\ldots,d_n} \in A$ is a nonzero (polynomial) multiple of the resultant $\mathrm{Res}_{d_0,\ldots,d_n} \in A$. As for the resultant, we will denote $\mathrm{Mac}_{d_0,\ldots,d_n}(a) = \mathrm{Mac}_{d_0,\ldots,d_n}(f_0,\ldots,f_n) = \mathrm{Mac}(f_0,\ldots,f_n)$ for $a = \phi^{-1}(f_0,\ldots,f_n)$. In the case where $n = 1$, we will have that $\mathrm{Mac}(f_0,f_1) = \mathrm{Syl}(f_0,f_1)$. Note that the image of the map represented by $\mathrm{Syl}(f_0,f_1)$ represents the degree $\hat{\rho} = d_0 + d_1 - 1$ part of the homogeneous ideal $\langle f_0, f_1\rangle$. Indeed, the columns are obtained by taking all monomial multiples of f_0, f_1 that result in a homogeneous equation of this degree. In the generalized construction, the columns of our matrix will represent polynomials in $\langle f_0,\ldots,f_n\rangle_{\hat{\rho}} \subset S_{\hat{\rho}}$ where

$$\hat{\rho} = d_0 + d_1 + \cdots + d_n - n. \tag{3.4.2}$$

More precisely, they will be monomial multiples of $f_0,\ldots,f_n$. In general, we will not multiply f_i with *all* monomials of degree $\hat{\rho} - d_i$, since this would not lead to a square matrix $\mathrm{Mac}_{d_0,\ldots,d_n}$ (and we cannot take the determinant). We denote the set of monomials of degree $\hat{\rho} - d_i$ by which we multiply f_i to obtain columns of $\mathrm{Mac}_{d_0,\ldots,d_n}$ by Σ_i. The set $\{\Sigma_0,\ldots,\Sigma_n\}$, indexing the columns of $\mathrm{Mac}_{d_0,\ldots,d_n}$, will correspond to a partitioning of the monomials of $S_{\hat{\rho}}$, indexing the rows of $\mathrm{Mac}_{d_0,\ldots,d_n}$. They are defined as follows:

$$\begin{aligned}
\Sigma'_n &= \{x^a \in S_{\hat{\rho}} \mid x_n^{d_n} \text{ divides } x^a\},\\
\Sigma'_{n-1} &= \{x^a \in S_{\hat{\rho}} \mid x_n^{d_n} \text{ does not divide } x^a \text{ but } x_{n-1}^{d_{n-1}} \text{ does }\},\\
&\ \ \vdots\\
\Sigma'_0 &= \{x^a \in S_{\hat{\rho}} \mid x_i^{d_i} \text{ does not divide } x^a \text{ for } i = 1,\ldots,n \text{ but } x_0^{d_0} \text{ does }\},
\end{aligned}$$

and $\Sigma_i = \{x^a / x_i^{d_i} \mid x^a \in \Sigma'_i\}$.

Example 3.4.3. Let $n = 2, d_0 = 1, d_1 = 3, d_2 = 2$. In this case, $\hat{\rho} = 4$ and we get

$$\begin{aligned}
\Sigma_2 &= \{x_0^2, x_0x_1, x_0x_2, x_1^2, x_1x_2, x_2^2\},\ \Sigma_1 = \{x_0, x_1, x_2\},\\
\Sigma_0 &= \{x_0^3, x_0^2x_1, x_0x_1^2, x_0^2x_2, x_0x_1x_2, x_1^2x_2\}.
\end{aligned}$$

The corresponding partitioning of the monomials in S_4 into Σ'_0, Σ'_1 and Σ'_2 is illustrated in Figure 3.4. In the figure, the monomial $x_0^{4-a_1-a_2}x_1^{a_1}x_2^{a_2}$ is identified with the lattice point (a_1, a_2). Denoting

$$\begin{aligned}
f_0 &= a_0x_0 + a_1x_1 + a_2x_2,\\
f_1 &= b_0x_0^3 + b_1x_0^2x_1 + b_2x_0^2x_2 + b_3x_0x_1^2 + b_4x_0x_1x_2 + b_5x_0x_2^2 + b_6x_1^3 + b_7x_1^2x_2\\
&\quad + b_8x_1x_2^2 + b_9x_2^2,\\
f_2 &= c_0x_0^2 + c_1x_0x_1 + c_2x_0x_2 + c_3x_1^2 + c_4x_1x_2 + c_5x_2^2,
\end{aligned}$$

we obtain the matrix $\mathrm{Mac}_{d_0,d_1,d_2}$ shown below.

$$
\begin{array}{c|cccccc|ccc|cccccc}
 & x_0^3 & x_0^2x_1 & x_0^2x_2 & x_0x_1^2 & x_0x_1x_2 & x_1^2x_2 & x_0 & x_1 & x_2 & x_0^2 & x_0x_1 & x_0x_2 & x_1^2 & x_1x_2 & x_2^2 \\
x_0^4 & a_0 & & & & & & b_0 & & & c_0 & & & & & \\
x_0^3x_1 & a_1 & a_0 & & & & & b_1 & b_0 & & c_1 & c_0 & & & & \\
x_0^3x_2 & a_2 & & a_0 & & & & b_2 & & b_0 & c_2 & & c_0 & & & \\
x_0^2x_1^2 & & a_1 & & a_0 & & & b_3 & b_1 & & c_3 & c_1 & & c_0 & & \\
x_0^2x_1x_2 & & a_2 & a_1 & & a_0 & & b_4 & b_2 & b_1 & c_4 & c_2 & c_1 & & c_0 & \\
x_0x_1^2x_2 & & & & a_2 & a_1 & a_0 & b_7 & b_4 & b_3 & & c_4 & c_3 & c_2 & c_1 & \\
\hline
x_0x_1^3 & & & & a_1 & & & b_6 & b_3 & & & c_3 & & c_1 & & \\
x_1^4 & & & & & & & & b_6 & & & & & c_3 & & \\
x_1^3x_2 & & & & & & a_1 & & b_7 & b_6 & & & & c_4 & c_3 & \\
\hline
x_0^2x_2^2 & & & a_2 & & & & b_5 & & b_2 & c_5 & & c_2 & & & c_0 \\
x_0x_1x_2^2 & & & & & a_2 & & b_8 & b_5 & b_4 & & c_5 & c_4 & & c_2 & c_1 \\
x_0x_2^3 & & & & & & & b_9 & & b_5 & & & c_5 & & & c_2 \\
x_1^2x_2^2 & & & & & & a_2 & & b_8 & b_7 & & & & c_5 & c_4 & c_3 \\
x_1x_2^3 & & & & & & & & b_9 & b_8 & & & & & c_5 & c_4 \\
x_2^4 & & & & & & & & & b_9 & & & & & & c_5
\end{array}
$$

Note that the columns of $\mathrm{Mac}_{d_0,d_1,d_2}$ are indexed by $\{\Sigma_0, \Sigma_1, \Sigma_2\}$ and the rows by $\{\Sigma_0', \Sigma_1', \Sigma_2'\}$ (recall that $\Sigma_i' = x_i^{d_i} \cdot \Sigma_i$). The column corresponding to $x_0^2x_2 \in \Sigma_0$ represents the polynomial $x_0^2x_2f_0$ in the monomial basis for S_4. $\triangle$

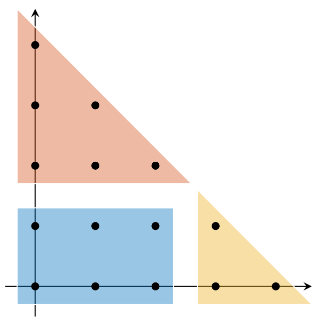


Figure 3.4: Illustration of the partitioning of S_4 into Σ_0' (blue), Σ_1' (yellow) and Σ_2' (orange) from Example 3.4.3.

Let us define the row vectors $\phi_{\Sigma_i}(x_0, \ldots, x_n) = (x^a \mid x^a \in \Sigma_i)$ where the ordering of the monomials is compatible with the indexing of the columns of $\mathrm{Mac}_{d_0,\ldots,d_n}$. That is, the columns are indexed by the vector

$$[\phi_{\Sigma_0}(x_0,\dots,x_n) \ \cdots \ \phi_{\Sigma_n}(x_0,\dots,x_n)].$$

In the same way, we define the row vectors $\phi_{\Sigma_i'}(x_0,\dots,x_n) = (x^a \mid x^a \in \Sigma_i')$ such that the order of the monomials is compatible with the row indexing of $\mathrm{Mac}_{d_0,\dots,d_n}$. Constructing the matrix $\mathrm{Mac}_{d_0,\dots,d_n}$ as illustrated in Example 3.4.3, one can check that $\mathrm{Mac}(x_0^{d_0},\dots,x_n^{d_n})$ is the identity matrix. This shows that $\det \mathrm{Mac}_{d_0,\dots,d_n} \in A$ is not the zero polynomial. Moreover, if $\zeta \in \mathbb{P}^n$ is such that $f_0(\zeta) = \cdots = f_n(\zeta) = 0$, then for any set of homogeneous coordinates $z \in \mathbb{C}^{n+1} \setminus \{0\}$ of ζ, we have that

$$[\phi_{\Sigma_0'}(z) \ \cdots \ \phi_{\Sigma_n'}(z)]\,\mathrm{Mac}(f_0,\dots,f_n) = [f_0(z)\phi_{\Sigma_0}(z) \ \dots \ f_n(z)\phi_{\Sigma_n}(z)] = 0.$$

This shows that if $f_0 = \cdots = f_n = 0$ has a solution in $\mathbb{P}^n$, $\det \mathrm{Mac}(f_0,\dots,f_n) = 0$, which implies that

$$\det \mathrm{Mac}_{d_0,\dots,d_n} \in \langle \mathrm{Res}_{d_0,\dots,d_n} \rangle$$

by the Nullstellensatz and the fact that $\mathrm{Res}_{d_0,\dots,d_n}$ is irreducible. Therefore, there is a nonzero polynomial E such that $\det \mathrm{Mac}_{d_0,\dots,d_n} = E \cdot \mathrm{Res}_{d_0,\dots,d_n}$. This polynomial E is called the *extraneous factor*. In his paper [Mac02], Macaulay identifies the extraneous factor as the determinant of a submatrix of $\mathrm{Mac}_{d_0,\dots,d_n}$, see also [CLO06, Chapter 3, §4].

In the construction of $\mathrm{Mac}_{d_0,\dots,d_n}$, the set Σ_0' consists of the $d_1 \cdots d_n$ elements

$$\Sigma_0' = \{x_0^{\hat{\rho}-a_1-\cdots-a_n} x_1^{a_1} \cdots x_n^{a_n} \mid a_i < d_i, i = 1,\dots,n\}.$$

Therefore, the number of elements in Σ_0' (and in Σ_0) is the Bézout number for the family $\mathcal{F}_S(d_1,\dots,d_n)$. We will see that this is no coincidence. In what follows, fix $(f_0,\dots,f_n) \in \mathcal{F}_S(d_0,\dots,d_n)$ and define $I = \langle f_1,\dots,f_n \rangle$. We partition the matrix $\mathrm{Mac}(f_0,\dots,f_n)$ into 4 submatrices as follows:

$$\mathrm{Mac}(f_0,\dots,f_n) = \begin{array}{cc} & \begin{array}{c|c} \Sigma_0 & \{\Sigma_1,\dots,\Sigma_n\} \end{array} \\ \begin{array}{c} \Sigma_0' \\ \hline \{\Sigma_1',\dots,\Sigma_n'\} \end{array} & \left[\begin{array}{c|c} M_{00} & M_{01} \\ \hline M_{10} & M_{11} \end{array}\right] \end{array}.$$

Here M_{00} and M_{11} are square matrices. Just like the Sylvester matrix, the matrix $\mathrm{Mac}(f_0,\dots,f_n)$ can be interpreted as a map

$$\mathrm{Mac}(f_0,\dots,f_n) : \Lambda_0 \times \Lambda_1 \times \cdots \times \Lambda_n \to \Lambda$$

where $\Lambda = S_{\hat{\rho}}$, $\Lambda_i = \mathrm{span}_{\mathbb{C}}(\Sigma_i)$, given by $\mathrm{Mac}(f_0,\dots,f_n)(q_0,\dots,q_n) = q_0 f_0 + \cdots q_n f_n$. The second block column of $\mathrm{Mac}(f_0,\dots,f_n)$ is the restriction of this map to $\Lambda_1 \times \cdots \times \Lambda_n$:

$$\begin{bmatrix} M_{01} \\ M_{11} \end{bmatrix} = \mathrm{Mac}(f_0,\dots,f_n)_{|\Lambda_1 \times \cdots \times \Lambda_n}.$$

Note that the image of $\mathrm{Mac}(f_0,\dots,f_n)_{|\Lambda_1\times\cdots\times\Lambda_n}$ is contained in $I_{\hat\rho}$. The following is the main result of this subsection. It uses some terminology from Subsection 3.2.2.

Theorem 3.4.2. *For any* $(d_0,\dots,d_n)\in\mathbb{N}_{>0}^{n+1}$, *let* $(f_0,\dots,f_n)\in\mathcal{F}_S(d_0,\dots,d_n)$. *Suppose that* $I=\langle f_1,\dots,f_n\rangle\subset S$ *is such that* $V_{\mathbb{P}^n}(I)=\{\zeta_1,\dots,\zeta_\delta\}$ *consists of* $\delta=d_1\cdots d_n$ *points with multiplicity 1 and the submatrix* M_{11} *of* $\mathrm{Mac}(f_0,\dots,f_n)$ *is invertible. Then*

1. $V_{\mathbb{P}^n}(I)\subset U_0$,
2. $\{x^a+I_\rho \mid x^a\in\Sigma_0\}$ *is a* $\mathbb{C}$*-basis for* $(S/I)_\rho$ *where* $\rho=\hat\rho-d_0$,
3. *the Schur complement* $M_{00}-M_{01}M_{11}^{-1}M_{10}$ *is the homogeneous multiplication map* $M_{f_0/x_0^{d_0}}:(S/I)_\rho\to(S/I)_\rho$ *in this basis,*
4. $\det\mathrm{Mac}(f_0,\dots,f_n)=\det(M_{11})\prod_{i=1}^{\delta}\frac{f_0}{x_0^{d_0}}(\zeta_i)$.

Proof. For the first statement, suppose that $\zeta\in V_{\mathbb{P}^n}(I)\in\mathbb{P}^n\setminus U_0$. For any set of homogeneous coordinates $z\in\mathbb{C}^{n+1}\setminus\{0\}$ for ζ, this gives

$$\begin{aligned}
[\phi_{\Sigma_0'}(z)\ \phi_{\Sigma_1'}(z)\ \cdots\ \phi_{\Sigma_n'}(z)]\begin{bmatrix}M_{01}\\M_{11}\end{bmatrix} &= [0\ \phi_{\Sigma_1'}(z)\ \cdots\ \phi_{\Sigma_n'}(z)]\begin{bmatrix}M_{01}\\M_{11}\end{bmatrix}\\
&= [\phi_{\Sigma_1'}(z)\ \cdots\ \phi_{\Sigma_n'}(z)]M_{11}\\
&= [f_1(z)\phi_{\Sigma_1}(z)\ \cdots\ f_n(z)\phi_{\Sigma_n}(z)]=0.
\end{aligned}$$

Here $\phi_{\Sigma_0'}(z)=0$ since $\Sigma_0'=x_0^{d_0}\cdot\Sigma_0$ and $\zeta\notin U_0$. This contradicts the assumption that M_{11} is invertible.

To show the second statement, note that $\rho,\hat\rho\in\mathrm{Reg}(I)$ by Theorem 3.2.3. Since $\mathrm{HF}_I(\hat\rho)=d_1\cdots d_n=\#(\Sigma_0)$ we have that the image of $\mathrm{Mac}(f_0,\dots,f_n)_{|\Lambda_1\times\cdots\times\Lambda_n}$, which has codimension $d_1\cdots d_n$ in $S_{\hat\rho}$ by the assumption that M_{11} is full rank, is $I_{\hat\rho}$. This also shows that the elements of $\{x^a+I_{\hat\rho}\mid x^a\in\Sigma_0'\}$ form a basis for $(S/I)_{\hat\rho}$. The second statement now follows from the fact that $M_{x_0^{d_0}}:(S/I)_\rho\to(S/I)_{\hat\rho}$ is an isomorphism (Lemma 3.2.1).

For the third statement, we define $M_{f_0/x_0^{d_0}}=M_{00}-M_{01}M_{11}^{-1}M_{10}$ and show that it is indeed multiplication with $f_0/x_0^{d_0}$ in $(S/I)_\rho$. For any set of homogeneous coordinates z of $\zeta\in V_{\mathbb{P}^n}(I)$ we observe that

$$\begin{aligned}
&[\phi_{\Sigma_0'}(z)\ \phi_{\Sigma_1'}(z)\ \cdots\ \phi_{\Sigma_n'}(z)]\begin{bmatrix}M_{00}&M_{01}\\M_{10}&M_{11}\end{bmatrix}\begin{bmatrix}\mathrm{id}&0\\-M_{11}^{-1}M_{10}&\mathrm{id}\end{bmatrix}\\
&\qquad= [\phi_{\Sigma_0'}(z)\ \phi_{\Sigma_1'}(z)\ \cdots\ \phi_{\Sigma_n'}(z)]\begin{bmatrix}M_{f_0/x_0^{d_0}}&M_{01}\\0&M_{11}\end{bmatrix}\\
&\qquad= [f_0(z)\phi_{\Sigma_0}(z)\ 0\ \cdots\ 0]\begin{bmatrix}\mathrm{id}&0\\-M_{11}^{-1}M_{10}&\mathrm{id}\end{bmatrix},
\end{aligned}$$

where 'id' are identity matrices of the appropriate size. It follows that

$$\phi_{\Sigma_0'}(z) M_{f_0/x_0^{d_0}} = f_0(z)\phi_{\Sigma_0}(z).$$

Using $\phi_{\Sigma_0'}(z) = x_0^{d_0}\phi_{\Sigma_0}(z)$ we obtain

$$\phi_{\Sigma_0}(z) M_{f_0/x_0^{d_0}} = \frac{f_0}{x_0^{d_0}}(z)\phi_{\Sigma_0}(z).$$

This shows that the eigenvalues of $M_{f_0/x_0^{d_0}}$ are indeed the evaluations of the rational function $f_0/x_0^{d_0}$ at the roots of I. We now show that the eigenvectors are also the correct ones. For any $h \in S_\rho$ such that $h(\zeta) \neq 0$ for all $\zeta \in V_{\mathbb{P}^n}(I)$, let $\mathrm{ev}_\zeta : (S/I)_\rho \to \mathbb{C}$ defined by $f + I_\rho \mapsto (f/h)(\zeta)$ be the corresponding element of $(S/I)^\vee$. We think of ev_ζ as a row vector, represented in the basis Σ_0 of $(S/I)_\rho$. Then $\phi_{\Sigma_0}(z) = h(z)\,\mathrm{ev}_\zeta$ together with Theorem 3.2.4 shows the third statement.

The fourth statement is obtained from

$$\det \mathrm{Mac}(f_0,\ldots,f_n) = \det\left(\begin{bmatrix} M_{00} & M_{01} \\ M_{10} & M_{11}\end{bmatrix}\begin{bmatrix} \mathrm{id} & 0 \\ -M_{11}^{-1}M_{10} & \mathrm{id}\end{bmatrix}\right) = \det\begin{bmatrix} M_{f_0/x_0^{d_0}} & M_{01} \\ 0 & M_{11}\end{bmatrix}.$$

□

Example 3.4.4. Consider the case where $n = 1$, $f_0 = x$, $f_1 = c_0 y^{d_1} + c_1 y^{d_1-1}x + \cdots + c_{d_1}x^{d_1}$ and we use $x_0 = y$, $x_1 = x$ for the definition of the Macaulay construction. We find $\hat{\rho} = d_1$ and

$$\Sigma_0 = \{y^{d_1-1}, xy^{d_1-2}, \ldots, x^{d_1-1}\}, \quad \Sigma_1 = \{1\}$$

which gives

$$\mathrm{Syl}(f_0,f_1) = \mathrm{Mac}(f_0,f_1) = \begin{array}{c} \\ y^{d_1} \\ xy^{d_1-1} \\ xy^{d_1-2} \\ \vdots \\ \hline x^{d_1} \end{array} \begin{array}{c} \begin{array}{cccc|c} y^{d_1-1} & xy^{d_1-2} & \cdots & x^{d_1-1} & 1 \end{array} \\ \left[\begin{array}{cccc|c} & & & & c_0 \\ 1 & & & & c_1 \\ & 1 & & & c_2 \\ & & \ddots & & \vdots \\ \hline & & & 1 & c_{d_1} \end{array}\right] \end{array}$$

and the Schur complement $M_{00} - M_{01}M_{11}^{-1}M_{10}$ is the Frobenius companion matrix of $f_1(x,1)$. △

The condition that M_{11} is invertible clearly imposes a determinantal condition on the coefficients of the f_i. This determinant is not the zero polynomial, which makes sure this condition holds for general members of $\mathcal{F}_S(d_0,\ldots,d_n)$ (see [Emi96, Lemma 4.4]). We make three remarks and end the subsection with an extension of Example 3.4.3.

Remark 3.4.2. Theorem 3.4.2 implies the following for systems of equations on $\mathbb{C}^n$. Suppose that for a member $(\hat{f}_1, \dots, \hat{f}_n) \in \mathcal{F}_R(d_1, \dots, d_n)$ with $R = \mathbb{C}[y_1, \dots, y_n]$ the homogenization $(f_1, \dots, f_n) \in \mathcal{F}_S(d_1, \dots, d_n)$ defines a zero-dimensional projective variety whose points have multiplicity one and for some $f_0 \in S_{d_0}$ the submatrix M_{11} of $\mathrm{Mac}(f_0, f_1, \dots, f_n)$ is invertible. Then

$$\{y_1^{a_1} \cdots y_n^{a_n} + \langle \hat{f}_1, \dots, \hat{f}_n \rangle \mid a_i < d_i, i = 1, \dots, n\}$$

is a basis for $R/\langle \hat{f}_1, \dots, \hat{f}_n \rangle$ and the Schur complement $M_{00} - M_{01}M_{11}^{-1}M_{01}$ is multiplication with $f_0(1, y_1, \dots, y_n)$ in $R/\langle \hat{f}_1, \dots, \hat{f}_n \rangle$ represented in this basis. This is what we observed in Example 3.4.4 in the case where $n = 1$. △

Remark 3.4.3. Note that in the situation of Theorem 3.4.2 the Schur complement can be written as the matrix product

$$M_{00} - M_{01}M_{11}^{-1}M_{10} = \begin{bmatrix} \mathrm{id} & -M_{01}M_{11}^{-1} \end{bmatrix} \begin{bmatrix} M_{00} \\ M_{10} \end{bmatrix},$$

where the first factor satisfies

$$\begin{bmatrix} \mathrm{id} & -M_{01}M_{11}^{-1} \end{bmatrix} \begin{bmatrix} M_{01} \\ M_{11} \end{bmatrix} = 0.$$

Since M_{11} is invertible, it follows that the kernel of the linear map $\begin{bmatrix} \mathrm{id} & -M_{01}M_{11}^{-1} \end{bmatrix}$ is the image of $\mathrm{Mac}(f_0, \dots, f_n)_{|\Lambda_1 \times \cdots \times \Lambda_n} : \Lambda_1 \times \cdots \times \Lambda_n \to \Lambda$, which is $I_{\hat{\rho}}$. That is, $\begin{bmatrix} \mathrm{id} & -M_{01}M_{11}^{-1} \end{bmatrix}$ represents a linear map $N : S_{\hat{\rho}} \to \mathbb{C}^\delta$ such that

$$0 \longrightarrow I_{\hat{\rho}} \longrightarrow S_{\hat{\rho}} \xrightarrow{N} \mathbb{C}^\delta \longrightarrow 0$$

is a short exact sequence, and $N_{x_0^{d_0}} : S_\rho \to \mathbb{C}^\delta$ given by $N_{x_0^{d_0}}(f) = N(x_0^{d_0} f)$ is onto. Such a map will give rise to a *homogeneous normal form*, a concept that we will define in Section 4.5. One can check that if, $d_0 = 1$ and $f_0 = x_i$ for some i, then

$$\begin{bmatrix} \mathrm{id} & -M_{01}M_{11}^{-1} \end{bmatrix} \begin{bmatrix} M_{00} \\ M_{10} \end{bmatrix}$$

is merely a 'column selection' of the matrix N. All this indicates that a homogeneous normal form with respect to I (in a large enough degree $\hat{\rho}$) gives us all the information we need to compute the homogeneous multiplication operators. △

Remark 3.4.4. Another, equivalent way to state that the kernel of N is the image of $\mathrm{Mac}(f_0, \dots, f_n)_{|\Lambda_1 \times \cdots \times \Lambda_n}$ is to say that N is the *cokernel map* of $\mathrm{Mac}(f_0, \dots, f_n)_{|\Lambda_1 \times \cdots \times \Lambda_n}$. The terminology used in numerical linear algebra literature is that N is the *left nullspace* of $\mathrm{Mac}(f_0, \dots, f_n)_{|\Lambda_1 \times \cdots \times \Lambda_n}$. Since the image of $\mathrm{Mac}(f_0, \dots, f_n)_{|\Lambda_1 \times \cdots \times \Lambda_n}$ is the same as the image of

$$S_{\hat{\rho}-d_1} \times \cdots \times S_{\hat{\rho}-d_n} \to S_{\hat{\rho}} \quad \text{with} \quad (q_1, \dots, q_n) \mapsto q_1 f_1 + \cdots + q_n f_n, \tag{3.4.3}$$

N may also be obtained as the cokernel of this map. Examples 3.4.1 and 3.4.3 illustrate that the two maps are the same for $n = 1, 2$. The definition of the map (3.4.3) seems slightly more natural or 'intuitive' than the one coming from the Macaulay resultant matrix, which restricts this map to subspaces of the $S_{\hat{\rho}-d_i}$ which give the same image. Computing the cokernel of (3.4.3) instead of $\mathrm{Mac}(f_0, \ldots, f_n)_{|\Lambda_1 \times \cdots \times \Lambda_n}$, although mathematically equivalent, gives better results numerically (even for generic systems of equations). We will illustrate this in Example 4.3.1. The use of cokernels of maps like (3.4.3) for polynomial root finding in affine and projective space is studied extensively in a numerical linear algebra context in the work of Dreesen, Batselier and De Moor [DBDM12, Dre13, Bat13, BDDM14]. $\triangle$

Example 3.4.5 (Example 3.4.3 continued)**.** Theorem 3.4.2 tells us that if the submatrix M_{11} of $\mathrm{Mac}(f_0, \ldots, f_n)$ is invertible, there cannot be any roots at infinity (the proof of this statement does not need the assumption of zero-dimensionality on $I = \langle f_0, \ldots, f_n \rangle$). This implies that if there are roots at infinity, $\det M_{11}$ must be zero. So the assumption that M_{11} is invertible fails when $f_1 = \cdots = f_n$ has solutions at 'infinity' (i.e., outside of U_0). However, this may not be the only case for which the condition is not satisfied. We investigate this for the matrix of Example 3.4.3. As we saw in Subsection 3.4.1, the equations $f_1 = f_2 = 0$ define solutions outside of U_0 if and only if the polynomial $\mathrm{Res}_\infty \in A$ vanishes, where Res_∞ is defined as

$$\mathrm{Res}_\infty = \mathrm{Res}_{3,2}(f_1(0, x_1, x_2), f_2(0, x_1, x_2)) = \det \begin{bmatrix} c_3 & & & b_6 & \\ c_4 & c_3 & & b_7 & b_6 \\ c_5 & c_4 & c_3 & b_8 & b_7 \\ & c_5 & c_4 & b_9 & b_8 \\ & & c_5 & & b_9 \end{bmatrix}.$$

Using Macaulay2, we find that

$$\det M_{11} = c_5 \left(b_9 c_3 c_4 - b_8 c_3 c_5 + b_6 c_5^2 \right) \mathrm{Res}_\infty .$$

This confirms that $\det M_{11}$ vanishes whenever $f_1 = f_2 = 0$ has roots 'at infinity', but it will also vanish when either $c_5 = 0$ or $b_9 c_3 c_4 - b_8 c_3 c_5 + b_6 c_5^2 = 0$. $\triangle$

Chapter 4

Truncated normal forms

This chapter introduces a new algebraic approach for solving zero-dimensional systems of polynomial equations. The key concept is that of a *truncated normal form*, which generalizes Gröbner and border bases (Section 3.3) as well as the resultant method described in Section 3.4. One of the main issues that is addressed by truncated normal forms is the following. Neither Gröbner/border bases nor resultants allow for a *way of choosing a basis* for the quotient algebra related to a zero-dimensional ideal *based on the numerical properties of the problem of computing multiplication operators in this basis*. This was mentioned as an open problem in [Mou07]. A solution is proposed in our first paper [TVB18], where the system is assumed to be a generic member of $\mathcal{F}_R(d_1, \ldots, d_n)$ in the sense that there are $d_1 \cdots d_n$ many roots in $\mathbb{C}^n$, counting multiplicities. The key idea is to let the basis be picked by a *QR factorization with optimal column pivoting*, which is a standard tool in numerical linear algebra. It was pointed out to the author by Tomas Pajdla that the bad numerical behavior of standard monomials coming from Gröbner bases for the computation of multiplication matrices was also noticed in the computer vision community. The authors of [BJA07, BJA08] use both QR and SVD techniques for basis selection on some problem-specific matrix constructions. The definition of truncated normal forms was first given in [TMVB18]. Next to developing the theory of the truncated normal form framework, the article proposes explicit algorithms for solving several families of systems, including $\mathcal{F}_R(d_1, \ldots, d_n)$, for which the algorithm is a reinterpretation of the algorithm in [TVB18]. Other families of systems considered in [TMVB18] are the polyhedral families discussed in Chapter 5, the homogeneous families $\mathcal{F}_S(d_1, \ldots, d_n)$ and multihomogeneous families. As mentioned above, the framework allows for a systematic way of selecting a basis for the quotient algebra which behaves well for numerical computations. As we will show in examples, these bases lead *rarely* to 'connected to 1' subspaces, let alone order ideals (see Subsection 3.3.2 for definitions). In a follow-up paper [MTVB19] some generalizations and modifications of the algorithms in [TMVB18] are proposed.
The content of this chapter is strongly based on the papers [TVB18, TMVB18,

MTVB19]. In Section 4.1 we give a motivating example for developing the framework of truncated normal forms, which is done in Section 4.2. We use the results of Section 4.2 to give an explicit numerical linear algebra based algorithm for solving generic members of $\mathcal{F}_R(d_1, \ldots, d_n)$ in Section 4.3. Section 4.4 discusses some ideas to make the algorithm more efficient and the use of non-monomial bases for the algebra R/I. In particular, we consider bases coming from using the SVD for basis selection and (tensor product) Chebyshev bases. Finally, Section 4.5 describes *homogeneous normal forms* for root finding in $\mathbb{P}^n$. The algorithms in this chapter focus on the isomorphic families $\mathcal{F}_R(d_1, \ldots, d_n)$ and $\mathcal{F}_S(d_1, \ldots, d_n)$. Generalizations to other (polyhedral) families, as introduced in [TMVB18, Section 4] and in [Tel20] for the homogeneous case, will be given in Chapter 5.

4.1 A motivating example

Let $R = \mathbb{C}[x, y]$ and consider the family $\mathcal{F}_R(2, 2)$ of polynomial systems with two equations in two unknowns of degree at most two. A member $(f_1, f_2) \in \mathcal{F}_R(2, 2)$ is given by

$$\begin{aligned} f_1 &= a_0 + a_1 x + a_2 y + a_3 x^2 + a_4 xy + a_5 y^2, \\ f_2 &= b_0 + b_1 x + b_2 y + b_3 x^2 + b_4 xy + b_5 y^2. \end{aligned}$$

For any values of $a_i, b_i \in \mathbb{C}$, these two polynomials generate an ideal $I = \langle f_1, f_2 \rangle \in R$ for which we want to compute $V_{\mathbb{C}^2}(I)$. The a_i, b_i are the variables of the coordinate ring $A = \mathbb{C}[a_0, \ldots, a_5, b_0, \ldots, b_5]$ of the affine variety $\mathbb{C}^{12}$ parametrizing our family. Motivated by the results of Subsection 3.1.1, we want to compute the multiplication maps $M_x : R/I \to R/I$ and $M_y : R/I \to R/I$ in some basis of R/I. With the appropriate genericity assumptions (see Subsection 3.1.2), we know that this basis should consist of four elements. Suppose we want to work with the basis $\mathcal{B} + I = \{b + I \mid b \in \mathcal{B}\}$ where $\mathcal{B} = \{1, x, y, xy\}$. If we can compute the representations

$$\begin{aligned} x^2 + I &= -c_{1,1} - c_{2,1} x - c_{3,1} y - c_{4,1} xy + I, \\ y^2 + I &= -c_{1,2} - c_{2,2} x - c_{3,2} y - c_{4,2} xy + I, \\ x^2 y + I &= -c_{1,4} - c_{2,4} x - c_{3,4} y - c_{4,4} xy + I, \\ xy^2 + I &= -c_{1,5} - c_{2,5} x - c_{3,5} y - c_{4,5} xy + I \end{aligned} \tag{4.1.1}$$

of x^2, y^2, x^2y, xy^2 modulo I (the indexing of the coefficients $c_{i,j} \in \mathbb{C}$ and the minus signs will soon make sense), then the multiplication matrices M_x, M_y in the basis $\mathcal{B} + I$ are given by

$$M_x = \begin{array}{c c} & \begin{array}{cccc} 1 & x & y & xy \end{array} \\ \begin{array}{c} 1 \\ x \\ y \\ xy \end{array} & \begin{bmatrix} 0 & -c_{1,1} & 0 & -c_{1,4} \\ 1 & -c_{2,1} & 0 & -c_{2,4} \\ 0 & -c_{3,1} & 0 & -c_{3,4} \\ 0 & -c_{4,1} & 1 & -c_{4,4} \end{bmatrix} \end{array}, \quad M_y = \begin{array}{c c} & \begin{array}{cccc} 1 & x & y & xy \end{array} \\ \begin{array}{c} 1 \\ x \\ y \\ xy \end{array} & \begin{bmatrix} 0 & 0 & -c_{1,2} & -c_{1,5} \\ 0 & 0 & -c_{2,2} & -c_{2,5} \\ 1 & 0 & -c_{3,2} & -c_{3,5} \\ 0 & 1 & -c_{4,2} & -c_{4,5} \end{bmatrix} \end{array}. \tag{4.1.2}$$

The coefficients $c_{i,j}$ depend, of course, on the specialization to $\mathbb{C}^{12}$ of the parameters a_i, b_i. In order to compute (4.1.1), we consider the so-called *resultant map*

$$\mathrm{res} : R_{\leq 1} \times R_{\leq 1} \to R_{\leq 3} \quad \text{given by} \quad \mathrm{res}(q_1, q_2) = q_1 f_1 + q_2 f_2.$$

Using the bases $\{1, x, y\}$ for $R_{\leq 1}$ and $\{1, x, y, xy, \ldots, y^3\}$ for $R_{\leq 3}$, this map is represented by

$$\mathrm{res} = \begin{array}{cc} & \begin{array}{cccccc} 1 & x & y & 1 & x & y \end{array} \\ \begin{array}{c} 1 \\ x \\ y \\ xy \\ x^2 \\ y^2 \\ x^3 \\ x^2y \\ xy^2 \\ y^3 \end{array} & \left[\begin{array}{cccccc} a_0 & & & b_0 & & \\ a_1 & a_0 & & b_1 & b_0 & \\ a_2 & & a_0 & b_2 & & b_0 \\ a_4 & a_2 & a_1 & b_4 & b_2 & b_1 \\ \hline a_3 & a_1 & & b_3 & b_1 & \\ a_5 & & a_2 & b_5 & & b_2 \\ & a_3 & & & b_3 & \\ & a_4 & a_3 & & b_4 & b_3 \\ & a_5 & a_4 & & b_5 & b_4 \\ & & a_5 & & & b_5 \end{array}\right] \end{array}.$$

Note that the columns of this matrix correspond to the polynomials

$$f_1, xf_1, yf_1, f_2, xf_2, yf_2 \in I \cap R_{\leq 3}.$$

In fact, from the definition of res it is clear that $\operatorname{im} \mathrm{res} \subset I \cap R_{\leq 3}$, so applying res to any column vector of length 6 gives us an element in $I \cap R_{\leq 3}$. Assuming that the considered member of $\mathcal{F}_R(2,2)$ is generic, the submatrix of res consisting of its last 6 rows is invertible (see Subsection 3.4.2) and we can find particularly nice elements of $I \cap R_{\leq 3}$ by computing

$$\begin{array}{cc} & \begin{array}{cccccc} 1 & x & y & 1 & x & y \end{array} \\ \begin{array}{c} 1 \\ x \\ y \\ xy \\ x^2 \\ y^2 \\ x^3 \\ x^2y \\ xy^2 \\ y^3 \end{array} & \left[\begin{array}{cccccc} a_0 & & & b_0 & & \\ a_1 & a_0 & & b_1 & b_0 & \\ a_2 & & a_0 & b_2 & & b_0 \\ a_4 & a_2 & a_1 & b_4 & b_2 & b_1 \\ \hline a_3 & a_1 & & b_3 & b_1 & \\ a_5 & & a_2 & b_5 & & b_2 \\ & a_3 & & & b_3 & \\ & a_4 & a_3 & & b_4 & b_3 \\ & a_5 & a_4 & & b_5 & b_4 \\ & & a_5 & & & b_5 \end{array}\right] \end{array} \begin{bmatrix} a_3 & a_1 & & b_3 & b_1 & \\ a_5 & & a_2 & b_5 & & b_2 \\ & a_3 & & & b_3 & \\ & a_4 & a_3 & & b_4 & b_3 \\ & a_5 & a_4 & & b_5 & b_4 \\ & & a_5 & & & b_5 \end{bmatrix}^{-1} = \begin{array}{cc} & \begin{array}{cccccc} g_1 & g_2 & g_3 & g_4 & g_5 & g_6 \end{array} \\ \begin{array}{c} 1 \\ x \\ y \\ xy \\ x^2 \\ y^2 \\ x^3 \\ x^2y \\ xy^2 \\ y^3 \end{array} & \left[\begin{array}{cccccc} c_{1,1} & & & & \cdots & c_{1,6} \\ c_{2,1} & & & & \cdots & c_{2,6} \\ c_{3,1} & & & & \cdots & c_{3,6} \\ c_{4,1} & & & & \cdots & c_{4,6} \\ \hline 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & \\ & & & & & 1 \end{array}\right] \end{array}.$$

This gives the polynomials $g_1, \ldots, g_6$, of which g_1, g_2, g_4, g_5 establish the representations (4.1.1). Notice that, in particular, $\mathcal{H} = \{g_1, g_2, g_4, g_5\}$ is a reduced B-border basis for

I, with $B = \mathrm{span}_{\mathbb{C}}(\mathcal{B}) \subset R$ (B has the connected to 1 property) and with respect to the monomial basis $\partial\mathcal{B} = \{x^2, y^2, x^2y, xy^2\}$ of $\partial B = B^+/B$. In this computation, we have computed two 'extra' rewriting rules modulo I, given by g_3, g_6.

An important observation is that we could play the same game for *any* $\mathcal{B}$ consisting of four monomials such that the square submatrix of res corresponding to the rows *not* indexed by $\mathcal{B}$ is invertible. We will denote the determinant of this submatrix by $D_{\mathcal{B}} \in A$, and the evaluation for a specific instance by $D_{\mathcal{B}}(f_1, f_2)$. Another restriction we impose on $\mathcal{B}$ is that the result of the computation allows us to construct multiplication matrices as in (4.1.1) and (4.1.2). For this we need that the monomials in $\partial\mathcal{B}$ (i.e. the monomials outside $\mathcal{B}$ obtained from multiplying the monomials in $\mathcal{B}$ with x and y) are contained in $R_{\leq 3}$. We conclude that we can pick any four element subset $\mathcal{B}$ of $\mathcal{W} = \{1, x, y, x^2, xy, y^2\}$ such that $D_{\mathcal{B}}(f_1, f_2) \neq 0$. The algorithm goes as follows. Let $\mathcal{V} = \{1, x, y, x^2, xy, y^2, x^3, x^2y, xy^2, y^3\}$ be the set of all monomials of degree at most 3. For any four element subset $\mathcal{B} \subset \mathcal{W}$ such that $D_{\mathcal{B}}(f_1, f_2) \neq 0$, construct the matrix of res such that its first 4 rows are indexed by $\mathcal{B}$:

$$\mathrm{res} = \begin{array}{c} \mathcal{B} \\ \mathcal{V}\backslash\mathcal{B} \end{array} \begin{bmatrix} M_{01} \\ M_{11} \end{bmatrix}.$$

Multiply res by M_{11}^{-1} (which makes sense because by construction $D_{\mathcal{B}}(f_1, f_2) = \det M_{11}$) to obtain

$$\mathrm{res}\, M_{11}^{-1} = \begin{array}{c} \mathcal{B} \\ \mathcal{V}\backslash\mathcal{B} \end{array} \begin{bmatrix} M_{01}M_{11}^{-1} \\ \mathrm{id} \end{bmatrix} = \begin{array}{c} \mathcal{B} \\ \mathcal{V}\backslash\mathcal{B} \end{array} \begin{bmatrix} C \\ \mathrm{id} \end{bmatrix}.$$

The columns of the result give rewriting rules analogous to (4.1.1) for $\mathcal{V} \setminus \mathcal{B}$ modulo I, which directly gives us the multiplication matrices in the basis $\mathcal{B} + I$ since $\partial\mathcal{B} \subset \mathcal{V} \setminus \mathcal{B}$. Indeed, all that is left to do is plug in the (negative of the) entries of the matrix C into M_x, M_y in the right place.

In Section 4.2 we will prove formally that this algorithm can indeed be used to compute the multiplication matrices M_x, M_y for any four element subset $\mathcal{B} \subset \mathcal{W}$ such that $D_{\mathcal{B}}(f_1, f_2) \neq 0$. If $B = \mathrm{span}_{\mathbb{C}}(\mathcal{B})$ is connected to 1, this gives a reduced B-border basis for I and the correctness of the algorithm follows from the theory of border bases. To show that this approach is indeed more general, we have computed $D_{\mathcal{B}}$ for all 15 four element subsets of $\mathcal{W}$ using Macaulay2. Each of these 15 polynomials in the ring A turns out to be nonzero, which means that for generic members of $\mathcal{F}_R(2, 2)$, any of these 15 possible choices of $\mathcal{B}$ works. Out of the 15 possible choices, only 5 satisfy the connected to 1 property. These configurations are shown in Figure 4.1. Among these five connected to 1 bases, there are only three order ideals. These are the three leftmost bases depicted in Figure 4.1. Note that the basis used in Example 3.1.2 is not in the picture. The computations in that example can be checked using the method described here.

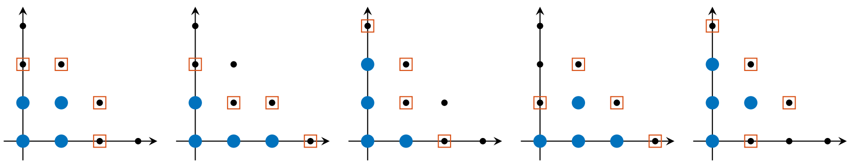


Figure 4.1: All possible subsets $\mathcal{B}$ (blue dots) of monomials of degree at most two for which B is connected to one. The border $\partial\mathcal{B}$ is indicated with small orange boxes.

Remark 4.1.1. In the case of $\mathcal{F}_R(2,2)$ we have shown that imposing the connected to 1 condition on the basis $\mathcal{B}$ reduces the number of possible choices of monomial bases of degree at most 2 from 15 to 5. To see how this scales with the degree of the equations, we have performed an analogous computation for the families $\mathcal{F}_R(2,3)$ and $\mathcal{F}_R(3,3)$. For $(f_1, f_2) \in \mathcal{F}_R(2,3)$, we consider the map $\mathrm{res} : R_{\leq 2} \times R_{\leq 1} \to R_{\leq 4}$ given by $\mathrm{res}(q_1, q_2) = q_1 f_1 + q_2 f_2$ and we compute the determinants $D_{\mathcal{B}}$ for all six element subsets $\mathcal{B}$ of the 10 monomials of degree at most 3. There are 210 such subsets, out of which 3 give a determinant $D_{\mathcal{B}} = 0$. These three 'bad' subsets[1] are

$$\{1, x, y, x^2, xy, y^2\}, \ x \cdot \{1, x, y, x^2, xy, y^2\}, \ y \cdot \{1, x, y, x^2, xy, y^2\}.$$

Among the other 207 subsets $\mathcal{B}$, which can be used as a basis $\mathcal{B} + I$ for R/I for generic members of $\mathcal{F}_R(2,3)$, there are only 19 subsets for which B is connected to 1, and only 6 of those are order ideals. For $\mathcal{F}_R(3,3)$, we consider the map $\mathrm{res} : R_{\leq 2} \times R_{\leq 2} \to R_{\leq 5}$ given by $\mathrm{res}(q_1, q_2) = q_1 f_1 + q_2 f_2$ and we compute the determinants $D_{\mathcal{B}}$ for all 5005 nine element subsets $\mathcal{B}$ of the 15 monomials of degree at most 4. Out of all these monomial bases, 4975 work for generic systems, of which 129 correspond to connected to 1 subspaces and 12 are order ideals. △

Now that we have established that there are, in general, 15 possible choices for $\mathcal{B}$, the question is *which one to pick*? The following numerical example makes it clear that, when computing in finite precision arithmetic, some choices may be significantly better than others.

Example 4.1.1. Consider the equations

$$f_1 = x + \frac{1}{3}y^2 - x^2, \quad f_2 = \frac{-1}{3}x + \frac{1}{3}x^2 + y^2,$$

for which $\langle f_1, f_2 \rangle$ equals the ideal I in Examples 3.1.6 and 3.3.4. This represents a member of $\mathcal{F}_R(2,2)$ which is non-generic in several ways. For instance, the roots have multiplicity greater than one. It is also non-generic in the sense that 13 out of 15 determinants $D_{\mathcal{B}}(f_1, f_2)$ vanish for this system. To make sure that we are dealing with generic equations, we perturb f_1 and f_2 slightly to obtain

$$f_1' = f_1 + e_1, \quad f_2' = f_2 + e_2,$$

[1] To see why these subsets cannot give bases for R/I, one can check that the vanishing of $f_1 \in R_{\leq 2}$ at all the points in $V_{\mathbb{C}^n}(I)$ implies that there cannot exist Lagrange polynomials supported in these monomials.

where $(e_1, e_2) \in \mathcal{F}_R(2,2)$ have random real coefficients which are all drawn from a normal distribution with mean 0 and standard deviation 10^{-7}. All determinants $D_{\mathcal{B}}(f_1', f_2')$ are nonzero. For all 15 choices of $\mathcal{B}$, we use Julia to compute the condition number $\kappa_{\mathcal{B}}$ of the matrix M_{11} from the algorithm explained above in double precision arithmetic. The result is

$$\begin{array}{lll}
\kappa_{\{1,x,y,xy\}} = 2.6 \cdot 10^0, & \kappa_{\{1,x,y,x^2\}} = 2.9 \cdot 10^8, & \kappa_{\{1,x,y,y^2\}} = 1.3 \cdot 10^7, \\
\kappa_{\{1,x,x^2,xy\}} = 1.8 \cdot 10^8, & \kappa_{\{1,y,xy,y^2\}} = 4.4 \cdot 10^8, & \kappa_{\{x,x^2,xy,y^2\}} = 1.7 \cdot 10^7, \\
\kappa_{\{x,y,x^2,xy\}} = 1.6 \cdot 10^7, & \kappa_{\{x,y,x^2,y^2\}} = 2.3 \cdot 10^7, & \kappa_{\{y,x^2,xy,y^2\}} = 8.4 \cdot 10^7, \\
\kappa_{\{x,y,xy,y^2\}} = 1.4 \cdot 10^8, & \kappa_{\{1,x^2,xy,y^2\}} = 1.1 \cdot 10^7, & \kappa_{\{1,y,x^2,y^2\}} = 1.1 \cdot 10^7, \\
\kappa_{\{1,x,x^2,y^2\}} = 9.2 \cdot 10^8, & \kappa_{\{1,y,x^2,xy\}} = 1.0 \cdot 10^0, & \kappa_{\{1,x,xy,y^2\}} = 1.7 \cdot 10^7.
\end{array}$$

Notice that for all choices of $\mathcal{B}$ except $\{1,x,y,xy\}$ and $\{1,y,x^2,xy\}$, the condition number is of order at least 10^7. This means that in the computation of C via $M_{01}M_{11}^{-1}$ we can expect to lose about 7 digits of accuracy (see Section B.1). Using $\mathcal{B} = \{1,x,y,xy\}$ or $\mathcal{B} = \{1,y,x^2,xy\}$ the multiplication matrices would be computed accurately up to machine precision. Note that this mirrors our conclusion in Example 3.3.4 that it is much better to stick with the basis $\mathcal{B} = \{1,x,y,xy\}$ instead of switching to $\mathcal{B} = \{1,x,y,x^2\}$ after perturbing the coefficients of f_1 and f_2 slightly. In fact, $\mathcal{B} = \{1,x,y,xy\}$ is the only basis for which B is connected to one and M_{11} is well-conditioned. Dropping the connected to 1 requirement, we see that there is another option $\mathcal{B} = \{1,y,x^2,xy\}$, for which the condition number of M_{11} is nearly perfect. $\triangle$

Example 4.1.1 shows that the choice of the right monomial basis $\mathcal{B}$ might be crucial for the accuracy with which we can compute the multiplication matrices. Let $\mathrm{res}_{\mathcal{W}}$ be the submatrix of res with rows indexed by the monomials in $\mathcal{W}$. We can formulate the problem of 'finding a good $\mathcal{B}$' as finding a submatrix of $\mathrm{res}_{\mathcal{W}}$ that is well-conditioned. This is a problem that can be solved by a standard algorithm in numerical linear algebra, called the *QR decomposition with optimal column pivoting* (see Section B.3).

We continue the discussion under the assumption that we chose the basis $\mathcal{B} = \{1,x,y,xy\}$ when we write down the matrices that are involved explicitly. Computing the matrix $C = M_{01}M_{11}^{-1}$ leads directly to a cokernel map $\mathcal{N}_{R_{\leq 3}} : R_{\leq 3} \to B$ given by

$$\mathcal{N}_{R_{\leq 3}} = \begin{array}{c} \\ 1 \\ x \\ y \\ xy \end{array} \begin{array}{c} \begin{array}{cccccccccc} 1 & x & y & xy & x^2 & y^2 & x^3 & x^2y & xy^2 & y^3 \end{array} \\ \begin{bmatrix} 1 & & & & -c_{1,1} & -c_{1,2} & -c_{1,3} & -c_{1,4} & -c_{1,5} & -c_{1,6} \\ & 1 & & & -c_{2,1} & -c_{2,2} & -c_{2,3} & -c_{2,4} & -c_{2,5} & -c_{2,6} \\ & & 1 & & -c_{3,1} & -c_{3,2} & -c_{3,3} & -c_{3,4} & -c_{3,5} & -c_{3,6} \\ & & & 1 & -c_{4,1} & -c_{4,2} & -c_{4,3} & -c_{4,4} & -c_{4,5} & -c_{4,6} \end{bmatrix} \end{array}.$$

To see that this is indeed the cokernel of res, recall that $C = M_{01}M_{11}^{-1}$ and

$$\begin{bmatrix} \mathrm{id} & -C \end{bmatrix} \mathrm{res} = \begin{bmatrix} \mathrm{id} & -M_{01}M_{11}^{-1} \end{bmatrix} \begin{bmatrix} M_{01} \\ M_{11} \end{bmatrix} = 0.$$

We have that $\ker \mathcal{N}_{R_{\leq 3}} = \mathrm{im}\,\mathrm{res} \subset I \cap R_{\leq 3}$ and $(\mathcal{N}_{R_{\leq 3}})_{|B} = \mathrm{id}_B$. As we will see, under the assumptions that I defines 4 points in $\mathbb{C}^2$ this implies that in fact we have the equality $\ker \mathcal{N}_{R_{\leq 3}} = I \cap R_{\leq 3}$. Therefore, $\mathcal{N}_{R_{\leq 3}}$ *rewrites* elements of $R_{\leq 3}$ modulo I as elements of B. A nice consequence is that the multiplication operators M_x and M_y in the basis $\mathcal{B} + I$ can be read off directly from $\mathcal{N}_{R_{\leq 3}}$: We define $\mathcal{N}_x : B \to B$ by $\mathcal{N}_x(b) = \mathcal{N}_{R_{\leq 3}}(xb)$ and $\mathcal{N}_y : B \to B$ by $\mathcal{N}_y(b) = \mathcal{N}_{R_{\leq 3}}(yb)$. This gives

$$\mathcal{N}_x = \begin{array}{c} \\ 1 \\ x \\ y \\ xy \end{array} \begin{array}{c} \begin{array}{cccc} 1 & x & y & xy \end{array} \\ \begin{bmatrix} 0 & -c_{1,1} & 0 & -c_{1,4} \\ 1 & -c_{2,1} & 0 & -c_{2,4} \\ 0 & -c_{3,1} & 0 & -c_{3,4} \\ 0 & -c_{4,1} & 1 & -c_{4,4} \end{bmatrix} \end{array}, \quad \mathcal{N}_y = \begin{array}{c} \\ 1 \\ x \\ y \\ xy \end{array} \begin{array}{c} \begin{array}{cccc} 1 & x & y & xy \end{array} \\ \begin{bmatrix} 0 & 0 & -c_{1,2} & -c_{1,5} \\ 0 & 0 & -c_{2,2} & -c_{2,5} \\ 1 & 0 & -c_{3,2} & -c_{3,5} \\ 0 & 1 & -c_{4,2} & -c_{4,5} \end{bmatrix} \end{array},$$

which are exactly the matrices of (4.1.2). This suggests a different (but equivalent) way of obtaining the multiplication matrices. First, compute a cokernel matrix $N : R_{\leq 3} \to \mathbb{C}^4$ of res (e.g. using the singular value decomposition, see Section B.2). The columns of N are indexed by the monomials in $\mathcal{V}$. Next, select a submatrix $N_{\mathcal{B}}$ of N indexed by a 4 element subset $\mathcal{B} \subset \mathcal{W}$ such that $N_{\mathcal{B}}$ is invertible ($N_{\mathcal{B}}$ is the restriction of the map N to the subspace $B = \operatorname{span}_{\mathbb{C}}(\mathcal{B}) \subset R_{\leq 2}$). If necessary, permute the columns of N such that the first 4 columns correspond to $N_{\mathcal{B}}$ and set

$$\mathcal{N}_{R_{\leq 3}} = N_{\mathcal{B}}^{-1} N : R_{\leq 3} \to B.$$

It is clear that after this procedure, $\ker \mathcal{N}_{R_{\leq 3}} = \operatorname{im} \operatorname{res}$ and $(\mathcal{N}_{R_{\leq 3}})_{|B} = \operatorname{id}_B$. The multiplication matrices can now be obtained as the matrices of $\mathcal{N}_x$ and $\mathcal{N}_y$, as defined above.

Just like in the first, equivalent approach, a choice of basis $\mathcal{B}$ has to be made. Again, this comes down to finding an invertible submatrix and for numerical stability reasons one should pick a well-conditioned submatrix using, for instance, QR with optimal pivoting.

4.2 A general framework for normal form methods

In this section we introduce *truncated normal forms* (TNFs) as defined in [TMVB18]. We consider a zero-dimensional ideal $I \subset R = \mathbb{C}[x_1, \ldots, x_n]$ such that $V(I) = V_{\mathbb{C}^n}(I) = \{z_1, \ldots, z_\delta\}$ consists of $\delta < \infty$ points and z_i has multiplicity μ_i. We have seen in Section 3.1 that this implies $\dim_{\mathbb{C}} R/I = \delta^+ = \mu_1 + \cdots + \mu_\delta$. In the same section, we also concluded that (numerical approximations of) the coordinates of the points in $V(I)$ can be computed via eigenvalue computations, once we know matrix representations of the multiplication operators

$$M_g : R/I \to R/I \quad \text{defined by} \quad M_g(f + I) = fg + I.$$

If $\mathcal{B} \subset R$ is a subset of δ^+ elements such that $\mathcal{B} + I = \{b + I \mid b \in \mathcal{B}\}$ is a basis for R/I, then the columns of a matrix representation of M_g in the basis $\mathcal{B} + I$ can be computed by *rewriting* $\{gb \mid b \in \mathcal{B}\}$ as a linear combination of the elements in $\mathcal{B}$

modulo the ideal I. A map $R \to B = \operatorname{span}_{\mathbb{C}}(\mathcal{B})$ with the right 'rewriting properties' is called a *normal form*.

Definition 4.2.1 (Normal form)**.** A *normal form* with respect to I is a $\mathbb{C}$-linear map $\mathcal{N} : R \to B$ where $B \subset R$ is a $\mathbb{C}$-vector subspace of dimension δ^+ such that

$$0 \longrightarrow I \longrightarrow R \xrightarrow{\mathcal{N}} B \longrightarrow 0 \tag{4.2.1}$$

is a short exact sequence of $\mathbb{C}$-vector spaces and $\mathcal{N}_{|B} = \mathrm{id}_B$.

Definition 4.2.1 imposes the natural condition of linearity over $\mathbb{C}$ on a normal form $\mathcal{N}$. It follows that, as vector spaces over $\mathbb{C}$, $B \simeq R/I$ (Theorem A.2.2). However, since $\mathcal{N}$ is a $\mathbb{C}$-linear map whose kernel is an ideal, it also identifies B with R/I as R-modules.

Lemma 4.2.1. *For a normal form $\mathcal{N} : R \to B$ with respect to I, define*

$$R \times B \to B \quad \textit{with} \quad (f, b) \mapsto f \cdot b = \mathcal{N}(fb). \tag{4.2.2}$$

Then (4.2.1) *is a short exact sequence of R-modules.*

Proof. We show that (4.2.2) satisfies the axioms of scalar multiplication (see Definition A.2.1). For all $f, g \in R$ and $b, b' \in B$ we have

1. $f \cdot (b + b') = \mathcal{N}(f(b + b')) = \mathcal{N}(fb) + \mathcal{N}(fb') = f \cdot b + f \cdot b'$,
2. $(f + g) \cdot b = \mathcal{N}((f + g)b) = \mathcal{N}(fb) + \mathcal{N}(gb) = f \cdot b + g \cdot b$,
3. $(fg) \cdot b = \mathcal{N}(fgb) = \mathcal{N}(f\mathcal{N}(gb) + f(gb - \mathcal{N}(gb)))$, and since $\mathcal{N} \circ \mathcal{N} = \mathcal{N}$ by $\mathcal{N}_{|B} = \mathrm{id}_B$, we have that $gb - \mathcal{N}(gb) \in \ker \mathcal{N} = I$, so that $f(gb - \mathcal{N}(gb)) \in I$ and $(fg) \cdot b = \mathcal{N}(f\mathcal{N}(gb)) = f \cdot (g \cdot b)$,
4. $1 \cdot b = \mathcal{N}(b) = b$.

The map $\mathcal{N}$ is also R-linear, since $\mathcal{N}(fg) = f \cdot \mathcal{N}(g)$ (the argument is similar to the one used in point 3 above). ☐

The property $\mathcal{N} \circ \mathcal{N} = \mathcal{N}$ used in the proof of Lemma 4.2.1 is a projection property, which is why normal forms are also called *ideal projectors*, see e.g. [DB04]. Notice that we have encountered normal forms before: the map $\mathcal{N}_{\mathcal{G}}$ of 'taking remainder upon division by a Gröbner basis $\mathcal{G}$' and the map $\mathcal{N}_{\mathcal{H}}$ of 'B-reduction along the subspace $L = \operatorname{span}_{\mathbb{C}}(\mathcal{H})$ for a B-border basis $\mathcal{H}$' both meet Definition 4.2.1. A direct consequence of Lemma 4.2.1 is that for a normal form $\mathcal{N} : R \to B$, 'multiplication with g' can be represented as the map $B \to B$ with $b \mapsto \mathcal{N}(gb)$.

As we remarked in Subsection 3.1.1, in order to compute the coordinates of the points in $V(I)$ it is sufficient to have a matrix representation for the maps $M_{x_i}, i = 1, \ldots, n$ representing multiplication with the coordinate functions. These maps are represented by $b \mapsto \mathcal{N}(x_i b)$, $b \in B$. It is therefore sufficient to compute the *restriction* of a normal form $\mathcal{N}$ to the finite-dimensional subspace

$$B^+ = B + x_1 \cdot B + \cdots + x_n \cdot B \subset R.$$

In practice, it will sometimes only be possible to compute $\mathcal{N}_{|B^+}$ from $\mathcal{N}_{|V}$ for some finite dimensional subspace $V \subset R$ containing B^+. This redundancy may force us to compute with larger matrices, but we can still extract the information we need. We therefore make the following definition.

Definition 4.2.2 (Truncated normal form (TNF))**.** Let B, V be finite dimensional $\mathbb{C}$-vector subspaces of R such that $B^+ \subset V$. A *truncated normal form* (TNF) on V with respect to I is a $\mathbb{C}$-linear map $\mathcal{N}_V : V \to B$ such that there is a normal form $\mathcal{N} : R \to B$ with respect to I such that $\mathcal{N}_{|V} = \mathcal{N}_V$.

Some obvious properties of a TNF $\mathcal{N}_V : V \to B$ with respect to I are

Property 1. The sequence $0 \longrightarrow I \cap V \longrightarrow V \xrightarrow{\mathcal{N}_V} B \longrightarrow 0$ is exact,

Property 2. $(\mathcal{N}_V)_{|B} = \mathrm{id}_B$,

Property 3. $\dim_{\mathbb{C}} B = \delta^+$.

It is not so straightforward that the converse statement is also true: TNFs are characterized by these properties.

Theorem 4.2.1. *Let B, V be finite dimensional $\mathbb{C}$-vector subspaces of R such that $B^+ \subset V$ and let $\mathcal{N}_V : V \to B$ be a $\mathbb{C}$-linear map. If $\mathcal{N}_V, V, B$ satisfy Properties 1-3 above, then $\mathcal{N}_V : V \to B$ is a TNF with respect to I.*

Before stating the proof of Theorem 4.2.1, it will be helpful to prove a lemma about the following construction. Consider a map $\mathcal{N}_V : V \to B$ with $B^+ \subset V$. For $u \in B$, we define a linear map $\mathcal{N}_u : R \to B$ by defining it on monomials first and extending it linearly. For a monomial $x^a \in R$ that can be written as $x_{i_1} \cdots x_{i_s}$ with $1 \le i_1 \le \cdots \le i_s \le n$ we set

$$\mathcal{N}_u(x_{i_1} \cdots x_{i_s}) = \mathcal{N}_V(x_{i_1}\mathcal{N}_V(x_{i_2}\mathcal{N}_V(\cdots \mathcal{N}_V(x_{i_s}u)\cdots))), \quad \mathcal{N}_u(1) = u. \tag{4.2.3}$$

Under the assumption that $(\mathcal{N}_V)_{|B} = \mathrm{id}_B$, the resulting $\mathbb{C}$-linear map $\mathcal{N}_u : R \to B$ has the following property.

Lemma 4.2.2. *Let $B, V \subset R$ be finite dimensional $\mathbb{C}$-vector subspaces of R such that $B^+ \subset V$ and let $\mathcal{N}_V : V \to B$ be a $\mathbb{C}$-linear map satisfying $(\mathcal{N}_V)_{|B} = \mathrm{id}_B$. For any $u \in B$, the $\mathbb{C}$-linear map $\mathcal{N}_u : R \to B$ obtained by extending (4.2.3) linearly is such that for any $f \in R$, $\mathcal{N}_u(f) + \langle \ker \mathcal{N}_V \rangle = fu + \langle \ker \mathcal{N}_V \rangle$ in $R/\langle \ker \mathcal{N}_V \rangle$.*

Proof. It suffices to show the lemma for monomials, so we can assume $f = x_{i_1} \cdots x_{i_s}$ with $1 \le i_1 \le \cdots \le i_s \le n$. For $s = 0$, the lemma holds trivially since $\mathcal{N}_u(1) = u$.

Since $(\mathcal{N}_V)_{|B} = \mathrm{id}_B$, we have that $\mathcal{N}_V(f) - f \in \ker \mathcal{N}_V$ for all $f \in V$. Hence, for $s = 1$ we have $\mathcal{N}_u(x_i) = \mathcal{N}_V(x_i u) = x_i u + h$ for some $h \in \ker \mathcal{N}_V$. For $s > 1$, the proof is by induction on s. Suppose the lemma holds for all monomials of degree $s-1$, then

$$\begin{aligned}\mathcal{N}_u(x_{i_1} \cdots x_{i_s}) &= \mathcal{N}_V(x_{i_1} \mathcal{N}_V(x_{i_2} \cdots x_{i_s})),\\ &= x_{i_1} \mathcal{N}_V(x_{i_2} \cdots x_{i_s}) + h \quad \text{ for some } h \in \ker \mathcal{N}_V.\end{aligned}$$

Since $\mathcal{N}_V(x_{i_2} \cdots x_{i_s}) = x_{i_2} \cdots x_{i_s} + h'$ for some $h' \in \langle \ker \mathcal{N}_V \rangle$ we have

$$\mathcal{N}_u(x_{i_1} \cdots x_{i_s}) = x_{i_1} \cdots x_{i_s} + x_{i_1} h' + h,$$

which concludes the proof. $\square$

Proof of Theorem 4.2.1. Our strategy is to construct explicitly a normal form $\mathcal{N} : R \to B$ satisfying $\mathcal{N}_{|V} = \mathcal{N}_V$. First, observe that from

$$0 \longrightarrow I \cap V \longrightarrow V \xrightarrow{\mathcal{N}_V} B \longrightarrow 0 \tag{4.2.4}$$

(Property 1) we have that $B \simeq V/(I \cap V)$ as $\mathbb{C}$-vector spaces. Since $\dim_{\mathbb{C}} B = \delta^+ = \dim_{\mathbb{C}} R/I$ (Property 3), the canonical inclusion $V/(I \cap V) \to R/I$ is an isomorphism. This gives an isomorphism $\iota : B \to R/I$, so that every residue class $f + I \in R/I$ has a representative $\iota^{-1}(f + I) \in B$. We define

$$u = \iota^{-1}(1 + I) \in B$$

such that $u + I = 1 + I$. We define the map $\mathcal{N} : R \to B$ as $\mathcal{N}_u$ from Lemma 4.2.2. That is, for a monomial $x_{i_1} \cdots x_{i_s} \in R$, $1 \le i_1 \le \cdots \le i_s \le n$ we set

$$\mathcal{N}(x_{i_1} \cdots x_{i_s}) = \mathcal{N}_V(x_{i_1} \mathcal{N}_V(x_{i_2} \mathcal{N}_V(\cdots \mathcal{N}_V(x_{i_s} u) \cdots))), \tag{4.2.5}$$

and $\mathcal{N}(1) = u$. We extend this map linearly to get a $\mathbb{C}$-linear map $\mathcal{N} : R \to B$.

We now show that $\mathcal{N} : R \to B$ is a normal form with respect to I. By Lemma 4.2.2, we have that $\mathcal{N}(f) + I = f + I$. Note that here we use Property 2. Using $V = B \oplus (I \cap V)$ (which follows from (4.2.4)), we get the following three statements.

- $\underline{\ker \mathcal{N} = I}$. If $f \in \ker \mathcal{N}$, then $f + I = 0 + I$. Conversely, if $f \in I$, then $\mathcal{N}(f) \in I$, and hence $\mathcal{N}(f) \in B \cap (I \cap V) = \{0\}$.
- $\underline{\mathcal{N}_{|B} = \mathrm{id}_B}$. For any $b \in B$, $\mathcal{N}(b) - b \in B \cap (I \cap V) = \{0\}$.
- $\underline{\mathcal{N}(R) = B}$. This follows directly from $\mathcal{N}_{|B} = \mathrm{id}_B$.

This shows that $\mathcal{N} : R \to B$ is a normal form with respect to I and hence $R = I \oplus B$. It remains to show that $\mathcal{N}_{|V} = \mathcal{N}_V$. For $f \in V$, we have that $\mathcal{N}(f) - f \in I$ and $\mathcal{N}_V(f) - f \in I \cap V$. Therefore $\mathcal{N}(f) - \mathcal{N}_V(f) \in B \cap I = \{0\}$. $\square$

Example 4.2.1. A TNF associated to a border basis normal form $\mathcal{N}_{\mathcal{H}} : R \to B$ (with B connected to 1) is $(\mathcal{N}_{\mathcal{H}})_{|B^+} = \mathcal{N}_{B^+} : B^+ \to B$, given by 'projection of B^+ onto B along $L = I \cap B^+$'. The short exact sequence looks like this:

$$0 \longrightarrow L \longrightarrow B^+ \xrightarrow{\mathcal{N}_{B^+}} B \longrightarrow 0.$$

A Gröbner basis gives a TNF by extending it to a border basis as in Example 3.3.5 and applying the same construction to obtain $\mathcal{N}_{B^+_{\prec}}$. The map $\mathcal{N}_{R_{\leq 3}} : R_{\leq 3} \to B$ from Section 4.1 is TNF by Theorem 4.2.1. $\triangle$

Remark 4.2.1. In the proof of Theorem 4.2.1 we extended the linear map $\mathcal{N}_V$ satisfying Properties 1-3 to a $\mathbb{C}$-linear map $\mathcal{N} : R \to \mathbb{C}$ by defining it on monomials as in (4.2.5). The definition seems to depend on the order of the variables $x_{i_1}, \ldots, x_{i_s}$ in which the monomial is expanded. To show that the map does not depend on this ordering, note that for each $b \in B$, by $\mathcal{N}_V(f) = f + h$ for some $h \in I \cap V$ there are $h_i, h_j \in I \cap V$ such that

$$\begin{aligned} \mathcal{N}_V(x_i \mathcal{N}_V(x_j b)) - \mathcal{N}_V(x_j \mathcal{N}_V(x_i b)) &= \mathcal{N}_V(x_i x_j b + x_i h_j - x_j x_i b - x_j h_i) \\ &= \mathcal{N}_V(x_i h_j - x_j h_i) \\ &= 0, \end{aligned}$$

where the last equality follows from $h_i, h_j \in I \cap V \Rightarrow x_i h_j - x_j h_i \in I$ and $x_i h_j - x_j h_i = x_i \mathcal{N}_V(x_j b)) - x_j \mathcal{N}_V(x_i b) \in V$. This means that in the proof of Theorem 4.2.1, the assumption that $1 \leq i_1 \leq \cdots \leq i_s \leq n$ was not strictly necessary: any other expansion of a monomial $x^a \in R$ would give the same map $\mathcal{N}$. The fact that for any $b \in B$, $\mathcal{N}_V(x_i \mathcal{N}_V(x_j b)) = \mathcal{N}_V(x_j \mathcal{N}_V(x_i b))$ corresponds to the pairwise commutativity of the multiplication operators $M_{x_i} \circ M_{x_j} = M_{x_j} \circ M_{x_i}$. $\triangle$

Note that once we have picked a basis $\mathcal{V}$ for V and $\mathcal{B}$ for B, a TNF $\mathcal{N}_V : V \to B$ is just a matrix. If we have computed such a matrix, it is straightforward to compute the multiplication matrices M_{x_i} in the basis $\mathcal{B} + I$ by computing the maps $b \mapsto \mathcal{N}_V(x_i b)$. In other words, we have reduced the root finding problem to the problem of computing a TNF with respect to I. To prove that a map $\mathcal{N}_V : V \to B$ (with $B^+ \subset V$) is a TNF, Theorem 4.2.1 shows that it suffices to show that is has Properties 1-3. In what follows, we will replace property 1 by a property that may be more convenient to check in practice.

In the following theorem, for $u \in R$ and an ideal $J \subset R$ we use the notation $(J : u) = \{f \in R \mid fu \in J\}$.

Theorem 4.2.2. *Let B, V be finite dimensional $\mathbb{C}$-vector subspaces of R such that $B^+ \subset V$ and let $\mathcal{N}_V : V \to B$ be a $\mathbb{C}$-linear map. If $\mathcal{N}_V, V, B$ are such that*

1. $\ker \mathcal{N}_V \subset I \cap V$ *and there is $u \in V$ such that $u + I$ is a unit in R/I,*
2. $(\mathcal{N}_V)_{|B} = \mathrm{id}_B$,

3. $\dim_{\mathbb{C}} B = \delta^+$,

then $\mathcal{N}_V : V \to B$ *is a TNF with respect to* I. *Moreover, we have* $I = (\langle \ker \mathcal{N}_V \rangle : u)$.

Proof. If $u + I$ is a unit in R/I for some $u \in V$, then $\mathcal{N}_V(u) + I$ is also a unit in R/I since $\mathcal{N}_V(u) - u \in \ker \mathcal{N}_V \subset I \cap V$, which implies $u + I = \mathcal{N}_V(u) + I$. Hence, we can pick an element $u \in B$ such that $u + I$ is a unit in R/I. We define the $\mathbb{C}$-linear map $\mathcal{N}_u : R \to B$ by extending (4.2.3) linearly. We consider the sequence of $\mathbb{C}$-vector spaces

$$0 \longrightarrow \ker \mathcal{N}_u \longrightarrow R \xrightarrow{\mathcal{N}_u} B \longrightarrow 0, \tag{4.2.6}$$

which we now show to be exact. Exactness at $\ker \mathcal{N}_u$ and R is clear. To show that $\mathcal{N}_u$ is surjective, we consider the $\mathbb{C}$-linear map $\phi : B \to R/I$ given by $\phi(b) = b + I$. By the assumption that $\ker \mathcal{N}_V \subset I \cap V$, Lemma 4.2.2 tells us that $\mathcal{N}_u(f) + I = fu + I$. Hence, we have that $\phi(\mathcal{N}_u(f)) = fu + I$. This shows that $\phi(\mathrm{im}\, \mathcal{N}_u) = R/I$ and hence $\dim_{\mathbb{C}}(\mathrm{im}\, \mathcal{N}_u) \geq \dim_{\mathbb{C}} R/I = \dim_{\mathbb{C}} B$, which implies $\mathrm{im}\, \mathcal{N}_u = B$.

The fact that $\mathcal{N}_u(f) + I = fu + I$ also shows that $\ker \mathcal{N}_u \subset I$. Indeed, if $\mathcal{N}_u(f) = 0$, then $fu + I = 0 + I$ which implies that $f \in I$ since $u + I$ is a unit. Exactness of the sequence (4.2.6) implies that $\dim_{\mathbb{C}} R/\ker \mathcal{N}_u = \dim_{\mathbb{C}} B = \dim_{\mathbb{C}} R/I$, which together with $\ker \mathcal{N}_u \subset I$ means that $I = \ker \mathcal{N}_u$.

We now define $\mathcal{N} : R \to B$ by $\mathcal{N}(f) = \mathcal{N}_u(fu^{-1})$ for any $u^{-1} \in R$ such that $u^{-1}u + I = 1 + I$. To show that $\mathcal{N}$ is a normal form with respect to I whose restriction to V is $\mathcal{N}_V$, we prove the following two things.

- $\underline{\ker \mathcal{N} = I}$. This follows from the fact that $\mathcal{N}_u(fu^{-1}) = 0$ is equivalent to $fu^{-1} \in I$, which is in turn equivalent to $fu^{-1}u + I = f + I = 0 + I$ in R/I.
- $\underline{\mathcal{N}_{|V} = \mathcal{N}_V}$. For $f \in V$ we have $\mathcal{N}_V(f) = f + h$ for some $h \in \ker \mathcal{N}_V \subset I \cap V$ and $\mathcal{N}(f) = \mathcal{N}_u(fu^{-1}) = fu^{-1}u + h'$ for some $h' \in \langle \ker \mathcal{N}_V \rangle \subset I$ (see Lemma 4.2.2). Therefore $\mathcal{N}(f) - \mathcal{N}_V(f) \in B \cap I = \{0\}$ by (4.2.6). In particular, this implies that $\mathcal{N}_{|B} = (\mathcal{N}_V)_{|B} = \mathrm{id}_B$.

This shows that $\mathcal{N}_V : V \to B$ is a TNF. It remains to show that $I = (\langle \ker \mathcal{N}_V \rangle : u)$. The inclusion $\ker \mathcal{N}_V \subset I \cap V$ implies $\langle \ker \mathcal{N}_V \rangle \subset I$ and thus $(\langle \ker \mathcal{N}_V \rangle : u) \subset (I : u) = I$ ($fu \in I$ implies $f \in I$ since $u + I$ is a unit in R/I). The opposite inclusion follows from the fact that if $f \in I$ then $\mathcal{N}_u(f) = 0$, and thus $0 = fu + h$ for some $h \in \langle \ker \mathcal{N}_V \rangle$ by Lemma 4.2.2. We conclude that $f \in (\langle \ker \mathcal{N}_V \rangle : u)$. $\square$

The following corollary of Theorem 4.2.2 will be important for the numerical stability of algorithms based on TNFs.

Corollary 4.2.1. *Let* V *be a finite dimensional* $\mathbb{C}$*-vector subspace of* R *and let* $W \subset V$ *be its largest subspace such that* $W^+ \subset V$ *(see Remark 4.2.2). If the space* V *and a* $\mathbb{C}$*-linear map* $N : V \to \mathbb{C}^{\delta^+}$ *satisfy the following properties:*

1. $\ker N \subset I \cap V$ *and there is* $u \in V$ *such that* $u + I$ *is a unit in* R/I,

2. $N_{|W} : W \to \mathbb{C}^{\delta^+}$ *is surjective,*

then for any δ^+*-dimensional subspace* $B \subset W$ *such that* $N_{|B}$ *is invertible,* $\mathcal{N}_V = (N_{|B})^{-1} \circ N : V \to B$ *is a TNF with respect to* I.

Proof. Note that surjectivity of $N_{|W}$ ensures that there exists some $B \subset W$ of dimension δ^+ such that $N_{|B}$ is invertible. It suffices to check that $\mathcal{N}_V = (N_{|B})^{-1} \circ N, B, V$ satisfy the assumptions of Theorem 4.2.2, which follows trivially from $\ker \mathcal{N}_V = \ker N$. $\square$

Remark 4.2.2 (Existence of W)**.** The vector space $W \subset V$ in Corollary 4.2.1 is

$$W = \{f \in V \mid x_i f \in V, i = 1, \dots, n\}.$$

To see this, note that W is closed under addition and scalar multiplication. Moreover, for each subspace T satisfying $W \subsetneq T \subset V$ we can find an element $t \in T \setminus W$ for which $x_i t \notin V$ for some i, which implies $T^+ \not\subset V$. We conclude that W is indeed the largest subspace of V such that $W^+ \subset V$. A different way of thinking about W that does not require taking elements was pointed out to the author by David Cox. Define W to be the sum of all subspaces $T \subset V$ such that $T^+ \subset V$ (this is a nonempty collection, containing $\{0\}$). By $(U+T)^+ = U^+ + T^+$, we see that $W^+ \subset V$, and W^+ is clearly the maximal such subspace. $\triangle$

The word *any* in Corollary 4.2.1 is very important: the space B is not required to come from a monomial order, to be spanned by an order ideal or to be connected to 1. The map $N : V \to \mathbb{C}^{\delta^+}$ from Corollary 4.2.1 can be thought of as a 'TNF in disguise': all we need to do to turn it into a TNF is to compose it with $N_{|B}^{-1}$ for any δ^+-dimensional subspace $B \subset R$ such that $N_{|B}$ is invertible. The terminology used in [TMVB18] is that N *covers* a TNF.

Definition 4.2.3. For a finite dimensional $\mathbb{C}$-vector subspace V of R, a map $N : V \to \mathbb{C}^{\delta^+}$ is said to *cover* a TNF $\mathcal{N}_V : V \to B$ with respect to I if there is an isomorphism $P : B \to \mathbb{C}^{\delta^+}$ such that $\mathcal{N}_V = P^{-1} \circ N$.

Proposition 4.2.1. *Let* V *be a finite dimensional* $\mathbb{C}$*-vector subspace of* R *and let* W *be as in Corollary 4.2.1. A map* $N : V \to \mathbb{C}^{\delta^+}$ *covers a TNF* $\mathcal{N}_V : V \to B$ *with respect to* I *for any* $B \subset W$ *such that* $N_{|B}$ *is invertible if and only if it satisfies the assumptions of Corollary 4.2.1.*

Proof. The 'if' direction is Corollary 4.2.1. For the 'only if' direction, suppose $N : V \to \mathbb{C}^{\delta^+}$ covers a TNF $\mathcal{N}_V : V \to B$ with respect to I, for any $B \subset W$ such that $N_{|B}$ is invertible. Then $N = P \circ \mathcal{N}_V$ for some isomorphism $P : B \to \mathbb{C}^{\delta^+}$. Since $N_{|B} = P$ and $B \subset W$, $N_{|W} : W \to \mathbb{C}^{\delta^+}$ is surjective. It follows from the definition of a TNF that $\ker N = I \cap V$ and for the normal form $\mathcal{N}$ such that $\mathcal{N}_{|V} = \mathcal{N}_V$ we have $\mathcal{N}(1) + I = 1 + I$, so $u = \mathcal{N}(1) \in B$ is such that $u + I$ is a unit in R/I. $\square$

A natural next question to ask is '*given a set of generators of I, how do we compute a map $N : V \to \mathbb{C}^{\delta^+}$ that covers a TNF with respect to I?* As we have seen, Gröbner and border bases are one way to go, and in Section 4.1 we hinted that Macaulay resultant matrices also lead to an example (at least in the case $n = s$). However, these techniques do not fully exploit the freedom for choosing B (Corollary 4.2.1). The goal of the next section is to present an algorithm that *does* exploit this, for generic members of $\mathcal{F}_R(d_1, \ldots, d_n)$.

4.3 Solving generic, dense systems

Although our goal in this section is to find solutions of a square polynomial system in affine space, some of the arguments need the homogeneous ideal obtained from homogenizing the affine equations. To avoid ambiguities we adopt our usual notation in this setting. Throughout this section, let $R = \mathbb{C}[y_1, \ldots, y_n]$ and let $(\hat{f}_1, \ldots, \hat{f}_n)$ be a generic member of $\mathcal{F}_R(d_1, \ldots, d_n)$ for $(d_1, \ldots, d_n) \in \mathbb{N}^n_{>0}$ in the sense that $V_{\mathbb{C}^n}(\hat{f}_1, \ldots, \hat{f}_n)$ consists of $\delta^+ = d_1 \cdots d_n$ points, counting multiplicities. Let $S = \mathbb{C}[x_0, \ldots, x_n]$ and let $f_i = \eta_{d_i}(\hat{f}_i)$ be the homogeneous polynomials obtained by homogenizing the $\hat{f}_i$. We denote $I = \langle f_1, \ldots, f_n \rangle \subset S$ and $I_0 = \mathscr{I}(U_0) = \langle \hat{f}_1, \ldots, \hat{f}_n \rangle \subset R$. We denote $(I_0)_{\leq d} = I_0 \cap R_{\leq d}$ for any $d \in \mathbb{N}$. This section is organized as follows. In Subsection 4.3.1 we discuss resultant maps and their close relation to TNFs. In Subsection 4.3.2 we present an algorithm for solving $\hat{f}_1 = \cdots = \hat{f}_n = 0$ under the assumptions that there are no solutions 'at infinity'. Finally, in Subsection 4.3.3 we show some numerical experiments.

4.3.1 Resultant maps

An effective way of computing a TNF starting from a set of generators of the ideal $I_0 \subset R$ is by using *resultant maps*.

Definition 4.3.1 (Resultant map)**.** For a tuple $(\hat{f}_1, \ldots, \hat{f}_s) \in R^s$ and finite dimensional $\mathbb{C}$-vector subspaces $V_1, \ldots, V_s, V \subset R$ such that $\hat{f}_i \cdot V_i \subset V, i = 1, \ldots, s$, the *resultant map* is the $\mathbb{C}$-linear map

$$\mathrm{res}_{\hat{f}_1, \ldots, \hat{f}_s} : V_1 \times \cdots \times V_s \to V \quad \text{given by} \quad \mathrm{res}_{\hat{f}_1, \ldots, \hat{f}_s}(\hat{q}_1, \ldots, \hat{q}_s) = \hat{q}_1 \hat{f}_1 + \cdots + \hat{q}_s \hat{f}_s.$$

We have encountered a resultant map before in Section 4.1. We will also consider resultant maps associated to elements of a graded ring S, which have a 'compatibility' property with respect to the grading.

Definition 4.3.2 (Graded resultant map)**.** Fix $d \in \mathbb{N}_{>0}$. For a tuple $(f_1, \ldots, f_s) \in S_{d_1} \times \cdots \times S_{d_s}$ and finite dimensional $\mathbb{C}$-vector subspaces $\Lambda_i \subset S_{d-d_i}, i = 1, \ldots, s$, $\Lambda = S_d$, the *graded resultant map* is the $\mathbb{C}$-linear map

$$\mathrm{res}_{f_1, \ldots, f_s} : \Lambda_1 \times \cdots \times \Lambda_s \to \Lambda \quad \text{given by} \quad \mathrm{res}_{f_1, \ldots, f_s}(q_1, \ldots, q_s) = q_1 f_1 + \cdots + q_s f_s.$$

Examples of graded resultant maps are the map ϕ_1 in the Koszul complex (3.2.2) and the map represented by $\mathrm{Mac}(f_0, \ldots, f_n)$ in Subsection 3.4.2 (the connection with Macaulay's matrix construction for computing resultants is why these maps are called *resultant maps*).

Recall that by Corollary 4.2.1, to show that a map $N : V \to \mathbb{C}^{\delta^+}$ covers a TNF with respect to I_0, it suffices to show that $\ker N \subset I_0 \cap V$, there is $u \in V$ such that $u + I_0$ is a unit in R/I_0 and $N_{|W}$ is onto $\mathbb{C}^{\delta^+}$, where $W \subset V$ is the largest subspace such that $W^+ \subset V$. A first indication that resultant maps could help us compute TNFs is the trivial observation that $\operatorname{im} \mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n} \subset I_0 \cap V$. This means that if $N : V \to V/\operatorname{im}\mathrm{res}$ is the cokernel map[2] of res, we have that $\ker N \subset I_0 \cap V$. Our task is to choose the spaces $V_1, \ldots, V_n$ and V for the resultant map

$$\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n} : V_1 \times \cdots \times V_n \to V$$

in such a way that the cokernel also satisfies the other criteria. One possible choice that works for generic members of $\mathcal{F}_R(d_1, \ldots, d_n)$ follows directly from Macaulay's construction. Let $d_0 = 1$, $\hat{\rho} = d_1 + \cdots + d_n - n + 1$ and let $\Lambda_0, \ldots, \Lambda_n, \Lambda$ be as defined in Subsection 3.4.2. Moreover, we let $V_i = \eta^{-1}_{\hat{\rho}-d_i}(\Lambda_i)$ be the image of *dehomogenization* restricted to Λ_i and $V = \eta^{-1}_{\hat{\rho}}(\Lambda)$. Note that $V_i \subset R_{\leq \hat{\rho}-d_i}$ and $V = R_{\leq \hat{\rho}}$.

Proposition 4.3.1. *Let $\hat{\rho}, V_1, \ldots, V_n, V$ be as defined above and consider the resultant map*

$$\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n} : V_1 \times \cdots \times V_n \to V.$$

If for some $f_0 \in S_1$, the submatrix M_{11} of $\mathrm{Mac}(f_0, \ldots, f_n)$ is invertible, then the corank of $\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ is δ^+ and any cokernel map $N : V \to \mathbb{C}^{\delta^+}$ of $\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ covers a TNF with respect to I_0.

Proof. Using the notation

$$\eta_{\hat{\rho}-d_1,\ldots,\hat{\rho}-d_n} = \eta_{\hat{\rho}-d_1} \times \cdots \times \eta_{\hat{\rho}-d_n} : V_1 \times \cdots \times V_n \to \Lambda_1 \times \cdots \times \Lambda_n$$

for 'component-wise' homogenization, we get the commuting diagram

$$\begin{tikzcd} V_1 \times \cdots \times V_n \arrow[r, "\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}"] \arrow[d, "\eta_{\hat{\rho}-d_1,\ldots,\hat{\rho}-d_n}"] & V \arrow[d, "\eta_{\hat{\rho}}"] \\ \Lambda_1 \times \cdots \times \Lambda_n \arrow[r, "\mathrm{res}_{f_1,\ldots,f_n}"] & \Lambda \end{tikzcd}$$

from which we see that $\operatorname{im}\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ and $\operatorname{im}\mathrm{res}_{f_1,\ldots,f_n}$ are isomorphic via $\eta_{\hat{\rho}}$. Since

$$\mathrm{res}_{f_1,\ldots,f_n} = \mathrm{Mac}(f_0, \ldots, f_n)_{|\Lambda_1 \times \cdots \times \Lambda_n} = \begin{bmatrix} M_{01} \\ M_{11} \end{bmatrix}$$

[2] The (canonical) cokernel map of a $\mathbb{C}$-linear map $\phi : V \to V'$ is the projection $\pi : V' \to V'/\operatorname{im}\phi$. We say that $\psi : V' \to V''$ is *a cokernel map* of ϕ if $\ker \psi = \operatorname{im}\phi$ and $\overline{\psi} : V'/\operatorname{im}\phi \to V''$ given by $\overline{\psi}(v' + \operatorname{im}\phi) = \psi(v')$ is an isomorphism.

and we are assuming that M_{11} is invertible, we have that $\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ has corank δ^+. Since $1 \in V$ and $1 + I_0$ is a unit in R/I_0, we only need to show that the restriction of a map $N : V \to \mathbb{C}^{\delta^+}$ such that $\ker N = \operatorname{im} \mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ to the subspace $W = R_{\leq \hat{\rho}-1}$ is onto $\mathbb{C}^{\delta^+}$. We can choose bases of $V_1, \ldots, V_n, V$ such that the resulting matrix representation of $\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ is

$$\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n} = \begin{bmatrix} M_{01} \\ M_{11} \end{bmatrix}.$$

Indeed, the rows are indexed by the 'dehomogenized versions' of the monomials in $\Sigma_0', \ldots, \Sigma_n'$ (in the notation of Subsection 3.4.2) and the columns by the dehomogenization of $\Sigma_0, \ldots, \Sigma_n$. A cokernel map of $\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ is given by

$$N = \begin{bmatrix} \mathrm{id} & -M_{01}M_{11}^{-1} \end{bmatrix}, \quad \begin{bmatrix} \mathrm{id} & -M_{01}M_{11}^{-1} \end{bmatrix} \begin{bmatrix} M_{01} \\ M_{11} \end{bmatrix} = 0.$$

From this observation it is clear that $N_{|B}$ is onto $\mathbb{C}^{\delta^+}$, where B is the $\mathbb{C}$-span of the δ^+ monomials in $\eta_{\hat{\rho}}^{-1}(\Sigma_0')$. Since $\Sigma_0' = x_0 \cdot \Sigma_0$, x_0 divides all monomials in Σ_0' and therefore all monomials in B are of degree $< \hat{\rho}$. It follows that $B \subset W$ and $N_{|W}(W) = \mathbb{C}^{\delta^+}$. $\square$

As noted in Remark 3.4.4, the image of the graded resultant map $\mathrm{res}_{f_1,\ldots,f_n} : \Lambda_1 \times \cdots \times \Lambda_n \to \Lambda$ with $\Lambda_1, \ldots, \Lambda_n, \Lambda$ coming from Macaulay's construction does not change when we replace Λ_i by $S_{\hat{\rho}-d_i}$. As the image remains unchanged, nothing happens to the cokernel map either. As a result, one may think that it is better to stick with the smaller spaces Λ_i from Macaulay's construction, since it leads to a cokernel computation of a smaller matrix. However, we observe in numerical experiments that the cokernel of the larger matrix is less sensitive to perturbations (see Appendix B).

Example 4.3.1. We consider a member $(\hat{f}_1, \hat{f}_2, \hat{f}_3) \in \mathcal{F}_R(8, 8, 8)$ whose coefficients are all real and drawn from a standard normal distribution. We construct matrices for two resultant maps

$$\mathrm{res}_{f_1,f_2,f_3} : \Lambda_1 \times \Lambda_2 \times \Lambda_3 \to \Lambda.$$

For the first map, $\Lambda = S_{22}$ and Λ_i is $\mathrm{span}_{\mathbb{C}}(\Sigma_i)$ coming from Macaulay's construction. For the second map, $\Lambda = S_{22}$ and Λ_i is the entire graded piece S_{14}. The corresponding matrices have sizes 2300×1788 and 2300×2040 respectively. These are also matrices for the resultant maps

$$\mathrm{res}_{\hat{f}_1,\hat{f}_2,\hat{f}_3} : V_1 \times V_2 \times V_3 \to V$$

where V_i is the dehomogenization of Λ_i and V is the dehomogenization of Λ. The singular values of these matrices are shown in Figure 4.2. The sensitivity of the cokernel of the matrix to perturbations can be measured by the smallest singular value that is considered 'numerically nonzero' (see Section B.2). This is the size of the minimal perturbation that enlarges the dimension of the cokernel by 1. The smaller this number, the more ill-conditioned the problem of computing the cokernel is. For

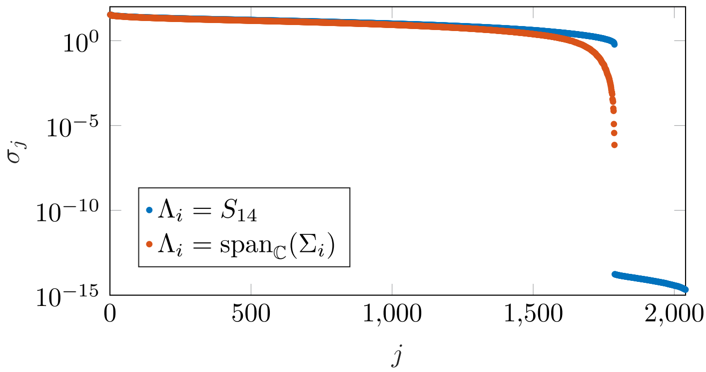


Figure 4.2: Singular values of two resultant maps with the same image.

$\Lambda_i = \mathrm{span}_{\mathbb{C}}(\Sigma_i)$ we expect the matrix to be of full rank: there are 1788 nonzero singular values. For $\Lambda_i = S_{14}$, the image of the map does not change so there are still 1788 nonzero singular values. However, now there are 252 singular values that are numerically zero: they are of the order $u \cdot \sigma_1$, where $u \approx 10^{-16}$ is the working precision and σ_1 is the largest singular value. This causes the dramatic 'jump' for the blue dots at $j = 1788$ in Figure 4.2. The ratio σ_{1788}/σ_1 is approximately $2.12 \cdot 10^{-8}$ for $\Lambda_i = \mathrm{span}_{\mathbb{C}}(\Sigma_i)$ and $1.76 \cdot 10^{-2}$ for $\Lambda_i = S_{14}$. $\triangle$

Proposition 4.3.1 implies that for a generic member $(\hat{f}_1, \ldots, \hat{f}_n)$, a TNF with respect to I_0 can be computed from the cokernel of the resultant map

$$\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n} : V_1 \times \cdots \times V_n \to V \tag{4.3.1}$$

where $V_i \subset R_{\leq \hat{\rho}-d_i}$ is the dehomogenization of $\Lambda_i = \mathrm{span}_{\mathbb{C}}(\Sigma_i) \subset S_{\hat{\rho}-d_i}$ and $V = R_{\leq \hat{\rho}} = \eta_{\hat{\rho}}^{-1}(S_{\hat{\rho}})$. By the discussion above, it is an easy corollary that for a generic member $(\hat{f}_1, \ldots, \hat{f}_n)$, a TNF with respect to I_0 can be computed from the cokernel of the resultant map (4.3.1) where each V_i is replaced by the larger space $R_{\leq \hat{\rho}-d_i}$. 'Genericity' here means that M_{11} is invertible. As we have seen in Example 3.4.5, this implies that the resultant

$$\mathrm{Res}_\infty = \mathrm{Res}_{d_1,\ldots,d_n}(f_1(0, x_1, \ldots, x_n), \ldots, f_n(0, x_1, \ldots, x_n))$$

does not vanish. However, the converse statement is not true: it might be that there are no solutions at infinity ($\mathrm{Res}_\infty \neq 0$), yet M_{11} is not invertible. The following proposition shows that $\mathrm{Res}_\infty \neq 0$ is the only condition we need for our cokernel computation to lead to a TNF.

Proposition 4.3.2. *Let $V_i = R_{\leq \hat{\rho}-d_i}, i = 1, \ldots, n$ and $V = R_{\leq \hat{\rho}}$. Consider the resultant map*

$$\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n} : V_1 \times \cdots \times V_n \to V.$$

If $\mathrm{Res}_\infty \neq 0$*, then the corank of* $\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ *is* δ^+ *and a cokernel map* $N : V \to \mathbb{C}^{\delta^+}$ *of* $\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ *covers a TNF with respect to* I_0*.*

Proof. Since $1 \in V$ and $\ker N \subset I_0 \cap V$ is immediate, we only have to show that $N_{|W}$ is onto $\mathbb{C}^{\delta^+}$, where $W = R_{\leq \hat{\rho}-1}$. As in the proof of Proposition 4.3.1, we have that $\mathrm{im}\,\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n} \simeq \mathrm{im}\,\mathrm{res}_{f_1,\ldots,f_n}$ where

$$\mathrm{res}_{f_1,\ldots,f_n} : \Lambda_1 \times \cdots \times \Lambda_n \to \Lambda$$

with $\Lambda_i = \eta_{d_i}(V_i) = S_{\hat{\rho}-d_i}$ and $\Lambda = \eta_{\hat{\rho}}(V) = S_{\hat{\rho}}$. The assumption $\mathrm{Res}_\infty \neq 0$ implies that $V_{\mathbb{P}^n}(I)$ is zero-dimensional (see Subsection 3.4.1). The statement about the corank of $\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ follows from $\mathrm{im}\,\mathrm{res}_{f_1,\ldots,f_n} = I_{\hat{\rho}}$, and by the proof of Theorem 3.2.2, $I_{\hat{\rho}}$ has codimension δ^+ in $S_{\hat{\rho}}$.

It also follows from the proof of Theorem 3.2.2 that $\mathrm{HF}_I(\rho) = \dim_{\mathbb{C}}(S/I)_\rho = \delta^+$ for $\rho = \hat{\rho} - 1$. Therefore, we can pick a set of δ^+ monomials $\mathcal{B}_\rho \subset S_\rho$ such that $\mathcal{B}_\rho + I_\rho$ is a basis for $(S/I)_\rho$. Since x_0 vanishes at none of the points in $V_{\mathbb{P}^n}(I)$, Lemma 3.2.1 tells us that, under the assumption that all multiplicities are 1 ($\delta = \delta^+$), $M_{x_0} : (S/I)_\rho \to (S/I)_{\rho+1}$ is an isomorphism of $\mathbb{C}$-vector spaces. However, this is also true for arbitrary multiplicities (Corollary 5.5.3). A consequence is that $\mathcal{B}_{\hat{\rho}} = x_0 \cdot \mathcal{B}_\rho = \{x_0 x^a \mid x^a \in \mathcal{B}_\rho\} \subset S_{\hat{\rho}}$ is such that $\mathcal{B}_{\hat{\rho}} + I_{\hat{\rho}}$ is a basis for $(S/I)_{\hat{\rho}}$. Let $\{g_1, \ldots, g_m\}$ be a basis for $I_{\hat{\rho}}$. Since $\mathrm{HF}_I(\hat{\rho}) = \delta^+$, we know that $m = \dim_{\mathbb{C}} S_{\hat{\rho}} - \delta^+$. We order the monomials $\mathcal{V}_{\hat{\rho}}$ of degree $\hat{\rho}$ such that the δ^+ monomials in $\mathcal{B}_{\hat{\rho}}$ come first and represent the inclusion $I_{\hat{\rho}} \to S_{\hat{\rho}}$ by the matrix

$$M = \begin{array}{c} \mathcal{B}_{\hat{\rho}} \\ \\ \mathcal{V}_{\hat{\rho}} \backslash \mathcal{B}_{\hat{\rho}} \end{array} \begin{bmatrix} | & & | \\ g_1 & \cdots & g_m \\ | & & | \end{bmatrix} = \begin{array}{c} \mathcal{B}_{\hat{\rho}} \\ \mathcal{V}_{\hat{\rho}} \backslash \mathcal{B}_{\hat{\rho}} \end{array} \begin{bmatrix} M_{01} \\ M_{11} \end{bmatrix}.$$

We claim that M_{11} is invertible. If not, there is a nonzero vector $v \in \mathbb{C}^m$ such that $M_{11} v = 0$. Since M is full rank ($\{g_1, \ldots, g_m\}$ is a basis), we must have $Mv \neq 0$. The vector Mv represents a polynomial in $I_{\hat{\rho}} \cap \mathrm{span}_{\mathbb{C}}(\mathcal{B}_{\hat{\rho}})$. Since $\mathcal{B}_{\hat{\rho}} + I_{\hat{\rho}}$ is a basis for $(S/I)_{\hat{\rho}}$, this leads to a contradiction. Since $\mathrm{im}\, M = I_{\hat{\rho}} = \mathrm{im}\,\mathrm{res}_{f_1,\ldots,f_n} \simeq \mathrm{im}\,\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$, we have that

$$N = \begin{array}{c} \begin{array}{cc} \eta_{\hat{\rho}}^{-1}(\mathcal{B}_{\hat{\rho}}) & \eta_{\hat{\rho}}^{-1}(\mathcal{V}_{\hat{\rho}} \backslash \mathcal{B}_{\hat{\rho}}) \end{array} \\ \begin{bmatrix} \mathrm{id} & -M_{01} M_{11}^{-1} \end{bmatrix} \end{array}$$

satisfies $NM = 0$. We conclude that $N : V \to \mathbb{C}^{\delta^+}$ represents a cokernel map of $\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ and the restriction of N to $B = \mathrm{span}_{\mathbb{C}}(\eta_{\hat{\rho}}^{-1}(\mathcal{B}_{\hat{\rho}}))$ is onto $\mathbb{C}^{\delta^+}$. By construction, x_0 divides every monomial in $\mathcal{B}_{\hat{\rho}}$, which implies $B \subset W$. □

Corollary 4.3.1. *If* $\mathrm{Res}_\infty \neq 0$*, then the image of* $\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n} : V_1 \times \cdots \times V_n \to V$ *with* V_i, V *as in Proposition 4.3.2 is* $(I_0)_{\leq \hat{\rho}}$*.*

Proof. By Proposition 4.3.2, a cokernel map $N : V \to \mathbb{C}^{\delta^+}$ covers a TNF with respect to I_0. Therefore

$$0 \longrightarrow I_0 \cap V \longrightarrow V \xrightarrow{N} \mathbb{C}^{\delta^+} \longrightarrow 0$$

is exact and $\ker N = \operatorname{im} \operatorname{res}_{\hat{f}_1,\dots,\hat{f}_n} = I_0 \cap V = (I_0)_{\leq \hat{\rho}}$. □

Remark 4.3.1. If $I = \langle f_1, \dots, f_n \rangle$ is zero-dimensional but $\mathrm{Res}_\infty = 0$ (there are *isolated* solutions at infinity), then a random affine change of coordinates $y_i \leftarrow c_{i0} + \sum_{j=1}^{n} c_{ij} y_j$ will make sure that the points at infinity move into the affine chart U_0, and Proposition 4.3.2 applies after performing this change of coordinates. △

Example 4.3.2. The resultant maps from Proposition 4.3.2 are often presented in a monomial basis for $V_1, \dots, V_n, V$. This leads to highly structured matrices with an interesting sparsity pattern. An example for $n = 3$ and $d_1 = 5, d_2 = 4, d_3 = 6$ is shown in Figure 4.3. The matrix has size 560×505. △

4.3.2 Algorithm

The following simple example illustrates the main steps in the algorithm presented in this subsection.

Example 4.3.3. We consider the polynomial system in Example 3.1.2. To be consistent with our notation of this section we replace x, y, f_1, f_2, I in that example by $y_1, y_2, \hat{f}_1, \hat{f}_2, I_0$ here. The equations become

$$\begin{aligned} \hat{f}_1 &= 7 + 3y_1 - 6y_2 - 4y_1^2 + 2y_1y_2 + 5y_2^2, \\ \hat{f}_2 &= -1 - 3y_1 + 14y_2 - 2y_1^2 + 2y_1y_2 - 3y_2^2. \end{aligned}$$

The resultant map from Proposition 4.3.2 is represented by

$$\operatorname{res}_{\hat{f}_1,\hat{f}_2}^{\top} = \begin{array}{c} \\ \hat{f}_1 \\ y_1\hat{f}_1 \\ y_2\hat{f}_1 \\ \hat{f}_2 \\ y_1\hat{f}_2 \\ y_2\hat{f}_2 \end{array} \begin{array}{c} \begin{array}{cccccccccc} 1 & y_1 & y_2 & y_1^2 & y_1y_2 & y_2^2 & y_1^3 & y_1^2y_2 & y_1y_2^2 & y_2^3 \end{array} \\ \begin{bmatrix} 7 & 3 & -6 & -4 & 2 & 5 & & & & \\ & 7 & & 3 & -6 & & -4 & 2 & 5 & \\ & & 7 & & 3 & -6 & & -4 & 2 & 5 \\ -1 & -3 & 14 & -2 & 2 & -3 & & & & \\ & -1 & & -3 & 14 & & -2 & 2 & -3 & \\ & & -1 & & -3 & 14 & & -2 & 2 & -3 \end{bmatrix} \end{array}.$$

Knowing the solutions of $\hat{f}_1 = \hat{f}_2 = 0$ (see Example 3.1.2), we can construct a cokernel matrix N whose rows represent 'evaluation at $z_i \in V_{\mathbb{C}^2}(I_0)$'. This gives

$$N = \begin{array}{c} \\ \mathrm{ev}_{(-2,3)} \\ \mathrm{ev}_{(3,2)} \\ \mathrm{ev}_{(2,1)} \\ \mathrm{ev}_{(-1,0)} \end{array} \begin{array}{c} \begin{array}{cccccccccc} 1 & y_1 & y_2 & y_1^2 & y_1y_2 & y_2^2 & y_1^3 & y_1^2y_2 & y_1y_2^2 & y_2^3 \end{array} \\ \begin{bmatrix} 1 & -2 & 3 & 4 & -6 & 9 & -8 & 12 & -18 & 27 \\ 1 & 3 & 2 & 9 & 6 & 4 & 27 & 18 & 12 & 8 \\ 1 & 2 & 1 & 4 & 2 & 1 & 8 & 4 & 2 & 1 \\ 1 & -1 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \end{bmatrix} \end{array}.$$

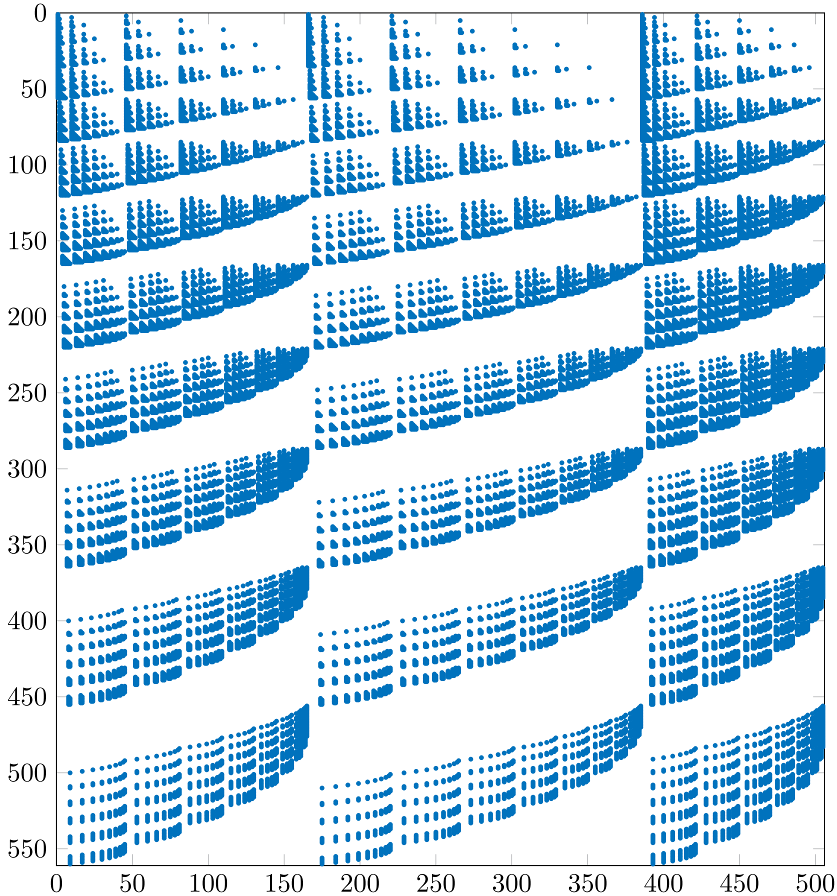


Figure 4.3: Nonzero pattern for the resultant map $\mathrm{res}_{\hat{f}_1,\hat{f}_2,\hat{f}_3} : R_{\leq 8} \times R_{\leq 9} \times R_{\leq 7} \to R_{\leq 13}$ for a generic member of $\mathcal{F}_R(5,4,6)$.

One can check that $N \,\mathrm{res}_{\hat{f}_1,\hat{f}_2} = 0$ and N has rank 4. This is of course cheating: we cannot construct a cokernel like this in practice. However, this construction will do for illustration purposes. We use the basis $\mathcal{B} = \{y_1, y_2, y_1^2, y_1 y_2\}$ ($\mathcal{B}$ in Example 3.1.2 corresponds to $\mathcal{B} + I_0$ here) and $B = \mathrm{span}_{\mathbb{C}}(\mathcal{B})$. The corresponding TNF is $\mathcal{N}_V = N_{|B}^{-1} N$. Defining $N_i : B \to B$ by $N_i(b) = N(y_i b)$ we find that

$$M_{y_i} : B \to B \quad \text{is given by} \quad M_{y_i}(b) = \mathcal{N}_V(y_i b) = (N_{|B}^{-1} N)(y_i b) = (N_{|B}^{-1} N_i)(b).$$

The maps $N_{|B}, N_1, N_2$ are the submatrices of N corresponding to $\mathcal{B}, y_1 \cdot \mathcal{B}, y_2 \cdot \mathcal{B}$. They are given by

$$N_{|B} = \begin{bmatrix} -2 & 3 & 4 & -6 \\ 3 & 2 & 9 & 6 \\ 2 & 1 & 4 & 2 \\ -1 & 0 & 1 & 0 \end{bmatrix}, \quad N_1 = \begin{bmatrix} 4 & -6 & -8 & 12 \\ 9 & 6 & 27 & 18 \\ 4 & 2 & 8 & 4 \\ 1 & 0 & -1 & 0 \end{bmatrix}, \quad N_2 = \begin{bmatrix} -6 & 9 & 12 & -18 \\ 6 & 4 & 18 & 12 \\ 2 & 1 & 4 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

One can check that $M_{y_1} = N_{|B}^{-1} N_1$ is indeed the matrix 'M_x' obtained in Example 3.1.2. △

Proposition 4.3.2 leads directly to Algorithm 4.1 for computing the multiplication operators M_{y_i}, $i = 1, \ldots, n$. There are other ways to tackle the actual implementation

Algorithm 4.1 Computes multiplication matrices for $(\hat{f}_1, \ldots, \hat{f}_n) \in \mathcal{F}_R(d_1, \ldots, d_n)$ such that $\mathrm{Res}_\infty \neq 0$

1: **procedure** MULTIPLICATIONMATRICES($\hat{f}_1, \ldots, \hat{f}_n$)
2: $\quad \hat{\rho} = d_1 + \cdots + d_n - n + 1$
3: $\quad \mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n} \leftarrow$ the resultant map $V_1 \times \cdots \times V_n \to V$ from Proposition 4.3.2
4: $\quad N \leftarrow \mathrm{coker}\, \mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$
5: $\quad N_{|W} \leftarrow$ submatrix of N corresponding to monomials of degree $< \hat{\rho}$
6: $\quad N_{|B} \leftarrow$ submatrix of $N_{|W}$ corresponding to an invertible submatrix
7: $\quad \mathcal{B} \leftarrow$ monomials corresponding to the columns of $N_{|B}$
8: $\quad$ **for** $i = 1, \ldots, n$ **do**
9: $\quad\quad N_i \leftarrow$ submatrix of N corresponding to $x_i \cdot \mathcal{B}$
10: $\quad\quad M_{y_i} \leftarrow (N_{|B})^{-1} N_i$
11: $\quad$ **end for**
12: $\quad$ **return** $M_{y_1}, \ldots, M_{y_n}$
13: **end procedure**

(see e.g. Section 4.4). We focus on the following choices in Algorithm 4.1 for now. In line 3, it is assumed that $\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ is constructed with respect to the monomial basis of $V = R_{\leq \hat{\rho}}$. The matrix has size

$$\dim_{\mathbb{C}} R_{\leq \hat{\rho}} \times \sum_{i=1}^{n} \dim_{\mathbb{C}} R_{\leq \hat{\rho} - d_i}$$

or in terms of binomial coefficients:

$$\binom{d_1 + \cdots + d_n + 1}{n} \times \sum_{i=1}^{n} \binom{d_1 + \cdots + d_{i-1} + d_{i+1} + \cdots + d_n + 1}{n}.$$

In line 4, we compute the cokernel (or left nullspace) of this matrix. This can be done, for instance, using the singular value decomposition or a rank revealing QR decomposition (see Section B.3). The result is a matrix of size $\delta^+ \times \dim_{\mathbb{C}} R_{\leq \hat{\rho}}$ or

$$d_1 \cdots d_n \times \binom{d_1 + \cdots + d_n + 1}{n},$$

whose columns are indexed by the monomials of $R_{\leq \hat{\rho}}$. In line 6 we can restrict $N_{|W}$ to *any* subspace $B \subset W$ such that $N_{|B}$ is invertible, by Corollary 4.2.1. Here we propose to select a submatrix of $N_{|W}$. It is crucial for numerical stability to select δ^+ columns that give a well-conditioned submatrix $N_{|B}$. One way to do this is to use QR with column pivoting (Section B.3). More concretely, if the column pivoted QR factorization of $N_{|W}$ is

$$N_{|W}\mathbf{P} = \mathbf{QR},$$

where $\mathbf{P}$ is a column pivoting matrix, $\mathbf{Q}$ is unitary and $\mathbf{R}$ is upper triangular, we set $N_{|B} = N_{|W}\mathbf{P}_{:,1:\delta^+}$. In lines 9 and 10, the multiplication maps M_{y_i} are computed as $b \mapsto \mathcal{N}_V(y_i b)$, where $\mathcal{N}_V = N_{|B}^{-1}N$ is a TNF covered by N.

Remark 4.3.2 (On the complexity of Algorithm 4.1)**.** Let us determine the asymptotic complexity of the different steps in Algorithm 4.1 in the simplified case where $d = d_1 = \cdots = d_n$. In line 4, we compute the cokernel of a matrix of size

$$p = \binom{nd+1}{n} = O\left(\frac{n^n}{n!}d^n\right) \text{ by } q = n\binom{(n-1)d+1}{n} = O\left(\frac{(n-1)^n}{(n-1)!}d^n\right).$$

Assuming this is done using the SVD, it requires $O(\min(p^2q, pq^2))$ flops [GVL12, §5.4.5]. For large enough n, d, we have $q > p$ and therefore, the cokernel computation has complexity $O(C_1(n)d^{3n})$ where

$$C_1(n) = \left(\frac{n^n}{n!}\right)^2 \frac{(n-1)^n}{(n-1)!}.$$

As $N_{|W}$ has size $d^n \times O(\frac{n^n}{n!}d^n)$, the column pivoted QR factorization in line 6 has complexity $O(C_2(n)d^{3n})$ where $C_2(n) = \frac{n^n}{n!}$, see [GVL12, §5.4.1]. Finally, the for loop in lines 8-11 takes $O(nd^{3n})$ floating point operations. From this rough analysis it is clear that, for large n, the dominant step in the algorithm in terms of computational complexity is the cokernel computation. We will propose some possible ways of reducing the complexity of this step in Subsection 4.4.1. Another important remark is that the increase in the complexity by performing a column pivoted QR factorization to make a 'numerically optimized' choice of basis $\mathcal{B}$ is negligible in comparison to the cost of the cokernel computation ($C_2(n) \ll C_1(n)$). Nevertheless, as we have mentioned before and as we will illustrate in Subsection 4.3.3, it is a very effective way to enhance the numerical stability of the algorithm. △

Remark 4.3.3. Let $f_0 = x_i$ and take M, N, B as in the proof of Proposition 4.3.2. We represent the monomial multiples $\{x^a f_0 = x^a x_i, x^a \in \mathcal{B}_{\hat{\rho}}\}$ in the monomial basis of $S_{\hat\rho}$ to obtain the matrix

$$\begin{array}{c} \mathcal{B}_{\hat\rho} \\ \\ \mathcal{V}_{\hat\rho}\setminus B_{\hat\rho} \end{array} \begin{bmatrix} | & & | \\ x^{a_1}x_i & \cdots & x^{a_{\delta^+}}x_i \\ | & & | \end{bmatrix} = \begin{array}{c} \mathcal{B}_{\hat\rho} \\ \mathcal{V}_{\hat\rho}\setminus B_{\hat\rho} \end{array} \begin{bmatrix} M_{00} \\ M_{10} \end{bmatrix}.$$

Then $N_{|B} = \mathrm{id}_B$, $\mathcal{N}_V = N$ and $N_{|B}^{-1}N_i = N_i$ is exactly the Schur complement $M_{00} - M_{01}M_{11}^{-1}M_{10}$. Note the strong analogy with the method described in Subsection 3.4.2. △

Once the multiplication matrices are computed, we have almost solved the system of equations $\hat f_1 = \cdots = \hat f_n = 0$: it remains to diagonalize the matrices $M_{y_1}, \ldots, M_{y_n}$. These matrices share a set of δ invariant subspaces, each associated to one of the isolated solutions in $V_{\mathbb{C}^n}(I_0)$ (see Subsection 3.1.3). In the case where each of the $\mu_i = 1$ (i.e., I_0 is radical and $\delta = \delta^+$), the matrices $M_{y_1}, \ldots, M_{y_n}$ have $\delta = \delta^+$ common eigenvectors. The M_{y_i} can be diagonalized simultaneously. We can compute the common eigenvectors by diagonalizing a generic linear combination M_h of the M_{y_i}. For $h = h_1x_1 + \cdots + h_nx_n, h_i \in \mathbb{C}$, set $M_h = h(M_{y_1}, \ldots, M_{y_n}) = \sum_{i=1}^n h_iM_{y_i}$. For generic h, all of the eigenvalues $h(z_j), j = 1, \ldots, \delta^+$ are distinct (see Lemma 3.1.1) and all invariant subspaces of M_h have dimension 1. We find $DM_hD^{-1} = \mathrm{diag}(h(z_1), \ldots, h(z_\delta))$. Applying the same transformation to the M_{y_i} gives $DM_{y_i}D^{-1} = \mathrm{diag}(z_{1i}, \ldots, z_{\delta i})$. Note that the order of the roots corresponding to the diagonal elements is the same for each i: it corresponds to the order of the evaluation functionals ev_{z_i} in the matrix D (see Subsection 3.1.1). We can then read off the coordinates of the δ roots from the diagonals of the $DM_{y_i}D^{-1}$.

An alternative is to compute the complex Schur form (see Section B.4) of M_h: $\mathbf{U}M_h\mathbf{U}^H = \mathbf{T}_h$, where $\mathbf{U}$ is a unitary matrix and $\mathbf{T}_h$ is upper triangular ($\cdot^H$ denotes the Hermitian transpose). The same transformation makes the M_{y_i} upper triangular: $\mathbf{U}M_{y_i}\mathbf{U}^H = \mathbf{T}_i$ and the solutions can be read off from the diagonals of the $\mathbf{T}_i$.

We note that a simultaneous diagonalization of a set of commuting matrices in the nondefective case is equivalent to the *tensor rank decomposition* or *canonical polyadic decomposition* (CPD) of a third order tensor [DL06]. It is possible to use tensor algorithms to refine the solutions obtained by the algorithm described above. The routine `cpd_gevd` in Tensorlab can be used for this computation [VDS+16]. In [VSDL17a], the problem of finding the coordinates of $z_1, \ldots, z_\delta$ from the cokernel map N is interpreted as a multidimensional harmonic retrieval (MHR) problem, which leads to a CPD computation closely related to the one described here. They establish the connection with multiplication matrices (in the context of the border basis approach in [Ste04]) and apply these methods in an overconstrained setting ($s > n$), where the coefficients of the polynomials may be contaminated by noise.

We should mention that in the recent work [BBV19], the authors show that computing the tensor rank decomposition via eigenvalue decompositions is in general unstable.

That is, it produces larger errors in the computed decomposition than predicted by the condition number of the tensor rank decomposition problem, as studied in [BV18b]. In practice, for generic systems, these errors are fortunately not too bad. As the authors of [BBV19] suggest, the output of the eigenvalue computation can be refined, if needed, to a satisfactory solution of the tensor decomposition problem.

In the case where some of the points in $V_{\mathbb{C}^n}(I_0)$ have multiplicity greater than 1, the invariant subspaces of a multiplication map M_h are revealed by (3.1.9) in Subsection 3.1.3. The *Jordan form* of M_h has the eigenvalues $h(z_i), i = 1, \ldots, \delta$ on its diagonal, where $h(z_i)$ occurs μ_i times. The computation of a Jordan form of a defective matrix in finite precision arithmetic is very tricky: the tiniest perturbation destroys the Jordan structure. However, the algorithm described in [Zen16], implemented in NAClab [ZL14], did show good results on some test cases.

A successful, alternative method is described in [CGT97]. We compute the Schur form of M_h: $\mathbf{U}^H M_h \mathbf{U} = \mathbf{T}_h$, with $\mathbf{U}$ orthogonal and $\mathbf{T}_h$ upper triangular. If there are solutions with multiplicity > 1, some elements on the diagonal of $\mathbf{T}_h$ appear multiple times. Next, we use a clustering of the diagonal elements of $\mathbf{T}_h$ and reorder the factorization to obtain $\mathbf{U}'$ orthogonal, $\mathbf{T}'_h$ upper triangular such that the diagonal elements are clustered and $(\mathbf{U}')^H M_h \mathbf{U}' = \mathbf{T}'_h$. The same transformation makes the M_{y_i} *block* upper triangular with δ diagonal blocks of size $\mu_i \times \mu_i, i = 1, \ldots, \delta$ corresponding to the clusters on the diagonal of $\mathbf{T}_h$. All of the diagonal blocks only have one eigenvalue, which is $h(z_i)$. For more details on this approach we refer to [CGT97]. Another approach based on the intersection of eigenspaces is given in [MT01] and [GT09].

In the follow-up paper [VSDL17b] of [VSDL17a], the authors relate the case of higher multiplicities to the *block term decomposition* for higher order tensors.

Remark 4.3.4. By the results of Subsection 3.1.1, the coordinates of the solutions may also be recovered from the *eigenvectors* of M_h. See for instance [CLO06, Chapter 2, §4 and Chapter 3, §6, Exercise 2], [Ste04, Page 52], [Cox20a, Page 50], [EM07, Section 4.7] or [CCC+05, Subsection 2.1.3]. The coordinates of the roots are the ratios between two entries of an eigenvector. This requires only one eigenvalue decomposition of M_h (the eigenvectors are usually computed by applying inverse iteration using the eigenvalues obtained from the Schur factorization [TBI97, Lecture 27]), instead of a Schur factorization $M_h = \mathbf{U}^H \mathbf{T} \mathbf{U}$ and $2n$ matrix-matrix multiplications $\Delta_{x_i} = \mathbf{U}^H M_{y_i} \mathbf{U}$. However, the speed-up is negligible compared to the other steps of the algorithm. Moreover, the coordinates can be obtained as the ratio between several different pairs of entries in the eigenvector, and if some solutions have very small or large coordinates, one should be careful which of these ratios to pick. In this thesis, we will work with the eigenvalues rather than the eigenvectors of the multiplication operators. △

Remark 4.3.5. Like in Section 3.3, it is possible to work over other fields than the complex numbers. In recent work by Avinash Kulkarni [Kul20], an adaptation of

Algorithm 4.1 is applied for solving systems of polynomial equations over the p-adics, using recent developments in 'p-numerical linear algebra'. $\triangle$

4.3.3 Numerical experiments

In this subsection we present some numerical experiments to illustrate the effectiveness of the TNF approach. We use Algorithm 4.1 from the previous subsection for computing the multiplication operators, where QR with pivoting is used for the basis selection (unless stated otherwise). For obtaining the solutions from these multiplication matrices, i.e., for diagonalizing them simultaneously, we use the Schur factorization of the multiplication operator corresponding to a generic linear form h. The algorithms are implemented in Matlab, version 2017a and executed in double precision arithmetic on an 8 GB RAM machine with an intel Core i7-6820HQ CPU working at 2.70 GHz. To measure the quality of the numerical approximations of the solutions, we use the *residual* as defined in Appendix C as a measure for the backward error. We should mention that the construction of the matrix $\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ in line 3 of Algorithm 4.1 is implemented in Fortran, because this step takes too much time in Matlab. We call the Fortran routine from Matlab using a MEX file. A Julia implementation of Algorithm 4.1 by Bernard Mourrain is accessible at `https://gitlab.inria.fr/AlgebraicGeometricModeling/AlgebraicSolvers.jl`.

Experiment 4.3.1 (Monomial bases for generic systems)**.** As a first experiment, for $n, d \in \mathbb{N}_{>0}$, we construct a generic member $\mathcal{F}_R(d, \ldots, d)$ (d is listed n times) by drawing the coefficients from a standard normal distribution. We apply Algorithm 4.1 and we look at the monomial basis $\mathcal{B}$ that is chosen using the QR algorithm with column pivoting. One could ask if it is (close to) being an order ideal or if it leads to a subspace $B \subset W$ that is connected to 1. The result for $n = 2, d = 15$ is shown in Figure 4.4, in comparison to the basis monomials used in Macaulay's construction (see Subsection 3.4.2), which is a nice order ideal. Note that, as before, we identify lattice points in the positive orthant with monomials of R. The basis chosen by the QR algorithm is not connected to 1, and it is not an order ideal. However, as we will see in Experiment 4.3.2, it leads to a tremendous improvement of the numerical behavior. We repeat the experiment, this time for $n = 3, d = 7$. The result, shown in Figure 4.5, is analogous. Finally, we have repeated the experiment 100 times for the families $n = 2, d = 30$ and $n = 3, d = 10$ and counted how many times each monomial of degree $\leq \rho = \hat{\rho} - 1$ occurred in the basis. The result is shown in Figure 4.6. $\triangle$

Experiment 4.3.2 (Improvement of QR bases with respect to Macaulay bases)**.** In this experiment, we show that choosing the basis $\mathcal{B}_{\mathrm{QR}}$ using QR with column pivoting leads to a great improvement of the accuracy of the computed solutions with respect to the basis $\mathcal{B}_{\mathrm{Mac}}$ coming from Macaulay's construction. We consider bivariate systems ($n = 2$) and for increasing values of d we generate generic members of $\mathcal{F}_R(d, d)$ with $R = \mathbb{C}[x, y]$ as in Experiment 4.3.1. The choice of B in step 6 of Algorithm 4.1 is made using either QR with column pivoting, resulting in the basis $\mathcal{B}_{\mathrm{QR}}$, or by selecting the

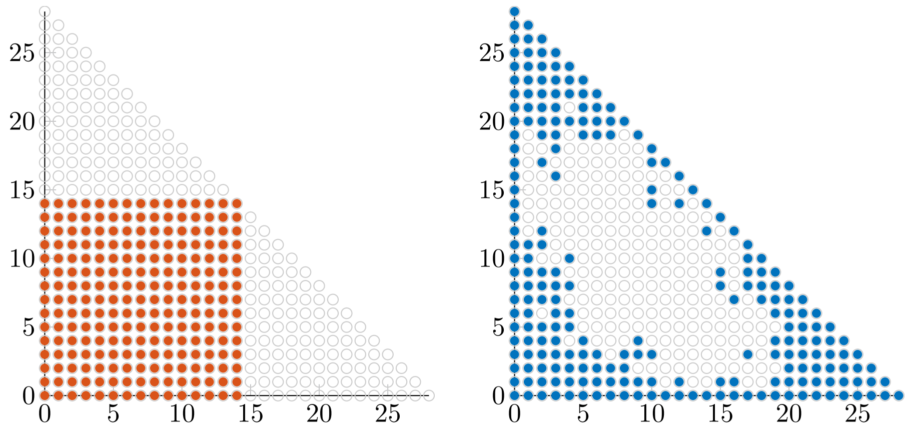


Figure 4.4: Monomials of degree $\leq \rho = \hat{\rho} - 1$ that are chosen to represent the quotient algebra associated to a generic member of $\mathcal{F}_{\mathbb{C}[x,y]}(15, 15)$ in the method of Subsection 3.4.2 (left) and in Algorithm 4.1 using QR with column pivoting (right).

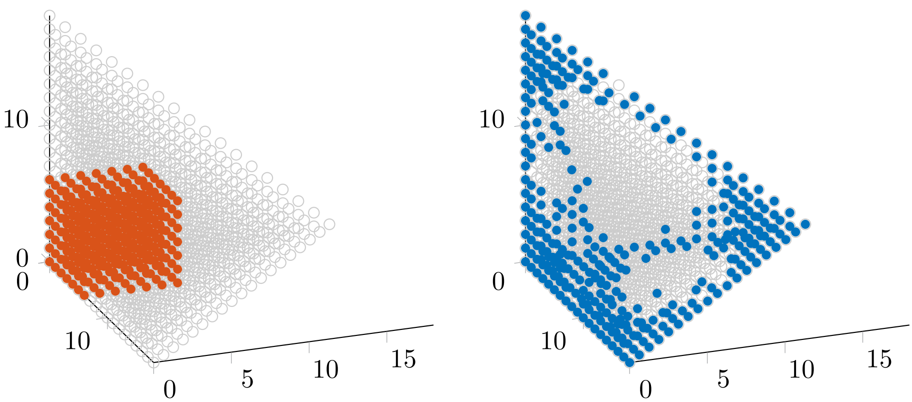


Figure 4.5: Monomials of degree $\leq \rho = \hat{\rho} - 1$ that are chosen to represent the quotient algebra associated to a generic member of $\mathcal{F}_{\mathbb{C}[x,y,z]}(7, 7, 7)$ in the method of Subsection 3.4.2 (left) and in Algorithm 4.1 using QR with column pivoting (right).

columns of N corresponding to the Macaulay basis

$$\mathcal{B}_{\text{Mac}} = \{x^{a_1} y^{a_2} \mid a_1 < d, a_2 < d\}.$$

For $d = 2, 3, \ldots, 20$, we compute the condition number κ of the matrix $N_{|B}$, the maximal ($r_{\max}$) and minimal ($r_{\min}$) residual of all solutions, and also the geometric mean of the residuals r_{mean} of all computed solutions. The results are reported in Figure 4.7. The figure shows the results averaged out over 10 experiments. That is, it shows the geometric mean of the condition numbers, minimal, maximal and

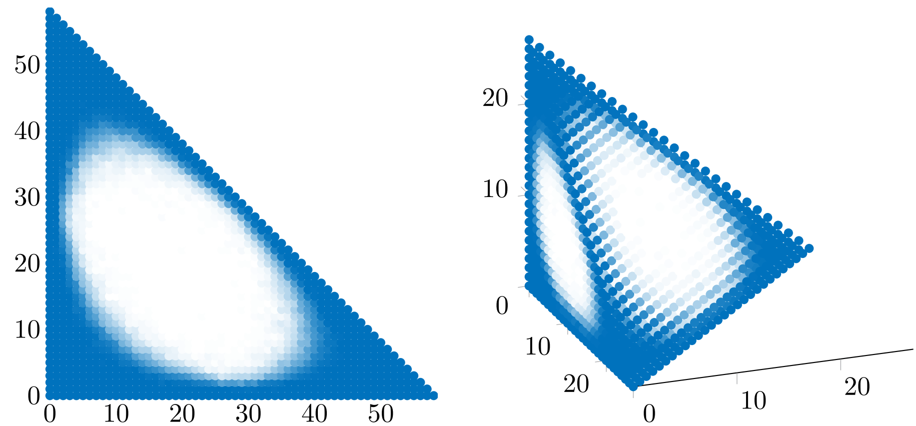


Figure 4.6: Illustration of how many times the monomials of degree $\leq \hat{\rho}$ are chosen to represent the quotient algebra associated to a generic member of $\mathcal{F}_{\mathbb{C}[x,y]}(30,30)$ (left) and $\mathcal{F}_{\mathbb{C}[x,y,z]}(10,10,10)$ (right) by Algorithm 4.1 using QR with column pivoting. The number of times the monomial is chosen is represented by the intensity of the color of the corresponding lattice point.

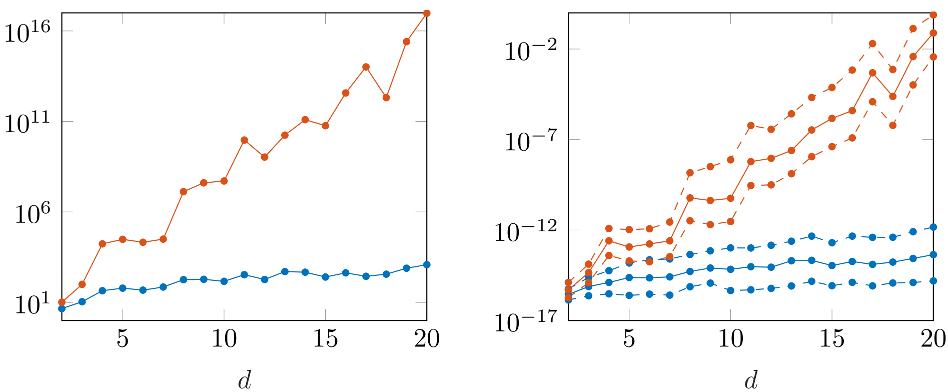


Figure 4.7: Average condition number of $N_{|B}$ (left) and $r_{\max}$, $r_{\min}$, r_{mean} (right) for the TNF solver using $\mathcal{B}_{\text{Mac}}$ (orange) and $\mathcal{B}_{\text{QR}}$ (blue) for solving generic members of $\mathcal{F}_R(d,d)$, $d = 2, 3, \ldots, 20$.

mean residuals of 10 different runs. It is clear that choosing the monomial basis using QR with column pivoting (illustrated in Experiment 4.3.1) leads to a significant improvement. The figure also shows that the large condition number of $N_{|B}$ for $\mathcal{B}_{\text{Mac}}$ is what's behind the loss of accuracy. $\triangle$

Experiment 4.3.3 (Comparison with PNLA)**.** PNLA is a Matlab package that can be used for several kinds of computations with multivariate polynomials, including

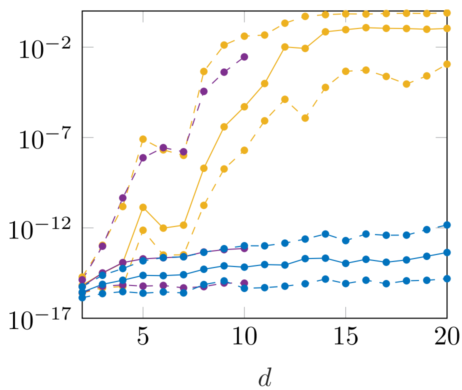


Figure 4.8: Values of $r_{\max}$, $r_{\min}$ and r_{mean} for the TNF solver (blue), `qdsparf` (yellow) and `sparf` (purple) in Experiment 4.3.3.

system solving. The software is available at `https://github.com/kbatseli/PNLA_MATLAB_OCTAVE`. The package implements the algorithms described in [Bat13]. The function `sparf` can be used for general purpose isolated affine root finding. It builds larger and larger resultant maps until the cokernel map can be used for computing all the roots. The function `qdsparf` is the 'quick and dirty' alternative for `sparf`, which is expected to be faster but in some cases less accurate. We compare both methods against Algorithm 4.1 using QR with column pivoting followed by Schur factorization for simultaneous diagonalization. The systems we solve are generic members of $\mathcal{F}_R(d,d)$ (i.e. we fix $n=2$) as in Experiment 4.3.1. We should note that the PNLA functions a priori do not make any assumptions on the system: it may have solutions at infinity, and it may even be overdetermined. It does not exploit the fact that the system is a general member of $\mathcal{F}_R(d,d)$. However, `sparf` and `qdsparf` should certainly be able to handle such systems. We compute the maximal ($r_{\max}$), minimal ($r_{\min}$) and mean (r_{mean}) residual of all computed solutions for $d = 2, 3, \ldots, 20$. The result, averaged out over 10 experiments, is shown in Figure 4.8. It is clear that the TNF solver gives better results. The function `sparf` gave errors for $d > 10$, so for this method results are only reported up to $d = 10$. As mentioned above, Figure 4.8 only shows the residuals of the *computed* solutions. In fact, the PNLA solvers do not compute numerical approximations for *all* roots of the system. The difference between the actual number of roots d^2 and the number of computed roots is denoted e_{TNF}, $e_{\texttt{sparf}}$, $e_{\texttt{qdsparf}}$ for the different solvers. These numbers are reported, together with the computation times t_{TNF}, $t_{\texttt{sparf}}$, $t_{\texttt{qdsparf}}$, in Table 4.1. $\triangle$

Experiment 4.3.4 (Comparison with Gröbner bases)**.** This is an experiment taken from [MTVB19, Subsection 6.3]. We have seen in Section 3.3 that Gröbner bases can be used to compute multiplication matrices. Let $\mathcal{G}$ be a Gröbner basis with respect to a given monomial order '$\prec$'. The set of standard monomials is denoted by $\mathcal{B}_\prec = \{x^{a_1}, \ldots, x^{a_{\delta+}}\}$. The j-th column of the multiplication matrix M_{x_i} is then given by $\mathcal{N}_\mathcal{G}(x_i x^{a_j})$. This gives an algorithm for finding the multiplication operators

d	t_{TNF}	$t_{\texttt{qdsparf}}$	$t_{\texttt{sparf}}$	e_{TNF}	$e_{\texttt{qdsparf}}$	$e_{\texttt{sparf}}$
2	$1.73 \cdot 10^{-3}$	$5.4 \cdot 10^{-3}$	$1.04 \cdot 10^{-2}$	0	0	0
3	$2.39 \cdot 10^{-3}$	$8.53 \cdot 10^{-3}$	$2.49 \cdot 10^{-2}$	0	0	0
4	$3.39 \cdot 10^{-3}$	$1.76 \cdot 10^{-2}$	$6.46 \cdot 10^{-2}$	0	0	0
5	$6.84 \cdot 10^{-3}$	$3.16 \cdot 10^{-2}$	0.14	0	0.2	0
6	$1.18 \cdot 10^{-2}$	$5.41 \cdot 10^{-2}$	0.29	0	0.1	0
7	$1.62 \cdot 10^{-2}$	$8.88 \cdot 10^{-2}$	0.54	0	0.1	0
8	$2.6 \cdot 10^{-2}$	0.14	0.96	0	0.3	0
9	$3.2 \cdot 10^{-2}$	0.2	1.68	0	0.6	0
10	$4.24 \cdot 10^{-2}$	0.29	2.84	0	0.9	0
11	$5.47 \cdot 10^{-2}$	0.39		0	2.5	
12	$7.17 \cdot 10^{-2}$	0.55		0	1.6	
13	$9.77 \cdot 10^{-2}$	0.72		0	3.1	
14	0.12	0.94		0	3.7	
15	0.15	1.17		0	5.7	
16	0.19	1.53		0	5.6	
17	0.23	1.85		0	9.5	
18	0.28	2.31		0	9.8	
19	0.34	2.9		0	10.4	
20	0.42	3.34		0	18.4	

Table 4.1: Average timing results and average number of missed solutions for the TNF solver, `qdsparf` and `sparf` in Experiment 4.3.3.

M_{x_i}. Table 4.2 summarizes the steps of the algorithm and gives the corresponding steps of the TNF algorithm. We use Faugère's FGb in Maple™ for step 1 [Fau10].

	TNF-QR algorithm	GB algorithm
1	Construct the resultant map and compute N	Compute a Gröbner basis $\mathcal{G}$ which induces a normal form $\mathcal{N}_{\mathcal{G}}$
2	QR with pivoting on $N_{\vert W}$ to find $N_{\vert B}$ corresponding to a basis $\mathcal{B}$ of R/I	Find a normal set $\mathcal{B}_{\prec}$ from $\mathcal{G}$
3	Compute the N_i and set $M_{x_i} = (N_{\vert B})^{-1} N_i$	Compute the multiplication matrices by applying the induced normal form $\mathcal{N}_{\mathcal{G}}$ on $x_i \cdot \mathcal{B}$

Table 4.2: Corresponding steps of the TNF algorithm and the Gröbner basis algorithm

This is considered state of the art software for computing Gröbner bases. The routine `fgb_gbasis` computes a Gröbner basis with respect to the degree reverse lexicographic ($\prec_{\text{drl}}$) monomial order. For step 2, we use the command `NormalSet` from the built-in Maple package `Groebner` to compute a normal set from this Gröbner basis. Step 3 is done using the command `MultiplicationMatrix` from the `Groebner` package.

An important note is that the Gröbner basis computation is performed in *exact arithmetic*. In this experiment we compare the speed of our algorithm with that of the Gröbner basis algorithm for computing the matrices M_{x_i}. This is, in a sense, comparing apples and oranges. Of course, a speed-up with respect to exact arithmetic is to be expected. The goal of this experiment is to *quantify* this speed-up. The price we pay for this speed-up (i.e. a numerical approximation error on the computed result) is quantified more in detail in different experiments. We note that the residuals for all tests in this experiment were no larger than 10^{-10}. TNFs may offer a numerical algebraic alternative for Gröbner basis computation if one is happy with accurate approximations of the multiplication matrices.

To compute the roots of the system, one can compute the eigenvalues of the approximate multiplication operators obtained via Algorithm 4.1, or of the *exact* multiplication operators obtained from a Gröbner basis, by using a numerical method. This solving step is not integrated in the comparison.

We perform two different experiments: one in which the coefficients are floating point numbers up to 16 digits of accuracy that are converted in Maple to rational numbers, and one in which the coefficients are integers, uniformly distributed between -50 and 50. We restrict Matlab to the use of only one core since Maple also uses only one. For different n, we construct a generic member of $\mathcal{F}_R(d,\dots,d)$ as in Experiment 4.3.1. We compare the computation time needed for finding the multiplication matrices using our algorithm with the time needed for the Gröbner basis algorithm as described in Table 4.2. The float coefficients are approximated up to 16 digits of accuracy by a rational number in Maple, before starting the computation. This results in rational numbers with large numerators and denominators, which makes the computation in exact arithmetic very time consuming. Results are shown in Table 4.3. We conclude that the TNF method using floating point arithmetic can lead to a huge reduction of the computation time in these situations and, with the right choice of basis for the quotient algebra, the loss of accuracy is very small.

We now construct a generic member of $\mathcal{F}_R(d,\dots,d)$ by drawing the coefficients from a discrete uniform distribution on the integers $-50,\dots,50$ for each of the n polynomials defining the system. Table 4.4 shows that the Gröbner basis method in exact precision is faster with these 'simple' coefficients, but the speed-up by using the TNF algorithm with floating point arithmetic is still significant.

$\triangle$

Experiment 4.3.5 (Comparison with homotopy solvers)**.** This experiment is taken from Subsection 8.5 in [TMVB18]. As mentioned in Chapter 1, a popular approach for solving polynomial equations numerically is *homotopy continuation*. We compare the speed and accuracy of our method to that of the homotopy implementations PHCpack (v2.4.64) [Ver99] and Bertini (v1.5.1) [BSHW13], which are considered state of the art. Later versions of these packages give similar results, as we will see in Chapter 6.

We use double precision arithmetic for all computations and standard settings for

n	d	t_{TNF}	t_{GB}	$t_{\mathrm{GB}}/t_{\mathrm{TNF}}$
2	2	$5.68 \cdot 10^{-4}$	$1.52 \cdot 10^{-2}$	26.76
2	3	$1.88 \cdot 10^{-3}$	$2.51 \cdot 10^{-2}$	13.34
2	4	$2.3 \cdot 10^{-3}$	$5.88 \cdot 10^{-2}$	25.57
2	5	$3.9 \cdot 10^{-3}$	0.19	47.96
2	6	$5.98 \cdot 10^{-3}$	0.48	79.55
2	7	$8.03 \cdot 10^{-3}$	1.16	143.89
2	8	$1.24 \cdot 10^{-2}$	2.85	229.04
2	9	$1.75 \cdot 10^{-2}$	6.19	354.39
2	10	$2.49 \cdot 10^{-2}$	14.27	573.24
3	2	$2.1 \cdot 10^{-3}$	$5.66 \cdot 10^{-2}$	27
3	3	$9.49 \cdot 10^{-3}$	1.82	191.54
3	4	$3.43 \cdot 10^{-2}$	52.19	1,520.51
3	5	0.12	893.38	7,186.04
4	2	$1.2 \cdot 10^{-2}$	1.31	109.76
4	3	0.27	910.96	3,391.25
5	2	0.15	59	398.27

Table 4.3: Timing results for the TNF algorithm (t_{TNF} (sec)) and the Gröbner basis algorithm in Maple (t_{GB} (sec)) for generic systems in n variables of degree d with floating point coefficients drawn from a normal distribution with zero mean and $\sigma = 1$.

Bertini and PHCpack apart from that. We use the command `solve_system` from the Matlab interface PHClab [GV08] for PHCpack and we run Bertini via the command `system` in Matlab, which calls the operating system to execute Bertini commands. By a *generic dense system* of degree d in n variables we mean a generic member $(f_1, \ldots, f_n) \in \mathcal{F}_R(d, \ldots, d)$, where R is the polynomial ring in n variables and d is listed n times. For the experiment we fix a value of n and generate generic dense systems of increasing degree d as in Experiment 4.3.1 to use as input for the different solvers.

Tables 4.5 up to 4.12 give detailed results of the experiment. The following notation is used in the tables. The number of solutions of the input system is δ (in this case, $\delta = \delta^+ = d^n$). The numbers $m_1, m_2 = n_1, n_2$ give the sizes of $(\mathrm{res}_{f_1,\ldots,f_n})^\top \in \mathbb{C}^{m_1 \times m_2}$ and $N \in \mathbb{C}^{n_1 \times n_2}$. The maximal residual of the solutions computed by the TNF solver is denoted by $r_{\max}$. By $e_{\mathrm{TNF}}, e_{\mathrm{phc}}, e_{\mathrm{brt}}$ we denote the number of 'missed' solutions for the TNF solver, PHCpack and Bertini respectively. This is equal to d^n minus the number of computed solutions. Since the homotopy methods use Newton refinement intrinsically, their computed solutions give residuals of the order of the unit roundoff. The values t_M, t_N, t_B, t_S denote the time for the construction of the resultant map (Fortran), the computation of its cokernel, the computation of the basis via QR together with the construction of the multiplication matrices and the time to compute the simultaneous Schur decomposition respectively. The total computation times are $t_{\mathrm{TNF}}, t_{\mathrm{phc}}$ and t_{brt} for the TNF solver ($t_{\mathrm{TNF}} = t_M + t_N + t_B + t_S$), PHCpack and Bertini respectively. All timings are in seconds. Tables 4.5 and 4.6 present the experiment for $n = 2$ variables, Tables 4.7 and 4.8 for $n = 3$, Tables 4.9 and 4.10 for

n	d	t_{TNF}	t_{GB}	$t_{\mathrm{GB}}/t_{\mathrm{TNF}}$
2	2	$6.09 \cdot 10^{-4}$	$1.1 \cdot 10^{-2}$	18.06
2	4	$2.3 \cdot 10^{-3}$	$1.82 \cdot 10^{-2}$	7.91
2	6	$8.75 \cdot 10^{-3}$	$3 \cdot 10^{-2}$	3.43
2	8	$1.24 \cdot 10^{-2}$	$8.1 \cdot 10^{-2}$	6.51
2	10	$2.48 \cdot 10^{-2}$	0.15	5.88
2	12	$4.24 \cdot 10^{-2}$	0.38	8.89
2	14	$6.73 \cdot 10^{-2}$	0.71	10.56
2	16	0.1	1.32	12.62
2	18	0.16	2.33	14.91
2	20	0.2	4.31	21.42
2	22	0.29	7.07	24.64
2	24	0.5	11.55	23.09
2	26	0.62	19.36	31.08
2	28	0.81	29.25	36.22
2	30	1.08	41.01	37.89
3	2	$2.47 \cdot 10^{-3}$	$1.74 \cdot 10^{-2}$	7.05
3	3	$9.82 \cdot 10^{-3}$	$6.1 \cdot 10^{-2}$	6.21
3	4	$3.17 \cdot 10^{-2}$	0.33	10.4
3	5	$9.38 \cdot 10^{-2}$	2.09	22.33
3	6	0.27	10.42	38.67
3	7	1.31	45.4	34.62
3	8	5.3	168.03	31.72
3	9	16.16	573.45	35.5
3	10	41.71	1,674	40.14
4	2	$1.27 \cdot 10^{-2}$	$5.8 \cdot 10^{-2}$	4.58
4	3	0.18	3.19	17.86
4	4	8.89	99.78	11.23
4	5	145.36	2,367.04	16.28
5	2	$9.32 \cdot 10^{-2}$	0.4	4.28
5	3	73.16	286.15	3.91

Table 4.4: Timing results for the TNF algorithm (t_{TNF} (sec)) and the Gröbner basis algorithm in Maple (t_{GB} (sec)) for generic systems in n variables of degree d with integer coefficients uniformly distributed between -50 and 50.

$n = 4$ and Tables 4.11 and 4.12 for $n = 5$.

We observe that our method has found numerical approximations for *all* d^n roots, with a residual no larger than order 10^{-9}. Due to the quadratic convergence of Newton's iteration, one refining step can be expected to result in a residual of the order of the unit roundoff. Table 4.5 shows that for 2 variables, up to degree $d = 61$, our method is the fastest. For $n = 3$ this is no longer the case but timings are comparable. For a larger number of variables, the matrix of the resultant map in the algorithms becomes very large and the cokernel computation is expensive, which makes the algebraic method slower than the continuation solvers.

An important note is that homotopy methods do not guarantee that all solutions are found. In fact, they lose some solutions for large systems. For $n = 2, d = 55$, Bertini

d	δ	m_1	$m_2{=}n_1$	n_2	$r_{\max}$	e_{TNF}	e_{phc}	e_{brt}
1	1	2	3	1	$1.28 \cdot 10^{-16}$	0	0	0
7	49	56	105	49	$2.06 \cdot 10^{-13}$	0	0	0
13	169	182	351	169	$2.18 \cdot 10^{-13}$	0	0	0
19	361	380	741	361	$5.28 \cdot 10^{-13}$	0	0	0
25	625	650	1,275	625	$1.21 \cdot 10^{-10}$	0	11	0
31	961	992	1,953	961	$5.23 \cdot 10^{-9}$	0	10	0
37	1,369	1,406	2,775	1,369	$4.05 \cdot 10^{-12}$	0	9	1
43	1,849	1,892	3,741	1,849	$1.74 \cdot 10^{-11}$	0	24	4
49	2,401	2,450	4,851	2,401	$1.57 \cdot 10^{-10}$	0	37	38
55	3,025	3,080	6,105	3,025	$1.84 \cdot 10^{-11}$	0	55	538
61	3,721	3,782	7,503	3,721	$3.26 \cdot 10^{-11}$	0	59	1,461

Table 4.5: Numerical results for PHCpack, Bertini and our method for dense systems in $n = 2$ variables of increasing degree d. The table shows matrix sizes, accuracy and number of solutions.

d	t_M	t_N	t_B	t_S	t_{TNF}	t_{phc}	t_{brt}
1	$1.48 \cdot 10^{-4}$	$5.5 \cdot 10^{-5}$	$2.96 \cdot 10^{-4}$	$3.6 \cdot 10^{-5}$	$5.35 \cdot 10^{-4}$	$5.6 \cdot 10^{-2}$	$1.41 \cdot 10^{-2}$
7	$7.88 \cdot 10^{-3}$	$1.68 \cdot 10^{-3}$	$3.76 \cdot 10^{-3}$	$2.78 \cdot 10^{-3}$	$1.61 \cdot 10^{-2}$	0.18	$8.65 \cdot 10^{-2}$
13	$4.65 \cdot 10^{-2}$	$1.03 \cdot 10^{-2}$	$1.66 \cdot 10^{-2}$	$2.81 \cdot 10^{-2}$	0.1	0.84	1.14
19	0.13	$5.69 \cdot 10^{-2}$	$5.34 \cdot 10^{-2}$	0.13	0.37	3.29	8.79
25	0.32	0.18	0.15	0.51	1.16	8.79	33.83
31	0.55	0.51	0.55	1.49	3.1	20.25	98.39
37	0.96	1.52	1.5	3.52	7.5	39.92	258.09
43	1.47	4.05	3.8	8.28	17.6	69.1	504.01
49	2.47	10.46	8.78	17.91	39.62	124.47	891.37
55	3.69	20.51	17.85	34.3	76.34	178.55	1,581.77
61	4.85	36.32	31.26	62.87	135.3	283.87	2,115.66

Table 4.6: Timing results for PHCpack, Bertini and our method for dense systems in $n = 2$ variables of increasing degree d.

d	δ	m_1	$m_2{=}n_1$	n_2	$r_{\max}$	e_{TNF}	e_{phc}	e_{brt}
1	1	3	4	1	$1.79 \cdot 10^{-16}$	0	0	0
3	27	105	120	27	$1.05 \cdot 10^{-14}$	0	0	0
5	125	495	560	125	$1.29 \cdot 10^{-12}$	0	0	0
7	343	1,365	1,540	343	$6.71 \cdot 10^{-12}$	0	0	0
9	729	2,907	3,276	729	$1.38 \cdot 10^{-10}$	0	3	0
11	1,331	5,313	5,984	1,331	$3.11 \cdot 10^{-11}$	0	0	0
13	2,197	8,775	9,880	2,197	$2.86 \cdot 10^{-11}$	0	5	0

Table 4.7: Numerical results for PHCpack, Bertini and our method for dense systems in $n = 3$ variables of increasing degree d. The table shows matrix sizes, accuracy and number of solutions.

gives up on 538 out of 3025 paths, so about 18% of the solutions is not found (using default settings). For the same problem, PHCpack loses 2% of the solutions.

In this experiment, we did not include a comparison with the relatively new Julia package HomotopyContinuation.jl [BT18]. The reason is that the software was not

d	t_M	t_N	t_B	t_S	t_{TNF}	t_{phc}	t_{brt}
1	$3.72 \cdot 10^{-4}$	$1.24 \cdot 10^{-4}$	$2.31 \cdot 10^{-3}$	$4.5 \cdot 10^{-5}$	$2.85 \cdot 10^{-3}$	$6.8 \cdot 10^{-2}$	$1.69 \cdot 10^{-2}$
3	$7.91 \cdot 10^{-3}$	$2.42 \cdot 10^{-3}$	$7.06 \cdot 10^{-3}$	$1.08 \cdot 10^{-3}$	$1.85 \cdot 10^{-2}$	0.14	$7.33 \cdot 10^{-2}$
5	$5.66 \cdot 10^{-2}$	$3.93 \cdot 10^{-2}$	$3.31 \cdot 10^{-2}$	$1.17 \cdot 10^{-2}$	0.14	0.68	0.63
7	0.23	1.13	0.12	$9.9 \cdot 10^{-2}$	1.57	3.42	4.11
9	0.68	14.43	0.65	0.63	16.4	12.21	17.29
11	1.77	44.79	3.91	3.98	54.46	39.08	70.66
13	5.81	183.67	16.07	15.35	220.9	97.28	210.34

Table 4.8: Timing results for PHCpack, Bertini and our method for dense systems in $n = 3$ variables of increasing degree d.

d	δ	m_1	$m_2{=}n_1$	n_2	$r_{\max}$	e_{TNF}	e_{phc}	e_{brt}
1	1	4	5	1	$1.24 \cdot 10^{-16}$	0	0	0
2	16	140	126	16	$1.13 \cdot 10^{-14}$	0	0	0
3	81	840	715	81	$3.84 \cdot 10^{-14}$	0	0	0
4	256	2,860	2,380	256	$1.52 \cdot 10^{-13}$	0	0	1

Table 4.9: Numerical results for PHCpack, Bertini and our method for dense systems in $n = 4$ variables of increasing degree d. The table shows matrix sizes, accuracy and number of solutions.

d	t_M	t_N	t_B	t_S	t_{TNF}	t_{phc}	t_{brt}
1	$1.1 \cdot 10^{-2}$	$2.83 \cdot 10^{-4}$	$1.83 \cdot 10^{-2}$	$8.43 \cdot 10^{-4}$	$3.04 \cdot 10^{-2}$	$6.82 \cdot 10^{-2}$	$1.76 \cdot 10^{-2}$
2	$1.12 \cdot 10^{-2}$	$4.29 \cdot 10^{-3}$	$1.08 \cdot 10^{-2}$	$5.94 \cdot 10^{-4}$	$2.69 \cdot 10^{-2}$	0.12	$6.32 \cdot 10^{-2}$
3	0.11	0.14	$5.76 \cdot 10^{-2}$	$5.55 \cdot 10^{-3}$	0.31	0.52	0.59
4	0.46	8.31	0.23	$5.41 \cdot 10^{-2}$	9.05	2.27	3.62

Table 4.10: Timing results for PHCpack, Bertini and our method for dense systems in $n = 4$ variables of increasing degree d.

d	δ	m_1	$m_2{=}n_1$	n_2	$r_{\max}$	e_{TNF}	e_{phc}	e_{brt}
1	1	5	6	1	$7.89 \cdot 10^{-17}$	0	0	0
2	32	630	462	32	$4.22 \cdot 10^{-14}$	0	0	0
3	243	6,435	4,368	243	$1.84 \cdot 10^{-12}$	0	0	0

Table 4.11: Numerical results for PHCpack, Bertini and our method for dense systems in $n = 5$ variables of increasing degree d. The table shows matrix sizes, accuracy and number of solutions.

d	t_M	t_N	t_B	t_S	t_{TNF}	t_{phc}	t_{brt}
1	$4.87 \cdot 10^{-4}$	$1.54 \cdot 10^{-4}$	$1.86 \cdot 10^{-3}$	$3 \cdot 10^{-5}$	$2.53 \cdot 10^{-3}$	$6.52 \cdot 10^{-2}$	$1.91 \cdot 10^{-2}$
2	$5.97 \cdot 10^{-2}$	$3.9 \cdot 10^{-2}$	$4.07 \cdot 10^{-2}$	$1.46 \cdot 10^{-3}$	0.14	0.26	0.24
3	1.21	69.38	0.53	$5.5 \cdot 10^{-2}$	71.18	2.42	4.74

Table 4.12: Timing results for PHCpack, Bertini and our method for dense systems in $n = 5$ variables of increasing degree d.

yet available at the time we wrote [TMVB18] and it is cumbersome to call it from Matlab. We remark that HomotopyContinuation.jl is a very promising package: the numerical path tracking happens amazingly fast and the implementation is also very robust. For instance, a generic member of $\mathcal{F}_{\mathbb{C}[x,y]}(50, 50)$ is solved within *less than a second* and often no solutions are lost: in 1000 runs, there was 1 missing solution for 178 systems, 2 missing solutions for 12 systems and no missing solutions for all 810 other systems (we used v1.4.1 for this). We will say more about this package and its performance in Chapter 6. $\triangle$

Experiment 4.3.6 (Intersecting plane curves of degree 170)**.** Experiment 4.3.5 shows that the TNF solver is robust for generic problems in 2 variables up to degree at least 61. In this experiment, we will push this much further: we use Algorithm 4.1 to solve generic members of $\mathcal{F}_{\mathbb{C}[x,y]}(170, 170)$. We also show what the algorithm can do for higher values of n. For this experiment we use a 128 GB RAM machine with a Xeon E5-2697 v3 CPU working at 2.60 GHz. Generic square systems in n variables of degree d are generated as in the previous experiments. The distribution of the residuals for all solutions for some values of n, d are shown in Figure 4.9. Note that plane curves of degree 170 are no problem for the TNF algorithm with pivoted QR for basis selection, while the classical Macaulay construction fails to give any meaningful results for $d = 20$ (Experiment 4.3.2) and homotopy solvers start missing solutions for $d \geq 40$ (or even smaller). The computation times for $n = 2, d = 100, 150, 170$ were approximately 53 minutes, 8 hours and 49 minutes and 19 hours respectively. Some other timings are

$$\begin{array}{cc|cc} n=3, d=15: & 8\,\text{min} & n=5, d=4: & 37\,\text{min} \\ n=3, d=20: & \text{1h } 1\,\text{min} & n=6, d=3: & \text{1h } 37\,\text{min} \\ n=3, d=23: & \text{3h } 26\,\text{min} & n=7, d=2: & 1\,\text{min} \\ n=4, d=8: & \text{8h } 12\,\text{min} & n=8, d=2: & \text{1h } 17\,\text{min} \end{array}$$

For higher degrees than $170, 23, 8, 4, 3, 2, 2$ for $n = 2, 3, 4, 5, 6, 7, 8$ respectively, the machine ran into memory problems. $\triangle$

4.4 Improvements and generalizations

This section is based on Sections 4 and 5 of [MTVB19]. In Subsection 4.4.1 we discuss two possible ways of reducing the computational complexity of computing the cokernel of a resultant map. We show with an experiment that this reduces the computation time significantly for $n > 2$. In Subsection 4.4.2, we discuss two natural ways of using non-monomial bases for constructing TNFs.

4.4.1 Fast cokernel computation

The TNF method for solving polynomial systems, like other algebraic approaches, has the important drawback that the complexity scales badly with the number n of

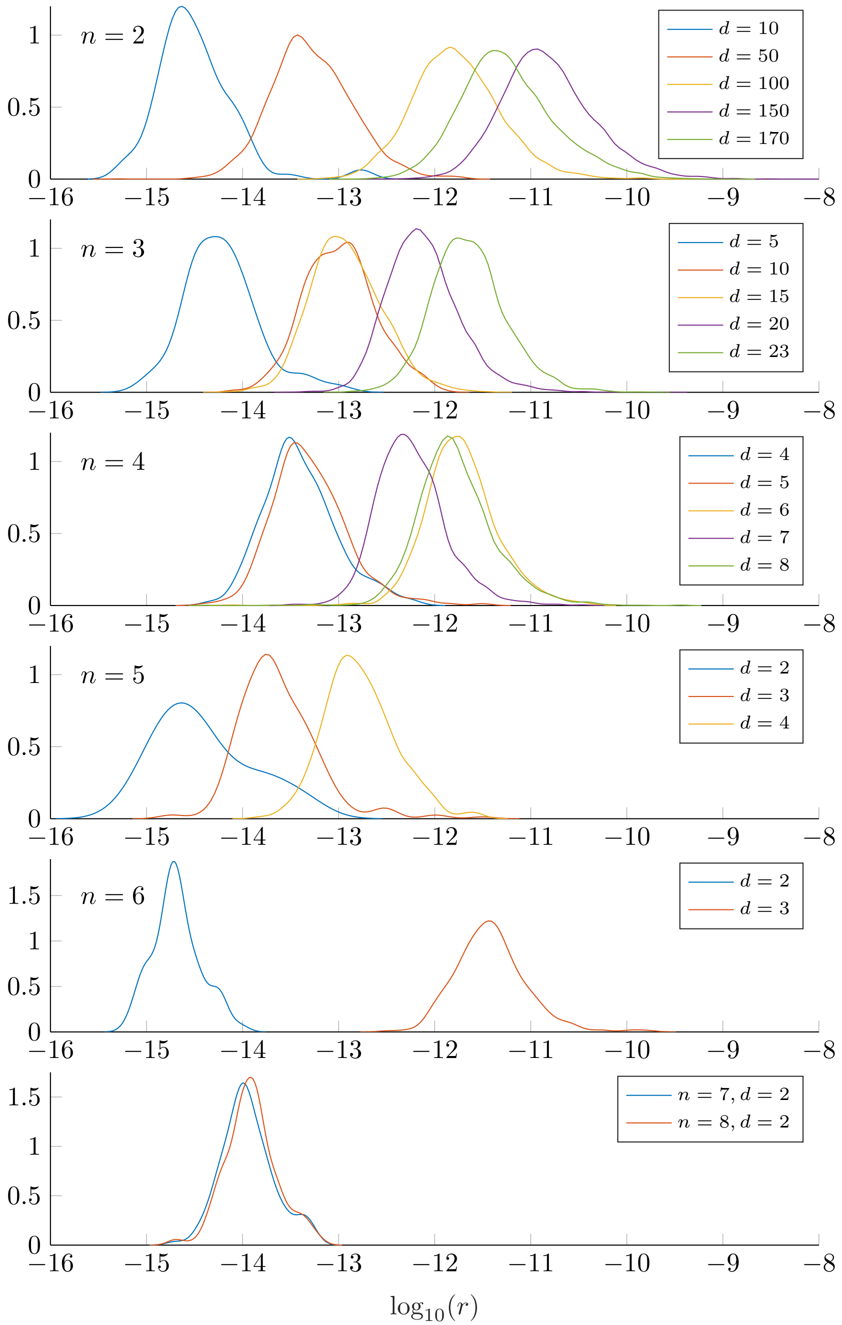


Figure 4.9: Density functions of the $\log_{10}$ of the residuals of all numerical solutions computed by Algorithm 4.1 for $n = 2, \ldots, 8$ and different values of d.

variables. This is due to the fact that the complexity of computing the cokernel map of the appropriate resultant map increases drastically with n. This is illustrated, for example, in Experiment 4.3.5. We describe two possible techniques to reduce this drastic increase of complexity. The first one computes the cokernel map degree by degree. This technique has also been exploited in [BDDM14]. The second one exploits the redundancy in the vector spaces V_i in the definition of the resultant map.

Computing the cokernel degree by degree

Let $I = \langle \hat{f}_1, \ldots, \hat{f}_s \rangle \subset R$ with $(\hat{f}_1, \ldots, \hat{f}_s) \in \mathcal{F}_R(d_1, \ldots, d_s)$. We consider a resultant map

$$\mathrm{res} = \mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_s} : V_1 \times \cdots \times V_s \to V$$

where $V = R_{\leq d}, V_i = R_{\leq d-d_i}$ for some degree d. Our aim is to compute a cokernel map of res. We define the resultant maps

$$\mathrm{res}_k : V_{1,k} \times \cdots \times V_{s,k} \to V(k), \quad k = 1, \ldots, d$$

such that $V(k) = R_{\leq k}$, $V_{i,k} = R_{\leq k-d_i}$ with the convention that $R_{\leq k} = \{0\}$ when $k < 0$. Let $N_k : V_k \to \mathbb{C}^{\delta_k}$ be a cokernel map of res_k. We have that $\mathrm{res}_d = \mathrm{res}$ and $N_d = N$ is the map we want to compute. Our aim here is to compute N_{k+1} from N_k in an efficient way. Note that $V(k) \subset V(k+1), V_{i,k} \subset V_{i,k+1}$. We write

$$\mathrm{res}_{k+1} : \prod_{i=1}^{s} V_{i,k} \times T_{k+1} \to V(k+1)$$

where $T_{k+1} \simeq \prod_{i=1}^{s} V_{i,k+1}/V_{i,k}$ and $(\mathrm{res}_{k+1})_{|\prod_{i=1}^{s} V_{i,k}} = \mathrm{res}_k$. Define

$$N_k \times \mathrm{id} : V(k) \times \frac{V(k+1)}{V(k)} \to \mathbb{C}^{\delta_k} \times \frac{V(k+1)}{V(k)} \quad \text{given by} \quad (v, w) \mapsto (N_k(v), w).$$

Furthermore, set $\mathrm{res}'_{k+1} = (\mathrm{res}_{k+1})_{|T_{k+1}}$. Here is what the matrices look like:

$$N_k \times \mathrm{id} = \begin{array}{c c} & \begin{array}{cc} V(k) & \frac{V(k+1)}{V(k)} \end{array} \\ \begin{array}{c} \mathbb{C}^{\delta_k} \\ \frac{V(k+1)}{V(k)} \end{array} & \left[\begin{array}{c|c} N_k & 0 \\ \hline 0 & \mathrm{id} \end{array}\right] \end{array}, \qquad \mathrm{res}_{k+1} = \begin{array}{c c} & \begin{array}{cc} \prod_{i=1}^{s} V_{i,k} & T_{k+1} \end{array} \\ \begin{array}{c} V(k) \\ \frac{V(k+1)}{V(k)} \end{array} & \left[\begin{array}{c|c} \mathrm{res}_k & A_{k+1} \\ \hline 0 & B_{k+1} \end{array}\right] \end{array},$$

where res'_{k+1} is represented as a block matrix $\begin{bmatrix} A_{k+1} \\ B_{k+1} \end{bmatrix}$. Finally, define

$$L_{k+1} : \mathbb{C}^{\delta_k} \times \frac{V(k+1)}{V(k)} \to \mathbb{C}^{\delta_{k+1}}$$

as the cokernel of $(N_k \times \mathrm{id}) \circ \mathrm{res}'_{k+1}$.

Theorem 4.4.1. *The map* $N_{k+1} = L_{k+1} \circ (N_k \times \mathrm{id})$ *is a cokernel map of* res_{k+1}*, i.e.* $\mathrm{im}\,\mathrm{res}_{k+1} = \ker N_{k+1}$ *and* N_{k+1} *is onto* $\mathbb{C}^{\delta_{k+1}}$.

Proof. We have the commutative diagram

$$\begin{tikzcd}
T_{k+1} \arrow[r, "(N_k \times \mathrm{id})\circ\ \mathrm{res}'_{k+1}"] & \mathbb{C}^{\delta_k} \times \frac{V(k+1)}{V(k)} \arrow[r, "L_{k+1}"] & \mathbb{C}^{\delta_{k+1}} \arrow[r] & 0 \\
\prod_{i=1}^{s} V_{i,k} \times T_{k+1} \arrow[u] \arrow[r, "\mathrm{res}_{k+1}"] & V(k) \times \frac{V(k+1)}{V(k)} \arrow[u, "N_k \times \mathrm{id}"] \arrow[ur, "N_{k+1}"'] & &
\end{tikzcd}$$

where the upward pointing arrow on the left is projection onto T_{k+1} and the top row is exact by the definition of L_{k+1}. The map $N_{k+1} = L_{k+1} \circ (N_k \times \mathrm{id})$ is onto since both L_{k+1} and $(N_k \times \mathrm{id})$ are onto.

If $(v_k, v_{k+1}) \in \mathrm{im}\,\mathrm{res}_{k+1} \subset V(k) \times V(k+1)/V(k)$, then for some $(w,t) \in \prod_{i=1}^{s} V_{i,k} \times T_{k+1}$,

$$\begin{aligned}
N_{k+1}(v_k, v_{k+1}) &= (L_{k+1} \circ (N_k \times \mathrm{id}) \circ \mathrm{res}_{k+1})(w,t) \\
&= (L_{k+1} \circ (N_k \times \mathrm{id}) \circ\ \mathrm{res}'_{k+1})(t) = 0.
\end{aligned}$$

This proves that $\mathrm{im}\,\mathrm{res}_{k+1} \subset \ker N_{k+1}$. For the opposite inclusion, take $(v_k, v_{k+1}) \in \ker N_{k+1}$. We have that $(N_k(v_k), v_{k+1}) \in \ker L_{k+1} = \mathrm{im}((N_k \times \mathrm{id}) \circ\ \mathrm{res}'_{k+1})$. Hence, for some $(w,t) \in \prod_{i=1}^{s} V_{i,k} \times T_{k+1}$, we have that

$$(N_k(v_k), v_{k+1}) = ((N_k \times \mathrm{id}) \circ \mathrm{res}'_{k+1})(t) = ((N_k \times \mathrm{id}) \circ \mathrm{res}_{k+1})(w,t).$$

This means that there is some element $\mathrm{res}_{k+1}(w,t) = (v'_k, v'_{k+1}) \in \mathrm{im}\,\mathrm{res}_{k+1}$ such that $N_k(v'_k) = N_k(v_k), v'_{k+1} = v_{k+1}$. Since $v_k - v'_k \in \ker N_k = \mathrm{im}\,\mathrm{res}_k$, we can find $w' \in \prod_{i=1}^{s} V_{i,k}$ such that $\mathrm{res}_{k+1}(w + w', t) = (v_k, v_{k+1})$. ☐

This means that if we have computed N_k, then we can compute N_{k+1} by computing the cokernel L_{k+1} of $(N_k \times \mathrm{id}) \circ \mathrm{Res}'_{k+1} = [N_k A_{k+1} \quad B_{k+1}]$ instead of the cokernel of Res_{k+1}. This reduces the computational complexity significantly for $n > 2$. We show some results in Experiment 4.4.1.

Reducing the size of the resultant map

We consider the case $n = s$ of square polynomial systems: $(\hat{f}_1, \ldots, \hat{f}_n) \in \mathcal{F}_R(d_1, \ldots, d_n)$. As explained above, a map N covering a TNF can be computed as the cokernel of the resultant map

$$\mathrm{res} = \mathrm{res}_{\hat{f}_1, \ldots, \hat{f}_n} : V_1 \times \cdots \times V_n \to V$$

from Proposition 4.3.2. We have seen in Section 4.2 that the Macaulay resultant construction gives a subspace of $V_1 \times \cdots \times V_n$ such that if we restrict res to this

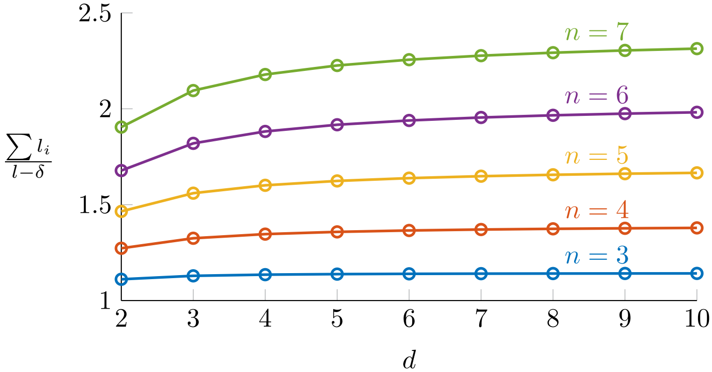


Figure 4.10: The ratio $(l_1 + \ldots + l_n)/(l - \delta)$ of the number of columns of res and res C for increasing values of $n = 3, 4, 5, 6, 7$ and degrees $d = 2, \ldots, 10$, in the context of Example 4.4.1.

subspace, it has the same image. In the case where this is a proper subspace, the matrix of res is column rank deficient. However, in the generic case, we know that the rank of res is $l - \delta$ where $l = \dim_{\mathbb{C}} V$ and $\delta = \dim_{\mathbb{C}} R/\langle \hat{f}_1, \ldots, \hat{f}_n \rangle$ (if some roots have multiplicities, δ should be replaced by δ^+ in our usual notation). This means that taking $l - \delta$ random linear combinations of the columns of res gives a matrix with the same rank and the same cokernel. This comes down to restricting res to a random linear subspace of $V_1 \times \cdots \times V_n$, instead of the very specific one from the Macaulay construction. We may hope that this procedure results in better numerical behaviour. Experiment 4.4.1 will show that it does. Let us denote $l_i = \dim_{\mathbb{C}}(V_i)$. By restricting to a random subspace of the right dimension, we reduce the number of columns of res from $l_1 + \ldots + l_n$ to $l - \delta$. In summary: instead of computing the cokernel of $\text{res} \in \mathbb{C}^{l \times (l_1 + \ldots + l_n)}$, we compute the cokernel of the product $\text{res} \cdot C \in \mathbb{C}^{l \times (l-\delta)}$ where $C \in \mathbb{C}^{(l_1 + \ldots + l_n) \times (l - \delta)}$ is a matrix with random entries (for instance, real and drawn from a normal distribution with zero mean and $\sigma = 1$). We note that this technique can be applied for any resultant map with some 'redundancy' in its domain. In particular, it also works for the methods in Section 5.3.

Example 4.4.1. For the resultant map associated to the family $\mathcal{F}_R(d_1, \ldots, d_n)$ where $d_i = d, i = 1, \ldots, n$ we have

$$l_i = \binom{(n-1)d + 1}{(n-1)(d-1)}, i = 1, \ldots, n, \quad l = \binom{nd + 1}{n(d-1) + 1}, \quad \delta = d^n.$$

The reduction in the number of columns is illustrated in Figure 4.10. △

Experiment 4.4.1 (Fast cokernel computation)**.** This is the experiment in Subsection 6.5 of [MTVB19]. It illustrates the two ways proposed in this subsection for reducing the complexity of the cokernel computation. We generate a generic system of degree d in n variables as in Experiment 4.3.1. Table 4.13 gives the results. In the table we present the computation times t and the maximal residuals r of three different algorithms: TNF stands for the standard TNF algorithm, FM stands for the algorithm that reduces the size of res by multiplying it with a random matrix C of the appropriate size and DBD represents the algorithm which computes the cokernel degree by degree. For all of the algorithms, we used a QR decomposition with optimal column pivoting for the basis selection. For $n = 2$, neither alternative gives any improvements. As

n	d	t_{TNF} (sec)	$t_{\text{TNF}}/t_{\text{FM}}$	$t_{\text{TNF}}/t_{\text{DBD}}$	r_{TNF}	r_{FM}	r_{DBD}
3	2	$1.57 \cdot 10^{-2}$	1.46	0.21	$8.95 \cdot 10^{-16}$	$2.19 \cdot 10^{-15}$	$8.44 \cdot 10^{-16}$
3	3	$4.67 \cdot 10^{-2}$	1.24	0.89	$3.02 \cdot 10^{-15}$	$4.65 \cdot 10^{-14}$	$1.55 \cdot 10^{-15}$
3	4	0.1	1.04	1.35	$1.19 \cdot 10^{-14}$	$2.76 \cdot 10^{-14}$	$8.76 \cdot 10^{-15}$
3	5	0.17	1.06	0.96	$1.43 \cdot 10^{-14}$	$5.14 \cdot 10^{-13}$	$4.92 \cdot 10^{-15}$
3	6	0.41	1.03	0.95	$5.16 \cdot 10^{-15}$	$9.48 \cdot 10^{-14}$	$7.06 \cdot 10^{-15}$
3	7	1.67	1.19	1.47	$8.82 \cdot 10^{-15}$	$1 \cdot 10^{-13}$	$4.05 \cdot 10^{-14}$
3	8	6.23	1.16	2.04	$1.19 \cdot 10^{-13}$	$6.71 \cdot 10^{-11}$	$5.64 \cdot 10^{-14}$
3	9	18.03	1.16	2.61	$2.3 \cdot 10^{-13}$	$6.58 \cdot 10^{-12}$	$2.54 \cdot 10^{-14}$
3	10	45.81	1.16	2.99	$1.56 \cdot 10^{-13}$	$5.67 \cdot 10^{-12}$	$7.08 \cdot 10^{-14}$
3	11	56.36	1.06	1.57	$1.16 \cdot 10^{-13}$	$1.81 \cdot 10^{-12}$	$2.14 \cdot 10^{-13}$
3	12	117.31	1.17	1.55	$1.83 \cdot 10^{-13}$	$3.21 \cdot 10^{-12}$	$8.35 \cdot 10^{-14}$
3	13	229.96	1.16	1.58	$3.16 \cdot 10^{-13}$	$8.87 \cdot 10^{-11}$	$2.03 \cdot 10^{-12}$
4	2	$3.81 \cdot 10^{-2}$	1.39	1.24	$1.36 \cdot 10^{-14}$	$2.35 \cdot 10^{-12}$	$2.74 \cdot 10^{-15}$
4	3	0.28	1.06	1.23	$1.55 \cdot 10^{-13}$	$2.91 \cdot 10^{-13}$	$1.67 \cdot 10^{-14}$
4	4	10.05	1.46	4.42	$5.82 \cdot 10^{-15}$	$1.36 \cdot 10^{-12}$	$1 \cdot 10^{-14}$
4	5	147.32	2.61	5.77	$9.97 \cdot 10^{-14}$	$6.6 \cdot 10^{-13}$	$5.47 \cdot 10^{-14}$
5	2	0.15	1.04	1.12	$3.58 \cdot 10^{-15}$	$9.38 \cdot 10^{-14}$	$1.8 \cdot 10^{-15}$
5	3	75.37	2.78	4.64	$1.97 \cdot 10^{-14}$	$1.83 \cdot 10^{-12}$	$3.49 \cdot 10^{-14}$
6	2	3.44	1.24	1.7	$1.91 \cdot 10^{-15}$	$2.46 \cdot 10^{-13}$	$3.66 \cdot 10^{-15}$
7	2	167.53	1.96	2.41	$1.69 \cdot 10^{-14}$	$4.01 \cdot 10^{-11}$	$3.07 \cdot 10^{-14}$

Table 4.13: Timing and relative error for the variants of the TNF algorithm presented in Subsection 4.4.1 for generic systems in n variables of degree d.

shown earlier, the TNF algorithm is very efficient as it is in this case. For $n > 2$ we see that both FM and DBD can make the algorithm significantly faster for sufficiently high degrees, and not much (or none) of the accuracy is lost. The biggest speed-up we achieved in the experiment is a factor 5.77 for $n = 4, d = 5$. Solving such a system takes about 17 seconds using Bertini and 11 seconds using PHCpack. PHCpack loses 2 out of 625 solutions. The DBD algorithm takes less than 26 seconds to find all solutions with a residual no larger than $\pm 10^{-14}$. The unmodified TNF algorithm takes 3 to 4 times as much time as the homotopy solvers for $n = 4, d = 4$ (see Experiment 4.3.5). The DBD algorithm is as fast as PHCpack, which is 1.6 times faster than Bertini in this case. The algorithms do not beat the homotopy solvers for larger numbers of variables, even in small degrees. For $n = 7, d = 2$, both homotopy packages solve the problem in less than 4 seconds, while the fastest version of the TNF solver takes more than a minute.

To compare the FM algorithm with the Macaulay resultant construction where the V_i are replaced by the span of a specific subset of monomials (see Subsection 3.4.2), we

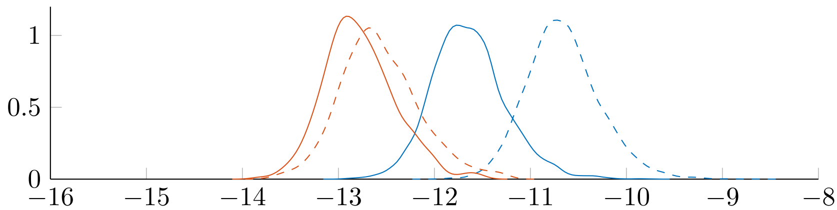


Figure 4.11: Distribution of the computed residuals for $n = 3, d = 23$ (blue) and $n = 5, d = 4$ (orange) using the standard TNF algorithm (solid line) and the DBD algorithm (dashed line).

used this construction to solve the case $n = 3, d = 13$ by computing a TNF from the corresponding resultant map. The obtained residual was $1.44 \cdot 10^{-4}$, which is roughly a factor 10^7 larger than r_{FM}.

On the machine used in Experiment 4.3.6, we see a speed-up factor $t_{\text{TNF}}/t_{\text{DBD}}$ of roughly 1.8 for the case $n = 3, d = 23$. For $n = 5, d = 4$, this factor is roughly 5.4, which reduces the computation time from about 37 minutes to about 7 minutes. The results are slightly less accurate, but the loss of precision for generic systems is not too bad. The distribution of the residuals is shown in Figure 4.11. $\triangle$

4.4.2 TNFs in non-monomial bases

In this subsection, we deal with different matrix representations of resultant maps and TNFs: we fix different bases for the vector spaces involved. Let $\delta = \dim_{\mathbb{C}} R/I$ for some zero-dimensional ideal $I \subset R$. For $\mathbb{C}^\delta$, we will use the standard basis $\{e_1, \ldots, e_\delta\}$. We denote $\mathcal{V} = \{v_1, \ldots, v_l\} \subset V$ for a basis of V ($l = \dim_{\mathbb{C}}(V)$) and $\mathcal{W} = \{w_1, \ldots, w_m\} \subset W$, $m < l$ for a basis of $W \subset V$, which is the largest subspace of V such that $W^+ \subset V$. Analogously, $\mathcal{B} = \{b_1, \ldots, b_\delta\}$ is a basis for B. For simplicity, we assume $\mathcal{W} \subset \mathcal{V}$. As per usual, to simplify the notation we will make no distinction between a matrix and the abstract linear map it represents.

Suppose we have a map $N : V \to \mathbb{C}^\delta$ which covers a TNF $\mathcal{N}_V : V \to B$ for some $B \subset W \subset V$. In practice, this means that we have a matrix representation of N with respect to a fixed basis $\mathcal{V}$ of V. Since N is usually computed as the cokernel of a resultant map res, using for instance the SVD, the basis $\mathcal{V}$ is usually induced by the basis used for V to represent res. Note that since we are assuming $\mathcal{W} \subset \mathcal{V}$, $N_{|W} : W \to \mathbb{C}^\delta$ is just a $\delta \times m$ submatrix of N consisting of the columns indexed by $\mathcal{W}$. In this case we write $N_{\mathcal{W}} = N_{|W}$. To recover $\mathcal{N}_V$ from N, all that is left to do is compute the matrix $N_{|B} : B \to \mathbb{C}^\delta$ with respect to a fixed basis $\mathcal{B} = \{b_1, \ldots, b_\delta\}$ of $B \subset W$. Then the matrix of $\mathcal{N}_V$ with respect to the bases $\mathcal{V}$ for V and $\mathcal{B}$ for B is $\mathcal{N}_V = (N_{|B})^{-1}N$. Note that if $\mathcal{B} \subset \mathcal{W}$, the matrix $N_{\mathcal{B}} = N_{|B}$ consists of a subset of δ columns of $N_{|W}$. Since $B \subset R$ is identified with R/I in the TNF framework, the set $\mathcal{B}$ of basis elements represents a basis $\mathcal{B} + I = \{b_1 + I, \ldots, b_\delta + I\}$ of R/I. Traditionally, e.g. in resultant and Gröbner basis contexts, but often for border bases as well, the b_i are monomials. In this section, we step away from this and show that it is sometimes natural to use non-monomial bases. The following three scenarios clearly lead to non-monomial bases of R/I.

1. The set $\mathcal{V}$ consists of monomials, but $B \subset W$ is computed using another procedure, such that $\mathcal{B} \not\subset \mathcal{W}$. An example is discussed below, where we use a SVD of $N_\mathcal{W}$ to select $\mathcal{B}$ instead of a QR decomposition.
2. The set $\mathcal{V}$ consists of non-monomial basis elements of V and $\mathcal{B} \subset \mathcal{W} \subset \mathcal{V}$. This happens, for instance, when $\mathcal{B}$ is chosen by performing a QR with optimal column pivoting on the matrix $N_\mathcal{W}$. The column pivoting comes down to a pivoting of the elements in $\mathcal{W}$, and $N_{|B}$ is simply a $\delta \times \delta$ submatrix $N_\mathcal{B}$ of $N_\mathcal{W}$. This situation is discussed below for a specific type of basis functions.
3. It is straightforward to combine these first two scenarios, such that $\mathcal{V}$ does not contain (only) monomials and $\mathcal{B} \not\subset \mathcal{W}$.

TNFs as orthogonal projectors

In the approach described in Section 4.3, the selection of a basis $\mathcal{B}$ happens via a column pivoted QR factorization of $N_{|W}$. We present an alternative basis selection using the singular value decomposition (SVD), which is another important tool from numerical linear algebra (see Section B.2). This provides a basis $\mathcal{B}$, which is not a monomial basis. Let $\mathcal{V} = \{x^a : a \in \mathscr{A}\}$ be a set of monomials corresponding to a finite set $\mathscr{A} \subset \mathbb{N}^n$ of lattice points such that $\mathcal{W} = \{x^{a_1}, \ldots, x^{a_m}\} \subset \mathcal{V}$ is a basis of W. We decompose

$$N_\mathcal{W} = \mathbf{U}\mathbf{S}\mathbf{V}^H$$

with $\cdot^H$ the Hermitian transpose. We split $\mathbf{S}$ and $\mathbf{V}$ into compatibly sized block columns:

$$N_\mathcal{W}\ [\mathbf{V}_1\ \mathbf{V}_2] = \mathbf{U}\ [\mathbf{S}_1\ \ 0]$$

with $\mathbf{S}_1$ diagonal and invertible ($N_{|W}$ is onto). In analogy with the QR case (where we would have the identity $N_\mathcal{W}[\mathbf{P}_1\ \mathbf{P}_2] = \mathbf{QR}$), we take

$$\mathcal{B} = [x^{a_1}\ \ \cdots\ \ x^{a_m}]\ \mathbf{V}_1, \tag{4.4.1}$$

such that $B = \operatorname{span}_\mathbb{C}(\mathcal{B}) \simeq \operatorname{im} \mathbf{V}_1$. Therefore

$$(N_\mathcal{W})_{|B} = N_{|B} = \mathbf{U}\ [\mathbf{S}_1\ \ 0]\ [\mathbf{V}_1\ \mathbf{V}_2]^H\ \mathbf{V}_1 = \mathbf{U}\mathbf{S}_1.$$

This tells us that the singular values of $N_{|B}$ are the singular values of $N_\mathcal{W}$ and $(\mathcal{N}_V)_{|W} = (N_{|B})^{-1} N_\mathcal{W} = \mathbf{V}_1^H$. Since $\ker N_\mathcal{W} = I \cap W \simeq \operatorname{im} \mathbf{V}_2 \subset \mathbb{C}^m$ and $\operatorname{im} \mathbf{V}_1 \perp \operatorname{im} \mathbf{V}_2$ by the properties of the SVD, we see that

$$(I \cap W) \perp B$$

with respect to the standard inner product in $\mathbb{C}^m$ and using coordinates w.r.t. $\mathcal{W}$. Equivalently, with this choice of B, $(\mathcal{N}_V)_{|W} = \mathbf{V}_1^H$ projects W orthogonally onto B. The obtained basis $\mathcal{B}$ is an orthonormal basis for the orthogonal complement B of $I \cap W$ in W. This makes B somehow a unique 'canonical' representation of R/I w.r.t. $\mathcal{W}$. Orthogonality is a favorable property for a projector, because the sensitivity of the image to perturbations of the input is minimal. Also, since $(\mathcal{N}_V)_{|W}(f) \perp (I \cap W), \forall f \in W$, $\|(\mathcal{N}_V)_{|W}(f)\|_2$ is a natural measure for the *distance* of f to the ideal in the basis $\mathcal{W}$, which is induced by the Euclidean distance in $\mathbb{C}^m$. We note that $\mathcal{N}_V$ does not project V orthogonally onto B. In order to have an orthogonal projector $(\mathcal{N}_V)_{|W'} : W' \to B$, one must take V large enough such that $W' \subset W \subset V$. Following this procedure, $\mathcal{B}$ is a non-monomial basis of B (or R/I) consisting of δ polynomials supported in $\mathcal{W}$. The above discussion shows that in some sense, $\mathcal{B} + I$ gives a 'natural' basis for R/I, given the freedom of choice provided by Corollary 4.2.1. Unlike the QR algorithm, there are no heuristics involved. For the root finding problem, we observe that $\mathcal{B}_{\text{SVD}}$ (4.4.1) has the same good numerical properties as $\mathcal{B}_{\text{QR}} = [x^{a_1} \ \cdots \ x^{a_m}] \ \mathbf{P}_1$.

Experiment 4.4.2 (SVD for basis selection)**.** We solve a generic member of $\mathcal{F}_{\mathbb{C}[x,y,z]}(8,8,8)$, constructed as in Experiment 4.3.1, using SVD for the basis selection. The computation time is about 6.17 seconds and the maximal residual of all the 512 solutions is $4.62 \cdot 10^{-14}$. This is comparable with the results for the QR basis selection, see for instance Table 4.13. An illustration of the real part of the surfaces defined by the generic equations is shown in Figure 4.12. △

TNFs from function values

We consider the square case ($n = s$) where $V = R_{\leq \hat{\rho}}$, $W = R_{\leq \hat{\rho}-1}$ with $\hat{\rho} = \sum_{i=1}^n d_i - n + 1$ and $R = \mathbb{C}[x_1, \ldots, x_n]$. The resultant map is $\text{res} : V_1 \times \cdots \times V_n \to V$ where $V_i = R_{\leq \hat{\rho} - d_i}$. Let $\{\phi_\ell(x)\} = \{\phi_\ell(x) \mid \ell \in \mathbb{N}\} \subset \mathbb{C}[x]$ be a family of orthogonal univariate polynomials[3] on an interval of $\mathbb{R}$, satisfying the recurrence relation $\phi_0(x) = 1$, $\phi_1(x) = a_0 x + b_0$ and

$$\phi_{\ell+1}(x) = (a_\ell x + b_\ell)\phi_\ell(x) + w_\ell \phi_{\ell-1}(x)$$

with $b_\ell, w_\ell \in \mathbb{C}$, $a_\ell \in \mathbb{C}^* = \mathbb{C} \setminus \{0\}$ so that $x\phi_\ell = \frac{1}{a_\ell}(\phi_{\ell+1} - b_\ell \phi_\ell - w_\ell \phi_{\ell-1}), \ell \geq 1$. For $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$, we define

$$\phi_\alpha(x) = \phi_\alpha(x_1, \ldots, x_n) = \prod_{i=1}^n \phi_{\alpha_i}(x_i).$$

We easily check that

$$x_i \phi_\alpha = \frac{1}{a_{\alpha_i}}(\phi_{\alpha+e_i} - b_{\alpha_i}\phi_\alpha - w_{\alpha_i}\phi_{\alpha-e_i})$$

[3] This is a family of polynomials which are orthogonal with respect to a scalar product $(f,g)_\mu = \int_{p_0}^{p_1} f(x)g(x)d\mu(x)$ for some positive measure $\mu(x)$ on the real interval $[p_0, p_1] \subset \mathbb{R}$. Such a family always satisfies a three term recurrence by Favard's theorem. See for instance [Sze39, Theorem 3.2.1].

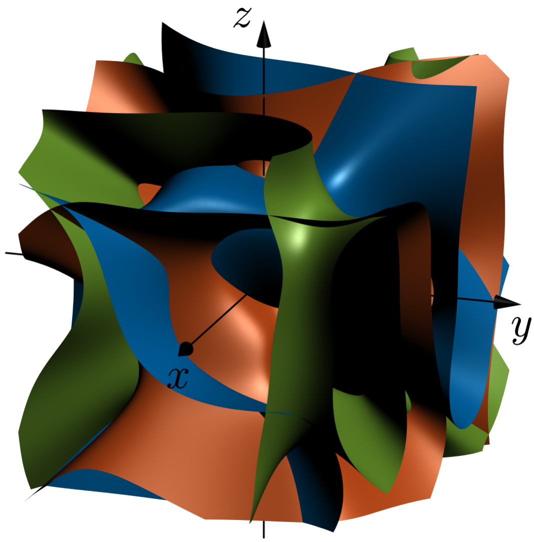


Figure 4.12: Real algebraic surfaces given by $f_i = 0$, $i = 1, \ldots, 3$ from Experiment 4.4.2.

where $e_i \in \mathbb{Z}^n$ is a vector with all zero entries except for a 1 in the i-th position and with the convention that if $\beta \in \mathbb{Z}^n$ has a negative component, $\phi_\beta = 0$. We consider the basis $\mathcal{V} = \{\phi_\alpha : |\alpha| \leq \hat{\rho}\}$ for V. The matrix of res can be constructed such that it has columns indexed by all monomial multiples $x^\alpha f_i$ such that $x^\alpha f_i \in V$ (we use monomial bases for the V_i, although we could use the functions ϕ_α here as well), and rows indexed by the basis $\mathcal{V}$. The corresponding cokernel matrix represents a map $N : V \to \mathbb{C}^\delta$ covering a TNF. The set $\mathcal{W} = \{\phi_\alpha : |\alpha| < \hat{\rho}\} \subset \mathcal{V}$ is a basis for W. The matrix $N_{|W} = N_{\mathcal{W}}$ is again a submatrix of columns indexed by $\mathcal{W}$. To compute a TNF, we have to compute an invertible matrix $N_{|B}$ from $N_{\mathcal{W}}$. If this is done using QR with pivoting, we have $\mathcal{B} = \{\phi_{\beta_1}, \ldots, \phi_{\beta_\delta}\} \subset \mathcal{W}$ and $N_{|B} = N_{\mathcal{B}}$ is the submatrix of $N_{\mathcal{W}}$ with columns indexed by $\mathcal{B}$. Let β_{ji} be the degree in x_i of ϕ_{β_j}. Then the j-th column of $N_i = N_{|x_i \cdot B}$ is given by

$$(N_i)_j = \frac{1}{a_{\beta_{ji}}}(N_{\phi_{\beta_j + e_i}} - b_{\beta_{ji}} N_{\phi_{\beta_j}} - w_{\beta_{ji}} N_{\phi_{\beta_j - e_i}})$$

where N_{ϕ_α} is the column of N corresponding to the basis function ϕ_α with the convention that an exponent α with a negative component gives a zero column. Recall that $M_{x_i} = (N_{|B})^{-1} N_i$ represents the multiplication with x_i in the basis $\mathcal{B} + I$ of R/I. Constructing the matrix res in this way can be done using merely function evaluations of the monomial multiples of the f_i by the properties of the orthogonal family $\{\phi_\ell\}$. This makes it particularly interesting to use bases for which there are fast $(O(d \log d))$

algorithms to convert a vector of function values to a vector of coefficients in the basis $\{\phi_\ell\}$. We now discuss the Chebyshev basis as an important example.

Recall that for the Chebyshev polynomials $\{T_\ell(x)\}$ of the first kind, the recurrence relation is given by $a_0 = 1$, $a_\ell = 2, \ell > 0$, $b_\ell = 0, \ell \geq 0$, $w_\ell = -1, \ell > 0$. We get a basis $\mathcal{B} = \{T_{\beta_1}, \dots, T_{\beta_\delta}\}$. In this basis we obtain

$$N_i = \frac{1}{2}(N_{\mathcal{B}_{+,i}} + N_{\mathcal{B}_{-,i}})$$

with $\mathcal{B}_{+,i} = \{T_{\beta_1+e_i}, \dots, T_{\beta_\delta+e_i}\}$ and $\mathcal{B}_{-,i} = \{T_{\beta_1-e_i}, \dots, T_{\beta_\delta-e_i}\}$ (negative exponents give a zero column by convention). Note that the expression is very simple here since the a_ℓ, b_ℓ, c_ℓ are independent of ℓ. We define

$$\omega_{k,d} = \cos\left(\frac{\pi(k+\frac{1}{2})}{d+1}\right), \quad k = 0, \dots, d.$$

Let $f = \sum_{\ell=0}^{d} c_\ell T_\ell$ be the representation in the Chebyshev basis of a polynomial $f \in \mathbb{C}[x]$ and define $f_k = f(\omega_{k,d})$. By the property of T_ℓ that $T_\ell(x) = \cos(\ell \arccos(x))$ for $x \in [-1, 1]$, we have

$$f_k = \sum_{\ell=0}^{d} c_\ell \cos\left(\frac{\ell\pi(k+\frac{1}{2})}{d+1}\right). \tag{4.4.2}$$

Comparing (4.4.2) to the definition of the (type III) discrete cosine transform (DCT) $(Z_k)_{k=0}^{d}$ of a sequence $(z_k)_{k=0}^{d}$ of $d+1$ complex numbers[4]

$$Z_k = \sqrt{\frac{2}{d+1}}\left(\frac{1}{\sqrt{2}}z_0 + \sum_{\ell=1}^{d} z_\ell \cos\left(\frac{\ell\pi(k+\frac{1}{2})}{d+1}\right)\right),$$

we see that

$$\sqrt{\frac{2}{d+1}}(f_0, f_1, \dots, f_d) = \mathrm{DCT}\left((\sqrt{2}c_0, c_1, \dots, c_d)\right).$$

We conclude that the coefficients c_k in the Chebyshev expansion can be computed from the function evaluations f_k via the inverse discrete cosine transform (IDCT), which is the DCT of type II:

$$z_k = \sqrt{\frac{2}{d+1}}\left(\sum_{\ell=0}^{d} Z_\ell \cos\left(\frac{k\pi(\ell+\frac{1}{2})}{d+1}\right)\right).$$

This gives

$$c_k = \left(\frac{1}{\sqrt{2}}\right)^{q_k}\left(\sqrt{\frac{2}{d+1}}\right)\tilde{c}_k$$

[4]We use the definitions of the discrete cosine transform that agree with the built in `dct` command in Matlab.

with $q_k = 1$ if $k = 0$, $q_k = 0$ otherwise and $(\tilde{c}_0, \dots \tilde{c}_d) = \mathrm{IDCT}((f_0, \dots, f_d))$. Let $T_\alpha = T_{\alpha_1}(x_1) \cdots T_{\alpha_n}(x_n) \in R, \alpha \in \mathbb{N}^n$. For a polynomial $f(x) = f(x_1, \dots, x_n) = \sum_\alpha c_\alpha T_\alpha(x)$ of degree d_i in x_i, this generalizes as follows. We define an n-dimensional array $(f_k)_{k_1=0,\dots,k_n=0}^{d_1,\dots,d_n}$ (this notation means that the index k_i ranges from 0 to d_i) of function values given by

$$f_k = f_{k_1,\dots,k_n} = f(\omega_{k,d}) = f(\omega_{k_1,d_1}, \dots, \omega_{k_n,d_n}).$$

We obtain another such array by performing an n-dimensional IDCT in the usual way: a series of 1-dimensional IDCTs along every dimension of the array. This gives $(\tilde{c}_\alpha)_{\alpha_1=0,\dots,\alpha_n=0}^{d_1,\dots,d_n}$ and the coefficients in the product Chebyshev basis are given by

$$c_\alpha = \left(\frac{1}{\sqrt{2}}\right)^{q_\alpha} \left(\prod_{i=1}^n \sqrt{\frac{2}{d_i+1}}\right) \tilde{c}_\alpha$$

with q_α the number of zero entries in α. This shows that the coefficients c_α needed to construct the matrix of res can be computed efficiently by taking an IDCT of an array of function values of the monomial multiples of the f_i.

A situation in which it is natural to use a product Chebyshev basis $\mathcal{V}$ for V is when $f_i = 0$ are (local) approximations of real transcendental (or higher degree algebraic) hypersurfaces. Chebyshev polynomials have remarkable interpolation and approximation properties on compact intervals of the real line, see [Tre19]. The multivariate product bases $\{T_\alpha\}$ inherit these properties for bounded boxes in $\mathbb{R}^n$. In [NNT15], bivariate, real intersection problems are solved by local Chebyshev approximation, and this is what is implemented in the `roots` command of Chebfun2 [TT13]. If the ideal I is expected to have many real solutions in a compact box of $\mathbb{R}^n$, it is probably a good idea to represent the generators in the Chebyshev basis. One reason is that functions with a lot of real zeros have 'nice coefficients' in this basis, whereas in the monomial basis, they do not.

Experiment 4.4.3 (TNFs in the Chebyshev basis)**.** This experiment comes from Subsection 6.7 in [MTVB19]. It illustrates the use of Chebyshev polynomials in the construction of a TNF. We construct a generic member of $\mathcal{F}_{\mathbb{C}[x_1,x_2]}(d,d)$ as follows. We define $f_1 = \sum_{|\alpha| \le d} c_{1,\alpha} T_\alpha, f_2 = \sum_{|\alpha| \le d} c_{2,\alpha} T_\alpha$ where $T_\alpha = T_{\alpha_1}(x_1) T_{\alpha_2}(x_2)$ and the $c_{i,\alpha}$ are drawn from a standard normal distribution. Since the zeros of T_i are all in the real interval $[-1,1]$, the real plane curves defined by f_1 and f_2 populate the box $[-1,1] \times [-1,1] \subset \mathbb{R}^2$. We expect a large number of real roots in this box. This is the situation in which we expect the Chebyshev basis to have good numerical properties. For $d = 20$, we computed the solutions using a TNF with QR for basis selection in the monomial basis and in the Chebyshev basis. The residuals of all 400 solutions are represented in Figure 4.13 in the form of a histogram. As expected, the Chebyshev TNF performs better. The TNF in the monomial basis still gives acceptable results: the largest residual is of order 10^{-6}. If we increase the degree to $d = 25$, the difference in performance grows. There are 625 solutions in this case. Results are shown in Figure 4.14 and the curves are depicted in Figure 4.15. Using monomials, one solution

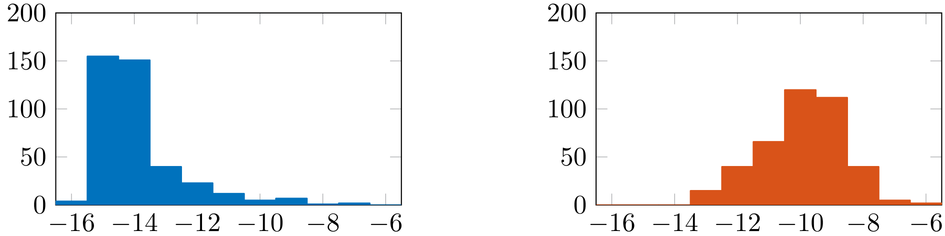


Figure 4.13: Histogram of $\log_{10}$ of the residuals of the computed solutions for a system as described in Experiment 4.4.3 of degree 20 using the Chebyshev basis (left) and the monomial basis (right).

has residual of order 10^{-1}. The quality of this approximate solution is so low that we have basically 'missed' this solution.

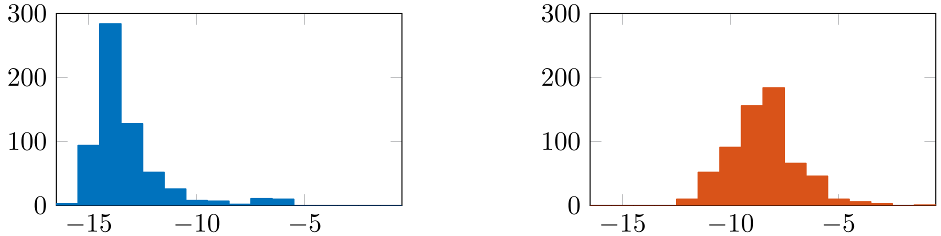


Figure 4.14: Histogram of $\log_{10}$ of the residuals of the computed solutions for a system as described in Experiment 4.4.3 of degree 25 using the Chebyshev basis (left) and the monomial basis (right).

$\triangle$

We conclude this subsection by noting that the monomials $\{x^\ell\}$ are a family of orthogonal polynomials on the complex unit circle and they satisfy the simple recurrence relation $x^{\ell+1} = x \cdot x^\ell$. This is an example of a so-called Szegő recurrence. Coefficients can be computed by taking a fast Fourier transform of equidistant function evaluations on the unit circle. Such a Szegő recurrence exists for all families of orthogonal polynomials on the unit circle and hence products of these bases can also be used in this context [Sze39].

4.5 Homogeneous normal forms

The kind of genericity that we had to assume in order for the methods of Section 4.3 to work is that $\mathrm{Res}_\infty \neq 0$. That is, the homogenized equations do not define any

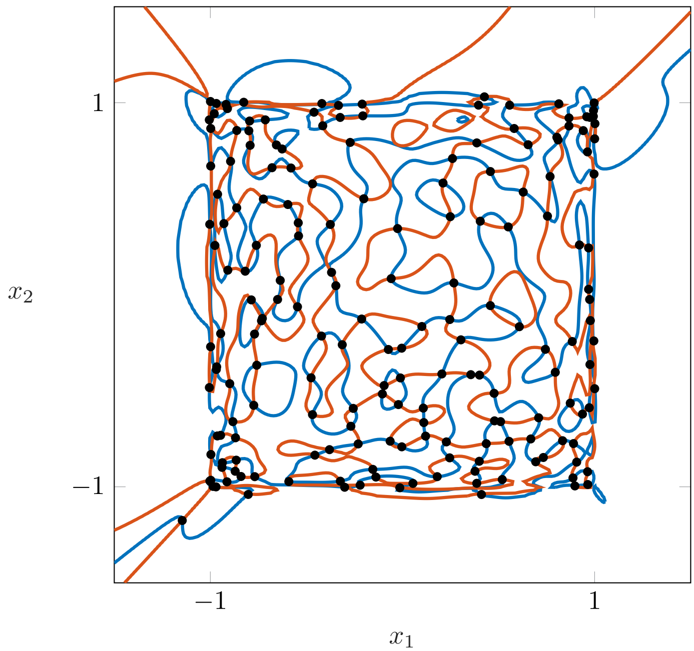


Figure 4.15: Real picture of a degree 25 system as described in Experiment 4.4.3.

solutions outside of $U_0 \subset \mathbb{P}^n$. We mentioned in Remark 4.3.1 that it is possible to weaken this assumption by applying a generic change of coordinates, such that it is enough to assume that the homogeneous ideal is zero-dimensional. However, such a generic change of coordinates may destroy some structure in the equations and it may induce extra rounding errors in the floating point computations. In this section, we introduce an elegant way to find $V_{\mathbb{P}^n}(I)$ for a zero-dimensional, homogeneous ideal $I \subset S = \mathbb{C}[x_0, \ldots, x_n]$, which possibly defines some isolated solutions at infinity. It uses *homogeneous normal forms*, which are the 'projective cousins' of truncated normal forms, as introduced in Section 4.2. In the homogeneous context, normal forms work on graded pieces of the ring S, the ideal I and the algebra S/I. As one would expect from the discussion in Section 3.2, we have to work with degrees that are 'large enough'. Recall that for $d, d_0 \in \mathbb{N}$, a homogeneous element $g \in S_{d_0}$ gives a multiplication map $M_g : (S/I)_d \to (S/I)_{d+d_0}$ given by $M_g(f + I_d) = fg + I_{d+d_0}$.

Definition 4.5.1 (Homogeneous normal form (HNF))**.** Let $I \subset S$ be a zero-dimensional homogeneous ideal such that $V_{\mathbb{P}^n}(I)$ consists of δ^+ points, counting multiplicities. Let $d, d_0 \in \mathbb{N}$ be such that $d, d + d_0 \in \operatorname{Reg}(I)$ and let $B \subset S_d$ be a $\mathbb{C}$-vector subspace. A *homogeneous normal form (HNF)* of degree $d + d_0$ w.r.t. I is a $\mathbb{C}$-linear map $\mathcal{N}_{d,d_0} : S_{d+d_0} \to B$ such that

$$0 \longrightarrow I_{d+d_0} \longrightarrow S_{d+d_0} \xrightarrow{\mathcal{N}_{d,d_0}} B \longrightarrow 0$$

is a short exact sequence and for some $h_0 \in S_{d_0}$ satisfying $V_{\mathbb{P}^n}(I) \cap V_{\mathbb{P}^n}(h_0) = \varnothing$,

$$\begin{tikzcd} B \arrow[r] & (S/I)_d \\ (S/I)_{d+d_0} \arrow[u, "\overline{\mathcal{N}}"] & (S/I)_d \arrow[l, "M_{h_0}"] \arrow[u, "\mathrm{id}"'] \end{tikzcd} \tag{4.5.1}$$

commutes, where $B \to (S/I)_d$ is given by $b \mapsto b + I_d$ and $\overline{\mathcal{N}}(f + I_{d+d_0}) = \mathcal{N}_{d,d_0}(f)$.

Remark 4.5.1. If $d_0 = 1$ and $h_0 = x_0$ (this implies that there are no roots at infinity), a TNF is recovered from a HNF by 'dehomogenizing' the vector spaces and maps that are involved. The commuting diagram (4.5.1) is the homogeneous variant of the condition '$(\mathcal{N}_V)_{|B} = \mathrm{id}_B$' on TNFs. △

Note that a HNF $\mathcal{N}_{d,d_0}$ always comes with a homogeneous polynomial $h_0 \in S_{d_0}$. We do not include h_0 in the notation $\mathcal{N}_{d,d_0}$ to keep the notation simple. Where it is important to specify what h_0 is, we will say that $\mathcal{N}_{d,d_0}$ is a HNF with respect to I *and* h_0. Intuitively, one can think of a HNF as a map that rewrites elements of S_{d+d_0} modulo the ideal and *divides* by h_0. We have seen in Lemma 3.2.1 that if all points in $V_{\mathbb{P}^n}(I)$ have multiplicity one and $V_{\mathbb{P}^n}(I) \cap V_{\mathbb{P}^n}(h_0) = \varnothing$, then M_{h_0} is an isomorphism. We will see later (Corollary 5.5.3) that this holds for higher multiplicities as well. Since $\overline{\mathcal{N}}$ is an isomorphism by definition, we have by (4.5.1) that $B \to (S/I)_d$ is an isomorphism. We conclude that by definition, a HNF identifies B with $(S/I)_d$ as a $\mathbb{C}$-vector space. Just like TNFs allow to compute affine multiplication operators as endomorphisms of B, HNFs can be used to find matrix representations of homogeneous multiplication maps. For a HNF $\mathcal{N}_{d,d_0}$ and a homogeneous polynomial $g \in S_{d_0}$, define $\mathcal{N}_g : S_d \to B$ by $\mathcal{N}_g(f) = \mathcal{N}_{d,d_0}(fg)$.

Proposition 4.5.1. *Let I, d, d_0, B be as in Definition 4.5.1. If $\mathcal{N}_{d,d_0}$ is a HNF with respect to I and $h_0 \in S_{d_0}$, then for any $g \in S_{d_0}$, $(\mathcal{N}_g)_{|B} : B \to B$ is similar to the map $M_{g/h_0} = M_{h_0}^{-1} \circ M_g$ from Theorem 3.2.4.*

Proof. We need to show that for some isomorphism $\nu : B \to (S/I)_d$, we have $(\mathcal{N}_g)_{|B} = \nu^{-1} \circ M_{h_0}^{-1} \circ M_g \circ \nu$. This follows directly from the commutative diagram

$$\begin{tikzcd} B \arrow[r, "\mathcal{N}_g"] \arrow[d, "\nu"] & B \arrow[r, "\nu"] & (S/I)_d \\ (S/I)_d \arrow[r, "M_g"] & (S/I)_{d+d_0} \arrow[u, "\overline{\mathcal{N}}"] \arrow[r, "M_{h_0}^{-1}"] & (S/I)_d \arrow[u, "\mathrm{id}"'] \end{tikzcd}$$

where $\nu(b) = b + I_d$ and the rectangle of isomorphisms on the right is exactly (4.5.1). □

As in the affine case, a first step in computing a HNF often consists of computing a map that is almost a HNF, but not quite.

Definition 4.5.2. Let I, d, d_0 be as in Definition 4.5.1. A $\mathbb{C}$-linear map $N : S_{d+d_0} \to \mathbb{C}^{\delta^+}$ *covers* a HNF $\mathcal{N}_{d,d_0} : S_{d+d_0} \to B$ with respect to I if there is an isomorphism $P : B \to \mathbb{C}^{\delta^+}$ such that $\mathcal{N}_{d,d_0} = P^{-1} \circ N$.

Proposition 4.5.2. *Let $I \subset S$ be a zero-dimensional homogeneous ideal such that $V_{\mathbb{P}^n}(I)$ consists of δ^+ points, counting multiplicities. Let $d, d_0 \in \mathbb{N}$ be such that $d, d + d_0 \in \mathrm{Reg}(I)$. A $\mathbb{C}$-linear map $N : S_{d+d_0} \to \mathbb{C}^{\delta^+}$ covers a HNF if and only if*

$$0 \longrightarrow I_{d+d_0} \longrightarrow S_{d+d_0} \xrightarrow{N} \mathbb{C}^{\delta^+} \longrightarrow 0 \tag{4.5.2}$$

is a short exact sequence.

Proof. If $\mathcal{N}_{d,d_0} = P^{-1} \circ N$ is a HNF for some isomorphism $P : B \to \mathbb{C}^{\delta^+}$, it is clear that (4.5.2) is exact. For the other implication, take $h_0 \in S_{d_0}$ such that $V_{\mathbb{P}^n}(I) \cap V_{\mathbb{P}^n}(h_0) = \varnothing$. We define the map

$$N_{h_0} : S_d \to \mathbb{C}^{\delta^+} \quad \text{by} \quad N_{h_0}(f) = N(f h_0). \tag{4.5.3}$$

Making use of the fact that $M_{h_0} : (S/I)_d \to (S/I)_{d+d_0}$ is an isomorphism, we find that $\ker N_{h_0} = I_d$ and that N_{h_0} is surjective, since every element of S_{d+d_0} can be written as $h_0 f$ modulo I_{d+d_0}. Therefore, we can find a subspace $B \subset S_d$ such that $(N_{h_0})_{|B}$ is invertible. For any such subspace, we set $\mathcal{N}_{d,d_0} = (N_{h_0})_{|B}^{-1} \circ N$. It is clear that

$$0 \longrightarrow I_{d+d_0} \longrightarrow S_{d+d_0} \xrightarrow{\mathcal{N}_{d,d_0}} B \longrightarrow 0$$

is exact. To show that

$$\begin{tikzcd}
B \arrow[r] & (S/I)_d \\
(S/I)_{d+d_0} \arrow[u, "\overline{\mathcal{N}}"] & (S/I)_d \arrow[l, "M_{h_0}"] \arrow[u, "\mathrm{id}"']
\end{tikzcd}$$

commutes, note that if $(\overline{\mathcal{N}} \circ M_{h_0})(f + I_d) = b \in B$, then $((N_{h_0})_{|B}^{-1} \circ N)(h_0 f) = b$, which means that $N_{h_0}(b) = N_{h_0}(f)$ and thus $f - b \in I_d$. We conclude that $f + I_d = b + I_d$. $\square$

The following is an immediate corollary of the proof of Proposition 4.5.2.

Corollary 4.5.1. *In the situation of Proposition 4.5.2, N covers a HNF $\mathcal{N}_{d,d_0} : S_{d+d_0} \to B$ with respect to I and h_0 for any $h_0 \in S_{d_0}$ such that $V_{\mathbb{P}^n}(I) \cap V_{\mathbb{P}^n}(h_0) = \varnothing$. Moreover, for any δ^+-dimensional subspace $B \subset S_d$ such that*

$$(N_{h_0})_{|B} : B \to \mathbb{C}^{\delta^+} \quad \textit{given by} \quad b \mapsto N(b h_0)$$

is invertible, $\mathcal{N}_{d,d_0} = (N_{h_0})_{|B}^{-1} \circ N$ is a HNF.

It follows from Proposition 4.5.1 and Corollary 4.5.1 that if we have computed a $\mathbb{C}$-linear map $N : S_{d+d_0} \to \mathbb{C}^{\delta^+}$ satisfying (4.5.2) for $d, d+d_0 \in \mathrm{Reg}(I)$, then for any $h_0 \in S_{d_0}$ which doesn't vanish at any of the roots of I and any δ^+-dimensional subspace B such that $(N_{h_0})_{|B}$ is invertible, we have that for any $g \in S_{d_0}$, 'multiplication with g/h_0' is given by

$$M_{g/h_0} = (N_{h_0})_{|B}^{-1} \circ (N_g)_{|B}$$

where $N_g : S_d \to \mathbb{C}^{\delta^+}$ is given by $N_g(f) = N(fg)$.

All of the statements above assumed that $d, d+d_0 \in \mathrm{Reg}(I)$. It turns out that if for some $d, d_0 \in \mathbb{N}$ we can find a map $N : S_{d+d_0} \to \mathbb{C}^{\delta^+}$ with the properties of N in Proposition 4.5.2, we can guarantee that $d, d+d_0 \in \mathrm{Reg}(I)$ if I is $\mathfrak{B}$-saturated.

Proposition 4.5.3. *Let $I \subset S$ be a zero-dimensional homogeneous ideal such that $I = (I : \mathfrak{B}^\infty)$ and such that $V_{\mathbb{P}^n}(I)$ consists of δ^+ points, counting multiplicities. If for $h_0 \in S_{d_0}$, the map $N : S_{d+d_0} \to \mathbb{C}^{\delta^+}$ is such that* (4.5.2) *is exact and N_{h_0} as defined in* (4.5.3) *is surjective, then $d, d+d_0 \in \mathrm{Reg}(I)$ and N covers a HNF with respect to I.*

Proof. The fact that $d+d_0 \in \mathrm{Reg}(I)$ follows from (4.5.2) and $I = (I : \mathfrak{B}^\infty)$. To show that $d \in \mathrm{Reg}(I)$, note that $I_d \subset \ker N_{h_0}$ and since I is $\mathfrak{B}$-saturated, $\mathrm{HF}_I(d) \leq \delta^+$ by Theorem 3.2.1. Therefore $\dim_{\mathbb{C}} S_d - \dim_{\mathbb{C}} I_d \leq \delta^+ = \dim_{\mathbb{C}} S_d - \dim_{\mathbb{C}} \ker N_{h_0}$, which implies $\dim_{\mathbb{C}} I_d \geq \dim_{\mathbb{C}} \ker N_{h_0}$. We conclude that $\ker N_{h_0} = I_d$ and $\mathrm{HF}_I(d) = \delta^+$. The fact that N covers a HNF follows from $d, d+d_0 \in \mathrm{Reg}(I)$ and Corollary 4.5.1. □

The following example shows what might go wrong if I is not saturated.

Example 4.5.1. Let $S = \mathbb{C}[x,y]$ and $I = \langle x^2, xy \rangle \subset S$. This is an ideal we considered earlier in Example 3.2.1. It is zero-dimensional and defines $\delta^+ = 1$ point with multiplicity one. Consider the $\mathbb{C}$-linear map $N : S_2 \to \mathbb{C}$ given by $N(x^2) = N(xy) = 0$ and $N(y^2) = 1$. We have that $\ker N = I_2$. Let $h_0 = y$, such that $N_{h_0}(x) = 0, N_{h_0}(y) = 1$. Note that N_{h_0} is onto $\mathbb{C}$. In this example $d = d_0 = 1$, and $d + d_0 \in \mathrm{Reg}(I)$ but d is not. △

Remark 4.5.2. The existence of a map as in Proposition 4.5.3 for generic $h_0 \in S_{d_0}$ with $d_0 = 1$ can be used to detect that the ideal I is '$(d+1)$-regular' in the (more commonly used) sense of *Castelnuovo-Mumford* regularity [Eis13, Chapter 20]. The criterion is strongly related to Theorem 1.10 in [BS87]. It implies, for instance, that $d+1, d+2, \ldots \in \mathrm{Reg}(I)$, which agrees with the observation in Example 4.5.1. A full discussion would take us too far off course. The reader is referred to Proposition 5.2 in [TMVB18] for details. △

In analogy with the affine case, our strategy to compute a map $N : S_{d+d_0} \to \mathbb{C}^{\delta^+}$ that covers a HNF is to compute a cokernel map of a resultant map whose image is I_{d+d_0}. In the homogeneous case, the construction of such a resultant map is trivial. If $I = \langle f_1, \ldots, f_s \rangle$, then I_d is the image of

$$\mathrm{res}_{f_1,\ldots,f_s} : \Lambda_1 \times \cdots \times \Lambda_s \to \Lambda$$

where $\Lambda_i = S_{d-d_i}, i = 1, \dots, s$, $\Lambda = S_d$. A case that is of special interest to us is the square case, where $s = n$. In this case, we know what $\mathrm{Reg}(I)$ is (Theorem 3.2.3). We set

$$\mathrm{res}_{f_1,\dots,f_n} : \Lambda_1 \times \cdots \times \Lambda_n \to \Lambda \tag{4.5.4}$$

where $\Lambda_i = S_{\hat{\rho}-d_i}, i = 1, \dots, n$, $\Lambda = S_{\hat{\rho}}$ with $\hat{\rho} = d_1 + \cdots + d_n - n + 1$. A cokernel map $N : S_{\hat{\rho}} \to \mathbb{C}^{\delta^+}$ of $\mathrm{res}_{f_1,\dots,f_n}$ satisfies the conditions of Proposition 4.5.2 with $d = \rho = \hat{\rho} - 1$ and $d_0 = 1$. This leads directly to Algorithm 4.2 for computing the homogeneous multiplication operators $M_{x_0/h_0}, \dots, M_{x_n/h_0}$.

Algorithm 4.2 Computes homogeneous multiplication matrices for $(f_1, \dots, f_n) \in \mathcal{F}_S(d_1, \dots, d_n)$ such that $I = \langle f_1, \dots, f_n \rangle \subset S$ is zero-dimensional

1: **procedure** HomogeneousMultiplicationMatrices($f_1, \dots, f_n$)
2: $\quad \hat{\rho} = d_1 + \cdots + d_n - n + 1$
3: $\quad \mathrm{res}_{f_1,\dots,f_n} \leftarrow$ the resultant map $\Lambda_1 \times \cdots \times \Lambda_n \to \Lambda$ from (4.5.4)
4: $\quad N \leftarrow \operatorname{coker} \mathrm{res}_{f_1,\dots,f_n}$
5: $\quad N_{h_0} \leftarrow$ matrix of the map $S_\rho \to \mathbb{C}^{\delta^+}$ where $f \mapsto N(fh_0)$
6: $\quad (N_{h_0})_{|B} \leftarrow$ invertible restriction of N_{h_0} to $B \subset S_\rho$, $\dim_{\mathbb{C}} B = \delta^+$
7: $\quad$ **for** $i = 0, \dots, n$ **do**
8: $\quad\quad (N_{x_i})_{|B} \leftarrow$ restriction of the map $S_\rho \to \mathbb{C}^{\delta^+}$ given by $f \mapsto N(x_i f)$ to B
9: $\quad\quad M_{x_i/h_0} \leftarrow (N_{h_0})_{|B}^{-1}(N_{x_i})_{|B}$
10: $\quad$ **end for**
11: $\quad$ **return** $M_{x_0/h_0}, \dots, M_{x_n/h_0}$
12: **end procedure**

In line 6, one should choose a subspace B that results in a well conditioned matrix for $(N_{h_0})_{|B}$. As in the affine case, QR with column pivoting or SVD are good options. In line 8, the same basis of B should of course be used for the product $(N_{h_0})_{|B}^{-1}(N_{x_i})_{|B}$ to make sense. The simultaneous diagonalization of the resulting matrices can happen in the same way as in the affine case. The following example shows how the use of Algorithm 4.2 instead of Algorithm 4.1 can be advantageous in the case of nearly degenerate (i.e. non-generic with respect to $\mathrm{Res}_\infty \neq 0$) systems.

Experiment 4.5.1. Let $R = \mathbb{C}[y_1, y_2, y_3]$ and $S = \mathbb{C}[x_0, x_1, x_2, x_3]$. In this experiment, we consider systems in $\mathcal{F}_R(5,5,5)$ and solve them using Algorithm 4.1 and, after homogenizing them to $\mathcal{F}_S(5,5,5)$, using Algorithm 4.2. The homogeneous solutions are dehomogenized for comparing the residuals. We generate the systems in the following way. First, we generate a generic member by assigning real coefficients drawn from a standard normal distribution to each monomial of degree at most 5. The result is $(\hat{f}_1, \hat{f}_2, \hat{f}_3) \in \mathcal{F}_R(5,5,5)$. Denote $\hat{f}_i = \sum_{|a| \leq 5} c_{i,a} y^a = \sum_{|a| \leq 4} c_{i,a} y^a + \hat{f}_{i,\infty}$. Let r_1, r_2 be fixed real numbers drawn from a standard normal distribution and let e be a real parameter. We define

$$\hat{f}_3(e) = \hat{f}_3 + r_1 \hat{f}_{1,\infty} + r_2 \hat{f}_{2,\infty} + (10^{-e} - 1)\hat{f}_{3,\infty}.$$

Note that as $e \to \infty$, the homogenized polynomials $f_i = \eta_5(\hat{f}_i)$ satisfy

$$f_3(0, x_1, x_2, x_3) = r_1 f_1(0, x_1, x_2, x_3) + r_2 f_2(0, x_1, x_2, x_3).$$

Therefore, the 25 solutions of $f_1(0, x_1, x_2, x_3) = f_2(0, x_1, x_2, x_3) = 0$ in $\mathbb{P}^2$ are solutions at infinity for $\hat{f}_1 = \hat{f}_2 = \hat{f}_3(e) = 0$ when $e \to \infty$. As the value of e grows from 0 to ∞, the system *degenerates*: 25 out of the 125 solutions in $\mathbb{C}^3$ move away towards infinity. We solve the systems $\hat{f}_1 = \hat{f}_2 = \hat{f}_3(e) = 0$ for $e = 0, 1, \ldots, 16$ using Algorithms 4.1 and 4.2 for the computation of the multiplication matrices. The maximal, minimal and geometric mean residuals are shown in Figure 4.16. The figure shows that as the

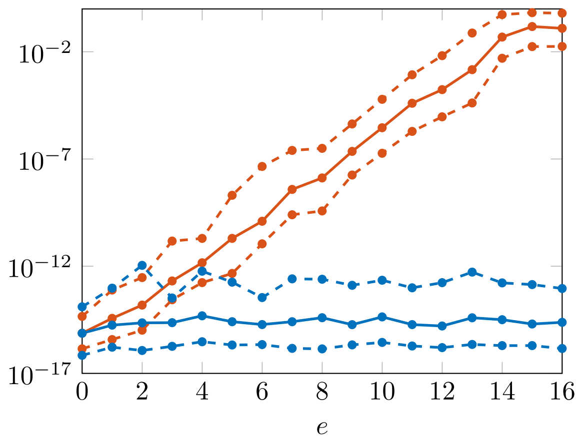


Figure 4.16: Maximal, minimal and geometric mean residual for the solutions computed using Algorithm 4.1 (orange) and Algorithm 4.2 (blue) for the parametrized system defined in Experiment 4.5.1.

system degenerates, the accuracy of the affine TNF solver gets worse and worse. This is due to the fact that *even the best* choice of subspace $B \subset W$ gives a large condition number for $N_{|B}$. Note that even though only 25 solutions move away to infinity, the accuracy is lost on *all* solutions (this can be seen from the minimal residual). Using Algorithm 4.2 corresponds to randomizing the affine patch in which we compute homogeneous coordinates. In such a random patch, the coordinates remain nice and there is no loss of precision at all. We note that the matrices of the resultant maps are *exactly the same*, at least when constructed in the compatible bases, and the complexity of both algorithms is roughly the same. $\triangle$

Chapter 5

Toric methods

In Chapters 3 and 4 we focussed on the families $\mathcal{F}_R(d_1, \ldots, d_n)$ and $\mathcal{F}_S(d_1, \ldots, d_n)$ and we proposed algebraic methods for solving generic members of these families. Here being 'generic' would mean 'defining the expected number of points in $\mathbb{C}^n$' or 'defining finitely many points in projective space'. Although any square polynomial system can be considered as a member of some $\mathcal{F}_R(d_1, \ldots, d_n)$, the systems encountered in applications often do not behave like a general member. For instance, there are often much less than $d_1 \cdots d_n$ solutions in $\mathbb{C}^n$. The reason is that the equations have some extra structure which cannot be detected from just looking at their degrees. In order to handle such systems correctly, they should be considered as members of some *smaller subfamily* of $\mathcal{F}_R(d_1, \ldots, d_n)$, which takes their special structure into account. The goal of this chapter is to propose methods for solving systems coming from a special type of such subfamilies, called *polyhedral families*. The families of type $\mathcal{F}_R(d_1, \ldots, d_n)$ (and hence also the isomorphic families $\mathcal{F}_S(d_1, \ldots, d_n)$) can be seen as polyhedral families, which means that we are in a more general setting. As we will see, the natural solution spaces for these families are *toric varieties*, of which $\mathbb{C}^n$ and $\mathbb{P}^n$ are examples. Taking the polyhedral structure into account may lead to *much* smaller matrices involved in the algorithms, such that it reduces the computational complexity significantly. The proposed methods are based on TNFs in the affine case, following Section 4 of [TMVB18]. In the 'homogeneous' setting, we use a generalization of HNFs to more general compact toric varieties X, working with homogeneous equations in the *Cox ring* of X. This approach is described in [Tel20].

The chapter is organized as follows. In Section 5.1 we describe polyhedral families and state a generalization of Bézout's theorem for square polyhedral families which counts the number of solutions for generic members. Section 5.2 discusses toric resultants and a Macaulay-like matrix construction from which multiplication matrices can be computed for polyhedral families. This is exploited in Section 5.3 to design a TNF algorithm for solving general members. Section 5.4 motivates the use of toric varieties as a natural solution space for polyhedral families. Finally, Section 5.5 describes the

Cox ring of a toric variety X and an algorithm for computing homogeneous coordinates using a toric version of HNFs. The material of this chapter is supported by a summary of some basic facts from polyhedral and toric geometry in Appendices D and E.

5.1 Polyhedral families and the BKK theorem

One of the reasons why Bézout's theorem is such a powerful result is that it gives us a way of bounding the number of solutions of a square polynomial system knowing *only the degree* of the equations. If the highest degree of all monomials appearing in f_i is d_i, the theorem guarantees that $f_1 = \cdots = f_n = 0$ has no more than $d_1 \cdots d_n$ isolated solutions in $\mathbb{C}^n$. However, often this bound is very pessimistic.

Example 5.1.1. The classical eigenvalue problem (see Section B.4) can be interpreted as a polynomial system given by

$$Ax = \lambda x, \quad c^\top x = 1,$$

where $A \in \mathbb{C}^{n \times n}$, the variables are $x_1, \ldots, x_n, \lambda$ and $c \in \mathbb{C}^n \setminus \{0\}$ is a vector used to normalize the eigenvectors. We know that for generic A, c there are n solutions to this system. However, the Bézout bound for $\mathcal{F}_{\mathbb{C}[x_1,\ldots,x_n,\lambda]}(2, \ldots, 2, 1)$ is 2^n. This gives a sequence of examples for which the asymptotic ratio $(n \to \infty)$ between the Bézout bound and the actual number of solutions is infinite. $\triangle$

One of the goals in this chapter is to sharpen Bézout's root count. In order to do so we will work with slightly more general objects than polynomials: we allow negative entries in the exponent vectors. This means we will be working in the ring

$$\mathbb{C}[M] = \mathbb{C}[t_1, t_1^{-1}, \ldots, t_n, t_n^{-1}] = \mathbb{C}[t_1^{\pm 1}, \ldots, t_n^{\pm 1}]$$

of *Laurent polynomials*. Here we let $M = \mathbb{Z}^n$ and the notation $\mathbb{C}[M]$ emphasizes that our Laurent polynomial ring is the semigroup algebra over M (see Definition E.1.3). An element $\hat{f} \in \mathbb{C}[M]$ can be written as

$$\hat{f} = \sum_{m \in M} c_m t^m, \tag{5.1.1}$$

where finitely many coefficients c_m are nonzero. Note that $R = \mathbb{C}[t_1, \ldots, t_n] \subset \mathbb{C}[M]$. Here are two motivations for working in the larger ring $\mathbb{C}[M]$.

1. As the title of this section suggests, we would like to associate polyhedral objects to polynomials and vice versa. More precisely, the exponents m in (5.1.1) will correspond to points in a *lattice polytope* in $M_\mathbb{R} = \mathbb{R}^n$ (see Section D.1). In this construction we would like to allow all lattice polytopes, not only those in the positive orthant.

2. In some sections of the previous chapters, we treated points 'at infinity' (i.e. points in $\mathbb{P}^n \setminus U_0$) as *special* points. The reason is that these are the points that lie outside of the affine chart U_0 with which we identified our original solution space $\mathbb{C}^n$. However, $\mathbb{P}^n$ is covered by $n+1$ open, dense affine charts U_i, whose complements have the easy description $x_i = 0$. In a sense, points outside of U_i are just as *special* as points outside of U_0. This corresponds to the intuition that if a blindfolded person were to 'throw a dart at $\mathbb{P}^n$', it would land on the intersection $U_0 \cap \cdots \cap U_n$ with probability 1. This intersection is exactly $(\mathbb{C}^*)^n$, whose coordinate ring is $\mathbb{C}[M]$ (see Example 2.1.12).

As suggested by point 2, Laurent polynomial systems define relations on the algebraic torus $(\mathbb{C}^*)^n$, which is a first justification for the title of this chapter. An ideal $I \subset \mathbb{C}[M]$ is called *zero-dimensional* if $V_{(\mathbb{C}^*)^n}(I)$ consists of finitely many points. The results in Subsection 3.1.1 generalize to the toric setting, where R should be replaced by $\mathbb{C}[M]$ and $\mathbb{C}^n$ by $(\mathbb{C}^*)^n$. We will now motivate why this is true and include an adapted version of the eigenvalue, eigenvector theorem. First of all, we note that the ring $\mathbb{C}[M]$ can be written as

$$\mathbb{C}[M] = R_{t_1\cdots t_n} = R[y]/\langle t_1 \cdots t_n y - 1\rangle = R[y_1, \ldots, y_n]/\langle t_1 y_1 - 1, \ldots, t_n y_n - 1\rangle$$

where $R_{t_1\cdots t_n}$ is the localization at $t_1 \cdots t_n$. This makes the fact that Laurent polynomial systems are really just polynomial systems explicit.

Example 5.1.2. The equation $t^{-1} + t - 5/2 = 0$ on $(\mathbb{C}^*)$ with solutions $t = 2$ and $t = 1/2$ is equivalent to the system $y + t - 5/2 = ty - 1 = 0$ on $\mathbb{C}^2$ with solutions $(t, y) = (2, 1/2)$ and $(t, y) = (1/2, 2)$. Another way to see the second formulation is by considering $V_{V_{\mathbb{C}^2}(ty-1)}(y + t - 5/2 + \langle ty - 1\rangle)$, where $y + t - 5/2 + \langle ty - 1\rangle$ corresponds to $t^{-1} + t - 5/2$ under the isomorphism $\mathbb{C}[t][y]/\langle ty - 1\rangle = \mathbb{C}[t, t^{-1}]$. △

Let $I = \langle \hat{f}_1, \ldots, \hat{f}_s\rangle \subset \mathbb{C}[M]$ be a zero-dimensional ideal. We may assume that $\hat{f}_i \in R \subset \mathbb{C}[M], i = 1, \ldots, s$. This is because any Laurent monomial t^m is a unit in $\mathbb{C}[M]$, hence multiplying the generators with a monomial does not change the ideal. In what follows we use some terminology given in Definition A.1.16. Thinking of $\mathbb{C}[M]$ as the localization $R_{t_1\cdots t_n} \supset R$, I is the extension I_{aff}^e of the ideal $I_{\text{aff}} = \langle \hat{f}_1, \ldots, \hat{f}_s\rangle \subset R$ (this is simply the ideal generated by the $\hat{f}_i$ in the subring $R \subset \mathbb{C}[M]$) in the localization $\mathbb{C}[M]$.

Lemma 5.1.1. *Let the zero-dimensional ideal $I = I_{\text{aff}}^e \subset \mathbb{C}[M]$ be the extension of $I_{\text{aff}} \subset R$ in the localization $\mathbb{C}[M] = R_{t_1\cdots t_n}$. The contraction $I^c = (I_{\text{aff}}^e)^c \subset R$ satisfies*

$$\begin{aligned} I^c = I \cap R &= (I_{\text{aff}} : (t_1 \cdots t_n)^\infty) \\ &= \{f \in R \mid (t_1 \cdots t_n)^\ell f \in I_{\text{aff}} \text{ for some } \ell \in \mathbb{N}\}. \end{aligned}$$

Moreover, the map

$$R/I^c \to \mathbb{C}[M]/I \quad \text{given by} \quad f + I^c \to f/1 + I \tag{5.1.2}$$

is an isomorphism of $\mathbb{C}$-algebras.

Proof. The first statement follows directly from the definition of contraction and some basic properties of localization, see e.g. [AM69, Proposition 3.11]. We now show that (5.1.2) is an isomorphism. Note that injectivity is clear. Moreover, injectivity of (5.1.2) implies that R/I^c is a finite-dimensional $\mathbb{C}$-vector space. To see this, note that $\dim_{\mathbb{C}} \mathbb{C}[M]/I < \infty$, since it is the coordinate ring of a zero-dimensional affine variety [CLO13, Chapter 5, §3, Theorem 6]. It remains to show that (5.1.2) is also surjective. Note that by the first statement, $(t_1 \cdots t_n)^\ell$ is not a zero divisor in R/I^c for all $\ell \in \mathbb{N}$. Therefore, 'multiplication with $(t_1 \cdots t_n)^\ell$' is injective and hence it is an isomorphism in R/I^c (here we use the fact that $\dim_{\mathbb{C}} R/I^c$ is finite). This means that for any $f/(t_1 \cdots t_n)^\ell + I \in \mathbb{C}[M]/I$, there is $g \in R$ such that $(t_1 \cdots t_n)^\ell g - f \in I^c$. Therefore

$$\frac{f}{(t_1 \cdots t_n)^\ell} - \frac{g}{1} \in I$$

and $f/(t_1 \cdots t_n)^\ell + I$ is the image of $g + I^c$ under (5.1.2). □

Example 5.1.3. For the ideal $I = \langle t^{-1} + t - 5/2 \rangle \subset \mathbb{C}[t, t^{-1}]$ from Example 5.1.2 we have that $\mathbb{C}[t, t^{-1}]/I \simeq \mathbb{C}[t]/\langle 1 + t^2 - 5/2t \rangle$. △

Recall that by Lemma 3.1.2, for any point set $\{z_1, \dots, z_\delta\} \subset \mathbb{C}^n$ there exists a set $\{\ell_1, \dots, \ell_\delta\} \subset R \subset \mathbb{C}[M]$ of Lagrange polynomials.

Theorem 5.1.1. *Let $I = \langle \hat{f}_1, \dots, \hat{f}_s \rangle \subset \mathbb{C}[M]$ be a zero-dimensional ideal such that $V_{(\mathbb{C}^*)^n}(I) = \{z_1, \dots, z_\delta\}$, where z_i has multiplicity μ_i. We have that*

$$\dim_{\mathbb{C}} \mathbb{C}[M]/I = \delta^+ = \mu_1 + \cdots + \mu_\delta$$

and for any $g \in \mathbb{C}[M]$, the $\mathbb{C}$-linear endomorphism $M_g : \mathbb{C}[M]/I \to \mathbb{C}[M]/I$ given by $M_g(f + I) = fg + I$ satisfies

$$\det(\lambda \, \mathrm{id}_{\mathbb{C}^{\delta^+}} - M_g) = \prod_{i=1}^{\delta} (\lambda - g(z_i))^{\mu_i}.$$

If $\delta = \delta^+$, the map M_g has left and right eigenpairs

$$(\mathrm{ev}_{z_i}, g(z_i)), \quad (g(z_i), \ell_i + I), \quad i = 1, \dots, \delta,$$

where $\{\ell_1, \dots, \ell_\delta\}$ is a set of Lagrange polynomials for $\{z_1, \dots, z_\delta\}$ and $\mathrm{ev}_{z_1}, \dots, \mathrm{ev}_{z_\delta}$ is the basis of $\mathbb{C}[M]/I$ dual to $\ell_1 + I, \dots, \ell_\delta + I$.

Proof. All statements follow immediately from applying the results of Subsections 3.1.1 and 3.1.3 and the fact that by Lemma 5.1.1, M_g is the map

$$R/I^c \to R/I^c \quad \text{given by} \quad f + I^c \mapsto fg^c + I^c$$

where $g^c + I^c$ is the inverse image of g under $R/I^c \to \mathbb{C}[M]/I$. □

The fact that finding the points defined by $I \subset \mathbb{C}[M]$ corresponds to finding the points defined by a $\langle t_1 \cdots t_n \rangle$-saturated ideal in R will come in handy in Section 5.3. We now turn back to the root counting problem. The family of polynomial systems in Example 5.1.1 parametrized by A and c is a subfamily of $\mathcal{F}_{\mathbb{C}[x_1,\ldots,x_n,\lambda]}(2,\ldots,2,1)$ with a different generic number of solutions. We could have suspected that these systems don't show the generic behavior of a dense family: not all monomials of degree up to 2 occur in the equations $Ax = \lambda x$. Indeed, the monomials $x_i x_j, 1 \leq i, j \leq n$ are missing. Motivated by this, rather than looking only at the degree, in this chapter we keep track of which monomials are present in our Laurent polynomials and which ones are not.

Definition 5.1.1 (Support)**.** The *support* of a Laurent polynomial $f = \sum_{m \in M} c_m t^m \in \mathbb{C}[M]$ is given by

$$\mathrm{Supp}(f) = \{m \in M \mid c_m \neq 0\}.$$

This allows us to define families of polynomial systems with fixed supports. In the following definition we use a straightforward generalization of Definition 3.1.3 where R is replaced by $\mathbb{C}[M]$.

Definition 5.1.2 (Families with fixed support)**.** Let $\mathscr{A}_i \subset M, i = 1,\ldots,s$ be finite subsets of the lattice M. The *family of (Laurent) polynomial systems supported in* $\mathscr{A}_1,\ldots,\mathscr{A}_s$ is the image of

$$\phi : \mathbb{C}^{|\mathscr{A}_1|} \times \cdots \times \mathbb{C}^{|\mathscr{A}_s|} \to \bigoplus_{m \in \mathscr{A}_1} \mathbb{C} \cdot t^m \times \cdots \times \bigoplus_{m \in \mathscr{A}_s} \mathbb{C} \cdot t^m,$$

where $|\cdot|$ denotes the cardinality and

$$\phi((c_{1,m})_{m \in \mathscr{A}_1}, \ldots, (c_{s,m})_{m \in \mathscr{A}_s}) = \left(\sum_{m \in \mathscr{A}_1} c_{1,m} t^m, \ldots, \sum_{m \in \mathscr{A}_s} c_{s,m} t^m \right).$$

We denote this family by

$$\mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1,\ldots,\mathscr{A}_s) = \{(\hat{f}_1,\ldots,\hat{f}_s) \in \mathbb{C}[M]^s \mid \mathrm{Supp}(\hat{f}_i) \subset \mathscr{A}_i, i = 1,\ldots,s\}.$$

We will focus on the *square case*, i.e. $n = s$. A remarkable fact is that the number of solutions in $(\mathbb{C}^*)^n$ of a generic member of $\mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1,\ldots,\mathscr{A}_n)$ depends only on the convex hull of the lattice point configurations $\mathscr{A}_1,\ldots,\mathscr{A}_n$. We will now make this precise.

Definition 5.1.3 (Newton polytope)**.** For $\hat{f} = \sum_{m \in M} c_m t^m \in \mathbb{C}[M]$ we embed the lattice $M = \mathbb{Z}^n$ in its associated real vector space $\mathbb{R}^n = M_{\mathbb{R}} = M \otimes_{\mathbb{Z}} \mathbb{R}$ and set

$$\mathrm{Newt}(\hat{f}) = \mathrm{Conv}(\{m \mid m \in \mathrm{Supp}(\hat{f}) \subset M_{\mathbb{R}}\}) \subset \mathbb{R}^n.$$

The convex polytope $\mathrm{Newt}(\hat{f})$ is called the *Newton polytope* of $\hat{f}$.

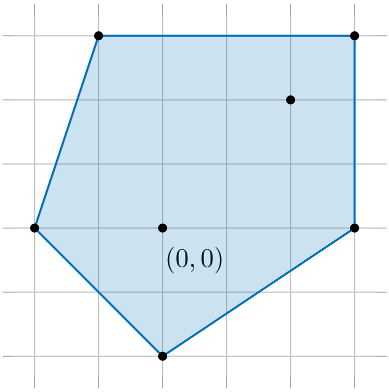


Figure 5.1: Newton polytope $\mathrm{Newt}(\hat{f})$ and support $\mathrm{Supp}(\hat{f})$ (black dots) of the Laurent polynomial $\hat{f}$ in Example 5.1.4.

For the definition of the convex hull and properties of convex polytopes, see Section D.1.

Example 5.1.4. Consider the case where $n = 2$ and

$$\hat{f} = c_0 + c_1 t_1^2 t_2^2 + c_2 t_1^3 + c_3 t_1^3 t_2^3 + c_4 t_1^{-1} t_2^3 + c_5 t_1^{-2} + c_6 t_2^{-2} \quad \in \mathbb{C}[t_1^{\pm 1}, t_2^{\pm 1}].$$

We assume that the coefficients c_i are nonzero. The Newton polytope, together with $\mathrm{Supp}(\hat{f})$, is shown in Figure 5.1. △

The following statement uses the notion of *mixed volume* of a set of polytopes, see Definition D.1.5.

Theorem 5.1.2 (BKK theorem)**.** *For n Laurent polynomials $\hat{f}_1, \cdots, \hat{f}_n \in \mathbb{C}[M]$, the number of isolated points in $V_{(\mathbb{C}^*)^n}(\hat{f}_1, \ldots, \hat{f}_n)$ is bounded by the mixed volume*

$$\mathrm{MV}(\mathrm{Newt}(\hat{f}_1), \ldots, \mathrm{Newt}(\hat{f}_n)).$$

Moreover, for a generic member $(\hat{f}_1, \ldots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \ldots, \mathscr{A}_n)$, the variety $V_{(\mathbb{C}^*)^n}(\hat{f}_1, \ldots, \hat{f}_n)$ consists of exactly $\mathrm{MV}(P_1, \ldots, P_n)$ *points, where*

$$P_i = \mathrm{Conv}(\mathscr{A}_i), \quad i = 1, \ldots, n.$$

Proof. See [Ber75] for the original proof. A proof based on homotopy continuation is given in [HS95], and a sketch of the proof can be found in [CLO06, Chapter 7, §5]. □

Note that for a member $(\hat{f}_1, \ldots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \ldots, \mathscr{A}_n)$, the property $\mathrm{Newt}(\hat{f}_i) = \mathrm{Newt}(\mathscr{A}_i) = P_i, i = 1, \ldots, n$ is a generic property: it only fails to hold if the coefficient of some $\hat{f}_i$ corresponding to a vertex of P_i is zero. Theorem 5.1.2 is sometimes called *Bernstein's theorem* because it was first proved (after experimental observation) by David Bernstein [Ber75]. The result was shown independently in the unmixed case, i.e. the case where $\mathscr{A}_1 = \cdots = \mathscr{A}_n$, by Kushnirenko [Kus76b, Kus76a]. Many different proofs of the theorem and its connections to toric geometry were given by Khovanskii, see e.g. [Kho77, Kho92]. For this reason, the theorem is also referred to as the *BKK theorem* (after Bernstein, Kushnirenko and Khovanskii), and the upper bound on the number of isolated solutions provided by the theorem is often called the *BKK number*.

Theorem 5.1.2 implies that the number of solutions of a generic member of the family $\mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \ldots, \mathscr{A}_n)$ only depends on the polytopes $\mathrm{Conv}(\mathscr{A}_i), i = 1, \ldots, s$. That is, if we only care about the number of solutions, we can consider (in general) larger families defined by a less 'fine grained' structure.

Definition 5.1.4 (Polyhedral families)**.** Let $P_1, \ldots, P_s \subset \mathbb{R}^n = M_{\mathbb{R}}$ be convex lattice polytopes. The *polyhedral family of (Laurent) polynomial systems* given by $P_1, \ldots, P_s$ is the family

$$\mathcal{F}_{\mathbb{C}[M]}(P_1, \ldots, P_s) = \mathcal{F}_{\mathbb{C}[M]}(P_1 \cap M, \ldots, P_s \cap M)$$

of systems supported in $\mathscr{A}_1 = P_1 \cap M, \ldots, \mathscr{A}_s = P_s \cap M$.

Note that if $\mathscr{A}_i \subset P_i \cap M, i = 1, \ldots, s$, then

$$\mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \ldots, \mathscr{A}_s) \subset \mathcal{F}_{\mathbb{C}[M]}(P_1, \ldots, P_s).$$

If $P_i \subset \mathbb{R}^n$ is contained in the positive orthant for all i, then there is some d_i for which $P_i \subset d_i \Delta_n$, where $\Delta_n = \mathrm{Conv}(0, e_1, \ldots, e_n)$ is the standard n-simplex in $\mathbb{R}^n$. For the numbers d_i, we have

$$\mathcal{F}_{\mathbb{C}[M]}(P_1, \ldots, P_s) \subset \mathcal{F}_R(d_1, \ldots, d_n).$$

This explains that when we look at total degree families, polyhedral families and families defined by supports, we are looking at smaller and smaller families with 'more structure'.

In what follows, by the *standard simplex* or *elementary simplex* in $\mathbb{R}^n$ we mean the convex hull of the standard basis vectors $e_1, \ldots, e_n$ and the origin in $\mathbb{R}^n$. We denote this polytope by $\Delta_n = \mathrm{Conv}(\{0, e_1, \ldots, e_n\}) \subset \mathbb{R}^n$.

Remark 5.1.1. Families defined by supports or polytopes are often called *sparse families* in the literature, whereas total degree families are called *dense families*. The reason is that these families take certain 'sparsity' patterns of the equations into account. However, especially for families defined by polytopes we prefer the terminology *polyhedral families*. The reason is that 'sparse' has the connotation of 'having only few terms', and many polytopes (that are not dilates of the standard simplex) have *many* lattice points. △

Remark 5.1.2. If there are natural numbers $d_1, \ldots, d_n \in \mathbb{N}_{\geq 0}$ such that $P_i = d_i \Delta_n$, then $\mathrm{MV}(P_1, \ldots, P_n) = d_1 \cdots d_n$ and the Bézout number agrees with the BKK number. △

Example 5.1.5. Consider the case where $n = 2$ and the square family $\mathcal{F} = \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}, \mathscr{A})$ with

$$\mathscr{A} = \{(0,0),(1,0),(0,1),(1,1)\} \subset \mathbb{Z}^2.$$

Note that $\mathcal{F} = \mathcal{F}_{\mathbb{C}[M]}(P,P)$ where $P \subset \mathbb{R}^2$ is the polytope $[0,1] \times [0,1] \subset \mathbb{R}^2$. This polytope is contained in the positive orthant in $\mathbb{R}^2$, so it makes sense to consider the solutions in $\mathbb{C}^2 \supset (\mathbb{C}^*)^2$. We have $\mathcal{F} \subset \mathcal{F}_R(2,2) = \mathcal{F}_{\mathbb{C}[M]}(2\Delta_2, 2\Delta_2)$, so that the Bézout bound on the number of solutions in $\mathbb{C}^2$ is 4. A member of $\mathcal{F}$ is given by $(\hat{f}_1, \hat{f}_2)$ with

$$\hat{f}_1 = a_0 + a_1 t_1 + a_2 t_2 + a_3 t_1 t_2, \quad \hat{f}_2 = b_0 + b_1 t_1 + b_2 t_2 + b_3 t_1 t_2. \tag{5.1.3}$$

For a generic member, the coefficients a_i, b_i are nonzero. The equations $\hat{f}_1 = \hat{f}_2 = 0$ on $\mathbb{C}^2$ are equivalent to

$$\hat{f}_1 = \hat{f}_2 - \frac{b_3}{a_3}\hat{f}_1 = 0.$$

But $\hat{f}_2 - b_3/a_3 \hat{f}_1$ is a linear equation, which means that after this rewriting step Bézout's theorem tells us that there can be at most 2 solutions in $\mathbb{C}^2$. This agrees with the BKK number. Indeed, applying the formula (D.1.3) for 2-dimensional mixed volume computations, we obtain $\mathrm{MV}(P,P) = 2$. △

In Example 5.1.5, the BKK number actually counts the number of solutions in $\mathbb{C}^2$ instead of $(\mathbb{C}^*)^2$. This is not always the case. To see this, we note that the mixed volume $\mathrm{MV}(P_1, \ldots, P_n)$ is invariant under translations of the polytopes $P_1, \ldots, P_n$ in the lattice [CLO06, Chapter 7, §4, Theorem 4.12]. This geometric observation corresponds to the algebraic fact that if one or more of the Laurent polynomials $\hat{f}_1, \ldots, \hat{f}_n$ are multiplied by a Laurent monomial, the solutions in $(\mathbb{C}^*)^n$ do not change (Laurent monomials are units in $\mathbb{C}[M]$). However, (assuming that the polytopes are contained in the positive orthant) the solutions in $\mathbb{C}^n$ do! We illustrate this briefly with an example and refer to [RW96, HS97, Roj99] for more details.

Example 5.1.6. Consider the support $\mathscr{A}$ from Example 5.1.5 and the family $\mathcal{F} = \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}, e_2 + \mathscr{A})$ where $e_2 + \mathscr{A} = \{m + (0,1) \mid m \in \mathscr{A}\}$. The BKK bound tells us that there are at most 2 solutions in the torus. However, in $\mathbb{C}^2$, there are generically three solutions. To see this, note that a member of $\mathcal{F}$ looks like $(\hat{f}_1, t_2 \hat{f}_2)$ with $\hat{f}_1, \hat{f}_2$ as in (5.1.3). Hence, for a generic member, we get 2 solutions in the torus satisfying $\hat{f}_1 = \hat{f}_2 = 0$ and an additional solution $(-a_0/a_1, 0)$ which is not in the torus. △

5.2 Toric resultants

In Section 3.4, we have seen that projective resultants provide many insights into the behavior of total degree families of polynomial systems. Moreover, they provide several ways of solving the equations. We called them *projective* resultants because they characterize exactly the members of an overdetermined family $\mathcal{F}_S(d_0, \ldots, d_n) \simeq$

$\mathcal{F}_R(d_0, \ldots, d_n)$ which define solutions in $\mathbb{P}^n$. In particular, $\mathrm{Res}_{d_0,\ldots,d_n}$ vanishes at members of $\mathcal{F}_R(d_0, \ldots, d_n)$ which define solutions in $(\mathbb{C}^*)^n$. There is a nice generalization of the projective resultant for the family $\mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_0, \ldots, \mathscr{A}_n)$ supported in $\mathscr{A}_0, \ldots, \mathscr{A}_n \subset M$ and the polyhedral family $\mathcal{F}_{\mathbb{C}[M]}(P_0, \ldots, P_n)$. Just as in our TNF construction for solving generic members of $\mathcal{F}_R(d_1, \ldots, d_n)$, these *toric* or *sparse resultants* will help us construct a TNF algorithm for solving generic members of polyhedral families. We note, without going into the details, that Gröbner and border basis techniques have also been adapted to work in the toric setting, see for instance [PU99] and Section 5 in [Mou99].
In Subsection 5.2.1 we give a definition of the toric resultant and list some of its properties. The interested reader is referred to [PS93, Stu94, GKZ94] for more details. In Subsection 5.2.2 we briefly discuss a construction due to Canny and Emiris [CE93] which is a toric variant of the Macaulay construction discussed in Subsection 3.4.2. A more complete introduction to these concepts can be found in [CLO06, Chapter 7]. Another nice overview with many references is given in [EM99b].

5.2.1 Definition and properties

Let $\mathscr{A}_0, \ldots, \mathscr{A}_n \subset M$ be finite subsets of the lattice M. We will assume for simplicity that the supports $\mathscr{A}_0, \ldots, \mathscr{A}_n$ affinely span the lattice M. That is,

$$M = \left\{ \sum_{m \in \mathscr{A}_0} c_{0,m} m + \cdots + \sum_{m \in \mathscr{A}_n} c_{n,m} m \mid c_{i,m} \in \mathbb{Z} \text{ and } \sum_{m \in \mathscr{A}_i} c_{i,m} = 0, i = 0, \ldots, n \right\}.$$

The family $\mathcal{F} = \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_0, \ldots, \mathscr{A}_n)$ is parametrized by

$$\mathbb{C}^p = \mathbb{C}^{|\mathscr{A}_0|} \times \cdots \times \mathbb{C}^{|\mathscr{A}_n|}.$$

In the case where $\mathscr{A}_i = d_i \Delta_n \cap M$ consists of all lattice points in a dilation of the elementary simplex, $\mathcal{F} = \mathcal{F}_R(d_0, \ldots, d_n)$. In this case $\mathrm{Res}_{d_0,\ldots,d_n}$ is a polynomial in the coordinate ring $A = \mathbb{C}[\mathbb{C}^p] = \mathbb{C}[\mathcal{F}]$ of the family which characterizes whether a member of $\mathcal{F}$ has a solution. Recall that the variables of A are the coefficients $c_{i,m}$ for $i = 0, \ldots, n, m \in \mathscr{A}_i$. For general $\mathscr{A}_0, \ldots, \mathscr{A}_n$, we would like to define a toric resultant $\mathrm{Res}_{\mathscr{A}_0,\ldots,\mathscr{A}_n} \in A$ which also has this property. Ideally, with the special choices of $\mathscr{A}_0, \ldots, \mathscr{A}_n$ above, we would like $\mathrm{Res}_{d_0,\ldots,d_n}$ and $\mathrm{Res}_{\mathscr{A}_0,\ldots,\mathscr{A}_n}$ to coincide. We let $Z_0(\mathscr{A}_0, \ldots, \mathscr{A}_n) \subset \mathbb{C}^p$ denote the set of members of $\mathcal{F}$ which have a solution in $(\mathbb{C}^*)^n$. The Zariski closure of this set is denoted by $Z(\mathscr{A}_0, \ldots, \mathscr{A}_n) = \overline{Z_0(\mathscr{A}_0, \ldots, \mathscr{A}_n)}$.

Theorem 5.2.1. *The variety $Z(\mathscr{A}_0, \ldots, \mathscr{A}_n) \subset \mathbb{C}^p \simeq \mathcal{F}$ is a proper, irreducible subvariety whose ideal $I_A(Z(\mathscr{A}_0, \ldots, \mathscr{A}_n))$ is generated by polynomials in A with coefficients in $\mathbb{Q}$.*

Proof. See [PS93, Proposition 2.3]. $\square$

Theorem 5.2.1 implies together with Theorem A.1.6 that the variety $Z(\mathscr{A}_0,\dots,\mathscr{A}_n)$ can be characterized by only one equation if and only if $\operatorname{codim}_{\mathbb{C}^p} Z(\mathscr{A}_0,\dots,\mathscr{A}_n)=1$.

Proposition 5.2.1. *For $i=0,\dots,n$, let $P_i=\operatorname{Conv}(\mathscr{A}_i)$ be the Newton polytope of $\hat f_i$ for a generic member of $\mathcal{F}$. We have that $\operatorname{codim}_{\mathbb{C}^p} Z(\mathscr{A}_0,\dots,\mathscr{A}_n)=1$ if and only if the following equivalent conditions hold:*

1. *for some $j\in\{0,\dots,n\}$, $\operatorname{MV}(P_0,\dots,P_{j-1},P_{j+1},\dots,P_n)\neq 0$,*
2. *for some $j\in\{0,\dots,n\}$, $\dim\sum_{i\in\mathscr{J}}P_i\geq|\mathscr{J}|$ for every subset $\mathscr{J}\subsetneq\{0,\dots,j-1,j+1,\dots,n\}$,*
3. *there exists a unique subset of $\{\mathscr{A}_0,\dots,\mathscr{A}_n\}$ which is* essential.[1]

Proof. See [PS93, Page 382] for the first two conditions and [Stu94, Corollary 1.1] for the third. □

Corollary 5.2.1. *Under the conditions of Proposition 5.2.1, there is a unique, up to sign, polynomial $\operatorname{Res}_{\mathscr{A}_0,\dots,\mathscr{A}_n}\in A$ with integer coefficients which is irreducible in $\mathbb{Z}[c_{i,m},i=0,\dots,n,m\in\mathscr{A}_i]\subset A$ such that*

$$I_A(Z(\mathscr{A}_0,\dots,\mathscr{A}_n))=\langle\operatorname{Res}_{\mathscr{A}_0,\dots,\mathscr{A}_n}\rangle.$$

In Section 3.4 we defined $\operatorname{Res}_{d_0,\dots,d_n}$ as an element of the coordinate ring A of $\mathcal{F}_S(d_0,\dots,d_n)$. Since $\mathcal{F}_S(d_0,\dots,d_n)\simeq\mathcal{F}_R(d_0,\dots,d_n)$ as affine varieties via homogenization, we can think of A as the coordinate ring of $\mathcal{F}_R(d_0,\dots,d_n)$ as well. In the following Proposition, we write $\operatorname{Res}_{d_0,\dots,d_n}(\hat f_0,\dots,\hat f_n)=\operatorname{Res}_{d_0,\dots,d_n}(\eta_{d_0}(\hat f_0),\dots,\eta_{d_n}(\hat f_n))$.

Proposition 5.2.2. *If $\mathscr{A}_i=d_i\Delta_n, d_i\in\mathbb{N}$ for $i=0,\dots,n$, we have that*

$$\operatorname{Res}_{\mathscr{A}_0,\dots,\mathscr{A}_n}=\operatorname{Res}_{d_0,\dots,d_n}\quad\textit{(up to sign)}.$$

Equivalently, $(\hat f_0,\dots,\hat f_n)\in Z(\mathscr{A}_0,\dots,\mathscr{A}_n)$ if and only if $\operatorname{Res}_{d_0,\dots,d_n}(\hat f_0,\dots,\hat f_n)=0$.

Proof. If $\hat f_0=\cdots=\hat f_n=0$ has a solution $t=(t_1,\dots,t_n)\in(\mathbb{C}^*)^n$, then $(1:t_1:\cdots:t_n)\in\mathbb{P}^n$ is a solution of the homogeneous system $f_0=\cdots=f_n=0$ obtained as $f_i=\eta_{d_i}(\hat f_i)$ and $\operatorname{Res}_{d_0,\dots,d_n}(\hat f_0,\dots,\hat f_n)=\operatorname{Res}_{d_0,\dots,d_n}(f_0,\dots,f_n)=0$. It follows that $\operatorname{Res}_{d_0,\dots,d_n}$ vanishes on $Z_0(\mathscr{A}_0,\dots,\mathscr{A}_n)$. Since $Z(\mathscr{A}_0,\dots,\mathscr{A}_n)$ is the Zariski closure of $Z_0(\mathscr{A}_0,\dots,\mathscr{A}_n)$, this implies that $\operatorname{Res}_{d_0,\dots,d_n}$ vanishes on $Z(\mathscr{A}_0,\dots,\mathscr{A}_n)$. Since $\operatorname{Res}_{d_0,\dots,d_n}$ is irreducible (Theorem 3.4.1) and $\operatorname{codim}_{\mathbb{C}^p} Z(\mathscr{A}_0,\dots,\mathscr{A}_n)=1$, the statement follows. □

[1] This is a condition on the affine lattices generated by subsets of $\{\mathscr{A}_0,\dots,\mathscr{A}_n\}$. We included this statement for completeness and refer to [Stu94] for a precise definition.

It is clear that if $(\hat{f}_0, \dots, \hat{f}_n) \in \mathcal{F} = \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_0, \dots, \mathscr{A}_n)$ has a solution in $(\mathbb{C}^*)^n$, then $(\hat{f}_0, \dots, \hat{f}_n) \in Z_0 \subset Z$. In general, the inclusion $Z_0 \subset Z$ is strict and the converse statement does not hold: $(\hat{f}_0, \dots, \hat{f}_n) \in \mathcal{F}$ might be in Z, even though it does not define any solutions in $(\mathbb{C}^*)^n$. In the projective case, we have seen that we can make this an 'if and only if' by considering a larger solution space, namely $\mathbb{P}^n \supset (\mathbb{C}^*)^n$. This generalizes nicely for toric resultants [GKZ94, Chapter 8, Proposition 1.5], where the appropriate solution space to consider is the projective toric variety X associated to the Minkowski sum $P_0 + \cdots + P_n$, where $P_i = \mathrm{Conv}(\mathscr{A}_i)$. We will say a few more things in this direction for readers who are familiar with toric geometry. The use of projective toric varieties as solution spaces for polyhedral families of systems will be motivated and explained in more detail in Sections 5.4 and 5.5.
In analogy with the projective case, a member $(\hat{f}_0, \dots, \hat{f}_n) \in \mathcal{F}$ is regarded as a global section s of the rank $n+1$ vector bundle with sheaf of sections $\mathscr{O}_X(D_{P_0}) \oplus \cdots \oplus \mathscr{O}_X(D_{P_n})$ on X, where D_{P_i} is the basepoint free Cartier divisor on X associated to the polytope P_i. The vector space of sections of this bundle is $\mathcal{F}_{\mathbb{C}[M]}(P_0, \dots, P_n) \supset \mathcal{F}$ and the toric resultant characterizes exactly when the zero locus of s on X is nonempty.
Just like in the projective case, toric resultants can be used to detect whether a square system $(\hat{f}_1, \dots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \dots, \mathscr{A}_n)$ defines solutions on the boundary of the torus $(\mathbb{C}^*)^n$ in the toric variety X associated to $P = P_1 + \cdots + P_n$. We describe briefly how that works. For more details, see for instance the appendix of [HS95]. Each facet Q of P is a Minkowski sum $Q = Q_1 + \cdots + Q_n$ where $Q_i \subset P_i$ is a face. Setting $\mathscr{A}_i(Q) = \mathscr{A}_i \cap Q_i$, we obtain a *face system* $(\hat{f}_1(Q), \dots, \hat{f}_n(Q)) \in \mathcal{F}_Q = \mathcal{F}_{\mathbb{C}[M_Q]}(\mathscr{A}_1(Q), \dots, \mathscr{A}_n(Q))$ in a lattice M_Q of rank $n-1$ given by

$$\hat{f}_i(Q) = \sum_{m \in \mathscr{A}_i(Q)} c_{i,m} t^m.$$

For these n equations in $\mathbb{C}[M_Q]$, the toric resultant $\mathrm{Res}_{\mathscr{A}_1(Q),\dots,\mathscr{A}_n(Q)} \in \mathbb{C}[\mathcal{F}_Q]$ vanishes at $(\hat{f}_1(Q), \dots, \hat{f}_n(Q))$ if and only if $(\hat{f}_1, \dots, \hat{f}_n)$ defines a solution on the torus invariant prime divisor $D_Q \subset X$ corresponding to the facet Q.

Remark 5.2.1. Following an analogous argument as in Remark 3.4.1 it is not hard to see that under the conditions of Proposition 5.2.1, the toric resultant $\mathrm{Res}_{\mathscr{A}_0,\dots,\mathscr{A}_n}$ is homogeneous in each group of variables $\{c_{i,m}, m \in \mathscr{A}_i\}$ of degree $\mathrm{MV}(\{P_j \mid j \neq i\})$. For a proof, see Proposition 1.6 in [GKZ94] (where it is assumed that P_i is full-dimensional for each i), or Corollary 2.4 in [PS93]. △

5.2.2 The Canny-Emiris construction

Just like in the projective case, toric resultants give rise to several methods for solving square systems in $\mathcal{F} = \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \dots, \mathscr{A}_n)$. In this subsection we discuss a construction due to Canny and Emiris [CE93] which gives a matrix $\mathrm{New}_{\mathscr{A}_0,\dots,\mathscr{A}_n}$ whose entries are variables of A (i.e. coefficients of a general system of $\mathcal{F}$) and whose determinant is a nonzero multiple of $\mathrm{Res}_{\mathscr{A}_0,\dots,\mathscr{A}_n}$. We also show how this leads to a way of obtaining multiplication operators using Schur complements. The authors of [CE93] call this matrix the *Newton matrix*, because of its relation to the Newton polytopes defining the associated polyhedral family. Explaining the details of the construction requires the introduction of concepts such as *polyhedral subdivisions* (of a special type) and a way of obtaining them via *lifting functions*. Since these will not play a role in the remainder of this text, this would lead us too far. We limit ourselves to a discussion of the main ideas and an example. For more information, see [CE93] or [CCC$^+$05, Chapter 7].

Consider the polytope $P = P_0 + \cdots + P_n$, where $P_i = \mathrm{Conv}(\mathscr{A}_i) \subset \mathbb{R}^n$. We will keep assuming that the supports $\mathscr{A}_0, \ldots, \mathscr{A}_n$ affinely span the lattice M, which implies that the polytope P is full-dimensional. We fix a sufficiently small, random vector $v \in \mathbb{R}^n$ and consider the polytope $P + v = \{m + v \mid m \in P\}$. Note that this is not a lattice polytope anymore. We define the subset

$$\mathcal{E} = (P + v) \cap M.$$

The lattice points in $\mathcal{E}$ are identified with Laurent monomials: $\mathcal{V} = \{t^m \mid m \in \mathcal{E}\}$. These will be the monomials indexing the rows of the matrix $\mathrm{New}_{\mathscr{A}_0,\ldots,\mathscr{A}_n}$ which we are about to construct. In analogy with the projective resultant, the set $\mathcal{V}$ is partitioned into subsets $\Sigma_0', \ldots, \Sigma_n'$ corresponding to $\hat{f}_0, \ldots, \hat{f}_n$, such that

$$|\Sigma_0'| = \mathrm{MV}(P_1, \ldots, P_n)$$

is the expected number of solutions of $\hat{f}_1 = \cdots = \hat{f}_n = 0$. The matrix $\mathrm{New}(\mathscr{A}_0, \ldots, \mathscr{A}_n)$ will be partitioned into block rows corresponding to $\Sigma_0', \ldots, \Sigma_n'$ and block columns corresponding to sets of Laurent monomials $\Sigma_0, \ldots, \Sigma_n \subset M$ of the same cardinality: $|\Sigma_i'| = |\Sigma_i|$. In particular, $\Sigma_0 = \Sigma_0'$. Denoting $V = \mathrm{span}_{\mathbb{C}}(\mathcal{V})$, the columns of the matrix will represent elements of $\langle \hat{f}_0, \ldots, \hat{f}_n \rangle \cap V \subset \mathbb{C}[M]$. More precisely, the columns in the block corresponding to Σ_i represent the polynomials $\{t^m \hat{f}_i \mid t^m \in \Sigma_i\}$ (which requires $t^m \cdot \Sigma_i \subset \mathcal{V}$ for all $m \in \mathscr{A}_i$). Using the short notation $\mathrm{New}_{\mathscr{A}_0,\ldots,\mathscr{A}_n}(\hat{f}_0, \ldots, \hat{f}_n) = \mathrm{New}(\hat{f}_0, \ldots, \hat{f}_n)$, we obtain a matrix of size $|\mathcal{E}| \times |\mathcal{E}|$ partitioned as follows:

$$\mathrm{New}(\hat{f}_0, \ldots, \hat{f}_n) = \begin{array}{c} \\ \Sigma_0' \\ \{\Sigma_1', \ldots, \Sigma_n'\} \end{array} \!\!\begin{array}{c} \begin{array}{c|c} \Sigma_0 & \{\Sigma_1, \ldots, \Sigma_n\} \end{array} \\ \left[\begin{array}{c|c} M_{00} & M_{01} \\ \hline M_{10} & M_{11} \end{array}\right] \end{array}. \tag{5.2.1}$$

Denoting $V_i = \mathrm{span}_{\mathbb{C}}(\Sigma_i)$, this matrix represents a resultant map

$$\mathrm{New}(\hat{f}_0, \ldots, \hat{f}_n) : V_0 \times \cdots \times V_n \to V$$

as in Definition 4.3.1, where the ring R is replaced by $\mathbb{C}[M]$. The restriction

$$\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n} = \mathrm{New}(\hat{f}_0, \ldots, \hat{f}_n)_{|V_1 \times \cdots \times V_n} = \begin{bmatrix} M_{01} \\ M_{11} \end{bmatrix} : V_1 \times \cdots \times V_n \to V$$

is such that $\operatorname{im}\operatorname{res}_{\hat{f}_1,\dots,\hat{f}_n} \subset I \cap V$, with $I = \langle \hat{f}_1, \dots, \hat{f}_n \rangle \in \mathbb{C}[M]$. As mentioned above, the matrix $\mathrm{New}_{\mathscr{A}_0,\dots,\mathscr{A}_n}$ is such that $\det(\mathrm{New}_{\mathscr{A}_0,\dots,\mathscr{A}_n})$ is a nonzero multiple of $\mathrm{Res}_{\mathscr{A}_0,\dots,\mathscr{A}_n}$ [CE93, Section 6]. By [Emi96, Lemma 4.4], the submatrix M_{11} in this construction is invertible for generic members of $\mathcal{F}$. Defining the Schur complement

$$M_{\hat{f}_0} = M_{00} - M_{01} M_{11}^{-1} M_{10}$$

and the (row vector valued) map

$$\phi_{\Sigma_0} : (\mathbb{C}^*)^n \to \mathbb{C}^\delta \quad \text{by} \quad \phi_{\Sigma_0}(z) = (z^m \mid t^m \in \Sigma_0),$$

a straightforward adaptation of the proof of Theorem 3.4.2 reveals that for $z \in V_{(\mathbb{C}^*)^n}(I)$,

$$\phi_{\Sigma_0}(z) M_{\hat{f}_0} = \hat{f}_0(z) \phi_{\Sigma_0}(z).$$

This shows, at least for the case where all $z \in V_{(\mathbb{C}^*)^n}(I)$ have multiplicity 1, by Theorem 5.1.1 that $M_{\hat{f}_0}$ represents the multiplication map

$$M_{\hat{f}_0} : \mathbb{C}[M]/I \to \mathbb{C}[M]/I \quad \text{given by} \quad M_{\hat{f}_0}(g + I) = \hat{f}_0 g + I,$$

where $\mathbb{C}[M]/I$ is identified with V_0.

Remark 5.2.2. In this construction, the basis used for the quotient ring $\mathbb{C}[M]/I$ corresponds to the Laurent monomials in $\Sigma_0 = \Sigma_0'$. These monomials correspond to the lattice points in the interior of so-called *mixed cells* in a coherent mixed subdivision of $P + v$. For this reason, this type of basis for $\mathbb{C}[M]/I$ is called a *mixed monomial basis* [PS96]. △

Remark 5.2.3. Note that the size of the matrix $\mathrm{New}_{\mathscr{A}_0,\dots,\mathscr{A}_n}$ only depends on the Newton polytopes $P_0, \dots, P_n$. This means that, in practice, the complexity of algorithms related to these resultant constructions often only depends on $P_0, \dots, P_n$, unless the sparsity of the matrix can be taken into account. △

Example 5.2.1. This is Example 7.2.5 in [CCC⁺05, Chapter 7]. Consider the support $\mathscr{A}$ from Example 5.1.5 and the family $\mathcal{F} = \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}, \mathscr{A}, \mathscr{A})$ $(n = 2)$. The polytope $P + v$ is depicted in Figure 5.2. We set

$$\mathcal{V} = \{1, t_1, t_1^2, t_2, t_1t_2, t_1^2t_2, t_2^2, t_1t_2^2, t_1^2t_2^2\},$$
$$\Sigma_0 = \{1, t_1t_2\}, \quad \Sigma_1 = \{1, t_1, t_2, t_1t_2\}, \quad \Sigma_2 = \{1, t_1, t_2\}.$$

A member of the family $\mathcal{F}$ is given by

$$\hat{f}_0 = a_0 + a_1t_1 + a_2t_2 + a_3t_1t_2,$$
$$\hat{f}_1 = b_0 + b_1t_1 + b_2t_2 + b_3t_1t_2,$$
$$\hat{f}_2 = c_0 + c_1t_1 + c_2t_2 + c_3t_1t_2.$$

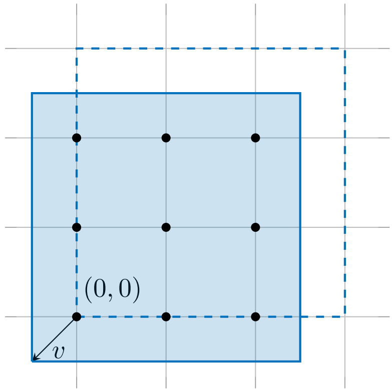


Figure 5.2: The polytope $P + v$ in Example 5.2.1 and the lattice points in $\mathcal{E}$ (black dots).

We obtain the 9×9 matrix

$$
\mathrm{New}_{\mathscr{A},\mathscr{A},\mathscr{A}} = \begin{array}{c c}
 & \begin{array}{cc|ccccccc} 1 & t_1t_2 & 1 & t_1 & t_2 & t_1t_2 & 1 & t_1 & t_2 \end{array} \\
\begin{array}{c} 1 \\ t_1t_2 \\ \hline t_1 \\ t_2 \\ t_1^2 \\ t_1^2t_2 \\ t_2^2 \\ t_1t_2^2 \\ t_1^2t_2^2 \end{array} &
\left[\begin{array}{cc|ccccccc}
a_0 & & b_0 & & & & c_0 & & \\
a_3 & a_0 & b_3 & b_2 & b_1 & b_0 & c_3 & c_2 & c_1 \\ \hline
a_1 & & b_1 & b_0 & & & c_1 & c_0 & \\
a_2 & & b_2 & & b_0 & & c_2 & & c_0 \\
 & & & b_1 & & & & c_1 & \\
 & a_1 & & b_3 & & b_1 & & c_3 & \\
 & & & & b_2 & & & & c_2 \\
 & a_2 & & & b_3 & b_2 & & & c_3 \\
 & a_3 & & & & b_3 & & &
\end{array}\right]
\end{array}
$$

which is partitioned as in (5.2.1). We note that for this example, viewing $(\hat{f}_0, \hat{f}_1, \hat{f}_2)$ as a member of $\mathcal{F}_R(2,2,2) \simeq \mathcal{F}_S(d_0, d_1, d_2)$, the Macaulay matrix $\mathrm{Mac}_{2,2,2}(\hat{f}_0, \hat{f}_1, \hat{f}_2)$ is a 15×15 matrix whose determinant *vanishes identically* on $\mathcal{F}$. Indeed, for any value of the parameters a_i, b_i, c_i, the homogeneous system $f_0 = f_1 = f_2 = 0$ defined by

$$
\begin{aligned}
f_0 &= a_0 x_0^2 + a_1 x_0 x_1 + a_2 x_0 x_2 + a_3 x_1 x_2, \\
f_1 &= b_0 x_0^2 + b_1 x_0 x_1 + b_2 x_0 x_2 + b_3 x_1 x_2, \\
f_2 &= c_0 x_0^2 + c_1 x_0 x_1 + c_2 x_0 x_2 + c_3 x_1 x_2
\end{aligned}
$$

has solutions $(0 : 1 : 0)$ and $(0 : 0 : 1)$ in $\mathbb{P}^2$. Note that this provides an explanation for the difference between the Bézout number of $\mathcal{F}_R(2,2)$ and the mixed volume for $\mathcal{F}_{\mathbb{C}[M]}(P, P)$ established in Example 5.1.5: the support *forces* two out of four solutions to lie 'at infinity'. $\triangle$

We conclude by pointing out that in [EC95], the authors propose an incremental version of the algorithm in [CE93] which produces a matrix with the same properties but usually of smaller size.

5.3 Truncated normal forms for polyhedral families

In this section, we consider a zero-dimensional ideal $I \subset \mathbb{C}[M]$ such that $\dim_{\mathbb{C}} \mathbb{C}[M]/I = \delta^+$ and its contraction $I^c = I \cap R \subset R$. By the results of Section 4.2 and the proof of Theorem 5.1.1, the coordinates of the points in $V_{(\mathbb{C}^*)^n}(I)$ can be computed via eigenvalue computations once we have computed a TNF with respect to I^c.

Theorem 5.3.1. *Let V be a finite dimensional $\mathbb{C}$-vector subspace of $R \subset \mathbb{C}[M]$ and let $W \subset V$ be its largest subspace such that $W^+ \subset V$. If the space V and a $\mathbb{C}$-linear map $N : V \to \mathbb{C}^{\delta^+}$ satisfy the following properties:*

1. $\ker N \subset I \cap V$ *and there is $u \in V$ such that $u + I$ is a unit in $\mathbb{C}[M]/I$,*
2. $N_{|W} : W \to \mathbb{C}^{\delta^+}$ *is surjective,*

then for any δ^+-dimensional subspace $B \subset W$ such that $N_{|B}$ is invertible, $\mathcal{N}_V = (N_{|B})^{-1} \circ N : V \to B$ is a TNF with respect to I^c.

Proof. Note that $I \cap V = I \cap R \cap V = I^c \cap V$ and for $u \in V \subset R$, $u + I$ is a unit in $\mathbb{C}[M]/I$ if and only if $u + I^c$ is a unit in R/I^c by Lemma 5.1.1. The theorem follows from Corollary 4.2.1. $\square$

In the terminology of Section 4.2, the map $N : V \to \mathbb{C}^{\delta^+}$ in Theorem 5.3.1 *covers* a TNF with respect to I^c. We will now derive one possible way of computing such a map $N : V \to \mathbb{C}^{\delta^+}$ as the cokernel of a resultant map in the case of square systems. The constructions we propose are strongly related to the Canny-Emiris construction from Subsection 5.2.2 and essentially, they only depend on the Newton polytopes of the Laurent polynomials defining the system. In what follows, we assume that $I = \langle \hat{f}_1, \ldots, \hat{f}_n \rangle$ where $(\hat{f}_1, \ldots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(P_1, \ldots, P_n)$ for some polytopes $P_1, \ldots, P_n \subset \mathbb{R}^n$. If the system one wants to solve is supported in $\mathscr{A}_1, \ldots, \mathscr{A}_n$, one should consider it as a member of $\mathcal{F}_{\mathbb{C}[M]}(P_1, \ldots, P_n)$ where $P_i = \operatorname{Conv}(\mathscr{A}_i)$. Since all Laurent monomials are units in $\mathbb{C}[M]$, we may assume $\hat{f}_i \in R, i = 1, \ldots, n$. We take $\hat{f}_0 \in \mathcal{F}_{\mathbb{C}[M]}(\Delta_n)$ to be any affine function in $\mathbb{C}[M]$ and set $P_0 = \Delta_n$. For the tuple $(\hat{f}_0, \ldots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(P_0, \ldots, P_n)$, we consider a Canny-Emiris construction as in Subsection 5.2.2. This gives a matrix

$$\operatorname{New}(\hat{f}_0, \ldots, \hat{f}_n) \in \mathbb{C}^{|\mathcal{E}| \times |\mathcal{E}|}$$

where $\mathcal{E} = (P_0 + \cdots + P_n + v) \cap M$ for some random small vector $v \in \mathbb{R}^n$. By our assumptions, we have that $\mathcal{V} = \{t^m \mid m \in \mathcal{E}\} \subset R$, and hence $V = \operatorname{span}_{\mathbb{C}}(\mathcal{V}) \subset R$. Recall from Subsection 5.2.2 that for any $\hat{f} \in V_i = \operatorname{span}_{\mathbb{C}}(\Sigma_i)$, we have $\hat{f}\hat{f}_i \in V$. Therefore, since for any $\hat{f}, \hat{f}' \in \mathbb{C}[M]$ we have that $\operatorname{Newt}(\hat{f}\hat{f}') = \operatorname{Newt}(\hat{f}) + \operatorname{Newt}(\hat{f}')$, we must have

$$V_i \subset \bigoplus_{m \in Q_i \cap M} \mathbb{C} \cdot t^m, \quad \text{where } Q_i = P_0 + \cdots + P_{i-1} + P_{i+1} + \cdots + P_n + v. \tag{5.3.1}$$

Recall that by restricting the map represented by $\operatorname{New}(\hat{f}_0, \ldots, \hat{f}_n)$ to $V_1 \times \cdots \times V_n$ we get a resultant map

$$\operatorname{res}_{\hat{f}_1,\ldots,\hat{f}_n} = \operatorname{New}(\hat{f}_0, \ldots, \hat{f}_n)_{|V_1 \times \cdots \times V_n} = \begin{bmatrix} M_{01} \\ M_{11} \end{bmatrix} : V_1 \times \cdots \times V_n \to V.$$

The following is the analogue of Proposition 4.3.1 in the toric case.

Proposition 5.3.1. *Let $(\hat{f}_0, \ldots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(P_0 = \Delta_n, P_1, \ldots, P_n)$, $I = \langle \hat{f}_1, \ldots, \hat{f}_n \rangle$ and consider the resultant map*

$$\operatorname{res}_{\hat{f}_1,\ldots,\hat{f}_n} = \operatorname{New}(\hat{f}_0, \ldots, \hat{f}_n)_{|V_1 \times \cdots \times V_n} : V_1 \times \cdots \times V_n \to V$$

with $V_i = \operatorname{span}_{\mathbb{C}}(\Sigma_i)$ and $V = \operatorname{span}_{\mathbb{C}}(\mathcal{V})$. If the submatrix M_{11} of $\operatorname{New}(\hat{f}_0, \ldots, \hat{f}_n)$ is invertible, then the corank of $\operatorname{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ is $\delta^+ = \dim_{\mathbb{C}} \mathbb{C}[M]/I$ and any cokernel map $N : V \to \mathbb{C}^{\delta^+}$ of $\operatorname{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ covers a TNF with respect to $I^c = I \cap R$.

Proof. Up to using Theorem 5.3.1, the fact that any monomial t^m in V corresponds to a unit $t^m + I$ in $\mathbb{C}[M]/I$ and $V_0 = \operatorname{span}_{\mathbb{C}}(\Sigma_0) \subset W$ since $V_0 + \Delta_n \subset V$ (see (5.3.1)), the proof is identical to the proof of Proposition 4.3.1. $\square$

By [Emi96, Lemma 4.4], the condition that M_{11} is invertible holds for generic members of $\mathcal{F}_{\mathbb{C}[M]}(P_0, \ldots, P_n)$. Better yet, it holds for generic members of any subfamily $\mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_0, \cdots, \mathscr{A}_n)$ such that $\operatorname{Conv}(\mathscr{A}_i) = P_i, i = 0, \ldots, n$. It follows from the fact that a cokernel map $N : V \to \mathbb{C}^{\delta^+}$ covers a TNF that $\ker N = \operatorname{im} \operatorname{res}_{\hat{f}_1,\ldots,\hat{f}_n} = I^c \cap V = I \cap V$. By Theorem 5.1.2, the number δ^+ in Proposition 5.3.1 is $\operatorname{MV}(P_1, \ldots, P_n)$.

As in the total degree case, we will use 'larger' resultant maps $V_1 \times \cdots \times V_n \to V$ in our TNF construction to stabilize the numerical computation of the cokernel. That is, we keep $V = \operatorname{span}_{\mathbb{C}}(\mathcal{V})$ and pick the subspaces $V_1, \ldots, V_n \subset R$ as large as possible such that $\hat{f}_i \cdot V_i \subset V$. We replace the inclusion in (5.3.1) by an equality:

$$V_i = \bigoplus_{m \in Q_i \cap M} \mathbb{C} \cdot t^m, \quad \text{where } Q_i = P_0 + \cdots + P_{i-1} + P_{i+1} + \cdots + P_n + v. \tag{5.3.2}$$

Corollary 5.3.1. *Let $(\hat{f}_1, \ldots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(P_1, \ldots, P_n)$ and consider the resultant map*

$$\operatorname{res}_{\hat{f}_1,\ldots,\hat{f}_n} : V_1 \times \cdots \times V_n \to V$$

with V_i *as in* (5.3.2) *and* $V = \mathrm{span}_{\mathbb{C}}(\mathcal{V})$*. For a generic member* $(\hat{f}_1, \ldots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(P_1, \ldots, P_n)$*, the corank of* $\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ *is* δ^+ *and any cokernel map* $N : V \to \mathbb{C}^{\delta^+}$ *of* $\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ *covers a TNF with respect to* I^c*.*

Proof. Let $\mathrm{res}'_{\hat{f}_1,\ldots,\hat{f}_n}$ be the resultant map from Proposition 5.3.1. By Proposition 5.3.1 and [Emi96, Lemma 4.4], for a generic member $(\hat{f}_0, \ldots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(P_0 = \Delta_n, P_1, \ldots, P_n)$ we have $\mathrm{im}\,\mathrm{res}'_{\hat{f}_1,\ldots,\hat{f}_n} = I \cap V$. Moreover, it is sufficient that $(\hat{f}_1, \ldots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(P_1, \ldots, P_n)$ be generic, since the coefficients of $\hat{f}_0$ are not involved in the matrix M_{11}. This implies

$$I \cap V = \mathrm{im}\,\mathrm{res}'_{\hat{f}_1,\ldots,\hat{f}_n} \subset \mathrm{im}\,\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n} \subset I \cap V.$$

Therefore $\mathrm{im}\,\mathrm{res}'_{\hat{f}_1,\ldots,\hat{f}_n} = \mathrm{im}\,\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$ and the cokernels of both maps agree. The statement now follows from Proposition 5.3.1. □

Remark 5.3.1. Since the subspaces V_i defining the resultant map of Corollary 5.3.1 depend on the random vector $v \in \mathbb{R}^n$, it is not straightforward to investigate what 'generic' means exactly in the context of this statement. We will be able to say more about this for a different construction in Section 5.5 via a homogeneous interpretation. △

Algorithm 5.3 Computes multiplication matrices for generic $(\hat{f}_1, \ldots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(P_1, \ldots, P_n)$

1: **procedure** MultiplicationMatrices($\hat{f}_1, \ldots, \hat{f}_n$)
2: $\quad v \leftarrow$ random small n-vector
3: $\quad \mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n} \leftarrow$ the resultant map $V_1 \times \cdots \times V_n \to V$ from Corollary 5.3.1
4: $\quad N \leftarrow \mathrm{coker}\,\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$
5: $\quad N_{|W} \leftarrow$ restriction of N to the largest subspace $W \subset V$ such that $W^+ \subset V$
6: $\quad N_{|B} \leftarrow$ any invertible restriction of $N_{|W}$ $(\dim_{\mathbb{C}} B = \delta^+)$
7: $\quad$ **for** $i = 1, \ldots, n$ **do**
8: $\qquad N_i \leftarrow N_{|t_i \cdot B}$
9: $\qquad M_{t_i} \leftarrow (N_{|B})^{-1} N_i$
10: $\quad$ **end for**
11: $\quad$ **return** $M_{t_1}, \ldots, M_{t_n}$
12: **end procedure**

In the notation of Algorithm 5.3, it is understood that if in line 6, the map $N_{|B}$ is represented in the basis $\mathcal{B}$ for B, then $N_{t_i \cdot B}$ in line 8 should be represented in the basis $t_i \cdot \mathcal{B}$. The choice of the subspace $B \subset W \subset V$ in line 6 can happen using the QR or SVD techniques proposed in the previous chapter.

Remark 5.3.2 (On the complexity of Algorithm 5.3)**.** The complexity analysis in Remark 4.3.2 can straightforwardly be adapted to Algorithm 5.3. In this case, the

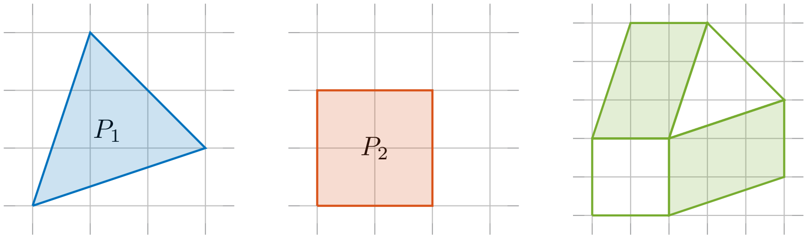


Figure 5.3: The polytopes P_1 (left), P_2 (center) from Example 5.3.1 and their Minkowski sum $P_1 + P_2$ (right).

sizes of the matrices depend on the number of monomials in $V_1, \ldots, V_n, V$ and on the mixed volume $\delta^+ = \mathrm{MV}(P_1, \ldots, P_n)$. The conclusion that the cokernel computation (line 4) dominates the computational cost of the algorithm holds in this case as well. In particular, the cost of using column pivoted QR or SVD on the matrix $N_{|W}$, which is usually much smaller than $\mathrm{res}_{\hat{f}_1,\ldots,\hat{f}_n}$, is negligible as compared to the cokernel computation, yet it is crucial for the numerical stability. We point out that the complexity of the cokernel computation in line 4 can be straightforwardly reduced by applying the second technique proposed in Subsection 4.4.1. $\triangle$

Example 5.3.1. Let $n = 2$, $\mathbb{C}[M] = \mathbb{C}[t_1^{\pm 1}, t_2^{\pm 1}]$ and consider the polynomials

$$\begin{aligned}\hat{f}_1 &= a_0 + a_1 t_1^3 t_2 + a_2 t_1 t_2^3,\\ \hat{f}_2 &= b_0 + b_1 t_1^2 + b_2 t_2^2 + b_3 t_1^2 t_2^2.\end{aligned}$$

The Newton polygons, together with their Minkowski sum, are shown in Figure 5.3. By applying the formula (D.1.3) we find that the BKK number for the system $\hat{f}_1 = \hat{f}_2 = 0$ is the area of the shaded regions in the right part of Figure 5.3, which is 12. Note that the Bézout bound is 16. Using $v = (-0.3, -0.4)$ we obtain the set $\mathcal{E}$ marked with black dots in Figure 5.4. The points in $\mathcal{E}$ correspond to the monomials t^m which span the $\mathbb{C}$-vector space V from Corollary 5.3.1. The $\mathbb{C}$-vector spaces V_1, V_2 are spanned by the monomials corresponding to the lattice points in $\Delta_2 + P_2 + v$ and $\Delta_2 + P_1 + v$ respectively. The diagram on the left side of Figure 5.5 illustrates the toric resultant map. For comparison, Figure 5.5 also shows an analogous picture for the total degree resultant map used in Algorithm 4.1. However, a cokernel of this resultant map does not yield a map that covers a TNF: the assumptions of Proposition 4.3.2 are not satisfied! The 4 'missing' solutions with respect to Bézout's bound lie, for any choice of a_i, b_i, on the line at infinity. However, the total degree resultant map can still be used in a homogeneous interpretation as in Algorithm 4.2. In Figure 5.5, the black dots in the blue and orange polytopes index the columns of the matrix representing the resultant map. The black dots in the purple polytopes index its rows. This means that in this example, the toric resultant map corresponds to an 29×18 matrix and the total degree resultant map has size 36×20. $\triangle$

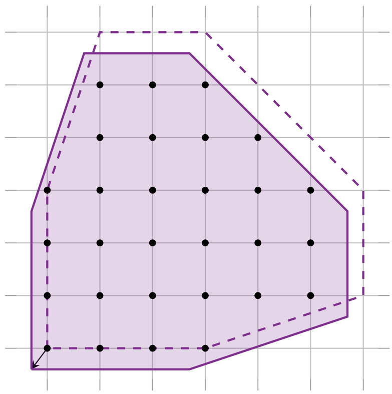


Figure 5.4: The polytope $\Delta_2 + P_1 + P_2 + v$ and its interior lattice points (black dots) corresponding to $\mathcal{E}$ from Example 5.3.1.

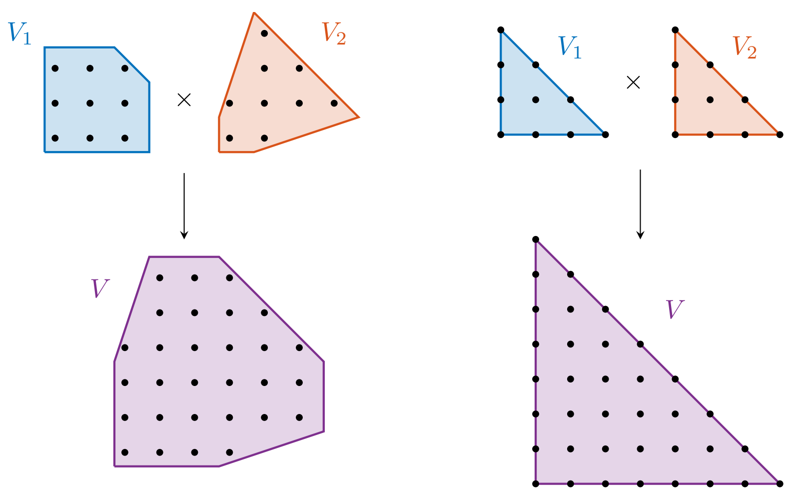


Figure 5.5: Illustration of the resultant maps from Corollary 5.3.1 (left) and Proposition 4.3.2 (right).

We conclude the section with some numerical experiments. In our algorithms, we have used the Schur decomposition for computing the coordinates of the solutions from the multiplication matrices. The machine used to perform the experiments is the same as in Subsection 4.3.3. The residual is measured as in Appendix C.

Experiment 5.3.1 ('Block' supports)**.** We consider $\mathcal{F}_{n,d} = \mathcal{F}_{\mathbb{C}[M]}(P, \ldots, P)$ (P is listed n times) where P is the hypercube $P = [0, d]^n$ for some $d \in \mathbb{N}$. This corresponds to polynomial systems $\hat{f}_1 = \cdots = \hat{f}_n = 0$ where the monomials occuring in $\hat{f}_i$ are

$$t^m = t_1^{m_1} \cdots t_n^{m_n} \quad \text{such that} \quad 0 \le m_i \le d, i = 1, \ldots, n.$$

An example for $d = 1, n = 3$ was given in Example 3.2.4. The Bézout bound for $\mathcal{F}_{n,d}$ is $(nd)^n$, whereas the BKK number is $n!d^n$. For different n and d, we solve generic members of $\mathcal{F}_{n,d}$ generated by drawing the coefficients from a real, standard normal distribution. We use a Matlab implementation of Algorithm 5.3, which calls Polymake [AGH+17] for all computations involving polytopes, except for computing the mixed volume, which is done using PHCpack [Ver99]. We compare the results with those of the function `qdsparf` and `sparf` from the PNLA package (see Experiment 4.3.3). Results for $n = 2$ and $n = 3$ are shown in Tables 5.1 and 5.2. The tables report computation time $t_\star$ (in seconds), number of computed solutions $\delta_\star$, maximal residual $r_{\max,\star}$ and geometric mean residual $r_{\text{mean},\star}$ for each solver $\star$.

d	t_{QR}	δ_{QR}	$r_{\max,\text{QR}}$	$r_{\text{mean},\text{QR}}$	t_{SVD}	δ_{SVD}	$r_{\max,\text{SVD}}$	$r_{\text{mean},\text{SVD}}$
1	2.977	2	$3.15 \cdot 10^{-16}$	$2.85 \cdot 10^{-16}$	2.913	2	$3.71 \cdot 10^{-16}$	$2.4 \cdot 10^{-16}$
2	2.954	8	$7.16 \cdot 10^{-15}$	$1.06 \cdot 10^{-15}$	2.968	8	$1.34 \cdot 10^{-14}$	$1.03 \cdot 10^{-15}$
3	2.981	18	$1.86 \cdot 10^{-14}$	$6.53 \cdot 10^{-15}$	2.941	18	$9.88 \cdot 10^{-15}$	$2.23 \cdot 10^{-15}$
4	2.939	32	$7.64 \cdot 10^{-13}$	$1.55 \cdot 10^{-14}$	3.248	32	$5 \cdot 10^{-15}$	$9.39 \cdot 10^{-16}$
5	2.981	50	$2.44 \cdot 10^{-14}$	$2.26 \cdot 10^{-15}$	2.935	50	$4.7 \cdot 10^{-15}$	$1.06 \cdot 10^{-15}$
6	2.942	72	$4.35 \cdot 10^{-14}$	$6.77 \cdot 10^{-15}$	3.000	72	$3.38 \cdot 10^{-15}$	$9.55 \cdot 10^{-16}$
7	3.068	98	$4.9 \cdot 10^{-14}$	$6.9 \cdot 10^{-15}$	3.195	98	$4.1 \cdot 10^{-15}$	$1.1 \cdot 10^{-15}$
8	3.075	128	$4.45 \cdot 10^{-13}$	$2.73 \cdot 10^{-15}$	3.098	128	$5.27 \cdot 10^{-15}$	$8.58 \cdot 10^{-16}$

d	t_{qdsparf}	δ_{qdsparf}	$r_{\max,\text{qdsparf}}$	$r_{\text{mean},\text{qdsparf}}$	t_{sparf}	δ_{sparf}	$r_{\max,\text{sparf}}$	$r_{\text{mean},\text{sparf}}$
1	0.004	2	$6.36 \cdot 10^{-17}$	$4.67 \cdot 10^{-17}$	0.010	2	$1.17 \cdot 10^{-16}$	$1.05 \cdot 10^{-16}$
2	0.012	8	$1.62 \cdot 10^{-11}$	$7.12 \cdot 10^{-15}$	0.055	8	$8.37 \cdot 10^{-11}$	$8.66 \cdot 10^{-15}$
3	0.038	19	0.46	$1.09 \cdot 10^{-12}$	0.287	18	0.44	$1.54 \cdot 10^{-12}$
4	0.095	33	0.14	$3.76 \cdot 10^{-13}$	0.728	32	$1.54 \cdot 10^{-2}$	$1.76 \cdot 10^{-13}$
5	0.146	50	$1.62 \cdot 10^{-5}$	$4.34 \cdot 10^{-13}$	1.910	50	$1.5 \cdot 10^{-7}$	$1.28 \cdot 10^{-13}$
6	0.494	75	0.37	$2.3 \cdot 10^{-11}$	4.708	72	$1.8 \cdot 10^{-2}$	$1 \cdot 10^{-12}$
7	0.684	100	0.3	$1.68 \cdot 10^{-10}$	12.356	97	$6.87 \cdot 10^{-2}$	$8.08 \cdot 10^{-11}$
8	0.840	129	$9.59 \cdot 10^{-2}$	$1.14 \cdot 10^{-12}$	22.612	128	$5.53 \cdot 10^{-6}$	$2.73 \cdot 10^{-13}$

Table 5.1: Results for a Matlab implementation of Algorithm 5.3 with QR/SVD for basis selection and the functions `qdsparf`, `sparf` from PNLA for the families $\mathcal{F}_{2,d}$ of Experiment 5.3.1.

Calling Polymake from Matlab causes some overhead (a little less than 3 seconds for $n = 2$), which can be seen from the fact that the computation time almost doesn't

d	t_{QR}	δ_{QR}	$r_{\max,\text{QR}}$	$r_{\text{mean},\text{QR}}$	t_{SVD}	δ_{SVD}	$r_{\max,\text{SVD}}$	$r_{\text{mean},\text{SVD}}$
1	4.562	6	$6.06 \cdot 10^{-16}$	$3.29 \cdot 10^{-16}$	4.741	6	$9.25 \cdot 10^{-16}$	$2.62 \cdot 10^{-16}$
2	4.834	48	$1.99 \cdot 10^{-14}$	$2.66 \cdot 10^{-15}$	4.655	48	$2.96 \cdot 10^{-15}$	$6.46 \cdot 10^{-16}$
3	6.105	162	$2.03 \cdot 10^{-12}$	$1.6 \cdot 10^{-14}$	5.842	162	$1.35 \cdot 10^{-13}$	$1.2 \cdot 10^{-15}$

d	t_{qdsparf}	δ_{qdsparf}	$r_{\max,\text{qdsparf}}$	$r_{\text{mean},\text{qdsparf}}$	t_{sparf}	δ_{sparf}	$r_{\max,\text{sparf}}$	$r_{\text{mean},\text{sparf}}$
1	0.030	6	$1.54 \cdot 10^{-13}$	$1.32 \cdot 10^{-14}$	0.149	6	$5.68 \cdot 10^{-15}$	$1.23 \cdot 10^{-15}$
2	0.472	48	$8.24 \cdot 10^{-6}$	$5.23 \cdot 10^{-11}$	6.896	48	$5.92 \cdot 10^{-7}$	$1.79 \cdot 10^{-12}$
3	26.551	172	0.76	$1.54 \cdot 10^{-3}$	128.489	161	0.54	$2.19 \cdot 10^{-8}$

Table 5.2: Results for a Matlab implementation of Algorithm 5.3 with QR/SVD for basis selection and the functions `qdsparf`, `sparf` from PNLA for the families $\mathcal{F}_{3,d}$ of Experiment 5.3.1.

increase for $n = 2$ and increasing d. This can be overcome using the recently developed Polymake interface in Julia [KLT20]. We therefore also implemented Algorithm 5.3 in Julia. With this implementation, solving a generic member of $\mathcal{F}_{2,8}$ takes on average 0.5 seconds. All solutions are found consistently with a residual no larger than $O(10^{-14})$. Numerical approximations of all 5000 solutions of a generic member of $\mathcal{F}_{2,50}$ are found within 17 minutes. The maximal residual is of order 10^{-12}. It takes 4 minutes and 32 seconds to solve $\mathcal{F}_{3,6}$ (1296 solutions), 4 minutes and 12 seconds to solve $\mathcal{F}_{4,2}$ (384 solutions) and 4 minutes and 52 seconds to solve $\mathcal{F}_{5,1}$ (120 solutions).

For $n = 2$, the incremental strategy of the PNLA solvers has to deal with two singular points on the line at infinity, whose multiplicities make up for the difference between the Bézout bound and the BKK number. For $n = 3$, there is a curve 'at infinity' (see Example 3.2.4), which makes things significantly more tricky (e.g. the Hilbert function does not stabilize). Note that these solvers sometimes miss a few solutions, and sometimes they return too many. For $n = 3, d > 3$, the solver `qdsparf` threw an error. $\triangle$

Experiment 5.3.2 (Molecule configurations)**.** In [EM99a], the authors study the use of toric resultants (and other algebraic techniques) for computing the possible configurations of a 6-atom molecule. The system of equations that needs to be solved is $\hat{f}_1 = \hat{f}_2 = \hat{f}_3 = 0$ with

$$\begin{aligned}\hat{f}_1 &= \beta_{11} + \beta_{12}t_2^2 + \beta_{13}t_3^2 + \beta_{14}t_2t_3 + \beta_{15}t_2^2t_3^2,\\ \hat{f}_2 &= \beta_{21} + \beta_{22}t_3^2 + \beta_{23}t_1^2 + \beta_{24}t_3t_1 + \beta_{25}t_3^2t_1^2,\\ \hat{f}_3 &= \beta_{31} + \beta_{32}t_1^2 + \beta_{33}t_2^2 + \beta_{34}t_1t_2 + \beta_{35}t_1^2t_2^2.\end{aligned}$$

Here the variables t_1, t_2, t_3 encode what the authors of [EM99a] call the *flap* angles of the molecule, and the parameters β_{ij} are computed from the fixed bond lengths and bond angles in the molecule. We are dealing with a family of square systems in $\mathbb{C}[M] = \mathbb{C}[t_1^{\pm 1}, t_2^{\pm 1}, t_3^{\pm 1}]$, where each equation only contains 2 out of the 3 variables. We denote this family by $\mathcal{F} = \mathcal{F}_{\mathbb{C}[M]}(P_1, P_2, P_3)$ where P_i is a 2-dimensional lattice polytope in $\mathbb{R}^3$. The BKK number for this family is $\mathrm{MV}(P_1, P_2, P_3) = 16$, whereas the classical Bézout number equals 64. Only the real solutions are physically meaningful. For the cyclohexane molecule, the coefficients (after contamination by noise) are given as the entries β_{ij} of the matrix

$$\beta = \begin{bmatrix} -310 & 959 & 774 & 1389 & 1313 \\ -365 & 755 & 917 & 1451 & 1269 \\ -413 & 837 & 838 & 1655 & 1352 \end{bmatrix}.$$

The Julia implementation of Algorithm 5.3 computes numerical approximations of all 16 solutions in less than half a second, with a maximal residual of order 10^{-15}. There are four real solutions, which correspond to the possible configurations of the molecule. Another interesting member of this family is one whose 16 solutions are all real. The coefficients are

$$\beta = \begin{bmatrix} -13 & -1 & -1 & 24 & -1 \\ -13 & -1 & -1 & 24 & -1 \\ -13 & -1 & -1 & 24 & -1 \end{bmatrix}.$$

Computation time and accuracy are roughly the same as for the cyclohexane problem. These results can be compared to Tables 1-3 in [EM99a], although the computations were performed on a different machine and the residual is measured using an absolute criterion. △

5.4 Solutions on toric varieties

We have seen in the previous chapters that the projective space $\mathbb{P}^n$ is a natural space to look for solutions of a member $(\hat{f}_1, \ldots, \hat{f}_n)$ of $\mathcal{F}_R(d_1, \ldots, d_n)$. Even though we might only be interested in solutions in the open subset $\mathbb{C}^2 \simeq U_0 \subset \mathbb{P}^n$, keeping track of what happens 'at infinity' has several benefits. For instance, it may allow us to explain the number of solutions in $\mathbb{C}^2$, by subtracting the number of solutions at infinity from the Bézout number. It is also natural from a numerical point of view to take roots at infinity into account, as the slightest perturbation inside $\mathcal{F}_R(d_1, \ldots, d_n)$ moves them into $\mathbb{C}^2$. However, if the systems we are interested in belong to a small subfamily of $\mathcal{F}_R(d_1, \ldots, d_n)$, extending the relations $\hat{f}_1 = \cdots = \hat{f}_n = 0$ to $\mathbb{P}^n$ may introduce solution components at infinity that do not seem so natural. For instance, they do not disappear or move into $\mathbb{C}^2$ upon perturbing the system, and they may even be independent of which member of the subfamily we consider. This happened in Example 3.2.4. In this section we will motivate the interpretation of more general toric varieties as a natural solution space for Laurent polynomial systems coming from the more general families $\mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \ldots, \mathscr{A}_n)$. That is, we will consider a projective toric variety X which, in many ways, is to $\mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \ldots, \mathscr{A}_n)$ what $\mathbb{P}^n$ is to $\mathcal{F}_R(d_1, \ldots, d_n)$. In particular, if $\mathscr{A}_i = d_i \Delta_n \cap M$ for some n-tuple $(d_1, \ldots, d_n) \in \mathbb{N}^n_{>0}$, then $X = \mathbb{P}^n$. For some background on toric varieties, see Appendix E.

5.4.1 Unmixed families

We first consider the case where $\mathscr{A}_1 = \cdots = \mathscr{A}_n = \mathscr{A}$ and $\mathscr{A}$ affinely spans the lattice M. The family $\mathcal{F}_{\mathscr{A}} = \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}, \ldots, \mathscr{A})$ ($\mathscr{A}$ is listed n times) is called the *unmixed family* supported in $\mathscr{A}$. Let $\mathscr{A} = \{m_0, \ldots, m_s\} \subset M$ and

$$\hat{f}_j = \sum_{i=0}^{s} c_{j,i} t^{m_i}, \quad j = 1, \ldots, n$$

such that $(\hat{f}_1, \ldots, \hat{f}_n) \in \mathcal{F}_{\mathscr{A}}$. Let $I = \langle \hat{f}_1, \ldots, \hat{f}_n \rangle \subset \mathbb{C}[M]$ be the corresponding ideal. We consider the resultant map

$$\mathrm{res}_{\mathscr{A}} = \mathrm{res}_{\hat{f}_1, \ldots, \hat{f}_n} : \mathbb{C} \times \cdots \times \mathbb{C} \to V_{\mathscr{A}},$$

where $V_{\mathscr{A}} = \bigoplus_{i=0}^{s} \mathbb{C} \cdot t^{m_i}$. The matrix of this map in the basis $\{t^{m_0}, \ldots, t^{m_s}\}$ for $V_{\mathscr{A}}$ is given by $(\mathrm{res}_{\mathscr{A}})_{ij} = c_{j,i}$ (for convenience, we start indexing the rows of $\mathrm{res}_{\mathscr{A}}$ by 0). We define the map

$$\phi_{\mathscr{A}} : (\mathbb{C}^*)^n \to \mathbb{P}(V_{\mathscr{A}}^\vee) \simeq \mathbb{P}^s \quad \text{given by} \quad t \mapsto (t^{m_0} : \cdots : t^{m_s}).$$

Here we write $\mathbb{P}(V_{\mathscr{A}}^\vee)$ for the projectivization[2] of the $\mathbb{C}$-vector space $V_{\mathscr{A}}^\vee$. Note that an element $w \in \mathbb{P}(V_{\mathscr{A}}^\vee)$ does not define a linear function on $V_{\mathscr{A}}$, but the set $\{\hat{f} \in V_{\mathscr{A}} \mid w(\hat{f}) = 0\}$ is well-defined. Hence, the statements $w(\hat{f}) = 0$ or $w(\hat{f}) \neq 0$ for $\hat{f} \in V_{\mathscr{A}}$ make sense. For the representative $(t^{m_0}, \ldots, t^{m_s}) \in V_{\mathscr{A}}^\vee \simeq \mathbb{C}^{s+1}$ of $\phi_{\mathscr{A}}(t) \in \mathbb{P}^s$ we have

$$\begin{bmatrix} t^{m_0} & \cdots & t^{m_s} \end{bmatrix} \begin{bmatrix} c_{1,0} & \cdots & c_{n,0} \\ \vdots & & \vdots \\ c_{1,s} & \cdots & c_{n,s} \end{bmatrix} = \begin{bmatrix} \hat{f}_1(t) & \cdots & \hat{f}_n(t) \end{bmatrix}.$$

If follows immediately that $\phi_{\mathscr{A}}(t) \circ \mathrm{res}_{\mathscr{A}} = 0$ if and only if $t \in V_{(\mathbb{C}^*)^n}(I)$. Using the fact that $\mathscr{A}$ affinely generates M, one can prove the following result.

Proposition 5.4.1. *The points in $V_{(\mathbb{C}^*)^n}(I)$ are in one-to-one correspondence with the points $w \in \operatorname{im} \phi_{\mathscr{A}} \subset \mathbb{P}^s$ such that $w(\hat{f}) = 0$ for all $\hat{f} \in \operatorname{im} \mathrm{res}_{\mathscr{A}}$.*

Let $u_0, \ldots, u_s$ be homogeneous coordinates on $\mathbb{P}^s$. A point $w = (u_0 : \cdots : u_s) \in \mathbb{P}^s$ is such that $w(\hat{f}) = 0$ for all $\hat{f} \in \operatorname{im} \mathrm{res}_{\mathscr{A}}$ if and only if

$$g_i = c_{i,0} u_0 + \cdots + c_{i,s} u_s = 0, \quad i = 1, \ldots, n. \tag{5.4.1}$$

In order to express the condition $(u_0 : \cdots : u_s) \in \operatorname{im} \phi_{\mathscr{A}}$ from Proposition 5.4.1 in terms of polynomial equations, we need to allow that w lies in the Zariski closure

[2] The *projectivization* of a $\mathbb{C}$-vector space V is $(V \setminus \{0\})/\sim$ where $v \sim w$ if $v = \lambda w$ for some $\lambda \in \mathbb{C}^*$.

$\overline{\operatorname{im}\phi_{\mathscr{A}}} \subset \mathbb{P}^s$. This is exactly the projective toric variety $X_{\mathscr{A}} \subset \mathbb{P}^s$ corresponding to $\mathscr{A}$. A point $u = (u_0 : \cdots : u_s) \in \mathbb{P}^s$ lies on $X_{\mathscr{A}}$ if and only if

$$w \in V_{\mathbb{P}^s}(I_{\mathscr{A}}),$$

where $I_{\mathscr{A}} = I_{\mathbb{C}[\mathbb{P}^s]}(X_{\mathscr{A}})$ is the toric ideal defining $X_{\mathscr{A}}$ (here we replace $\mathscr{A}$ by $\mathscr{A} \times \{1\}$ in order to obtain a homogeneous toric ideal, see Appendix E). In order to put these conditions together, we regard the equations (5.4.1) as equations *on* $X_{\mathscr{A}}$. That is, we consider their images $g_i + I_{\mathscr{A}}$ in the coordinate ring $\mathbb{C}[X_{\mathscr{A}}] = \mathbb{C}[\mathbb{P}^s]/I_{\mathscr{A}}$ of $X_{\mathscr{A}}$. Since the points in $V_{(\mathbb{C}^*)^n}(I)$ correspond to points on $X_{\mathscr{A}}$ on which the g_i vanish, they correspond to points in

$$V_{X_{\mathscr{A}}}(I_L) \subset X_{\mathscr{A}} \quad \text{where} \quad I_L = \langle g_1 + I_{\mathscr{A}}, \ldots, g_n + I_{\mathscr{A}} \rangle \subset \mathbb{C}[X_{\mathscr{A}}].$$

Note that I_L is an ideal generated by *linear forms* in $\mathbb{C}[X_{\mathscr{A}}]_1$ (in the grading induced by the standard grading on $\mathbb{C}[\mathbb{P}^s]$). By the assumption that $\mathscr{A}$ affinely generates the lattice M, the map $\phi_{\mathscr{A}}$ embeds the torus $(\mathbb{C}^*)^n$ in $X_{\mathscr{A}}$, which establishes the chain of inclusions

$$V_{(\mathbb{C}^*)^n}(I) \subset V_{X_{\mathscr{A}}}(I_L) \subset X_{\mathscr{A}}.$$

Since $X_{\mathscr{A}}$ is strictly larger than $\operatorname{im}\phi_{\mathscr{A}}$, the inclusion $V_{(\mathbb{C}^*)^n}(I) \subset V_{X_{\mathscr{A}}}(I_L)$ might be strict.

Example 5.4.1. Let $\mathscr{A} = \Delta_2 \cap M$, $\hat{f}_1 = 1 + t_1 + t_2, \hat{f}_2 = 2 + t_1 + t_2$. Then $X_{\mathscr{A}} = \mathbb{P}^2$, $\mathbb{C}[X_{\mathscr{A}}] = \mathbb{C}[u_0, u_1, u_2]$ and $I_L = \langle u_0 + u_1 + u_2, 2u_0 + u_1 + u_2 \rangle$. We have $V_{(\mathbb{C}^*)^2}(I) = \varnothing$ but $V_{X_{\mathscr{A}}}(I_L) = (0 : 1 : -1)$. △

However, by Proposition 5.4.1 we have the equality $V_{(\mathbb{C}^*)^n}(I) = V_{X_{\mathscr{A}}}(I_L) \cap T_{X_{\mathscr{A}}}$, where $T_{X_{\mathscr{A}}} = \operatorname{im}\phi_{\mathscr{A}}$. If the coefficients $c_{j,i}$ are generic, (5.4.1) defines a linear subvariety of codimension n in $\mathbb{P}^s$. Since $X_{\mathscr{A}}$ has dimension n, we may expect that $V_{X_{\mathscr{A}}}(I_L)$ consists of finitely many points. The number of points in the intersection of an n-dimensional projective variety X and a general linear space of codimension n is what we defined to be its *degree* (Definition 2.2.10). This gives the expected number of points in $V_{X_{\mathscr{A}}}(I_L)$ a nice geometric interpretation: it is the degree of the projective toric variety $X_{\mathscr{A}}$.

Theorem 5.4.1. *The degree of $X_{\mathscr{A}}$ is $n!\mathrm{Vol}_n(\mathrm{Conv}(\mathscr{A}))$.*

Proof. See [Kho92], [Sot11, Subsection 3.1.2] or [Sot17, Lemma 2.11]. □

A corollary of Theorem 5.4.1 is Kushnirenko's theorem, which states that the number of isolated points in $V_{(\mathbb{C}^*)^n}(I)$ is at most $n!\mathrm{Vol}_n(\mathrm{Conv}(\mathscr{A}))$. Kushnirenko's theorem is implied by Theorem 5.1.2 since for a polytope $P \subset \mathbb{R}^n$, $\mathrm{MV}(P, \ldots, P) = n!\mathrm{Vol}_n(P)$.

Example 5.4.2. Consider the Laurent polynomials

$$\hat{f}_1 = 3 - 2t_1 - 2t_2 + t_1 t_2,$$
$$\hat{f}_2(e) = (4 - e) - t_1 - (3 - e)t_2 + t_1 t_2,$$

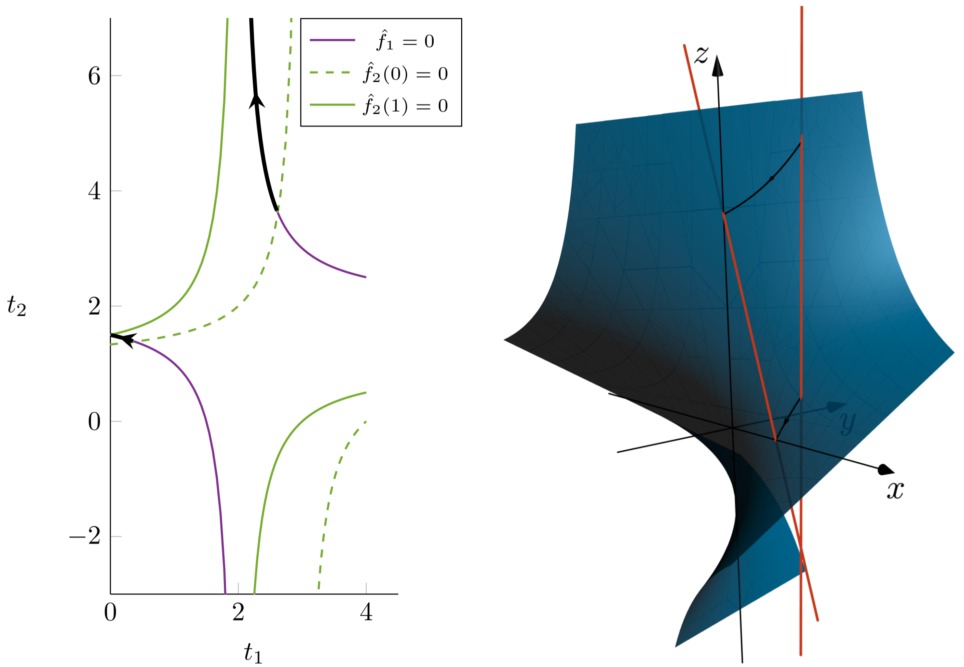


Figure 5.6: Paths traced out by the solutions of $\hat{f}_1 = \hat{f}_2(e) = 0$ from Example 5.4.2 for $e \in [0,1]$ in the torus (left) and on $X_{\mathscr{A}} \simeq \mathbb{P}^1 \times \mathbb{P}^1$ (right).

in the ring $\mathbb{C}[M] = \mathbb{C}[t_1^{\pm 1}, t_2^{\pm 1}]$ where e is a parameter for which we will consider the values $e \in [0,1] \subset \mathbb{R}$. For all values of e, $(\hat{f}_1, \hat{f}_2(e)) \in \mathcal{F}_{\mathscr{A}}$ with $\mathscr{A} = [0,1]^2 \cap M$, and for $e \in [0,1]$, $\mathrm{Newt}(\hat{f}_1) = \mathrm{Newt}(\hat{f}_2(e)) = [0,1]^2 \subset \mathbb{R}^2$. For $e = 0$, there are 2 solutions in the torus, which is the BKK number for $\mathcal{F}_{\mathscr{A}}$. For $e = 1$, $V_{(\mathbb{C}^*)^2}(\hat{f}_1, \hat{f}_2(1)) = \varnothing$. As e increases from 0 to 1, the solutions move out of the torus. For one of them, the t_1-coordinate becomes 0. The other one shoots off to 'infinity'. This is illustrated on the left part of Figure 5.6. The projective toric variety $X_{\mathscr{A}}$ in this example is the closure of the image of the map $\phi_{\mathscr{A}}$ given by $(t_1, t_2) \mapsto (1 : t_1 : t_2 : t_1 t_2)$. This is equal to the image of the Segre embedding $\mathbb{P}^1 \times \mathbb{P}^1 \to \mathbb{P}^3$, which is the variety $V_{\mathbb{P}^3}(u_1 u_2 - u_0 u_3)$ of rank 1 matrices of size 2×2. A picture of this variety in the chart U_2 ($u_2 \neq 0$) is shown in the right part of Figure 5.6 ($X_{\mathscr{A}}$ is the blue surface). Here the coordinates $x = u_0/u_2, y = u_1/u_2, z = u_3/u_2$ were used, and $X_{\mathscr{A}} \cap U_2 \subset U_2 \simeq \mathbb{C}^3$ is described by $V_{\mathbb{C}^3}(y - xz)$. The ideal $I_L \subset \mathbb{C}[X_{\mathscr{A}}]$ in this example is given by

$$I_L(e) = \langle 3u_0 - 2u_1 - 2u_2 + u_3 + I_{\mathscr{A}}, (4-e)u_0 - u_1 - (3-e)u_2 + u_3 + I_{\mathscr{A}} \rangle,$$

where $I_{\mathscr{A}} = \langle u_1 u_2 - u_0 u_3 \rangle$. Hence $V_{X_{\mathscr{A}}}(I_L(e))$ is the intersection of $X_{\mathscr{A}}$ and a line in $\mathbb{P}^3$ that moves as e increases from 0 to 1. This is illustrated in Figure 5.6 by the moving orange line, which traces out two paths on $X_{\mathscr{A}}$. As $e \to 1$, these paths move out of $\operatorname{im} \phi_{\mathscr{A}} \simeq (\mathbb{C}^*)^2$ and they end up in the boundary of $(\mathbb{C}^*)^2$ in $X_{\mathscr{A}}$. $\triangle$

The projective toric variety $X_{\mathscr{A}}$ obtained from the set of lattice points $\mathscr{A}$ may not be a normal variety. This is a property we would like to have, so in general we will need to associate a different toric variety to $\mathscr{A}$. For $\mathscr{A} = \{m_0, \ldots, m_s\} \subset M$ such that $\dim \operatorname{Conv}(\mathscr{A}) = n$, we consider an element $\hat{f} \in \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A})$ and show that it has a well-defined zero set on the normal toric variety X_P associated to the polytope $P = \operatorname{Conv}(A)$. We define

$$\mathscr{T} = \{i \in \{0, \ldots, s\} : m_i \text{ is a vertex of } P\}.$$

By Proposition E.2.4, we know that there is some $\ell \in \mathbb{N}$ such that the dilation ℓP is very ample. We fix such an ℓ (the construction will not depend on which ℓ we choose) and define

$$\mathscr{A}_i = \ell P \cap M - \ell m_i, \quad \text{for all } i \in \mathscr{T}.$$

We obtain the normal affine toric varieties $Y_i = Y_{\mathscr{A}_i}, i \in \mathscr{T}$. Each of these affine toric varieties corresponds to a saturated affine semigroup $\mathsf{S}_i = \mathbb{N}\mathscr{A}_i \subset M$, or equivalently, to a cone $\sigma_i^\vee = \operatorname{Cone}(\mathsf{S}_i) \subset N_{\mathbb{R}}$. For each $i \in \mathscr{T}$, we set

$$f^{\sigma_i} = t^{-m_i}\hat{f} \quad \in \mathbb{C}[\mathsf{S}_i].$$

This gives a function $\hat{f}^{\sigma_i} : Y_i \to \mathbb{C}, i \in \mathscr{T}$. Recall that the affine toric varieties $Y_i, i \in \mathscr{T}$ glue together along the open subsets

$$Y_{ij} = (Y_i)_{t^{m_j - m_i}} = \operatorname{MaxSpec}(\mathbb{C}[\mathsf{S}_i]_{t^{m_j - m_i}})$$

to obtain the toric variety X_P. In the gluing, the open subsets $Y_{ij} \subset Y_i, Y_{ji} \subset Y_j$ are identified via the isomorphisms

$$\phi_{ij} : Y_{ij} \to Y_{ji} \quad \text{given by} \quad \phi_{ij}^* : \mathbb{C}[\mathsf{S}_j]_{t^{m_i - m_j}} \simeq \mathbb{C}[\mathsf{S}_i]_{t^{m_j - m_i}}.$$

Note that $f^{\sigma_i} = \phi_{ij}^*(f^{\sigma_j}/t^{m_i - m_j})$, which implies that for $p \in Y_{ij}$, $f^{\sigma_i}(p) = 0$ if and only if $f^{\sigma_j}(\phi_{ij}(p)) = 0$. In other words, although f^{σ_i} and f^{σ_j} define different functions on Y_{ij} and Y_{ji}, *their zero sets are identified under the gluing*. Let $U_{\sigma_i} \simeq Y_i$ be the open subset of X_P identified with Y_i. We define the *divisor of zeros* of $\hat{f}$ as

$$\operatorname{div}_0(\hat{f}) = \{p \in X_P \mid f^{\sigma_i}(p) = 0 \text{ for any } i \in \mathscr{T} \text{ such that } p \in U_{\sigma_i}\}.$$

It is not hard to check that $\operatorname{div}_0(\hat{f}) \cap (\mathbb{C}^*)^n = V_{(\mathbb{C}^*)^n}(\hat{f})$. Indeed since $(\mathbb{C}^*)^n \subset U_{\sigma_i}$ for all $i \in \mathscr{T}$, for $p \in (\mathbb{C}^*)^n$ we have $p \in \operatorname{div}_0(\hat{f})$ if and only if for any $i \in \mathscr{T}$,

$$f^{\sigma_i}(p) = 0 \Leftrightarrow (t^{-m_i}\hat{f})(p) = 0 \Leftrightarrow \hat{f}(p) = 0.$$

Hence, the divisor of zeros contains $V_{(\mathbb{C}^*)^n}(\hat{f})$ and can be seen as an extension of the relation $\hat{f} = 0$ on $(\mathbb{C}^*)^n$ to a relation on $X_P \supset (\mathbb{C}^*)^n$. For the reader familiar with vector bundles, what we did here was describe the interpretation of $\hat{f}$ as a global section of the line bundle with sheaf of sections $\mathscr{O}_{X_P}(D_P)$ associated to the ample divisor D_P corresponding to P.

It is now straightforward to define a zero set on X_P for members of the family $\mathcal{F}_{\mathbb{C}[M]}(\mathscr{A},\dots,\mathscr{A})$:

$$V_{X_P}(\hat{f}_1,\dots,\hat{f}_n)=\operatorname{div}_0(\hat{f}_1)\cap\cdots\cap\operatorname{div}_0(\hat{f}_n).$$

In this setting, for each $i\in\mathscr{T}$ we get a vector valued function

$$F^{\sigma_i}:Y_i\to\mathbb{C}^n\quad\text{given by}\quad p\mapsto(f_1^{\sigma_i}(p),\dots,f_n^{\sigma_i}(p)).$$

These functions do not glue to a function on X_P, but they have a well defined zero set, which is $V_{X_P}(\hat{f}_1,\dots,\hat{f}_n)$.

Remark 5.4.1. In the case where $\hat{f}=\sum_{i=0}^{s}c_it^{m_i}\in\mathcal{F}_{\mathbb{C}[M]}(\mathscr{A})$ for $\mathscr{A}=P\cap M$ and P is very ample, $\operatorname{div}_0(\hat{f})$ is exactly the zero locus of $c_0u_0+\cdots+c_su_s+I_{\mathscr{A}}$ on $X_{\mathscr{A}}$. The reason is that in this case, $X_{\mathscr{A}}$ is an embedding of X_P. △

5.4.2 Mixed families

Let $\mathscr{A}_1,\dots,\mathscr{A}_n\subset M=\mathbb{Z}^n$ and $P_j=\operatorname{Conv}(\mathscr{A}_j), j=1,\dots,n$. We set $P=P_1+\cdots+P_n$ and assume that $\dim P=n$. We show that a member $(\hat{f}_1,\dots,\hat{f}_n)\in\mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1,\dots,\mathscr{A}_n)$ has a well-defined zero set on the normal toric variety X_P associated to the Minkowski sum P. Note that if $\mathscr{A}_1=\cdots=\mathscr{A}_n$, P is a dilate of each P_j, which implies that P and P_j have the same normal fan, and hence $X_P=X_{P_j}$. This is the same normal toric variety we were considering in the unmixed case.

We first argue that each of the $\hat{f}_j\in\mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_j)$ separately has a well-defined zero set on X_P. For that we will need a result about polytopes. The proof of the next proposition uses some tools from [CLS11, Chapter 6] that were not introduced in this thesis. We include it for completeness and illustrate it with an example. The statement uses the following terminology. We say that a polytope $Q'\subset\mathbb{R}^n$ is an $\mathbb{N}$-*Minkowski summand* of a polytope $Q\subset\mathbb{R}^n$ if there is $Q''\subset\mathbb{R}^n$ such that $Q'+Q''=\ell Q$ for some $\ell\in\mathbb{N}$.

Proposition 5.4.2. *Let $P_j,P\subset M_{\mathbb{R}}=\mathbb{R}^n$ be lattice polytopes. If (and only if) P_j is an $\mathbb{N}$-Minkowski summand of P, then for each full dimensional cone $\sigma\in\Sigma_P(n)$ there is a unique vertex $m_\sigma\in P_j\cap M$ such that $P_j-m_\sigma\subset\sigma^\vee$. Moreover, the cone corresponding to a vertex $m\in P_j\cap M$ in the normal fan Σ_{P_j} of P_j is*

$$\sigma_m=\bigcup_{\substack{\sigma\in\Sigma_P(n)\\ m_\sigma=m}}\sigma.$$

Proof. Since P_j is an $\mathbb{N}$-Minkowski summand of P, P_j corresponds to a torus invariant basepoint free Cartier divisor D_{P_j} on X_P [CLS11, Corollary 6.2.15]. Therefore there exist $(a_\rho)_{\rho\in\Sigma_P(1)}$ such that

$$P_j=\{m\in M_{\mathbb{R}}\mid\langle u_\rho,m\rangle+a_\rho\geq 0,\ \text{for all }\rho\in\Sigma_P(1)\},$$

where u_ρ is the primitive ray generator of the ray $\rho \in \Sigma_P(1)$. This implies, by Theorems 4.2.8 and 6.1.7 in [CLS11] that for each $\sigma \in \Sigma_P(n)$ there is a unique vertex $m_\sigma \in P_j \cap M$ such that $\langle u_\rho, m_\sigma \rangle = a_\rho$ for each $\rho \in \sigma(1)$. It follows easily that $P_j - m_\sigma \subset \sigma^\vee$. If for some (different) lattice point $m \in P_j \cap M$ such that $P_j - m \subset \sigma^\vee$, then both $m_\sigma - m$ and $m - m_\sigma$ are contained in $\sigma^\vee$, but $\sigma^\vee$ is pointed since σ is full-dimensional. We conclude that $m = m_\sigma$. The statement about the normal fan of P is Proposition 6.2.5 in [CLS11]. □

Example 5.4.3. Consider the polytopes P_1, P_2, P shown in Figure 5.7. The normal

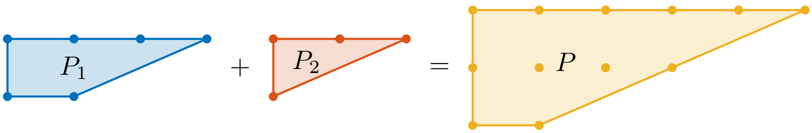


Figure 5.7: Polytopes from Example 5.4.3.

fan Σ_P of P is shown in Figure 5.8, together with a picture of the dual cones of the maximal cones in Σ_P (the cones $\sigma_1^\vee$ and $\sigma_4^\vee$ overlap). Note that $X_{P_2} \neq X_{P_1} = X_P$, since P_2 has a different normal fan. However, P_2 is a Minkowski summand of P, so we can apply proposition 5.4.2. Figure 5.7 defines the polytopes up to translation in the lattice. We fix P_2 as the polytope with vertices $m_1 = (0,0), m_2 = (0,1)$ and $m_3 = (2,1)$. For the maximal cones $\sigma_1, \ldots, \sigma_4 \in \Sigma_P$, we have that the vertices m_σ from Proposition 5.4.2 are given by

$$m_{\sigma_1} = m_1, \quad m_{\sigma_2} = m_2, \quad m_{\sigma_3} = m_3, \quad m_{\sigma_4} = m_1.$$

Moreover, the normal fan Σ_{P_2} looks like Σ_P, but with the cones σ_1 and σ_4 'merged together'. △

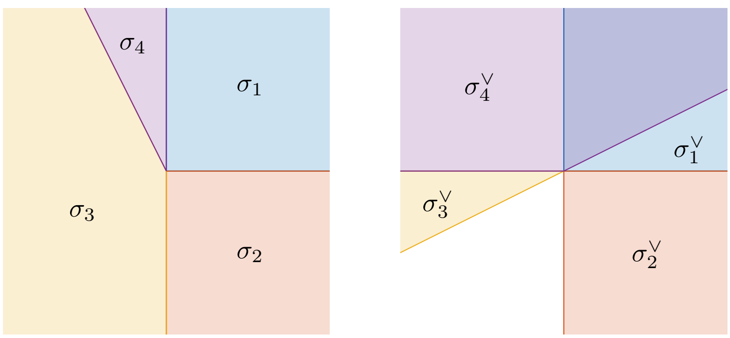


Figure 5.8: Normal fan Σ_P of the polytope P from Example 5.4.3 (left) and the dual cones of the maximal cones in $\Sigma_P(2)$ (right).

As in the unmixed case, let $\mathscr{T}$ be the set indexing the vertices of P and the cones in $\Sigma_P(n)$. For each $j = 1, \dots, n$ and each $i \in \mathscr{T}$, by Proposition 5.4.2 there is a vertex $m_{j,i} \in P_j \cap M$ for which

$$f_j^{\sigma_i} = t^{-m_{j,i}} \hat{f}_j \quad \in \mathbb{C}[\mathsf{S}_i].$$

One can check that this gives again a well-defined zero set

$$\mathrm{div}_0(\hat{f}_j) = \{p \in X_P \mid f_j^{\sigma_i}(p) = 0 \text{ for any } i \in \mathscr{T} \text{ such that } p \in U_{\sigma_i}\}.$$

Doing this for each of the $\hat{f}_j$, we obtain

$$V_{X_P}(\hat{f}_1, \dots, \hat{f}_n) = \mathrm{div}_0(\hat{f}_1) \cap \cdots \cap \mathrm{div}_0(\hat{f}_n).$$

This should be viewed as a natural extension of the relations $\hat{f}_1 = \cdots = \hat{f}_n = 0$ from $(\mathbb{C}^*)^n$ to $X_P \supset (\mathbb{C}^*)^n$. Note that $V_{X_P}(\hat{f}_1, \dots, \hat{f}_n) \cap (\mathbb{C}^*)^n = V_{(\mathbb{C}^*)^n}(\hat{f}_1, \dots, \hat{f}_n)$. In this discussion, each of the $\hat{f}_j$ was viewed as a global section of the line bundle with sheaf of sections $\mathscr{O}_{X_P}(D_{P_j})$, where D_{P_j} is the basepoint free Cartier divisor from the proof of Proposition 5.4.2, and $(\hat{f}_1, \dots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \dots, \mathscr{A}_n)$ as a global section of the rank n vector bundle with sheaf of sections $\mathscr{O}_{X_P}(D_{P_1}) \oplus \cdots \oplus \mathscr{O}_{X_P}(D_{P_n})$.

Remark 5.4.2. The set $V_{X_P}(\hat{f}_1, \dots, \hat{f}_n)$ defined above can be given the structure of a *subscheme* of X, whose local equations in $U_\sigma, \sigma \in \Sigma_P(n)$ are given by $f_1^\sigma, \dots, f_n^\sigma$. For a point $\zeta \in V_{X_P}(\hat{f}_1, \dots, \hat{f}_n) \cap U_\sigma$, the *multiplicity* of $\mathbb{Z}$ as a point of $V_{X_P}(\hat{f}_1, \dots, \hat{f}_n)$ is defined as the multiplicity of the corresponding point in the affine variety $V_{U_\sigma}(f_1^\sigma, \dots, f_n^\sigma)$, see Subsection 3.1.3. △

Example 5.4.4. Let $(\hat{f}_1, \dots, \hat{f}_n) \in \mathcal{F}_R(d_1, \dots, d_n) = \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \dots, \mathscr{A}_n)$ with $\mathscr{A}_i = d_i\Delta_n \cap M$. Then, since the toric variety corresponding to a simplex is the projective space $\mathbb{P}^n$, we find $X_P = \mathbb{P}^n$. Let σ_0 be the cone in Σ_P corresponding to the vertex $(0, \dots, 0) \in \Delta_n$ and let σ_i be the cone corresponding to the vertex $e_i \in \Delta_n \cap M$. Then $f_j^{\sigma_i}$ is exactly the dehomogenization with respect to x_i of $f_j = \eta_{d_j}(\hat{f}_j)$. That is, $f_j^{\sigma_i} = f_j(x_0/x_i, \dots, x_{i-1}/x_i, , 1, x_{i+1}/x_i, \dots, x_n/x_i)$ and

$$\mathbb{C}[\mathsf{S}_i] \simeq \mathbb{C}\left[\frac{x_0}{x_i}, \dots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \dots, \frac{x_n}{x_i}\right].$$

△

Example 5.4.5. Consider the system of Laurent polynomial equations $\hat{f}_1 = \hat{f}_2 = 0$ given by

$$\hat{f}_1 = 1 + t_1 + t_2 + t_1t_2 + t_1^2t_2 + t_1^3t_2,$$
$$\hat{f}_2 = 1 + t_2 + t_1t_2 + t_1^2t_2.$$

We think of this system as a member of $\mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \mathscr{A}_2) = \mathcal{F}_{\mathbb{C}[M]}(P_1, P_2)$ where $\mathbb{C}[M] = \mathbb{C}[t_1^{\pm 1}, t_2^{\pm 1}]$, $\mathscr{A}_i = \mathrm{Conv}(P_i), i = 1, 2$ and P_1, P_2 are the polytopes from Example 5.4.3.

The BKK number for this family is $\mathrm{MV}(P_1, P_2) = 3$. However, there is only one solution in the torus, namely $(t_1, t_2) = (-1, -1)$. We will show that $V_{X_P}(\hat{f}_1, \hat{f}_2)$ consists of 3 points, where X_P is the toric variety associated to the polytope P from Example 5.4.3. In order to do this, let us see what the equations look like on the affine charts U_{σ_1} and U_{σ_4} of X_P. We set $\mathsf{S}_1 = \sigma_1^\vee \cap M = \mathbb{N}\{(1,0),(0,1)\}$ and $\mathsf{S}_4 = \sigma_4^\vee \cap M = \mathbb{N}\{(-1,0),(2,1)\}$ and

$$\begin{aligned} Y_1 &= \mathrm{MaxSpec}(\mathbb{C}[\mathsf{S}_1]) = \mathrm{MaxSpec}(\mathbb{C}[t_1, t_2]) \simeq \mathbb{C}^2, \\ Y_4 &= \mathrm{MaxSpec}(\mathbb{C}[\mathsf{S}_4]) = \mathrm{MaxSpec}(\mathbb{C}[t_1^{-1}, t_1^2 t_2]) = \mathrm{MaxSpec}(\mathbb{C}[u_1, u_2]) \simeq \mathbb{C}^2. \end{aligned}$$

On $Y_1 \simeq \mathbb{C}^2$, the equations remain unchanged ($m_{\sigma_1} = (0,0)$ for both P_1 and P_2 and we are using the coordinates on $(\mathbb{C}^*)^2$ as coordinates on $\mathbb{C}^2$):

$$\begin{aligned} f_1^{\sigma_1} &= 1 + t_1 + t_2 + t_1 t_2 + t_1^2 t_2 + t_1^3 t_2, \\ f_2^{\sigma_1} &= 1 + t_2 + t_1 t_2 + t_1^2 t_2. \end{aligned}$$

We find that $V_{Y_1}(f_1^{\sigma_1}, f_2^{\sigma_1}) = \{(0,-1),(-1,-1)\}$. Hence, next to the point $(-1,-1)$ in the torus, we pick up the point $(0,-1)$ on the boundary of the torus in X_P. This point lies on the one-dimensional torus orbit corresponding to $\sigma_1 \cap \sigma_2$ (see Theorem E.2.3). On $Y_4 \simeq \mathbb{C}^2$, the equations become

$$\begin{aligned} f_1^{\sigma_4} &= u_1 + 1 + u_1^3 u_2 + u_1^2 u_2 + u_1 u_2 + u_2, \\ f_2^{\sigma_4} &= 1 + u_1^2 u_2 + u_1 u_2 + u_2. \end{aligned}$$

We get $V_{Y_4}(f_1^{\sigma_4}, f_2^{\sigma_4}) = \{(0,-1),(-1,-1)\}$. To see how these points are related to the points in $V_{Y_1}(f_1^{\sigma_1}, f_2^{\sigma_1})$ note that the gluing isomorphism $\phi_{14} : Y_{14} \to Y_{41}$ with $Y_{14} = Y_1 \setminus V_{Y_1}(t_1)$ and $Y_{41} = Y_4 \setminus V_{Y_4}(u_1)$ is given by

$$\phi_{14}(t_1, t_2) = (t_1^{-1}, t_1^2 t_2).$$

We see that the point $(-1,-1) \in V_{Y_1}(f_1^{\sigma_1}, f_2^{\sigma_1})$ is mapped to $(-1,-1) \in V_{Y_4}(f_1^{\sigma_4}, f_2^{\sigma_4})$, so these two solutions correspond to the same solution on X_P, but the other solutions $(0,-1) \in Y_1 \setminus Y_{14}$ and $(0,-1) \in Y_4 \setminus Y_{41}$ represent distinct points on X_P. We conclude that we have found 3 points in $V_{X_P}(\hat{f}_1, f_2)$. One of them lies in both U_{σ_1} and U_{σ_4}, one of them lies in U_{σ_1}, but not in U_{σ_4}, and one of them lies in U_{σ_4}, but not in U_{σ_1}.

The toric variety X_P in this example is a *Hirzebruch surface* $\mathscr{H}_2$. We will use this toric variety and this (Laurent) polynomial system as a running example. △

Although it is instructive to see how a system $(\hat{f}_1, \dots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \dots, \mathscr{A}_n)$ defines a subvariety of the abstract toric variety X_P by 'moving the polytopes around' to see the local equations, it would be nice to have a *global description* of the variety $V_{X_P}(\hat{f}_1, \dots, \hat{f}_n)$. For this we need *global coordinates on* X_P. Example 5.4.4 shows that when $X_P = \mathbb{P}^n$, this can be realized by homogenizing the equations. The construction presented and used in the following section generalizes this nicely and enables us to compute 'homogeneous coordinates' of the points defined by $(\hat{f}_1, \dots, \hat{f}_n) \in$

$\mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \ldots, \mathscr{A}_n)$ on X_P via a generalization of homogeneous normal forms as introduced in Section 4.5.

Example 5.4.5 deals with a system for which the BKK number is a strict upper bound: the number of solutions in $(\mathbb{C}^*)^2$ is strictly smaller than the mixed volume of the Newton polytopes. However, taking the boundary of the torus in X_P into account we can see where these 'missing' solutions are. There's a generalization of Theorem 5.1.2 behind this, which nicely demonstrates another way in which X_P is for $\mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \ldots, \mathscr{A}_n)$ what $\mathbb{P}^n$ is for $\mathcal{F}_R(d_1, \ldots, d_n)$ by comparing it to the homogeneous version of Bézout's theorem (Theorem 3.2.2).

Theorem 5.4.2 (Toric BKK theorem)**.** *Let $(\hat{f}_1, \ldots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \ldots, \mathscr{A}_n)$ and let X_P be the toric variety of the polytope $P = P_1 + \cdots + P_n$, where $P_i = \mathrm{Conv}(\mathscr{A}_i)$. If $V_{X_P}(\hat{f}_1, \ldots, \hat{f}_n)$ consists of $\delta^+ < \infty$ points on X_P, counting multiplicities, then δ^+ is given by $\mathrm{MV}(P_1, \ldots, P_n)$. For generic choices of the coefficients of the $\hat{f}_i$, the number of roots in the torus $T_{X_P} \simeq T_N = (\mathbb{C}^*)^n$ is exactly equal to $\mathrm{MV}(P_1, \ldots, P_n)$ and they all have multiplicity one.*

Proof. See [Ful93, §5.5]. □

5.5 Cox rings and homogeneous normal forms

Although the variety $V_{(\mathbb{C}^*)^n}(\hat{f}_1, \ldots, \hat{f}_n)$ of a member $(\hat{f}_1, \ldots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \ldots, \mathscr{A}_n)$ may *not* consist of the BKK number many isolated points, this will be true for the system obtained by applying the slightest random perturbation to the nonzero coefficients of the $\hat{f}_i$ (such that the resulting system still lives in the same family), see Theorem 5.1.2. In fact, this is true for the more general perturbations for which the system does not leave the possibly larger family $\mathcal{F}_{\mathbb{C}[M]}(P_1, \ldots, P_n)$, where $P_i = \mathrm{Conv}(\mathscr{A}_i)$. Two ways in which such a perturbation may enlarge the number of isolated solutions in $(\mathbb{C}^*)^n$ are:

1. a solution with multiplicity $\mu > 1$ breaks up into μ isolated solutions,
2. a positive dimensional component of $V_{(\mathbb{C}^*)^n}(\hat{f}_1, \ldots, \hat{f}_n)$ breaks up into a number of isolated solutions.

This corresponds to two types of 'non-genericity' for a member $(\hat{f}_1, \ldots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(\mathscr{A}_1, \ldots, \mathscr{A}_n)$. The first one does not pose a problem for computing multiplication matrices, and there are ways for obtaining the coordinates from these matrices (see the discussion at the end of Subsection 4.3.2). The second phenomenon will not be our focus in this thesis, although there are ways of dealing with positive dimensional components using TNFs [MTVB19, Section 3]. The type of non-genericity we will address in this section is one that comes from the toric interpretation of the BKK theorem (Theorem 5.4.2). Let $P = P_1 + \cdots + P_n$. The number of solutions in $(\mathbb{C}^*)^n$ may increase upon perturbing $\hat{f}_1, \ldots, \hat{f}_n$ if

3. one or more solutions in $V_{X_P}(\hat{f}_1, \ldots, \hat{f}_n) \setminus (\mathbb{C}^*)^n$ move out of the boundary, into the torus.

We will focus on the case where $V_{X_P}(\hat{f}_1, \ldots, \hat{f}_n)$ consists of finitely many points on X_P. This means we do not allow positive dimensional components, but we do allow solutions on the boundary $X_P \setminus (\mathbb{C}^*)^n$. We have studied this case for $X_P = \mathbb{P}^n$ in the previous chapters to develop methods for computing homogeneous coordinates of isolated solutions. This made it possible to deal with solutions on the boundary in a robust way, especially with solutions on the part of the boundary corresponding to the hyperplane 'at infinity', see Section 4.5. For this we exploited the fact that $V_{\mathbb{P}^n}(\hat{f}_1, \ldots, \hat{f}_n)$ has a *global description* given by $V_{\mathbb{P}^n}(I)$ where $I = \langle f_1, \ldots, f_s \rangle \subset S$ is a homogeneous ideal generated by $f_i = \eta_{d_i}(\hat{f}_i) \in S$. A point $\zeta \in V_{\mathbb{P}^n}(I)$ can be described by a set of homogeneous coordinates, which is given by a point in $\mathbb{C}^{n+1} \setminus \{0\}$ in the fiber of ζ (i.e. the inverse image of ζ) under

$$\pi : \mathbb{C}^{n+1} \setminus \{0\} \to \mathbb{P}^n \quad \text{with} \quad \pi(x_0, \ldots, x_n) = (x_0 : \cdots : x_n).$$

This map π has the property that all fibers are orbits of the group action

$$(\mathbb{C}^*) \times (\mathbb{C}^{n+1} \setminus \{0\}) \to (\mathbb{C}^{n+1} \setminus \{0\}) \quad \text{given by} \quad \lambda \cdot (x_0, \ldots, x_n) = (\lambda x_0, \ldots, \lambda x_n),$$

under which the affine variety $V_{\mathbb{C}^{n+1}}(I)$ is stable. The sets of homogeneous coordinates for all points in $V_{\mathbb{P}^n}(I)$ can be obtained via eigenvalue computations (Subsection 3.2.2). In this section, we will generalize this in the following ways. The toric variety X_P can be constructed as the image of a map $\pi : \mathbb{C}^k \setminus Z \to X_P$ whose fibers on an open subset $U \subset X_P$ are orbits of an algebraic group action $G \times (\mathbb{C}^k \setminus Z) \to (\mathbb{C}^k \setminus Z)$. The equations $\hat{f}_1, \ldots, \hat{f}_n$ can be *homogenized* to obtain homogeneous elements $f_1, \ldots, f_n$ in a graded ring S, where the grading is such that for a homogeneous element $f \in S$, $V_{\mathbb{C}^k}(f)$ is stable under the G-action (which extends to $\mathbb{C}^k$). This will be the subject of Subsection 5.5.1. After that, in Subsection 5.5.2 we describe a notion of *regularity* for homogeneous ideals in S. This will be the right notion to use for generalizing the projective eigenvalue, eigenvector theorem to the toric setting, which we do in Subsection 5.5.3. Finally, in subsection 5.5.4 we describe homogeneous normal forms in this context, show how they can be used for computing homogeneous multiplication maps and provide an algorithm for computing homogeneous coordinates of the points in $V_{X_P}(\hat{f}_1, \ldots, \hat{f}_n)$. For all this, we will make the assumption that all points in $V_{X_P}(\hat{f}_1, \ldots, \hat{f}_n)$ have multiplicity 1 for simplicity. All results extend to the case with higher multiplicities. This, together with some results regarding the regularity will be discussed in Subsection 5.5.5. The results of this section are strongly based on the paper [Tel20] and on a recent collaboration of the author with Matías Bender [BT20a].

5.5.1 The Cox ring of a complete toric variety

In this subsection we describe the construction of a toric variety X as the image of a quotient map

$$\pi : \mathbb{C}^k \setminus Z \to X,$$

where $Z \subset \mathbb{C}^k$ is a subvariety and π is invariant under an algebraic group action $G \times (\mathbb{C}^k \setminus Z) \to (\mathbb{C}^k \setminus Z)$. This construction is described by Cox in [Cox95], and it is referred to as the *Cox construction*. We should mention that the result had been described in the analytic category by Audin, Delzant and Kirwan, see [Aud12, Chapter 6] and references therein. The reader may be familiar with the construction for $X = \mathbb{P}^n$, where $k = n+1$, $Z = \{0\}$ and $G = \mathbb{C}^*$. Some background in toric geometry beyond Appendix E is assumed in this subsection. The reader is referred to [CLS11, Chapters 1-4] or [Ful93].

Consider the algebraic torus $(\mathbb{C}^*)^n$ of dimension n. Its character and cocharacter lattices are denoted by $M = \mathrm{Hom}_{\mathbb{Z}}((\mathbb{C}^*)^n, \mathbb{C}^*) \simeq \mathbb{Z}^n$ and $N = \mathrm{Hom}_{\mathbb{Z}}(M, \mathbb{Z}) \simeq \mathbb{Z}^n$ respectively. Let $\Sigma = \Sigma_P$ be the normal fan in $N_{\mathbb{R}}$ of a full dimensional lattice polytope $P \subset M_{\mathbb{R}} = M \otimes_{\mathbb{Z}} \mathbb{R}$. We will denote the set of cones of dimension d in Σ by $\Sigma(d)$. The corresponding toric variety X is compact.[3] We will sometimes denote $X = X_\Sigma = X_P$ to emphasize the correspondence between X and its fan or polytope. Before introducing the Cox construction for general compact X, we will work out the example of $X = \mathbb{P}^2$.

Example 5.5.1. The projective plane $\mathbb{P}^2$ is defined as

$$\mathbb{P}^2 = \frac{\mathbb{C}^3 \setminus \{0\}}{\mathbb{C}^*}, \tag{5.5.1}$$

where the quotient is by the action $\mathbb{C}^* \times (\mathbb{C}^3 \setminus \{0\}) \to (\mathbb{C}^3 \setminus \{0\})$ given by $(\lambda, (x_1, x_2, x_3)) \mapsto (\lambda x_1, \lambda x_2, \lambda x_3)$. This action extends trivially to an action on $\mathbb{C}^3$. Subvarieties of $\mathbb{P}^2$ are given by homogeneous ideals in the polynomial ring $S = \mathbb{C}[x_1, x_2, x_3]$. Here 'homogeneous' is with respect to the $\mathbb{Z}$-grading

$$S = \bigoplus_{\alpha \in \mathbb{Z}} S_\alpha,$$

which is such that for $f \in S$ homogeneous, $V_{\mathbb{C}^3}(f)$ is stable under the $\mathbb{C}^*$-action. Equivalently, $V_{\mathbb{C}^3}(f)$ is a union of $\mathbb{C}^*$-orbits. In the ring S, the ideal $\mathfrak{B} = \langle x_1, x_2, x_3 \rangle$ plays a special role: its variety in $\mathbb{P}^2$ is the empty set. The interplay between the algebra and geometry in this construction is illustrated by the following table.

Algebra		Geometry
S	$\xrightarrow{\mathrm{MaxSpec}(\cdot)}$	$\mathbb{C}^3$
$\mathfrak{B}$	$\xrightarrow{V_{\mathbb{C}^3}(\cdot)}$	$\{0\}$
$\mathbb{Z}$	$\xrightarrow{\mathrm{Hom}_{\mathbb{Z}}(\cdot, \mathbb{C}^*)}$	$\mathbb{C}^*$

[3]The construction presented here works for more general normal toric varieties coming from fans Σ whose rays span $\mathbb{R}^n$ [Cox95].

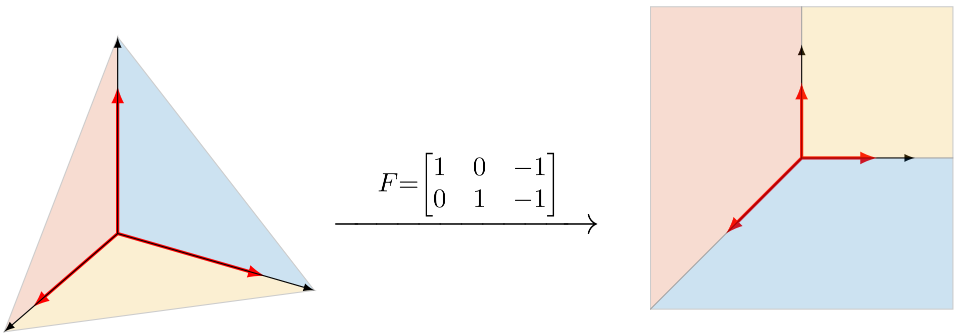


Figure 5.9: An illustration of the $\mathbb{Z}$-linear map $F : N' \to N$ from Example 5.5.1. The ray generators of $\Sigma'(1), \Sigma(1)$ are depicted as red arrows and the two dimensional cones are colored in blue, orange and yellow.

For the purpose of generalizing this construction, we make the following observation. The quotient (5.5.1) comes from a toric morphism $\pi : \mathbb{C}^3 \setminus \{0\} \to \mathbb{P}^2$ which is constant on $\mathbb{C}^*$-orbits. A toric morphism comes from a lattice homomorphism $N' \to N$ that is *compatible* with fans Σ' and Σ in $N'_{\mathbb{R}}$ and $N_{\mathbb{R}}$ respectively (see [CLS11, Section 3.3]). In our case Σ' is the fan of $\mathbb{C}^3 \setminus \{0\}$ and Σ is the fan of $\mathbb{P}^2$. The lattices are $N' = \mathbb{Z}^3$ and $N = \mathbb{Z}^2$, and the morphism π comes from $F : N' \to N$ where F is a 2×3 integer matrix whose columns are the primitive ray generators of $\Sigma(1)$. The fans and the matrix F are shown in Figure 5.9. The *compatibility* of the map F with the fans Σ' and Σ comes down to the fact that each cone of Σ' is mapped (under the $\mathbb{R}$-map $F_{\mathbb{R}} = F \otimes_{\mathbb{Z}} \mathbb{R}$ associated to F) into a cone of Σ. In Figure 5.9 the 2-dimensional cones have matching colors according to this association. Note that the three dimensional cone $\sigma = \mathrm{Cone}(e_1, e_2, e_3)$ of the positive orthant in $\mathbb{R}^3$ is not mapped to a cone of Σ under $F_{\mathbb{R}}$. Therefore, this cone does not belong to Σ'. Taking this three dimensional cone out of the positive orthant corresponds to taking the origin out of $\mathbb{C}^3$. Hence $\mathbb{C}^3 \setminus \{0\} = X_{\Sigma'}$. △

In what follows, it is instructive to keep Example 5.5.1 in mind as a reference. Let $\Sigma(1) = \{\rho_1, \dots, \rho_k\}$ and let $u_i \in N$ be the primitive ray generator of ρ_i. We collect the u_i in a matrix

$$F = [u_1 \ \cdots \ u_k] \in \mathbb{Z}^{n \times k}.$$

This gives a lattice homomorphism $F : N' \to N$ where $N' = \mathbb{Z}^k$. Consider the fan given by the positive orthant in $\mathbb{R}^k$ and all its faces. We let Σ' be the subfan of all the cones whose image under $F_{\mathbb{R}}$ is contained in a cone of Σ. By construction, the lattice homomorphism F is compatible with the fans Σ' and Σ in $N'_{\mathbb{R}}$ and $N_{\mathbb{R}}$ respectively. It follows that F gives a toric morphism $\pi : X_{\Sigma'} \to X_{\Sigma}$, where $X_{\Sigma'} = \mathbb{C}^k \setminus Z$ and Z is a union of coordinate subspaces. We now give a description of Z as the affine variety of a radical monomial ideal. Let $S = \mathbb{C}[x_1, \dots, x_k]$ be the coordinate ring of $\mathbb{C}^k$ and for each $\sigma \in \Sigma$, consider the monomial

$$x^{\hat{\sigma}} = \prod_{\rho_i \not\subset \sigma} x_i,$$

where the product ranges over all $i \in \{1, \ldots, k\}$ such that $\rho_i \not\subset \sigma$. Then $Z = V_{\mathbb{C}^k}(\mathfrak{B})$ with

$$\mathfrak{B} = \left\langle x^{\hat{\sigma}} \mid \sigma \in \Sigma \right\rangle = \left\langle x^{\hat{\sigma}} \mid \sigma \in \Sigma(n) \right\rangle.$$

The morphism π is an extension of a map of tori $\pi_{|(\mathbb{C}^*)^k} : (\mathbb{C}^*)^k \to (\mathbb{C}^*)^n$, which has an easy description based on the matrix F. It is given by the Laurent monomial map

$$\pi|_{(\mathbb{C}^*)^k} = F \otimes_{\mathbb{Z}} \mathbb{C}^* : (\mathbb{C}^*)^k \to (\mathbb{C}^*)^n \quad \text{where} \quad (z_1, \ldots, z_k) \mapsto (z^{F_{1,:}}, \ldots, z^{F_{n,:}}). \tag{5.5.2}$$

This uses the short notation $z^a = z_1^{a_i} \cdots z_k^{a_k}$ and $F_{i,:}$ for the i-th row of F. The kernel of $\pi|_{(\mathbb{C}^*)^k}$ (as a group homomorphism) is given by

$$G = \{g \in (\mathbb{C}^*)^k \mid g^{F_{1,:}} = \cdots = g^{F_{n,:}} = 1\}. \tag{5.5.3}$$

This is a subgroup $G \subset (\mathbb{C}^*)^k$ which acts on $\mathbb{C}^k$ by

$$(g_1, \ldots, g_k) \cdot (x_1, \ldots, x_k) \mapsto (g_1 x_1, \ldots, g_k x_k)$$

(this is the restriction of the action of $(\mathbb{C}^*)^k$ on $\mathbb{C}^k$ to G) and the morphism π is constant on G-orbits in $\mathbb{C}^k \setminus Z$. The following theorem uses some terminology for GIT (geometric invariant theory) quotients from [CLS11, Section 5.0].

Theorem 5.5.1. *The morphism $\pi : \mathbb{C}^k \setminus Z \to X_\Sigma$ coming from $F = [u_1 \ \cdots \ u_k]$ is an almost geometric quotient for the action of G on $\mathbb{C}^k \setminus Z$. Moreover, the open subset $U \subset X_\Sigma$ for which $\pi|_{\pi^{-1}(U)}$ is a geometric quotient is such that $(X_\Sigma \setminus U)$ has codimension at least 3 in X_Σ.*

Proof. See [Cox95, Theorem 2.1] or [CLS11, Theorem 5.1.11]. □

Here is a longer, equivalent formulation of Theorem 5.5.1 which uses less terminology.

Theorem 5.5.2. *Consider the action of the group G in (5.5.3) on $\mathbb{C}^k \setminus Z$. There is a one-to-one correspondence*

$$\{ \text{ closed } G\text{-orbits in } \mathbb{C}^k \setminus Z \ \} \leftrightarrow \{ \text{ points in } X_\Sigma \ \}.$$

Moreover, there is an open subset $U \subset X_\Sigma$ which is such that $\operatorname{codim}_{X_\Sigma}(X_\Sigma \setminus U) \geq 3$ *for which there is a one-to-one correspondence*

$$\{ \ G\text{-orbits in } \pi^{-1}(U) \ \} \leftrightarrow \{ \text{ points in } U \ \}.$$

These correspondences are realized by the toric morphism $\pi : \mathbb{C}^k \setminus Z \to X_\Sigma$ coming from $F = [u_1 \ \cdots \ u_k]$.

Remark 5.5.1. The open subset $U \subset X_\Sigma$ in Theorems 5.5.1 and 5.5.2 is the toric variety corresponding to the largest simplicial[4] subfan of Σ, see for instance the proof of Theorem 5.1.11 in [CLS11]. The fact that $X_\Sigma \setminus U$ has codimension at least 3 in X_Σ corresponds to the fact that all cones of dimension ≤ 2 are simplicial. If Σ is a simplicial fan, then $\pi : \mathbb{C}^k \setminus Z \to X_\Sigma$ is a *geometric quotient*, meaning that the nicest possible correspondence holds: G-orbits in $\mathbb{C}^k \setminus Z$ are points in X_Σ. △

Example 5.5.2. The matrix F, the variety Z and the ideal $\mathfrak{B}$ for $X_\Sigma = \mathbb{P}^2$ were given in Example 5.5.1. One can check that $\pi_{|(\mathbb{C}^*)^3}$ is given by $(t_1, t_2, t_3) \mapsto (t_1 t_3^{-1}, t_2 t_3^{-1})$ with kernel $G = \{(g_1, g_2, g_3) \in (\mathbb{C}^*)^3 \mid g_1 = g_2 = g_3\} \simeq \mathbb{C}^*$. The (real part of the) closure of three G-orbits in $\mathbb{C}^3$ are shown in Figure 5.10. This corresponds to the familiar fact that points in $\mathbb{P}^2$ are lines through the origin in $\mathbb{C}^3$. We now consider the

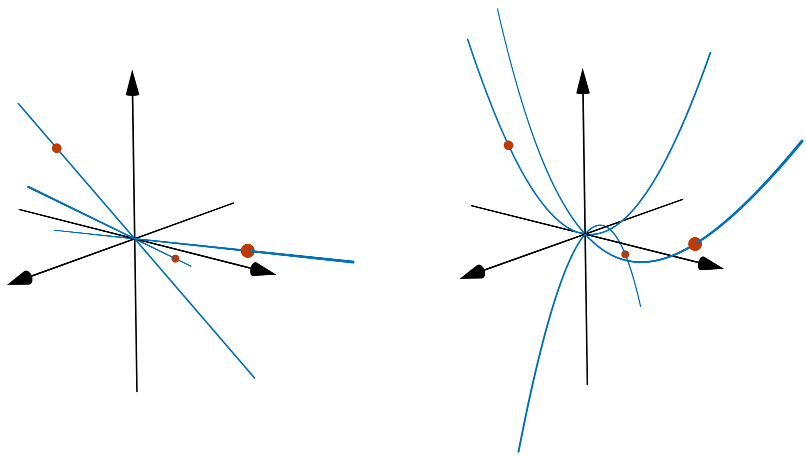


Figure 5.10: Real G-orbits (closures in $\mathbb{R}^3$) of three points (orange dots) in the quotient construction of $\mathbb{P}^2$ (left) and $\mathbb{P}_{(1,2,1)}$ (right).

complete fan in $\mathbb{R}^2$ whose rays are given by

$$F = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -2 \end{bmatrix}.$$

For this example $Z = \{0\}$ and $G = \{(\lambda, \lambda^2, \lambda) \mid \lambda \in \mathbb{C}^*\} \simeq \mathbb{C}^*$. Some orbits are shown in the right part of Figure 5.10. This is the toric variety corresponding to the *weighted projective space* $\mathbb{P}_{(1,2,1)}$. The figure suggests that we can think of points in $\mathbb{P}_{(1,2,1)}$ as 'curves through the origin in $\mathbb{C}^3$'. △

In order to associate the ring S (with its distinguished ideal $\mathfrak{B}$) to our toric variety X_Σ, we will equip it with a grading such that homogeneous elements in S define varieties in $\mathbb{C}^k$ which are stable under the action of G. The grading will be by the

[4] A convex polyhedral cone is *simplicial* if it is generated by an $\mathbb{R}$-linearly independent set. A fan is simplicial if all its cones are. See [CLS11, Definitions 1.2.16 and 3.1.18].

(divisor) class group $\mathrm{Cl}(X_\Sigma)$ of X_Σ, which is the group of Weil divisors modulo linear equivalence [CLS11, Chapter 4] (this group is sometimes written as $A_{n-1}(X)$, see for instance [EH16, Section 1.2]). For toric varieties, the class group is easy to describe explicitly. Let $D_1, \ldots, D_k$ be the torus invariant prime divisors on X_Σ corresponding to $\rho_1, \ldots, \rho_k \in \Sigma(1)$ respectively. These are the closures of the codimension 1 torus orbits of X_Σ, see Theorem E.2.3. The divisors $D_1, \ldots, D_k$ generate the free group of *torus invariant Weil divisors*

$$\mathrm{Div}_T(X_\Sigma) = \{\sum_{i=1}^k a_i D_i \mid a_i \in \mathbb{Z}\} \simeq \mathbb{Z}^k.$$

Characters $m \in M$ give rational functions t^m on X_Σ whose divisor is given by

$$\mathrm{div}(t^m) = \sum_{i=1}^k \langle u_i, m\rangle D_i \in \mathrm{Div}_T(X_\Sigma),$$

see [Ful93, page 61]. Identifying $\mathrm{Div}_T(X_\Sigma) \simeq \mathbb{Z}^k$, there is a short exact sequence [Ful93, page 63]

$$0 \longrightarrow M \xrightarrow{F^\top} \mathbb{Z}^k \longrightarrow \mathrm{Cl}(X_\Sigma) \longrightarrow 0 \tag{5.5.4}$$

where the first map is $F^\top = \mathrm{div}$ and the second map takes a torus invariant divisor to its class in $\mathrm{Cl}(X_\Sigma)$. Note that taking $\mathrm{Hom}_\mathbb{Z}(-, \mathbb{C}^*)$ of (5.5.4) gives us back the map of tori $\pi_{|(\mathbb{C}^*)^k} : (\mathbb{C}^*)^k \to (\mathbb{C}^*)^n$ from the geometric construction discussed above. This shows that the group G is $G = \mathrm{Hom}_\mathbb{Z}(\mathrm{Cl}(X_\Sigma), \mathbb{C}^*) \subset (\mathbb{C}^*)^k$. The sequence (5.5.4) shows that $\mathrm{Cl}(X_\Sigma) \simeq \mathbb{Z}^k / \operatorname{im} F^\top$ and every element of $\mathrm{Cl}(X_\Sigma)$ can be written as the class $[D]$ of some torus invariant divisor $D = \sum_{i=1}^k a_i D_i \in \mathrm{Div}_T(X_\Sigma)$. For an element $\alpha = [\sum_{i=1}^k a_i D_i] \in \mathrm{Cl}(X_\Sigma)$, we define the $\mathbb{C}$-vector subspace

$$S_\alpha = \bigoplus_{F^\top m + a \geq 0} \mathbb{C} \cdot x^{F^\top m + a} \subset S,$$

where the sum ranges over all $m \in M$ satisfying $\langle u_i, m\rangle + a_i \geq 0$ (here $\langle \cdot, \cdot \rangle$ denotes the usual pairing between $N \simeq \mathbb{Z}^n$ and its dual $M \simeq \mathbb{Z}^n$), for $i = 1, \ldots, k$. One can check that this definition is independent of the chosen representative for α: setting $a' = a + F^\top m'$ for some $m' \in M$ gives the same vector subspace S_α. We consider the grading

$$S = \bigoplus_{\alpha \in \mathrm{Cl}(X_\Sigma)} S_\alpha \tag{5.5.5}$$

on the ring S. The ring S with its *irrelevant ideal* $\mathfrak{B}$ and the grading (5.5.5) is called the *homogeneous coordinate ring, total coordinate ring* or *Cox ring* of X_Σ. If $f = \sum_{F^\top m + a \geq 0} c_m x^{F^\top m + a} \in S_\alpha$ is homogeneous of degree α, then for $g \in G \subset (\mathbb{C}^*)^k$ we have

$$f(g \cdot x) = \sum_{F^\top m + a \geq 0} c_m (g \cdot x)^{F^\top m + a} = g^a f(x),$$

where we use that $g^{F^\top m} = 1$ by definition of G. It follows that the set

$$V_{X_\Sigma}(f) = \{\zeta \in X_\Sigma \mid f(z) = 0 \text{ for some } z \in \pi^{-1}(\zeta)\}$$

is well defined. The generalized definition for homogeneous ideals $I \subset S$ is

$$V_{X_\Sigma}(I) = \{\zeta \in X_\Sigma \mid f(z) = 0 \text{ for some } z \in \pi^{-1}(\zeta) \text{ for all } f \in I\}.$$

The set $V_{X_\Sigma}(I)$ has a scheme structure and we will say more about the local defining equations soon. We generalize the table from Example 5.5.1 for compact toric varieties X_Σ and add some terminology.

	Algebra		Geometry	
Cox ring	S	$\xrightarrow{\text{MaxSpec}(\cdot)}$	$\mathbb{C}^k$	total coordinate space
irrelevant ideal	$\mathfrak{B}$	$\xrightarrow{V_{\mathbb{C}^k}(\cdot)}$	Z	base locus
class group	$\text{Cl}(X_\Sigma)$	$\xrightarrow{\text{Hom}_\mathbb{Z}(\cdot,\mathbb{C}^*)}$	G	reductive group

We point out that under our assumption that Σ is complete, all of the graded pieces $S_\alpha, \alpha \in \text{Cl}(X_\Sigma)$ are finite dimensional $\mathbb{C}$-vector spaces [CLS11, Proposition 4.3.8].

Remark 5.5.2. In this construction, there is a one-to-one correspondence between

1. the variables $x_1, \ldots, x_k$ of S,
2. the rays $\rho_1, \ldots, \rho_k$ of $\Sigma(1)$,
3. the columns $u_1, \ldots, u_k$ of F,
4. the torus invariant prime divisors $D_1, \ldots, D_k$,
5. the facets of P.

We have that $D_i = V_{X_\Sigma}(x_i)$ and $\pi(x) \in D_i \Leftrightarrow x_i = 0$. △

Example 5.5.3. Let $X_\Sigma = \mathbb{P}^2$. The class group $\text{Cl}(\mathbb{P}^2)$ is given by

$$\mathbb{Z}^3 \Big/ \text{im} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix} \simeq \mathbb{Z}\,.$$

Using, for instance, the identification $\mathbb{Z}^3/\operatorname{im} F^\top \to \mathbb{Z}$ given by $(a_1, a_2, a_3) + \operatorname{im} F^\top = (0, 0, a_1 + a_2 + a_3) + \operatorname{im} F^\top \mapsto a_1 + a_2 + a_3 \in \mathbb{Z}$ (the divisors $a_1 D_1 + a_2 D_2 + a_3 D_3$ and $(a_1 + a_2 + a_3) D_3$ are linearly equivalent), we see that the $\mathbb{Z}$-grading on S is the usual grading of the homogeneous coordinate ring of $\mathbb{P}^2$: $\deg(x_1^{a_1} x_2^{a_2} x_3^{a_3}) = a_1 + a_2 + a_3$ and the graded piece

$$S_{[dD_3]} = \bigoplus_{\substack{m_1 \geq 0 \\ m_2 \geq 0 \\ d - m_1 - m_2 \geq 0}} \mathbb{C} \cdot x_1^{m_1} x_2^{m_2} x_3^{d - m_1 - m_2}$$

is spanned by monomials of ‘degree’ d, in the classical sense. △

Example 5.5.4. Consider the Hirzebruch surface $\mathscr{H}_2$ (see Example 5.4.5). The associated fan Σ is shown in Figure 5.8. The matrix F is

$$F = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 2 & -1 \end{bmatrix}.$$

The Cox ring $S = \mathbb{C}[x_1, x_2, x_3, x_4]$ is graded by $\mathrm{Cl}(\mathscr{H}_2) \simeq \mathbb{Z}^4/\operatorname{im} F^\top \simeq \mathbb{Z}^2$, with $\deg(x^a) = \deg(x_1^{a_1} x_2^{a_2} x_3^{a_3} x_4^{a_4}) = (a_1 - 2a_2 + a_3, a_2 + a_4)$. The reductive group and base locus are given by $G = \{(\lambda, \mu, \lambda, \lambda^2\mu) \mid (\lambda, \mu) \in (\mathbb{C}^*)^2\} \subset (\mathbb{C}^*)^4$ and $Z = V_{\mathbb{C}^4}(x_1, x_3) \cup V_{\mathbb{C}^4}(x_2, x_4) \subset \mathbb{C}^4$ respectively. Since $\mathscr{H}_2$ is smooth, it is simplicial (in the notation from above $U = \mathscr{H}_2$). $\triangle$

With this information, we are able to get more insight in Theorems 5.5.1 and 5.5.2 by explicitly describing the restriction of π to the affine subvarieties in an affine open cover of $\mathbb{C}^k \setminus Z$. For a cone $\sigma \in \Sigma$, consider the cone $\sigma' = \mathrm{Cone}(e_i \mid \rho_i \in \sigma) \subset \mathbb{R}^k$. Note that $(F_\mathbb{R})_{|\sigma'}$ sends σ' into σ, so $\sigma' \in \Sigma'$ and $(F_\mathbb{R})_{|\sigma'} : \sigma' \to \sigma$ gives a morphism of affine toric varieties $\pi_\sigma = \pi_{|U_{\sigma'}} : U_{\sigma'} \to U_\sigma \subset X_\Sigma$. Here

$$U_{\sigma'} = (\mathbb{C}^k)_{x^{\hat\sigma}} = \{x \in \mathbb{C}^k \mid x^{\hat\sigma} \neq 0\}$$

and by construction $X_\Sigma \setminus Z = \bigcup_{\sigma \in \Sigma} U_{\sigma'}$. The map π_σ corresponds to a map of coordinate rings

$$\pi_\sigma^* : \mathbb{C}[U_\sigma] \to \mathbb{C}[U_{\sigma'}] = S_{x^{\hat\sigma}} \quad \text{given by} \quad t^m \mapsto x^{F^\top m}.$$

In particular, in the grading on $S_{x^{\hat\sigma}}$ induced by the grading (5.5.5) on S, we see that π_σ^* factors as $\pi_\sigma^* : \mathbb{C}[U_\sigma] \xrightarrow{\sim} (S_{x^{\hat\sigma}})_0 \to S_{x^{\hat\sigma}}$ (see the proof of Theorem 5.1.11 in [CLS11]). For the reader who is familiar with invariant theory, we note that since the elements of degree 0 in $S_{x^{\hat\sigma}}$ are precisely the G-invariant elements [CLS11, Exercise 5.3.1], the morphism π_σ corresponds to the inclusion $\mathbb{C}[U_{\sigma'}]^G \to \mathbb{C}[U_{\sigma'}]$, which shows that U_σ is a GIT quotient of $U_{\sigma'}$ by the action of G.

Now that we understand the coordinate rings of the affine charts of X_Σ, we are ready to discuss (de-)homogenization. For some degrees $\alpha \in \mathrm{Cl}(X_\Sigma)$, there is a nice, canonical way of *dehomogenizing* homogeneous elements $f \in S_\alpha$ to obtain an element $f^\sigma \in \mathbb{C}[U_\sigma]$ for each $\sigma \in \Sigma_P(n)$. These degrees are the classes of special divisors, called *Cartier divisors.*

Definition 5.5.1 (Cartier divisors and the Picard group)**.** A torus invariant divisor $D = \sum_{i=1}^k a_i D_i \in \mathrm{Div}_T(X_\Sigma)$ is called *Cartier* if it is *locally principal* (see [CLS11, Definition 4.0.12]). Equivalently, D is Cartier if for each $\sigma \in \Sigma$ there is $m_\sigma \in M$ such that $\langle u_i, m_\sigma\rangle + a_i = 0$ for all i such that $\rho_i \in \sigma$. Moreover, for $\sigma \in \Sigma(n)$, m_σ is unique (see [CLS11, Theorem 4.2.8] or [Ful93, §3.3]). The *Picard group* $\mathrm{Pic}(X_\Sigma) \subset \mathrm{Cl}(X_\Sigma)$ is the group of Cartier divisors modulo linear equivalence.

For each $\alpha \in \mathrm{Pic}(X_\Sigma)$ and each $\sigma \in \Sigma(n)$, take any representative $\alpha = [\sum_{i=1}^k a_i D_i] = [D]$ and let $m_\sigma \in M$ be as in Definition 5.5.1. We define

$$x^{\hat\sigma,\alpha} = x^{F^\top m_\sigma + a}$$

(note that this doesn't depend on the choice of representative). For $f \in S_\alpha$ and $\sigma \in \Sigma(n)$, we set

$$f^\sigma = \frac{f}{x^{\hat{\sigma},\alpha}} \quad \in (S_{x^{\hat{\sigma}}})_0 = \mathbb{C}[U_\sigma]. \tag{5.5.6}$$

It is instructive to check that for $X_\Sigma = \mathbb{P}^n$, this corresponds to what we defined as *dehomogenization* η_α^{-1} to the affine charts $U_0, \ldots, U_n$.

We now discuss the 'inverse' operation of *homogenization*. For that we consider the following scenario. We take $\hat{f}_1, \ldots, \hat{f}_s \in \mathbb{C}[M]$. Let $P_i \subset \mathbb{R}^n$ be the Newton polytope of $\hat{f}_i$ for $i = 1, \ldots, s$. Let $P = P_1 + \ldots + P_s$ be the Minkowski sum of all these polytopes. We will assume that P is full-dimensional. The normal fan $\Sigma = \Sigma_P$ of P defines a complete, normal toric variety $X = X_\Sigma$ (we drop the index Σ for simplicity of notation). To each of the polytopes P_i, we associate a torus invariant divisor $D_{P_i} \in \mathrm{Div}_T(X)$ as follows. Let $a_i = (a_{i,1}, \ldots, a_{i,k}) \in \mathbb{Z}^k$ be such that

$$a_{i,j} = \min_{\mathbb{Z}} c \;\text{ s.t. }\; P_i \subset \{m \in M_\mathbb{R} \mid \langle u_j, m\rangle + c \geq 0\}.$$

The divisors $D_{P_i} = \sum_{j=1}^k a_{i,j} D_j$ obtained in this manner are Cartier (they are also basepoint free, see Subsection 5.5.2). We denote $\alpha_i = [D_{P_i}] \in \mathrm{Pic}(X)$. In order to send the $\hat{f}_i$ to the Cox ring S of X, we observe that by construction

$$\hat{f}_i \in \bigoplus_{m \in P_i \cap M} \mathbb{C} \cdot t^m \simeq \bigoplus_{F^\top m + a_i \geq 0} \mathbb{C} \cdot t^m \simeq \bigoplus_{F^\top m + a_i \geq 0} \mathbb{C} \cdot x^{F^\top m + a_i} = S_{\alpha_i}.$$

This gives a canonical way of *homogenizing*[5] the $\hat{f}_i$:

$$\hat{f}_i = \sum_{F^\top m + a_i \geq 0} c_{m,i} t^m \mapsto f_i = \sum_{F^\top m + a_i \geq 0} c_{m,i} x^{F^\top m + a_i} \quad \in S_{\alpha_i}.$$

Dehomogenizing to an affine chart U_σ for $\sigma \in \Sigma(n)$ yields $f_i^\sigma \in \mathbb{C}[U_\sigma]$. One can check that these are exactly the elements of $\mathbb{C}[U_\sigma]$ we obtained in Subsection 5.4.2. Moreover, for each $\sigma \in \Sigma(n)$ we have $V_X(f_i) \cap U_\sigma = V_{U_\sigma}(f_i^\sigma)$ and hence

$$V_X(f_i) = \mathrm{div}_0(\hat{f}_i).$$

The homogeneous elements $f_1, \ldots, f_s$ generate the homogeneous ideal $I = \langle f_1, \ldots, f_s\rangle \subset S$, whose zero locus satisfies

$$V_X(I) = V_X(\hat{f}_1, \ldots, \hat{f}_s).$$

[5] This map can be defined for any torus invariant divisor $\sum_{i=1}^k a_i D_i$ and it identifies the graded pieces of S with the vector spaces of global sections of divisoral sheaves on X [CLS11, Proposition 4.3.2, Proposition 5.3.7]:

$$S_{[\sum_{i=1}^k a_i D_i]} \simeq H^0\left(X, \mathscr{O}_X\left(\sum_{i=1}^k a_i D_i\right)\right) = \bigoplus_{F^\top m + a \geq 0} \mathbb{C} \cdot t^m.$$

This is a subvariety (in fact, it's a subscheme) of X which is locally given by the ideal

$$\mathscr{I}(U_\sigma) = \langle f_1^\sigma, \ldots, f_s^\sigma \rangle \subset \mathbb{C}[U_\sigma],$$

for $\sigma \in \Sigma(n)$. The multiplicity of a point $\zeta \in V_X(I) \cap U_\sigma$ is given by its multiplicity as a point of $V_{U_\sigma}(\mathscr{I}(U_\sigma))$. We conclude that the ideal I gives a *global description* of the zero set defined by extending $\hat{f}_1 = \cdots = \hat{f}_s = 0$ to X. We will work with the following assumptions on the ideal I.

Assumption 1. *$V_X(I)$ is zero-dimensional. We denote $V_X(I) = \{\zeta_1, \ldots, \zeta_\delta\} \subset X$.*

Assumption 2. *$V_X(I) \subset U \subset X$, where U is the largest simplicial open subset of X.*

Assumption 3. *I defines a reduced subscheme of $U \subset X$. That is, all points $\zeta_i \in V_X(I)$ have multiplicity one.*

The first assumption we will need throughout the text. The second assumption makes sure that the points in $V_X(I)$ have 'nice homogeneous coordinates'. That is, it implies that $\pi^{-1}(\zeta_i) = G \cdot z$ for any $z \in \pi^{-1}(\zeta_i)$, so that any homogeneous $f \in S$ vanishes at ζ_i if and only if it vanishes (as a function on $\mathbb{C}^k$) on the entire preimage $\pi^{-1}(\zeta_i)$. For $\zeta \in U \subset X$, we say that any point $z \in \pi^{-1}(\zeta)$ is a set of *homogeneous coordinates* for ζ. It is clear that whenever X is simplicial, Assumption 2 is automatically satisfied. This includes all examples where $n = 2$. For $n = 3$, U is the complement of finitely many points in X: one point for each vertex of P corresponding to a non-simplicial, full dimensional cone of Σ_P. It follows that Assumption 2 is automatically satisfied also when $n = s = 3$, since 'face systems' corresponding to vertices do not contribute any solutions (see for instance the appendix in [HS95]). We will say a few things about what we can do without Assumption 3 in Subsection 5.5.5.

Example 5.5.5. Consider the Laurent polynomials $\hat{f}_1, \hat{f}_2 \in \mathbb{C}[t_1^{\pm 1}, t_2^{\pm 1}]$ given by

$$\begin{aligned} \hat{f}_1 &= 1 + t_1 + t_2 + t_1 t_2 + t_1^2 t_2 + t_1^3 t_2, \\ \hat{f}_2 &= 1 + t_2 + t_1 t_2 + t_1^2 t_2. \end{aligned}$$

These are the equations from Example 5.4.5, which we view as relations on the Hirzebruch surface $\mathscr{H}_2$. The polytopes and fan are shown in Figures 5.7 and 5.8. The matrix F was given in Example 5.5.4. As we have seen in Example 5.4.5, the BKK bound for the system $\hat{f}_1 = \hat{f}_2 = 0$ equals 3 and the point $(-1, -1)$ is the unique solution (with multiplicity 1) in $(\mathbb{C}^*)^2$. The divisor D_{P_2} is given by $D_{P_2} = D_4$ (i.e. $a_{2,1} = a_{2,2} = a_{2,3} = 0, a_{2,4} = 1$, or $a_2 = (0,0,0,1)^\top$). The homogenization of the monomials t^m in $\hat{f}_2$ is given by $F^\top m + a_2$:

$$F^\top \begin{bmatrix} 0 & 0 & 1 & 2 \\ 0 & 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 2 \\ 0 & 1 & 1 & 1 \\ 0 & 2 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

which gives

$$f_2 = x_4 + x_2x_3^2 + x_1x_2x_3 + x_1^2x_3 \in S_{[D_4]}.$$

Analogously, using $D_{P_1} = D_3 + D_4$ we find

$$f_1 = x_3x_4 + x_1x_4 + x_2x_3^3 + x_1x_2x_3^2 + x_1^2x_2x_3 + x_1^3x_2 \in S_{[D_3+D_4]}.$$

We now see that for $I = \langle f_1, f_2 \rangle \subset S$, the vanishing locus $V_X(I)$ on X consists of three points, with homogeneous coordinates

$$z_1 = (-1,-1,1,1),\ z_2 = (0,-1,1,1),\ z_3 = (1,-1,0,1).$$

We see that $\hat{f}_1 = \hat{f}_2 = 0$ defines 3 isolated points on X, which confirms what we observed in Example 5.4.5. The ideal I satisfies Assumptions 1-3. Note that $\pi(z_1)$ is the toric solution $(-1,-1)$ (π denotes the quotient $\pi : \mathbb{C}^4 \setminus Z \to \mathscr{H}_2$) and the other solutions are on the boundary of the torus: $\pi(z_2) \in D_1, \pi(z_3) \in D_3$. Figure 5.11 illustrates what is going on in the total coordinate space $\mathbb{C}^4$ of $\mathscr{H}_2$. In order to make a picture, we consider the 3-dimensional slice given by $x_4 = 1$ of $\mathbb{C}^4$ (note that this contains all the solutions). In this slice, $f_1 = 0$ and $f_2 = 0$ define surfaces whose real parts are shown as the blue and orange surfaces in Figure 5.11. These surfaces intersect in the intersections of the orbits $G \cdot z_i$ with $\{x_4 = 1\}$, which are shown as black curves (it looks like there are six curves in the intersection, but these actually belong together two by two). The representatives z_1, z_2, z_3 are shown as red dots. $\triangle$

With all this terminology introduced, we are now ready to give a specific formulation of our goal in this section. Given $\hat{f}_1, \ldots, \hat{f}_s \in \mathbb{C}[M]$ such that $I = \langle f_1, \ldots, f_s \rangle$ satisfies Assumptions 1-3 (or maybe only Assumptions 1 and 2), we want to compute homogeneous coordinates of the points in $V_X(I)$ via eigenvalue computations. More specifically, we want to generalize the results of Subsection 3.2.2 and Section 4.5 to the multigraded, toric setting. For that, a first thing we need to do is give an appropriate definition of *regularity* in the multigraded case. This will be discussed in the next subsection.

5.5.2 Multigraded regularity

Many of the results of this section are taken from [Tel20, Section 4]. Let S be an E-graded ring. The *regularity* of a graded S-module measures its complexity (for instance, in terms of the degree of minimal generators). A classical notion of regularity (in the case where $E = \mathbb{Z}$) is that of *Castelnuovo-Mumford regularity*, see for instance [Eis13, Section 20.5] or [BS87], whose definition requires minimal free resolutions and would take us too far. Castelnuovo-Mumford regularity has been studied in a multigraded context by Maclagan and Smith in [MS03]. The zero-dimensional case is further investigated in [ŞS16], where the authors start from a subscheme of X and investigate the regularity of the 'nicest' corresponding graded S-module. Some more results in a multiprojective setting can be found in [BFT18, SVTW06].

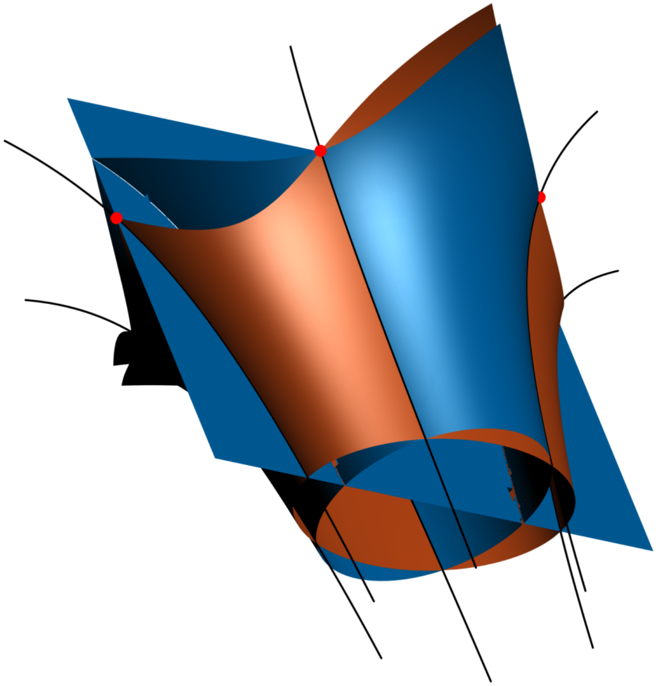


Figure 5.11: Illustration of the affine varieties defined by f_1 and f_2 from Example 5.5.5 in a 3-dimensional slice of the 4-dimensional total coordinate space of $\mathscr{H}_2$.

Let $X = X_\Sigma$ be a toric variety corresponding to a complete fan Σ, which is the normal fan of a full-dimensional polytope $P \subset M_\mathbb{R} \simeq \mathbb{R}^n$ as in Subsection 5.5.1. We consider a homogeneous ideal $I = \langle f_1, \ldots, f_s \rangle \subset S$ which we require to satisfy Assumptions 1-3 from Subsection 5.5.1. We denote $V_X(I) = \{\zeta_1, \ldots, \zeta_\delta\} \subset U$ (here U is the open subset of Remark 5.5.1) and we let $z_j \in \mathbb{C}^k \setminus Z$ be a set of homogeneous coordinates for ζ_j. In our case, the *regularity* (as defined below) of the homogeneous ideal I in the Cox ring S of X will determine in which graded piece S_α of S we can work to define our multiplication maps in Subsection 5.5.3. The 'larger' this graded piece (i.e. the larger the dimension of S_α as a $\mathbb{C}$-vector space), the larger the matrices involved in the presented algorithm in Subsection 5.5.4 will be. We will define homogeneous Lagrange polynomials and show how they are related to multigraded regularity. As in Subsections 3.1.1 and 3.2.2, these Lagrange polynomials and their dual basis will have a nice interpretation as eigenvectors of multiplication maps. For $\alpha \in \mathrm{Cl}(X)$, we denote $n_\alpha = \dim_\mathbb{C}(S_\alpha)$. Since X is complete, $n_\alpha < \infty, \forall \alpha \in \mathrm{Cl}(X)$ [CLS11, Proposition 4.3.8]. We will sometimes work with the $\mathfrak{B}$-saturated ideal J corresonding to I. We set $J = (I : \mathfrak{B}^\infty) \subset S$, which is itself homogeneous. For $\alpha \in \mathrm{Cl}(X)$, let

$S_\alpha = \bigoplus_{i=1}^{n_\alpha} \mathbb{C} \cdot x^{b_i}, b_i \in \mathbb{N}^k$ and consider the map

$$\Phi_\alpha : \mathbb{C}^k \setminus Z \dashrightarrow \mathbb{P}^{n_\alpha - 1} \simeq \mathbb{P}(S_\alpha^\vee) : (x_1, \ldots, x_k) \mapsto (x^{b_1} : \ldots : x^{b_{n_\alpha}}). \tag{5.5.7}$$

The map Φ_α is constant on G-orbits. The reason for the dashed arrow in (5.5.7) is the following. There may be points in $z \in \mathbb{C}^k \setminus Z$ for which $z^{b_i} = 0, i = 1, \ldots, n_\alpha$. For these points, the image of Φ_α is not defined. We say that $\zeta \in X$ is a *basepoint* of S_α if $\pi^{-1}(\zeta)$ contains such a point z. Note that if $\zeta \in U$, ζ is a basepoint of S_α if and only if $z^{b_i} = 0, i = 1, \ldots, n_\alpha$ for all $z \in \pi^{-1}(\zeta)$. We say that $\alpha \in \mathrm{Cl}(X)$ is *basepoint free* if Φ_α has no basepoints. If α is basepoint free, by the universal property of a good categorical quotient [CLS11, Theorem 5.0.6] the map Φ_α factors as $\Phi_\alpha = \phi_\alpha \circ \pi$, with $\phi_\alpha : X \to \mathbb{P}^{n_\alpha - 1}$. The following lemma is straightforward and we omit the proof.

Lemma 5.5.1. *Let $\alpha = [D] \in \mathrm{Cl}(X)$ be such that no ζ_j is a basepoint of S_α. For generic $h \in S_\alpha$, we have $V_X(h) \cap V_X(I) = \varnothing$ (h does not vanish at any of the points $\zeta_j \in V_X(I)$).*

Note that in particular, the condition of Lemma 5.5.1 is always satisfied for basepoint free α. The grading on S defines a grading on the quotient S/I: $(S/I)_\alpha = S_\alpha / I_\alpha$. It follows from Lemma 5.5.1 that for any $\alpha = [D] \in \mathrm{Cl}(X)$ such that no ζ_j is a basepoint of S_α, the following $\mathbb{C}$-linear map is well defined for generic $h \in S_\alpha$:

$$\psi_\alpha : (S/I)_\alpha \to \mathbb{C}^\delta : f + I_\alpha \mapsto \left(\frac{f}{h}(\zeta_1), \ldots, \frac{f}{h}(\zeta_\delta) \right). \tag{5.5.8}$$

Here we write $(f/h)(\zeta_j)$ for $f(z_j)/h(z_j)$. This notation makes sense because the evaluation does not depend on the choice of representative z_j of $G \cdot z_j$. We fix such a generic $h \in S_\alpha$. We will now investigate for which $\alpha \in \mathrm{Cl}(X)$ the map ψ_α defines coordinates on $(S/I)_\alpha$, that is, for which α it is an isomorphism (note that this is independent of the choice of h satisfying $V_X(h) \cap V_X(I) = \varnothing$). It is clear that for this to happen, we need $\dim_{\mathbb{C}}((S/I)_\alpha) = \delta$. The dimension of the graded parts of S/I is given by the multigraded analog of the Hilbert function [ŞS16].

Definition 5.5.2 (Hilbert function)**.** For a homogeneous ideal I in the Cox ring S of X, the Hilbert function of I is given by $\mathrm{HF}_I : \mathrm{Cl}(X) \to \mathbb{N} : \alpha \mapsto \dim_{\mathbb{C}}((S/I)_\alpha)$.

In order to state a necessary and sufficient condition for surjectivity of ψ_α, we will introduce a homogeneous analog of the Lagrange polynomials introduced in Subsection 3.1.1.

Definition 5.5.3 (homogeneous Lagrange polynomials)**.** Let $\alpha \in \mathrm{Cl}(X)$ be such that no ζ_j is a basepoint of S_α and let $h \in S_\alpha$ be such that $V_X(h) \cap V_X(I) = \varnothing$. A set of elements $\ell_1, \ldots, \ell_\delta \in S_\alpha$ is called a set of *homogeneous Lagrange polynomials* of degree α with respect to h if for $j = 1, \ldots, \delta$,

1. $\zeta_i \in V_X(\ell_j), i \neq j$,
2. $\zeta_j \in V_X(h - \ell_j)$.

In terms of the homogeneous coordinates z_j, a set of homogeneous Lagrange polynomials satisfies $\ell_j(z_i) = 0, i \neq j$ and $\ell_j(z_j) = h(z_j), j = 1, \dots, \delta$. In what follows, we use the same function h to define ψ_α and a set of homogeneous Lagrange polynomials. The following lemma from [BT20a] will be useful.

Lemma 5.5.2. *Let $I \subset S$ be such that $V_X(I) = \{\zeta_1, \dots, \zeta_\delta\} \subset U \subset X$ is zero-dimensional (I satisfies Assumptions 1-2, but not necessarily Assumption 3). We have that, as varieties (meaning not necessarily as schemes),*

$$V_{\mathbb{C}^k}(I : \mathfrak{B}^\infty) = \overline{\pi^{-1}(\zeta_1) \cup \cdots \cup \pi^{-1}(\zeta_\delta)} = \overline{\pi^{-1}(\zeta_1)} \cup \cdots \cup \overline{\pi^{-1}(\zeta_\delta)},$$

where the closures are taken in $\mathbb{C}^k$.

Proof. By [CLO13, Chapter, §4, Theorem 10 (iii)] we have that, as varieties,

$$V_{\mathbb{C}^k}(I : \mathfrak{B}^\infty) = \overline{V_{\mathbb{C}^k}(I) \setminus Z},$$

where $Z = V_{\mathbb{C}^k}(\mathfrak{B})$. The lemma will follow from

$$V_{\mathbb{C}^k}(I) \setminus Z = \pi^{-1}(\zeta_1) \cup \cdots \cup \pi^{-1}(\zeta_\delta).$$

The inclusion '$\supset$' needs $V_X(I) \subset U$. The other inclusion follows from $z \in V_{\mathbb{C}^k}(I) \setminus Z \Rightarrow \pi(z) \in V_X(I)$, and is satisfied also when $V_X(I)$ contains points outside of U. $\square$

Lemma 5.5.2 implies by the Nullstellensatz that the radical of $J = (I : \mathfrak{B}^\infty)$ is the vanishing ideal of the union of the orbits:

$$\sqrt{J} = \{f \in S \mid f(z) = 0, \text{ for all } z \in \pi^{-1}(\zeta_1) \cup \cdots \cup \pi^{-1}(\zeta_\delta)\}. \tag{5.5.9}$$

Recall that in the projective case, for an ideal satisfying Assumptions 1-3 we have $J = \sqrt{J}$ (Proposition 3.2.2). This is not true in the more general setting we are considering here. Here's an example.

Example 5.5.6. Consider the weighted projective space $X = \mathbb{P}_{(1,2,1)}$ with coordinate ring $\mathbb{C}[x, y, z]$ where y has degree 2 and x, z have degree 1. The fan is described in Example 5.5.2. The irrelevant ideal is $\mathfrak{B} = \langle x, y, z\rangle$. The homogeneous ideal $I = \langle x^2, xz, z^2\rangle$ defines $V_X(I)$ consisting of 1 point with multiplicity 1 and X is simplicial, so Assumptions 1-3 are satisfied. Moreover, in this example we have $I = (I : \mathfrak{B}^\infty) = J$. However, $I = J$ is not radical: $\sqrt{J} = \langle x, z\rangle \not\subset J$. $\triangle$

Proposition 5.5.1. *Consider $I \subset S$ such that Assumptions 1-3 are satisfied. Let $\alpha \in \mathrm{Cl}(X)$ be such that no ζ_j is a basepoint of S_α. Then*

1. *ψ_α is injective if and only if $I_\alpha = (\sqrt{J})_\alpha$. In this case $\mathrm{HF}_I(\alpha) \leq \delta$,*
2. *ψ_α is surjective if and only if there exists a set of homogeneous Lagrange polynomials of degree α. In this case $\mathrm{HF}_I(\alpha) \geq \delta$.*

Proof. Let $f, h \in S_\alpha$ such that $V_X(h) \cap V_X(I) = \varnothing$. If ψ_α is injective, then $f \in (\sqrt{J})_\alpha \Rightarrow \psi_\alpha(f + I_\alpha) = 0 \Rightarrow f \in I_\alpha$. So $(\sqrt{J})_\alpha \subset I_\alpha$ and the other inclusion is trivial. Conversely, if $I_\alpha = (\sqrt{J})_\alpha$, then $\psi_\alpha(f + I_\alpha) = 0 \Rightarrow f \in (\sqrt{J})_\alpha \Rightarrow f \in I_\alpha$, so ψ_α is injective. The corresponding statement about HF_I follows easily.
If ψ_α is surjective, take $\ell_j \in \psi_\alpha^{-1}(e_j)$. Conversely, if $\ell_j, j = 1, \ldots, \delta$ is a set of homogeneous Lagrange polynomials of degree α, $\psi_\alpha(\ell_j + I_\alpha) = e_j$ and ψ_α is surjective. Again, the statement about HF_I follows easily. □

Corollary 5.5.1. *Consider $I \subset S$ such that Assumptions 1-3 are satisfied. If $\alpha \in \mathrm{Pic}(X)$ is ample*[6] *and I is radical, then ψ_α is injective.*

Proof. In this case, by the Nullstellensatz we have

$$I = I_S(V_{\mathbb{C}^k}(I)) = I_S(\overline{G \cdot z_1} \cup \cdots \cup \overline{G \cdot z_\delta} \cup Z')$$

where $Z' \subset Z$. Take $f \in (\sqrt{J})_\alpha$. Since any polynomial in S_α for α ample vanishes on Z ($S_\alpha \subset \mathfrak{B}$, see e.g. [Sop05]), f vanishes on $Z' \subset Z$. Therefore $f \in I_\alpha$ and $(\sqrt{J})_\alpha \subset I_\alpha \subset (\sqrt{J})_\alpha$. Now apply Proposition 5.5.1. □

Example 5.5.7. Consider the ideal I from Example 5.5.5. We computed the primary decomposition of I over the rationals using Macaulay2 [EGSS01]. This gives

$$I = \langle x_1 + x_3, x_2x_3^2 + x_4\rangle \cap \langle x_1, x_2x_3^2 + x_4\rangle \cap \langle x_3, x_1^2x_2 + x_4\rangle \cap \langle x_2, x_4\rangle.$$

All primary components are prime, which implies that I is radical. This decomposition of I corresponds to the decomposition of the associated affine variety $V_{\mathbb{C}^k}(I) = \overline{G \cdot z_1} \cup \overline{G \cdot z_2} \cup \overline{G \cdot z_3} \cup Z'$ with orbit representatives $z_1 = (-1, -1, 1, 1), z_2 = (0, -1, 1, 1), z_3 = (1, -1, 0, 1)$ and $Z' = V(x_2, x_4) \subset Z$. △

The following proposition shows that the existence of homogeneous Lagrange polynomials of degree $\alpha \in \mathrm{Cl}(X)$ is equivalent to the fact that the points $\Phi_\alpha(z_j)$ span a linear space of dimension $\delta - 1$ in $\mathbb{P}^{n_\alpha - 1}$. Let $p_j \in \mathbb{C}^{n_\alpha}$ be a set of homogeneous coordinates (in the standard sense) of $\Phi_\alpha(z_j) \in \mathbb{P}^{n_\alpha - 1}$ and define the matrix $L_\alpha = [p_1 \ \cdots \ p_\delta] \in \mathbb{C}^{n_\alpha \times \delta}$.

Proposition 5.5.2. *Consider $I \subset S$ such that Assumptions 1-3 are satisfied. Let $\alpha \in \mathrm{Cl}(X)$ be such that no ζ_j is a basepoint of S_α. There exists a set of Lagrange polynomials of degree α if and only if L_α has rank δ.*

Proof. The rank of L_α is δ if and only if there exists a left inverse matrix $L_\alpha^\dagger \in \mathbb{C}^{\delta \times n_\alpha}$ such that $L_\alpha^\dagger L_\alpha = \mathrm{id}_\delta$ is the $\delta \times \delta$ identity matrix. We will show that this is equivalent to the existence of a set of homogeneous Lagrange polynomials of degree α. Suppose

[6] A divisor D and its degree $\alpha = [D]$ are called *very ample* if D is basepoint free and $X \to \mathbb{P}(\Gamma(X, \mathscr{O}_X(D))^\vee)$ is a closed embedding. If kD (or $k\alpha$) is very ample for some $k \geq 1$, then D (or α) is called *ample*. See [CLS11, Chapter 6] for definitions and properties.

that $L_\alpha^\dagger$ exists. The rows of $L_\alpha^\dagger$ should be interpreted as elements of S_α represented in the basis $\{x^{b_1}, \dots, x^{b_{n_\alpha}}\}$. The columns of L_α are elements of $S_\alpha^\vee$ represented in the dual basis. Let the j-th row of $L_\alpha^\dagger$ correspond to $\tilde{\ell}_j \in S_\alpha$. It is clear from $L_\alpha^\dagger L_\alpha = \mathrm{id}_\delta$ that

$$\langle \tilde{\ell}_j, p_i \rangle = \tilde{\ell}_j(z_i) = \begin{cases} 1 & i = j, \\ 0 & \text{otherwise.} \end{cases}$$

By Lemma 5.5.1, there is $h \in S_\alpha$ such that $h(z_j) \neq 0, j = 1, \dots, \delta$. Then $\ell_j = h(z_j)\tilde{\ell}_j, j = 1, \dots, \delta$ are a set of homogeneous Lagrange polynomials. Conversely, if a set of homogeneous Lagrange polynomials exists, construct a matrix $\tilde{L}_\alpha^\dagger$ by plugging the coefficients of ℓ_j into the j-th row. Then there is $h \in S_\alpha$ such that $\tilde{L}_\alpha^\dagger L_\alpha = \mathrm{diag}(h(z_1), \dots, h(z_\delta))$ is an invertible diagonal matrix. The left inverse is $L_\alpha^\dagger = \mathrm{diag}(h(z_1), \dots, h(z_\delta))^{-1} \tilde{L}_\alpha^\dagger$. □

An important property of the homogeneous evaluation maps in Subsection 3.2.2 was that, for degrees in the regularity, they are isomorphisms. In order to generalize this, we make the following definition.

Definition 5.5.4 (Regularity)**.** Consider $I \subset S$ satisfying Assumptions 1-3 and let $J = (I : \mathfrak{B}^\infty)$. The *regularity* $\mathrm{Reg}(I) \subset \mathrm{Cl}(X)$ of I is the subset of degrees $\alpha \in \mathrm{Cl}(X)$ for which no ζ_j is a basepoint of S_α and the following equivalent conditions are satisfied:

1. ψ_α is an isomorphism,
2. $\mathrm{HF}_I(\alpha) = \delta$ and $I_\alpha = (\sqrt{J})_\alpha$,
3. $\mathrm{HF}_I(\alpha) = \delta$ and there exists a set of homogeneous Lagrange polynomials of degree α,
4. $I_\alpha = (\sqrt{J})_\alpha$ and there exists a set of homogeneous Lagrange polynomials of degree α.

Example 5.5.8. We continue Example 5.5.7. The polytope $P = P_1 + P_2$ (shown in Figure 5.7) has 12 lattice points. Therefore $n_\alpha = 12$, with $\alpha = [D_P] \in \mathrm{Pic}(X)$. Since $\delta = 3$, L_α is a 12×3 matrix. Its rows are indexed by the monomials spanning S_α, and its columns by the representatives z_j. The transpose is given by

$$L_\alpha^\top = \begin{array}{c} \begin{array}{cccccccccccc} x_3x_4^2 & x_1x_4^2 & x_2x_3^3x_4 & x_1x_2x_3^2x_4 & x_1^2x_2x_3x_4 & x_1^3x_2x_4 & x_2^2x_3^5 & x_1x_2^2x_3^4 & x_1^2x_2^2x_3^3 & x_1^3x_2^2x_3^2 & x_1^4x_2^2x_3 & x_1^5x_2^2 \end{array} \\ \begin{bmatrix} 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{c} z_1 \\ z_2 \\ z_3 \end{array} .$$

Consider $h = 39(x_3x_4^2 - x_1x_4^2) \in S_\alpha$ and note that $h(z_j) \neq 0$ for all j. A set of homogeneous Lagrange polynomials w.r.t. h is given by

$$\frac{2\,\tilde{L}_\alpha^\dagger}{13} = \begin{array}{c} \begin{array}{cccccccccccc} x_3x_4^2 & x_1x_4^2 & x_2x_3^3x_4 & x_1x_2x_3^2x_4 & x_1^2x_2x_3x_4 & x_1^3x_2x_4 & x_2^2x_3^5 & x_1x_2^2x_3^4 & x_1^2x_2^2x_3^3 & x_1^3x_2^2x_3^2 & x_1^4x_2^2x_3 & x_1^5x_2^2 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 2 & -2 & 0 & 0 & -2 & 2 & -2 & 2 & 0 \\ 2 & 0 & -2 & -1 & 1 & 0 & 2 & 1 & -1 & 1 & -1 & 0 \\ 0 & 2 & 0 & 1 & -1 & -2 & 0 & -1 & 1 & -1 & 1 & 2 \end{bmatrix} \end{array} \begin{array}{c} \\ \ell_1 \\ \ell_2 \\ \ell_3 \end{array} ,$$

which is related to the pseudo inverse of L_α by

$$L_\alpha^\dagger = \operatorname{diag}(h(z_1), h(z_2), h(z_3))^{-1}\tilde{L}_\alpha^\dagger = \operatorname{diag}(1/78, 1/39, 1/39)\tilde{L}_\alpha^\dagger.$$

To check that $I_\alpha = (\sqrt{J})_\alpha$ we compute $\mathrm{HF}_I(\alpha) = \mathrm{HF}_{\sqrt{J}}(\alpha) = 3$. Because $I \subset \sqrt{J}$, we conclude $\alpha \in \mathrm{Reg}(I)$. In fact, in this example I is radical and α is ample, so $I_\alpha = (\sqrt{J})_\alpha$ follows from Corollary 5.5.1. $\triangle$

The following proposition shows that, in the case where $X = \mathbb{P}^n$ and I is a zero-dimensional homogeneous ideal whose projective variety consists of isolated points with multiplicity 1, Definition 5.5.4 agrees with Definition 3.2.4.

Proposition 5.5.3. *Let $I = \langle f_1, \ldots, f_s \rangle \subset S$ be such that Assumptions 1-3 are satisfied. We have that $J_\alpha = (I : \mathfrak{B}^\infty)_\alpha = (\sqrt{(I : \mathfrak{B}^\infty)})_\alpha = (\sqrt{J})_\alpha$ for all $\alpha \in \mathrm{Pic}(X)$.*

Proof. The inclusion $J \subset \sqrt{J}$ holds for all degrees. We sketch a proof of the opposite inclusion, which is very similar to the proof of Proposition 3.2.2. For $g \in (\sqrt{J})_\alpha$ with $\alpha \in \mathrm{Pic}(X)$, we consider the dehomogenization g^σ as in (5.5.6). Since g^σ vanishes at all the points $\zeta \in V_X(I) \cap U_\sigma$ (Lemma 5.5.2) we have that

$$g^\sigma = h_1^\sigma f_1^\sigma + \cdots + h_s^\sigma f_s^\sigma,$$

for some $h_i^\sigma \in \mathbb{C}[U_\sigma]$, which implies that there is some $\ell \in \mathbb{N}$ such that $(x^{\hat{\sigma}})^\ell g \in I$ for all $\sigma \in \Sigma(n)$. Hence $g \in (I : \mathfrak{B}^\infty)_\alpha = J_\alpha$. $\square$

Remark 5.5.3. Proposition 5.5.3 implies that $\sqrt{J}$ in Definition 5.5.4 may be replaced by J when X is smooth (because in this case $\mathrm{Cl}(X) = \mathrm{Pic}(X)$). In particular, this holds when $X = \mathbb{P}^n$. $\triangle$

What we will prove in the next subsection is that, in analogy with Subsection 3.2.2, if $\alpha, \alpha + \alpha_0$ are in the regularity, then 'multiplication of elements in $(S/I)_\alpha$ with elements of degree α_0' has some nice properties. It makes sense to require α_0 to be such that S_{α_0} has some nonzero elements. We define the following submonoid of the class group:

$$\mathrm{Cl}(X)_+ = \{\alpha \in \mathrm{Cl}(X) \mid \alpha = [\sum_{i=1}^k a_i D_i] \text{ with } a_i \geq 0, i = 1, \ldots, k\}.$$

These are the divisor classes represented by *effective divisors*. This is sometimes called the *weight monoid* of S. Note that $S_\alpha = \{0\}$ for $\alpha \in \mathrm{Cl}(X) \setminus \mathrm{Cl}(X)_+$. Since all points of X are basepoints of S_α for $\alpha \in \mathrm{Cl}(X) \setminus \mathrm{Cl}(X)_+$, we have that $\mathrm{Reg}(I) \subset \mathrm{Cl}(X)_+$. The following definition helps to reduce the length of some statements in what follows and was suggested to the author by Matías Bender.

Definition 5.5.5 (Regularity pair)**.** Let $I \subset S$ be such that $V_X(I) = \{\zeta_1, \ldots, \zeta_\delta\}$ and Assumptions 1-3 are satisfied. A tuple $(\alpha, \alpha_0) \in \mathrm{Cl}(X)_+^2$ is called a *regularity pair* for I if $\alpha, \alpha + \alpha_0 \in \mathrm{Reg}(I)$ and no ζ_j is a basepoint of S_{α_0}.

In general, characterizing the regularity $\mathrm{Reg}(I)$ is a hard problem. This is a topic of ongoing research. There are some things we can say in the case where I is generated by $s = n$ elements (i.e. the square case), and some general properties are known. These results are listed in Subsection 5.5.5. For now, we assume that we can compute a regularity pair (α, α_0) and show what we can do under this assumption.

5.5.3 Toric eigenvalue-eigenvector theorem

The material presented here can be found in Section 5 of [Tel20]. Throughout this subsection, $I \subset S$ is a homogeneous ideal satisfying Assumptions 1-3. We denote the points in $V_X(I)$ by $V_X(I) = \{\zeta_1, \ldots, \zeta_\delta\}$. For $\alpha, \alpha_0 \in \mathrm{Cl}(X)_+$, a homogeneous element $g \in S_{\alpha_0}$ defines a $\mathbb{C}$-linear map

$$M_g : (S/I)_\alpha \to (S/I)_{\alpha+\alpha_0} : f + I_\alpha \mapsto gf + I_{\alpha+\alpha_0}$$

representing 'multiplication with g'. Just as in the affine and projective case, these multiplication maps will be the key ingredient to formulate our root finding problem as a linear algebra problem. We state a toric version of the eigenvalue, eigenvector theorem and show how the eigenvalues can be used to recover homogeneous coordinates of the solutions and equations for the corresponding G-orbits. Our main result uses the following lemma.

Lemma 5.5.3. *Let $(\alpha, \alpha_0) \in \mathrm{Cl}(X)_+^2$ be a regularity pair for I. Then for $h_0 \in S_{\alpha_0}$ such that $V_X(h_0) \cap V_X(I) = \varnothing$, $M_{h_0} : (S/I)_\alpha \to (S/I)_{\alpha+\alpha_0} : f + I_\alpha \mapsto h_0 f + I_{\alpha+\alpha_0}$ is an isomorphism of vector spaces.*

Proof. Let ψ_α be given as in (5.5.8) for some $h \in S_\alpha$. We can take $hh_0 \in S_{\alpha+\alpha_0}$ to define $\psi_{\alpha+\alpha_0}$. Then $\psi_{\alpha+\alpha_0} \circ M_{h_0} = \psi_\alpha$ shows that M_{h_0} is invertible. □

For $\alpha \in \mathrm{Reg}(I)$, a set of Lagrange polynomials $\ell_j, j = 1, \ldots, \delta$ of degree α with respect to $h \in S_\alpha$ gives a basis $\{\ell_j + I_\alpha\}_{j=1,\ldots,\delta}$ for $(S/I)_\alpha$. The dual basis is given by

$$\mathrm{ev}_{\zeta_j} : (S/I)_\alpha \to \mathbb{C} \quad \text{with} \quad \mathrm{ev}_{\zeta_j}(f + I_\alpha) = \frac{f}{h}(\zeta_j).$$

Note that $\psi_\alpha = (\mathrm{ev}_{\zeta_1}, \ldots, \mathrm{ev}_{\zeta_\delta})$. The following theorem is a generalization of Theorem 3.2.4. The proofs are identical.

Theorem 5.5.3 (Toric eigenvalue, eigenvector theorem)**.** *Let $(\alpha, \alpha_0) \in \mathrm{Cl}(X)^2_+$ be a regularity pair for I and let $h_0 \in S_{\alpha_0}$ be such that $V_X(h_0) \cap V_X(I) = \varnothing$. For any $g \in S_{\alpha_0}$, the $\mathbb{C}$-linear map $M_{g/h_0} = M_{h_0}^{-1} \circ M_g : (S/I)_\alpha \to (S/I)_\alpha$ has eigenpairs*

$$\left(\frac{g}{h_0}(\zeta_j), \ell_j + I_\alpha\right), \qquad \left(\mathrm{ev}_{\zeta_j}, \frac{g}{h_0}(\zeta_j)\right), \quad j = 1, \dots, \delta,$$

where the $\ell_j + I_\alpha$ are cosets of homogeneous Lagrange polynomials of degree α and the ev_{ζ_j} are the dual basis of $(S/I)^\vee_\alpha$.

Proof. The map M_{h_0} is an isomorphism by Lemma 5.5.3. We define ψ_α, $\psi_{\alpha+\alpha_0}$ as in (5.5.8) with $h \in S_\alpha$, $hh_0 \in S_{\alpha+\alpha_0}$ respectively. A straightforward computation shows that $\psi_{\alpha+\alpha_0} \circ M_{h_0}(\ell_j + I_\alpha) = e_j$. Analogously, we have $\psi_{\alpha+\alpha_0} \circ M_g(\ell_j + I_\alpha) = \frac{g}{h_0}(\zeta_j) e_j$. It follows that $h_0(z_j) M_g(\ell_j + I_\alpha) = g(z_j) M_{h_0}(\ell_j + I_\alpha)$, and therefore

$$M_{g/h_0}(\ell_j + I_\alpha) = \frac{g}{h_0}(\zeta_j)(\ell_j + I_\alpha),$$

which proves the statement about the right eigenpairs, since the $\ell_j + I_\alpha$ are linearly independent. For the statement about the left eigenpairs, note that for any $f \in S_\alpha$

$$\mathrm{ev}_{\zeta_j} \circ M_{g/h_0}(f + I_\alpha) = \mathrm{ev}_{\zeta_j} \circ M_{h_0}^{-1}(gf + I_{\alpha+\alpha_0})$$

and since M_{h_0} is an isomorphism, there is $\tilde{f} \in S_\alpha$ such that $gf - h_0\tilde{f} \in I_{\alpha+\alpha_0}$. Therefore, for each $\zeta_j \in V_X(I)$ we have

$$\frac{gf - h_0\tilde{f}}{h_0 h}(\zeta_j) = 0 \Rightarrow \frac{\tilde{f}}{h}(\zeta_j) = \frac{g}{h_0}(\zeta_j)\frac{f}{h}(\zeta_j)$$

and thus, since $M_{h_0}^{-1}(gf + I_{\alpha+\alpha_0}) = \tilde{f} + I_\alpha$, we have

$$\mathrm{ev}_{\zeta_j} \circ M_{g/h_0}(f + I_\alpha) = \mathrm{ev}_{\zeta_j}(\tilde{f} + I_\alpha) = \frac{g}{h_0}(z_j)\,\mathrm{ev}_{\zeta_j}(f + I_\alpha).$$

The ev_{ζ_j} are linearly independent, so this concludes the proof. □

Remark 5.5.4. The condition '$h_0 \in S_{\alpha_0}$ such that $V_X(h_0) \cap V_X(I) = \varnothing$' in Lemma 5.5.3 and Theorem 5.5.3 holds for generic elements of S_{α_0}. △

Theorem 5.5.3 suggests a strategy for achieving our goal, which is to compute (approximations of) the homogeneous coordinates z_j of the points ζ_j in $V_X(I)$. For a regularity pair $(\alpha, \alpha_0) \in \mathrm{Cl}(X)^2_+$, we consider all monomials $x^{b_i} \in S_{\alpha_0}, i = 1, \dots, n_{\alpha_0}$. For each of these monomials, we compute the multiplication map $M_{x^{b_i}/h_0}$ in some basis. The eigenvalues, by Theorem 5.5.3, are

$$\lambda_{ij} = \frac{z_j^{b_i}}{h_0(z_j)}, \quad i = 1, \dots, n_{\alpha_0}, \quad j = 1, \dots, \delta.$$

After simultaneous diagonalization (or simultaneous upper-triangularization) of the matrices $M_{x^{b_i}/h_0}$, we can construct a table

$$\begin{array}{c|ccc}
 & \cdots & x^{b_i}/h_0 & \cdots \\
\hline
\vdots & & \vdots & \\
z_j & \cdots & \lambda_{ij} & \cdots \\
\vdots & & \vdots &
\end{array}$$

whose columns are indexed by the multiplication maps $M_{x^{b_i}/h_0}$ and filled with their eigenvalues. The order in which the eigenvalues are plugged into the columns corresponds to the ordering of the shared eigenvectors. Up to the factor $h_0(z_j)^{-1}$ we have computed the evaluation of n_{α_0} monomials at a set of homogeneous coordinates z_j for ζ_j. Intuitively, if S_{α_0} has 'enough' monomials, we should be able to recover the homogeneous coordinates from our table. The rest of this subsection is dedicated to the problem of finding the coordinates of z_j from the eigenvalues λ_{ij}. Before we continue, we illustrate how the construction works for our running example.

Example 5.5.9. We consider again the curves on the Hirzebruch surface from Example 5.5.5. Take $\alpha = (1,2), \alpha_0 = (0,1), h_0 = x_4 \in S_{\alpha_0}$. Recall that $V_X(h_0) \cap V_X(I) = \varnothing$. One can check that (α, α_0) is a regularity pair (and we will prove this, see Theorem 5.5.7). The monomials in S_{α_0} are $x_4, x_2x_3^2, x_1x_2x_3, x_1^2x_2$. We use the bases

$$\begin{aligned}
(S/I)_\alpha &= \mathrm{span}_{\mathbb{C}}(x_3x_4^2 + I_\alpha, x_1x_4^2 + I_\alpha, x_1x_2x_3^2x_4 + I_\alpha), \\
(S/I)_{\alpha+\alpha_0} &= \mathrm{span}_{\mathbb{C}}(x_3x_4^3 + I_{\alpha+\alpha_0}, x_1x_4^3 + I_{\alpha+\alpha_0}, x_1x_2x_3^2x_4^2 + I_{\alpha+\alpha_0})
\end{aligned}$$

to construct matrices of the multiplication maps. To construct $M_{x_2x_3^2}$ we use

$$\begin{aligned}
x_2x_3^2 \cdot (x_3x_4^2 + I_\alpha) &= -x_3x_4^3 + I_{\alpha+\alpha_0} \\
x_2x_3^2 \cdot (x_1x_4^2 + I_\alpha) &= x_1x_2x_3^2x_4^2 + I_{\alpha+\alpha_0} \\
x_2x_3^2 \cdot (x_1x_2x_3^2x_4 + I_\alpha) &= -x_1x_2x_3^2x_4^2 + I_{\alpha+\alpha_0}
\end{aligned}
\qquad M_{x_2x_3^2} = \begin{bmatrix} -1 & & \\ & & \\ & 1 & -1 \end{bmatrix}.$$

One can check that in these bases, M_{x_4} is the identity matrix. The matrices of $M_{x^{b_i}/h_0}$ for all monomials x^{b_i} of degree α_0 are

$$M_{x_4/x_4} = \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix}, \qquad M_{x_2x_3^2/x_4} = \begin{bmatrix} -1 & & \\ & & \\ & 1 & -1 \end{bmatrix},$$

$$M_{x_1x_2x_3/x_4} = \begin{bmatrix} & & \\ & & \\ 1 & -1 & 1 \end{bmatrix}, \qquad M_{x_1^2x_2/x_4} = \begin{bmatrix} & & \\ & -1 & \\ -1 & & -1 \end{bmatrix}.$$

After the eigenvalue computations, we obtain the following table.

	x_4/x_4	$x_2x_3^2/x_4$	$x_1x_2x_3/x_4$	$x_1^2x_2/x_4$
$z_1 = (-1,-1,1,1)$	1	-1	1	-1
$z_2 = (0,-1,1,1)$	1	-1	0	0
$z_3 = (1,-1,0,1)$	1	0	0	-1

△

Let $S_{\alpha_0} = \bigoplus_{i=1}^{n_{\alpha_0}} \mathbb{C} \cdot x^{b_i}$ where $\alpha_0 \in \mathrm{Cl}(X)_+$ is such that no ζ_j is a basepoint of S_{α_0}. Our goal in what follows is to show how the eigenvalues of the $M_{x^{b_i}/h_0}$ lead directly to a set of defining equations of $G \cdot z_j, j = 1, \ldots, \delta$ if α_0 is 'large enough'. We now specify what we mean by 'defining equations' and 'large enough'.

For every cone $\sigma \in \Sigma_P$ (recall that this is the normal fan of a full-dimensional lattice polytope $P \subset M_{\mathbb{R}} \simeq \mathbb{R}^n$), we define $U_{\sigma'} = \mathbb{C}^k \setminus V(x^{\hat{\sigma}}) = \mathrm{MaxSpec}(S_{x^{\hat{\sigma}}})$. These open subsets of $\mathbb{C}^k$ appeared also in Subsection 5.5.1. By Assumption 3, the orbit $G \cdot z_j$ is contained in $U_{\sigma'}$ for some simplicial cone $\sigma \in \Sigma_P$. Moreover, $G \cdot z_j$ is closed in $\mathbb{C}^k \setminus Z$, which implies that it is closed in $U_{\sigma'}$. What we are looking for is an ideal of $\mathbb{C}[U_{\sigma'}] = S_{x^{\hat{\sigma}}}$ whose variety is $G \cdot z_j$.

Let D_{α_0} be a representative divisor for α_0: $\alpha_0 = [D_{\alpha_0}] = [\sum_{i=1}^k a_{0,i} D_i]$. Since $\alpha_0 \in \mathrm{Cl}(X)_+$, we may assume that $a_{0,i} \geq 0, i = 1, \ldots, k$. Let $P_0 \subset M_{\mathbb{R}}$ be the polytope $\{m \in M_{\mathbb{R}} \mid F^\top m + a_0 \geq 0\}$ with $a_0 = (a_{0,1}, \ldots, a_{0,k})$. If D_{α_0} is Cartier and basepoint free, then for every $\sigma \in \Sigma_P$ there is $m_\sigma \in P_0 \cap M$ such that

$$\langle u_i, m_\sigma \rangle + a_{0,i} = 0, \quad \forall \rho_i \in \sigma(1), \tag{5.5.10}$$

see [CLS11, Theorem 6.1.7]. If D_{α_0} is not Cartier and basepoint free, such an m_σ does not exist for every cone $\sigma \in \Sigma_P$. We will denote the subset of cones for which $m_\sigma \in P_0 \cap M$ satisfying (5.5.10) exists by $\widetilde{\Sigma}_P \subset \Sigma_P$. This set is nonempty since $\{0\} \in \widetilde{\Sigma}_P$. We write $P_0 \cap M = \{m_1, \ldots, m_{n_{\alpha_0}}\}$, $b_i = F^\top m_i + a_0$ and $b_\sigma = F^\top m_\sigma + a_0$. For all $\sigma \in \widetilde{\Sigma}_P$ we denote $P_0 \cap M - m_\sigma = \{m_1 - m_\sigma, \ldots, m_{n_{\alpha_0}} - m_\sigma\}$ (note that $0 \in P_0 \cap M - m_\sigma$) and

$$\sigma^\vee = \{m \in M_{\mathbb{R}} \mid \langle u, m \rangle \geq 0, \forall u \in \sigma\}, \quad \sigma^\perp = \{m \in M_{\mathbb{R}} \mid \langle u, m \rangle = 0, \forall u \in \sigma\}.$$

We partition $P_0 \cap M - m_\sigma$ into

$$\mathcal{M}_\sigma^\perp = (P_0 \cap M - m_\sigma) \cap \sigma^\perp \quad \text{and} \quad \mathcal{M}_\sigma = (P_0 \cap M - m_\sigma) \setminus \mathcal{M}_\sigma^\perp.$$

These sets depend on α_0, although it is not explicit in the notation. The inclusion

$$\mathbb{N}\mathcal{M}_\sigma + \mathbb{Z}\mathcal{M}_\sigma^\perp = \left\{ \sum_{m \in \mathcal{M}_\sigma} c_m m + \sum_{m \in \mathcal{M}_\sigma^\perp} d_m m \mid c_m \in \mathbb{N}, d_m \in \mathbb{Z} \right\} \subset \sigma^\vee \cap M$$

is clear. In what follows, we will show that if equality holds for some simplicial $\sigma \in \widetilde{\Sigma}_P$, then α_0 is 'large enough' to recover equations for $G \cdot z$ from the evaluations of $x^{b_i}/h_0, i = 1, \ldots, n_{\alpha_0}$ at $\zeta = \pi(z)$ for each point $z \in U_{\sigma'} \setminus V_{\mathbb{C}^k}(h_0)$ (or, equivalently, $\zeta \in U_\sigma \setminus V_X(h_0) = \pi(U_{\sigma'} \setminus V_{\mathbb{C}^k}(h_0))$). To illustrate the idea and the notation, we first apply this to our running example.

Example 5.5.10. We consider again the Hirzebruch surface $X = \mathscr{H}_2$ and its $\mathbb{Z}^2$-graded Cox ring S. As in Example 5.5.9, let $\alpha_0 = (0, 1) = [D_4] \in \mathrm{Cl}(X)$. That is, we choose $D_{\alpha_0} = D_4$ and $a_0 = (0, 0, 0, 1)$ $(a_{0,1} = a_{0,2} = a_{0,3} = 0, a_{0,4} = 1)$. For the reader's convenience, the fan Σ_P of X (with its cones labeled in consistency with the

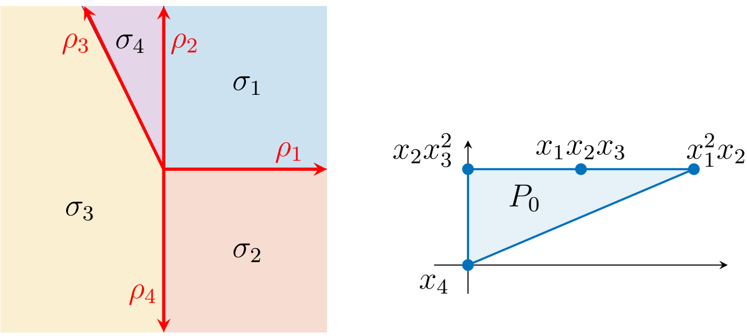


Figure 5.12: Fan of the Hirzebruch surface $\mathscr{H}_2$ (left) and the polytope P_0 from Example 5.5.10 (right).

previous examples) is shown once more in the left part of Figure 5.12. The polytope P_0, whose lattice points correspond to the monomials in S of degree α_0, is shown in the right part of the same figure. The polytope P, of which Σ_P is the normal fan, is shown in Figure 5.7. Since $\alpha_0 \in \mathrm{Pic}(X)$ is basepoint free, m_σ satisfying (5.5.10) exists for each $\sigma \in \Sigma_P$. In other words, in this example $\widetilde{\Sigma}_P = \Sigma_P$. For a selection of cones in $\widetilde{\Sigma}_P = \Sigma_P$, the sets $\mathcal{M}_\sigma, \mathcal{M}_\sigma^\perp, \sigma^\vee \cap M$ and $\mathbb{N}\mathcal{M}_\sigma + \mathbb{Z}\mathcal{M}_\sigma^\perp$ are shown in Table 5.3. One can check that the equality $\sigma^\vee \cap M = \mathbb{N}\mathcal{M}_\sigma + \mathbb{Z}\mathcal{M}_\sigma^\perp$ holds for $\sigma = \sigma_2, \sigma_3, \rho_1, \rho_3, \rho_4, \{0\}$, and it fails for the other cones of Σ_P. We will see that this implies that, in order for it to be possible to recover the homogeneous coordinates of a point $\zeta \in X$ from the evaluations of x^{b_i}/h_0 at ζ, for $x^{b_i} \in S_{\alpha_0}$ and $h_0 \in S_{\alpha_0}$ such that $\zeta \notin V_X(h_0)$, it is sufficient that

$$\zeta \in \bigcup_{\sigma \in \{\sigma_2, \sigma_3, \rho_1, \rho_3, \rho_4, \{0\}\}} U_\sigma = X \setminus D_2,$$

where the last equality follows from the orbit-cone correspondence (Theorem E.2.3). Note that if $\zeta \in D_2$, then all monomials in S_{α_0}, except for x_4, vanish at ζ. Knowing only the evaluation of these monomials at ζ, we do not have sufficient information to recover the first and third homogeneous coordinates. $\triangle$

Theorem 5.5.4. *Let $z \in U_{\sigma'}$ for a simplicial cone $\sigma \in \widetilde{\Sigma}_P$ such that $\zeta = \pi(z)$ is not a basepoint of S_{α_0}. Take $h_0 \in S_{\alpha_0}$ such that $\zeta \notin V_X(h_0)$ and let $\lambda_i = z^{b_i}/h_0(z), i = 1, \ldots, n_{\alpha_0}$ be the evaluations of x^{b_i}/h_0 at ζ. If α_0 is such that $\sigma^\vee \cap M = \mathbb{N}\mathcal{M}_\sigma + \mathbb{Z}\mathcal{M}_\sigma^\perp$, then $G \cdot z \subset U_{\sigma'}$ is the subvariety*

$$V_{U_{\sigma'}}\left(x^{b_i - b_\sigma} - \lambda_i \frac{h_0(x)}{x^{b_\sigma}}, i = 1, \ldots, n_{\alpha_0}\right) \subset U_{\sigma'}.$$

We will use the following lemma in the proof of Theorem 5.5.4.

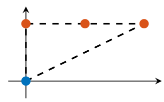
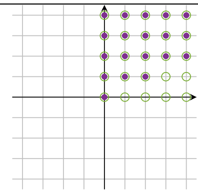
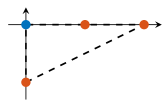
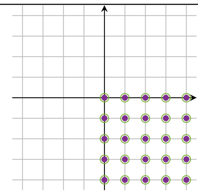
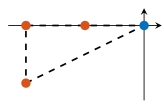
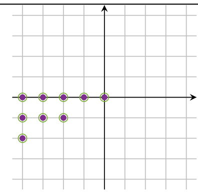
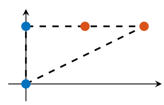
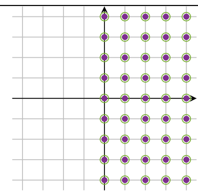
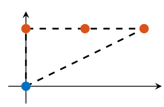
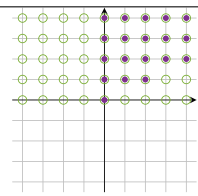
σ	m_σ	b_σ	$\mathcal{M}_\sigma^\perp, \mathcal{M}_\sigma$	$\sigma^\vee \cap M,\ \mathbb{N}\mathcal{M}_\sigma + \mathbb{Z}\mathcal{M}_\sigma^\perp$
σ_1	$(0,0)$	$(0,0,0,1)$		
σ_2	$(0,1)$	$(0,1,2,0)$		
σ_3	$(2,1)$	$(2,1,0,0)$		
ρ_1	$(0,0)$	$(0,0,0,1)$		
ρ_2	$(0,0)$	$(0,0,0,1)$		

Table 5.3: Sets of lattice points corresponding to α_0 and some cones of Σ_P in Example 5.5.10.

Lemma 5.5.4. *Let $\sigma \in \widetilde{\Sigma}_P$ be a simplicial cone. For any point $z \in U_{\sigma'}$, the orbit $G \cdot z$ is the subvariety*

$$G \cdot z = V_{U_{\sigma'}}\left(x^{F^\top m} - z^{F^\top m}, m \in \sigma^\vee \cap M\right) \subset U_{\sigma'}.$$

If $\sigma^\vee \cap M = \mathbb{N}\{m_1, \ldots, m_\kappa\} + \mathbb{Z}\{m_{\kappa+1}, \ldots, m_s\}$, then

$$V_{U_{\sigma'}}\left(x^{F^\top m} - z^{F^\top m}, m \in \sigma^\vee \cap M\right) = V_{U_{\sigma'}}\left(x^{F^\top m_i} - z^{F^\top m_i}, i = 1, \ldots, s\right).$$

Proof. Note that $x^{F^\top m} - z^{F^\top m} \in S_{x^{\hat{\sigma}}} = \mathbb{C}[U_{\sigma'}], \forall m \in \sigma^\vee \cap M$ and $m_{\kappa+1}, \ldots, m_s \in \sigma^\perp \cap M$. The first statement is shown in the proof of Theorem 2.1 in [Cox95]. For the second statement, the inclusion '$\subset$' is obvious. To show the opposite inclusion, take $m \in \sigma^\vee \cap M$ and write $m = c_1 m_1 + \ldots + c_s m_s$ with $c_1, \ldots, c_\kappa \in \mathbb{N}$, $c_{\kappa+1}, \ldots, c_s \in \mathbb{Z}$. Then

$$x^{F^\top m} = \prod_{i=1}^{\kappa} (x^{F^\top m_i})^{c_i} \prod_{j=\kappa+1}^{s} (x^{F^\top m_j})^{c_j}$$

and if $x^{F^\top m_i} = z^{F^\top m_i}, i = 1, \ldots, s$, it follows that $x^{F^\top m} = z^{F^\top m}$. □

Proof of Theorem 5.5.4. It follows from Lemma 5.5.4 that $G \cdot z$ is the variety of

$$\langle x^{F^\top (m_i - m_\sigma)} - z^{F^\top (m_i - m_\sigma)} \mid i = 1, \ldots, n_{\alpha_0} \rangle = \langle x^{b_i - b_\sigma} - z^{b_i - b_\sigma} \mid i = 1, \ldots, n_{\alpha_0} \rangle.$$

Write $h_0(x) = \sum_{i=1}^{n_{\alpha_0}} c_i x^{b_i}, c_i \in \mathbb{C}$. It is easy to check that

$$\left(\mathrm{id}_{n_{\alpha_0}} - \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_{n_{\alpha_0}} \end{bmatrix} \begin{bmatrix} c_1 & \ldots & c_{n_{\alpha_0}} \end{bmatrix} \right) \begin{bmatrix} x^{b_1 - b_\sigma} - z^{b_1 - b_\sigma} \\ \vdots \\ x^{b_{n_{\alpha_0}} - b_\sigma} - z^{b_{n_{\alpha_0}} - b_\sigma} \end{bmatrix} = \begin{bmatrix} x^{b_1 - b_\sigma} - \lambda_1 \frac{h_0(x)}{x^{b_\sigma}} \\ \vdots \\ x^{b_{n_{\alpha_0}} - b_\sigma} - \lambda_{n_{\alpha_0}} \frac{h_0(x)}{x^{b_\sigma}} \end{bmatrix}.$$

Now, if $x \in G \cdot z$ it is clear that $x^{b_i - b_\sigma} - \lambda_i (h_0(x)/x^{b_\sigma}) = 0, i = 1, \ldots, n_{\alpha_0}$. For the other implication, we observe that if for some $x \in U_{\sigma'}$, $x^{b_i - b_\sigma} - \lambda_i (h_0(x)/x^{b_\sigma}) = 0, i = 1, \ldots, n_{\alpha_0}$, then the vector $(x^{b_i - b_\sigma} - z^{b_i - b_\sigma})_{i=1,\ldots,n_{\alpha_0}}$ is a multiple μv of the eigenvector $v = (\lambda_1, \ldots, \lambda_{n_{\alpha_0}})^\top$ of the rank-one matrix $(\lambda_i c_j)_{1 \le i,j \le n_{\alpha_0}}$. For $b_i = b_\sigma$ we have $\lambda_i \neq 0$, yet $x^{b_i - b_\sigma} - z^{b_i - b_\sigma} = 0$. We conclude $\mu = 0$. Hence $x^{b_i - b_\sigma} - z^{b_i - b_\sigma} = 0$ and $x \in G \cdot z$ by Lemma 5.5.4. □

In what follows, we derive a set of simple, non-homogeneous binomial equations on $\mathbb{C}^k$ defining a subvariety of $G \cdot z$.

Theorem 5.5.5. *Let $z \in U_{\sigma'}$ with $\sigma \in \widetilde{\Sigma}_P$ simplicial be such that $\pi(z)$ is not a basepoint of S_{α_0} and $\sigma^\vee \cap M = \mathbb{N}\mathcal{M}_\sigma + \mathbb{Z}\mathcal{M}_\sigma^\perp$. For generic $h_0 \in S_{\alpha_0}$ satisfying $h_0(z) \neq 0$, the affine variety*

$$Y_z = V_{\mathbb{C}^k}\left(x^{b_i} - \frac{z^{b_i}}{h_0(z)}, i = 1, \ldots, n_{\alpha_0} \right) \subset \mathbb{C}^k$$

is nonempty and $Y_z \subset G \cdot z$.

The proof of Theorem 5.5.5 uses the following lemma.

Lemma 5.5.5. *If $\alpha_0 \in \mathrm{Cl}(X)_+$, then α_0 is not a torsion element of* $\mathrm{Cl}(X)$.

Proof. Suppose $\ell\alpha_0 = 0$ for some $\ell \in \mathbb{N}_{>0}$. Then $F^\top m + \ell a_0 = 0$ for some $m \in M$, and therefore $F^\top(m/\ell) + a_0 = 0$. Since Σ_P is complete, this means that $P_0 = \{m/\ell\}$

and P_0 either has 1 lattice point if $m/\ell \in M$, or it has none. The latter situation is excluded by $\alpha_0 \in \mathrm{Cl}(X)_+$, since we can assume $0 \in P_0 \cap M$. Hence we have $m/\ell = m' \in M$ such that $F^\top m' + a_0 = 0$, which shows that $\alpha_0 = 0$. $\square$

Proof of Theorem 5.5.5. Since α_0 is not a torsion element of $\mathrm{Cl}(X)$ (Lemma 5.5.5), we have the exact sequence

$$0 \longrightarrow \mathbb{Z} \longrightarrow \mathrm{Cl}(X) \longrightarrow \mathrm{Cl}(X)/(\mathbb{Z} \cdot \alpha_0) \longrightarrow 0$$

where $\mathbb{Z} \to \mathrm{Cl}(X)$ sends $\ell \mapsto \ell\alpha_0 \in \mathrm{Cl}(X)$. Taking $\mathrm{Hom}_{\mathbb{Z}}(-, \mathbb{C}^*)$ shows that $G \to \mathbb{C}^* : g \mapsto g^{a_0}$ is surjective (because $\mathbb{C}^*$ is divisible). Therefore we can find $g \in G$ such that $g^{a_0} = h_0(z)^{-1}$ and thus $h_0(g \cdot z) = 1$. Every $x \in Y_z$ satisfies $x^{b_i} - (g \cdot z)^{b_i} = 0, i = 1, \ldots, n_{\alpha_0}$: this follows from $(g \cdot z)^{b_i} = z^{b_i}/h_0(z)$. In particular, $x^{b_\sigma} = (g \cdot z)^{b_\sigma} \neq 0$ ($z \in U_{\sigma'}$ and hence $g \cdot z \in U_{\sigma'}$ since $U_{\sigma'}$ is G-invariant) and therefore x satisfies $x^{b_i - b_\sigma} = (g \cdot z)^{b_i - b_\sigma}, i = 1, \ldots n_{\alpha_0}$. By Lemma 5.5.4 it follows that $g \cdot z \in Y_z \subset G \cdot z$. $\square$

Recall that $I \subset S$ is an ideal satisfying Assumptions 1-3 with $V_X(I) = \{\zeta_1, \ldots, \zeta_\delta\}$, $z_j \in \mathbb{C}^k \setminus Z$ is a set of homogeneous coordinates of ζ_j and we took α_0 such that no ζ_j is a basepoint of S_{α_0}. We have the following immediate corollary of Theorems 5.5.4 and 5.5.5.

Corollary 5.5.2. *Let $\lambda_{ij} = z_j^{b_i}/h_0(z_j)$ be the j-th eigenvalue of the i-th multiplication map $M_{x^{b_i}/h_0}$, $i = 1, \ldots, n_{\alpha_0}$, $j = 1, \ldots, \delta$. Assume that α_0 is such that, for $j = 1, \ldots, \delta$, $z_j \in U_{\sigma_j'}$ for a simplicial cone $\sigma_j \in \widetilde{\Sigma}_P$ satisfying $\sigma_j^\vee \cap M = \mathbb{N}\mathcal{M}_{\sigma_j} + \mathbb{Z}\mathcal{M}_{\sigma_j}^\perp$. For each j, we have that*

$$G \cdot z_j = V_{U_{\sigma_j'}}\left(x^{b_i - b_{\sigma_j}} - \lambda_{ij}\frac{h_0(x)}{x^{b_{\sigma_j}}}, i = 1, \ldots, n_{\alpha_0}\right) \subset U_{\sigma_j'}$$

and for any point $z_j' \in Y_{z_j} = V_{\mathbb{C}^k}(x^{b_i} - \lambda_{ij}, i = 1, \ldots, n_{\alpha_0}) \subset U_{\sigma_j'}$, we have $\pi(z_j') = \zeta_j$.

Corollary 5.5.2 implies that we can find homogeneous coordinates of the solutions from the eigenvalues λ_{ij} by solving a system of binomial equations

$$\{x^{b_i} - \lambda_{ij}, i = 1, \ldots, n_{\alpha_0}\} \tag{5.5.11}$$

provided that P_0 'has enough lattice points'. Concretely, for every point $\zeta_j \in V_X(I)$ there has to be a cone $\sigma_j \in \widetilde{\Sigma}_P$ such that $\zeta_j \in U_{\sigma_j}$ and $\sigma_j^\vee \cap M = \mathbb{N}\mathcal{M}_{\sigma_j} + \mathbb{Z}\mathcal{M}_{\sigma_j}^\perp$. Note that if all solutions are in the torus, then $\zeta_j \in U_\sigma$ for $\sigma = \{0\} \in \widetilde{\Sigma}_P$ and this condition translates to the fact that $\mathbb{Z}(P_0 \cap M - m) = M$ for some $m \in P_0 \cap M$. If P_0 is very ample, then $\widetilde{\Sigma}_P = \Sigma_P$ and $\sigma^\vee \cap M = \mathbb{N}\mathcal{M}_\sigma + \mathbb{Z}\mathcal{M}_\sigma^\perp$ holds for all $\sigma \in \Sigma_P$ [CLS11, Proposition 1.3.16].

We conclude this subsection with a discussion on how to solve the systems of binomial equations (5.5.11). Note that by Corollary 5.5.2, in this context it is enough to consider the system 'solved' once we have found *one point* on the variety $Y_{z_j}, j = 1, \dots, \delta$. This can be done using Newton iteration with the necessary adaptations. For instance, it should take the possibility of divergence into account and use a good criterion for convergence. For those ζ_j that are in the torus of X, there is a more clever way of doing this. For these solutions, all eigenvalues $\lambda_{ij}, i = 1, \dots, n_{\alpha_0}$ are nonzero. The method we describe here is suggested by Lemma 3.2 in [HS95]. Let $A = [b_1 \ \cdots \ b_{n_{\alpha_0}}] \in \mathbb{Z}^{k \times n_{\alpha_0}}$ be the matrix of exponents and compute its Smith normal form: $\mathbf{P}A\mathbf{Q} = \mathbf{S}$ with $\mathbf{P}, \mathbf{Q}$ unimodular and $\mathbf{S} = [\mathrm{diag}(s_1, \dots, s_r, 0, \dots, 0) \ 0] \in \mathbb{Z}^{k \times n_{\alpha_0}}$, where $s_i | s_{i+1}$. We make the substitution of variables $x_\ell = y_1^{\mathbf{P}_{1\ell}} \cdots y_k^{\mathbf{P}_{k\ell}}$ to obtain the equivalent system of equations given by $y^{\mathbf{P}b_i} = \lambda_{ij}$. Applying the invertible transformation given by the matrix $\mathbf{Q}$, this simplifies to

$$y_\ell^{s_\ell} = \prod_{i=1}^{n_{\alpha_0}} \lambda_{ij}^{\mathbf{Q}_{i\ell}}, \ \ell = 1, \dots, r \quad \text{and} \quad 1 = \prod_{i=1}^{n_{\alpha_0}} \lambda_{ij}^{\mathbf{Q}_{i\ell}}, \ r < \ell \leq k.$$

This imposes no conditions on $y_\ell, \ell > r$, so we can put $y_\ell = 1, \ell > r$. Taking the logarithm then shows that

$$\log y = [\log y_1 \ \cdots \ \log y_k] = [w \ 0_{k-r}]$$

where $w = [\log \lambda_{1j} \ \cdots \log \lambda_{n_{\alpha_0} j}][\mathbf{Q}_{:,1} \ \cdots \ \mathbf{Q}_{:,r}]\mathrm{diag}(1/s_1, \dots, 1/s_r)$ and 0_{k-r} is a row vector of length $k - r$ with zero entries. To find the homogeneous coordinates, we only need to invert our change of coordinates and the logarithm:

$$\log x = [\log x_1 \ \cdots \ \log x_k] = \log y \ \mathbf{P}, \qquad x_\ell = e^{\log x_\ell}, \ell = 1, \dots, k.$$

Taking the logarithm has some advantages for the implementation: it reduces all computations to some matrix multiplications and it may prevent overflow. Since the exponent matrix A is the same for all binomial systems (5.5.11), we can solve all systems for which this technique applies together by performing only one Smith normal form computation and a series of (small) matrix-matrix multiplications. We gather the eigenvalues λ_{ij} in a size $\delta^* \times n_{\alpha_0}$ matrix

$$\Lambda_{ji} = \lambda_{ij} = \frac{z_j^{b_i}}{h_0(z_j)},$$

whose $\delta^* \leq \delta$ rows correspond to the solutions z_j for which all λ_{ij} are nonzero (these are the solutions in the torus). This is a selection of the rows of the table we saw before, e.g., in Example 5.5.9. The resulting algorithm is Algorithm 5.4. After computing the Smith normal form, in line 3 we compute the entry-wise logarithm of the matrix Λ. In the next lines, we execute the steps explained above. In line 5, $0_{\delta^*,k-r}$ is a $\delta^* \times (k - r)$ matrix filled with zeros. The algorithm returns a set of homogeneous coordinates for each of the solutions represented by the rows of Λ.

As indicated before, Algorithm 5.4 fails for solutions on the boundary of the torus, for which some of the λ_{ij} are zero. We mentioned Newton iteration as an alternative.

Algorithm 5.4 Solves the binomial systems given by the exponents in A and the rows of $\Lambda \in \mathbb{C}^{\delta^* \times n_{\alpha_0}}$

procedure SolveBinomialSystem(A, Λ)
  $\mathbf{P}, \mathbf{Q}, \mathbf{S} \leftarrow$ Smith normal form of A
  $\log \Lambda \leftarrow (\log(\Lambda_{ij}))_{1 \le i \le \delta^*, 1 \le j \le n_{\alpha_0}}$
  $W \leftarrow \log \Lambda \, [\mathbf{Q}_{:,1} \;\cdots\; \mathbf{Q}_{:,r}] \operatorname{diag}(1/s_1, \dots, 1/s_r)$
  $\log Y \leftarrow [W \; 0_{\delta^*, k-r}]$
  $\log Z \leftarrow \log y\mathbf{P}$
  for $j = 1, \dots, \delta^*$ **do**
    $z'_j \leftarrow (e^{(\log Z)_{j1}}, \dots, e^{(\log Z)_{jk}})$
  end for
  return $z'_1, \dots, z'_{\delta^*}$
end procedure

There are other possibilities for dealing with this, such as dropping the equations in (5.5.11) for which $\lambda_{ij} = 0$ (which should be tested numerically using some robust criterion), and using the Smith normal form approach to solve for the remaining variables only. Note that if one is only interested in computing the solutions in the torus, computing the homogeneous coordinates for the solutions on the boundary can be skipped. We do not go into more detail here.

Now that we have presented *what to do* with the multiplication maps $M_{x^{b_i}/h_0}$ once we have them (i.e. find their eigenvalues and apply Algorithm 5.4), the next subsection will discuss *how to compute* the $M_{x^{b_i}/h_0}$.

5.5.4 Toric homogeneous normal forms

In this subsection we generalize the framework of homogeneous normal forms to the toric setting. With the definitions of the regularity and the homogeneous multiplication maps from Subsections 5.5.2 and 5.5.3, the proofs are identical to those in Section 4.5.

Definition 5.5.6 (Homogeneous normal form (HNF))**.** Let $I \subset S$ be a homogeneous ideal satisfying Assumptions 1-3. Let $(\alpha, \alpha_0) \in \mathrm{Cl}(X)_+^2$ be a regularity pair and let $B \subset S_d$ be a $\mathbb{C}$-vector subspace. A *homogeneous normal form (HNF)* of degree $\alpha + \alpha_0$ w.r.t. I is a $\mathbb{C}$-linear map $\mathcal{N}_{\alpha,\alpha_0} : S_{\alpha+\alpha_0} \to B$ such that

$$0 \longrightarrow I_{\alpha+\alpha_0} \longrightarrow S_{\alpha+\alpha_0} \xrightarrow{\mathcal{N}_{\alpha,\alpha_0}} B \longrightarrow 0$$

is a short exact sequence and for some $h_0 \in S_{\alpha_0}$ satisfying $V_X(h_0) \cap V_X(I) = \varnothing$,

$$\begin{tikzcd} B \arrow[r] & (S/I)_\alpha \\ (S/I)_{\alpha+\alpha_0} \arrow[u, "\overline{\mathcal{N}}"] & (S/I)_\alpha \arrow[l, "M_{h_0}"'] \arrow[u, "\mathrm{id}"'] \end{tikzcd}$$

commutes, where $B \to (S/I)_\alpha$ is given by $b \mapsto b + I_\alpha$ and $\overline{\mathcal{N}}(f + I_{\alpha+\alpha_0}) = \mathcal{N}_{\alpha,\alpha_0}(f)$.

In Definition 5.5.6, the maps id and $\overline{\mathcal{N}}$ are isomorphisms of $\mathbb{C}$-vector spaces. We have seen (Lemma 5.5.3) that M_{h_0} is an isomorphism as well, hence $B \simeq (S/I)_\alpha$ via $b \mapsto b + I_\alpha$. Definition 5.5.6 should be slightly adapted when we want to consider the more general case where the points in $V_X(I)$ are allowed to have multiplicities. More precisely, we need a different notion of *regularity*. We will say a few things about this in Subsection 5.5.5 but stick to the case where all points have multiplicity 1 for now.

Just like in the projective case, if we want to specify the function $h_0 \in S_{\alpha_0}$ in Definition 5.5.6, we say that $\mathcal{N}_{\alpha,\alpha_0}$ is a HNF with respect to I *and* h_0. The way homogeneous multiplication matrices are obtained from homogeneous normal forms should come as no surprise. For a HNF $\mathcal{N}_{\alpha,\alpha_0}$ and $g \in S_{\alpha_0}$ we define $\mathcal{N}_g : S_\alpha \to B$ by $\mathcal{N}_g(f) = \mathcal{N}_{\alpha,\alpha_0}(fg)$.

Proposition 5.5.4. *Let I, α, α_0, B be as in Definition 5.5.6. If $\mathcal{N}_{\alpha,\alpha_0}$ is a HNF with respect to I and $h_0 \in S_{\alpha_0}$, then for any $g \in S_{\alpha_0}$, $(\mathcal{N}_g)_{|B} : B \to B$ is similar to the map $M_{g/h_0} = M_{h_0}^{-1} \circ M_g$ from Theorem 5.5.3.*

Proof. The proof is identical to that of Proposition 4.5.1. ☐

Definition 5.5.7. Let I, α, α_0, B be as in Definition 5.5.6. A $\mathbb{C}$-linear map $N : S_{\alpha+\alpha_0} \to \mathbb{C}^\delta$ *covers* a HNF $\mathcal{N}_{\alpha,\alpha_0} : S_{\alpha+\alpha_0} \to B$ with respect to I if there is an isomorphism $P : B \to \mathbb{C}^\delta$ such that $\mathcal{N}_{\alpha,\alpha_0} = P^{-1} \circ N$.

Proposition 5.5.5. *Let $I \subset S$ be a zero-dimensional homogeneous ideal satisfying Assumptions 1-3. Let $(\alpha, \alpha_0) \in \mathrm{Cl}(X)_+^2$ be a regularity pair. A $\mathbb{C}$-linear map $N : S_{\alpha+\alpha_0} \to \mathbb{C}^\delta$ covers a HNF if and only if*

$$0 \longrightarrow I_{\alpha+\alpha_0} \longrightarrow S_{\alpha+\alpha_0} \xrightarrow{N} \mathbb{C}^\delta \longrightarrow 0 \qquad (5.5.12)$$

is a short exact sequence. In this case, N covers a HNF $\mathcal{N}_{\alpha,\alpha_0} : S_{\alpha+\alpha_0} \to B$ with respect to I and h_0 for any $h_0 \in S_{\alpha_0}$ such that $V_X(h_0) \cap V_X(I) = \varnothing$ and for any δ-dimensional subspace $B \subset S_\alpha$ such that

$$(N_{h_0})_{|B} : B \to \mathbb{C}^\delta$$

is invertible, where $N_{h_0} : S_\alpha \to \mathbb{C}^\delta$ is given by $N_{h_0}(f) = N(h_0 f)$. The HNF $\mathcal{N}_{\alpha,\alpha_0}$ is given by $\mathcal{N}_{\alpha,\alpha_0} = (N_{h_0})_{|B}^{-1} \circ N$.

Proof. Note that N_{h_0} is surjective by Lemma 5.5.3, so there is some δ-dimensional $\mathbb{C}$-vector subspace for which the restriction $(N_{h_0})_{|B}$ is invertible. The proof of the proposition is identical to that of Proposition 4.5.2. ☐

We conclude from Proposition 5.5.5 that if for a regularity pair (α, α_0) we have computed a $\mathbb{C}$-linear map $N : S_{\alpha+\alpha_0} \to \mathbb{C}^\delta$ such that (5.5.12) is exact, then for any

$h_0 \in S_{\alpha_0}$ such that $V_X(h_0) \cap V_X(I) = \varnothing$ and any $B \subset S_\alpha$ such that $(N_{h_0})_{|B}$ is invertible, we have that for any $g \in S_{\alpha_0}$, 'multiplication with g/h_0' is given by

$$M_{g/h_0} = (N_{h_0})_{|B}^{-1} \circ (N_g)_{|B},$$

where $N_g : S_\alpha \to \mathbb{C}^\delta$ is given by $N_g(f) = N(fg)$.

As in the projective case, we compute a map $N : S_{\alpha+\alpha_0} \to \mathbb{C}^\delta$ such that (5.5.12) is exact as a cokernel map of a map whose image is $I_{\alpha+\alpha_0}$. To this end, we extend the definition of a graded resultant map (Definition 4.3.2) to the toric case.

Definition 5.5.8 (Graded resultant map)**.** Fix $\alpha \in \mathrm{Cl}(X)_+$. For a tuple $(f_1, \ldots, f_s) \in S_{\alpha_1} \times \cdots \times S_{\alpha_s}$ with $\alpha_i \in \mathrm{Cl}(X)_+$ and finite dimensional $\mathbb{C}$-vector subspaces $\Lambda_i \subset S_{\alpha-\alpha_i}, i = 1, \ldots, s$, $\Lambda = S_\alpha$, the *graded resultant map* is the $\mathbb{C}$-linear map

$$\mathrm{res}_{f_1,\ldots,f_s} : \Lambda_1 \times \cdots \times \Lambda_s \to \Lambda \quad \text{given by} \quad \mathrm{res}_{f_1,\ldots,f_s}(q_1, \ldots, q_s) = q_1 f_1 + \cdots + q_s f_s.$$

Suppose $(\alpha, \alpha_0) \in \mathrm{Cl}(X)_+^2$ is a regularity pair for $I = \langle f_1, \ldots, f_s \rangle$. The graded resultant map

$$\mathrm{res}_{f_1,\ldots,f_s} : \Lambda_1 \times \cdots \times \Lambda_s \to \Lambda \quad \text{with} \quad \Lambda = S_{\alpha+\alpha_0}, \Lambda_i = S_{\alpha+\alpha_0-\deg(f_i)} \tag{5.5.13}$$

has the property that $\mathrm{im}\,\mathrm{res}_{f_1,\ldots,f_s} = I_{\alpha+\alpha_0}$. A cokernel map $N : \Lambda \to \mathbb{C}^\delta$ therefore satisfies (5.5.12) and covers a HNF by Proposition 5.5.5. This leads to Algorithm 5.5 for computing the homogeneous multiplication matrices in the toric setting. The

Algorithm 5.5 Computes homogeneous multiplication matrices for $I = \langle f_1, \ldots, f_s \rangle \subset S$ satisfying Assumptions 1-3

1: **procedure** HOMOGENEOUSMULTIPLICATIONMATRICES$(f_1, \ldots, f_s, (\alpha, \alpha_0))$
2: $\quad \mathrm{res}_{f_1,\ldots,f_s} \leftarrow$ the resultant map $\Lambda_1 \times \cdots \times \Lambda_s \to \Lambda$ from (5.5.13)
3: $\quad N \leftarrow \mathrm{coker}\,\mathrm{res}_{f_1,\ldots,f_s}$
4: $\quad h_0 \leftarrow$ generic element of S_{α_0}
5: $\quad N_{h_0} \leftarrow$ matrix of the map $S_\alpha \to \mathbb{C}^\delta$ where $f \mapsto N(h_0 f)$
6: $\quad (N_{h_0})_{|B} \leftarrow$ invertible restriction of N_{h_0} to $B \subset S_\alpha$, $\dim_{\mathbb{C}} B = \delta$
7: $\quad$ **for** $i = 1, \ldots, n_{\alpha_0}$ **do**
8: $\quad\quad (N_{x^{b_i}})_{|B} \leftarrow$ restriction of the map $S_\alpha \to \mathbb{C}^\delta$ given by $f \mapsto N(x^{b_i} f)$ to B
9: $\quad\quad M_{x^{b_i}/h_0} \leftarrow (N_{h_0})_{|B}^{-1} (N_{x^{b_i}})_{|B}$
10: $\quad$ **end for**
11: $\quad$ **return** $M_{x^{b_1}/h_0}, \ldots, M_{x^{b_{n_{\alpha_0}}}/h_0}$
12: **end procedure**

algorithm takes homogeneous generators for I and a regularity pair as its input. It returns the multiplication matrices corresponding to all monomials of degree α_0. The usual remarks concerning the basis choice in line 6 apply. Note that Algorithm 5.5 also provides a generalization of Algorithm 4.2 in the non-square case.

The case we are particularly interested in is that where $n = s$ and $I = \langle f_1, \ldots, f_n \rangle$ where f_i comes from *homogenizing* $\hat{f}_i$ and $X = X_P$ where $P = P_1 + \ldots + P_n$ is the sum of $P_i = \mathrm{Newt}(\hat{f}_i), i = 1, \ldots, n$. In this case $\alpha_i = \deg(f_i) \in \mathrm{Pic}(X)$ is basepoint free. We will show in Subsection 5.5.5 that if $V_X(I)$ is zero-dimensional, $\alpha = \alpha_1 + \cdots + \alpha_n \in \mathrm{Reg}(I)$. Moreover, for any basepoint free $\alpha_0 \in \mathrm{Pic}(X)$, $\alpha + \alpha_0 \in \mathrm{Reg}(I)$. Hence $(\alpha, \alpha_0) \in \mathrm{Cl}(X)_+^2$ is a regularity pair. We observe in experiments that it is too strict to require α_0 to be basepoint free and contained in $\mathrm{Pic}(X)$. In practice, one can work with any α_0 such that

$$(\alpha, \alpha_0) \text{ is a regularity pair and Corollary 5.5.2 applies for } \alpha_0. \tag{5.5.14}$$

Concretely, the polytope P_0 associated to α_0 should have 'enough lattice points', see the discussion following Corollary 5.5.2. This leads to Algorithm 5.6 for computing the homogeneous coordinates of the solutions. Lines 6 and 7 use some notation introduced

Algorithm 5.6 Computes homogeneous coordinates on $X = X_{P_1+\cdots+P_n}$ of the solutions of $(\hat{f}_1, \ldots, \hat{f}_n) \in \mathcal{F}_{\mathbb{C}[M]}(P_1, \cdots, P_n)$ where $I = \langle f_1, \ldots, f_n \rangle \subset S$ satisfies Assumptions 1-3

1: **procedure** SolveHomogeneous($\hat{f}_1, \ldots, \hat{f}_n$)
2: $\quad f_1, \ldots, f_n \leftarrow$ homogenize $\hat{f}_1, \ldots, \hat{f}_n$
3: $\quad \alpha \leftarrow \deg(f_1) + \cdots + \deg(f_n) \in \mathrm{Pic}(X)$
4: $\quad \alpha_0 \leftarrow$ element of $\mathrm{Cl}(X)_+$ such that (5.5.14) is satisfied
5: $\quad \{M_{x^{b_i}/h_0}\} \leftarrow$ HomogeneousMultiplicationMatrices($f_1, \ldots, f_n, (\alpha, \alpha_0)$)
6: $\quad A \leftarrow$ exponent matrix in $\mathbb{Z}^{k \times n_{\alpha_0}}$ of the monomials in S_{α_0}
7: $\quad \Lambda \leftarrow$ eigenvalues of $M_{x^{b_1}/h_0}, \ldots, M_{x^{b_{n_{\alpha_0}}}/h_0}$ such that $\Lambda_{ji} = \lambda_{ij}$
8: $\quad$ **return** SolveBinomialSystem(A, Λ)
9: **end procedure**

in Subsection 5.5.3. In Line 8, Algorithm 5.4 is used to solve the binomial systems (5.5.11). As we have mentioned before, the Smith normal form based algorithm will only work for solutions in the torus. For the other solutions, one has to adapt the solving method. For simplicity, in Algorithm 5.6 we assume that SolveBinomialSystem takes care of this. The approach taken in the experiments below is the same as in [Tel20, Algorithm 1]. The Smith normal form method is used for a solution ζ_j for which

$$\min_{1 \le i \le n_{\alpha_0}} |\lambda_{ij}| > \left(\sum_{i=1}^{n_{\alpha_0}} |\lambda_{ij}|^2 \right)^{1/2} \mathrm{tol}, \tag{5.5.15}$$

where tol is a predefined tolerance. For solutions not satisfying (5.5.15), an adapted Newton iteration is applied for solving the corresponding binomial system. Once the homogeneous coordinates $z_j = (z_{j,1}, \ldots, z_{j,k})$ are computed, we can obtain the coordinates of these solutions in the torus via the Laurent monomial map (5.5.2). The following is Remark 6.1 in [Tel20].

Remark 5.5.5. We briefly discuss the complexity of Algorithm 5.5 as compared to Algorithm 5.3. The first step in both algorithms is to compute the cokernel of a resultant map res. Since for both algorithms the monomials indexing the vector spaces V and Λ in the definition of res are the lattice points contained in a slightly enlarged (and shifted) version of the polytope $P = P_1 + \ldots + P_n$, this step takes roughly the same computation time for both algorithms. Even though the Cox ring has dimension $k > n$, the dimensions of its graded pieces correspond to the lattice points contained in n-dimensional polytopes. *The grading of S by the class group is such a fine grading that it's almost like we are only* implicitly *working with k variables instead of n.* This is an important observation, because for larger problems, the computation of the cokernel of res is the most expensive step of the algorithm. Next, both algorithms compute the multiplication matrices from this cokernel. This is more expensive for Algorithm 5.5: there are more multiplication maps. Another important difference is that for the TNF algorithm, the eigenvalues of the multiplication maps immediately give the coordinates of the solutions, whereas Algorithm 5.6 processes these eigenvalues to find the homogeneous coordinates by solving binomial systems of equations. We conclude that Algorithm 5.6 is computationally more expensive overall. This should be considered the price that is payed for being more robust in nearly degenerate situations, which is our main reason for developing the algorithm. However, the increase of complexity is not dramatic: systems with thousands of solutions can be solved within reasonable time (see the experiments below). $\triangle$

We conclude the subsection with some experiments illustrating the effectiveness of Algorithm 5.6. They are taken from [Tel20, Section 7]. We use a Matlab implementation of Algorithm 5.6. As in Section 5.3, we call Polymake from Matlab for all computations involving polytopes, except for the mixed volume computation, which is done using PHCpack. To reduce the overhead caused by calling Polymake through Matlab we have implemented an *online* and an *offline* version of the algorithm. The offline version takes the polytope information as an input. The online version computes everything from the input polynomials and automatically generates an α_0 whose lattice points affinely generate M. The basis selection is done using the SVD and all eigenvalue computations use the Schur factorization. The experiments were executed on the same machine. To measure the quality of an approximate solution, we compute the residual of the dehomogenized solutions as detailed in Appendix C. The goal of the experiments is to show that Algorithm 5.6 meets our objectives: it finds *all* solutions with *good accuracy* within reasonable time. In particular, it does so for (nearly) degenerate systems with solutions on or near the exceptional divisors of X that cannot be solved by other state of the art solvers.

Experiment 5.5.1 (Points on $\mathscr{H}_2$)**.** We finish our running example by using Algorithm 5.6 to compute homogeneous coordinates of the solutions of the system defined in Example 5.4.5. We use tol $= 10^{-12}$, $\alpha = \alpha_1 + \alpha_2$. For $\alpha_0 = \alpha_2$, Algorithm 5.6 finds three solutions. All three residuals are of order 10^{-16}.

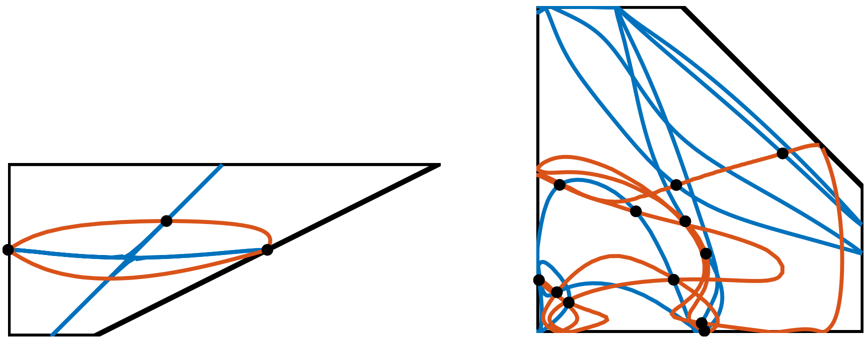


Figure 5.13: Left: images in P of the real part of $V(f_1)$ and $V(f_2)$ from Example 5.4.3 under the moment map μ. The images of the computed real solutions are shown as black dots. Right: same picture for a different system.

To illustrate the results, we use the *moment map*

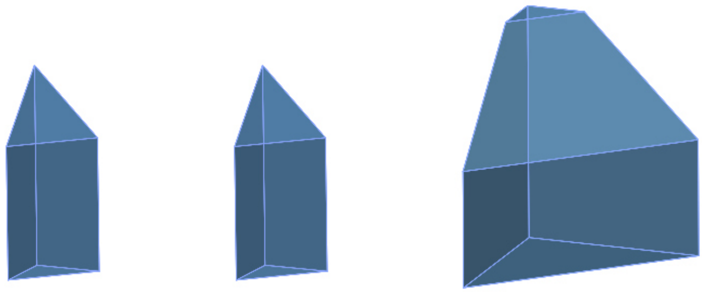
$$\mu : \mathbb{C}^k \setminus Z \to P : x \mapsto \frac{1}{\sum_{m \in P \cap M} |x^{F^\top m + a}|} \sum_{m \in P \cap M} |x^{F^\top m + a}| m,$$

where $|\cdot|$ denotes the modulus. The map μ is constant on G-orbits and takes a point $x \in \mathbb{C}^k \setminus Z$ to a convex combination of the lattice points of P. It has the property that torus invariant prime divisors are sent to their corresponding facets and $(\mathbb{C}^*)^k$ is sent to the interior of P. More information can be found in [Ful93, Section 4.2] and [Sot17, Section 2]. Figure 5.13 shows that two of the computed solutions lie on divisors and one is in the torus. The image under μ of all of the solutions must lie on an intersection of the images of $V(f_1) \setminus Z, V(f_2) \setminus Z$ (but not all intersections correspond to solutions). As an illustration, we have included the same picture for a system with more solutions in the right part of the same figure. The polytopes for this system are $P_1 = [0,4] \times [0,4]$ and $P_2 = 5\Delta_2$ where Δ_2 is the standard simplex. There are $\delta = 40$ solutions, 12 of them are real. $\triangle$

Experiment 5.5.2 (A problem from computer vision)**.** The author is grateful to Tomas Pajdla and Zuzana Kukelova for suggesting this example. One of the so-called 'minimal problems' in computer vision is the problem of estimating radial distortion from eight point correspondences in two images. In [KP07], Kukelova and Pajdla propose a formulation of this problem as a system of 3 polynomial equations in 3 unknowns. The Newton polytopes are visualized in Figure 5.14. The mixed volume is $\delta = \mathrm{MV}(P_1, P_2, P_3) = 17$ and the matrix of facet normals is

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

so the Cox ring S has dimension 6. We assign random real coefficients drawn from a standard normal distribution to all lattice points in the polytopes and solve the system



(a) P_1 (b) P_2 (c) P_3

Figure 5.14: Newton polytopes of the equations of the eight point radial distortion problem.

using Algorithm 5.6. We first run the offline version, which generates the polytope P_0. In this case, P_0 is the standard simplex. All 17 solutions are found with a residual of order 10^{-16} within ± 0.1 s (using the online version of the algorithm). To show the robustness of Algorithm 5.6 in the nearly degenerate case, i.e. the case where there are solutions on or near the torus invariant prime divisors, we perform the following experiment. Consider the lattice points

$$\mathscr{F}_3 = \{m \in P_1 \cap M \mid \langle u_3, m\rangle + 3 = 0\}, \qquad \mathscr{G}_3 = (P_1 \cap M) \setminus \mathscr{F}_3.$$

The points in $\mathscr{F}_3$ are the lattice points on the facet of P_1 corresponding to $u_3 = (-1,-1,-1)$. Set

$$\hat{g}_i = \sum_{m \in \mathscr{F}_3} c_{m,i} t^m + \sum_{m \in \mathscr{G}_3} c_{m,i} t^m, \qquad i = 1, 2$$

with $c_{m,i}$ real numbers drawn from a standard normal distribution. Now let $\hat{f}_1 = \hat{g}_1$ and

$$\hat{f}_2(e) = \sum_{m \in \mathscr{F}_3} (10^{-e} c_{m,2} + (1 - 10^{-e}) c_{m,1}) t^m + \sum_{m \in \mathscr{G}_3} c_{m,2} t^m, \quad e \in [0, \infty).$$

The equation $\hat{f}_2 = 0$ is parametrized by the real parameter e. The third equation $\hat{f}_3 = 0$ is chosen randomly. When $e = 0$, $\hat{f}_2 = \hat{g}_2$ and the system is generic, as before. When $e \to \infty$, the part of $\hat{f}_2$ corresponding to $\mathscr{F}_3$ converges to the part of $\hat{f}_1$ corresponding to $\mathscr{F}_3$, meaning that there will be solutions 'at infinity' on the divisor D_3. We solve the system for $e = 0, 1/2, 1, 3/2, \ldots, 16$ and compute both the maximal residual $r_{\max}$ and the minimal residual $r_{\min}$ for the 17 solutions found by Algorithm 5.6 with tol $= 10^{-4}$ and the solutions found by Algorithm 5.3. The result of the experiment is shown in Figure 5.15. Note that not only the residuals of the solutions approaching the divisor deteriorate for the TNF algorithm. Accuracy is lost on *all* solutions. The reason is that even for the 'best' basis selected by this algorithm, the computation of the classical multiplication matrices is ill-conditioned because the system is nearly degenerate. Looking at the computed Cox coordinates, we see that for three of the solutions, the coordinate x_3 goes to zero as e increases, so 3 out of 17 solutions approach the divisor D_3.

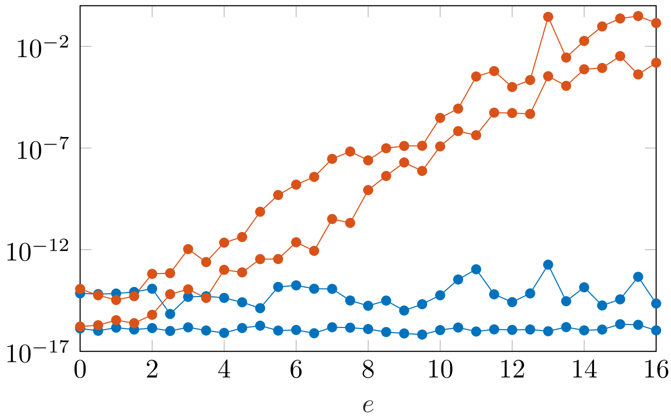


Figure 5.15: Minimal and maximal residual for different values of the parameter e for the parametrized eight point radial distortion problem, for Algorithm 5.6 (blue) and Algorithm 5.3 (orange).

One can perform the same experiment for any other facet of P_1. However, in order to find the solutions on the divisors, the polytope P_0 must be large enough and it might not be sufficient that its lattice points generate the lattice (Corollary 5.5.2). Repeating the same experiment, but this time using $\mathcal{F}_2$ instead of $\mathcal{F}_3$, the solutions in the torus are still found with good accuracy by Algorithm 5.6. Accuracy is lost on the solutions approaching D_2. The reason is that the standard simplex does not 'show' this facet. Using $P_0 = \mathrm{Conv}((0,0,0),(1,0,0),(0,1,0),(0,0,1),(1,0,1),(0,1,1),(0,0,2))$ we find homogeneous coordinates of all solutions. $\triangle$

Experiment 5.5.3 (Generic problems)**.** To give an idea of the computation time and the type of systems Algorithm 5.6 can handle, we perform the following experiment. Consider the parameters n, NZ, $d_{\max} \in \mathbb{N} \setminus \{0\}$. For $j = 1, \ldots, n$ we generate a set $\mathscr{A}_j \subset \mathbb{Z}^n$ of NZ lattice points by selecting NZ points in $\mathbb{N}^n$ with coordinates drawn uniformly from $\{0, 1, \ldots, d_{\max}\}$ and shifting these points by substracting the first point from all other points. Then for each $m \in \mathscr{A}_j$ we generate a random real number $c_{m,j}$ drawn from a standard normal distribution and we set

$$\hat{f}_j = \sum_{m \in \mathscr{A}_j} c_{m,j} t^m.$$

If two or more points $m \in \mathscr{A}_j$ coincide, we add the $c_{m,j}$ together, so NZ is an upper bound for the number of terms in $\hat{f}_j$. We use Algorithm 5.6 to compute the Cox

n	NZ	$d_{\max}$	δ	k	n_{α_0}	OFFLINE			ONLINE		
						t	D_{mean}	$D_{\max}$	t	D_{mean}	$D_{\max}$
2	20	10	144	12	3	1.9e+1	15	14	2.0e-1	15	14
2	20	20	505	14	4	2.4e+1	14	12	1.9e+0	14	11
2	20	30	1268	15	3	5.8e+1	14	12	1.9e+1	14	12
2	20	40	2390	16	3	2.6e+2	14	11	1.4e+2	14	13
2	20	50	3275	16	3	3.7e+2	14	12	2.3e+2	14	11
2	20	60	4469	12	3	7.8e+2	11	7	5.2e+2	11	8
2	40	30	1522	15	3	9.5e+1	14	11	3.4e+1	14	10
2	60	30	1670	15	4	1.2e+2	14	12	5.3e+1	14	12
2	200	30	1672	10	3	1.1e+2	15	10	6.0e+1	15	9
3	5	3	18	21	4	2.2e+1	14	12	1.1e-1	15	13
3	5	5	136	36	4	3.9e+1	14	9	6.3e-1	14	13
3	10	5	190	60	5	3.5e+1	15	7	2.1e+0	15	11
3	10	7	592	63	5	1.3e+2	14	10	3.2e+1	15	7
4	5	3	81	106	6	6.9e+1	14	11	3.7e+1	14	11

Table 5.4: Results for generic systems with mixed supports.

coordinates of the solutions of the resulting system and their image under (5.5.2). In Table 5.4 we report the number of solutions δ, the dimension k of the Cox ring, the number n_{α_0} for the automatically generated α_0, and, for both the offline and the online solver, the maximal residual $r_{\max}$, the geometric mean of the residuals of all solutions r_{mean} and the computation time t (in seconds). The residuals are represented by $D_{\text{mean}} = \lceil -\log_{10} r_{\text{mean}} \rceil$ and $D_{\max} = \lceil -\log_{10} r_{\max} \rceil$. It follows from Bernstein's second theorem [Ber75, HS95] that solutions on divisors can only occur if the polytopes involved have common tropisms corresponding to positive dimensional faces. An important case in which this may happen is the unmixed case in which all input polytopes are equal. We repeat the experiment, but this time we keep the supports $\mathscr{A} = \mathscr{A}_1 = \ldots = \mathscr{A}_n$ fixed. Table 5.5 shows some results. Of course, for this type of systems, the dimension of the Cox ring (or, equivalently, the number of facets of the Minkowski sum of the input polytopes) is lower and the system of binomial equations from Corollary 5.5.2 is easier to solve. $\triangle$

n	NZ	$d_{\max}$	δ	k	n_{α_0}	OFFLINE			ONLINE		
						t	D_{mean}	$D_{\max}$	t	D_{mean}	$D_{\max}$
2	20	60	3638	7	3	5.8e+2	13	11	3.8e+2	13	10
3	10	10	834	14	6	3.5e+2	13	12	1.9e+2	13	12
4	6	3	15	7	8	3.3e+1	15	15	8.4e-1	15	14
4	6	4	28	6	11	4.3e+1	14	13	5.4e+0	15	14
4	6	5	216	9	7	5.7e+2	12	11	2.7e+2	12	11
4	6	6	339	8	6	1.5e+3	6	4	2.0e+3	6	5
5	6	3	10	6	8	7.5e+1	15	14	1.0e+1	15	15

Table 5.5: Results for generic systems with unmixed supports.

5.5.5 More on regularity and fat points

In this subsection we discuss how the results from the previous subsections generalize to the case where some of the points in $V_X(I)$ have multiplicity > 1 and we state some results about the regularity $\mathrm{Reg}(I)$. The presented material is taken from [Tel20] and from [BT20a]. Throughout the subsection, $I = \langle f_1, \ldots, f_s \rangle \subset S$ is a homogeneous ideal such that $V_X(I)$ is zero-dimensional, consisting of the δ points $\{\zeta_1, \ldots, \zeta_\delta\}$ with multiplicities $\mu_1, \ldots, \mu_\delta$ respectively. We set $\delta^+ = \mu_1 + \cdots + \mu_\delta$. We say that $V_X(I)$ has *degree* δ^+. We will also assume that the f_i are homogeneous of degree $\deg(f_i) = \alpha_i \in \mathrm{Pic}(X)$. The fan of X is denoted by Σ and as before, we let $J = (I : \mathfrak{B}^\infty)$. When we need the extra assumption that all points have multiplicity 1 ($\delta = \delta^+$), we will say that $V_X(I)$ is *reduced.* In the non-reduced case, Definition 5.5.4 for the regularity of I is not the right one to use.

Definition 5.5.9 (Regularity (general case))**.** Let $I \subset S$ be such that $V_X(I) = \{\zeta_1, \ldots, \zeta_\delta\}$ is zero-dimensional of degree δ^+ and let $J = (I : \mathfrak{B}^\infty)$. The *regularity* $\mathrm{Reg}(I) \subset \mathrm{Cl}(X)$ of I is

$$\mathrm{Reg}(I) = \{\alpha \in \mathrm{Cl}(X) \mid \mathrm{HF}_I(\alpha) = \delta^+, I_\alpha = J_\alpha, \text{ no } \zeta_j \text{ is a basepoint of } S_\alpha\}.$$

We say that $(\alpha, \alpha_0) \in \mathrm{Cl}(X)_+^2$ is a *regularity pair* if $\alpha, \alpha + \alpha_0 \in \mathrm{Reg}(I)$ and no ζ_j is a basepoint of S_{α_0}.

Although we change the definition slightly, we keep the same notation for $\mathrm{Reg}(I)$ as before. The new definition does not change anything for statements about degrees in $\mathrm{Reg}(I) \cap \mathrm{Pic}(X)$ in the reduced case, for which the two definitions coincide (Proposition 5.5.3). We will add a remark where there is danger for confusion. For $\alpha \in \mathrm{Pic}(X)$ and $f \in S_\alpha$, we denote f^σ for the *dehomogenization* of f with respect to the affine chart $U_\sigma \subset X$ as in (5.5.6). The ideal defined by I in $\mathbb{C}[U_\sigma]$ (i.e. the sections of the ideal sheaf $\mathscr{I}$ on U_σ) is denoted by $\mathscr{I}(U_\sigma) = \langle f_1^\sigma, \ldots, f_s^\sigma \rangle \subset \mathbb{C}[U_\sigma]$. For $\sigma \in \Sigma(n), \alpha \in \mathrm{Pic}(X)$, we denote the operation of 'dehomogenization modulo the ideal I' by

$$\eta_{\alpha,\sigma}^{-1} : (S/I)_\alpha \to \mathbb{C}[U_\sigma]/\mathscr{I}(U_\sigma) \quad \text{where} \quad \eta_{\alpha,\sigma}^{-1}(f + I_\alpha) = f^\sigma + \mathscr{I}(U_\sigma).$$

Note that this is well-defined since for any $f \in I_\alpha$ we can find homogeneous $g_i \in S_{\alpha - \alpha_i}$ such that $f = g_1 f_1 + \cdots + g_s f_s$ and $f^\sigma = g_1^\sigma f_1^\sigma + \cdots + g_s^\sigma f_s^\sigma \in \mathscr{I}(U_\sigma)$.

Lemma 5.5.6. *For $\alpha \in \mathrm{Reg}(I) \cap \mathrm{Pic}(X)$ we have that $f \in I_\alpha$ if and only if $f^\sigma \in \mathscr{I}(U_\sigma)$ for all $\sigma \in \Sigma(n)$.*

Proof. It is clear that $f \in I_\alpha$ implies $f^\sigma \in \mathscr{I}(U_\sigma)$, for all $\sigma \in \Sigma(n)$. Conversely, suppose that $f^\sigma \in \mathscr{I}(U_\sigma)$ for all $\sigma \in \Sigma(n)$. Then we can find $g_1^\sigma, \ldots, g_s^\sigma$ such that

$$f^\sigma = g_1^\sigma f_1^\sigma + \cdots + g_s^\sigma f_s^\sigma.$$

Note that this is an equality in the localization $S_{x^{\hat{\sigma}}}$, and clearing denominators shows that there is $\ell \in \mathbb{N}$ such that $(x^{\hat{\sigma}})^\ell f \in I$. Since $\mathfrak{B} = \langle x^{\hat{\sigma}} \mid \sigma \in \Sigma(n) \rangle$, we have that $f \in (I : \mathfrak{B}^\infty) = J$. Since $f \in S_\alpha$ and $\alpha \in \mathrm{Reg}(I)$, this implies $f \in I_\alpha$. $\square$

For each $\zeta_i \in V_X(I)$, let $\sigma_i \in \Sigma(n)$ be such that $\zeta_i \in U_{\sigma_i}$. We will use an embedding of U_{σ_i} in an affine space $\mathbb{C}^{n_{\sigma_i}}$ in order to apply the theory developed in Subsection 3.1.3. We denote the coordinate ring of this affine space $\mathbb{C}^{n_{\sigma_i}}$ by R_{σ_i}. The embedding $U_{\sigma_i} \to \mathbb{C}^{n_{\sigma_i}}$ gives an isomorphism $\mathbb{C}[U_{\sigma_i}]/\mathscr{I}(U_{\sigma_i}) \simeq R_{\sigma_i}/I_{\sigma_i}$ for some zero-dimensional ideal $I_{\sigma_i} \subset R_{\sigma_i}$. We denote the $\mathbb{C}$-vector space of differential operators on $\mathbb{C}^{n_{\sigma_i}}$ by $\mathscr{D}_{\sigma_i}$. The point ζ_i corresponds to some primary ideal $Q_i \subset R_{\sigma_i}$ containing I_{σ_i}, which gives a closed subspace $D_i \subset \mathscr{D}_{\sigma_i}$ of dimension $\dim_{\mathbb{C}} D_i = \mu_i$ (Theorem 3.1.3). Let $\partial_{i1}, \ldots, \partial_{i\mu_i}$ be a consistently ordered basis for D_i. Note that $\mathrm{ev}_{\zeta_i} \circ \partial_{ij}$ gives an element of $(R_{\sigma_i}/I_{\sigma_i})^\vee \simeq (\mathbb{C}[U_{\sigma_i}]/\mathscr{I}(U_{\sigma_i}))^\vee$ (see the discussion following Theorem 3.1.3). For $\alpha \in \mathrm{Pic}(X)$, $i = 1, \ldots, \delta$ and $j = 1, \ldots, \mu_i$ we define

$$v_{ij,\alpha} : (S/I)_\alpha \to \mathbb{C} \quad \text{with} \quad v_{ij,\alpha} = \mathrm{ev}_{\zeta_i} \circ \partial_{ij} \circ \eta^{-1}_{\alpha,\sigma_i}.$$

Consider the map $\psi_\alpha : (S/I)_\alpha \to \mathbb{C}^{\delta^+}$ given by

$$\psi_\alpha(f + I_\alpha) = (v_{ij,\alpha}(f + I_\alpha) \mid i = 1, \ldots, \delta, j = 1, \ldots, \mu_i). \tag{5.5.16}$$

If $V_X(I)$ is reduced, this is the map ψ_α from (5.5.8) up to an invertible diagonal scaling.

Proposition 5.5.6. *For $\alpha \in \mathrm{Reg}(I) \cap \mathrm{Pic}(X)$, the map ψ_α from (5.5.16) is an isomorphism of $\mathbb{C}$-vector spaces.*

Proof. The condition $v_{ij,\alpha}(f + I_\alpha) = 0, i = 1, \ldots, \mu_i$ is independent from the choice of $\sigma_i \in \Sigma(n)$ such that $\zeta_i \in U_{\sigma_i}$. We have that $v_{ij,\alpha}(f+I_\alpha) = 0, i = 1, \ldots, \delta, j = 1, \ldots, \mu_i$ if and only if $f^\sigma \in \mathscr{I}(U_\sigma)$ for all $\sigma \in \Sigma(n)$. Because $\alpha \in \mathrm{Reg}(I)$, Lemma 5.5.6 applies. We conclude that ψ_α is an injective map between $\mathbb{C}$-vector spaces of the same dimension. The proposition follows. $\square$

Theorem 5.5.6 (Toric eigenvalue, eigenvector theorem (non-reduced case))**.** *Let $I \subset S$ be such that $V_X(I) = \{\zeta_1, \ldots, \zeta_\delta\} \subset U$ is zero-dimensional, where ζ_i has multiplicity μ_i. Let $(\alpha, \alpha_0) \in \mathrm{Pic}(X)^2$ be a regularity pair. For any $g \in S_{\alpha_0}$ and a generic $h_0 \in S_{\alpha_0}$, consider the linear map $M_g \circ M_{h_0}^{-1} : (S/I)_{\alpha+\alpha_0} \to (S/I)_{\alpha+\alpha_0}$. We have*

$$\det(\lambda \mathrm{id}_{\mathbb{C}^{\delta^+}} - M_g \circ M_{h_0}^{-1}) = \prod_{i=1}^{\delta} \left(\lambda - \frac{g}{h_0}(\zeta_i)\right)^{\mu_i}.$$

Proof. The map M_{h_0} is invertible by Corollary 5.5.3 below. Our strategy is to prove that there exist $\mathbb{C}$-linear maps L_{h_0} and L_g such that $L_{h_0} \circ \psi_{\alpha+\alpha_0} \circ M_g = L_g \circ \psi_{\alpha+\alpha_0} \circ M_{h_0}$ where L_{h_0} is invertible and

$$\det(\lambda \mathrm{id}_{\mathbb{C}^{\delta^+}} - L_{h_0}^{-1} \circ L_g) = \prod_{i=1}^{\delta} \left(\lambda - \frac{g}{h_0}(\zeta_i)\right)^{\mu_i}. \tag{5.5.17}$$

Recall that $v_{ij,\alpha+\alpha_0} = \mathrm{ev}_{\zeta_i} \circ \partial_{ij} \circ \eta_{\alpha+\alpha_0,\sigma_i}$ and hence

$$v_{ij,\alpha+\alpha_0}(gf + I_{\alpha+\alpha_0}) = (\mathrm{ev}_{\zeta_i} \circ \partial_{ij})(g^{\sigma_i} f^{\sigma_i} + \mathscr{I}(U_{\sigma_i})).$$

Viewing ∂_{ij} as a differential operator on $\mathbb{C}[U_{\sigma_i}]$, by Leibniz' rule we have

$$\partial_{ij}(h_0^{\sigma_i} g^{\sigma_i} f^{\sigma_i}) = \sum_{b \in \mathbb{N}^\ell} \partial_b(h_0^{\sigma_i}) s_b(\partial_{ij})(g^{\sigma_i} f^{\sigma_i}) = \sum_{b \in \mathbb{N}^\ell} \partial_b(g^{\sigma_i}) s_b(\partial_{ij})(h_0^{\sigma_i} f^{\sigma_i}).$$

Composing with ev_{ζ_i}, by consistent ordering of the ∂_{ij}, as in (3.1.8) we get

$$\underbrace{\begin{bmatrix} h_0^{\sigma_i}(\zeta_i) & & & \\ c_{i2}^{(1)} & h_0^{\sigma_i}(\zeta_i) & & \\ \vdots & & \ddots & \\ c_{i\mu_i}^{(1)} & c_{i\mu_i}^{(2)} & \dots & h_0^{\sigma_i}(\zeta_i) \end{bmatrix}}_{L_{i,h_0}} \begin{bmatrix} v_{i1,\alpha+\alpha_0} \\ v_{i2,\alpha+\alpha_0} \\ \vdots \\ v_{i\mu_i,\alpha+\alpha_0} \end{bmatrix} \circ M_g$$

$$= \underbrace{\begin{bmatrix} g^{\sigma_i}(\zeta_i) & & & \\ d_{i2}^{(1)} & g^{\sigma_i}(\zeta_i) & & \\ \vdots & & \ddots & \\ d_{i\mu_i}^{(1)} & d_{i\mu_i}^{(2)} & \dots & g^{\sigma_i}(\zeta_i) \end{bmatrix}}_{L_{i,g}} \begin{bmatrix} v_{i1,\alpha+\alpha_0} \\ v_{i2,\alpha+\alpha_0} \\ \vdots \\ v_{i\mu_i,\alpha+\alpha_0} \end{bmatrix} \circ M_{h_0}.$$

Putting all the equations together for $i = 1, \dots, \delta$, we get

$$\begin{bmatrix} L_{1,h_0} & & & \\ & L_{2,h_0} & & \\ & & \ddots & \\ & & & L_{\delta,h_0} \end{bmatrix} \circ \psi_{\alpha+\alpha_0} \circ M_g = \begin{bmatrix} L_{1,g} & & & \\ & L_{2,g} & & \\ & & \ddots & \\ & & & L_{\delta,g} \end{bmatrix} \circ \psi_{\alpha+\alpha_0} \circ M_{h_0}, \quad (5.5.18)$$

which is the desired relation $L_{h_0} \circ \psi_{\alpha+\alpha_0} \circ M_g = L_g \circ \psi_{\alpha+\alpha_0} \circ M_{h_0}$. Indeed, by construction, $h_0^{\sigma_i}(\zeta_i) \neq 0, \forall i$ and $\frac{g^{\sigma_i}}{h_0^{\sigma_i}}(\zeta_i) = \frac{g}{h_0}(\zeta_i)$, so L_{h_0} is invertible and (5.5.17) is satisfied. $\square$

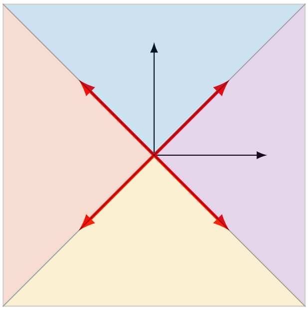
Remark 5.5.6. In the proof of Theorem 5.5.6 we represented $\psi_{\alpha+\alpha_0} : (S/I)_{\alpha+\alpha_0} \to \mathbb{C}^{\delta^+}$ as a vector of linear functionals $v_{ij,\alpha+\alpha_0}$. Fixing bases for $(S/I)_\alpha$ and $(S/I)_{\alpha+\alpha_0}$, (5.5.18) can be written as the matrix equation

$$L_{h_0} V M_g = L_g V M_{h_0},$$

where V represents $\psi_{\alpha+\alpha_0}$ in the chosen basis. We now relate Theorem 5.5.6 to Theorem 5.5.3. We have that

$$V(M_g M_{h_0}^{-1})V^{-1} = L_{h_0}^{-1} L_g \quad \text{and so} \quad V M_{h_0}(M_{h_0}^{-1} M_g) M_{h_0}^{-1} V^{-1} = L_{h_0}^{-1} L_g.$$

If $V_X(I)$ is reduced ($\delta = \delta^+$), then $L_{h_0}^{-1} L_g$ is a diagonal matrix containing the evaluations of the rational function $(g/h_0)(\zeta_i)$, which are the eigenvalues of $M_{h_0}^{-1} \circ M_g$



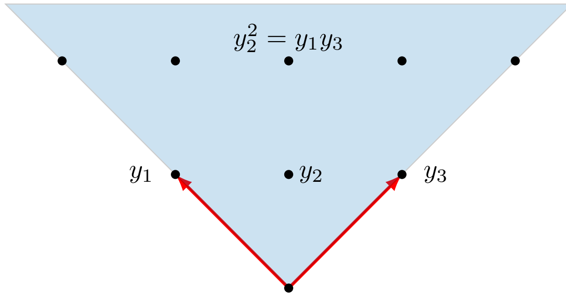


Figure 5.16: Illustration of the fan Σ (left) of the toric variety from Example 5.5.11, and of the semigroup algebra $\mathbb{C}[U_{\sigma_1}] \simeq \mathbb{C}[y_1, y_2, y_3]/\langle y_2^2 - y_1y_3\rangle$ corresponding to the (dual cone of the) blue cone (right).

corresponding to the left eigenvectors given by the functionals $(\mathrm{ev}_{\zeta_i} \circ \partial_0 \circ \eta_{\alpha+\alpha_0,\sigma_i}) \circ M_{h_0}$ (these are the rows of VM_{h_0} in the matrix representation). We observe that

$$(\mathrm{ev}_{\zeta_i} \circ \partial_0 \circ \eta^{-1}_{\alpha+\alpha_0,\sigma_i}) \circ M_{h_0}(f + I_\alpha) = \frac{h_0}{x^{\hat{\sigma}_i,\alpha_0}}(\zeta_i)\frac{f}{x^{\hat{\sigma}_i,\alpha}}(\zeta_i) = \frac{h_0 h}{x^{\hat{\sigma}_i,\alpha+\alpha_0}}(\zeta_i)\,\mathrm{ev}_{\zeta_i}(f + I_\alpha),$$

where h, ev_{ζ_i} are as in (the proof of) Theorem 5.5.3. △

Example 5.5.11. Let $n = 2$, $\mathbb{C}[M] = \mathbb{C}[t_1^{\pm 1}, t_2^{\pm 1}]$ and consider the equations

$$\hat{f}_1 = t_1 - t_2^{-1} + t_2 + t_1^{-1}, \quad \hat{f}_2 = 2t_1 + t_2^{-1} - t_2 - t_1^{-1}.$$

There are no solutions of $\hat{f}_1 = \hat{f}_2 = 0$ in $(\mathbb{C}^*)^2$ (note that $\hat{f}_1 + \hat{f}_2$ is a unit in $\mathbb{C}[M]$). The mixed volume of the Newton polygons P_1, P_2 is 4 and the associated toric variety X corresponds to the fan Σ depicted in Figure 5.16. We arrange the primitive ray generators of $\Sigma(1)$ in the matrix

$$F = [u_1\ u_2\ u_3\ u_4] = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \end{bmatrix}.$$

Our equations homogenize to

$$f_1 = x_1^2x_4^2 - x_3^2x_4^2 + x_1^2x_2^2 + x_2^2x_3^2, \quad f_2 = 2x_1^2x_4^2 + x_3^2x_4^2 - x_1^2x_2^2 - x_2^2x_3^2$$

in the Cox ring S of X. The degrees are $\alpha_1 = \alpha_2 = [\sum_{i=1}^4 D_i]$. Here $V_X(f_1, f_2)$ consists of two points, each with multiplicity two. These points correspond to the orbits of

$$z_1 = (0, 1, 1, 1), \quad z_2 = (1, 1, 0, \sqrt{-1}).$$

Let $\alpha = 2\alpha_1$ and $\alpha_0 = \alpha_1$. One can check that in the bases

$$\mathcal{B}_\alpha = \{x_3^4x_4^4 + I_\alpha, x_1x_2x_3^3x_4^3 + I_\alpha, x_1x_2^3x_3^3x_4 + I_\alpha, x_1^4x_2^4 + I_\alpha\}$$
$$\mathcal{B}_{\alpha+\alpha_0} = \{x_2^2x_3^6x_4^4 + I_{\alpha+\alpha_0}, x_1x_2^3x_3^5x_4^3 + I_{\alpha+\alpha_0}, x_1x_2^5x_3^5x_4 + I_{\alpha+\alpha_0}, x_1^4x_2^6x_3^2 + I_{\alpha+\alpha_0}\}$$

of $(S/I)_\alpha$ and $(S/I)_{\alpha+\alpha_0}$ respectively, multiplication with $x_2^2x_3^2,\ x_1x_2x_3x_4 \in S_{\alpha_0}$ looks like this:

$$M_{x_2^2x_3^2} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad M_{x_1x_2x_3x_4} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Let σ_1 be the blue cone in Figure 5.16. The ideal $\mathscr{I}(U_{\sigma_1}) = \langle f_1^{\sigma_1}, f_2^{\sigma_1}\rangle \subset \mathbb{C}[U_{\sigma_1}]$ corresponds to the ideal

$$I_{\sigma_1} = \langle y_1y_3 - y_2^2, y_2^2 - y_1 + y_3 + 1, 2y_2^2 + y_1 - y_3 - 1\rangle \subset \mathbb{C}[y_1, y_2, y_3] = R_{\sigma_1}.$$

The ordering of the variables y_i of R_{σ_1} is clarified in the right part of Figure 5.16. Only the solution ζ_1 corresponding to the orbit of z_1 is contained in U_{σ_1}, which explains that $\dim_{\mathbb{C}} \mathbb{C}[y_1, y_2, y_3]/I_{\sigma_1} = 2$. It has coordinates $(y_1, y_2, y_3) = (1, 0, 0)$, and a consistently ordered basis for D_1 is $\{\partial_{(0,0,0)}, \partial_{(0,1,0)}\}$. This gives

$$v_{11,\alpha+\alpha_0} = \mathrm{ev}_{\zeta_1} \circ \eta_{\alpha+\alpha_0,\sigma_1}, \quad v_{12,\alpha+\alpha_0} = \mathrm{ev}_{\zeta_1} \circ \frac{\partial}{\partial y_2} \circ \eta_{\alpha+\alpha_0,\sigma_1}.$$

Representing this in the basis $\mathcal{B}_{\alpha+\alpha_0}$ we get

$$\begin{bmatrix} v_{11,\alpha+\alpha_0} \\ v_{12,\alpha+\alpha_0} \end{bmatrix} = \begin{bmatrix} \mathrm{ev}_{\zeta_1} \circ \eta_{\alpha+\alpha_0,\sigma_1} \\ \mathrm{ev}_{\zeta_1} \circ \frac{\partial}{\partial y_2} \circ \eta_{\alpha+\alpha_0,\sigma_1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix},$$

which follows from $\eta_{\alpha+\alpha_0,\sigma_1}(\mathcal{B}_{\alpha+\alpha_0}) = \{y_1, y_1y_2, y_1^2y_2, y_1y_2^4\}$. For any $g \in S_{\alpha_0}$, the lower triangular matrix $L_{1,g}$ is given by

$$L_{1,g} = \begin{bmatrix} g^{\sigma_1}(\zeta_1) & 0 \\ \frac{\partial g^{\sigma_1}}{\partial y_2}(\zeta_1) & g^{\sigma_1}(\zeta_1) \end{bmatrix},$$

which gives, for $g = x_1x_2x_3x_4, h_0 = x_2^2x_3^2$,

$$L_{1,g} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \qquad L_{1,h_0} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

which follows from $g^{\sigma_1} \sim y_2$, $h_0^{\sigma_1} \sim y_1$. This gives for the rows of (5.5.18) corresponding to ζ_1:

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

In order to complete this equation with the rows corresponding to ζ_2, one has to work in the chart corresponding to either the purple or the yellow cone in Figure 5.16. △

Example 5.5.12 (27 lines on a cubic surface)**.** The author is grateful to Marta Panizzut and Sascha Timme for bringing this example to his attention. A classical result in intersection theory states that a general cubic surface in $\mathbb{P}^3$ given by

$$\begin{aligned} c_0w^3 + c_1w^2z + c_2wz^2 + c_3z^3 + c_4w^2y + c_5wyz + c_6yz^2 + c_7wy^2 \\ + c_8y^2z + c_9y^3 + c_{10}w^2x + c_{11}wxz + c_{12}xz^2 + c_{13}wxy + c_{14}xyz \\ + c_{15}xy^2 + c_{16}wx^2 + c_{17}x^2z + c_{18}x^2y + c_{19}x^3 = 0 \end{aligned}$$

contains 27 lines, see for instance [EH16, Subsection 6.2.1]. As detailed in [PSS19, Section 4], these lines correspond to the solutions of the polynomial system given by $\hat{f}_1 = \cdots = \hat{f}_4 = 0$ with

$$\begin{aligned} \hat{f}_1 &= c_0t^3 + c_1t^2v + c_2tv^2 + c_3v^3 + c_4t^2 + c_5tv + c_6v^2 + c_7t + c_8v + c_9, \\ \hat{f}_2 &= c_0s^3 + c_1s^2u + c_2su^2 + c_3u^3 + c_{10}s^2 + c_{11}su + c_{12}u^2 + c_{16}s + c_{17}u + c_{19}, \\ \hat{f}_3 &= 3c_0st^2 + 2c_1stv + c_2sv^2 + c_1t^2u + 2c_2tuv + 3c_3uv^2 + 2c_4st + c_5sv + c_{10}t^2 \\ &\quad + c_5tu + c_{11}tv + 2c_6uv + c_{12}v^2 + c_7s + c_{13}t + c_8u + c_{14}v + c_{15}, \\ \hat{f}_4 &= 3c_0s^2t + c_1s^2v + 2c_1stu + 2c_2suv + c_2tu^2 + 3c_3u^2v + c_4s^2 + 2c_{10}st + c_5su \\ &\quad + c_{11}sv + c_{11}tu + c_6u^2 + 2c_{12}uv + c_{13}s + c_{16}t + c_{14}u + c_{17}v + c_{18}. \end{aligned}$$

The mixed volume $\mathrm{MV}(P_1, P_2, P_3, P_4) = 45$ (with $P_i = \mathrm{Newt}(\hat{f}_i)$), yet we know that for generic parameter values $c_0, \ldots, c_{19}$, there are only 27 solutions in $(\mathbb{C}^*)^4$. The relations defined by $\hat{f}_1, \ldots, \hat{f}_4$ on $(\mathbb{C}^*)^4$ extend naturally to a toric compactification $X = X_\Sigma \supset (\mathbb{C}^*)^4$, where X is the toric variety coming from the fan Σ that we will now describe. We define

$$F = \begin{bmatrix} 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \end{bmatrix} = [u_1\ u_2\ u_3\ u_4\ u_5\ u_6] \quad \text{and} \quad a = \begin{bmatrix} 0 \\ 0 \\ 6 \\ 6 \\ 0 \\ 0 \end{bmatrix}$$

and the convex polytope $P = P_1 + \cdots + P_4 \subset \mathbb{R}^4$ is given by

$$P = \{m \in \mathbb{R}^4 \mid F^\top m + a \geq 0\}.$$

The fan Σ is the normal fan of P. It has 6 rays, whose primitive generators u_i are the columns of F. Theorem 5.4.2 states that the maximal number of isolated solutions of $f_1 = \cdots = f_4 = 0$ on X is 45. Solving a generic instance of our system using the algorithm, we find that there are in fact 45 isolated solutions on X (counting multiplicities), of which 18 are on the boundary $X \setminus (\mathbb{C}^*)^4$. Figure 5.17 shows the computed coordinates. The figure suggests clearly that there are indeed 27 solutions in the torus, and 18 solutions that are on the intersection of the 3rd and 4th torus invariant prime divisors, which we will denote by $D_3, D_4 \subset X$. These are the

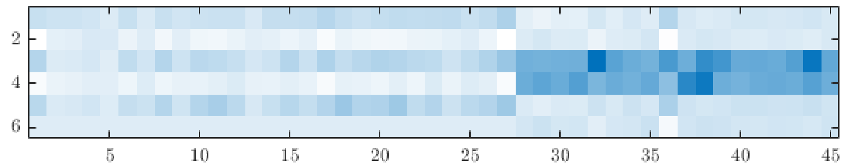


Figure 5.17: Absolute value of the computed homogeneous coordinates of 45 solutions. The i-th row corresponds to the i-th torus invariant prime divisor, associated to the ray generated by u_i, and the j-th column corresponds to the j-th computed solution. Dark colors correspond to small absolute values.

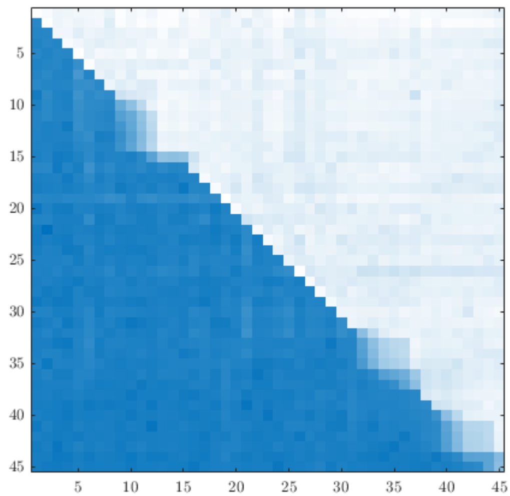


Figure 5.18: Absolute values of the entries of the block upper triangularized form of one of the homogeneous multiplication matrices $M_{x^{b_i}/h_0}$ in Example 5.5.12. Dark colors correspond to small absolute values.

divisors corresponding to u_3 and u_4. In fact, having a closer look at the intermediate computations, there should be only 3 solutions on $D_3 \cap D_4$, each with multiplicity 6. These multiplicities become apparent when the $\mathbf{U}'$ matrix in the ordered Schur factorization of a generic linear combination of the $M_{x^{b_i}/h_0}$ brings the matrix $M_{x^{b_1}/h_0}$ into *block upper triangular* instead of *upper triangular* form (see the discussion at the end of Subsection 4.3.2). One of the matrices $\mathbf{U}' M_{x^{b_i}/h_0} (\mathbf{U}')^H$ is shown in Figure 5.18. We now explicitly compute the three solutions on the boundary by solving the *face system*[7] corresponding to u_3 and u_4:

$$\begin{aligned}
(\hat{f}_1)_{u_3,u_4}(s,u,t,v) &= c_0t^3 + c_1t^2v + c_2tv^2 + c_3v^3,\\
(\hat{f}_2)_{u_3,u_4}(s,u,t,v) &= c_0s^3 + c_1s^2u + c_2su^2 + c_3u^3,\\
(\hat{f}_3)_{u_3,u_4}(s,u,t,v) &= 3c_0st^2 + 2c_1stv + c_2sv^2 + c_1t^2u + 2c_2tuv + 3c_3uv^2,\\
(\hat{f}_4)_{u_3,u_4}(s,u,t,v) &= 3c_0s^2t + c_1s^2v + 2c_1stu + 2c_2suv + c_2tu^2 + 3c_3u^2v.
\end{aligned}$$

One can see from these equations that $D_3 \cap D_4 \simeq \mathbb{P}^1 \times \mathbb{P}^1$, with coordinates $(s:u)$ and $(t:v)$ on the first and second copy of $\mathbb{P}^1$ respectively. The bidegrees of the equations are $(0,1),(1,0),(1,2),(2,1)$. We now interpret $(\hat{f}_1)_{u_3,u_4}$ as an equation on $\mathbb{P}^1$ and consider its three roots $(t_j^*:v_j^*)$, $j=1,2,3$ (for which we can write down explicit expressions) and we define $\zeta_j = ((t_j^*:v_j^*),(t_j^*:v_j^*)) \in \mathbb{P}^1\times\mathbb{P}^1$. It is clear that $(\hat{f}_1)_{u_3,u_4}(\zeta_j) = (\hat{f}_2)_{u_3,u_4}(\zeta_j) = 0$. If we substitute $s=t, u=v$ in $(\hat{f}_3)_{u_3,u_4}, (\hat{f}_4)_{u_3,u_4}$ we find that

$$(\hat{f}_3)_{u_3,u_4}(t,v,t,v) = (\hat{f}_4)_{u_3,u_4}(t,v,t,v) = 3(\hat{f}_1)_{u_3,u_4}(s,u,t,v).$$

From this it is clear that also $(\hat{f}_3)_{u_3,u_4}(\zeta_j) = (\hat{f}_4)_{u_3,u_4}(\zeta_j) = 0$, $j=1,\dots,3$, and we have identified the three solutions on $D_3\cap D_4$. △

The rest of this subsection is devoted to some results related to the regularity $\mathrm{Reg}(I)$. The first result is perhaps the most conclusive one. The strategy of proof is strongly related to that of Theorem 3 in [Mas16].

Theorem 5.5.7. *Let $I = \langle f_1,\dots,f_n\rangle \subset S$ with $f_i \in S_{\alpha_i}$ such that $\alpha_i \in \mathrm{Pic}(X)$ is basepoint free and $V_X(I)$ is zero-dimensional. For any basepoint free $\alpha_0 \in \mathrm{Pic}(X)$, the degree $\beta = \sum_{i=1}^n \alpha_i + \alpha_0$ belongs to the regularity $\mathrm{Reg}(I)$. In particular, $\sum_{i=1}^n \alpha_i \in \mathrm{Reg}(I)$.*

Proof. The proof requires some tools from homological algebra that were not introduced in this text. We present a sketch. Details can be found in [BT20a]. Let $\mathscr{O}_X$ be the structure sheaf of X and let $\mathscr{O}_{\mathcal{Z}}$ be the structure sheaf of $\mathcal{Z} = V_X(I)$ (this is the coherent sheaf associated to the S-module S/I, see [Cox95, §3]). Consider the Koszul complex of sheaves

$$\mathcal{K}(f_1,\dots,f_n): 0 \to \mathscr{K}_n \to \cdots \to \mathscr{K}_1 \to \mathscr{O}_X \quad \text{with} \quad \mathscr{K}_j = \bigoplus_{\substack{\mathscr{T}\subset\{1,\dots,n\}\\ |\mathscr{T}|=j}} \mathscr{O}_X(-\sum_{i\in\mathscr{T}}\alpha_i),$$

where $\mathscr{K}_1 = \bigoplus_{i=1}^n \mathscr{O}_X(-\alpha_i) \to \mathscr{O}_X$ is given locally on U_σ by $(g_1,\dots,g_n) \mapsto g_1f_1^\sigma + \cdots + g_nf_n^\sigma$. By Exercise 17.20 in [Eis13] and the fact that X is locally

[7] This is the system defined by the terms of the $\hat{f}_i$ with exponents m for which both $\langle u_3, m\rangle$ and $\langle u_4, m\rangle$ are minimized. This gives a system of equations in a 2-dimensional lattice which can be interpreted as the restriction of the original system to the dense torus of $D_3\cap D_4$. See for instance [HS95].

Cohen Macaulay [CLS11, Theorem 9.2.9], $\mathcal{K}(f_1,\dots,f_n)$ is a free resolution of $\mathscr{O}_{\mathcal{Z}}$, meaning that $\mathcal{K}(f_1,\dots,f_n) \to \mathscr{O}_{\mathcal{Z}} \to 0$ is an exact sequence of sheaves. Tensoring with $\mathscr{O}_X(\beta)$ preserves exactness ($\beta = \sum_{i=1}^n \alpha_i + \alpha_0$ is Cartier, so $\mathscr{O}_X(\beta)$ is invertible). Taking global sections then gives the sequence

$$0 \to S_{(\beta - \sum_{i=1}^n \alpha_i)} \to \cdots \to \bigoplus_{i=1}^n S_{(\beta-\alpha_i)} \to S_\beta \to H^0(X, \mathscr{O}_{\mathcal{Z}}) \to 0, \tag{5.5.19}$$

by [CLS11, Proposition 5.3.7] and the fact that $H^0(X, \mathscr{O}_{\mathcal{Z}} \otimes_{\mathscr{O}_X} \mathscr{O}_X(\beta)) = H^0(X, \mathscr{O}_{\mathcal{Z}})$ because $\mathcal{Z}$ is zero-dimensional. The exactness of this complex will follow from [GKZ94, Chapter 2, Lemma 2.4], which states that it is enough to show that the higher order sheaf cohomologies vanish for all terms in the finite sequence $\mathcal{K}(f_1,\dots,f_n) \to \mathscr{O}_{\mathcal{Z}} \to 0$. We have that $H^p(X, \mathscr{K}_j \otimes_{\mathscr{O}_X} \mathscr{O}_X(\beta)) = 0$ for $j = 1,\dots,n$ and $p > 0$, by Demazure vanishing [CLS11, Theorem 9.2.3]. The vanishing of $H^p(X, \mathscr{O}_{\mathcal{Z}} \otimes_{\mathscr{O}_X} \mathscr{O}_X(\beta)) = H^p(X, \mathscr{O}_{\mathcal{Z}})$ for $p > 0$ is proved by using Serre's criterion [Har77, Chapter III, Theorem 3.7] ($\mathcal{Z}$ is zero-dimensional so it's affine). Exactness of (5.5.19) implies that

$$H^0(X, \mathscr{O}_{\mathcal{Z}}) \simeq S_\beta / \operatorname{im}\left(\bigoplus_{i=1}^n S(\beta - \alpha_i) \to S_\beta \right) = (S/I)_\beta.$$

It follows that $\mathrm{HF}_I(\beta) = \delta^+$. The fact that $I_\beta = J_\beta$ follows from the last sequence in [CLS11, Theorem 9.5.7], which shows that J_β / I_β is the kernel of $(S/I)_\beta \to H^0(X, \mathscr{O}_{\mathcal{Z}}(\beta)) = H^0(X, \mathscr{O}_{\mathcal{Z}})$. □

Theorem 5.5.7 guarantees that the regularity for a square system is nonempty and it gives some degrees in $\mathrm{Pic}(X)$ which must be contained in it. However, we see in practice that the regularity is larger. For instance, it is often possible to choose $\alpha_0 \in \mathrm{Cl}(X)_+ \setminus \mathrm{Pic}(X)$ in Theorem 5.5.7 without leaving the regularity.

We now prove a result that has been used earlier in this chapter and previous chapters.

Lemma 5.5.7. *Let $I \subset S$ be such that $V_X(I)$ is zero-dimensional. For any $\alpha_0 \in \mathrm{Cl}(X)_+$ and $h_0 \in S_{\alpha_0}$ such that $V_X(h_0) \cap V_X(I) = \varnothing$, we have that the image of h_0 in S/J and in $S/\sqrt{J}$ is not a zero divisor, where $J = (I : \mathfrak{B}^\infty)$.*

Proof. Let $J = Q_1 \cap \cdots \cap Q_\ell$ be a minimal primary decomposition. Since J is $\mathfrak{B}$-saturated, $V_{\mathbb{C}^k}(Q_i) \not\subset V_{\mathbb{C}^k}(\mathfrak{B})$ for every $i = 1,\dots,\ell$. Indeed, if $V_{\mathbb{C}^k}(Q_i) \subset V_{\mathbb{C}^k}(\mathfrak{B})$ then $\mathfrak{B}^{\ell'} \subset Q_i$ for some $\ell' \in \mathbb{N}$. Take $f \in \bigcap_{j \neq i} Q_j$ such that $f \notin Q_i$. Then $f \notin J$, but $b^{\ell'} f \in J$ for all $b \in \mathfrak{B}$. This contradicts $J = (J : \mathfrak{B}^\infty)$.
Consider $h_0 \in S_{\alpha_0}$ such that the image of h_0 in S/J is a zero divisor. We can find $f \notin J$ such that $h_0 f \in J$. Therefore, there is a primary ideal Q_i in the decomposition of J such that $f \notin Q_i$. Since Q_i is primary, this implies $h_0^q \in Q_i$ for some q, and so $h_0 \in \sqrt{Q_i}$. As $V_{\mathbb{C}^k}(Q_i) \not\subset V_{\mathbb{C}^k}(\mathfrak{B})$, we conclude that $V_X(h_0) \cap V_X(I) \neq \varnothing$. This shows that if $V_X(h_0) \cap V_X(I) = \varnothing$, $h_0 + J$ is not a zero divisor in S/J. To show the statement for $\sqrt{J}$, we note that $h_0 f \in \sqrt{J}$ implies that $h_0^q f^q \in J$ for some q and hence $f^q \in J$, which implies $f \in \sqrt{J}$. □

Corollary 5.5.3. *Let $I \subset S$ be such that $V_X(I)$ is zero-dimensional. For a regularity pair $(\alpha, \alpha_0) \in \mathrm{Cl}(X)_+^2$ and an element $h_0 \in S_{\alpha_0}$ such that $V_X(h_0) \cap V_X(I) = \varnothing$, we have that $M_{h_0} : (S/I)_\alpha \to (S/I)_{\alpha+\alpha_0}$ is an isomorphism of $\mathbb{C}$-vector spaces.*

Proof. By assumption, $\mathrm{HF}_I(\alpha) = \mathrm{HF}_I(\alpha + \alpha_0)$, so it suffices to show that M_{h_0} is injective. This follows immediately from $\alpha, \alpha + \alpha_0 \in \mathrm{Reg}(I)$ and Lemma 5.5.7. □

Remark 5.5.7. It is a straightforward consequence of Lemma 5.5.7 that Corollary 5.5.3 also holds for regularity pairs with respect to Definition 5.5.4. △

We now state a possibly useful proposition which guarantees that once we have found $\alpha \in \mathrm{Reg}(I)$, in order to 'jump' to another degree in the regularity, all we need to check is the value of the Hilbert function.

Proposition 5.5.7. *Let $I \subset S$ be such that $V_X(I)$ is zero-dimensional. If $\alpha \in \mathrm{Reg}(I)$, $\alpha_0 \in \mathrm{Cl}(X)_+$ is such that no ζ_j is a basepoint of S_{α_0} and $\mathrm{HF}_I(\alpha + \alpha_0) = \delta^+$, then $\alpha + \alpha_0 \in \mathrm{Reg}(I)$.*

Proof. By Lemma 5.5.7, $M_{h_0} : (S/J)_\alpha \to (S/J)_{\alpha+\alpha_0}$ is injective for generic h_0. Therefore $\mathrm{HF}_J(\alpha + \alpha_0) \geq \mathrm{HF}_J(\alpha) = \mathrm{HF}_I(\alpha) = \delta^+$. Since $I \subset J$ we also have $\mathrm{HF}_J(\alpha + \alpha_0) \leq \mathrm{HF}_I(\alpha + \alpha_0) = \mathrm{HF}_I(\alpha) = \delta^+$. We conclude that $I_{\alpha+\alpha_0} = J_{\alpha+\alpha_0}$. □

We consider the question for which $\alpha \in \mathrm{Cl}(X)$ we have $\mathrm{HF}_I(\alpha) = \delta^+$ in the case where $V_X(I)$ is a complete intersection, i.e., where $I = \langle f_1, \ldots, f_n \rangle$ is generated by n elements. We prove some results that are implied by Theorem 5.5.7 but their proofs do not require the same advanced tools. A formula for the mixed volume that will be useful is (see [ŞS16, Theorem 3.16])

$$\mathrm{MV}(P_1, \ldots, P_n) = \sum_{\ell=0}^{n} (-1)^{n-\ell} \sum_{\substack{\mathscr{T} \subset \{1,\ldots,n\} \\ |\mathscr{T}|=\ell}} |(P_0 + P_{\mathscr{T}}) \cap M| , \tag{5.5.20}$$

for any lattice polytope $P_0 \subset \mathbb{R}^n$ corresponding to a torus invariant, basepoint free Cartier divisor D_{P_0} on X. Some of the proofs of the following statements make use of the Koszul complex and its properties, see Subsection A.2.5. The following theorem generalizes Theorem 3.16 in [ŞS16] in the case where Z is small enough. It is Theorem 4.2 in [Tel20].

Theorem 5.5.8. *Let $I = \langle f_1, \ldots, f_n \rangle \subset S$ be such that $V_X(I)$ is a zero-dimensional subscheme of $U \subset X$ of degree δ^+. Let $\alpha_i = \deg(f_i) \in \mathrm{Pic}(X)$ be the basepoint free degrees of the generators. If* $\operatorname{codim} Z \geq n$ *then for all basepoint free $\alpha_0 \in \mathrm{Pic}(X)_+$, $\mathrm{HF}_I(\alpha_0 + \alpha_1 + \ldots + \alpha_n) = \delta^+$.*

Proof. Consider the Koszul complex

$$0 \to S(-\sum_{i=1}^{n} \alpha_i) \to \bigoplus_{\substack{\mathscr{T} \subset \{1,\dots,n\} \\ |\mathscr{T}|=n-1}} S(-\alpha_{\mathscr{T}}) \to \cdots \to \bigoplus_{\substack{\mathscr{T} \subset \{1,\dots,n\} \\ |\mathscr{T}|=2}} S(-\alpha_{\mathscr{T}}) \to \bigoplus_{i=1}^{n} S(-\alpha_i) \to S$$

where $\alpha_{\mathscr{T}} = \sum_{i \in \mathscr{T}} \alpha_i$ and $S(-\alpha)$ is the Cox ring with twisted grading: $S(-\alpha)_\beta = S(\beta - \alpha)$. Since the orbit closures $\overline{G \cdot z_j}$ have dimension $k - n$ [SR17, Theorem 4.22] and by assumption $\dim(Z) \leq k - n$, the f_i form a regular sequence in S (S is Cohen-Macaulay). Hence the Koszul complex is exact. Restricting to the degree $\beta = \alpha_0 + \alpha_1 + \ldots + \alpha_n$ part we get

$$0 \to S_{\alpha_0} \to \bigoplus_{\substack{\mathscr{T} \subset \{1,\dots,n\} \\ |\mathscr{T}|=n-1}} S_{\beta-\alpha_{\mathscr{T}}} \to \cdots \to \bigoplus_{\substack{\mathscr{T} \subset \{1,\dots,n\} \\ |\mathscr{T}|=2}} S_{\beta-\alpha_{\mathscr{T}}} \to \bigoplus_{i=1}^{n} S_{\beta-\alpha_i} \to S_\beta.$$

Let P_0 be the polytope corresponding to the basepoint free degree $\alpha_0 \in \mathrm{Pic}(X)$, we have

$$\dim_{\mathbb{C}}(S_{\alpha_0+\alpha_{\mathscr{T}}}) = |(P_0 + P_{\mathscr{T}}) \cap M|$$

with $P_{\mathscr{T}} = \sum_{i \in \mathscr{T}} P_i$ for any subset $\mathscr{T} \subset \{0, \dots, n\}$. Counting dimensions we get

$$\dim_{\mathbb{C}}((S/I)_\beta) = \sum_{\ell=0}^{n} (-1)^{n-\ell} \sum_{\substack{\mathscr{T} \subset \{1,\dots,n\} \\ |\mathscr{T}|=\ell}} |(P_0 + P_{\mathscr{T}}) \cap M|,$$

and the right hand side is the formula (5.5.20) for the mixed volume $\delta^+ = \mathrm{MV}(P_1, \dots, P_n)$ (Theorem 5.4.2). $\square$

Note that the conditions of Theorem 5.5.8 are satisfied by all toric surfaces ($n = 2$). Here is an analogous result (Theorem 4.3 from [Tel20]) for the case where the system is 'unmixed' (in some sense) and the corresponding polytope is normal.

Theorem 5.5.9. *Let $I = \langle f_1, \dots, f_n \rangle \subset S$ such that $V_X(I)$ is a zero-dimensional subscheme of X of degree δ^+. Let $\alpha_i = \deg(f_i) \in \mathrm{Pic}(X)$ be the basepoint free degrees of the generators. If there is a basepoint free degree $\alpha_\star \in \mathrm{Pic}(X)$ corresponding to a normal polytope, such that $\alpha_i = t_i \alpha_\star$ for positive integers t_i, then $\mathrm{HF}_I(t\alpha_\star) = \delta^+$ for $t \geq \sum_{i=1}^{n} t_i$.*

Proof. The assumption on α_i implies that $P_i = t_i P_\star + m_i$ for a normal polytope $P_\star$, lattice points m_i and positive integers t_i. We can assume without loss of generality that $m_i = 0, i = 1, \dots, n$. We consider the embedding $X_{\mathscr{A}} \subset \mathbb{P}^{|\mathscr{A}|-1}$ of X where $\mathscr{A} = P_\star \cap M$. More precisely, $X_{\mathscr{A}}$ is the image of $\Phi_{\alpha_\star}$ [CLS11, Proposition 5.4.7]. Let $u_m, m \in \mathscr{A}$ be homogeneous coordinates on $\mathbb{P}^{n_{\alpha_\star}-1} = \mathbb{P}^{|\mathscr{A}|-1}$. The toric ideal of $X_{\mathscr{A}}$ is denoted $I_{\mathscr{A}} \subset \mathbb{C}[u_m, m \in \mathscr{A}]$ and the $\mathbb{Z}$-graded coordinate ring of $X_{\mathscr{A}}$ is $\mathbb{C}[X_{\mathscr{A}}] = \mathbb{C}[u_m, m \in \mathscr{A}]/I_{\mathscr{A}}$. By [CLS11, Theorem 5.4.8], we have $S_{\alpha_i} \simeq \mathbb{C}[X_{\mathscr{A}}]_{t_i}$ and $f_i \in S_{\alpha_i}$ corresponds to an element $h_i + I_{\mathscr{A}} \in \mathbb{C}[X_{\mathscr{A}}]_{t_i}$. We define the homogeneous ideal

$I' = \langle h_1 + I_{\mathscr{A}}, \dots, h_n + I_{\mathscr{A}} \rangle \subset \mathbb{C}[X_{\mathscr{A}}]$. By assumption, I' defines a 0-dimensional subscheme of $X_{\mathscr{A}}$, so $h_1 + I_{\mathscr{A}}, \dots, h_n + I_{\mathscr{A}}$ is a regular sequence in $\mathbb{C}[X_{\mathscr{A}}]$ (the ring $\mathbb{C}[X_{\mathscr{A}}]$ is arithmetically Cohen-Macaulay [CLS11, Exercise 9.2.8]). As a consequence (Theorem A.2.6), the corresponding Koszul complex

$$0 \to K_n \to K_{n-1} \to \cdots \to K_2 \to K_1 \to \mathbb{C}[X_{\mathscr{A}}] \quad \text{with} \quad K_t = \bigoplus_{\substack{\mathscr{T} \subset \{1,\dots,n\} \\ |\mathscr{T}|=t}} \mathbb{C}[X_{\mathscr{A}}](-\sum_{i \in \mathscr{T}} t_i)$$

is exact. Since $P_\star$ is a normal polytope, we have $\dim_{\mathbb{C}}(\mathbb{C}[X_{\mathscr{A}}]_t) = |tP_\star \cap M|$. Counting dimensions and using the same formula as before for $\delta^+ = \mathrm{MV}(P_1, \dots, P_n) = \mathrm{MV}(t_1 P_\star, \dots, t_n P_\star)$ we find that $\dim_{\mathbb{C}}((\mathbb{C}[X_{\mathscr{A}}]/I')_t) = \delta^+$ for $t \geq \sum_{i=1}^n t_i$. Combining this with $(\mathbb{C}[X_{\mathscr{A}}]/I')_t \simeq (S/I)_{t\alpha_\star}$ (see [CLS11, Theorem 5.4.8]) we get the desired result. $\square$

We note that in the case where X is a product of projective spaces, stronger bounds than those of Theorem 5.5.8 and Theorem 5.5.9 are known [BFT18].

Theorem 5.5.8 exploits the fact that when the base locus is small, an ideal $\langle f_1, \dots, f_n \rangle$ behaves like a complete intersection in $\mathbb{C}^k$. Here's another series of results that makes use of this to prove a conjecture in [Tel20] in some special cases. Recall that $U \subset X$ is the largest simplicial open subset of X (see Remark 5.5.1).

Theorem 5.5.10. *Let X be such that the base locus $Z \subset \mathbb{C}^k$ satisfies $\operatorname{codim}_{\mathbb{C}^k} Z > n$. If $I = \langle f_1, \dots, f_n \rangle \subset S$ is a homogeneous ideal such that $V_X(I) \subset U$ is zero-dimensional, then $I = (I : \mathfrak{B}^\infty)$.*

Proof. By assumption, $V_{\mathbb{C}^k}(I) \setminus Z$ is a finite union of fibers of $\pi_{|\pi^{-1}(U)}$, where $\pi : \mathbb{C}^k \setminus Z \to X$ is the quotient map from the Cox construction. The closure of each fiber in $\mathbb{C}^k$ has dimension $k - n$, and by the assumption $\operatorname{codim} V_{\mathbb{C}^k}(\mathfrak{B}) > n$, we conclude $\operatorname{codim}_{\mathbb{C}^k} V_{\mathbb{C}^k}(I) = n$. Consider a primary decomposition

$$I = Q_1 \cap \cdots \cap Q_s.$$

Suppose $f \in (I : \mathfrak{B}^\infty) \setminus I$. This implies, in particular, that $f \notin Q_i$ for some i. Since $f \in (I : \mathfrak{B}^\infty)$, for any $b \in \mathfrak{B}$ we have that $b^\ell f \in Q_i$ for some $\ell \in \mathbb{N}$. Because Q_i is primary and $f \notin Q_i$, we find that $\mathfrak{B} \subset \sqrt{Q_i}$. However, by the unmixedness theorem [Eis13, Corollary 18.14] and the fact that S is Cohen-Macaulay, the associated prime $\sqrt{Q_i}$ has codimension n. Hence, we arrive at a contradiction and conclude that $I = (I : \mathfrak{B}^\infty)$. $\square$

Theorem 5.5.10 implies that one of the conditions for being in the regularity is satisfied for *all* degrees $\alpha \in \mathrm{Cl}(X)$ in the special case where the base locus $Z = V_{\mathbb{C}^k}(\mathfrak{B})$ is very small.

Corollary 5.5.4. *Let X be such that the base locus $Z \subset \mathbb{C}^k$ satisfies $\operatorname{codim}_{\mathbb{C}^k} Z > n$. Let $I = \langle f_1, \dots, f_n \rangle \subset S$ be a homogeneous ideal such that $\deg(f_i) = \alpha_i \in \operatorname{Pic}(X)$ is* *basepoint free and $V_X(I) \subset U$ is zero-dimensional, then $\alpha_0 + \alpha_1 + \cdots + \alpha_n \in \operatorname{Reg}(I)$ for any $\alpha_0 \in \operatorname{Cl}(X)_+$ such that $\ell\alpha_0 \in \operatorname{Pic}(X)$ is basepoint free for some $\ell \in \mathbb{N}$ and $V_X(h_0) \cap V_X(I) = \varnothing$ for some $h_0 \in S_{\alpha_0}$.*

Proof. Let $\alpha = \alpha_1 + \cdots + \alpha_n$. By Theorem 5.5.10, $I = J = (I : \mathfrak{B}^\infty)$ and we only need to show that $\operatorname{HF}_{S/I}(\alpha + \alpha_0) = \delta^+$, where δ^+ is the degree of $V_X(I)$. Let $\alpha_0 \in \operatorname{Cl}(X)_+$ be such that $V_X(h_0) \cap V_X(I) = \varnothing$ for some $h_0 \in S_{\alpha_0}$. By Lemma 5.5.7, h_0 is not a zero-divisor in $S/I = S/J$. By Theorem 5.5.7, we know that $\operatorname{HF}_{S/I}(\alpha) = \delta^+$. Since $M_{h_0} : (S/I)_\alpha \to (S/I)_{\alpha+\alpha_0}$ is injective, we see that $\operatorname{HF}_{S/I}(\alpha) \leq \operatorname{HF}_{S/I}(\alpha + \alpha_0)$. By assumption, there is $\ell \in \mathbb{N}$ such that $\ell\alpha_0 \in \operatorname{Pic}(X)$ and $\ell\alpha_0$ is basepoint free, so by Theorem 5.5.7 we find $\operatorname{HF}_{S/I}(\alpha + \ell\alpha_0) = \delta^+$. Using the same reasoning as before for the map $M_{h_0^{\ell-1}} : (S/I)_{\alpha+\alpha_0} \to (S/I)_{\alpha+\ell\alpha_0}$ we get

$$\delta^+ = \operatorname{HF}_{S/I}(\alpha) \leq \operatorname{HF}_{S/I}(\alpha + \alpha_0) \leq \operatorname{HF}_{S/I}(\alpha + \ell\alpha_0) = \delta^+. \qquad \square$$

Note that if X is simplicial, then each Weil divisor is $\mathbb{Q}$-Cartier [CLS11, Proposition 4.2.7], hence for every $\alpha_0 \in \operatorname{Cl}(X)$, there is $\ell \in \mathbb{N}$ such that $\ell\alpha_0 \in \operatorname{Pic}(X)$. By [BC94, Proposition 2.8], the only toric varieties satisfying the conditions of Corollary 5.5.4 are the so-called *fake weighted projective spaces*. These are the simplicial toric varieties corresponding to simplices. We can use Corollary 5.5.4 to prove a conjecture proposed in [Tel20] for this special class of toric varieties. We prove a helpful lemma first. Let $\operatorname{Pic}(X)_+ = \operatorname{Pic}(X) \cap \operatorname{Cl}(X)_+$.

Lemma 5.5.8. *If X is a toric variety associated to a lattice simplex in $\mathbb{R}^n$, we have that every element $\alpha \in \operatorname{Pic}(X)_+$ is basepoint free.*

Proof. Let Σ be the fan of X. Since Σ is the normal fan of a simplex, the Cox ring has $n+1$ variables and the base locus is $Z = \{0\}$. Therefore, it suffices to show that for any element $\alpha \in \operatorname{Pic}(X)_+$, there are $\ell_1, \ldots, \ell_{n+1} \in \mathbb{N}$ such that $x_j^{\ell_j} \in S_\alpha$. Let $\alpha = [\sum_{i=1}^{n+1} a_i D_i]$ with $a_i \in \mathbb{N}$. Since $\alpha \in \operatorname{Pic}(X)$ and any collection of n rays in $\Sigma(1)$ corresponds to an n-dimensional cone in $\Sigma(n)$, for $j = 1, \ldots, n+1$ there is $m_j \in M$ such that $\langle u_i, m_j \rangle + a_i = 0$, for all $i \neq j$. Hence

$$\langle u_i, -m_j \rangle \geq 0, i \neq j, \quad \text{which means} \quad m_j \in -\sigma_j^\vee,$$

where σ_j is the cone of Σ whose rays are generated by $u_i, i \neq j$. Since Σ is a complete fan and all its cones are pointed, we must also have $u_j \in -\sigma_j$, which implies $\langle u_j, m_j \rangle \geq 0$ and thus $\langle u_j, m_j \rangle + a_j \geq 0$. It follows that $x_j^{\ell_j} \in S_\alpha$ with $\ell_j = \langle u_j, m_j \rangle + a_j$. $\square$

The following is a direct consequence.

Corollary 5.5.5. *Let X be a toric variety associated to a lattice simplex in $\mathbb{R}^n$ (i.e. X is a* fake weighted projective space*) and let $I = \langle f_1, \ldots, f_n \rangle \subset S$ be a homogeneous ideal such that $V_X(I)$ is zero-dimensional and $\deg(f_i) = \alpha_i \in \operatorname{Pic}(X)$ is basepoint free. Then $\alpha_0 + \alpha_1 + \cdots + \alpha_n \in \operatorname{Reg}(I)$ for all $\alpha_0 \in \operatorname{Cl}(X)_+$ such that no ζ_j is a basepoint of S_{α_0}.*

Proof. A toric variety coming from a simplex is simplicial. Its fan has $k = n+1$ rays and the base locus satisfies $Z = \{0\}$. Moreover, any element of $\mathrm{Pic}(X)_+$ is basepoint free (Lemma 5.5.8), so for any $\alpha_0 \in \mathrm{Cl}(X)_+$ and any ℓ such that $\ell\alpha_0 \in \mathrm{Pic}(X)_+$, $\ell\alpha_0$ is basepoint free. Now apply Corollary 5.5.4. $\square$

Remark 5.5.8. The assumptions of Corollary 5.5.5 are satisfied for all weighted projective spaces. $\triangle$

Corollary 5.5.5 is Conjecture 1 in [Tel20] with the extra assumption that X is a fake weighted projective space. The conjecture is false in general. A counter example is given in [BT20a]. However, Theorem 5.5.7 shows that the conjecture holds for any toric variety with the extra assumption that $\alpha_0 \in \mathrm{Pic}(X)$ is basepoint free.

Chapter 6

Homotopy continuation

In this chapter we switch gears and consider a completely different approach to the problem of solving a system of polynomial equations. The presented material is mostly taken from [TVBV19]. Homotopy continuation is an important tool in numerical algebraic geometry. It is used for, among others, isolated polynomial root finding and for the numerical decomposition of algebraic varieties into irreducible components. We revisit the fundamental task of a polynomial homotopy algorithm, which is the numerical tracking of a smooth path in a homotopy, and propose a new algorithm for doing this in a robust way. For introductory texts on numerical algebraic geometry and homotopy continuation, we refer to [AG12, Li97, SVW01, SVW05, WS05] and references therein.

Let Y be an affine variety of dimension n with coordinate ring $R = \mathbb{C}[Y]$ and let $h_i, i = 1, \ldots, n$ be elements of $R[t] = \mathbb{C}[Y \times \mathbb{C}] = \mathbb{C}[Y] \otimes_{\mathbb{C}} \mathbb{C}[t]$. The h_i define the map

$$H : Y \times \mathbb{C} \to \mathbb{C}^n$$

given by $H(x,t) = (h_i(x,t))_{i=1}^n$. Such a map H should be thought of as a family of morphisms $Y \to \mathbb{C}^n$ parametrized by t, which defines a *homotopy* with continuation parameter t. This gives the *solution variety*

$$Z = H^{-1}(0) = \{(x,t) \in Y \times \mathbb{C} \mid h_i(x,t) = 0, i = 1, \ldots, n\} \subset Y \times \mathbb{C}.$$

We will limit ourselves to the cases $Y = \mathbb{C}^n, R = \mathbb{C}[x_1, \ldots, x_n]$ and $Y = (\mathbb{C}^*)^n, R = \mathbb{C}[x_1^{\pm 1}, \ldots, x_n^{\pm 1}]$. In both cases, we will use the coordinates $x = (x_1, \ldots, x_n)$ on Y. Note that for every fixed parameter value $t^* \in \mathbb{C}$, $H_{t^*} : Y \to \mathbb{C}^n : x \mapsto H(x, t^*)$ represents a system of n (Laurent) polynomial equations in n variables with solutions $H_{t^*}^{-1}(0) \subset Y$. Typically, for some parameter value $t_0 \in \mathbb{C}$, H_{t_0} is a *start system* with known, isolated and regular (i.e. multiplicity 1) solutions and for some other $t_1 \neq t_0$, H_{t_1} represents a *target system* in which we are interested. Suppose we know a point $(z_0, t_0) \in Z$. The task of a homotopy continuation algorithm is to track the point

$(z_0, t_0) \in Z$ to a point $(z_1, t_1) \in Z$ along a smooth continuous path

$$\{(x(s), \Gamma(s)), s \in [0,1]\} \subset Z$$

with $\Gamma : [0,1] \to \mathbb{C}$ and $x(s) \in Y, s \in [0,1]$ such that $\Gamma(0) = t_0, x(0) = z_0, \Gamma(1) = t_1, x(1) = z_1$. This is assuming that such a path exists. In practice there may be singular points on the path (e.g. *path crossing*), which may cause trouble for numerical path tracking. We will see an example in Section 6.1. In the cases we are interested in, issues may arise when the parameter s approaches 1. That is, there is a continuous path

$$\{(x(s), \Gamma(s)), s \in [0,1)\} \subset Z$$

which 'escapes' from $Y \times \mathbb{C}$ when $s \to 1$. For example, solutions may move to infinity or out of the algebraic torus. Many tools have been developed for dealing with such situations [HV98, MSW90, MSW92b, PV10]. In this text, we do not focus on this kind of difficulties. Existing techniques can be incorporated in the algorithms we present. We will work with $s \in [0,1)$ in some of our definitions to take these issues into account. We will mainly restrict ourselves to paths of the form $\{(x(t), t), t \in [0,1)\}$ (i.e. $\Gamma(s) = s$), but other Γ will be useful for constructing illustrative examples.

In typical constructions, such as linear homotopies for polynomial system solving, H is randomized such that the paths that need to be tracked do not contain singular points with probability one for $s \in [0,1)$ [WS05, Lemma 7.1.2]. This implies for example that all paths are disjoint. However, there might be singularities very near the path in the parameter space. In this situation, the coordinates in Y along the path may become very large, which causes scaling problems[1], or two different paths may be very near to each other for some parameter values. The latter phenomenon causes *path jumping*, which is considered one of the main problems for numerical path trackers. Path jumping occurs when along the way, the solution that is being tracked 'jumps' from one path to another. The typical reason is that starting from a point in $H_{t^*}^{-1}(0)$, the *predictor* step in the path tracking algorithm returns a point in $Y \times \{t^* + \Delta t\}$ which, according to the *corrector* step, is a numerical approximation of a point in $H_{t^*+\Delta t}^{-1}(0)$ which is on a different path than the one being tracked. We will say more about predictors and correctors in Section 6.1. It is clear that path jumping is more likely to occur in the case where two or more paths come near each other. Ideally, a numerical path tracker should take small steps Δt in such 'difficult' regions and larger steps where there's no risk for path jumping. There have been many efforts to design such *adaptive stepsize* path trackers [GS04, KX94, SC87]. However, the state of the art homotopy software packages such as PHCpack [Ver99], Bertini [BSHW13] and HomotopyContinuation.jl [BT18] still suffer from path jumping, as we will show in our experiments. We should mention that the algorithm presented in [Tim20] will soon be implemented in HomotopyContinuation.jl and shows some very promising results in terms of both robustness and computation time. A typical way to adjust the stepsize

[1] Scaling problems caused by large coordinates can be resolved by using homogeneous coordinates, after a projective transformation [Mor09]. These issues are addressed in a different way in [Tim20, Subsection 2.2].

is by an *a posteriori* step control. This is represented schematically (in a simplified way) by Figure 6.1. In the figure, $0 < \beta < 1$ is a real constant, the $\| \cdot \|$ should be

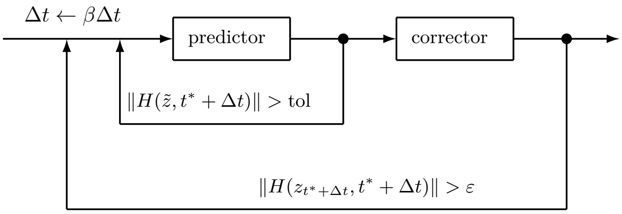


Figure 6.1: Two feedback loops in a predictor-corrector method for a posteriori step control.

interpreted as a relative measure of the backward error and $\tilde{z}$ is the predicted solution which is refined to $z_{t^*+\Delta t}$ by the corrector. If $\mathrm{tol} \leq \varepsilon$, then the corrector stage is not needed. If $\mathrm{tol} = \infty$, then the first feedback loop never happens. Such extreme choices for tol are not recommended. With well chosen values for tol and ε, the second feedback loop never occurs, as Newton's method converges to the required accuracy of ε in just a couple of steps. This type of feedback loops is implemented in, e.g., PHCpack [Ver99] and Bertini [BHSW08].

Certified path trackers have been developed to prevent path jumping [BL13, BC13, vdH15, XBY18], but they require more computational effort. Moreover, the certification assumes that the coefficients of the input systems are exact rational numbers, as stated in [BL13].

In this thesis, we propose an adaptive stepsize path tracking algorithm that is robust yet efficient. As opposed to standard methods, we use *a priori* step control: we compute the appropriate stepsize *before* taking the step. We use Padé approximants [BJGM96] of the solution curve $x(t)$ in the predictor step, not only to generate a next approximate solution, but also to detect nearby singularities in the parameter space. In the case of type $(L, 1)$ approximants (see Section 6.2 for a definition), this is a direct application of Fabry's ratio theorem (Theorem 6.2.2). The Padé approximants are computed from the series expansion of $x(t)$. We use the iterative, symbolic-numeric algorithm from [BV18a] to compute this series expansion. Let $\mathbb{C}[[t]]$ be the ring of formal power series in the variable t with coefficients in $\mathbb{C}$. For an appropriate starting value $x^{(0)}(t) \in \mathbb{C}[[t]]$, we prove 'second order convergence' of this iteration in the sense that $x(t) - x^{(k)}(t) = 0 \mod \langle t^{2k} \rangle$ where $x^{(k)}(t) \in \mathbb{C}[[t]]$ is the approximate series solution after the k-th iteration and $\langle \cdot \rangle$ denotes the ideal generated in the power series ring $\mathbb{C}[[t]]$ (see Proposition 6.3.1). We use information contained in the Padé approximant to determine a trust region for the predictor and use this as a first criterion to compute the adaptive stepsize. A second criterion is based on an estimate for the distance to the most nearby path and a standard approximation error estimate for the Padé approximant.

We note that Padé approximants have been used before in path tracking algorithms [GS04, SC87]. In these articles, their use has been limited to type $(2, 1)$ Padé approximants (see later for a definition) and they have not been used as nearby singularity detectors.

The chapter is organized as follows. In Section 6.1, we describe numerical path tracking algorithms for smooth paths in general and give some examples. In Section 6.2 we discuss fractional power series solutions and Padé approximants. Section 6.3 discusses the algorithmic aspects of computing power series solutions. Our path tracking algorithm is described in Section 6.4 and implemented in version 2.4.72 of PHCpack, which is available on github. We show the algorithm's effectiveness through several numerical experiments in Section 6.5. We compare with the built-in path tracking routines in (previous versions of) PHCpack [Ver99], Bertini [BSHW13] and HomotopyContinuation.jl [BT18].

6.1 Tracking smooth paths

Let $H(x, t) : Y \times \mathbb{C} \to \mathbb{C}^n$ be as above where Y is either $\mathbb{C}^n$ or $(\mathbb{C}^*)^n$. We denote $Z = H^{-1}(0)$ and we assume that $\dim Z = 1$. To avoid ambiguities, we will denote t for the coordinate on $\mathbb{C}$ in $Y \times \mathbb{C}$ and $t^* \in \mathbb{C}$ for points in $\mathbb{C}$. We define the projection map $\Pi : Z \to \mathbb{C} : (x, t) \mapsto t$. By [WS05, Theorem 7.1.1] Π is a branched covering of $\mathbb{C}$ with ramification locus $\mathcal{S}$ consisting of a finite set of points in $\mathbb{C}$, such that the fiber $\Pi^{-1}(t^*)$ consists of a fixed number $\deg \Pi = \delta \in \mathbb{N}$ of points in Z if and only if $t^* \in \mathbb{C} \setminus \mathcal{S}$. Let

$$J_H(x, t) = \left(\frac{\partial h_i}{\partial x_j} \right)_{1 \le i, j \le n}$$

be the Jacobian matrix of H with respect to the x_j.

Definition 6.1.1. Let H, Z be defined as above. Let $\Gamma : [0, 1] \to \mathbb{C}$ and let $\hat{\Gamma} = \{(x(s), \Gamma(s)), s \in [0, 1)\} \subset Z$ be a continuous path in Z. We say that $\hat{\Gamma}$ is smooth if $J_H(x, t)$ is invertible for all $(x, t) \in \hat{\Gamma}$.

If $\hat{\Gamma} = \{(x(s), \Gamma(s)) \mid s \in [0, 1)\} \subset Z$ is continuous with $\Gamma([0, 1)) \cap \mathcal{S} = \varnothing$, then $\hat{\Gamma} \subset \Pi^{-1}(\mathbb{C} \setminus \mathcal{S})$ is smooth. In this case, Γ is called a *smooth parameter path*. In more down to earth terms, $\Gamma = \Pi(\hat{\Gamma})$ is smooth if $\{\Gamma(s), s \in [0, 1)\} \subset \mathbb{C}$ contains only parameter values t^* for which H_{t^*} represents a (Laurent) polynomial system with the expected number of regular solutions.

Example 6.1.1. Consider the homotopy taken from [KX94] defined by

$$H(x, t) = x^2 - (t - 1/2)^2 - p^2 \tag{6.1.1}$$

where $p \in \mathbb{R}$ is a parameter which we take to be 0.1 in this example. It is clear that a generic fiber $\Pi^{-1}(t^*)$ consists of the two points

$$\pm\sqrt{(t^* - 1/2)^2 + p^2}$$

and the ramification locus is $\mathcal{S} = \{1/2 \pm p\sqrt{-1}\}$. Note that $J_H = \frac{\partial H}{\partial x}$ is equal to zero at $\Pi^{-1}(t^*)$ for $t^* \in \mathcal{S}$. We consider three different parameter paths:

$$\Gamma_1 : s \mapsto s, \quad \Gamma_2 : s \mapsto s - 4ps(s-1)\sqrt{-1}, \quad \Gamma_3 : s \mapsto s + 0.2\sin(\pi s)\sqrt{-1}.$$

In Figure 6.2 these paths are drawn in the complex plane. The background colour at $t^* \in \mathbb{C}$ in this figure corresponds to the absolute value of J_H evaluated at a point in $\Pi^{-1}(t^*)$: dark (blue) regions correspond to a small value. For each Γ_i, we

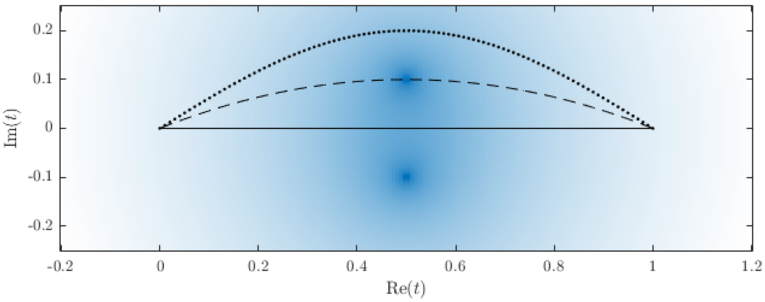


Figure 6.2: The image of $[0, 1]$ under Γ_1 (full line), Γ_2 (dashed line) and Γ_3 (dotted line) as defined in Example 6.1.1.

track two different paths in Z for $s \in [0, 1]$ starting at $(z_0^{(1)}, 0) = (\sqrt{1/4 + p^2}, 0)$ and $(z_0^{(2)}, 0) = (-\sqrt{1/4 + p^2}, 0)$ respectively. The result is shown in Figure 6.3. Denote the

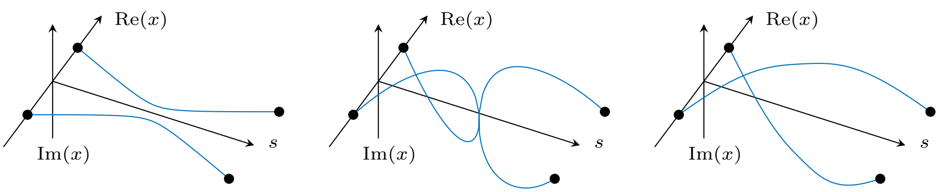


Figure 6.3: Solution curves with respect to s using, from left to right, Γ_1, Γ_2 and Γ_3.

corresponding paths on Z by $\hat{\Gamma}_j^{(i)} = \{(x^{(i)}(s), \Gamma_j(s)), s \in [0, 1]\}$ where $x^{(i)}(0) = z_0^{(i)}$. Since Γ_1 and Γ_3 do not hit any singular points in the parameter space (Figure 6.2), the corresponding paths $\hat{\Gamma}_j^{(i)}$ are disjoint and smooth. The paths corresponding to Γ_2, on the other hand, cross a singularity. They intersect at $s = 1/2$, as can be seen from Figure 6.3. We conclude that Γ_2 is not smooth. $\triangle$

An important application of smooth path tracking is the solution of systems of polynomial equations. The typical setup is the following. Define

$$F : Y \to \mathbb{C}^n : x \mapsto (f_1(x), \ldots, f_n(x))$$

with $f_i \in R$. We want to compute $F^{-1}(0)$, that is, all points $x \in Y$ such that $f_i(x) = 0, i = 1, \ldots, n$. The homotopy approach to this problem is to construct $H : Y \times \mathbb{C} \to \mathbb{C}^n$ such that $H_1 : x \mapsto H(x, 1)$ satisfies $Z_1 = H_1^{-1}(0) = F^{-1}(0)$ (the *target system* is equivalent to F) and the *start system* $G = H_0 : x \mapsto H(x, 0)$ is such that $Z_0 = G^{-1}(0)$ is easy to compute and contains the expected number δ of regular solutions. Moreover, H has the additional property that $\Gamma : [0, 1) \to \mathbb{C} : s \mapsto s$ is a smooth parameter path.

Example 6.1.2 (Straight line homotopies)**.** A typical construction that meets these criteria is given by a *straight line homotopy* between G and F, i.e.

$$H(x, t) = (1 - t)G(x) + \gamma t F(x),$$

where γ is a random nonzero complex constant, used to guarantee (with probability 1) that $\Gamma : s \mapsto s$ is smooth. This is called the γ-*trick*, see for instance [WS05, Page 18]. △

The number δ is equal to, for example, the Bézout number in the case of total degree homotopies, or the mixed volume of the Newton polytopes in the case of polyhedral homotopies [WS05, HS95, VVC94]. We denote

$$Z_0 = G^{-1}(0) = \{z_0^{(1)}, \ldots, z_0^{(\delta)}\}$$

and by smoothness of Γ, we have that

$$Z_{t^*} = H_{t^*}^{-1}(0) = \{z_{t^*}^{(1)}, \ldots, z_{t^*}^{(\delta)}\}$$

consists of δ distinct points in Y for $t^* \in [0, 1)$ and the paths $\{(z_{t^*}^{(i)}, t^*), t^* \in [0, 1)\}$ are smooth and disjoint. Depending on the given system F, Z_1 may consist of fewer than δ points, or it might even consist of infinitely many points. Two or more paths may approach the same point as $t^* \to 1$ or paths may diverge to infinity. As stated in the introduction to this chapter, several *end games* have been developed to deal with this kind of situations [HV98, MSW90, MSW92b, PV10]. We will focus here on the path tracking before the paths enter the end game operating region. We assume, for simplicity that this region is $[t_{\mathrm{EG}}, 1]$, for t_{EG} a parameter value 'near' 1. Algorithm 6.7 is a simple template algorithm for smooth path tracking. With a slight abuse of notation, we use $z_{t^*}^{(i)}$ both for actual points on the path and 'satisfactory' numerical approximations of the $z_{t^*}^{(i)}$.

The algorithm uses several auxiliary procedures. The *predictor* (line 6) computes a point $\tilde{z} \in Y$ and a stepsize Δt such that $\tilde{z}$ is an approximation for $z_{t^*+\Delta t}^{(i)}$. Some

Algorithm 6.7 Template path tracking algorithm with a priori step control

1: **procedure** $\textsc{Track}(H, Z_0)$
2: $\quad Z_1 \leftarrow \varnothing$
3: $\quad$ **for** $z_0^{(i)} \in Z_0$ **do**
4: $\quad\quad t^* \leftarrow 0$
5: $\quad\quad$ **while** $t^* < t_{\mathrm{EG}}$ **do**
6: $\quad\quad\quad (\tilde{z}, \Delta t) \leftarrow \textsc{Predict}(H, z_{t^*}^{(i)}, t^*)$
7: $\quad\quad\quad z_{t^*+\Delta t}^{(i)} \leftarrow \textsc{Correct}(H, \tilde{z}, t^* + \Delta t)$
8: $\quad\quad\quad t^* \leftarrow t^* + \Delta t$
9: $\quad\quad$ **end while**
10: $\quad\quad z_1^{(i)} \leftarrow \mathrm{endgame}(H, z_{t^*}^{(i)}, t^*)$
11: $\quad\quad Z_1 \leftarrow Z_1 \cup \{z_1^{(i)}\}$
12: $\quad$ **end for**
13: $\quad$ **return** Z_1
14: **end procedure**

existing predictors use an Euler step (tangent predictor) or higher order integrating techniques such as RK4.[2] Intuitively, the computed stepsize Δt should be small in 'difficult' regions. Algorithms that take this into account are called *adaptive stepsize* algorithms. Our main contribution is the adaptive stepsize predictor algorithm which we present in detail in Section 6.4. Our predictor computes an appropriate stepsize *before* the step is taken (a priori step control). The *corrector* step (line 7) then refines $\tilde{z}$ to a satisfactory numerical approximation of $z_{t^*+\Delta t}^{(i)}$. Typically, satisfactory means that the residual (see Appendix C) of $z_{t^*+\Delta t}^{(i)}$ is of size $\pm$ the unit roundoff. The endgame procedure in line 10 finishes the path tracking by performing an appropriate end game.

6.2 Puiseux series and Padé approximants

In this section we introduce some aspects of Puiseux series solutions and Padé approximants that are relevant for this text. References are provided for the reader who is interested in a more detailed treatment. In a first subsection we introduce Puiseux series. This will give us insight in the local behavior of the fibers of $\Pi : Z \to \mathbb{C}$ near the branch locus $\mathcal{S}$. In the second subsection, we discuss Padé approximants with an emphasis on how they behave in the presence of these kinds of singularities. Since we assume smoothness of the path, as described in the previous section, we will not construct series approximations at singularities in our algorithm. The Padé approximant at a regular point is influenced by nearby singular points, and it can be used to estimate their location.

[2] Some higher order predictors need several previous points on the path in order to compute $\tilde{z}$. The predictor we present in this algorithm uses only the last computed point, hence the notation in Algorithm 6.7.

Let $\mathbb{C}[[t]]$ be the ring of formal power series in the variable t and let $\mathfrak{m} = \langle t \rangle$ be its maximal ideal. We denote $\mathbb{C}[t]_{\leq d} \simeq \mathbb{C}[[t]]/\mathfrak{m}^{d+1}$ for the $\mathbb{C}$-vector space of polynomials of degree at most d. For $f, g \in \mathbb{C}[[t]]$, the notation $f = g + O(t^{d+1})$ means that $f - g \in \mathfrak{m}^{d+1}$. The field of fractions of $\mathbb{C}[t]$ is denoted by $\mathbb{C}(t)$.

6.2.1 Puiseux series

Let $R = \mathbb{C}[x_1^{\pm 1}, \dots, x_n^{\pm 1}]$ be the ring of Laurent polynomials in n variables and let $Y = (\mathbb{C} \setminus \{0\})^n$ be the n-dimensional algebraic torus. We consider a homotopy given by $H(x,t) : Y \times \mathbb{C} \to \mathbb{C}^n$:

$$H(x,t) = (h_1(x,t), \dots, h_n(x,t))$$

with $h_i \in R[t]$. We will denote

$$h_i = \sum_{\hat{q} \in \mathscr{A}_i} c_{\hat{q}} x^q t^{k_q}$$

where $\hat{q} = (q, k_q) \in \mathbb{Z}^n \times \mathbb{N}$ represents the exponent of a Laurent monomial in $R[t]$, $c_{\hat{q}} \in \mathbb{C}^*$ and $\mathscr{A}_i \subset \mathbb{Z}^n \times \mathbb{N}$ is the support of h_i. A *series solution at* $t^* = 0$ of $H(x,t)$ is a parametrization of the form

$$\begin{cases} x_j(s) = a_j s^{\omega_j} \left(1 + \sum_{\ell=1}^{\infty} a_{j\ell} s^{\ell}\right), & j = 1, \dots, n \\ t(s) = s^m \end{cases} \tag{6.2.1}$$

with $m \in \mathbb{N} \setminus \{0\}$, $\omega = (\omega_1, \dots, \omega_n) \in \mathbb{Z}^n$, $a = (a_1, \dots, a_n) \in (\mathbb{C}^*)^n, a_{j\ell} \in \mathbb{C}$ and such that $H(x(s), t(s)) = H(x_1(s), \dots, x_n(s), t(s)) \equiv 0$ and there is a real $\varepsilon > 0$ such that the series $x_j(s)$ converge for $0 < |s| \leq \varepsilon$. Such a series representation can be found for all irreducible components of $Z = H^{-1}(0)$ intersecting but not contained in the hyperplane $\{t = 0\}$ (see for instance [HV98, MM12, Mau80, MSW92b]). Substituting (6.2.1) in a monomial of h_i we get

$$x(s)^q t(s)^{k_q} = a^q s^{\langle \omega, q \rangle + m k_q} (1 + O(s))$$

where $\langle \cdot, \cdot \rangle$ is the usual pairing in $\mathbb{Z}^n$. It follows that the lowest order term in the series $h_i(x(s), t(s))$ has exponent $\min_{\hat{q} \in \mathscr{A}_i} (\langle \omega, q \rangle + m k_q)$. Denoting $\hat{\omega} = (\omega, m) \in \mathbb{Z}^{n+1}$ and

$$\partial_{\hat{\omega}} \mathscr{A}_i = \{\hat{q} \in \mathscr{A}_i | \langle \hat{\omega}, \hat{q} \rangle = \min_{\hat{q} \in \mathscr{A}_i} (\langle \hat{\omega}, \hat{q} \rangle)\}, \qquad \partial_{\hat{\omega}} h_i = \sum_{\hat{q} \in \partial_{\hat{\omega}} \mathscr{A}_i} c_{\hat{q}} x^q t^{k_q},$$

the vanishing of the lower order terms of $H(x(s), t(s))$ gives

$$\partial_{\hat{\omega}} h_i(a, 1) = \sum_{\hat{q} \in \partial_{\hat{\omega}} \mathscr{A}_i} c_{\hat{q}} a^q = 0, \qquad i = 1, \dots, n.$$

We note three things.

1. The set $\partial_{\hat{\omega}}\mathscr{A}_i$ contains at least two exponents, since none of the $c_{\hat{q}}$ are zero and $a \in (\mathbb{C}^*)^n$. It follows that $\partial_{\hat{\omega}}\mathscr{A}_i$ corresponds to a positive dimensional face $Q_{\hat{\omega}}$ of the convex hull P_i of $\mathscr{A}_i$. Since it is defined by $\hat{\omega} = (\omega, m)$ with $m \in \mathbb{N} \setminus \{0\}$, $Q_{\hat{\omega}}$ is contained in the *lower hull* of P_i (the facet normal points in the positive t-direction).

2. The point $(a, 1) \in (\mathbb{C}^*)^{n+1}$ is a solution of the *face system* corresponding to $\hat{\omega}$:
$$\partial_{\hat{\omega}} h_1(a, 1) = \cdots = \partial_{\hat{\omega}} h_n(a, 1) = 0.$$

3. The algorithm to compute more terms of the series is a generalization of the Newton-Puiseux procedure for algebraic plane curves and can be found, for instance, in [Mau80].

For $t = 0$, $H_0 = H(x, 0)$ represents a square polynomial system in the x_i and a series solution at $t = 0$ corresponds to a solution $x(0)$ of this system. If $\omega = 0$, $H(a, 0) = 0$ and hence $a \in (\mathbb{C}^*)^n$ is a toric solution. If one of the coordinates of ω, say ω_j is nonzero, then $x_j(s)$ is either zero for $s = 0$ ($\omega_j > 0$) or escapes to infinity as $s \to 0$ ($\omega_j < 0$).

Remark 6.2.1. A *series solution at* $t = t^*$, $t^* \in \mathbb{C}$ of $H(x, t)$ can be obtained from a series solution around $t = 0$ of $H'(x, t) = H(x, t + t^*)$. It satisfies $H(x(s), t(s)) = 0$ and has the form
$$\begin{cases} x_j(s) = a_j s^{\omega_j} \left(1 + \sum_{\ell=1}^{\infty} a_{j\ell} s^{\ell}\right), & j = 1, \ldots, n \\ t(s) = t^* + s^m \end{cases}.$$
△

Substituting $s = t^{1/m}$ in the coordinate functions we get
$$x_j(t) = a_j t^{\omega_j/m} \left(1 + \sum_{\ell=1}^{\infty} a_{j\ell} t^{\ell/m}\right), \quad j = 1, \ldots, n \tag{6.2.2}$$
which is a *Puiseux series* of order ω_j/m. We think of $x_j(t)$ as a function of a complex variable t, convergent by assumption in the punctured disk $0 < |t| \leq \varepsilon^m$. Then $t^* = 0$ is either a regular point if (6.2.2) is a Taylor series, a pole if it is a Laurent series with strictly negative powers, or a branch point if non integer fractional powers occur. Since in a regular point t^*, the $x_j(t)$ are Taylor series, they will have convergence radii equal to the distance to the nearest singular point t_s. The corresponding series solution(s) of $H(x, t)$ around $t = t_s$ will give the type of singularity. The discussion in this subsection shows that $t = t_s$ is either a branchpoint or a pole.

Example 6.2.1. Consider the algebraic plane curve given by $H(x, t) = tx^3 + 2x^2 + t$. The Newton polygon is given in the left part of Figure 6.4. The faces of the lower hull are indicated with bold blue lines. The facet normals are also shown in the figure (not to scale). From the discussion above, the parameters of any series solution $(x(s), t(s))$ must be such that $x(s) = as^{\omega}(1 + O(s)), t(s) = s^m$ with $\hat{\omega} = (\omega, m)$ equal to one of these facet normals. Furthermore, the constant a must be a nonzero solution of the face system $\partial_{\hat{\omega}} H(a, 1) = 0$. For $\hat{\omega}_1 = (-1, 1)$, the face equation is $tx^3 + 2x^2 = 0$ with nonzero solution $a = -2$ for $t = 1$. We expect a series solution $x_1(t) = -2t^{-1} + O(1)$. There are no other nonzero solutions to the face equation, so we consider the next facet normal. The vector $\hat{\omega}_2 = (1, 2)$ gives face equation $2x^2 + 1$ with two nonzero solutions $\pm\sqrt{-2}/2$. This gives $x_2 = \sqrt{-2t}/2 + O(t)$ and $x_3 = -\sqrt{-2t}/2 + O(t)$. The real parts of the solution curves are shown in the right part of Figure 6.4. △

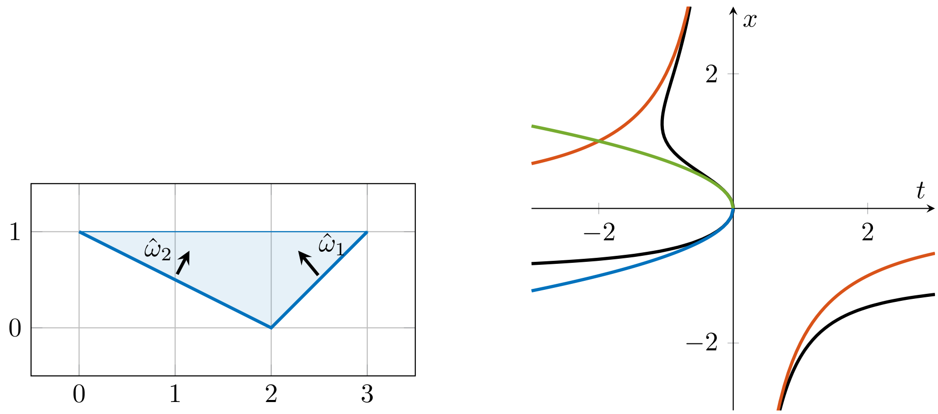


Figure 6.4: Left: Newton polygon of $H(x, t)$ from Example 6.2.1. Right: the curve $H(x, t) = 0$ (black), and the first term of the series expansions x_1 (orange), x_2 (green) and x_3 (blue).

6.2.2 Padé approximants

In this subsection we discuss Padé approximants and the way they behave in the presence of poles and branch points. An extensive treatment of Padé approximants can be found in [BJGM96]. We will limit ourselves to the definition and the properties that are relevant to the heuristics of our algorithm. The following definition uses some notation from [BJGM96].

Definition 6.2.1 (Padé approximant)**.** Let $x(t) = \sum_{\ell=0}^{\infty} c_\ell t^\ell \in \mathbb{C}[[t]]$. The type (L, M) Padé approximant of $x(t)$ is

$$[L/M]_x = \frac{p(t)}{q(t)} \in \mathbb{C}(t)$$

such that $p(t) \in \mathbb{C}[t]_{\leq L}$ and $q(t) \in \mathbb{C}[t]_{\leq M}$ is a unit in $\mathbb{C}[[t]]$, with

$$[L/M]_x - x \in \mathfrak{m}^k \tag{6.2.3}$$

for k maximal.

Informally, Padé approximants are rational functions agreeing with the Maclaurin series of a function x up to a degree that is as large as possible. They are generalizations of truncated Maclaurin series, which are type $(L,0)$ Padé approximants. Just like Maclaurin expansions are specific instances of Taylor expansions, it is straightforward to define Padé approximants around points $t = t^*$ in the complex plane different from 0. Without loss of generality, we consider only approximants around $t^* = 0$, since the general case reduces to this case after a simple change of coordinates. The type (L,M) Padé approximant is known to exist and it is unique. Multiplying the condition (6.2.3) by q yields

$$p(t) - x(t)q(t) \in \mathfrak{m}^k \quad \text{or equivalently,} \quad p(t) = x(t)q(t) + O(t^k) \tag{6.2.4}$$

for k maximal. Writing $p(t) = a_0 + a_1 t + \ldots + a_L t^L$, $q(t) = b_0 + b_1 t + \ldots + b_M t^M$ and equating terms of the same degree, this gives k linear conditions on the a_i, b_i, which can always be satisfied for $k \leq M + L + 1$. So for the linearized condition (6.2.4), k is at least $M+L+1$. Computing the a_i and b_i in practice is a nontrivial task. Difficulties are, for instance, degenerate situations where $\deg(p) < L$ or $\deg(q) < M$ and the presence of so-called *Froissart doublets* (spurious pole-zero pairs [Tre19, Chapter 27]). Some of the issues are discussed in [BJGM96, Chapter 2] and in [BM15, GGT13, IA13]. In [GGT13], a robust algorithm is proposed for computing Padé approximants. We will use this algorithm to compute Padé approximants from the coefficients c_i in our algorithm, presented in Section 6.4. The algorithm we use to compute the c_i is discussed in the next section.

What's important for our purpose is that a Padé approximant can be used to detect singularities of $x(t)$ of the types we are interested in (poles and branch points) close to $t^* = 0$, even for relatively small L and M. The idea is to compute Padé approximants of the coordinate functions $x_i(t)$ from local information on the path (the series coefficients c_i) and use them as a *radar* for detecting difficulties near the path. We are now going to motivate this. Since we intend to use Padé approximants to detect only *nearby* singularities, a natural first class to consider is the type $(L,1)$ approximants. We allow the approximant to have only one singularity, and hope that it chooses to place this singularity near the actual nearest singularity to capture the nearby non-analytic behaviour. Here is a powerful result due to Beardon [Bea68].

Theorem 6.2.1. *Let $x_j(t)$ be analytic in $\{t^* \in \mathbb{C} \mid |t^*| \leq r\}$. An infinite subsequence of $\{[L/1]_{x_j}\}_{L=0}^{\infty}$ converges to $x_j(t)$ uniformly in $\{t^* \in \mathbb{C} \mid |t^*| \leq r\}$.*

Proof. We refer to [Bea68] or [BJGM96, Theorem 6.1.1] for a proof. □

This applies in our case as follows. Suppose that $(a,0) \in Y \times \mathbb{C}$ is a regular point of the variety $Z = H^{-1}(0)$ and the irreducible component of Z containing $(a,0)$ is not contained in $\{(x,t^*) \in Y \times \mathbb{C} \mid t^* = 0\}$. Then there is a holomorphic function

$x : \mathbb{C} \to Y$ such that $x(0) = a$ and $H(x(t^*), t^*) = 0$ for t^* in some nonempty open neighborhood of 0 (see for instance Theorem A.3.2 in [WS05]). That is, if a is a regular solution of H_0, then the corresponding power series solution (6.2.1) consists of n Taylor series $x_j(t)$. The function $x(t)$ can be continued analytically in a disk with radius r if no singularities lie within a distance r from the origin. Theorem 6.2.1 makes the following statement precise. For large enough degrees L of the numerator of the Padé approximant, the $[L/1]_{x_j}$ are expected to approximate the coordinate functions $x_j(t)$ in a disk centered at the origin with radius $\pm$ the distance to the most nearby singularity. The fact that for sufficiently large L, the pole of $[L/1]_{x_j}$ is expected to give an indication of the *distance* to the nearest singularity (also if it is a branch point) can be seen as follows. Write $x_j(t) = \sum_{\ell=0}^{\infty} c_\ell t^\ell$ for the Maclaurin expansion of the coordinate function $x_j(t)$. Then a simple computation shows that if $c_L \neq 0$,

$$[L/1]_{x_j} = c_0 + c_1 t + \ldots + c_{L-1} t^{L-1} + \frac{c_L t^L}{1 - c_{L+1} t / c_L}.$$

Hence the pole of $[L/1]_{x_j}$ is c_L/c_{L+1} (or it is ∞ if $c_{L+1} = 0$). For large L, the modulus $|c_L/c_{L+1}|$ can be considered an approximation of the limit

$$\lim_{L \to \infty} \left| \frac{c_L}{c_{L+1}} \right|$$

if this limit exists. Also, if this limit exists it is a well-known expression for the convergence radius of the power series $x_j(t) = \sum_{\ell=0}^{\infty} c_\ell t^\ell$, which is the distance to the nearest singularity. Since the main application we have in mind is polynomial system solving, in which the homotopy is usually 'randomized', in practice this limit exists and for reasonably small L, $|c_L/c_{L+1}|$ is a satisfactory approximation of the convergence radius of the power series. Theorem 6.2.1 suggests that more is true: it can be expected that the ratio c_L/c_{L+1} is a reasonable estimate for the actual location of the most nearby singularity. This is Fabry's ratio theorem [Fab96]; see also [Bie55, Die57, Sue02].

Theorem 6.2.2. *If the coefficients of the power series $x_j(t) = \sum_{\ell=0}^{\infty} c_\ell t^\ell$ satisfy $\lim_{L\to\infty} c_L/c_{L+1} = t_s$, then $t = t_s$ is a singular point of the sum of this series. The point $t = t_s$ belongs to the boundary of the circle of convergence of the series.*

Proof. See [Fab96]. □

We now briefly discuss the behaviour of type (L, M) Padé approximants in the presence of poles and branch points and end the section with two illustrative examples.

Padé approximants and nearby poles

Since Padé approximants are rational functions, it is reasonable to expect that they can capture this kind of behaviour quite well. The following theorem, due to de Montessus [dM02], gives strong evidence of this intuition.

Theorem 6.2.3. *Suppose $x_j(t)$ is meromorphic in the disk $\{t^* \in \mathbb{C} \mid |t^*| \leq r\}$, with m distinct poles $z_1, \ldots, z_m \in \mathbb{C}$ in the punctured disk $\{t^* \in \mathbb{C} \setminus \{0\} \mid |t^*| < r\}$. Furthermore, suppose that μ_i is the multiplicity of the pole z_i and $\sum_{i=1}^m \mu_i = M$. Then $\lim_{L\to\infty}[L/M]_{x_j} = x_j$ on any compact subset of $\{t^* \in \mathbb{C} \mid |t^*| \leq r, t^* \neq z_i, i = 1, \ldots, \mu\}$.*

Proof. This is Theorem 6.2.2 in [BJGM96]. □

Loosely speaking, this tells us that the poles of $[L/M]_{x_j}$, for large enough L, will converge to the M most nearby poles of $x_j(t)$ (counting multiplicities), if these are the only singularities encountered in the disk $\{t^* \in \mathbb{C} \mid |t^*| \leq r\}$. For the $[L/1]_{x_j}$ approximant, this means that convergence may be expected beyond the nearest singularity if this is a simple pole, and the pole of $[L/1]_{x_j}$ will approximate the actual nearby pole. This may be considered as a practical approach to analytic continuation [Tre20]. Padé approximants also give answers to the inverse problem: the asymptotic behaviour of the poles of $[L/M]_{x_j}$ as $L \to \infty$ can be used to describe meromorphic continuations of the function $x_j(t)$. We do not give any details here, the interested reader is referred to [Gon81, Sue85, VLLP79].

Padé approximants and nearby branch points

Many singularities encountered in polynomial homotopy continuation are not poles, but branch points. This situation is more subtle since the Padé approximant, being a rational function, cannot have branch points. For an intuitive description of the behavior of Padé approximants for functions with multi-valued continuations, the reader may consult [BJGM96, Section 2.2]. The conclusion is that the poles and zeros of $[L/M]_{x_j}$ are expected to delineate a 'natural' branch cut. The authors also describe some ways to estimate the location and winding number of branch points using Padé approximants. We should also mention that there are convergence results in the presence of branch points which involve potential theory. We refer to [Sta97] for some important results for convergence of sequences of Padé approximants with $L, M \to \infty$, $L/M \to 1$ (so-called *near-diagonal* sequences). These results are beyond the scope of this chapter, mainly because we will limit ourselves to *near-polynomial* approximants: we allow only a small number of poles (often we even take $M = 1$) and we will estimate nearby singularities directly from $[L/M]_{x_j}$. This is an unusual choice, since near-diagonal approximants tend to show better behavior for the approximation of algebraic functions (see, e.g. [NST18, Section 6.2]). The reason for this choice will become clear in Example 6.2.3.

We will show in experiments that in this way, even for small L, we can predict at least the order of magnitude of the distance to the nearest branch point, which is enough to ring an alarm when this distance gets small, and often we can do much better.

The reason for limiting ourselves to a small number of parameters $L + M$ and for not trying to compute a very accurate location of the nearest branch point and its winding number is of course efficiency. Moreover, for the purpose of this thesis a local approximation of the coordinate functions and a rough estimate of the nearest singularity suffice. The above mentioned techniques to compute more information about nearby branch points may be powerful for approximation of algebraic curves in compact regions of the complex plane and for computing monodromy groups. We leave this as future research.

Example 6.2.2 (Padé approximants for function approximation and singularity detection)**.** We consider again the homotopy (6.1.1) from Example 6.1.1. Let us first take $p = 0.15$ and consider the smooth parameter path Γ_3. It is clear that the singularity $z_+ = 1/2 + p\sqrt{-1} \in S$ is the closest singularity to nearly every point in $\Gamma_3([0,1])$. As s moves closer to $1/2$, it moves closer to z_+. To show how this causes difficulties for the local approximation using Padé approximants, we have performed the following experiment. For several points t^* on the parameter path $\Gamma_3([0,1])$ we have plotted the contour in $\mathbb{C}$ where the absolute value of the difference between $x(t) = \sqrt{(t-1/2)^2 + p^2}$ and its type $(6,1)$ Padé approximation around t^* equals 10^{-4}. The result is shown in Figure 6.5. It is clear that the local approximation can be 'trusted' in a much larger region if the singularity is far away.

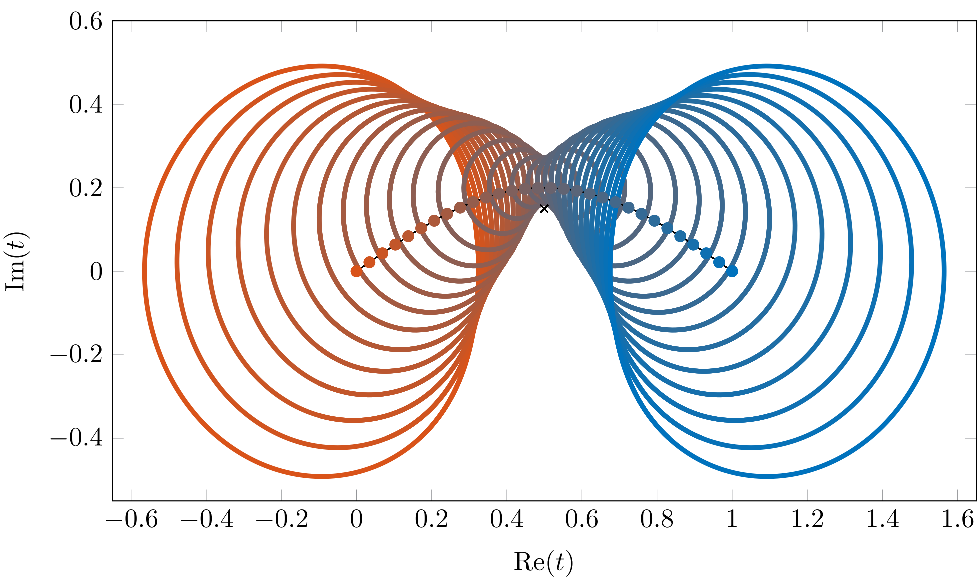


Figure 6.5: Contours of the approximation error as described in Section 6.2.2. The colour of the contours correspond to the color of the dots on the parameter path they correspond to. The singularity z_+ is shown as a small black cross.

We now investigate the behaviour of the pole of $[L/1]_x$ as we move along the path. We consider the four cases defined by $p = 0.15, 0.19$ and $L = 2, 6$. The results are shown in Figure 6.6. The figure shows that as we move closer to $\Gamma(0.5)$ on the path, the pole

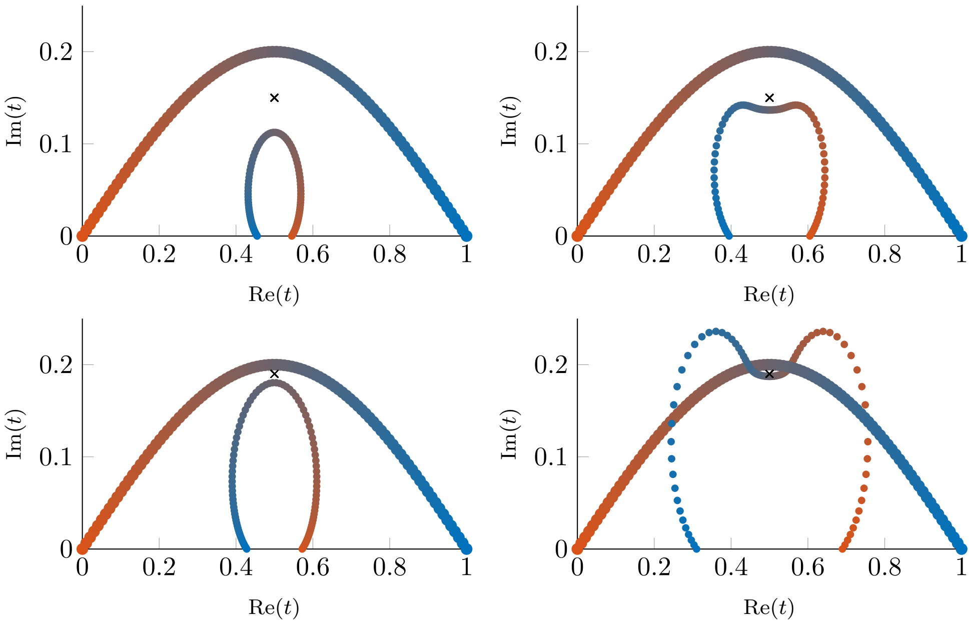


Figure 6.6: The path $\Gamma_3([0,1])$ and the corresponding path described by the pole of the type $(L,1)$ Padé approximant (associated points on the two paths have been given the same color) for $p = 0.15$ (first row), $p = 0.19$ (second row), $L = 2$ (left column), $L = 6$ (right column).

of the Padé approximant moves closer to the actual branch point. What's important is that in the trouble region of the path (s close to 0.5), the pole of $[L/1]_x$ is fairly close to z_+. It gives, at least, an indication of the order of magnitude of the distance to z_+. Another way to see this is that on a point of the path near to z_+, the $(L,1)$ Padé approximant is not so much influenced by the presence of z_-. For instance, at $t = 0$, the pole is real because z_+ and z_- are complex conjugates and they are located at the same distance from $\Gamma_3(0)$. For t^* near $\Gamma_3(0.5)$, the pole has a relatively large positive imaginary part. Comparing the first row to the second row in the figure shows that this effect gets stronger when a singularity moves closer to the path. Comparing the left column to the right column we see that the approximation of z_+ gets better as L increases, which is to be expected. If we use Γ_1 instead of Γ_3, for whatever p, the branch points z_+ and z_- will have the same distance to each point of the path. The result is that the $(L,1)$ Padé approximant will have poles on the real line. For $L = 4$, $p = 0.001$, $t^* \in [0,1]$, the pole is contained in the real interval $[0.4997, 0.5003]$, so the local difficulties are detected. However, in this specific situation, it is more natural to use type $(L,2)$ approximants. The result for $L = 6$, $p = 0.05$ is shown in Figure 6.7. We note that in a randomized homotopy, it is not to be expected that at

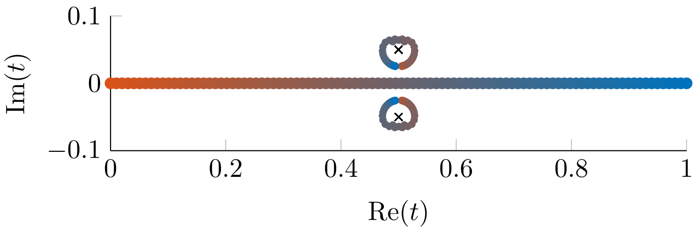


Figure 6.7: The path $\Gamma_1([0,1])$ and the corresponding paths described by the poles of the type $(6,2)$ Padé approximant (associated points on the two paths have been given the same color) for $p = 0.05$.

a general point of the path two poles are equally important. As we move along the path, the most important singularity may change, and the type $(L,1)$ approximant can be expected to relocate its pole accordingly. △

Example 6.2.3 (Near-diagonal VS near-polynomial approximants)**.** Consider the algebraic function $x(t) = \sqrt{(t+1.01)(t^2 - t + 37/4)}$ with branch points

$$\mathcal{S} = \{-1.01, 1/2 + 3\sqrt{-1}, 1/2 - 3\sqrt{-1}\}.$$

For $\ell = 1, \ldots, 13$, we compute both the type (ℓ,ℓ) and the type $(2\ell-1,1)$ Padé approximant (around $t = 0$) of $x(t)$ using a Matlab implementation of the algorithm in [GGT13]. For all these approximants we compute

1. the minimum of the distances of the poles of the Padé approximant to the branch point -1.01,
2. the difference between the smallest modulus of the poles of the Padé approximant and the modulus of the nearest branch point, which is 1.01,
3. an estimate for the approximation error (the infinity norm of a discretized approximation) of $x(t)$ on the disk $|t| \leq 1/2$ in the complex plane.

Results are shown in Figure 6.8. The right part of the figure shows that the diagonal approximants behave better for function approximation. However, for small ℓ, the near-polynomial approximants are competitive. For the type $(2\ell-1,1)$ approximant, the first two quantities coincide since the pole is real. For the (ℓ,ℓ) case, the first quantity is a lower bound for the second one. This is illustrated by the difference between the dashed and the full blue line in Figure 6.8. What happens is the following. One of the poles of the type (ℓ,ℓ) approximant approximates the branch point 1.01, but some other pole indicates that there could be a branch point with smaller modulus. This is illustrated in Figure 6.9 for $\ell = 3, 4$ (for $\ell = 4$, one of the poles of the (ℓ,ℓ) approximant lies close to that of the $(2\ell-1,1)$ approximant and the corresponding dot is nearly invisible). The pole of the type $(3,3)$ approximant that is closest to the origin actually comes from a Froissart doublet which was not detected using the default settings in the algorithm of [GGT13]. As a consequence, this spurious pole

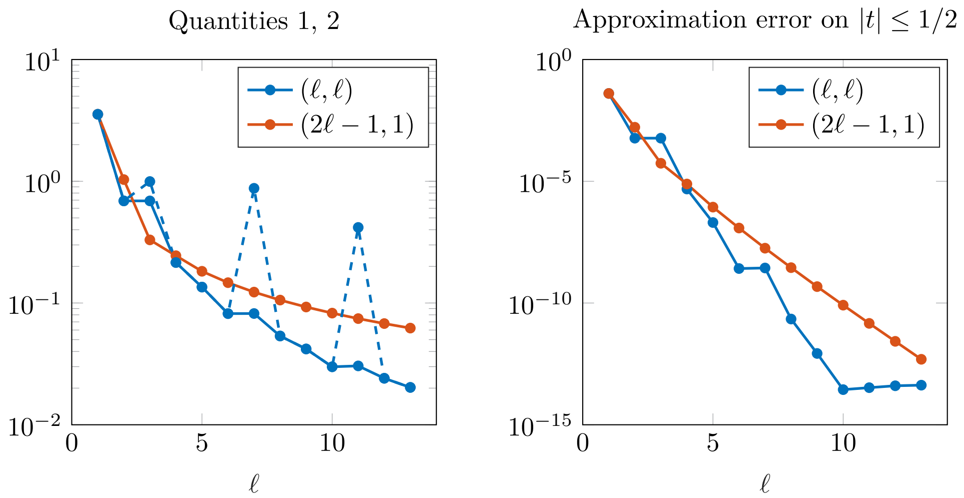


Figure 6.8: Results of the experiment in Example 6.2.3.

would tell us that a singularity is nearby such that only a small step can be taken (see Subsection 6.4.1), while the actual branch point is quite far away. Detecting such Froissart doublets is often tricky. Since we will use only low orders, the approximation quality of the $(L, 1)$ approximant suffices for our purpose. Moreover, this example shows that they are more robust for estimating the distance to the nearest singularity. We will use this type of approximants for our default settings.

$\triangle$

6.3 Computing power series solutions

In this section we present the algorithm for computing a power series solution of $H(x,t) = (h_1(x,t), \dots, h_n(x,t))$ at $t^* = 0$ proposed in [BV18a] and prove a result of convergence. An analogous result for the case $n = 1$ can be found in [Lip76]. We will consider the situation where the series solution has the form (6.2.1) with parameters satisfying $\omega_i \geq 0$. Futhermore, we assume that the winding number m is known. If this is not the case, m can be computed by using, for instance, monodromy loops. Note that it is sufficient to consider the case where $m = 1$, since if m is known and $m > 1$ we can consider the homotopy

$$\hat{H}(x, \tau) = (h_1(x, \tau^m), \dots, h_n(x, \tau^m))$$

with power series solution

$$\begin{cases} x_j(s) = a_j s^{\omega_j} \left(1 + \sum_{\ell=1}^{\infty} a_{j\ell} s^\ell\right), \quad j = 1, \dots, n \\ \tau(s) = s \end{cases}.$$

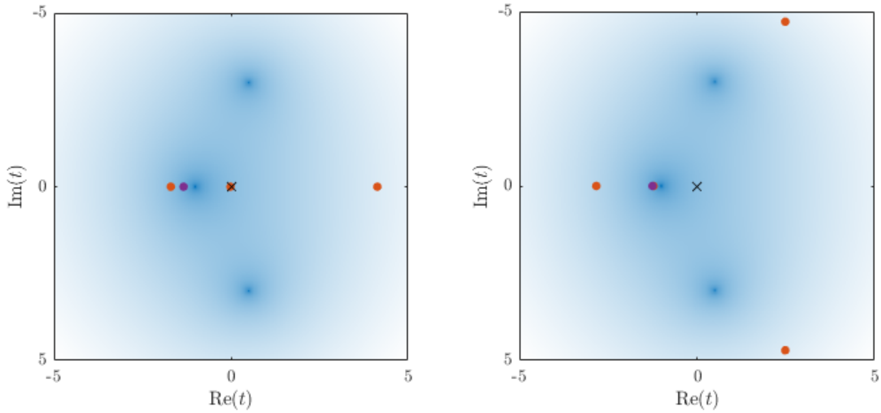


Figure 6.9: Poles of the type (ℓ, ℓ) approximant (orange dots) and pole of the type $(2\ell-1, 1)$ approximant (purple dot) for $\ell = 3, 4$ (left and right respectively). The origin is indicated with a black cross. The background color corresponds to $|x(t)|$ (dark regions correspond to small absolute values).

Therefore, we can avoid introducing the extra parameter s and the unknown power series solution is given by

$$x_j(t) = a_j t^{\omega_j} \left(1 + \sum_{\ell=1}^{\infty} a_{j\ell} t^{\ell}\right), j = 1, \ldots, n. \tag{6.3.1}$$

We think of $H(x,t)$ as a column vector $[h_1 \ \cdots \ h_n]^\top$ in $R[[t]]^n \simeq R^n[[t]]$ and the Jacobian matrix $J_H(x,t)$ is considered an element of $R[[t]]^{n\times n} \simeq R^{n\times n}[[t]]$. For any $h(x,t) \in R[[t]]^n$, plugging in $y(t) \in \mathbb{C}[[t]]^n$ gives $h(y(t),t) \in \mathbb{C}[[t]]^n$, and the same can be done for $J(x,t) \in R[[t]]^{n\times n}$, which gives $J(y(t),t) \in \mathbb{C}[[t]]^{n\times n}$.

Definition 6.3.1. Let $\star$ be either $\mathbb{C}^n$ or $\mathbb{C}^{n\times n}$. For $v = \sum_{\ell=0}^{\infty} v_\ell t^\ell \in \star[[t]] \setminus \{0\}$, the *order* of v is

$$\operatorname{ord}(v) = \min_{\ell}\{v_\ell \neq 0\},$$

where $v_\ell \in \star, \ell \in \mathbb{N}$. For $w \neq v \in \star[[t]]$ we denote $v = w + O(t^k)$ if $\operatorname{ord}(v-w) \geq k$. For $v = 0$, we define $\operatorname{ord}(v) = \infty$.

Note that this means that for a vector or matrix v with power series entries, $v = O(t^k)$ if and only if every entry of v is in $\mathfrak{m}^k$, where $\mathfrak{m}$ is the maximal ideal of $\mathbb{C}[[t]]$. With elementwise addition and multiplication in $\mathbb{C}[[t]]^n$ and the usual addition and multiplication in $\mathbb{C}[[t]]^{n\times n}$, it is clear that for $v, w \in \star[[t]]$, $\operatorname{ord}(v) = \operatorname{ord}(-v)$, $\operatorname{ord}(v+w) \geq \min(\operatorname{ord}(v), \operatorname{ord}(w))$ and $\operatorname{ord}(vw) \geq \operatorname{ord}(v) + \operatorname{ord}(w)$. For the product rule, equality holds if $\star = \mathbb{C}^n$. Matrix-vector multiplication $\mathbb{C}[[t]]^{n\times n} \times \mathbb{C}[[t]]^n \to \mathbb{C}[[t]]^n$ is defined in the usual way and for $M \in \mathbb{C}[[t]]^{n\times n}, v \in \mathbb{C}[[t]]^n$ we have $\operatorname{ord}(Mv) \geq \operatorname{ord}(M) + \operatorname{ord}(v)$.

Given $x^{(0)}(t) = (x_1^{(0)}(t), \dots, x_n^{(0)}(t)) \in \mathbb{C}[[t]]^n$, fix positive integers $w_k \in \mathbb{N} \setminus \{0\}$ and consider the sequence $\{x^{(k)}(t)\}_{k\geq 0}$ defined by

$$\begin{aligned} \tilde{x}^{(k+1)}(t) &= x^{(k)}(t) - J_H(x^{(k)}(t),t)^{-1}H(x^{(k)}(t),t) = \sum_{\ell=0}^{\infty} b_\ell t^\ell, \\ x^{(k+1)}(t) &= \sum_{\ell=0}^{w_k-1} b_\ell t^\ell \end{aligned} \tag{6.3.2}$$

where we assume that $J_H(x^{(k)}(t),t)$ is a unit in $\mathbb{C}[[t]]^{n\times n}$ for all k and this is equivalent to assuming that $J_H(x^{(k)}(0),0)$ is invertible for all $k \geq 0$. The iteration is clearly based on the well-known Newton-Raphson iteration for approximating a root of a nonlinear system of equations. The following proposition specifies the statement that the iteration has similar 'quadratic' convergence properties. It is related to a multivariate version of Hensel lifting, see for instance [Eis13, Exercise 7.26].

Proposition 6.3.1. *Let $H(x,t) : Y \times \mathbb{C} \to \mathbb{C}$ be a homotopy with power series solution $x(t) \in \mathbb{C}[[t]]^n$ given by* (6.3.1) *and let $\{x^{(k)}(t)\}_{k\geq 0}$ be a sequence generated as in* (6.3.2). *If $J_H(x^{(k)}(t),t)$ is a unit in $\mathbb{C}[[t]]^{n\times n}$ for all $k \geq 0$ then*

$$\operatorname{ord}(x^{(k+1)}(t) - x(t)) \geq \min(2\operatorname{ord}(x^{(k)}(t) - x(t)), w_k), \quad k \geq 0.$$

Proof. We know that $x(t) = (x_1(t), \dots, x_n(t))^\top \in \mathbb{C}[[t]]^n$ satisfies $H(x(t),t) = 0$. Take $x^{(k)}(t) \in \mathbb{C}[[t]]^n$ and define $e^{(k)}(t) = x^{(k)}(t) - x(t)$. We have

$$0 = H(x^{(k)}(t) - e^{(k)}(t),t) = H(x^{(k)}(t),t) - J_H(x^{(k)}(t),t)e^{(k)}(t) + O(t^{2\operatorname{ord}(e^{(k)}(t))}). \tag{6.3.3}$$

By assumption, $J_H(x^{(k)}(t),t)$ is a unit and thus $\operatorname{ord}(J_H(x^{(k)}(t),t)^{-1}) = 0$. We now multiply (6.3.3) from the left with $J_H(x^{(k)}(t),t)^{-1}$ and we get (using $e^{(k)}(t) = x^{(k)}(t) - x(t)$)

$$-J_H(x^{(k)}(t),t)^{-1}H(x^{(k)}(t),t) + (x^{(k)}(t) - x(t)) = O(t^{2\operatorname{ord}(e^{(k)}(t))}).$$

It follows that $\tilde{x}^{(k+1)}(t) - x(t) = O(t^{2\operatorname{ord}(e^{(k)}(t))})$. So we find that

$$\operatorname{ord}(e^{(k+1)}(t)) \geq \min(2\operatorname{ord}(e^{(k)}(t)), w_k). \qquad \square$$

It follows that if $e^{(0)}(t)$ has order ≥ 1, the iteration converges to the solution $x(t)$ and the order of the error doubles in every iteration, as long as the truncation orders w_k allow for it. Also, if $\operatorname{ord}(e^{(0)}(t)) \geq 1$, $H(x^{(0)}(0),0) = 0$ and thus $\operatorname{ord}(H(x^{(0)}(t),t)) \geq 1$. It follows that the term $-J_H(x^{(k)}(t),t)^{-1}H(x^{(k)}(t),t)$ has order at least 1 and so the constant terms of $x^{(1)}$ and $x^{(0)}$ agree. This stays true for the following iterations as well. We conclude that if $\operatorname{ord}(e^{(0)}(t)) \geq 1$, the assumption that $J_H(x^{(k)}(t),t)$ is a unit for all k translates to the assumption that $x^{(0)}(0) = a$ is a regular solution of the polynomial system defined by H_0. If we want to compute a series solution that is accurate up to order w, and $\operatorname{ord}(e^{(0)}(t)) = r \geq 1$, we set $w_k = \min(r2^k, w)$ and execute $\lceil \log_2(w/r) \rceil$ steps of the iteration. We denote

$$\begin{aligned}
J_H(x^{(k)}(t), t) &= J_0^{(k)} + J_1^{(k)} t + J_2^{(k)} t^2 + \ldots, \\
H(x^{(k)}(t), t) &= H_0^{(k)} + H_1^{(k)} t + H_2^{(k)} t^2 + \ldots, \\
\Delta x^{(k)}(t) &= -J_H(x^{(k)}(t), t)^{-1} H(x^{(k)}(t), t) = d_0^{(k)} + d_1^{(k)} t + d_2^{(k)} t^2 + \ldots.
\end{aligned}$$

We have to compute the first w_k terms of $\tilde{x}^{(k+1)}(t) = x^{(k)}(t) + \Delta x^{(k)}(t)$. The equation

$$-J_H(x^{(k)}(t), t)\Delta x^{(k)}(t) = H(x^{(k)}(t), t)$$

gives

$$\begin{aligned}
J_0^{(k)} d_0^{(k)} &= -H_0^{(k)}, \\
J_0^{(k)} d_1^{(k)} + J_1^{(k)} d_0^{(k)} &= -H_1^{(k)}, \\
J_0^{(k)} d_2^{(k)} + J_1^{(k)} d_1^{(k)} + J_2^{(k)} d_0^{(k)} &= -H_2^{(k)}, \\
&\vdots \\
J_0^{(k)} d_{w_k-1}^{(k)} + J_1^{(k)} d_{w_k-2}^{(k)} + \ldots + J_{w_k-1}^{(k)} d_0^{(k)} &= -H_{w_k-1}^{(k)}.
\end{aligned}$$

It is an immediate corollary from Proposition 6.3.1 that if $\operatorname{ord}(e^{(0)}(t)) = r \geq 1$, then $d_i^{(k)} = 0, i = 0, \ldots, w_{k-1} - 1$ and hence $H_i^{(k)} = 0, i = 0, \ldots, w_{k-1} - 1$. It follows that we only have to solve

$$\begin{aligned}
J_0^{(k)} d_{w_{k-1}}^{(k)} &= -H_{w_{k-1}}^{(k)}, \\
&\vdots \\
J_0^{(k)} d_{w_k-1}^{(k)} + J_1^{(k)} d_{w_k-2}^{(k)} + \ldots + J_{w_k - w_{k-1} - 1}^{(k)} d_{w_{k-1}}^{(k)} &= -H_{w_k-1}^{(k)}.
\end{aligned} \tag{6.3.4}$$

and this can be done equation by equation, via backsubstitution. In practice, we will use these results as in Algorithm 6.8, where we assume that $r = 1, t^* \in \mathbb{C}$, $x^{(0)} \in \mathbb{C}^n \subset \mathbb{C}[[t]]^n$ such that $H(x^{(0)}, t^*) = 0$.

6.4 A robust algorithm for tracking smooth paths

In this section we show how the results of the previous sections lead to a smooth path tracking algorithm. More specifically, we propose a new adaptive stepsize predictor for homotopy path tracking. We will use $\Gamma(s) = s$ and assume that this is a smooth parameter path for simplicity, but the generalization to different parameter paths is straightforward. The aim of this section is to motivate the heuristics and to present and analyze the algorithm. In the next section we will show some convincing experiments.

Algorithm 6.8 Computes the power series solution of $H(x,t)=0$ corresponding to $x^{(0)}$ around $t=t^*$.

procedure ComputeSeries$(H, t^*, w, x^{(0)})$
  $H \leftarrow H(x, t+t^*)$
  $k \leftarrow 0$
  while $k < \lceil \log_2(w) \rceil$ **do**
    $w_k \leftarrow \min(2^k, w)$
    Compute $x^{(k+1)}$ by solving (6.3.4)
    $k \leftarrow k+1$
  end while
  return $\{x_1^{(k)}(t), \dots, x_n^{(k)}(t)\}$
end procedure

We will use Padé approximants for the prediction. The stepsize computation is based on two criteria. That is, we compute two candidate stepsizes $\{\Delta t_1, \Delta t_2\}$ based on two different estimates of the largest 'safe' stepsize. The eventual value of Δt that is returned by the predictor (line 6 in Algorithm 6.7) is $\min(\Delta t_1, \Delta t_2, t_{\mathrm{EG}} - t^*)$. For the first criterion we estimate the distance to the nearest point of a different path in $Y \times \{t^*\}$. This estimate is only accurate if we are actually in a difficult region. Comparing this to an estimate for the Padé approximation error we compute Δt_1 such that the predicted point $\tilde{z}$ is much closer to the correct path than to the nearest different path. The value of Δt_2 is an estimate for the radius of the 'trust region' of the Padé approximant, which is influenced by nearby singularities in the parameter space (see Section 6.2). We discuss these two criteria in detail in the first subsection. In the second subsection we present the algorithm.

6.4.1 Adaptive stepsize: two criteria

The values of Δt_1 and Δt_2 are computed from an estimate of the distance to the nearest different path, the approximation error of the Padé approximant for small stepsizes and an estimate for some global 'trust radius' of the Padé approximants. We discuss these estimates first and then turn to the computation of Δt_1 and Δt_2 from these estimates.

Distance to the nearest path

We will use $\|\cdot\|$ to denote the euclidean 2-norm for vectors and the induced operator norm for matrices. Consider the homotopy $H : Y \times \mathbb{C} \to \mathbb{C}^n$. Suppose that for some $t^* \in [0,1)$ we have $H(z_{t^*}^{(1)}, t^*) = H(z_{t^*}^{(2)}, t^*) = 0$, so $z_{t^*}^{(1)} \neq z_{t^*}^{(2)} \in Z_{t^*}$ lie on two different solution paths. We assume that $z_{t^*}^{(1)}$ is close to $z_{t^*}^{(2)}$. Denote $\Delta z = z_{t^*}^{(2)} - z_{t^*}^{(1)} \in \mathbb{C}^n$ and think of Δz as a column vector. Our goal here is to estimate $\|\Delta z\|$. Neglecting higher order terms, we get

$$H(z_{t^*}^{(2)}, t^*) \approx H(z_{t^*}^{(1)}, t^*) + J_H(z_{t^*}^{(1)}, t^*)\Delta z + \frac{v}{2}, \quad v = \begin{bmatrix} \langle \mathcal{H}_1(z_{t^*}^{(1)}, t^*)\Delta z, \Delta z\rangle \\ \vdots \\ \langle \mathcal{H}_n(z_{t^*}^{(1)}, t^*)\Delta z, \Delta z\rangle \end{bmatrix} \tag{6.4.1}$$

where

$$(\mathcal{H}_i(x,t))_{j,k} = \frac{\partial^2 h_i}{\partial x_j \partial x_k}, \quad 1 \le j,k \le n$$

are the Hessian matrices of the individual equations and $\langle \cdot, \cdot \rangle$ is the usual inner product in $\mathbb{C}^n$. To simplify the notation, we denote $\mathcal{H}_i = \mathcal{H}_i(z_{t^*}^{(1)}, t^*)$ and $J_H = J_H(z_{t^*}^{(1)}, t^*)$. The Hessian matrices are Hermitian, so they have a unitary diagonalization (see Remark B.4.2) $\mathcal{H}_i = \mathbf{U}_i \mathbf{T}_i \mathbf{U}_i^H$ where the $\mathbf{T}_i$ are diagonal matrices and the $\mathbf{U}_i$ are unitary matrices with eigenvectors of $\mathcal{H}_i$ in their columns. We may write $\Delta z = \mathbf{U}_i w_i$ for some coefficient vector w_i such that $\|w_i\| = \|\Delta z\|$. We have

$$\langle \mathcal{H}_i \Delta z, \Delta z\rangle = \langle \mathbf{T}_i w_i, w_i\rangle.$$

Let $\sigma_{k,\ell} = \sigma_\ell(\mathcal{H}_k)$ be the ℓ-th singular value of $\mathcal{H}_k$. The absolute values of the diagonal entries of $\mathbf{T}_i$ are exactly these singular values, so that

$$|\langle \mathcal{H}_i \Delta z, \Delta z\rangle| \le \sigma_{i,1}\|w_i\|^2 = \sigma_{i,1}\|\Delta z\|^2.$$

It follows easily that

$$\|v\| \le \sqrt{\sigma_{1,1}^2 + \ldots + \sigma_{n,1}^2}\,\|\Delta z\|^2.$$

Since $\|J_H \Delta z\| \ge \sigma_n(J_H)\|\Delta z\|$ and by (6.4.1) we have $\|J_H \Delta z\| \approx \|v\|/2$, it follows that

$$\|\Delta z\| \gtrsim \frac{2\sigma_n(J_H)}{\sqrt{\sigma_{1,1}^2 + \ldots + \sigma_{n,1}^2}}. \tag{6.4.2}$$

Intuitively, the 'more regular' the Jacobian at $(z_{t^*}^{(1)}, t^*)$, the larger the lower bound (6.4.2) becomes. On the other hand, the 'larger the curvature' of Z at $(z_{t^*}^{(1)}, t^*)$, the smaller the upper bound (6.4.2) becomes. Motivated by (6.4.2), we make the following definition.

Definition 6.4.1. For $z_{t^*}^{(i)} \in Z_{t^*}, t^* \in [0,1)$, set $J_H = J_H(z_{t^*}^{(i)}, t^*)$ and $\sigma_{k,\ell} = \sigma_\ell(\mathcal{H}_k(z_{t^*}^{(i)}, t^*))$ and define

$$\eta_{i,t^*} = \frac{2\sigma_n(J_H)}{\sqrt{\sigma_{1,1}^2 + \ldots + \sigma_{n,1}^2}}.$$

The numbers η_{i,t^*} are estimates for the distance to the most nearby different path. To make sure the prediction error $\|x(t^* + \Delta t) - \tilde{x}(t^* + \Delta t)\|$ (where $\tilde{x}(t)$ is the coordinatewise Padé approximant) is highly unlikely to cause path jumping, we will solve $\|x(t^* + \Delta t) - \tilde{x}(t^* + \Delta t)\| = \beta_1 \eta_{i,t^*}$ for a small fraction $0 < \beta_1 \ll 1$ to compute an adaptive stepsize Δt. We now discuss how to estimate $\|x(t^* + \Delta t) - \tilde{x}(t^* + \Delta t)\|$.

Approximation error of the Padé approximant

Without loss of generality, we take the current value of t to be zero and consider Padé approximants around $t^* = 0$ as in Section 6.2. Suppose that we have computed a type (L, M) Padé approximant $[L/M]_{x_j} = p_j(t)/q_j(t)$ of a coordinate function $x_j(t)$ around 0. Given a small real stepsize Δt, we want to estimate the error

$$|e_j(\Delta t)| = \left|\frac{p_j(\Delta t)}{q_j(\Delta t)} - x_j(\Delta t)\right| = \left|\frac{a_0 + a_1\Delta t + \ldots + a_L\Delta t^L}{b_0 + b_1\Delta t + \ldots + b_M\Delta t^M} - x_j(\Delta t)\right|. \tag{6.4.3}$$

From Definition 6.2.1 we know that $e_j(t) \in \mathfrak{m}^k$ (where usually $k = L + M + 1$), so (6.4.3) can be written as $|e_{0,j}\Delta t^k + e_{1,j}\Delta t^{k+1} + \ldots|$ with $e_{0,j} \neq 0$. For small Δt, the first term is expected to dominate the sum and so $|e_j(\Delta t)| \approx |e_{0,j}\Delta t^k|$. This estimate is also used in [GS04] for the case $L = 2, M = 1$ and a similar strategy is common to estimate the error in a power series approximation. An alternative is to use an estimate for the 'linearized' error

$$|q_j(\Delta t)e_j(\Delta t)| = |p_j(\Delta t) - x_j(\Delta t)q_j(\Delta t)| \tag{6.4.4}$$

which is equal to

$$|(b_0 + b_1\Delta t + \ldots + b_M\Delta t^M)(e_{0,j}\Delta t^k + e_{1,j}\Delta t^{k+1} + \ldots)| \simeq |b_0 e_{0,j}\Delta t^k|.$$

Since $q_j(t)$ is a unit in $\mathbb{C}[[t]]$, $b_0 \neq 0$ and we can scale p_j and q_j such that $b_0 = 1$ and the estimates of (6.4.3) and (6.4.4) coincide. Taking $b_0 = 1$, the constant $e_{0,j}$ is the coefficient of t^k in $(a_0 + a_1 t + \ldots + a_L t^L) - (1 + b_1 t + \ldots + b_M t^M)(c_0 + c_1 t + \ldots)$, which is easily seen to be

$$e_{0,j} = a_k - (c_k + b_1 c_{k-1} + \ldots + b_M c_{k-M}) \tag{6.4.5}$$

where $a_k = 0$ if $k > L$ and $c_j = 0$ for $j < 0$. Doing this for all j and assuming that k is the same for all coordinates we get an estimate

$$\|x(\Delta t) - \left(\frac{p_1(\Delta t)}{q_1(\Delta t)}, \ldots, \frac{p_n(\Delta t)}{q_n(\Delta t)}\right)\| \approx \|e_0\||\Delta t|^k,$$

with $e_0 = (e_{0,1}, \ldots, e_{0,n})$.

Trust region for the Padé approximant

As discussed in Subsection 6.2.2 and illustrated in Example 6.2.2, branchpoints in the parameter space that are close to the parameter path cause problems for the Padé approximation. If none of the poles of $[L/M]_{x_j}$ are close to a current parameter value on the path, we may be able to take a reasonably large step forward without getting into difficulties. However, since we take L and M to be small, we cannot expect the approximants $[L/M]_{x_j}$ to have already converged in a disk with radius the distance to the nearest singularity. Nor can we expect that the poles of $[L/M]_{x_j}$ are very good approximations of the actual singularities. Taking the distance D to the most nearby pole of $[L/M]_{x_j}$ as an estimate for the convergence radius is a very rough estimate in this case. However, we observe that D does give an estimate of the order of magnitude of the region in which $[L/M]_{x_j}$ is a satisfactory approximation. The conclusion is that we do not use D itself, but $\beta_2 D$ where $0 < \beta_2 < 1$ is a safety factor.

The candidate stepsizes Δt_1 and Δt_2

We now use the ingredients presented above to compute two candidate stepsizes Δt_1 and Δt_2. For Δt_1, we use the estimate η_{i,t^*} for the distance to the nearest path and the estimate $\|e_0\||\Delta t|^k$ for the approximation error of the Padé approximant. The heuristic is that we want the approximation error to be only a small fraction of the estimated distance to the nearest path, so that the predicted point $\tilde{z}$ is much closer to the path being tracked than to the nearest different path. That is, we solve

$$\|e_0\||\Delta t_1|^k = \beta_1 \eta_{i,t^*}$$

for Δt_1, where $\beta_1 > 0$ is a small factor. Since the attraction basins of Newton correction can behave in unexpected ways, it is best to take β_1 to be fairly small, for instance $\beta_1 = 0.005$. This gives

$$\Delta t_1 = \sqrt[k]{\frac{\beta_1 \eta_{i,t^*}}{\|e_0\|}}.$$

Both the estimates η_{i,t^*} and $\|e_0\||\Delta t|^k$ are only accurate in case trouble is near (they are based on lowest order approximations). If the resulting Δt_1 is large, the only thing this tells us is that we are not on a difficult point on the path with high probability. The second candidate stepsize, Δt_2, will make sure we don't take a step that is too large in this situation. At the same time, Δt_2 will be small when singularities in the parameter space are near the current point on the path. Let D be the distance to the nearest pole out of all the poles of the $[L/M]_{x_j}, j = 1, \ldots, n$. We set

$$\Delta t_2 = \beta_2 D$$

where $0 < \beta_2 < 1$ is a safety factor which should not change the order of magnitude, for instance $\beta_2 = 0.5$.

Example 6.4.1. As mentioned above, the estimate η_{i,t^*} for the distance to the nearest different path is only accurate when another path is actually near. If this is not the case, Δt_1 may be too large and we need Δt_2 to make sure the resulting stepsize is still safe. To see that it is not enough to take only Δt_2 into account, consider the homotopy

$$H(x,t) = (x - (t - (a + b\sqrt{-1}))^2)(x + (t - (a + b\sqrt{-1}))^2), \quad t \in [0,1],$$

with $a, b \in \mathbb{R}$, $0 < a < 1$ and $|b|$ small. The paths corresponding to the two solutions are smooth and can be analytically continued in the entire complex plane: there are no singular points in $x_1(t), x_2(t)$. However, for $t = a + b\sqrt{-1}$ the two solutions coincide. By the assumptions on a and b, this value of t lies close to the parameter path $[0, 1]$. Intuitively, the singularity of the Jacobian $J_H = \partial H/\partial x$ is canceled by a zero of $\partial H/\partial t$: along the solution paths we have

$$\frac{dx}{dt} = \frac{-\frac{\partial H}{\partial t}}{\frac{\partial H}{\partial x}} = \frac{4(t-(a+b\sqrt{-1}))^3}{2x} = \frac{4(t-(a+b\sqrt{-1}))^3}{\pm 2(t-(a+b\sqrt{-1}))^2} = \pm 2(t-(a+b\sqrt{-1})).$$

For $t = a$, the solutions are $x_1 = -b^2, x_2 = b^2$, so for small b, the paths are very close to each other. The type $(1, 1)$ Padé approximant will have no poles (or very large ones due to numerical artefacts), so taking only this criterion into account would allow us to take large steps. However, the estimate (6.4.2) at $t = a$ gives $|\Delta z| \approx 4b^2/2$, which is exactly the distance to the nearest different path. $\triangle$

6.4.2 Path tracking algorithm

We are now ready to present the path tracking algorithm. Since our contribution is in the predictor step (line 6 in Algoritm 6.7), we focus on this part. The predictor algorithm is Algorithm 6.9 below. It is straightforward to embed this predictor algorithm in the template Algorithm 6.7.

Algorithm 6.9 Predictor algorithm

1: **procedure** Predict($H, z_{t^*}^{(i)}, t^*, L, M, \beta_1, \beta_2, t_{\mathrm{EG}}$)
2: $\{x_1(t), \ldots, x_n(t)\} \leftarrow \textsc{ComputeSeries}(H, t^*, L + M + 2, z_{t^*}^{(i)})$
3: $D \leftarrow \infty$
4: compute η_{i,t^*} as in Definition 6.4.1
5: **for** $j = 1, \ldots, n$ **do**
6: $\quad p_j, q_j \leftarrow \textsc{PadéApprox}(x_j(t), L, M)$
7: $\quad$ compute $e_{0,j}$ using (6.4.5)
8: $\quad D \leftarrow \min(D, \min\{|z| \mid q_j(z) = 0\})$
9: **end for**
10: $e_0 \leftarrow (e_{0,1}, \ldots, e_{0,n})$
11: $\Delta t_1 \leftarrow \sqrt[k]{\frac{\beta_1 \eta_{i,t^*}}{\|e_0\|}}$
12: $\Delta t_2 \leftarrow \beta_2 D$
13: $\Delta t \leftarrow \min(\Delta t_1, \Delta t_2, t_{\mathrm{EG}} - t^*)$
14: $\tilde{z} \leftarrow (p_1(\Delta t)/q_1(\Delta t), \ldots, p_n(\Delta t)/q_n(\Delta t))$
15: **return** $\tilde{z}, \Delta t$
16: **end procedure**

We briefly discuss some of the steps in Algorithm 6.9. In line 2, Algorithm 6.8 is used. The point around which we compute the series is t^*, the current parameter value on the path. The parameter $w = L + M + 2$ is the number of coefficients needed to compute the Padé approximant of type (L, M) and the approximation error estimate. The starting value of the power series is the constant vector $x^{(0)} = z_{t^*}^{(i)}$, satisfying $H(z_{t^*}^{(i)}, t^*) = 0$ such that $\mathrm{ord}(e^{(0)}) > 0$ and convergence is guaranteed. In step 6, the type (L, M) Padé approximant of the coordinate function $x_j(t)$ is computed using the algorithm of [GGT13]. Algorithm 6.9 has some more input parameters than the predictor in the template algorithm. We will usually take M very small (and often 1), motivated by the conclusions of Section 6.2. The value of L is chosen, for instance, such that $L + M + 2$ is a power of 2 e.g. $L = 5, M = 1$, because of the quadratic convergence property of the iteration in Algorithm 6.8 proved in Proposition 6.3.1. Reasonable values for β_1, β_2 are $\beta_1 = 0.005, \beta_2 = 0.5$ as stated before. The parameter t_{EG} is the beginning of the endgame operating region as in Section 6.1.

Figure 6.10 shows a summary of our a priori adaptive step control algorithm: Newton's method is followed by the Padé approximant computation and the differentiation to calculate the Hessians is followed by the singular value decompositions.

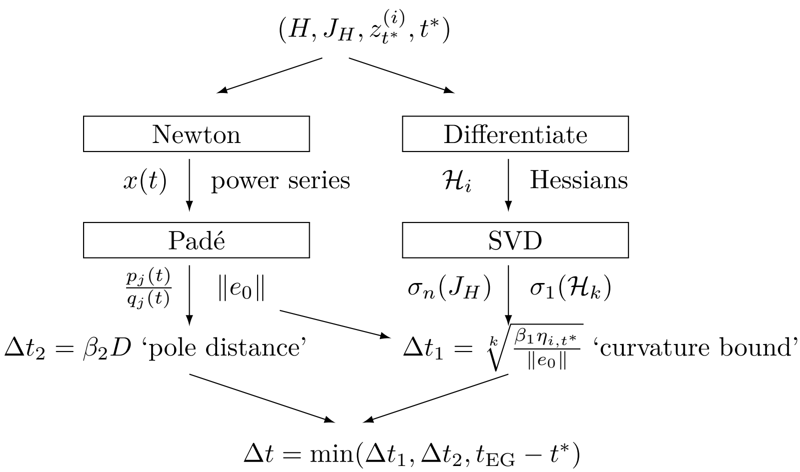


Figure 6.10: Schematic summary of an a priori adaptive step control algorithm.

Remark 6.4.1. We conclude this section with a remark on the complexity of Algorithm 6.9 as a function of the number of variables n in comparison with a posteriori step size control algorithms. As is detailed in Subsection 4.3 in [TVBV19], taking one step in the predictor-corrector scheme with Algorithm 6.9 can be expected to be *at most* $O(n \log(n))$ times more expensive than a standard step control using Newton iteration in the corrector and a predictor which runs in $O(n)$ time (e.g. a fourth order extrapolator). This is assuming that the Padé parameters L and M behave as $O(n)$, which is quite restrictive. For a full complexity analysis, one should take into account that because of the 'a priori' strategy, (virtually) none of the steps have to be *re-taken*, making the feedback loops in Figure 6.1 unnecessary. As we will see in the next section, Algorithm 6.9 also allows us to track some paths using only *very few steps*, even for problems with high degrees. In [Tim20], the step size candidate Δt_1 is replaced by a different heuristic which is cheaper to compute and it complements the step size Δt_2 in a similar way. This way the computational cost is reduced significantly while the reliability seems to be maintained. This algorithm will soon be the default in the Julia package HomotopyContinuation.jl [BT18]. $\triangle$

6.5 Numerical experiments

In this section we show some numerical experiments to illustrate the effectiveness of the techniques proposed in this chapter. Algorithm 6.9 is implemented in PHCpack (v2.4.72), available at `https://github.com/janverschelde/PHCpack`, and in Julia. In the experiments, our implementations are compared with the state of the art. We will use the following short notations for the different solvers in our experiments:

`brt_DP`	Bertini v1.6 using double precision (MPTYPE = 0) [BSHW13],
`brt_AP`	Bertini v1.6 using adaptive precision (MPTYPE = 2) [BHSW08],
`HC.jl`	HomotopyContinuation.jl v1.1 [BT18],
`phc -p`	The `phc -p` command of PHCpack v2.4.72 [Ver99],
`phc -u`	Our algorithm, used in PHCpack v2.4.72 via `phc -u`,
`Padé.jl`	Our algorithm, implemented in Julia.

We use default double precision settings for all these solvers, except `brt_AP`, for which we use default adaptive precision settings. The experiments in all but the last subsection are performed on an 8 GB RAM machine with an intel Core I7-6820HQ CPU working at 2.70 GHz (this is the machine that was used for most experiments in previous chapters as well). We restrict all solvers to the use of only one core for all the experiments, unless stated otherwise. We will use $\Gamma : [0,1] \mapsto \mathbb{C} : s \mapsto s$, which will be a smooth parameter path as defined in Section 6.1 by the constructions in the experiments. Therefore, the parameter s will not occur in this section and paths are of the form $\{(x(t), t), t \in [0,1)\} \subset X \times [0,1)$. In all experiments, we use $\beta_1 = 0.005, \beta_2 = 0.5$. To measure the quality of a numerical solution of a system of polynomial equations, we compute its residual as explained in Appendix C.

Experiment 6.5.1 (A family of hyperbolas)**.** Consider again the homotopy (6.1.1) from Example 6.1.1, which represents a family of hyperbolas parametrized by the real parameter p. Recall that the ramification locus is $\mathcal{S} = \{1/2 + p\sqrt{-1}\}$. We will consider $p \neq 0$ here, such that $[0,1]$ is a smooth parameter path. The smaller $|p|$, the closer the branchpoints move to the line segment $[0,1]$. Figure 6.11 shows that as the value of $p > 0$ decreases, the two solution paths approach each other for parameter values $t^* \approx 0.5$ which causes danger for path jumping. This is confirmed by our experiments. Table 6.1 shows the results. We used $L = 5, M = 1$ in `phc -u`. The Julia implementation `HC.jl` checks whether the starting solutions are (coincidentally)

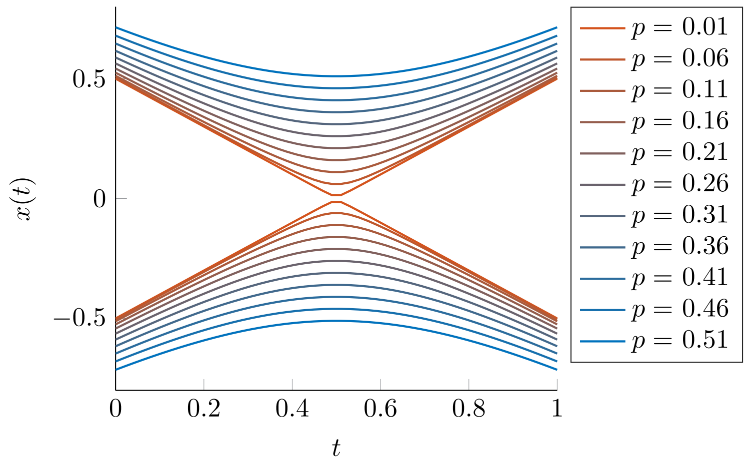


Figure 6.11: Family of hyperbolas from Experiment 6.5.1.

Solver \ k	1	2	3	4	5	6	7
`brt_DP`	✓	✓	✓	✗	✗	✗	✗
`brt_AP`	✓	✓	✓	✓	✗	✗	✗
`HC.jl`	✓	✗	✗	✗	✗	✗	✗
`phc -p`	✓	✗	✗	✗	✗	✗	✗
`phc -u`	✓	✓	✓	✓	✓	✓	✓

Table 6.1: Results of Experiment 6.5.1 for $p = 10^{-k}, k = 1, \dots, 7$. A '✗' indicates that path jumping happened.

solutions of the target system. For this reason, with this solver, we track for $t \in [0.1, 1]$. △

Experiment 6.5.2 (Wilkinson polynomials)**.** As a second experiment, consider the Wilkinson polynomial $W_d(x) = \prod_{i=1}^{d}(x-i)$ for $d \in \mathbb{N}_{>0}$. When $d > 10$, it is notoriously hard to compute the roots of these polynomials numerically when they are presented in the standard monomial basis. For Bertini and HomotopyContinuation.jl, we use the blackbox solvers to find the roots of the $W_d(x)$. In PHCpack, we use

$$H(x,t) = (x^d - 1)(1-t) + \gamma W_d(x)t$$

with γ a random complex number.[3] The case $d = 12$ is illustrated in Figure 6.12. We use default settings for other solvers and $L = 5, M = 1$ in our algorithm to solve $W_d(x)$ for $d = 10, \dots, 19$. The results are reported in Table 6.2. The number e is the number of failures, i.e. d minus the number of distinct solutions (up to a certain tolerance) returned by each solver with residual $< 10^{-9}$, and T is the computation

[3]The other solvers use $\Gamma(s) = 1 - s$ by default. This is not important here.

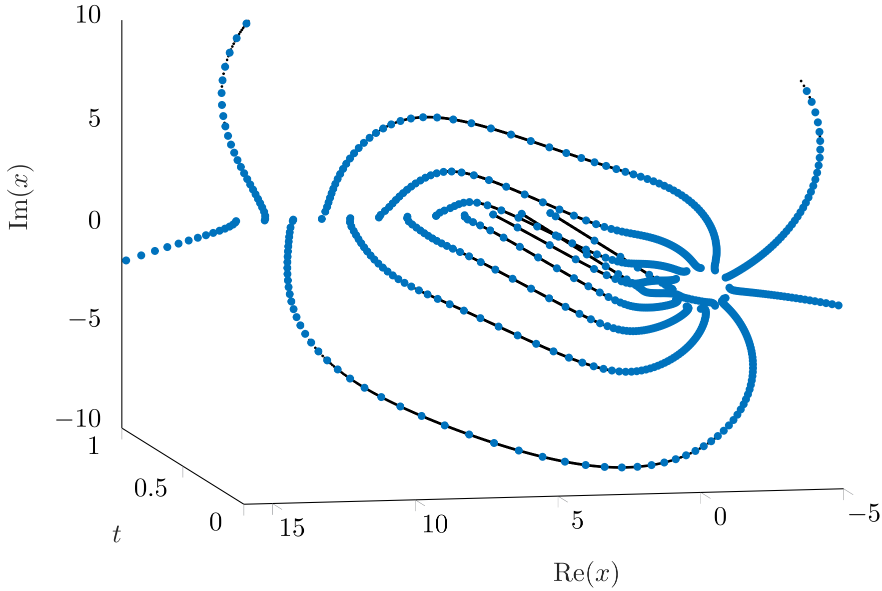


Figure 6.12: Solution paths for a random linear homotopy as in Experiment 6.5.2 connecting the 12th roots of unity to the roots of $W_{12}(x)$. The blue dots are the numerical approximations of points on the paths computed by our algorithm using $L = M = 1$.

d	`phc -p`		`HC.jl`		`brt_DP`		`brt_AP`		`phc -u`		
	e	T	e	T	e	T	e	T	e	T	#
10	5	8.0e-3	0	2.5e-3	0	4.5e-2	0	2.5e-2	0	4.0e-2	23-42
11	7	2.9e-2	0	3.6e-3	0	1.9e-1	0	1.4e+0	0	5.2e-2	12-45
12	9	3.4e-2	0	6.7e-3	0	1.5e-1	0	2.0e+0	0	6.9e-2	12-50
13	10	3.5e-2	0	4.1e-3	0	3.2e-1	0	2.8e+0	0	1.1e-1	35-54
14	11	2.4e-2	1	6.2e-3	0	4.8e-1	0	3.8e+0	0	1.0e-1	12-69
15	13	1.7e-2	1	9.0e-3	15	1.5e-2	15	1.6e-2	0	1.2e-1	43-63
16	15	2.1e-2	6	6.7e-3	16	1.6e-2	16	1.4e-2	0	1.7e-1	12-74
17	16	1.6e-2	10	3.2e-3	17	1.8e-2	17	1.3e-2	0	1.9e-1	11-73
18	18	6.0e-3	11	1.4e-2	18	1.8e-2	18	1.4e-2	0	2.4e-1	57-81
19	18	1.8e-2	13	7.0e-3	19	1.8e-2	19	1.4e-2	0	2.6e-1	12-83

Table 6.2: Results for Experiment 6.5.2.

time in seconds. The column indexed by '#' gives the minimum and maximum number of steps on a path for our solver. We conclude this experiment with a brief comparison with certified tracking algorithms. For $W_4(x)$, the algorithm[4] proposed in [BL13] takes 6261.6 steps for the path starting at $z_0 = -1$ (this is averaged out over 5 experiments with random, rational γ). For $W_{15}(x)$ the certified tracking algorithm of [XBY18] (which is specialized for the univariate case) takes on average 790 steps per path. $\triangle$

Experiment 6.5.3 (Generic polynomial systems)**.** In this experiment, we consider random, square polynomial systems and solve them using the different homotopy continuation packages and the algorithm proposed in this chapter. We now specify what 'random' means. Fix n and $d \in \mathbb{N} \setminus \{0\}$. A *generic polynomial system* of dimension n and degree d is given by a generic member of the square family $\mathcal{F}_R(d, \ldots, d)$. That is, we generate

$$f_i(x) = \sum_{|a| \leq d} c_{i,a} x^a \in R = \mathbb{C}[x_1, \ldots, x_n], \quad i = 1, \ldots, n,$$

where $c_{i,a}$ are complex numbers whose real and imaginary parts are drawn from a standard normal distribution to obtain

$$F : \mathbb{C}^n \to \mathbb{C}^n : x \mapsto (f_1(x), \ldots, f_n(x)).$$

The *solutions* of F are the points in the fiber $F^{-1}(0) \subset \mathbb{C}^n$, and by Bézout's theorem, there are d^n such points. In order to find these solutions, we track the paths of the homotopy

$$H(x, t) = G(x)(1 - t) + \gamma F(x)t, \quad t \in [0, 1]$$

where γ is a random complex constant and

$$G : \mathbb{C}^n \to \mathbb{C}^n : x \mapsto (x_1^d - 1, \ldots, x_n^d - 1)$$

represents the *start system* with d^n known, regular solutions. Results are given in Table 6.3. In the table, n and d are as in the discussion above and e is the number of failures (i.e. d^n minus the number of successfully computed solutions, as in Experiment 6.5.2). For `phc -u`, the column indexed by '#' gives the minimum and maximum number of steps on a path, and the column indexed by h gives the ratio of the number of steps for which $\Delta t = \Delta t_1$ is the first candidate stepsize. In this experiment, we took $L = 5, M = 1$ and we set the maximum stepsize to be 0.5. Note that even for this type of generic systems, the 'difficulty' of the paths (based on the number of steps needed) can vary strongly. The case $n = 1, d = 300$ is not supported by `HC.jl`, because only one byte is used to represent the degree. Note that `HC.jl` performs extremely well in all other cases in this experiment, both in terms of speed and robustness. The extra comparative experiment in the next subsection will show that, for difficult (non-generic) paths, our heuristic shows better results (this was also shown in Experiments 6.5.1 and 6.5.2). $\triangle$

[4] We use a Macaulay2 implementation, available at `http://people.math.gatech.edu/~aleykin3/RobustCHT/` to perform these experiments.

n	d	phc -p		HC.jl		brt_DP		brt_AP		phc -u			
		e	T	e	T	e	T	e	T	e	T	#	h
1	20	0	5.0e+0	0	1.7e-3	0	3.1e-2	0	7.5e-2	0	4.2e-2	6-16	0.09
	50	0	2.6e-2	0	6.3e-3	0	1.3e-1	0	2.3e+0	0	2.4e-1	5-27	0.07
	100	2	9.1e-2	0	1.1e-2	49	5.3e-1	0	1.2e+1	0	8.9e-1	4-27	0.13
	200	2	2.7e-1	0	3.2e-2	97	1.6e+0	1	4.5e+1	0	2.9e+0	5-25	0.13
	300	5	6.6e-1	×	×	221	2.8e+0	27	3.3e+2	0	8.3e+0	4-49	0.13
2	10	0	1.8e-1	0	1.5e-2	0	3.8e-1	0	2.4e+0	0	2.1e+0	8-37	0.10
	20	2	2.2e+0	0	8.9e-2	0	1.4e+1	0	1.2e+2	0	2.6e+1	8-55	0.13
	30	8	1.2e+1	0	3.3e-1	0	9.9e+1	0	2.0e+3	0	1.3e+2	8-68	0.13
	40	22	3.7e+2	0	9.1e-1	68	3.5e+2	0	7.8e+3	0	4.2e+2	6-57	0.15
	50	39	8.7e+2	0	2.3e+0	12	1.4e+3	0	3.4e+4	0	1.0e+3	7-57	0.14
3	5	0	3.5e-1	0	3.0e-2	0	7.0e-1	0	7.0e-1	0	4.8e+0	9-55	0.09
	9	1	8.5e+0	0	2.3e-1	0	2.1e+1	0	4.8e+1	0	9.8e+1	8-56	0.10
	13	4	6.8e+1	0	1.5e+0	0	2.3e+2	0	1.0e+3	0	8.3e+2	8-85	0.11

Table 6.3: Results for Experiment 6.5.3.

Experiment 6.5.4 (Clustered solutions)**.** Homotopies that cause danger for path jumping are such that for some parameter value t^* on the path, the map $H(x,t^*)$ is a polynomial system with some solutions that are clustered together. Motivated by this, we construct the following experiment. Let n_c be a parameter representing the number of solution clusters and let CS represent the 'cluster size'. We consider the set of clusters $\{C_1, \ldots, C_{n_c}\}$ where $C_i = \{z_{i,1}, \ldots, z_{i,\mathrm{CS}}\} \subset \mathbb{C}$ is a set of complex numbers that are 'clustered' in the following sense. Take $c_i = e^{\frac{i-1}{n_c}2\pi\sqrt{-1}}$ and for a real parameter α, we define

$$z_{i,j} = c_i + \alpha u^{1/\mathrm{CS}} e^{\frac{j-1}{\mathrm{CS}}2\pi\sqrt{-1}},$$

where u is the unit roundoff ($\approx 10^{-16}$ in double precision arithmetic). Define the polynomial

$$E(x) = \prod_{i=1}^{n_c}\left(\prod_{j=1}^{\mathrm{CS}}(x - z_{i,j})\right).$$

The situation is illustrated in Figure 6.13 for $n_c = \mathrm{CS} = 5$, $\alpha = 100$. For $\alpha = 1$, we know from classical perturbation theory of univariate polynomials that the roots of $E(x)$ look like the roots of a slightly perturbed version of a polynomial whose n_c roots are the cluster centers, which have multiplicity CS. We will use $\alpha \geq 10$, such that the roots of $E(x)$ are not 'numerically singular'. Let $d = n_c\mathrm{CS}$. Let $G(x) = x^d - 1$ and let $F(x)$ be a polynomial of degree d with random complex coefficients, with real and imaginary part drawn from a standard normal distribution. We consider the homotopy

$$H(x,t) = (1-t)(1/2-t)G(x) + \gamma_1 t(1-t)E(x) + \gamma_2 t(1/2-t)F(x), \quad t \in [0,1]$$

where γ_1 and γ_2 are random complex constants. $G(x)$ represents the start system with starting solutions the d-th roots of unity. By tracking the homotopy H, the polynomial $G(x)$ is continuously transformed into the random polynomial $F(x)$, passing through the polynomial $(\gamma_1/4)E(x)$ (for $t^* = 1/2$) with clustered solutions. The *success rate*

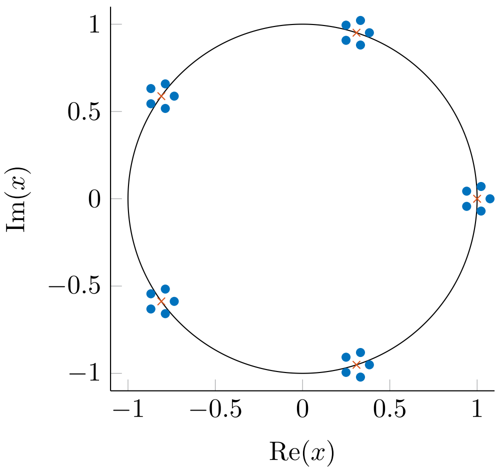


Figure 6.13: Roots (blue dots) and cluster centers (orange crosses) of $E(x)$ constructed as in Experiment 6.5.4 with $n_c = \mathrm{CS} = 5$, $\alpha = 100$.

(SR) of a numerical path tracker for solving this problem is defined as follows. Let $\hat{d}$ be the number of points among the solutions of $F(x)$ that coincide with a point returned by the path tracker up to a certain tolerance (e.g. 10^{-6}). We set $\mathrm{SR} = \hat{d}/d$. For fixed α, n_c, CS, we track 10 homotopies $H(x,t)$ constructed as above with different random γ_i using `HC.jl` and `Padé.jl`. We compute the average success rate for these 10 runs. Results are reported below. For each problem, the best average success rate is highlighted in blue.

$n_c = 5$

α	Solver \ CS	1	2	3	4	5
10	`HC.jl`	1.0	0.740	0.100	0.060	0.080
	`Padé.jl`	1.0	0.990	0.993	0.995	0.988
100	`HC.jl`	1.0	1.0	0.627	0.985	0.980
	`Padé.jl`	1.0	1.0	1.0	0.985	0.996
1000	`HC.jl`	1.0	1.0	1.0	1.0	1.0
	`Padé.jl`	1.0	1.0	0.987	1.0	1.0

$n_c = 10$

α	Solver \ CS	1	2	3	4	5
10	`HC.jl`	1.0	0.095	0.083	0.078	0.504
	`Padé.jl`	1.0	0.995	1.0	1.0	0.990
100	`HC.jl`	1.0	0.530	0.673	0.982	1.0
	`Padé.jl`	1.0	1.0	0.997	0.988	1.0
1000	`HC.jl`	1.0	0.995	0.990	1.0	0.310
	`Padé.jl`	1.0	0.995	0.997	1.0	0.992

△

Experiment 6.5.5 (Benchmark Problems)**.** Parallel computations were applied for the problems in this section. For two families of structured polynomial systems, our experiments show that no path failures and no path jumps occur, even when the number of solution paths goes past one million.

The program for this experiment is available in the MPI folder of PHCpack, available in its source code distribution on github, under the current name `mpi2padcon`. The code was executed on two 22-core 2.2 GHz Intel Xeon E5-2699 processors in a CentOS

Linux workstation with 256 GB RAM. The number of processes for each run equals 44. The root node manages the distribution of the start solutions and the collection of the end paths. In a static work load assignment, the other 43 processes each track the same number of paths.

The `katsura` family of systems is named after the problem posed by Katsura [Kat94], see [Kat90] for a description of its relevance to applications. The `katsura`-n problem consists of n quadratic equations and one linear equation. The number of solutions equals 2^n, which is the Bézout number. Table 6.4 summarizes the characteristics and wall clock times on `katsura`-n, for n ranging from 12 to 20. While the times with HOM4PS-2.0para [LT09] are much faster than in Table 6.4, Table 3 of [LT09] reports 2 and 4 path jumpings respectively for `katsura`-19 and `katsura`-20. In the runs with the MPI version for our code, no path failures and no path jumping happened. The good results we obtained required the use of homogeneous coordinates. When tracking the paths first in affine coordinates, we observed large values for the coordinates, which forced too small step sizes, which then resulted in path failures. Although the defining equations are nice quadrics, the condition numbers of the solutions gradually increase as n grows. For example, for $n = 20$, the largest condition number of the Jacobian matrix was of the order 10^7, observed for 66 solutions. Table 6.4 reports the number of real solutions in the column with header #real and the number of solutions with nonzero imaginary part under the header #imag.

n	#sols	#real	#imag	wall clock time (seconds)	
12	4,096	582	3,514	7.925E+01	1m 19s
13	8,192	900	7,292	2.081E+02	3m 28s
14	16,384	1,606	14,778	5.065E+02	8m 27s
15	32,768	2,542	30,226	1.456E+03	24m 16s
16	65,536	4,440	61,096	4.156E+03	1h 9m 16s
17	131,072	7,116	123,956	1.001E+04	2h 46m 50s
18	262,144	12,458	249,686	2.308E+04	6h 24m 15s
19	524,288	20,210	504,078	5.696E+04	15h 49m 20s
20	1,048,576	35,206	1,013,370	1.317E+05	36h 34m 11s

Table 6.4: Wall clock time on 44 processes on the `katsura` problem, in a static workload balancing schedule with one manager node and 43 worker nodes. Only the workers track solution paths.

Another interesting class of polynomial systems [Noo89] was introduced to the computer algebra community by [Gat90]. The n-dimensional system consists of n cubic equations and originated from a model of a neural network. The Bézout bound on the number of solutions is attained. Although the permutation symmetry could be exploited with a symmetric homotopy, using the algorithms in [VC94], this was not done for the computations summarized in Table 6.5. We used homogeneous coordinates in the runs.

The formulation of the polynomials in [Noo89] depends on one parameter c, which was set to 1.1. The number of real solutions is reported in Table 6.5 in the column with header #real and the number of solutions with nonzero imaginary part is under the header #imag. Because every new equation is of degree three and the number of

n	#sols	#real	#imag	wall clock time (seconds)	
10	59,029	21	59,008	3.478E+03	57m 58s
11	177,125	23	177,102	1.594E+04	4h 25m 37s
12	531,417	25	531,392	7.202E+04	20h 0m 17s
13	1,594,297	27	1,594,270	3.030E+05	84h 9m 58s

Table 6.5: Wall clock time on 44 processes, in a static workload balancing schedule with one manager node and 43 worker nodes. Only the worker nodes track solution paths.

paths triples, the wall clock time increases more than in the previous benchmark. As before, no path failures and no path jumping happened. $\triangle$

Chapter 7

Conclusion and future work

In this thesis we have addressed the problem of solving systems of polynomial equations with finitely many solutions using several different approaches.
A first class of methods, referred to as *algebraic methods* in Subsection 1.3.1, is based on *eigenvalue-eigenvector theorems* relating the eigenstructure of a commuting family of matrices to the set of solutions. These matrices represent multiplication with some function in the quotient algebra associated to the system. Such methods require the computation of *rewriting rules* modulo the ideal. We have shown that, in a numerical context, it is crucial to use a representation of the algebra for which the problem of computing these rewriting rules is a well-conditioned problem. This gives rise to *truncated normal forms* (TNFs) in a natural way. TNFs provide a general framework for normal forms, leading to a class of algorithms containing both Gröbner and border basis methods. Unlike the monomial representations that are used in the literature, which are usually restricted to be induced by a monomial ordering or to be connected (to 1), our methods use more general, possibly non-monomial bases for the quotient algebra leading to significantly more robust numerical algorithms. We have presented explicit constructions for solving 'generic' square polynomial systems inspired by the theory of projective and toric resultants. The homogeneous counterpart of the TNF framework provides a variant of the algorithms for finding zero-dimensional solution sets on $\mathbb{P}^n$ or another compact toric variety X. For this we use a global description of the solutions by homogeneous ideals in the Cox ring S of X. *Homogeneous normal forms* (HNFs) provide rewriting rules for a graded piece of S modulo the corresponding graded piece of the ideal. We developed the necessary theory for generalizing the standard eigenvalue-eigenvector theorem to the toric setting. The approach gives rise to some questions regarding the *multigraded regularity* of such ideals, some of which were answered in this thesis.
Next to these methods based on algebraic techniques, we have also considered homotopy continuation methods (see Subsection 1.3.2). These methods are very important and popular. One of the reasons is that their complexity scales relatively well with the

dimension of the solution space. However, homotopy continuation algorithms depend on some choices of heuristics and thresholds which should be chosen carefully in order for the methods to be reliable. In particular, bad choices may lead to the occurrence of *path jumping*, which may be fatal if one's objective is to find *all* solutions of a system. We have revisited the core steps of the standard predictor-corrector scheme for continuation algorithms and proposed a new method for *a priori adaptive step size control* which proves to be reliable and significantly less prone to path jumping than state of the art implementations.

7.1 Contributions

In this section we highlight the contributions of the different chapters to the field of solving systems of polynomial equations.

CHAPTER 3

While the results of Chapter 3 are well known, some of the material is presented in a slightly non-standard way in order to emphasize the analogy with some of the new results in later chapters. We hope that this may be a valuable resource for further development and improvement of the techniques presented in Chapters 4 and 5.

CHAPTER 4

The results of Chapter 4 are published in [TVB18, TMVB18, MTVB19].

- In Section 4.2 we define the natural concept of a *truncated normal form* (TNF) and prove complete characterizations in terms of properties that are relatively easy to check or prove in practice. The main results are Theorems 4.2.1 and 4.2.2.
- In Subsection 4.3.1 we prove that for square, dense systems a TNF can be computed as the cokernel of a resultant map (Proposition 4.3.2).
- In Subsection 4.3.2 we present an algorithm for solving generic dense systems with an automated choice of basis for the quotient algebra (Algorithm 4.1).
- Subsection 4.3.3 contains many numerical experiments, showcasing the strengths of Algorithm 4.1 in comparison to state of the art software.
- In Section 4.4 we illustrate the flexibility of the TNF framework by proposing the use of different representations of the quotient algebra and some options for efficient computation of TNFs. In particular, we illustrate the use of the SVD for basis selection and TNFs in a product Chebyshev basis.
- Section 4.5 introduces and characterizes *homogeneous normal forms* (HNFs) in the projective setting and presents an explicit construction (Algorithm 4.2) for

zero-dimensional, square homogeneous systems. Proposition 4.5.2 is the main result.

CHAPTER 5

Most results of Chapter 5 are published in [TMVB18, Tel20]. Some of the results can be found in [BT20a].

- Section 5.3 generalizes Theorem 4.2.2 to the toric case and presents a TNF algorithm for solving systems which are generic members of a square polyhedral family. We prove an explicit TNF construction from the cokernel of a resultant map based on a toric resultant matrix construction from Canny and Emiris (Corollary 5.3.1), which leads to Algorithm 5.3.

- In Subsection 5.5.2 we define a notion of multigraded regularity of homogeneous ideals in the Cox ring of a toric variety X defining finitely many points on X with multiplicity 1 ($V_X(I)$ is a reduced, zero-dimensional subscheme). We also define homogeneous Lagrange polynomials and prove several connections between these polynomials and the regularity (Propositions 5.5.1 and 5.5.2) and some properties of the ideal I (Lemma 5.5.2 and Proposition 5.5.3).

- We prove a toric version of the eigenvalue-eigenvector theorem in Subsection 5.5.3. The main result is Theorem 5.5.3. In addition, we prove conditions under which the eigenvalues of homogeneous multiplication matrices can be used directly to obtain points on the solution orbits (Theorems 5.5.4 and 5.5.5).

- In Subsection 5.5.4 we generalize HNFs to the toric setting. The main result is Proposition 5.5.5, which is used to design algorithms 5.5 and 5.6. Algorithm 5.6 is tested in several experiments which show that it can deal with degenerate systems of equations in a robust way.

- In Subsection 5.5.5 we generalize the toric eigenvalue-eigenvector theorem to the non-reduced case: we allow multiplicities. The result is Theorem 5.5.6. In addition we prove several properties of the regularity of a homogeneous zero-dimensional ideal in the Cox ring. The main results are Theorem 5.5.7 and Corollary 5.5.5, which both imply weaker versions of a conjecture in [Tel20].

CHAPTER 6

The results of this chapter are submitted for publication [TVBV19].

- In Subsection 6.2.2 we highlight some of the properties of Padé approximants in the context of homotopy continuation and illustrate with examples that they can be used as ‘radars’ for finding singularities along the solution paths.

- In Section 6.4 we propose a new algorithm (Algorithm 6.9) that uses these insights.

- The experiments in Section 6.5 show that with this algorithm we accomplished our goal of designing a path tracker that is significantly more robust with respect to path jumping than the state of the art implementations.

7.2 Future directions

We conclude by listing some open challenges for future research. Throughout the text, for each solution space X and in each different context, we have assumed that the given equations define finitely many points in X. In the case where there are positive dimensional solution components, the isolated solutions can still be recovered via the *regular* eigenvalues of a *singular* hidden variable resultant pencil. Because of recent advances such as [HMP19], solving such singular eigenvalue problems is now tractable and these connections could be exploited to develop *eigenvalue methods for computing a numerical irreducible decomposition.*

An important scenario is when there are finitely many solutions in $(\mathbb{C}^*)^n$, but positive dimensional components are sitting in the boundary of the torus in the toric compactification X. If we know in which torus invariant prime divisor such a component is located, we can get rid of it by performing a *numerical saturation with respect to one of the variables in the Cox ring.*

In this text, the constructions we proposed for solving square (Laurent) polynomial systems were based on resultant matrix constructions. This has the important drawback that even for moderate dimensions of the solution space (i.e. $n = 4, 5, \ldots$), the size of these matrices gets *much* bigger than the number of solutions to the system. This establishes the need for *TNF constructions which operate on smaller vector spaces V but with the same good numerical properties.*

Toric varieties arise naturally as the solution space of Laurent polynomial systems when the equations are presented in a monomial basis. In applications, the equations may arise as approximations of functions on a bounded real interval or they may come from previous numerical computations with real data. In these cases, it is well-known that it is better to work with, for instance, Chebyshev or Legendre bases instead of monomials [Tre19]. A natural question to ask is *what are the properties of varieties parametrized by Chebyshev polynomials and of families of polynomial systems with generic coefficients in a (tensor product) Chebyshev basis?*

The insight that the problem of diverging paths in homotopy continuation methods can be circumvented by tracking the paths in (multi-)projective space has led to great advances, see for instance [Wam93]. It can often prevent a lot of 'wasted' computation time and it can sometimes help us understand the affine solution count for a family of systems. The idea is to track a set of homogeneous coordinates of each solution by slicing the corresponding orbits in the total coordinate space with a generic linear space. We could ask to what extent this approach can be generalized to arbitrary toric varieties: can we *track paths in a complete toric variety X by tracking a representative in its total coordinate space?*

Appendix A

Commutative algebra

In this appendix we summarize some results and definitions from commutative algebra to support the material in this thesis. All of this information and much more can be found in the books [AM69, Rei95, Eis13, Rot10].

A.1 Rings and ideals

A.1.1 Elementary definitions

We will limit ourselves to a special type of rings, namely those for which multiplication is commutative and has a neutral element.

Definition A.1.1 (Commutative ring with identity)**.** A *commutative ring with identity* is a set R together with two binary operations '+' and '·', called *addition* and *multiplication*, such that R is closed under '+' and '·' and for all $f, g, h \in R$

1. $(R, +)$ is an abelian group: $(f + g) + h = f + (g + h)$, $f + g = g + f$, there is $0 \in R$ such that $f + 0 = f, \forall f \in R$ and for each $f \in R$ there is $-f \in R$ such that $f + (-f) = 0$,
2. $(fg)h = f(gh)$,
3. $f(g + h) = fg + fh$,
4. $fg = gf$,
5. there is $1 \in R$ such that $1f = f$, $\forall f \in R$,

where $f \cdot g$ is denoted by fg.

From now on, R is a commutative ring with identity.

Example A.1.1 (Fields)**.** A *field* is a commutative ring K with identity such that for each $f \in K \setminus \{0\}$ there exists $f^{-1} \in K$ satisfying $ff^{-1} = f^{-1}f = 1$. The simplest examples are $\mathbb{Q}, \mathbb{R}, \mathbb{C}$, finite fields $\mathbb{F}_q$ and the field of p-adic numbers $\mathbb{Q}_p$. $\triangle$

Example A.1.2 (Polynomial rings)**.** A very important example in the context of this thesis is the ring of polynomials in n variables $x_1, \ldots, x_n$ over a ring A. We will mostly consider the case where $A = \mathbb{C}$. This ring is denoted by $R = \mathbb{C}[x_1, \ldots, x_n]$ and its elements are of the form

$$f = \sum_{a \in \mathbb{N}^n} c_a x^a$$

with $c_a \in \mathbb{C}$, $x^a = x_1^{a_1} \cdots x_n^{a_n}$ for $a = (a_1, \ldots, a_n) \in \mathbb{N}^n$ and finitely many c_a are nonzero. Elements of the form x^a for some $a \in \mathbb{N}^n$ are called *monomials* of R. $\triangle$

Definition A.1.2 (Ring homomorphism)**.** Let R, R' be commutative rings with identity. A *ring homomorphism* is a map $\phi : R \to R'$ such that for any $f, g \in R$,

1. $\phi(f + g) = \phi(f) + \phi(g)$,
2. $\phi(fg) = \phi(f)\phi(g)$,
3. $\phi(1) = 1$.

The last condition of Definition A.1.2 is dropped when R does not have an identity element for '$\cdot$'.

Definition A.1.3 (A-algebra)**.** Let R and A be commutative rings with identity. R is an A-algebra if there is a ring homomorphism $\phi : A \to R$. If R and R' are A-algebras with homomorphisms $\phi : A \to R$ and $\phi' : A \to R'$, then a ring homomorphism $\psi : R \to R'$ is called an *A-algebra homomorphism* if it satisfies $\psi \circ \phi = \phi'$.

Note that if R and R' are A-algebras with homomorphisms $\phi : A \to R$ and $\phi' : A \to R'$, for an A-algebra homomorphism $\psi : R \to R'$ we have $\psi(\phi(a)f) = \phi'(a)\psi(f)$ for all $a \in A$ and all $f \in R$.

Example A.1.3 ($\mathbb{C}$-algebras)**.** The most important for us is the case where $A = \mathbb{C}$ and $\phi : \mathbb{C} \to R$ is the inclusion. An example is the polynomial ring $\mathbb{C}[x_1, \ldots, x_n]$. If R, R' are $\mathbb{C}$-algebras with respect to the inclusion of $\mathbb{C}$ in R, R', then $\mathbb{C}$-algebra homomorphisms $\psi : R \to R'$ are ring homomorphisms which are constant on $\mathbb{C}$: $\psi(cf) = \psi(c)\psi(f) = c\psi(f)$, $\forall c \in \mathbb{C}, f \in R$. If R is a $\mathbb{C}$-vector space which is also a ring, then R is a $\mathbb{C}$-algebra if scalar multiplication $\mathbb{C} \times R \to R$ is the restriction of multiplication $R \times R \to R$ to $\mathbb{C} \times R$. $\triangle$

For an A-algebra R with homomorphism $\phi : A \to R$ and $f_1, \ldots, f_s \in R$, we define

$$A[f_1, \ldots, f_s] = \{ \text{ finite sums } \sum_{a \in \mathbb{N}^a} \phi(c_a) f^a \mid c_a \in A\} \subset R,$$

where $f^a = f_1^{a_1} \cdots f_s^{a_s}$ for $a = (a_1, \ldots, a_s) \in \mathbb{N}^s$.

Definition A.1.4 (Finite generation)**.** An A-algebra R with homomorphism $\phi : A \to R$ is *finitely generated* (over A) if there is a finite set $\{f_1, \ldots, f_s\} \subset R$ such that $R = A[f_1, \ldots, f_s]$. In this case, the set $\{x_1, \ldots, x_s\}$ is called a set of A-generators of R.

Example A.1.4. The polynomial ring $R = \mathbb{C}[x_1, \ldots, x_n]$ is finitely generated as a $\mathbb{C}$-algebra: it is generated by the coordinate functions $\{x_1, \ldots, x_n\}$. △

Definition A.1.5 (Ideal)**.** A subset $I \subset R$ is called an *ideal* if

1. $0 \in I$,
2. for all $f, g \in I$, $f + g \in I$,
3. for all $g \in R$ and $f \in I$, $gf \in I$.

For any subset $P \subset R$, we denote $\langle P \rangle$ for the smallest ideal containing P.

Example A.1.5 (Sums, products, intersections, quotients of ideals)**.** If $I, J \subset R$ are ideals, then so are

1. $I + J = \{f + g \mid f \in I, g \in J\}$,
2. $IJ = \langle fg \mid f \in I, g \in J \rangle$,
3. $I \cap J$,
4. $(I : J) = \{f \in R \mid gf \in I \text{ for all } g \in J\}$.

△

Definition A.1.6 (Finitely generated ideals)**.** An ideal $I \subset R$ is called *finitely generated* if there are $f_1, \ldots, f_s \in R$ such that

$$I = \{g_1 f_1 + \cdots + g_s f_s \mid g_1, \ldots, g_s \in R\}.$$

In this case $\{f_1, \ldots, f_s\}$ is called a *set of generators* or a *basis* for the ideal I and we denote $I = \langle f_1, \ldots, f_s \rangle$.

Definition A.1.7 (Noetherian rings)**.** A ring R is called *Noetherian* if all its ideals $I \subset R$ are finitely generated.

Theorem A.1.1 (Hilbert's basis theorem)**.** *If a ring R is Noetherian, then so is the polynomial ring $R[x]$.*

Proof. See [AM69, Theorem 7.5]. □

Corollary A.1.1. *The polynomial ring $R = \mathbb{C}[x_1, \ldots, x_n]$ is Noetherian.*

Proof. The only ideals in $\mathbb{C}$ are $\{0\}$ and $\mathbb{C}$. These are generated by 0 and 1 respectively. The corollary follows by induction on n. □

Definition A.1.8 (Prime, maximal and radical ideals)**.** An ideal $I \subset R$ is called *prime* if $I \subsetneq R$ and $fg \in I$ implies that $f \in I$ or $g \in I$. It is called *maximal* if $I \subsetneq R$ and, when for another ideal $J \subset R$ we have $I \subsetneq J$, then $J = R$. An ideal $I \subset R$ is called *radical* if

$$I = \sqrt{I} = \{f \in R \mid f^m \in I \text{ for some } m \in \mathbb{N}\}.$$

The subset $\sqrt{I} \subset R$ is itself a radical ideal ($\sqrt{\sqrt{I}} = \sqrt{I}$) called the *radical* of I.

Another special type of ideals in R, called *primary ideals*, can be used to decompose ideals in a way similar to the decomposition of an integer as the product of powers of prime numbers.

Definition A.1.9. An ideal $I \subset R$ is called *primary* if for all $f, g \in R$, $fg \in I$ implies that either $f \in I$ or $g^m \in I$ for some $m \in \mathbb{N}$.

Theorem A.1.2 (Primary decomposition)**.** *Let R be Noetherian. For every ideal $I \subset R$ there exist primary ideals $Q_1, \dots, Q_s$ such that*

$$I = Q_1 \cap \cdots \cap Q_s.$$

Proof. See [AM69, Theorem 7.13] for the general statement or [CLO13, Chapter 4, §8, Theorem 4] for the case where R is a polynomial ring. □

The ideals of R are the subrings (in general, without identity) which play the role of normal subgroups in a group: they can be used to construct quotients.

A.1.2 Quotient rings

Definition A.1.10 (Quotient ring)**.** Let $I \subset R$ be an ideal. The *quotient ring* of R by I is the set

$$\{f + I \mid f \in R\}/\sim$$

modulo the equivalence relation $f + I \sim g + I \Leftrightarrow f - g \in I$, with operations

$$(f + I) + (g + I) = (f + g) + I, \quad (f + I)(g + I) = fg + I.$$

One can check that these operations are well defined and that the quotient ring R/I is indeed a commutative ring with identity element $1 + I$. Moreover, if $I = R = \langle 1 \rangle$ then $R/I = \{0\}$ and $1 = 0$, if $I = \langle 0 \rangle$ then $R/I = R$. Here's a definition for some special elements in a commutative ring with identity.

Definition A.1.11 (Units and nilpotents)**.** An element $f \in R$ is called a *unit* if there exists $g \in R$ such that $fg = 1$. It is called a *nilpotent element* or a *nilpotent* if $f^m = 0$ for some $m \in \mathbb{N}$.

Example A.1.6. If $R = \mathbb{Z}$ and $I = \langle 4 \rangle$ then $R/I = \mathbb{Z}/4\mathbb{Z}$ is the ring of integers modulo 4. In R/I, $2 + I$ is a nilpotent since $(2 + I)^2 = 0$ and $3 + I$ is a unit since $(3 + I)^2 = 1 + I = 1$. △

Example A.1.7. If $R = \mathbb{C}[x]$ and $I = \langle x \rangle$, then $R/I \simeq \mathbb{C}$ and $f + I \in R/I$ corresponds to $f(0) \in \mathbb{C}$. All elements $f + I, f \neq 0$ are units in R/I and there are no nonzero nilpotent elements. △

A commutative ring R is called an *integral domain* if for $f, g \in R$, $fg = 0$ implies $f = 0$ or $g = 0$ (in other words, R has no zero divisors). The ring R is called *nilpotent free* if it has no nonzero nilpotent elements. Some special ideals give rise to some special quotients.

Proposition A.1.1. *Let $I \subset R$ be a proper ideal. The quotient ring R/I is*

1. *nilpotent free if and only if I is radical,*
2. *an integral domain if and only if I is prime,*
3. *a field if and only if I is maximal.*

Proof. The quotient R/I is nilpotent free if $(f + I)^m = 0$ implies $f + I = 0 \Leftrightarrow f \in I$. This proves $\sqrt{I} \subset I$ and the reverse inclusion is obvious. The second statement follows from the fact that R/I is an integral domain if and only if $fg + I = 0$ implies $f + I = 0$ or $g + I = 0$ which is equivalent to $fg \in I \Rightarrow f \in I$ or $g \in I$. The third statement is Proposition 6.7 in [Rot10]. □

Corollary A.1.2. *Every maximal ideal is prime, and every prime ideal is radical.*

Proof. If I is maximal, then R/I is a field. In particular, it is an integral domain, so I is prime by Proposition A.1.1. If I is prime, R/I is a domain, and hence $f^k \in I$ implies $f \in I$ or $f^{k-1} \in I$. If $f \notin I$, then we must have $f^{k-2} \in I$, $f^{k-3} \in I, \ldots$ which leads to a contradiction. □

Two ideals $I, J \subset R$ are called *coprime* if $I + J = R$ (see Example A.1.5 for the definition of a sum of ideals). The following important theorem allows us to decompose a quotient ring R/I into 'simpler' quotient rings if $I = I_1 \cap \cdots \cap I_s$ where the ideals I_i are pairwise coprime.

Theorem A.1.3 (Chinese remainder theorem)**.** *Let $I_1, \ldots, I_s$ be ideals of R that are pairwise coprime and let $I = I_1 \cap \cdots \cap I_s$. Then we have*

$$R/I \simeq R/I_1 \times \cdots \times R/I_s$$

via the canonical ring homomorphism $f + I \mapsto (f + I_1, \ldots, f + I_s)$.

Proof. See [Lan02, Corollary 2.2, page 95]. □

A.1.3 Krull's principal ideal theorem

Definition A.1.12 (Height of a prime ideal)**.** The *height* of a prime ideal $\mathfrak{p} \subset R$, denoted $\mathrm{ht}(\mathfrak{p})$, is the supremum n of the lengths of all chains of prime ideals

$$\mathfrak{p}_0 \subsetneq \mathfrak{p}_1 \subsetneq \cdots \subsetneq \mathfrak{p}_n = \mathfrak{p} \subsetneq R.$$

Definition A.1.13 (Krull dimension)**.** The Krull dimension of R, denoted $\dim R$, is the supremum of the heights of all prime ideals of R.

Theorem A.1.4. *Let R be an integral domain which is a finitely generated $\mathbb{C}$-algebra. Then for any prime ideal $\mathfrak{p} \subset R$ we have*

$$\mathrm{ht}(\mathfrak{p}) + \dim R/\mathfrak{p} = \dim R.$$

Proof. See [Har77, Chapter I, Theorem 1.8A]. □

Another special class of ideals are those which can be generated by only one element. Such ideals are called *principal.*

Theorem A.1.5 (Krull's principal ideal theorem)**.** *Let $R = \mathbb{C}[x_1, \dots, x_n]$ and let $f \in R$ be a non-constant polynomial. Then for every minimal prime ideal $\mathfrak{p}$ containing the principal ideal $\langle f \rangle$ we have $\mathrm{ht}(\mathfrak{p}) = 1$.*

Proof. See [AM69, Corollary 11.7]. □

A nonzero element $f \in R$ is called *irreducible* if f is not a unit and $f = f_1 \cdots f_s$ implies that for all i, f_i is either a unit or $f = uf_i$ where u is a unit. A ring R is called a *unique factorization domain* if for all $f \in R \setminus \{0\}$ such that f is not a unit, f can be written 'essentially uniquely' as a product of irreducibles. For a precise definition the reader can consult [Rot10, Section 6.2]. The following is Proposition 1.12A in Chapter 1 of [Har77].

Proposition A.1.2. *A Noetherian integral domain R is a unique factorization domain if and only if every prime ideal $\mathfrak{p}$ such that $\mathrm{ht}(\mathfrak{p}) = 1$ is principal.*

As an application of these results, we can show that affine hypersurfaces have the special property that they can always be defined by only one equation. This uses some notation from Section 2.1.

Theorem A.1.6. *An affine variety $Y \subset \mathbb{C}^n$ is pure-dimensional of dimension $n-1$ if and only if $Y = V(f)$ for some $f \in R \setminus \{0\}$.*

Proof. Suppose $Y = V(f)$. Let $f = f_1 \cdots f_s$ be a decomposition of f into non constant irreducible polynomials. Then $Y = V(f_1) \cup \cdots \cup V(f_s)$ is a decomposition of Y into irreducible components. Each of these components has codimension one, since by Krull's principal ideal theorem A.1.5, the ideal $\langle f_i \rangle$ has height 1 and therefore $\dim V(f_i) = \dim R/\langle f_i \rangle = n-1$ (see Theorem A.1.4). Conversely, if Y is pure-dimensional of dimension $n-1$, then all its irreducible components $Y_1, \ldots, Y_s$ have dimension $n-1$, and their vanishing ideals $I(Y_i)$ have height 1 by Theorem A.1.4. Since R is a unique factorization domain, these vanishing ideals are principal by Proposition A.1.2, so $I(Y_i) = \langle f_i \rangle$, where f_i is irreducible. It follows that $Y = V(f_1 \cdots f_s)$. □

A.1.4 Localization

The way in which the field of rational numbers $\mathbb{Q}$ is constructed from the integers $\mathbb{Z}$ can be generalized straightforwardly to arbitrary integral domains.

Definition A.1.14 (Field of fractions)**.** Let R be an integral domain with identity. The *field of fractions* $K(R)$ of R is

$$\{f/g \mid f \in R, g \in R \setminus \{0\}\}/\sim$$

where $f_1/g_1 \sim f_2/g_2 \Leftrightarrow f_1g_2 - f_2g_1 = 0$, with operations

$$f_1/g_1 + f_2/g_2 = (f_1g_2 + f_2g_1)/(g_1g_2), \quad (f_1/g_1)(f_2/g_2) = (f_1f_2)/(g_1g_2).$$

One checks that $K(R)$ is indeed a field with zero element $0 = 0/1$ and identity element $1 = 1/1$. Note that the operations are not well defined if R is not an integral domain, because $R \setminus \{0\}$ is not closed under multiplication. Also, to check that the relation $\sim$ in Definition A.1.14 is transitive, we need the property that R has no zero divisors. However, the construction can be generalized to arbitrary commutative rings with identity by slightly modifying the definition of the equivalence relation and the set of possible 'denominators'.

Definition A.1.15 (Localization)**.** Let $T \subset R$ be a multiplicatively closed subset of R, that is, $1 \in T$ and T is closed under multiplication. The *localization* $T^{-1}R$ of R at T is

$$\{f/g \mid f \in R, g \in T\}/\sim$$

where $f_1/g_1 \sim f_2/g_2 \Leftrightarrow t(f_1g_2 - f_2g_1) = 0$ for some $t \in T$, with operations

$$f_1/g_1 + f_2/g_2 = (f_1g_2 + f_2g_1)/(g_1g_2), \quad (f_1/g_1)(f_2/g_2) = (f_1f_2)/(g_1g_2),$$

where $f_1, f_2 \in R$, $g_1, g_2 \in T$.

It is a standard exercise in commutative algebra to check that $\sim$ from Definition A.1.15 is indeed an equivalence relation and that the operations from the definition are well defined and give $T^{-1}R$ the structure of a commutative ring with identity. Note that there is a natural homomorphism

$$R \to T^{-1}R : f \mapsto f/1,$$

which is injective when R is an integral domain. Here are some important examples of localization.

Example A.1.8. Let R be an integral domain and $T = R \setminus \{0\}$, then $T^{-1}R = K(R)$ is the field of fractions of R. △

Example A.1.9. Let $f \in R \setminus \{0\}$ and $T = \{f^\ell\}_{\ell \in \mathbb{N}}$. Then $T^{-1}R$ is denoted by R_f:

$$R_f = \left\{ \frac{g}{f^\ell} \mid \ell \in \mathbb{N}, g \in R \right\} \Big/ \sim .$$

The ring R_f is called the *localization of R at f*. △

Example A.1.10. Let $\mathfrak{p} \subset R$ be a prime ideal. The set $T = R \setminus \mathfrak{p}$ is multiplicatively closed. The localization $T^{-1}R$ is denoted by $R_\mathfrak{p}$:

$$R_\mathfrak{p} = \left\{ \frac{f}{g} \mid g \in R \setminus \mathfrak{p}, f \in R \right\} \Big/ \sim .$$

The unique maximal ideal of the ring $R_\mathfrak{p}$ is the image of $\mathfrak{p}$ under $R \to R_\mathfrak{p}$. The ring $R_\mathfrak{p}$ is called the *localization of R at $\mathfrak{p}$*. △

Definition A.1.16 (Extension and contraction)**.** The *extension* $I^e \subset T^{-1}R$ of an ideal $I \subset R$ in the localization $T^{-1}R$ of R at T is the ideal generated by the image of I under $R \to T^{-1}R$. The *contraction* $I^c \subset R$ of an ideal $I \subset T^{-1}R$ is the preimage of I under $R \to T^{-1}R$. That is, I^c is the largest ideal of R whose image under $R \to T^{-1}R$ is contained in I.

A.2 Modules over rings

Throughout this section, R is a commutative ring with identity. Some of the material presented here is taken from [CLO06, Chapter 6], which contains a more complete introduction to R-modules and related subjects from a computational perspective.

A.2.1 Elementary definitions

Definition A.2.1 (R-module)**.** A *module over R* or *R-module* is a set M together with a binary operation (addition) under which it is an abelian group and an operation $R \times M \to M$, written $(f, m) \mapsto fm$, $f \in R, m \in M$, of R on M (scalar multiplication), satisfying for all $f, g \in R$, $m, m' \in M$:

1. $f(m+m') = fm + gm'$,
2. $(f+g)m = fm + gm$,
3. $(fg)m = f(gm)$,
4. $1m = m$, with 1 the identity element of R.

Here are some examples.

Example A.2.1. Abelian groups are $\mathbb{Z}$-modules. If K is a field, then K-modules are the vector spaces over K. Modules are to a commutative ring with identity what vector spaces are to a field. △

Example A.2.2. Perhaps the simplest example of a module over R is the set of s-vectors of elements in R with the usual addition and scalar multiplication. We denote this set by R^s. In particular, R itself is an R-module ($s = 1$). It is also not difficult to see that any finite subset $\{m_1, \ldots, m_\ell\} \subset R^s$ gives an R-module

$$R\{m_1, \ldots, m_\ell\} = \langle m_1, \ldots, m_\ell \rangle = \{f_1 m_1 + \ldots + f_\ell m_\ell \in R^s \mid f_1, \ldots f_\ell \in R\}.$$

If $M = \langle m_1, \ldots m_\ell \rangle$, we say that M is *generated* by $\{m_1, \ldots, m_\ell\}$. △

Example A.2.3. Let $R = \mathbb{C}[x_1, \ldots, x_n]$, any polynomial ideal $I \subset R$ is a module over R. △

Example A.2.4. Different algebraic structures lead to different notions of 'generators'. The ring $R = \mathbb{C}[x_1, \ldots, x_n]$ is generated, as an R-module (and as an ideal), by $\{1\}$. As a $\mathbb{C}$-algebra, it is generated by $\{x_1, \ldots, x_n\}$. As a $\mathbb{C}$-module (i.e. as a $\mathbb{C}$-vector space), it is infinitely generated. △

Example A.2.5. Let A be an $m \times n$ matrix with entries in R. It is easy to show that the set

$$\ker A = \{m \in R^n : Am = 0\}$$

is a module over R. Also, the set

$$\operatorname{im} A = \{Am' : m' \in R^n\}$$

is a module, given by $R\langle m_1, \ldots, m_n \rangle$ where m_i is the i-th column of A. △

Example A.2.6 (Direct sum of modules)**.** The direct sum $M \oplus N$ of two R-modules M and N is the set of all ordered pairs (m, n), $m \in M$ and $n \in N$. Such a direct sum $M \oplus N$ is an R-module under component-wise sum and scalar multiplication. We can think of R^m as the direct sum $R \oplus \ldots \oplus R$ with m summands equal to R. △

Example A.2.7 (Quotient module)**.** If $N \subset M$ is a submodule, then the quotient

$$M/N = \{m + N \mid m \in M\}/\sim$$

where $m + N \sim m' + N$ if $m - m' \in N$ is an R-module with $R \times M/N \to M/N$ given by $(f, m+N) \mapsto fm + N$. △

Definition A.2.2 (R-linear independence)**.** Let M be an R-module. A set $\{m_1, \ldots, m_\ell\} \subset M$ is called *R-linearly independent* if $f_1 m_1 + \ldots + f_\ell m_\ell = 0$, $f_1, \ldots, f_\ell \in R$ implies $f_1 = \ldots = f_\ell = 0$.

Unlike vector spaces, modules may have minimal generating sets that are not R-linearly independent.

Example A.2.8. An easy example is the ideal $\langle f, g\rangle \subset \mathbb{C}[x, y]$ where f does not divide g and vice versa. Indeed, the R-linear combination $gf - fg = 0$ shows that the set of generators is R-linearly dependent, yet f nor g can be left out without shrinking the ideal. △

In analogy with the theory of vector spaces, we use the following notion of a *basis*.

Definition A.2.3. A subset $F \subset M$ of an R-module M is called a *module basis* (or simply *basis*) of M if F generates M and F is an R-linearly independent set.

A set of generators for an ideal is also referred to as a *basis* (see Definition A.1.6). In the previous example, this means that $\{f, g\}$ is a basis for I as an ideal of R, but not as an R-module. The example showed that not every minimal set of generators is a module basis. Sadly, even more is true. Many modules do not admit a module basis. The ideal $\langle f, g\rangle$ from before is an example.

Definition A.2.4 (Free module)**.** An R-module M that admits a module basis is called a *free module*.

Example A.2.9. The module R^s is free for any $s \geq 1$ and its standard basis is given by $\{e_1, \ldots, e_s\}$, where $e_i \in R^s$ has the zero element in all but the i-th entry, which is 1. However, not every submodule of R^m is free. For instance, consider the module $\langle (f, 0), (g, 0)\rangle \subset \mathbb{C}[x, y]^2$. △

The following theorem implies, together with Hilbert's basis theorem (Theorem A.1.1) that if $R = \mathbb{C}[x_1, \ldots, x_n]$, every submodule $M \subset R^s$, $s \geq 1$ is finitely generated.

Theorem A.2.1. *A commutative ring R is Noetherian if and only if every submodule of a finitely generated R-module is finitely generated.*

Proof. See [Rot10, Proposition 7.23]. □

Proposition A.2.1. *For an R-module M, a set $F \subset M$ is a module basis if and only if every $m \in M$ can be written in exactly one way as an R-linear combination of the elements in F.*

Proof. Let $F = \{m_1, \ldots, m_\ell\}$ and suppose

$$m = f_1 m_1 + \cdots + f_\ell m_\ell = f_1' m_1 + \cdots + f_\ell' m_\ell$$

for some $f_i, f_i' \in R$. Then $(f_1 - f_1')m_1 + \cdots + (f_\ell - f_\ell')m_\ell = 0$ implies $f_i = f_i'$, $i = 1, \ldots, \ell$ since F is a basis. □

We now define structure preserving maps between R-modules.

Definition A.2.5 (R-module homomorphism)**.** An *R-module homomorphism* between two R-modules M and N is an R-linear map between M and N. This means that for a homomorphism $\phi : M \to N$ we have for all $f \in R$ and for all $m, m' \in M$ that

$$\phi(fm + m') = f\phi(m) + \phi(m').$$

Example A.2.10 (Modules of homomorphisms)**.** Let M, N be R-modules. The set of all R-module homomorphisms $M \to N$ is denoted $\mathrm{Hom}_R(M, N)$. This is itself an R-module: for $\phi, \phi' \in \mathrm{Hom}_R(M, N)$, $f \in R$,

$$(f\phi)(m) = f\phi(m), \quad (\phi + \phi')(m) = \phi(m) + \phi'(m), \quad , \forall m \in M.$$

Here are some examples.

- $\mathrm{Hom}_R(R^n, R) \simeq R^n$ where $\phi : R^n \to R$ corresponds to $(\phi(e_1), \ldots, \phi(e_n)) \in R^n$.
- $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}^n, \mathbb{C}^*) \simeq (\mathbb{C}^*)^n$ where $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$ is thought of as a multiplicative abelian group. Here $\phi : \mathbb{Z}^n \to \mathbb{C}^*$ corresponds to $(\phi(e_1), \ldots, \phi(e_n))$.
- $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}/n\mathbb{Z}, \mathbb{C}^*) \simeq \{\exp(\frac{2\pi\sqrt{-1}k}{n})\}_{k=0,\ldots,n-1}$.

$\triangle$

Example A.2.11 (Tensor product of modules)**.** Let M, M', N be R-modules. A mapping $\psi : M \times M' \to N$ is called *bilinear* if for each $m \in M$, $m' \mapsto \psi(m, m')$ is an R-module homomorphism and for each $m' \in M'$, $m \mapsto \psi(m, m')$ is an R-module homomorphism. There exists a module $M \otimes_R M'$, called the *tensor product* of M and M', and a bilinear mapping $\otimes : M \times M' \to M \otimes_R M'$, unique up to isomorphism, satisfying the following *universal property*. For each R-module N and each bilinear mapping $\psi : M \times M' \to N$ there is a unique R-module homomorphism $\theta : M \otimes_R M' \to N$ which makes the following diagram commute.

$$\begin{tikzcd} M \times M' \arrow[r, "\otimes"] \arrow[dr, "\psi"'] & M \otimes_R M' \arrow[d, "\theta"] \\ & N \end{tikzcd}$$

The module $M \otimes_R M'$ is generated as an R-module by elements $m \otimes m' = \otimes(m, m'), m \in M, m' \in M'$. These elements are called the *elementary tensors* of $M \otimes_R M'$. Here are some examples.

- $\mathbb{C}[x] \otimes_{\mathbb{C}} \mathbb{C}[y] \simeq \mathbb{C}[x, y]$ with $\otimes(f(x), g(y)) = f(x)g(y)$ and $\theta(x^\ell y^m) = \psi(x^\ell, y^m)$.
- $\mathbb{C}[x] \otimes_{\mathbb{C}[x]} \mathbb{C}[x] \simeq \mathbb{C}[x]$.

- $R \otimes_R M \simeq M$ for every R-module M with $\otimes$ the map defining the R-module structure and $\theta(m) = \psi(1, m)$ for all $m \in M$.
- $\mathbb{Z}^n \otimes_{\mathbb{Z}} \mathbb{C}^* = (\mathbb{C}^*)^n$ where $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$ is thought of as a multiplicative abelian group, with $a \otimes c = (a_1, \ldots, a_n) \otimes c = (c^{a_1}, \ldots, c^{a_n})$ and $\theta(c_1, \ldots, c_n) = \psi(e_1, c_1) + \cdots + \psi(e_n, c_n)$.
- $\mathbb{Z}^n \otimes_{\mathbb{Z}} \mathbb{R} = \mathbb{R}^n$ where $\mathbb{R}$ is thought of as an abelian group under element-wise addition, with $a \otimes r = (a_1, \ldots, a_n) \otimes r = (ra_1, \ldots, ra_n)$ and $\theta(r_1, \ldots, r_n) = \psi(e_1, r_1) + \cdots + \psi(e_n, r_n)$.

$\triangle$

If M is free, any R-module homomorphism $\phi : M \to N$ is specified completely by the image of the basis elements. This follows directly from Proposition A.2.1. If also N is free, the image can be represented in a unique way as an R-linear combination of basis elements. The following proposition follows easily.

Proposition A.2.2. *Let $\phi : R^n \to R^m$ be any R-module homomorphism. There exists an $m \times n$ matrix A with entries in R such that $\phi(m) = Am, \forall m \in R^n$. Conversely, any such matrix defines an R-module homomorphism $\phi : R^n \to R^m$.*

For any R-module homomorphism $\phi : M \to N$ the kernel $\ker \phi$ and the image $\operatorname{im} \phi$ are defined in the usual way and ϕ is called an isomorphism if it is both one-to-one and onto. It is easy to check that both $\ker \phi$ and $\operatorname{im} \phi$ are R-modules.

Proposition A.2.3. *Consider an ordered s-tuple $(m_1, \ldots, m_s)$ of elements $m_i \in M$ of an R-module M. The set of all $(f_1, \ldots, f_s) \in R^s$ such that $f_1 m_1 + \ldots f_s m_s = 0$ is an R-submodule of R^s. This module is called the* first syzygy module *of $(m_1, \ldots, m_s)$ and it is denoted by $\operatorname{Syz}(m_1, \ldots, m_s)$.*

Proof. This follows from $\operatorname{Syz}(f_1, \ldots, f_s) = \ker(\phi : R^s \to R)$ with $\phi(f_1, \ldots, f_s) = f_1 m_1 + \ldots f_s m_s$. $\square$

A.2.2 Exact sequences

Definition A.2.6 (Exact sequence)**.** A sequence of R-modules and homomorphisms

$$\cdots \longrightarrow M_{i+1} \xrightarrow{\phi_{i+1}} M_i \xrightarrow{\phi_i} M_{i-1} \longrightarrow \cdots$$

is called *exact* at M_i if $\operatorname{im} \phi_{i+1} = \ker \phi_i$. The entire sequence is an *exact sequence* if it is exact at each M_i which is not at the beginning nor at the end of the sequence.

Let M and N be two R-modules. It follows directly from the definition of exactness that a homomorphism $\phi : M \to N$ is onto if and only if the sequence

$$M \xrightarrow{\phi} N \longrightarrow 0$$

is exact. Analogously, $\phi : M \to N$ is one-to-one if and only if

$$0 \longrightarrow M \xrightarrow{\phi} N$$

is exact and ϕ is an isomorphism if and only if

$$0 \longrightarrow M \xrightarrow{\phi} N \longrightarrow 0$$

is exact. A frequently encountered type of exact sequence involves only three nonzero modules.

Definition A.2.7 (Short exact sequence)**.** A *short exact sequence* of R-modules is an exact sequence of the form

$$0 \longrightarrow M' \xrightarrow{\phi} M \xrightarrow{\psi} M'' \longrightarrow 0.$$

The following theorem, together with the discussion above, shows that exact sequences provide a very compact way of writing down properties of modules and homomorphisms between them.

Theorem A.2.2 (First isomorphism theorem)**.** *Let M, N be R-modules and let $\phi : M \to N$ be an R-module homomorphism. Then $M/\ker\phi \to \operatorname{im}\phi$ given by $m + \ker\phi \mapsto \phi(m)$ is an R-module isomorphism.*

Example A.2.12. If $I \subset R$ is an ideal of R and

$$0 \longrightarrow I \xrightarrow{i} R \xrightarrow{\phi} M \longrightarrow 0$$

is a short exact sequence of R-modules where $i : I \to R$ is inclusion, then $M \simeq R/I$. $\triangle$

The following theorem is frequently used in this thesis.

Theorem A.2.3. *Let $V_i, 0 \leq i \leq \ell$ be finite dimensional vector spaces over a field K and let*

$$0 \longrightarrow V_\ell \xrightarrow{\phi_\ell} V_{\ell-1} \xrightarrow{\phi_{\ell-1}} \cdots \xrightarrow{\phi_2} V_1 \xrightarrow{\phi_1} V_0 \longrightarrow 0$$

be an exact sequence of K-vector spaces. The alternating sum of the dimensions of the V_i satisfies:

$$\sum_{i=0}^{\ell} (-1)^i \dim_K(V_i) = 0,$$

where $\dim_K(\cdot)$ *denotes the dimension as a K-vector space.*

Proof. To prove this, we only need that for a linear map $\phi : V \to W$ between finite dimensional vector spaces it holds that

$$\dim_K(V) = \dim_K(\ker\phi) + \dim_K(\operatorname{im}\phi).$$

Applying this to an exact sequence of vector spaces, using $\ker\phi_i = \operatorname{im}\phi_{i+1}$ the theorem follows. ☐

A.2.3 Free resolutions

If an R-module M is finitely generated and given by $\langle m_1, \dots, m_s\rangle$, we have an onto map $\phi_0 : R^s \to M$ given by $(f_1, \dots, f_s) \mapsto f_1 m_1 + \dots f_s m_s$ and a corresponding exact sequence

$$R^s \xrightarrow{\phi_0} M \longrightarrow 0.$$

Suppose R is Noetherian, e.g. $R = \mathbb{C}[x_1, \dots, x_n]$. Since $\operatorname{Syz}(m_1, \dots, m_s)$ is a submodule of R^s, it is finitely generated (Theorem A.2.1). It follows that $\operatorname{Syz}(m_1, \dots, m_s)$ can be generated by $\{m'_1, \dots, m'_{s'}\}$. This gives $\phi_1 : R^{s'} \to R^s$ with $\phi_1(f_1, \dots, f_{s'}) = f_1 m'_1 + \dots + f_{s'} m'_{s'}$. The image is given by $\operatorname{im}\phi_1 = \operatorname{Syz}(m_1, \dots, m_s)$. Our exact sequence extends to

$$R^{s'} \xrightarrow{\phi_1} R^s \xrightarrow{\phi_0} M \longrightarrow 0.$$

Next, we can consider the syzygy module $\operatorname{Syz}(m'_1, \dots, m'_{s'})$. This is called the *second syzygy module*, and it will again by finitely generated. One can imagine that this process can be continued. It gives rise to a *free resolution* of the module M.

Definition A.2.8 (Free resolution)**.** Let M be an R-module. An exact sequence of the form

$$\cdots \longrightarrow F_2 \xrightarrow{\phi_2} F_1 \xrightarrow{\phi_1} F_0 \xrightarrow{\phi_0} M \xrightarrow{0}$$

where $F_i \simeq R^{s_i}, i = 0, 1, \dots$ are free R-modules is called a *free resolution* of M. A free resolution for which $F_{\ell+1} = F_{\ell+2} = \dots = 0$ for some $\ell \geq 0$ is called *finite* of length ℓ. In that case we write it down as

$$0 \longrightarrow F_\ell \xrightarrow{\phi_\ell} \cdots \longrightarrow F_2 \xrightarrow{\phi_2} F_1 \xrightarrow{\phi_1} F_0 \xrightarrow{\phi_0} M \longrightarrow 0$$

Note that in this notation, $\ker\phi_0$ is the first syzygy module for some choice of generators for M, $\ker\phi_1$ is the syzygy module of the first syzygy module, ... We say that $\ker\phi_i$ is the $(i+1)$-st syzygy module of M. It turns out that in the case that is most interesting to us, a finite free resolution always exists. The following result is due to Hilbert.

Theorem A.2.4 (Hilbert Syzygy Theorem)**.** *Let $R = K[x_1, \dots, x_n]$ where K is a field. Every finitely generated R-module has a finite free resolution of length at most n.*

Proof. A proof based on Groebner bases for modules is given in [CLO06, Chapter 6, §2]. ☐

A.2.4 Graded rings, modules and resolutions

The polynomial rings in this thesis often come with a *grading*. The definition of the grading is important for the geometric context: different gradings on the same ring associate the ring to completely different geometric objects. In this section, S is a $\mathbb{C}$-algebra with respect to the inclusion $\mathbb{C} \subset S$.

Definition A.2.9 (Graded $\mathbb{C}$-algebras)**.** Let E be an abelian group. An E-graded $\mathbb{C}$-algebra is a $\mathbb{C}$-algebra S with direct sum decomposition

$$S = \bigoplus_{\alpha \in E} S_\alpha$$

into $\mathbb{C}$-vector spaces $S_\alpha \subset S$ such that $S_\alpha \cdot S_{\alpha'} \subset S_{\alpha+\alpha'}$ (meaning that for any $f \in S_\alpha, g \in S_{\alpha'}$, $fg \in S_{\alpha+\alpha'}$). The $\mathbb{C}$-vector spaces S_α are called the *graded or homogeneous parts* of S. An element $f \in S_\alpha$ is called *homogeneous of degree* α. We denote $\deg(f) = \alpha$.

Note that Definition A.2.9 also makes sense in the case where E is just a monoid. When we work over a graded ring we will switch notation from R to S to emphasize this.

Remark A.2.1. Since S is a commutative ring with identity element $1 \in \mathbb{C} \subset S$ and $1 \cdot f = f, \forall f \in S$, we must have $\deg(1) = 0 \in E$. Moreover, since the S_α are $\mathbb{C}$-vector spaces we have $\mathbb{C} \subset S_0$. Also, S_0 is a commutative ring with identity and each of the S_α is an S_0-module. △

Example A.2.13. Let $S = \mathbb{C}[x_0, x_1, x_2, x_3]$. We make S into a $\mathbb{Z}$-graded ring by setting $\deg(x_0) = \deg(x_1) = \deg(x_2) = \deg(x_3) = 1$. This is the standard grading where, for instance, the polynomial $f = x_0^2x_2^2 - x_0x_1x_2x_3$ has degree 4. Here $S_0 = \mathbb{C}$. Note that $S_\alpha = 0$ for $\alpha < 0$. The submonoid $\{\alpha \in E \mid S_\alpha \neq 0\}$ is called the *weight monoid* of S. Another way to make S into a $\mathbb{Z}$ graded ring is to set $\deg(x_0) = \deg(x_1) = \deg(x_2) = 1, \deg(x_3) = 2$. With respect to this grading, $f = x_0^2x_2^2 - x_0x_1x_2x_3$ is *not* homogeneous and $g = x_0x_3 - x_0x_1x_2$ is homogeneous of degree 3. We now make S into a $\mathbb{Z}^2$-graded ring by setting $\deg(x_0) = \deg(x_1) = (1, 0)$ and $\deg(x_2) = \deg(x_3) = (0, 1)$. In this grading, f is homogeneous with degree $\deg(f) = (2, 2)$. △

Definition A.2.10 (Homogeneous ideal)**.** Let S be an E-graded $\mathbb{C}$-algebra. A *homogeneous ideal* of S is an ideal I that can be generated by homogeneous elements. That is, $I = \langle f_1, \dots, f_s \rangle$ with $f_i \in S_{\alpha_i}$ for some $\alpha_i \in E$.

Definition A.2.11 (Graded S-modules)**.** Let E be an abelian group and let S be an E-graded $\mathbb{C}$-algebra. An S-module M is called *graded* if it has a decomposition

$$M = \bigoplus_{\alpha \in E} M_\alpha$$

into $\mathbb{C}$-vector spaces $M_\alpha \subset M$ such that $S_\alpha \cdot M_{\alpha'} \subset M_{\alpha+\alpha'}$ (meaning that for any $f \in S_\alpha, m \in M_{\alpha'}$, $fm \in M_{\alpha+\alpha'}$). The $\mathbb{C}$-vector spaces M_α are called the *graded or homogeneous parts* of M. An element $m \in M_\alpha$ is called *homogeneous of degree* α. We denote $\deg(m) = \alpha$.

Note that the group E is not explicitly mentioned when we say that an S-module is graded. The reason is that it is implicit from the grading on S. We will sometimes say that S is *graded*, rather than E-graded, when the group E is clear from the context or not important.

Example A.2.14. The ring S is a graded S-module. Every homogeneous ideal $I \subset S$ is a graded S-module with $I_\alpha = I \cap S_\alpha$. A free S-module is a graded S-module with $(S^s)_\alpha = (S_\alpha)^s$. For a homogeneous ideal $I \subset S$, the quotient ring S/I is a graded S-module with $(S/I)_\alpha = S_\alpha / I_\alpha$ as a quotient of $\mathbb{C}$-vector spaces. △

Example A.2.15 (Twisted modules)**.** Let M be a graded S-module. For $\alpha' \in E$, consider the module

$$M(\alpha') = \bigoplus_{\alpha \in E} M(\alpha')_\alpha = \bigoplus_{\alpha \in E} M_{\alpha+\alpha'}.$$

This is a graded S-module, which is said to be the module M with grading *twisted* by α'. △

Example A.2.16. Let M and M' be graded S-modules. The direct sum $M \oplus M'$ is a graded S-module with $(M \oplus M')_\alpha = M_\alpha \oplus M_\alpha$ as a direct sum of $\mathbb{C}$-vector spaces. If $M' \subset M$ is a submodule, then the quotient module M/M' is a graded S-module with $(M/M')_\alpha = M_\alpha / M'_\alpha$ as a quotient of vector spaces. △

Definition A.2.12 (Twisted free graded S-modules)**.** A *twisted free graded S-module* is an S-module of the form

$$S(\alpha_1) \oplus \cdots \oplus S(\alpha_s), \quad \alpha_i \in E,$$

where $S(\alpha_i)$ is S with grading twisted by $\alpha_i \in E$.

Definition A.2.13 (Graded homomorphism)**.** Let M, M' be graded S-modules and let $\phi : M \to M'$ be a module homomorphism. The homomorphism ϕ is called *graded* of degree α if $\phi(M_{\alpha'}) \subset M'_{\alpha+\alpha'}$ for all $\alpha' \in E$.

Example A.2.17. The degree of a graded morphism $\phi : M \to M'$ can be 'changed' by twisting the degree of, say, M. For instance, the homomorphism $\phi : S \to S$ given by $g \mapsto fg$ for some $f \in S_\alpha, f \neq 0$ is graded of degree α. The homomorphism $\phi' : S(-\alpha) \to S$ given by $g \mapsto fg$ for some $f \in S_\alpha$ has degree zero. △

We will mainly be interested in graded homomorphisms of degree 0. The reason is the following. Suppose $F_0, \ldots, F_\ell$ are graded S-modules such that for each $\alpha \in E$,

$\dim_{\mathbb{C}}(F_i)_\alpha$ is easy to compute for $i = 0, \dots, \ell$. Moreover, suppose that M is some other graded S-module for which we want to compute $\dim_{\mathbb{C}} M_\alpha$. If

$$0 \longrightarrow F_\ell \xrightarrow{\phi_\ell} \cdots \longrightarrow F_2 \xrightarrow{\phi_2} F_1 \xrightarrow{\phi_1} F_0 \xrightarrow{\phi_0} M \longrightarrow 0 \tag{A.2.1}$$

is an exact sequence and the ϕ_i are homomorphisms of degree 0, we can restrict the sequence to the degree α part to obtain an exact sequence of vector spaces

$$0 \longrightarrow (F_\ell)_\alpha \xrightarrow{\phi_\ell} \cdots \longrightarrow (F_2)_\alpha \xrightarrow{\phi_2} (F_1)_\alpha \xrightarrow{\phi_1} (F_0)_\alpha \xrightarrow{\phi_0} M_\alpha \longrightarrow 0.$$

Theorem A.2.3 now allows us to compute $\dim_{\mathbb{C}} M_\alpha$. This raises the question which graded S-modules F are such that F_α is easy to compute. In our context, these will be exactly the twisted free graded S-modules from Definition A.2.12. Motivated by this, we will give exact sequences of the form (A.2.1) where the F_i are twisted free graded S-modules a name.

Definition A.2.14 (Graded resolution)**.** Let M be a graded S-module. A *graded resolution* of M is an exact sequence of the form

$$\cdots \longrightarrow F_1 \xrightarrow{\phi_1} F_0 \xrightarrow{\phi_0} M \longrightarrow 0,$$

where each F_i is a twisted free graded S-module and each of the ϕ_i is a graded homomorphism of degree 0. If $F_{\ell+1} = F_{\ell+2} = \cdots = 0$ for some $\ell \geq 0$ the resolution is called *finite* of length ℓ.

Again, in the cases which are of interest to us, a finite graded resolution always exists. Here is a graded version of Theorem A.2.4.

Theorem A.2.5 (Graded Hilbert Syzygy Theorem)**.** *Let $S = K[x_1, \dots, x_n]$ be a $\mathbb{Z}$-graded K-algebra where K is a field. Every finitely generated S-module has a finite graded resolution of length at most n.*

Proof. See [CLO06, Chapter 6, §3]. □

The graded resolutions in this text all arise from a so-called *Koszul complex.* This example is important enough to dedicate a separate subsection to it.

A.2.5 The Koszul complex

Definition A.2.15 (Complex of R-modules)**.** A sequence $\mathcal{K}$ of R-modules and homomorphisms

$$\mathcal{K}: \quad \cdots \longrightarrow M_{i+1} \xrightarrow{\phi_{i+1}} M_i \xrightarrow{\phi_i} M_{i-1} \longrightarrow \cdots$$

is called a *complex* or *chain complex* of R-modules if $\phi_i \circ \phi_{i+1} = 0, \forall i$.

Note that an exact sequence of R-modules is always a complex, but the converse statement is not true.

Example A.2.18. Let $I = \langle f_1, f_2 \rangle \subset R$ for some $f_1, f_2 \neq 0$. The map $d_1 : R^2 \to R$ defined by $d_1(g_1, g_2) = g_1 f_1 + g_2 f_2$ has image I. An obvious element in $\ker d_1$ is $(-f_2, f_1)$. Consider the map $d_2 : R \to R^2$ given by $g \mapsto (-gf_2, gf_1)$. This gives the complex

$$0 \longrightarrow R \xrightarrow{d_2} R^2 \xrightarrow{d_1} R \longrightarrow 0.$$

In fact, this is our first example of a so-called *Koszul complex* (we will give a definition below). The maps of this complex can be represented by matrices with entries in R:

$$d_1 = \begin{bmatrix} f_1 & f_2 \end{bmatrix}, \quad d_2 = \begin{bmatrix} -f_2 \\ f_1 \end{bmatrix} \quad \text{and} \quad d_1 \circ d_2 = 0.$$

It is easy to see that if $I \neq R$, the complex is not an exact sequence: d_1 is not onto. However, this can be remedied by extending the complex to

$$0 \longrightarrow R \xrightarrow{d_2} R^2 \xrightarrow{d_1} R \longrightarrow R/I \longrightarrow 0 \tag{A.2.2}$$

where $R \to R/I$ is the canonical map $f \mapsto f + I$. If R is an integral domain, exactness at every module of the complex but R^2 is clear:

- $\ker d_2 = 0$ since R is an integral domain,
- $\operatorname{im} d_1 = I = \ker(R \to R/I)$,
- $\operatorname{im}(R \to R/I) = R/I$.

Exactness at R^2 may fail: if there is a non-unit $g \in R \setminus \{0\}$ such that $f_1 = gf_1'$, $f_2 = gf_2'$, then $(-f_2', f_1') \in \ker d_1 \setminus \operatorname{im} d_2$. We will soon describe a sufficient condition on f_1, f_2 such that (A.2.2) is exact. Note that when this happens, (A.2.2) is a free resolution of R/I which was very easy to construct. In particular, in this case $\operatorname{im} d_2 = \operatorname{Syz}(f_1, f_2)$ is generated by $(-f_2, f_1)$. We say that $\operatorname{Syz}(f_1, f_2)$ consists only of *trivial syzygies.* △

It is instructive to repeat the same construction for an ideal generated by three elements.

Example A.2.19. Let $I = \langle f_1, f_2, f_3 \rangle \subset R$ with $f_1, f_2, f_3 \in R \setminus \{0\}$. Starting from the map $d_1 : R^3 \to R$ given by the matrix $[f_1 \ f_2 \ f_3]$ we will construct a 'candidate complex' for a free resolution of the R-module R/I. Some trivial elements of $\ker d_1 = \operatorname{Syz}(f_1, f_2, f_3)$ are $(-f_2, f_1, 0), (-f_3, 0, f_1)$ and $(0, -f_3, f_2)$. We define $d_2 : R^3 \to R^3$ by

$$(g_1, g_2, g_3) \mapsto g_1(-f_2, f_1, 0) + g_2(-f_3, 0, f_1) + g_3(0, -f_3, f_2).$$

By construction $d_1 \circ d_2 = 0$, which can also be seen from the matrix representation

$$\begin{bmatrix} f_1 & f_2 & f_2 \end{bmatrix} \begin{bmatrix} -f_2 & -f_3 & 0 \\ f_1 & 0 & -f_3 \\ 0 & f_1 & f_2 \end{bmatrix} = 0.$$

Taking a closer look at the definition of d_2, we find again at least one trivial element in its kernel: $(f_3, -f_2, f_1) \in \ker d_2$. This gives us the next map in our chain complex: $d_3 : R \to R^3$ is given by $d_3(g) = g(f_3, -f_2, f_1)$. This results in the complex

$$0 \longrightarrow R \xrightarrow{d_3} R^3 \xrightarrow{d_2} R^3 \xrightarrow{d_1} R \longrightarrow 0. \tag{A.2.3}$$

We will see that under certain assumptions, the augmented complex

$$0 \longrightarrow R \xrightarrow{d_3} R^3 \xrightarrow{d_2} R^3 \xrightarrow{d_1} R \longrightarrow R/I \longrightarrow 0. \tag{A.2.4}$$

gives a free resolution of R/I. △

We will now introduce some notation which allows to extend the constructions in Examples A.2.18 and A.2.19 for s elements $f_1, \ldots, f_s$ of our ring R. That is, we will define a complex

$$\mathcal{K}(f_1, \ldots, f_s) : \quad 0 \longrightarrow K_s \xrightarrow{d_s} K_{s-1} \xrightarrow{d_{s-1}} \cdots \xrightarrow{d_2} K_1 \xrightarrow{d_1} R \longrightarrow 0 \tag{A.2.5}$$

where $K_1, \ldots, K_s$ are free R-modules and $\operatorname{im} d_1 = I = \langle f_1, \ldots, f_s \rangle$. To this end, let $e_1, \ldots, e_s$ be symbols and let K_1 be the free module

$$K_1 = \bigoplus_{i=1}^{s} R \cdot e_i$$

of rank s generated by the e_i ($\{e_1, \ldots, e_s\}$ is an R-module basis for K_1). It is clear what the definition of $d_1 : K_1 \to R$ should be:

$$d_1(g_1 e_1 + \ldots + g_s e_s) = g_1 f_1 + \cdots + g_s f_s.$$

This map is completely defined by the image of the basis elements $d_1(e_i) = f_i$ by extending R-linearly. For bases of the remaining modules K_ℓ, we use the symbols $e_{i_1} \wedge \cdots \wedge e_{i_\ell}$ where $1 \le i_1 \le \cdots \le i_\ell \le s$. We set

$$K_\ell = \bigoplus_{1 \le i_1 \le \cdots \le i_\ell \le s} R \cdot e_{i_1} \wedge \cdots \wedge e_{i_\ell}.$$

For the reader who is familiar with exterior products, we will discuss the intuition behind this notation in Remark A.2.2. The important thing is that this notation allows for an elegant definition of the maps $d_\ell : K_\ell \to K_{\ell-1}$, generalizing d_1, d_2, d_3 from Examples A.2.18 and A.2.19. We set

$$d_\ell(e_{i_1} \wedge \cdots \wedge e_{i_\ell}) = \sum_{j=1}^{\ell} (-1)^{j-1} f_{i_j} e_{i_1} \wedge \cdots \wedge \widehat{e_{i_j}} \wedge \cdots \wedge e_{i_\ell}, \tag{A.2.6}$$

where the hat on $\widehat{e_{i_j}}$ indicates that the symbol e_{i_j} is omitted. To check that these definitions make (A.2.5) into a complex, we need to show that $d_{\ell-1} \circ d_\ell = 0$. This follows from

$$(d_{\ell-1} \circ d_\ell)(e_{i_1} \wedge \cdots \wedge e_{i_\ell}) = \sum_{j=1}^{\ell} (-1)^{j-1} f_{i_j} d_{\ell-1}(e_{i_1} \wedge \cdots \wedge \widehat{e_{i_j}} \wedge \cdots \wedge e_{i_\ell})$$

which is equal to

$$\sum_{j=1}^{\ell} (-1)^{j-1} f_{i_j} \Big(\sum_{k=1}^{j-1} (-1)^{k-1} f_{i_k} e_{i_1} \wedge \cdots \wedge \widehat{e_{i_k}} \wedge \cdots \wedge \widehat{e_{i_j}} \wedge \cdots \wedge e_{i_\ell} \\ + \sum_{k=j+1}^{\ell} (-1)^{k-2} f_{i_k} e_{i_1} \wedge \cdots \wedge \widehat{e_{i_j}} \wedge \cdots \wedge \widehat{e_{i_k}} \wedge \cdots \wedge e_{i_\ell} \Big)$$

which is indeed zero because the term corresponding to the ordered tuple (j,k), $1 \leq j,k \leq \ell$, $k \neq j$ and the term corresponding to (k,j) have opposite sign. The complex (A.2.5) is called the *Koszul complex* of the ordered tuple $(f_1, \ldots, f_s)$.

Example A.2.20. For $s = 2$, The Koszul complex $\mathcal{K}(f_1, f_2)$ is

$$\mathcal{K}(f_1, f_2): \quad 0 \longrightarrow K_2 \xrightarrow{d_2} K_1 \xrightarrow{d_1} R \longrightarrow 0$$

with $K_1 = R \cdot e_1 \oplus R \cdot e_2 \simeq R^2$, $K_2 = R \cdot e_1 \wedge e_2 \simeq R$ and

$$\begin{aligned} d_2(e_1 \wedge e_2) &= f_1 e_2 - f_2 e_1, \\ d_1(e_1) &= f_1, \quad d_1(e_2) = f_2. \end{aligned}$$

△

Example A.2.21. For $s = 3$, The Koszul complex $\mathcal{K}(f_1, f_2, f_3)$ is

$$\mathcal{K}(f_1, f_2, f_3): \quad 0 \longrightarrow K_3 \xrightarrow{d_3} K_2 \xrightarrow{d_2} K_1 \xrightarrow{d_1} R \longrightarrow 0$$

with

$$\begin{aligned} K_1 &= R \cdot e_1 \oplus R \cdot e_2 \oplus R \cdot e_3 \simeq R^3, \\ K_2 &= R \cdot e_1 \wedge e_2 \oplus R \cdot e_1 \wedge e_3 \oplus R \cdot e_2 \wedge e_3 \simeq R^3, \\ K_3 &= R \cdot e_1 \wedge e_2 \wedge e_3 \simeq R \end{aligned}$$

and

$$\begin{aligned} d_3(e_1 \wedge e_2 \wedge e_3) &= f_1 e_2 \wedge e_3 - f_2 e_1 \wedge e_3 + f_3 e_1 \wedge e_2, \\ d_2(e_1 \wedge e_2) &= f_1 e_2 - f_2 e_1, \\ d_2(e_1 \wedge e_3) &= f_1 e_3 - f_3 e_1, \\ d_2(e_2 \wedge e_3) &= f_2 e_3 - f_3 e_2, \\ d_1(e_i) &= f_i, \quad i = 1, 2, 3. \end{aligned}$$

△

Remark A.2.2. Another way to define the Koszul complex is via the so-called *dual complex*

$$\mathcal{K}^\vee(f_1,\ldots,f_s): \quad 0 \longrightarrow R^\vee \xrightarrow{d_1^\vee} K_1^\vee \xrightarrow{d_2^\vee} \cdots \xrightarrow{d_{s-1}^\vee} K_{s-1}^\vee \xrightarrow{d_s^\vee} K_s^\vee \longrightarrow 0,$$

where $R^\vee = \operatorname{Hom}_R(R,R) \simeq R$, $K_\ell^\vee = \operatorname{Hom}_R(K_\ell,R) \simeq K_\ell$ and the map $d_\ell^\vee$ sends $(\phi : K_{\ell-1} \to R) \in K_{\ell-1}^\vee$ to $\phi \circ d_\ell \in K_\ell^\vee$ where d_ℓ is as defined above. If $v_i : K_1 \to R$ is the map that sends e_i to 1 and e_j to 0 for $j \neq i$, then $K_1^\vee$ has basis $v_1,\ldots,v_s$ and $K_\ell^\vee$ has basis

$$\{v_{i_1} \wedge \cdots \wedge v_{i_\ell}\}_{1 \leq i_1 \cdots \leq i_\ell \leq s}$$

where $v_{i_1} \wedge \cdots \wedge v_{i_\ell} \in K_\ell^\vee$ is the map that sends $e_{i_1} \wedge \cdots \wedge e_{i_\ell}$ to 1 and all other basis elements to 0. Then the maps $d_\ell^\vee$ are defined simply by $d_1^\vee(1) = f_1v_1 + \ldots + f_sv_s$ and

$$d_{\ell+1}^\vee(v_{i_1} \wedge \cdots \wedge v_{i_\ell}) = \sum_{j=1}^{s} f_jv_j \wedge v_{i_1} \wedge \cdots \wedge v_{i_\ell},$$

where the *wedge product* $v_i \wedge v_j$ has the usual algebraic properties of being *alternating* $(v_i \wedge v_i = 0)$ and anti-commutative $(v_i \wedge v_j = -v_j \wedge v_i)$. As an example, consider the case $s = 3$. We have $d_1^\vee(1) = f_1v_1 + f_2v_2 + f_3v_3$ and

$$\begin{aligned}
d_2^\vee(v_1) &= f_1v_1 \wedge v_1 + f_2v_2 \wedge v_1 + f_3v_3 \wedge v_1 = -f_2v_1 \wedge v_2 - f_3v_1 \wedge v_3,\\
d_2^\vee(v_2) &= f_1v_1 \wedge v_2 + f_2v_2 \wedge v_2 + f_3v_3 \wedge v_2 = f_1v_1 \wedge v_2 - f_3v_2 \wedge v_3,\\
d_2^\vee(v_3) &= f_1v_1 \wedge v_3 + f_2v_2 \wedge v_3 + f_3v_3 \wedge v_3 = f_1v_1 \wedge v_3 + f_2v_2 \wedge v_3,\\
d_3^\vee(v_1 \wedge v_2) &= f_1v_1 \wedge v_1 \wedge v_2 + f_2v_2 \wedge v_1 \wedge v_2 + f_3v_3 \wedge v_1 \wedge v_2 = f_3v_1 \wedge v_2 \wedge v_3,\\
d_3^\vee(v_1 \wedge v_3) &= f_1v_1 \wedge v_1 \wedge v_3 + f_2v_2 \wedge v_1 \wedge v_3 + f_3v_3 \wedge v_1 \wedge v_3 = -f_2v_1 \wedge v_2 \wedge v_3,\\
d_3^\vee(v_2 \wedge v_3) &= f_1v_1 \wedge v_2 \wedge v_3 + f_2v_2 \wedge v_2 \wedge v_3 + f_3v_3 \wedge v_2 \wedge v_3 = f_1v_1 \wedge v_2 \wedge v_3,
\end{aligned}$$

which shows that the matrices of $(d_i)^\vee$ from the dual Koszul complex are exactly the transposes of the matrices representing d_i in $\mathcal{K}(f_1,\ldots,f_s)$. △

If $I = \langle f_1,\ldots,f_s\rangle \neq R$, there is no hope for $\mathcal{K}(f_1,\ldots,f_s)$ to be exact. We will often consider the so-called *augmented Koszul complex* given by

$$\hat{\mathcal{K}}(f_1,\ldots,f_s): \quad 0 \longrightarrow K_s \xrightarrow{d_s} K_{s-1} \xrightarrow{d_{s-1}} \cdots \xrightarrow{d_2} K_1 \xrightarrow{d_1} R \longrightarrow R/I \longrightarrow 0. \quad \text{(A.2.7)}$$

It turns out that this complex is an exact sequence under some easy-to-describe conditions on the f_i. Recall that an element $f \in R$ is called a *zero divisor* in R if there is some $g \neq 0$ such that $fg = 0$.

Definition A.2.16 (Regular sequence)**.** Let R be a commutative ring with identity. A sequence $f_1,\ldots,f_s \in R$ is called a *regular sequence* if $\langle f_1,\ldots,f_s\rangle \neq R$, f_1 is not a zero divisor in R and $f_i + \langle f_1,\ldots,f_{i-1}\rangle$ is not a zero divisor in the quotient ring $R/\langle f_1,\ldots,f_{i-1}\rangle$, for $i = 2,\ldots,s$.

Theorem A.2.6. *Let R be a commutative ring with identity and let $f_1, \dots, f_s \in R$ be a regular sequence. Then the augmented Koszul complex $\hat{\mathcal{K}}(f_1, \dots, f_s)$ is a free resolution of R/I.*

Proof. See [Lan02, Chapter XXI, Theorem 4.6]. □

Remark A.2.3. The property of regularity may depend on the order of the elements $f_1, \dots, f_s$, but the exactness of the Koszul complex does not. Here's an example taken from [Ben19, page 41]. Let $R = \mathbb{C}[x, y, z]$ and consider the polynomials $f = z, g = x(z+1), h = y(z+1)$. One can check that f, g, h is a regular sequence, but g, h, f is not. △

When we are working in a graded setting, we will twist the gradings of the free modules K_ℓ such that all homomorphisms d_ℓ have degree zero. In this way we hope to obtain graded resolutions via the Koszul complex. If S is an E-graded $\mathbb{C}$-algebra and $I = \langle f_1, \dots, f_s \rangle$ is a homogeneous ideal generated by homogeneous elements f_i of degree $\deg(f_i) = \alpha_i$, then the Koszul complex $\mathcal{K}(f_1, \dots, f_s)$ is defined as

$$\mathcal{K}(f_1, \dots, f_s): \quad 0 \longrightarrow K_s \xrightarrow{d_s} K_{s-1} \xrightarrow{d_{s-1}} \cdots \xrightarrow{d_2} K_1 \xrightarrow{d_1} S \longrightarrow 0, \tag{A.2.8}$$

where

$$K_\ell = \bigoplus_{1 \le i_1 \le \cdots \le i_\ell \le s} S(-\alpha_{i_1} - \cdots - \alpha_{i_\ell}) \cdot e_{i_1} \wedge \cdots \wedge e_{i_\ell}.$$

and $d_\ell : K_\ell \to K_{\ell-1}$ are defined as in (A.2.6). The augmented Koszul complex $\hat{\mathcal{K}}(f_1, \dots, f_s)$ is defined as in (A.2.7), with R replaced by S.

Example A.2.22. Let S be an E-graded $\mathbb{C}$-algebra and let $f_1, f_2 \in S$ be homogeneous of degree α_1, α_2 respectively. The Koszul complex $\mathcal{K}(f_1, f_2)$ is

$$0 \longrightarrow S(-\alpha_1 - \alpha_2) \xrightarrow{d_2} \begin{matrix} S(-\alpha_1) \\ \oplus \\ S(-\alpha_2) \end{matrix} \xrightarrow{d_1} S \longrightarrow 0.$$

Note that an element $(g_1, g_2) \in (S(-\alpha_1) \oplus S(-\alpha_2))_\alpha$ is sent to $g_1 f_1 + g_2 f_2 \in S_\alpha$ under d_1, so d_1 is indeed of degree 0. The same can be checked for d_2. △

A.2.6 Localization of modules

The definition of localization (see Definition A.1.15) can be generalized to R-modules.

Definition A.2.17 (Localization of R-modules)**.** Let $T \subset R$ be a multiplicatively closed subset of R, that is, $1 \in T$ and T is closed under multiplication. The *localization* $T^{-1}M$ of an R-module M at T is the $T^{-1}R$-module

$$\{m/g \mid m \in M, g \in T\}/\sim$$

where $m_1/g_1 \sim m_2/g_2 \Leftrightarrow t(g_2m_1 - g_1m_2) = 0$ in M for some $t \in T$, with operations

$$m_1/g_1 + m_2/g_2 = (g_1m_1 + g_2m_2)/(g_1g_2) \quad \text{and} \quad (f/g_1)(m/g_2) = (fm)/(g_1g_2),$$

where $m_1, m_2, m \in M$, $g_1, g_2 \in T$, $f \in R$.

An R-module homomorphism $\phi : M \to M'$ can be 'localized' to obtain a $T^{-1}R$-module homomorphism $T^{-1}\phi : T^{-1}M \to T^{-1}M'$ by setting

$$T^{-1}\phi(m/g) = \phi(m)/g.$$

Note that this construction behaves nicely with respect to composition: $T^{-1}(\phi \circ \psi) = T^{-1}\phi \circ T^{-1}\psi$. The operation of localizing R-modules and homomorphisms between them has the special property of preserving exactness. The following is Proposition 3.3 in [AM69].

Proposition A.2.4. *Let $M'' \xrightarrow{\psi} M \xrightarrow{\phi} M'$ be an exact sequence of R-modules and let $T \subset R$ be a multiplicatively closed subset, then*

$$T^{-1}M'' \xrightarrow{T^{-1}\psi} T^{-1}M \xrightarrow{T^{-1}\phi} T^{-1}M'$$

is an exact sequence of $T^{-1}R$-modules.

Example A.2.23. Let $I \subsetneq R$ be an ideal and let $A = R/I$ be the corresponding quotient ring. For $f \in R \backslash I$, the localization A_f of A at f as an R-module is isomorphic to the localization A_{f+I} of A at $f + I$ as an (R/I)-module via

$$\frac{g+I}{f^\ell} \mapsto \frac{g+I}{f^\ell + I}.$$

It follows from this observation, $0 \to I \to R \to A \to 0$ and Proposition A.2.4 that $A_{f+I} \simeq R_f/I_f$, where I_f is the image of I under $R \to R_f$. △

The localization $T^{-1}R$ has the obvious structure of an R-module. The tensor product of R-modules $T^{-1}R \otimes_R M$ can be given the structure of a $T^{-1}R$-module by setting

$$(f/g) \cdot (f'/g' \otimes m) = (ff')/(gg') \otimes m.$$

This allows to describe the localization as a tensor product of R-modules. This is Proposition 3.5 in [AM69].

Proposition A.2.5. *Let $T \subset R$ be a multiplicatively closed subset and let M be an R-module. The homomorphism*

$$T^{-1}R \otimes_R M \to T^{-1}M$$

given by $f/g \otimes m \mapsto (fm)/g$ is an isomorphism of $T^{-1}R$-modules.

Appendix B

Numerical linear algebra

In this appendix we give a brief introduction to the methods and concepts of *numerical linear algebra* that are used in this thesis. We discuss some of the most important matrix factorizations and their use for solving linear systems of equations and for computing eigenvalues. Numerical linear algebra algorithms are at the heart of countless methods for solving problems in applied mathematics. While further improving and specializing these algorithms is still an active area of research today, the state of the art implementations (e.g. the LAPACK library [ABB+99]) are able to solve linear systems and eigenvalue problems in a *backward stable* way. This makes the tools very powerful, and it is a great motivation for trying to reformulate any computational problem as a problem of numerical linear algebra. We limit ourselves to conceptual descriptions and give full references for algorithmic details. The book of Trefethen and Bau [TBI97] contains a great first introduction to some of the fundamental concepts. A more complete, encyclopedic treatment can be found in the book by Golub and Van Loan [GVL12].

We work with finite dimensional vector spaces over $\mathbb{C}$. A matrix $A \subset \mathbb{C}^{m \times n}$ is a 2-dimensional array, whose entries are denoted by

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix} = (A_{ij})_{1 \leq i \leq m, 1 \leq j \leq n}.$$

The *Euclidean 2-norm* of $v = (v_1, \ldots, v_m) \in \mathbb{C}^m$ is

$$\|v\|_2 = \sqrt{\overline{v_1}v_1 + \cdots + \overline{v_m}v_m},$$

where $\bar{\cdot}$ denotes complex conjugation. For an n-dimensional $\mathbb{C}$-vector space V, an m-dimensional $\mathbb{C}$-vector space W and any norms $\|\cdot\|_V$, $\|\cdot\|_W$ on V and W respectively, the *induced operator norm* of a $\mathbb{C}$-linear map $A : V \to W$ (which we think of as a matrix) is

$$\|A\|_{V,W} = \sup_{x \in V \setminus \{0\}} \frac{\|Ax\|_W}{\|x\|_V}.$$

If $\|\cdot\|_V$ and $\|\cdot\|_W$ are the Euclidean 2-norms on V and W respectively, we denote $\|\cdot\|_2$ for the induced operator norm. To keep the notation unambiguous, we will use bold characters for the matrices in standard factorizations. For instance, since Q, R, S, U, V are reserved for polynomial rings, primary ideals, vector spaces, open subsets of a variety, ... we will use $\mathbf{Q}, \mathbf{R}, \mathbf{S}, \mathbf{U}, \mathbf{V}$ in the QR factorization and the SVD.

B.1 Conditioning and stability

We know from our first course in linear algebra that if $A \in \mathbb{C}^{m \times m}$ is a square, nonsingular matrix and $b \in \mathbb{C}^m$ is any column vector, then $Ax = b$ has exactly one solution. This solution is given by $x = A^{-1}b$. We could view x as the image of a function $f(A, b) = x$, which takes an invertible matrix $A \in \mathbb{C}^{m \times m}$ and a vector $b \in \mathbb{C}^m$ and returns the solution of $Ax = b$. In exact arithmetic, this is where the story ends. When we want to 'compute' the solution of $Ax = b$ in *floating point arithmetic*[1], the intermediate results of the computations are replaced by nearby *machine numbers*, causing *rounding errors* in the computed solution $\tilde{x}$. Moreover, if A and b cannot be represented exactly on the computer, their entries are replaced by machine numbers before the computations even start. Abstractly, we can think of our numerical algorithm which computes the approximation $\tilde{x}$ as an operator $\hat{f}$ such that $\hat{f}(A, b) = \tilde{x}$, where $\hat{f}$ 'approximates' f. Hopefully, we will still have the 'approximate equalities' $A\tilde{x} \approx b$ and, more ambitiously, $\tilde{x} \approx x$. To establish whether the numerical algorithm $\hat{f}$ did a good job, we need a way of *measuring* the 'magnitude' of the errors $A\tilde{x} - b$ and $\tilde{x} - x$, and a way of deciding whether the obtained errors are *satisfactory*. A good criterion for deciding this takes into account that the computer treats our original problem as a slightly perturbed version of the problem, and the solution x may be very sensitive to such perturbations. All of this is captured by the fundamental concepts of forward error, backward error, condition and stability in numerical analysis.

In general, we may think of a *problem* as a function $f : V \to W$ between normed vector spaces. We use the notation $\|v\|$, $\|w\|$ for the norms of $v \in V$, $w \in W$. The vector space V is the *space of data* and W is the *space of solutions*. 'Solving' the problem with data $v \in V$ corresponds to computing $f(v)$. Suppose $f(v) = w$ and consider a perturbation $\delta v \in V$, which should be thought of as a vector with small norm. The *sensitivity* of f at v can be measured by

$$\frac{\|f(v) - f(v + \delta v)\|}{\|\delta v\|}. \tag{B.1.1}$$

[1] We assume familiarity with the floating point number system. Introductions can be found, for instance, in [Hig02, Chapter 2] or [TBI97, Lecture 13].

This number clearly depends on δv, so the actual measure for the local sensitivity should be the supremum of (B.1.1) over all small perturbations δv. In the context of floating point arithmetic, it is natural to measure the distances between $f(v), f(v+\delta v)$ and $v, v+\delta v$ *relatively* with respect to the norms $\|f(v)\|$ and $\|v\|$. Motivated by these considerations, as a measure for the sensitivity of f to perturbations on v, we define the *relative condition number*

$$\kappa_f(v) = \lim_{\varepsilon\to 0} \sup_{\|\delta v\|\le\varepsilon} \left(\frac{\|f(v+\delta v) - f(v)\|}{\|f(v)\|} \Big/ \frac{\|\delta v\|}{\|v\|} \right).$$

The condition number $\kappa_f(v)$ may depend *strongly* on v, meaning that some *instances* of the problem f may be much more sensitive to perturbations than others. We say that the problem f is *well-conditioned* at v if $\kappa_f(v)$ is small, and that it is *ill-conditioned* at v if $\kappa_f(v)$ is large. What 'small' and 'large' mean may depend on the specific problem and on the accuracy with which one wants to compute. Note that *conditioning is a property of the problem*, not of an algorithm for solving the problem (approximately).

Example B.1.1 (Matrix-vector product)**.** Fix a matrix $A \subset \mathbb{C}^{m\times m}$ and let $f : \mathbb{C}^m \to \mathbb{C}^m$ be given by $f(x) = Ax$. For any norm $\|\cdot\|$ on $\mathbb{C}^m$ we find that

$$\kappa_f(x) = \lim_{\varepsilon\to 0} \sup_{\|\delta x\|\le\varepsilon} \left(\frac{\|A\delta x\|}{\|Ax\|} \Big/ \frac{\|\delta x\|}{\|x\|} \right) = \|A\| \frac{\|x\|}{\|Ax\|},$$

where $\|A\|$ is the operator norm induced by $\|\cdot\|$ on $\mathbb{C}^m$. The same formula holds for non-square matrices. If A is invertible this gives a global bound for the condition number by using $\|x\|/\|Ax\| \le \|A^{-1}\|$:

$$\kappa_f(x) \le \|A\|\|A^{-1}\|, \quad \text{for all } x \in \mathbb{C}^m. \tag{B.1.2}$$

The number $\kappa(A) = \|A\|\|A^{-1}\|$ is a very important constant, called *the condition number of* A. We will characterize the vectors x for which the bound (B.1.2) is attained, i.e. for which $\|x\|/\|Ax\| = \|A^{-1}\|$ in Section B.2. △

Example B.1.2 (Solving linear systems)**.** Fix an invertible matrix $A \subset \mathbb{C}^{m\times m}$ and let $f_A : \mathbb{C}^m \to \mathbb{C}^m$ be the function sending $b \in \mathbb{C}^m$ to the solution of $Ax = b$, that is $f_A(b) = A^{-1}b$. By the results of Example B.1.1 we find that $\kappa_{f_A}(b) \le \kappa(A)$ and the bound is attained where $\|b\|/\|A^{-1}b\| = \|A\|$. We conclude that the sensitivity of the problem of solving a system of linear equations to perturbations on the right hand side b is measured by the condition number of A. Let us now fix $b \in \mathbb{C}^m$ and consider the function $f_b : \mathbb{C}^{m\times m} \to \mathbb{C}^m$ such that $f_b(A) = A^{-1}b$. Denoting $f_b(A) = x$ and $f_b(A+\delta A) = x + \delta x$, we have that

$$(A+\delta A)(x+\delta x) = b.$$

Since in the definition of the condition number we take the limit $\lim_{\|\delta A\|\to 0}$, we can drop the doubly infinitesimal term $(\delta A)(\delta x)$ to find that $\delta A x + A\delta x = 0$. Hence $\|\delta x\| \le \|A^{-1}\|\|\delta A\|\|x\|$, which gives

$$\left(\frac{\|\delta x\|}{\|x\|} \Big/ \frac{\|\delta A\|}{\|A\|} \right) \le \kappa(A).$$

There are perturbations δA for which this bound is attained [TBI97, Exercise 3.6], so we get $\kappa_{f_b}(A) = \kappa(A)$. This shows that also the sensitivity of the problem of solving a linear system $Ax = b$ with respect to perturbations in A is governed by the condition number of A. We conclude that if the data (A, b) of the linear system $Ax = b$ is perturbed by a relative error of size u, where u is the unit round-off (for instance, $u = 2^{-52} \approx 10^{-16}$ in double precision arithmetic), then the order of magnitude of the perturbation on the exact solution x of $Ax = b$ is at most the order of magnitude of $\kappa(A)u$. In fact, the relative size of the perturbation on x is of the same order of magnitude as $\kappa(A)u$, except in some very special situations. These observations are the motivation for the general rule of thumb in numerical linear algebra that *when one wants to compute $A^{-1}b$ in floating point arithmetic, one generally loses* $\log_{10}(\kappa(A))$ *decimal digits of accuracy, that is*

$$\frac{\|\delta x\|}{\|x\|} \approx \kappa(A)u.$$

The condition number $\kappa(A)$ of a matrix plays a very important role in this thesis. In what follows, we will always consider $\kappa(A)$ with respect to the Euclidean 2-norm $\|\cdot\|_2$. △

Example B.1.3. Consider the linear systems $Ax = b$ and $A(x + \delta x) = b + \delta b$ where

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1+\varepsilon \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \quad \delta b = \begin{bmatrix} 0 \\ \varepsilon \end{bmatrix}.$$

The solutions are

$$x = \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \quad x + \delta x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

As $\epsilon \to 0$, the condition number of A with respect to $\|\cdot\|_2$ behaves like ε^{-1}, whereas $\|\delta b\|_2 = \varepsilon$. △

The condition number relates the relative size of perturbations on the input to the relative size of the resulting perturbations on the output of a problem $f : V \to W$. Related to these two types of perturbations there are two ways of measuring the *error* of a point $\tilde{x} \in W$ as an approximation for $f(v) \in W$. In what follows, we fix a norm $\|\cdot\|$, which is usually taken to be the Euclidean 2-norm $\|\cdot\|_2$.

Definition B.1.1 (Relative forward error)**.** For a problem $f : V \to W$ and a point $v \in V$, the *relative forward error* of a point $\tilde{x} \in W$ as an approximation for $f(v) = x \in W$ is

$$\frac{\|x - \tilde{x}\|}{\|x\|}.$$

Definition B.1.2 (Relative backward error)**.** For a problem $f : V \to W$ and a point $v \in V$, the *relative backward error* of a point $\tilde{x} \in W$ as an approximation for $f(v) = x \in W$ is the smallest $\varepsilon \in \mathbb{R}_{\geq 0}$ such that there exists $\tilde{v} \in V$ with

$$\frac{\|v - \tilde{v}\|}{\|v\|} \leq \varepsilon \quad \text{and} \quad f(\tilde{v}) = \tilde{x}.$$

The relative forward error of $\tilde{x}$ is small if the approximate solution is close to the actual solution (in a relative sense). The relative backward error of $\tilde{x}$ is small if the approximate solution is the exact solution of a slightly perturbed problem instance. For the example of solving $Ax = b$, $\tilde{x}$ has a small relative backward error if there is a slightly perturbed vector $\tilde{b} \in \mathbb{C}^m$ close to b such that $A\tilde{x} = \tilde{b}$. The relative backward error for this example can be measured by

$$\frac{\|\tilde{b} - b\|}{\|b\|} = \frac{\|A\tilde{x} - b\|}{\|b\|}.$$

In the following definition, we use the notation $\hat{f} : V \to W$ for a numerical algorithm that 'approximates' a problem $f : V \to W$. That is, given a point $v \in V$, $\hat{f}(v)$ is an approximation for $f(v)$. We say that a positive real number a is of size $O(\varepsilon)$ if a has 'order of magnitude' ε. In practice, this means that a is bounded by $C^{-1}\varepsilon \leq a \leq C\varepsilon$, for a 'not too large' constant C (e.g. $C = 100$).

Definition B.1.3 (Forward stability)**.** An algorithm $\hat{f} : V \to W$ is called *forward stable* if for any $v \in V$, $\hat{f}(v)$ has a relative forward error of size $O(u)$.

Remark B.1.1. Different authors use different definitions for various notions of stability. For instance, in [TBI97, Lecture 14], a forward stable algorithm in the sense of Definition B.1.3 is called *accurate* and the definition of forward stability in [Hig02, §1.6] takes the condition number into account. In the spirit of [Bul06, Chapter 2, Subsection 7.3], Definition B.1.3 emphasizes that *forward* stability is measured by the *forward* error. △

The sensitivity of the problem f to perturbations may depend strongly on the input v. Since we are working in floating point arithmetic, there is not much we can do about this: after a single floating point operation it is as if we were dealing with a slightly perturbed problem. Therefore, it seems too strict to ask of our numerical algorithms to be forward stable. It makes more sense to impose a small *backward error*.

Definition B.1.4 (Backward stability)**.** An algorithm $\hat{f} : V \to W$ is called *backward stable* if for any $v \in V$, $\hat{f}(v)$ has a relative backward error of size $O(u)$.

A backward stable algorithm finds the exact solution to a slightly perturbed version of the problem that one wants to solve. If the problem instance we are interested in is ill-conditioned, then the forward error might still be large. The factor between backward and forward error depends on the conditioning of f at $v \in V$. One has the rule of thumb

$$\text{relative forward error } = O(\kappa_f(v) \cdot \text{ relative backward error}).$$

Note the similarity with the definition of $\kappa_f(v)$. This means that for a backward stable algorithm the forward error satisfies

$$\text{relative forward error } = O(\kappa_f(v) \cdot u).$$

Backward stability is the type of stability one usually aims for in numerical linear algebra, and there are backward stable algorithms for the fundamental problems of solving a linear system of equations $Ax = b$ and solving an eigenvalue problem $Ax = \lambda x$.

Unlike conditioning, *stability is a property of a method or algorithm*, not that of a problem. Usually, the condition of a problem is out of our hands, but (backward) stability is what we aim for in designing our algorithms. Somewhat confusingly, often algorithms are unstable because they reformulate the problem to one that is mathematically equivalent, *but much more ill-conditioned.* A typical example is that of solving a linear least squares problem via the normal equations [TBI97, Lecture 19], and we will encounter some examples related to the problem of polynomial system solving in this thesis as well.

B.2 Singular value decomposition

An important class of problems in numerical linear algebra is that of computing *matrix factorizations.* In general, this means that for a matrix $A \in \mathbb{C}^{m \times n}$ we compute a set of matrices $A_1, \ldots, A_k$ such that $A = A_1 A_2 \cdots A_k$ and this *decomposition* or *factorization* of A either helps us do further computations with A or it reveals some properties of A that are of interest to us. Perhaps the most powerful of all such factorizations is the *singular value decomposition.*

For a matrix $A \in \mathbb{C}^{m \times n}$, let $A^H \in \mathbb{C}^{n \times m}$ be its Hermitian transpose. That is,

$$A = (A_{ij})_{1 \leq i \leq m, 1 \leq j \leq n}, \quad A^H = ((A^H)_{ij})_{1 \leq i \leq n, 1 \leq j \leq m} \text{ with } (A^H)_{ij} = \overline{A_{ji}}$$

and $\overline{a + b\sqrt{-1}} = a - b\sqrt{-1}, a, b \in \mathbb{R}$ denotes complex conjugation. We think of row and column vectors as matrices of size $1 \times m$ and $m \times 1$ respectively, and the pairing $\mathbb{C}^m \times \mathbb{C}^m \to \mathbb{C}$ given by $(v, w) \mapsto v^H w$ is the scalar product that induces the Euclidean 2-norm $\|v\|_2 = \sqrt{v^H v}$ on $\mathbb{C}^m$. We recall the following definition

Definition B.2.1 (Unitary matrix)**.** A matrix $A \in \mathbb{C}^{m \times m}$ is called *unitary* if $A^H A = \mathrm{id}_m$, where id_m is the identity matrix of size $m \times m$. Equivalently, A is unitary if $A^{-1} = A^H$.

Definition B.2.2 (Singular value decomposition (SVD))**.** For a matrix $A \in \mathbb{C}^{m \times n}$, a decomposition $A = \mathbf{U}\mathbf{S}\mathbf{V}^H$ is called a *singular value decomposition (SVD)* of A if

1. $\mathbf{U} \in \mathbb{C}^{m \times m}$ and $\mathbf{V} \in \mathbb{C}^{n \times n}$ are unitary,
2. $\mathbf{S} \in \mathbb{R}^{m \times n}$ is diagonal with nonnegative entries $\sigma_i = \mathbf{S}_{ii} \in \mathbb{R}_{\geq 0}$ on its diagonal such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n)} \geq 0$.

The numbers $\sigma_1, \ldots, \sigma_{\min(m,n)}$ are called *singular values* of A.

Theorem B.2.1 (Existence and uniqueness of the SVD). *Any matrix $A \in \mathbb{C}^{m\times n}$ has a singular value decomposition. Moreover, the singular values σ_i are uniquely determined. If $m = n$ and $\sigma_i \neq \sigma_j$ for $i \neq j$, then the columns of $\mathbf{U}$ and $\mathbf{V}$ in the SVD $A = \mathbf{U}\mathbf{S}\mathbf{V}^H$ are unique up to complex sign.*

Proof. See [TBI97, Theorem 4.1] or the analogous proof for real matrices in [GVL12, Section 2.5]. □

By 'unique up to complex sign' in Theorem B.2.1 we mean up to multiplication with a complex number $e^{\sqrt{-1}\theta}$ of modulus 1. It is convenient to have the notation $\sigma_1 = \sigma_{\max}$ and $\sigma_{\min(m,n)} = \sigma_{\min}$ for the largest and smallest singular value of the matrix A. We write $u_i = \mathbf{U}_{:,i}$ for the i-th column of $\mathbf{U}$ and likewise for the columns v_i of $\mathbf{V}$. Let r be the largest index such that $\sigma_r \neq 0$. The SVD $A = \mathbf{U}\mathbf{S}\mathbf{V}^H$ can be written as

$$A = \begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix}\begin{bmatrix}\mathbf{S}_1 & 0_{r,n-r}\\ 0_{m-r,r} & 0_{m-r,n-r}\end{bmatrix}\begin{bmatrix}\mathbf{V}_1^H\\ \mathbf{V}_2^H\end{bmatrix} = \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^H, \tag{B.2.1}$$

with $0_{k,\ell}$ the zero matrix of size $k \times \ell$, $\mathbf{U}_1 = [u_1 \ \cdots \ u_r], \mathbf{U}_2 = [u_{r+1} \ \cdots \ u_m], \mathbf{V}_1 = [v_1 \ \cdots \ v_r], \mathbf{V}_2 = [v_{r+1} \ \cdots \ v_n]$, $\mathbf{S}_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$. To distinguish the factorizations

$$A = \mathbf{U}\mathbf{S}\mathbf{V}^H \quad \text{and} \quad A = \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^H$$

they are sometimes called the *full* SVD and the *thin* or *reduced* SVD of A. In this text, when we talk about *the* SVD we always have the full SVD in mind. If we know the SVD of A, we have an orthonormal basis for all fundamental subspaces of A and we know its rank: from (B.2.1) we see that

1. the rank of A is the number of nonzero singular values, r,
2. $\ker A = \operatorname{span}_{\mathbb{C}}(v_{r+1}, \ldots, v_n) = \operatorname{im} \mathbf{V}_2$, the columns of $\mathbf{V}_2$ are an orthonormal basis for the *kernel* or *(right) nullspace* of A,
3. $\operatorname{im} A = \operatorname{span}_{\mathbb{C}}(u_1, \ldots, u_r) = \operatorname{im} \mathbf{U}_1$, the columns of $\mathbf{U}_1$ are an orthonormal basis for the *image*, *range* or *column space* of A,
4. $\operatorname{coker} A = \operatorname{span}_{\mathbb{C}}(u_{r+1}, \ldots, u_m) = \operatorname{im} \mathbf{U}_2$, i.e. $\ker \mathbf{U}_2^H = \operatorname{im} A$, the columns of $\mathbf{U}_2$ are an orthonormal basis for the *cokernel* or *left nullspace* of A,
5. $\operatorname{coim} A = \operatorname{span}_{\mathbb{C}}(v_1, \ldots, v_r) = \operatorname{im} \mathbf{V}_1$, the columns of $\mathbf{V}_1$ are an orthonormal basis for the *coimage*, *corange* or *row space* of A.

The SVD also allows to write A as a sum of r rank one matrices

$$A = \sigma_1 u_1 v_1^H + \cdots + \sigma_r u_r v_r^H$$

and the famous *Eckart-Young theorem* [EY36] guarantees that for $r' \leq r$,

$$A_{r'} = \sigma_1 u_1 v_1^H + \cdots + \sigma_{r'} u_{r'} v_{r'}^H$$

is the *best rank r' approximation of A*. By this we mean that it minimizes $\|A - A_{r'}\|$ over the rank r' matrices $A_{r'}$ where the norm can be, for instance, the operator 2-norm or the Frobenius norm.

A SVD of the matrix $A = \mathbf{U}\mathbf{S}\mathbf{V}^H$ gives immediately an SVD for its inverse (if $m = n$ and $r = m$) and its Hermitian transpose:

$$A^{-1} = \mathbf{V}\mathbf{S}^{-1}\mathbf{U}^H, \quad A^H = \mathbf{V}\mathbf{S}^\top\mathbf{U}^H.$$

The SVD also reveals some important norms of A. It is not difficult to show that $\|A\|_2 = \sigma_{\max}$ and $\|A\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$, where $\|\cdot\|_F$ denotes the Frobenius norm. An important direct consequence is that the condition number $\kappa(A)$ relative to the Euclidean 2-norm $\|\cdot\|_2$ is given by $\|A\|_2\|A^{-1}\|_2 = \sigma_{\max}/\sigma_{\min}$.

The SVD can be computed in a backward stable way [TBI97, Lecture 31], in the sense that for the computed matrices $\tilde{\mathbf{U}}, \tilde{\mathbf{S}}, \tilde{\mathbf{V}}$ we have

$$\frac{\|A - \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^H\|_2}{\|A\|_2} = O(u).$$

The complexity of the algorithms is $O(mn^2)$. The Eckart-Young theorem implies that if the computed singular values $\tilde{\sigma}_{r+1}, \ldots \tilde{\sigma}_{\min}$ (using a backward stable algorithm) are of size $O(\varepsilon)$, then there is a matrix *very close to A* (at distance $O(\varepsilon)$) of rank r. The SVD therefore provides us with a good method for *deciding on the numerical rank*. What is usually done is the following. The numerical rank of A is set to be the largest index r such that

$$\tilde{\sigma}_r > \mathrm{tol} \cdot \tilde{\sigma}_1,$$

where tol is some tolerance, typically 10-1000 times the unit round-off. This numerical rank decision allows to partition the SVD as in (B.2.1), where the zero matrix $0_{m-r,n-r}$ now contains the 'numerically zero' singular values on its diagonal and the submatrices are replaced by their numerical approximations $\tilde{\mathbf{U}}_1, \tilde{\mathbf{U}}_2, \tilde{\mathbf{S}}_1, \tilde{\mathbf{V}}_1, \tilde{\mathbf{V}}_2$. These matrices contain numerical approximations for the fundamental subspaces of A. Suppose that the *gap* $\gamma = \tilde{\sigma}_r - \tilde{\sigma}_{r+1}$ between the last 'numerically nonzero' and the first 'numerically zero' singular value is small, such that $\tilde{\sigma}_r$ is larger than $\mathrm{tol}\tilde{\sigma}_1$ but not much. Then it is clear that A is nearly as close to being rank $r-1$ as it is to being rank r, and the partitioning (B.2.1) is very sensitive to the value of tol. In this case, it is tricky to decide the numerical rank. Also, since the dimension of the numerical approximations for fundamental subspaces depends on the numerical rank, these spaces are harder to compute. Intuitively, we see that the *conditioning* of computing the numerical rank and the fundamental subspaces of A depends on the gap γ. This intuition can be made precise [Ste91, Ste06], but we will not go into detail here. If one wants to compute, for instance, the cokernel $\mathbf{U}_2$ of the matrix A and use it for further numerical computations, one should make sure that the gap γ is large enough, such that not too much accuracy is lost in the numerical computation of $\tilde{\mathbf{U}}_2$.

Remark B.2.1. Once the SVD of an invertible square matrix $A = \mathbf{U}\mathbf{S}\mathbf{V}^H$ is computed, it can be used to solve the linear system $Ax = b$ via $x = \mathbf{V}\mathbf{S}^{-1}\mathbf{U}^H b$.

Note that the inversion of the diagonal matrix $\mathbf{S}$ is trivial. This leads to a backward stable algorithm for solving linear systems, but it is not the most efficient one. Cheaper alternatives use LU factorization with pivoting followed by forward- and back substitution [TBI97, Lecture 21] or QR factorization followed by back substitution, as explained in the next section. △

Remark B.2.2. Note that a unitary matrix $\mathbf{Q}$ 'is its own SVD' in the sense that $\mathbf{U} = \mathbf{Q}$, $\mathbf{S} = \mathbf{V} = \mathrm{id}$. All of its singular values are 1 and it has a *perfect condition number* $\kappa(\mathbf{Q}) = 1$. Algorithms of numerical linear algebra make use of this fact all the time: often they repeatedly apply unitary transformations to a given matrix A, which is key to prove their backward stability. △

B.3 QR factorization

Another important matrix factorization is the *QR factorization*. It can be used in many important algorithms as an alternative for SVD, and it can be computed in a backward stable way using roughly half as many floating point operations.

Definition B.3.1 (QR factorization)**.** For a matrix $A \in \mathbb{C}^{m\times n}$, a decomposition $A = \mathbf{QR}$ is called a *QR factorization* of A if

1. $\mathbf{Q} \in \mathbb{C}^{m\times m}$ is unitary,
2. $\mathbf{R} \in \mathbb{C}^{m\times n}$ is upper triangular, meaning that $\mathbf{R}_{ij} = 0$ for $i > j$.

Theorem B.3.1 (Existence of a QR factorization)**.** *Every matrix $A \in \mathbb{C}^{m\times n}$ has a QR factorization.*

Proof. The proof follows almost immediately from the Gram-Schmidt orthogonalization process. See [TBI97, Theorem 7.1]. □

If we assume that $m \geq n$ and A has rank n, then a QR decomposition of A can be written as

$$A = \mathbf{QR} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R}_1 \\ 0_{m-n,n} \end{bmatrix} = A = \mathbf{Q}_1\mathbf{R}_1$$

and the diagonal entries of $\mathbf{R}$ can be chosen real and positive. With these constraints, the factorization $Q = \mathbf{Q}_1\mathbf{R}_1$ is unique, and it is called the *reduced QR factorization* of A [TBI97, Theorem 7.2]. As mentioned above, there are backward stable algorithms for computing a QR decomposition in the sense that for the computed matrices $\tilde{\mathbf{Q}}, \tilde{\mathbf{R}}$ we have that

$$\frac{\|A - \tilde{\mathbf{Q}}\tilde{\mathbf{R}}\|_2}{\|A\|_2} = O(u).$$

A good way to go is to use Householder reflectors and/or Givens rotations to systematically 'create zeros' in the matrix A by applying orthogonal transformations until it becomes upper triangular, see for instance [TBI97, Lecture 10], [Hig02, Chapter 19] or [GVL12, Section 5.2].

Remark B.3.1. As noted in Remark B.2.1, the QR decomposition can be used to solve linear systems of equations in a backward stable way [TBI97, Lecture 16]. If $A \in \mathbb{C}^{m\times m}$ is invertible, then the solution of $Ax = b$ can be found via the equivalent system $\mathbf{R}x = \mathbf{Q}^H b$, which can be solved via back substitution. △

If the rank of A is n, then the columns of the matrix $\mathbf{Q}_1$ form an orthonormal basis for the image (or column space) of A. Unfortunately, the assumption on the rank cannot be dropped, not even when we replace 'the columns of $\mathbf{Q}_1$' by 'a subset of the columns of $\mathbf{Q}_1$. This is shown by the following example, taken from [GVL12, Subsection 5.4.1].

Example B.3.1. Consider the matrix

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{with QR factorization} \quad A = \mathbf{QR} = \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

It is clear that no subset of the columns of $\mathbf{Q}$ is a basis for $\operatorname{im} A$. △

A solution for this is given by a generalization of the QR decomposition, in which it is allowed to permute the columns of A [GVL12, Subsection 5.4.1].

Definition B.3.2 (QR factorization with column pivoting)**.** For a matrix $A \in \mathbb{C}^{m\times n}$, a decomposition $A\mathbf{P} = \mathbf{QR}$ of $A\mathbf{P}$ is called a *column pivoted QR factorization* of A if

1. $\mathbf{P}$ is a column permutation matrix. That is, its columns are given by $\mathbf{P}_{:,i} = (\mathrm{id}_m)_{:,\pi(i)}$, for some permutation π in the symmetric group of order m,
2. $\mathbf{Q} \in \mathbb{C}^{m\times m}$ is unitary,
3. $\mathbf{R} \in \mathbb{C}^{m\times n}$ is upper triangular, meaning that $\mathbf{R}_{ij} = 0$ for $i > j$.

Definition B.3.3 (Rank-revealing QR decomposition)**.** For a matrix $A \in \mathbb{C}^{m\times n}$ of rank r, a column pivoted QR decomposition $A\mathbf{P} = \mathbf{QR}$ is a *rank revealing QR decomposition* if

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{bmatrix} \in \mathbb{C}^{m\times n},$$

where $\mathbf{R}_{11} \in \mathbb{C}^{r\times r}$ is upper triangular and invertible and $\mathbf{R}_{12} \in \mathbb{C}^{r\times(n-r)}$.

It can be shown [HP92] that a rank-revealing QR decomposition exists for any matrix $A \in \mathbb{C}^{m\times n}$. A rank revealing QR decomposition has some of the nice properties of the SVD: it reveals the rank r, it gives an orthonormal basis for $\operatorname{im} A$ (these are the first r columns of $\mathbf{Q}$), the rows of $\mathbf{RP}^\top$ form a basis of the row space of A and a basis for $\ker A$ is given by the columns of the $n \times (n-r)$ matrix

$$\mathbf{P} \begin{bmatrix} -\mathbf{R}_{11}^{-1}\mathbf{R}_{12} \\ \mathrm{id}_{n-r} \end{bmatrix},$$

which can be seen from

$$A\mathbf{P}\begin{bmatrix}-\mathbf{R}_{11}^{-1}\mathbf{R}_{12}\\ \mathrm{id}_{n-r}\end{bmatrix} = \mathbf{Q}\begin{bmatrix}\mathbf{R}_{11} & \mathbf{R}_{12}\\ 0_{m-r,r} & 0_{m-r,n-r}\end{bmatrix}\begin{bmatrix}-\mathbf{R}_{11}^{-1}\mathbf{R}_{12}\\ \mathrm{id}_{n-r}\end{bmatrix} = 0. \tag{B.3.1}$$

Example B.3.2. For the matrix in Example B.3.1, a rank-revealing QR decomposition is given by

$$A\mathbf{P} = \begin{bmatrix}1&1&1\\0&0&1\\0&0&1\end{bmatrix}\begin{bmatrix}1&0&0\\0&0&1\\0&1&0\end{bmatrix} = \begin{bmatrix}1&0&0\\0&1&1\\0&1&-1\end{bmatrix}\begin{bmatrix}1&1&1\\0&1&0\\0&0&0\end{bmatrix} = \mathbf{QR}.$$

△

Rank revealing QR decompositions are used for solving rank deficient least squares problems, matrix approximation problems and subset selection problems. See [CH92] for an overview. A naive, brute force approach to the problem of computing a rank-revealing QR decomposition is to try all possible permutations of the columns of A and compute a standard QR decomposition of the permuted matrices. The complexity is, of course, combinatorial. A lot of research has been conducted on finding a heuristic, non-combinatorial algorithm for computing a rank revealing QR decomposition. A first and in many cases effective heuristic for choosing the column permutation $\mathbf{P}$ to yield a rank revealing QR permutation was proposed by Businger and Golub in [BG65]. The columns are pivoted in such a way that, heuristically, the diagonal of $\mathbf{R}_{11}$ contains 'large' elements. This has the effect that the matrix $\mathbf{R}_{11}$ is heuristically well-conditioned. This is important in case one plans, for instance, to use the factorization for a kernel computation as in (B.3.1). Indeed, we have seen that the accuracy with which $\mathbf{R}_{11}^{-1}\mathbf{R}_{12}$ can be computed is governed by the condition number $\kappa(\mathbf{R}_{11})$. We will call the factorization $A\mathbf{P} = \mathbf{QR}$ computed using the column pivoting strategy of [BG65] the *QR decomposition with optimal column pivoting.* Although this strategy of pivoting works quite well in practice, it does not *guarantee* to find a rank-revealing QR decomposition. A well known example by Kahan [Kah66, Example 3.1] shows that it might fail. Other algorithms have been designed to circumvent these problems, see for instance [CH92, HP92, CI94, GE96] and references therein.

B.4 Eigenvalue problems

Next to linear system solving, the *eigenvalue problem* is a fundamental problem in numerical linear algebra. Recall that for $\mathbb{C}$-vector space V and an endomorphism $A : V \to V$, a *right eigenpair* (or simply *eigenpair*) of A is a tuple

$$(\lambda, v) \in \mathbb{C} \times (V \setminus \{0\}) \quad \text{such that} \quad A(v) = \lambda v.$$

Here, we will think of V as $\mathbb{C}^m$ and of A as a matrix in $\mathbb{C}^{m\times m}$, such that an eigenpair is $(\lambda, v) \in \mathbb{C}\times(\mathbb{C}^m\setminus\{0\})$ satisfying $Av = \lambda v$. With this notation, λ is called an *eigenvalue* of A, and v is a corresponding (right) *eigenvector*. The eigenvalues are precisely the roots of the *characteristic polynomial* of A, which is $\chi_A(\lambda) = \det(\lambda \mathrm{id}_m - A)$. One can easily check that $\chi_A(\lambda)$ is monic of degree m, so A has m eigenvalues $\lambda_1, \ldots, \lambda_m$ 'counting multiplicities'. The multiplicity of an eigenvalue λ_i as a root of $\chi_A(\lambda)$ is called the *algebraic multiplicity* of the eigenvalue. For each eigenvalue λ_i, let v_i be a corresponding eigenvector. The equations $Av_i = \lambda_i v_i$ can be arranged into the matrix equation

$$A\mathbf{V} = \mathbf{V}\Delta \quad \text{where} \quad \mathbf{V} = \begin{bmatrix} v_1 & \cdots & v_m \end{bmatrix}, \Delta = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{bmatrix}.$$

If the eigenvectors $v_1, \ldots, v_m$ are linearly independent, this gives the factorization $A = \mathbf{V}\Delta\mathbf{V}^{-1}$.

Definition B.4.1 (Eigenvalue decomposition)**.** For a matrix $A \in \mathbb{C}^{m\times m}$ and an invertible matrix $\mathbf{V} \in \mathbb{C}^{m\times m}$, a decomposition $A = \mathbf{V}\Delta\mathbf{V}^{-1}$ is called an *eigenvalue decomposition* of A if Δ is a diagonal matrix.

It is well known that not every matrix $A \in \mathbb{C}^{m\times m}$ has an eigenvalue decomposition. Those that do are called *nondefective* or *diagonalizable*. These are exactly the matrices for which the algebraic multiplicity of λ_i equals the *geometric multiplicity* of λ_i, which is defined as

$$\dim_{\mathbb{C}} \ker(\lambda_i \mathrm{id}_m - A).$$

If A is nondefective, then the eigenvalue decomposition $A = \mathbf{V}\Delta\mathbf{V}^{-1}$ shows that when represented in the basis corresponding to the eigenvectors $v_1, \ldots, v_m$, A behaves like a diagonal matrix Δ.

Remark B.4.1. A *left eigenpair* of A is a tuple $(w, \lambda) \in (\mathbb{C}^m \setminus \{0\}) \times \mathbb{C}$ such that $w^H A = \lambda w^H$. A vector w coming from such a left eigenpair is called a *left eigenvector*. By definition, (λ, w) is a right eigenpair of A^H if and only if $(w, \overline{\lambda})$ is a left eigenpair of A. Note that if A is nondefective, then $A = \mathbf{V}\Delta\mathbf{V}^{-1}$ gives $A^H = \mathbf{V}^{-H}\Delta^H\mathbf{V}^\top$ where $\mathbf{V}^{-H} = (\mathbf{V}^{-1})^H = (\mathbf{V}^H)^{-1}$. This shows that the left eigenvectors of A are given by the columns of $\mathbf{V}^{-H}$. $\triangle$

Definition B.4.2 (Similarity)**.** A matrix $A \in \mathbb{C}^{m\times m}$ is called *similar* to a matrix $B \in \mathbb{C}^{m\times m}$ if there is an invertible matrix $\mathbf{V} \in \mathbb{C}^{m\times m}$ such that $A = \mathbf{V}B\mathbf{V}^{-1}$.

If A is similar to B, then A and B have the same eigenvalues. Moreover, they occur with the same algebraic and geometric multiplicities [TBI97, Theorem 24.3]. This amounts to saying that the *eigenstructure* of a linear map is independent of the basis in which it is represented. A transformation $A \to \mathbf{V}^{-1}A\mathbf{V}$ is called a *similarity transformation.*

An important observation related to algorithms for computing the eigenvalues of a general matrix $A^{m\times m}$ is that *any such algorithm must be of an iterative nature.* By this we mean that the algorithm may iteratively compute better and better approximations of the eigenvalues, but it can never, even in exact arithmetic, compute the eigenvalues in finite time. This is prohibited by the famous Abel-Ruffini theorem which states that there is no general expression in radicals for the roots of a (univariate) polynomial of degree 5 or higher. By the fact that the univariate root finding problem can be translated to an eigenvalue problem (see Example 3.1.1), the existence of a direct algorithm for computing eigenvalues would contradict this theorem. The key idea of many of the most successful eigenvalue solvers is to *apply a sequence of similarity transformations* to the matrix A such that the result converges to a structured matrix from which we can read off the eigenvalues. An example would be to choose invertible matrices $\mathbf{V}_1, \mathbf{V}_2, \dots$ in such a way that the sequence

$$A \to \mathbf{V}_1^{-1} A \mathbf{V}_1 \to \mathbf{V}_2^{-1}\mathbf{V}_1^{-1} A \mathbf{V}_1 \mathbf{V}_2 \to \cdots$$

converges to a diagonal matrix Δ. In every step of the sequence, the eigenstructure is maintained, and after sufficiently many (say k) steps, numerical approximations of the eigenvalues can be read off the diagonal of $\mathbf{V}_k \cdots \mathbf{V}_1 A \mathbf{V}_1^{-1} \cdots \mathbf{V}_k^{-1}$. Although this illustrates the idea, this is not what is usually done in practice. There are two problems with this approach. First of all, we have seen that not every matrix is diagonalizable. Secondly, if some of the $\mathbf{V}_i$ along the way are ill-conditioned, the result will be contaminated by rounding errors. This indicates that we need a different matrix factorization which exists for all matrices $A \in \mathbb{C}^{m\times m}$, reveals the eigenstructure of A and, preferably, is such that it can be computed (approximately) by applying *unitary* similarity transformations of the form $\mathbf{U}^{-1} A \mathbf{U}$. This is where the *Schur factorization* comes into play.

Definition B.4.3 (Schur decomposition)**.** For a matrix $A \in \mathbb{C}^{m\times m}$, a decomposition $A = \mathbf{U}\mathbf{T}\mathbf{U}^H$ is called a *Schur decomposition* if $\mathbf{T}$ is upper triangular and $\mathbf{U}$ is unitary.

It is clear that if $A = \mathbf{U}\mathbf{T}\mathbf{U}^H = \mathbf{U}\mathbf{T}\mathbf{U}^{-1}$ is a Schur decomposition, then A is similar to $\mathbf{T}$ and $\mathbf{T}$ has the eigenvalues of A on its diagonal.

Theorem B.4.1. *Every matrix $A \in \mathbb{C}^{m\times m}$ has a Schur decomposition.*

Proof. See [GVL12, Theorem 7.1.3] or [TBI97, Theorem 24.9]. □

Remark B.4.2. In the special case where A is *normal* (i.e. $A^H A = A A^H$), the Schur decomposition coincides with the eigenvalue decomposition. That is, $A = \mathbf{U}\mathbf{T}\mathbf{U}^H = \mathbf{V}\Delta\mathbf{V}^{-1}$ where T is diagonal and $\mathbf{V}$ is unitary. For this reason, normal matrices are especially nice for eigenvalue decompositions. They can be diagonalized by a *unitary* similarity transformation or, equivalently, they have an orthogonal set of m eigenvectors. △

The columns of $\mathbf{U}$ in the Schur factorization $A = \mathbf{U}\mathbf{T}\mathbf{U}^H$ are called the *Schur vectors* of A. The matrix $\mathbf{U}$ can be chosen such that the eigenvalues appear on the diagonal of $\mathbf{T}$ in *any order* [GVL12, Theorem 7.1.3]. If they are ordered in such a way that

$$\mathbf{T} = \begin{bmatrix} \mathbf{R}_{11} & \cdots & \mathbf{R}_{1k} \\ & \ddots & \vdots \\ & & \mathbf{R}_{kk} \end{bmatrix}$$

where $\mathbf{R}_{ii}$ is an upper triangular matrix with only one distinct eigenvalue and the eigenvalue of $\mathbf{R}_{ii}$ is different from the eigenvalue of $\mathbf{R}_{jj}$ for all $j \neq i$, then there is a similarity transformation $\mathbf{V}^{-1}\mathbf{T}\mathbf{V} = \tilde{\Delta}$ such that

$$\tilde{\Delta} = \begin{bmatrix} \mathbf{R}_{11} & & \\ & \ddots & \\ & & \mathbf{R}_{kk} \end{bmatrix}$$

is a block diagonal matrix with upper triangular blocks $\mathbb{R}_{ii} \in \mathbb{C}^{\mu_i \times \mu_i}$ on its diagonal [GVL12, Theorem 7.1.6]. If we write $\mathbf{V}_1$ for the submatrix of $\mathbf{UV}$ given by its first μ_1 columns and $V_1 = \operatorname{im} \mathbf{V}_1$, we see that $A\mathbf{V}_1 = \mathbf{V}_1\mathbf{R}_{11}$, and thus for any $v \in V_1, Av \in V_1$. For this reason, V_1 is called an *invariant subspace of A*. We obtain k invariant subspaces $V_1, \ldots, V_k$ in this way, of dimensions $\mu_1, \ldots, \mu_k$ respectively.

Most general purpose eigenvalue solvers proceed by approximating the Schur decomposition of A by a sequence of unitary similarity transformations

$$A \to \mathbf{U}_1^H A \mathbf{U}_1 \to \mathbf{U}_2^H \mathbf{U}_1^H A \mathbf{U}_1 \mathbf{U}_2 \to \cdots$$

which converges to an upper triangular matrix $\mathbf{T}$. In this process, usually A is first brought into so-called *upper-Hessenberg* form $A \to H$, which takes only finitely many similarity transformations, such that the remaining similarity transformations can exploit this upper-Hessenberg structure. The step $A \to H$ has complexity $O(m^3)$. One similarity transformation on H takes $O(m^2)$ floating point operations, and usualy $O(m)$ transformations are needed to reach convergence. In total, this makes the complexity of the step $H \to T$ equal to $O(m^3)$. The overall complexity is thus $O(m^3)$ as well. The method is backward stable, in the sense that with enough iterations, the computed matrices $\tilde{\mathbf{U}}, \tilde{\mathbf{T}}$ are such that

$$\frac{\|A - \tilde{\mathbf{U}}\tilde{\mathbf{T}}\tilde{\mathbf{U}}^H\|_2}{\|A\|_2} = O(u).$$

For more details, the reader can consult [TBI97, Part V] or [GVL12, Chapter 7].

We should mention that a different type of techniques based on *Krylov subspace iteration* is powerful for solving large, sparse eigenvalue problems [TBI97, Lectures 33, 34, 36]. Also, there are several variants of the eigenvalue problem which are important in applications. Here are a few examples.

1. The *generalized eigenvalue problem* consists of computing $(\lambda, v) \in \mathbb{C} \times (\mathbb{C}^m \setminus \{0\})$ such that $Av = \lambda Bv$ for $A, B \in \mathbb{C}^{m \times m}$. A backward stable algorithm is given

by the famous QZ algorithm, which computes a generalization of the Schur decomposition [GVL12, Section 7.7].

2. The *nonlinear eigenvalue problem* consists of computing $(\lambda, v) \in \Omega \times (\mathbb{C}^m \setminus \{0\})$ with $\Omega \subset \mathbb{C}$ some compact domain such that $F(\lambda)v = 0$ for a matrix valued analytic function $F : \Omega \to \mathbb{C}^{m \times m}$. An important subclass of problems is given by the case where $F = A_0 + A_1 x + \cdots + A_d x^d$ is a polynomial with matrix coefficients. These problems are commonly solved via linearization or contour integration techniques, see [GT17] for a modern overview. This is the kind of eigenvalue problem that is encountered when solving a polynomial system via the *hidden variable resultant method*. See Subsection 3.4.2 for a brief discussion and references.

3. The *multi-parameter eigenvalue problem* can be formulated as follows. Given $A_{ij} \in \mathbb{C}^{m_i \times m_i}, 1 \leq i \leq n, 0 \leq j \leq n$, find $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ and $v_i \in (\mathbb{C}^{m_i} \backslash \{0\}), i = 1, \ldots, n$ such that

$$\begin{aligned}
(A_{10} + A_{11}\lambda_1 + \cdots + A_{1n}\lambda_n)v_1 &= 0,\\
(A_{20} + A_{21}\lambda_1 + \cdots + A_{2n}\lambda_n)v_2 &= 0,\\
&\vdots\\
(A_{n0} + A_{n1}\lambda_1 + \cdots + A_{nn}\lambda_n)v_n &= 0.
\end{aligned}$$

The classical method for solving such a problem is given by the *Delta-method* [Atk72]. More recently, methods based on homotopy continuation have been developed [DYY16, RLY18]. It turns out that the problem of solving polynomial systems may be reformulated as a multi-parameter eigenvalue problem. This observation was used in the bivariate setting in [PH16, BvDD$^+$17].

Appendix C

Error measures

The goal of this appendix is to describe and motivate the way that the quality of an approximate solution for a (Laurent) polynomial system is assessed in this thesis. Let $\mathbb{C}[M] = \mathbb{C}[t_1^{\pm 1}, \ldots, t_n^{\pm 1}]$ be the ring of Laurent polynomials in n variables and consider s nonzero elements $\hat{f}_1, \ldots, \hat{f}_s \in \mathbb{C}[M]$. We denote

$$\hat{f}_i = \sum_{a \in \mathbb{Z}^n} c_{i,a} t^a.$$

Let $\tilde{z} \in (\mathbb{C}^*)^n$ be a numerical approximation for a solution of $\hat{f}_1 = \cdots = \hat{f}_s = 0$. In the context of the first four chapters of this thesis, where the $\hat{f}_i$ are polynomials in $\mathbb{C}[t_1, \ldots, t_n] \subset \mathbb{C}[M]$, we will allow the coordinates of $\tilde{z}$ to be zero. We come back to this later.

A first observation is that by 'solving' the system, in this thesis we usually mean finding approximations of *all* solutions. As discussed in Section B.1, the best way to assess the quality of the result of a numerical computation is by measuring the *relative backward error*. This means that we should find a way to 'measure' the distance of the (Laurent) polynomial system $\hat{f}_1 = \ldots = \hat{f}_s = 0$ to a system $\tilde{f}_1 = \cdots = \tilde{f}_s = 0$ for which *all* our computed solutions are *all exact* solutions. Suppose $\{\tilde{z}_1, \ldots, \tilde{z}_\delta\}$ is a set of approximate solutions for $\hat{f}_1 = \ldots = \hat{f}_s = 0$. We could think of $(\hat{f}_1, \ldots, \hat{f}_s)$ as a generic member of some family $\mathcal{F}$ of systems, such that it has the generic number δ of solutions of that family, and measure (according to some metric on $\mathcal{F}$) the distance of $(\hat{f}_1, \ldots, \hat{f}_s)$ to a different system $(\tilde{f}_1, \ldots, \tilde{f}_s) \in \mathcal{F}$ for which $\tilde{f}_i(\tilde{z}_j) = 0$ for all i and j. Here's an example which shows that this is too ambitious.

Example C.0.1. Let $(\hat{f}_1, \hat{f}_2) \in \mathcal{F}_R(3,3)$ where $R = \mathbb{C}[t_1, t_2] \subset \mathbb{C}[t_1^{\pm 1}, t_2^{\pm 1}]$. We assume that $(\hat{f}_1, \hat{f}_2)$ is generic in the sense of Bézout's theorem (Theorem 3.1.2), which means that there are 9 solutions $V_{\mathbb{C}^2}(\hat{f}_1, \hat{f}_2) = \{z_1, \ldots, z_9\}$. However, if we perturb the points in $V_{\mathbb{C}^2}(\hat{f}_1, \hat{f}_2)$ just a little bit to obtain (possibly very good) approximations

$\tilde{z}_1, \ldots, \tilde{z}_9$ for the solutions, there is generically *no* member of $\mathcal{F}_R(3,3)$ whose solutions are $\{\tilde{z}_1, \ldots, \tilde{z}_9\}$. The reason is that the 9 intersection points of two general cubics are special: they make all maximal minors of a 10×9 bivariate Vandermonde matrix vanish. It is highly unlikely that the numerical solutions computed by some numerical algorithm land on this subvariety of $(\mathbb{C}^2)^9$. $\triangle$

There are some special families for which the observation in Example C.0.1 does not really pose a problem. An example is given by the families $\mathcal{F}_{\mathbb{C}[t]}(d)$ of univariate polynomials of degree at most d. Surprisingly enough, finding good measures for the backward error of an approximate set of roots of a univariate polynomial is still a topic of research today [MVD15, TVB20, TTVB20]. Another example is $\mathcal{F}_{\mathbb{C}[t_1,t_2]}(2,2)$, the family of systems given by two quadratic equations in two variables. Almost all configurations of four points in $\mathbb{C}^2$ are the variety of a member of $\mathcal{F}_{\mathbb{C}[t_1,t_2]}(2,2)$.

Because of this issue, instead of computing the relative backward error for a set of solutions we will limit ourselves to computing it for each approximate solution $\tilde{z}$ individually. The idea is to compute (Laurent) polynomials $\Delta\hat{f}_i$ with ‘small’ coefficients such that the perturbed functions $\tilde{f}_i = \hat{f}_i + \Delta\hat{f}_i$ satisfy $\tilde{f}_i(\tilde{z}) = (\hat{f}_i + \Delta\hat{f}_i)(\tilde{z}) = 0$, $i = 1, \ldots, s$. The relative backward error will be a measure of the *size* of the coefficients of $\Delta\hat{f}_i$, relative to the size of the coefficients of $\hat{f}_i$. Let us now make this precise. We look for polynomials of the form

$$\Delta\hat{f}_i = \sum_{c_{i,a} \neq 0} \Delta c_{i,a} t^a = \sum_{c_{i,a} \neq 0} \varepsilon_{i,a} c_{i,a} t^a,$$

where the sum ranges over all a such that $c_{i,a} \neq 0$, such that the parameters $\varepsilon_{i,a}$ have small modulus and $(\hat{f}_i + \Delta\hat{f}_i)(\tilde{z}) = 0$. Note that

$$|\varepsilon_{i,a}| = \frac{|\Delta c_{i,a}|}{|c_{i,a}|}$$

is the *relative* size of the perturbation on the coefficient $c_{i,a}$. A possible measure for the relative backward error of $\tilde{z}$ is

$$\begin{aligned} r(\tilde{z}) = \quad \min_{\varepsilon \in \mathbb{C}^m} \qquad & \frac{1}{s}\|\varepsilon\|_1 = \frac{1}{s}\sum_{i=1}^{s} \sum_{c_{i,a} \neq 0} |\varepsilon_{i,a}|, \\ \text{subject to} \qquad & \hat{f}_i(\tilde{z}) + \sum_{c_{i,a} \neq 0} \varepsilon_{i,a} c_{i,a} \tilde{z}^a = 0, \quad i = 1, \ldots, s, \end{aligned} \tag{C.0.1}$$

Where m denotes the total number of parameters $\varepsilon_{i,a}$. The complex optimization problem C.0.1 may seem hard to solve at first sight. Fortunately, it is not. For $i = 1, \ldots, s$ let b_i be such that

$$|c_{i,b_i} \tilde{z}^{b_i}| = \max_{c_{i,a} \neq 0} |c_{i,a} \tilde{z}^a|.$$

Note that

$$\varepsilon_{i,b_i} = \frac{-\hat{f}_i(\tilde{z})}{c_{i,b_i} \tilde{z}^{b_i}}, \quad \varepsilon_{i,a} = 0, a \neq b_i, \quad i = 1, \ldots, s$$

satisfies the constraint of (C.0.1) and gives

$$\frac{1}{s}\|\varepsilon\|_1 = \frac{1}{s}\sum_{i=1}^{s}\left|\frac{\hat{f}_i(\tilde{z})}{c_{i,b_i}\tilde{z}^{b_i}}\right| = \frac{1}{s}\sum_{i=1}^{s}\frac{|\hat{f}_i(\tilde{z})|}{\max_{c_{i,a}\neq 0}|c_{i,a}\tilde{z}^a|}.$$

This immediately leads to the upper bound

$$r(\tilde{z}) \leq \frac{1}{s}\sum_{i=1}^{s}\frac{|\hat{f}_i(\tilde{z})|}{\max_{c_{i,a}\neq 0}|c_{i,a}\tilde{z}^a|}.$$

We now prove that this is also a lower bound. Collecting the $c_{i,a}\tilde{z}^a$ in a vector v and the $\varepsilon_{i,a}$ in a subvector ε_i of ε, the constraint of (C.0.1) can be written as $v^\top \varepsilon_i = -\hat{f}_i(\tilde{z})$. By submultiplicativity of the matrix 1-norm we get

$$\|v^\top\|_1\|\varepsilon_i\|_1 \geq |\hat{f}_i(\tilde{z})|.$$

Making use of the fact that the 1-norm of a matrix is its maximal absolute column sum, we get that $\|v^\top\|_1 = \|v\|_\infty = \max_{c_{i,a}\neq 0}|c_{i,a}\tilde{z}^a|$. We conclude that

$$r(\tilde{z}) = \frac{1}{s}\sum_{i=1}^{s}\frac{|\hat{f}_i(\tilde{z})|}{\max_{c_{i,a}\neq 0}|c_{i,a}\tilde{z}^a|}.$$

This confirms the intuition that the residual can be measured by evaluating the $\hat{f}_i$ at $\tilde{z}$ and checking 'how zero' the result actually is, *relative to the size of the terms in the sum.* The following chain of inequalities now follows trivially:

$$\frac{1}{s}\sum_{i=1}^{s}\frac{|\hat{f}_i(\tilde{z})|}{\sum_{c_{i,a}\neq 0}|c_{i,a}\tilde{z}^a|} \leq r(\tilde{z}) \leq \frac{1}{s}\sum_{i=1}^{s}\frac{m_i|\hat{f}_i(\tilde{z})|}{\sum_{c_{i,a}\neq 0}|c_{i,a}\tilde{z}^a|}, \tag{C.0.2}$$

with m_i the number of nonzero terms in $\hat{f}_i$. In practice, the lower- and upper bound for $r(\tilde{z})$ in (C.0.2) are of the same order of magnitude. This means that they are as good an indication of the relative backward error as $r(\tilde{z})$.

In the case where the $\hat{f}_i$ are polynomials or have solutions in $(\mathbb{C}^*)^n$ with coordinates that are very close to zero, the terms of $\hat{f}_i(\tilde{z})$ may become very small such that taking the relative error gives awkward results. Consider for instance the case where $\hat{f}_1 = t_1$ and $\hat{f}_2 = t_2$. For the approximate solution $\tilde{z} = (10^{-16}, 10^{-16})$ of $\hat{f}_1 = \hat{f}_2 = 0$, the lower bound in (C.0.2) evaluates to 1 (so does $r(\tilde{z})$ and the upper bound in (C.0.2)). However, $\tilde{z}$ seems like a perfectly acceptable numerical approximation of the actual solution $(0, 0)$. To avoid this kind of situations, we use a slightly modified version of the lower bound in (C.0.2) to compute the residual.

Definition C.0.1 (Residual)**.** For $\hat{f}_1, \ldots, \hat{f}_s \in \mathbb{C}[M]$ and $\tilde{z} \in (\mathbb{C}^*)^n$, we define the *residual* of $\tilde{z}$ as a solution of $\hat{f}_1 = \cdots = \hat{f}_s = 0$ as

$$\frac{1}{s}\sum_{i=1}^{s}\frac{|\hat{f}_i(\tilde{z})|}{\sum_{c_{i,a}\neq 0}|c_{i,a}\tilde{z}^a| + 1}.$$

The term '+1' in the denominator of the residual in Definition C.0.1 makes the criterion a *mixed criterion*: it depends on the magnitude of $\sum_{c_{i,a}\neq 0} |c_{i,a}\tilde{z}^a|$ whether it behaves like a relative or an absolute measure. We note that Definition C.0.1 is the measure for the residual that was used in [TVB18, TMVB18, MTVB19, Tel20].

Appendix D

Polytopes, cones and fans

The algebraic and geometric properties of a normal toric variety are encoded in the combinatorics of the associated fan. As a consequence, polyhedral geometry is an important tool for studying certain families of polynomial systems. In this appendix, we recall the basic properties of polytopes, cones and fans that are relevant in this context. All of what is discussed here and more can be found in [CLS11, Sections 1.2, 2.2, 2.3, 3.1], where some results are stated without proof but their implications in toric geometry are highlighted. See also [Ful93, Sections 1.2, 1.4, 1.5] and [Oda89, Appendix A]. A nice introduction to convex (lattice) polytopes can be found in [CLO06, Chapter 7, §1] and its exercises. A standard reference for convex polytopes in a much more general context is [Grü13].

D.1 Polytopes

Let $M \simeq \mathbb{Z}^n$ be an n-dimensional lattice (by a *lattice* we mean a free abelian group of finite rank) and let $M_\mathbb{R} = M \otimes_\mathbb{Z} \mathbb{R} \simeq \mathbb{R}^n$ be the associated real vector space. The dual lattice of M is $N = \mathrm{Hom}_\mathbb{Z}(M, \mathbb{Z}) \simeq \mathbb{Z}^n$ and the dual vector space of $M_\mathbb{R}$ is $(M_\mathbb{R})^\vee = \mathrm{Hom}_\mathbb{R}(M_\mathbb{R}, \mathbb{R}) = N \otimes_\mathbb{Z} \mathbb{R} = N_\mathbb{R} \simeq \mathbb{R}^n$. We denote

$$\langle \cdot, \cdot \rangle : N_\mathbb{R} \times M_\mathbb{R} \to \mathbb{R}, (u, m) \mapsto \langle u, m \rangle$$

for the natural pairing between $N_\mathbb{R}$ and $M_\mathbb{R}$ and its restriction to the lattice $N \times M \to \mathbb{Z}$. This is the usual dot product on $\mathbb{R}^n$ and its restriction to $\mathbb{Z}^n$.

For a subset $\mathscr{A} \subset V$ of an $\mathbb{R}$-vector space V, the *convex hull* of $\mathscr{A}$, denoted by $\mathrm{Conv}(\mathscr{A})$, is the set of sums $\sum_{m \in \mathscr{A}} c_m m$ where the coefficients $c_m \in \mathbb{R}_{\geq 0}$ are nonnegative, finitely many c_m are nonzero and $\sum_{m \in \mathscr{A}} c_m = 1$.

Definition D.1.1 (Polytope)**.** A *polytope* P in $M_{\mathbb{R}}$ is the convex hull of a finite set of points $\mathscr{A} = \{m_1, \dots, m_k\}$ in $M_{\mathbb{R}}$:

$$P = \operatorname{Conv}(\mathscr{A}) = \left\{ \sum_{i=1}^{k} c_i m_i \in M_{\mathbb{R}} \mid c_i \in \mathbb{R}, \sum_{i=1}^{k} c_i = 1, c_i \geq 0 \right\} \subset M_{\mathbb{R}}.$$

If $\mathscr{A} \subset M$, P is called a *lattice polytope*.

Note that we define a polytope to be convex. The reason is that we will not encounter non-convex polytopes in this text. The dimension $\dim P$ of a polytope $P \subset M_{\mathbb{R}}$ is defined as the dimension of the smallest affine subspace of $M_{\mathbb{R}}$ containing P. A polytope in $M_{\mathbb{R}}$ is said to be *full-dimensional* if $\dim P = n$.

A point $u \in N_{\mathbb{R}} \backslash \{0\}$ and a scalar $a \in \mathbb{R}$ give a hyperplane

$$H_{u,a} = \{m \in M_{\mathbb{R}} : \langle u, m \rangle + a = 0\}$$

and a closed half-space

$$H_{u,a}^+ = \{m \in M_{\mathbb{R}} : \langle u, m \rangle + a \geq 0\}.$$

Definition D.1.2 (Faces of a polytope)**.** Take $u \in N_{\mathbb{R}} \backslash \{0\}$, $a \in \mathbb{R}$ and let $P \subset M_{\mathbb{R}}$ be a convex polytope, the set $H_{u,a} \cap P$ is a *face* of P if $P \subset H_{u,a}^+$ and $a = -\min_{m \in P} \langle u, m \rangle$. We say that P is a face of P by convention.

A face Q of a polytope is again a polytope, so what we mean by the *dimension* $\dim Q$ of Q should be clear. The *codimension* of a face $Q \subset P$ is $\dim P - \dim Q$. A face of codimension 1 in P is called a *facet*, a face of dimension 1 is an *edge* and a face of dimension 0 is a *vertex*. A hyperplane $H_{u,a}$ for which $H_{u,a} \cap P$ is a face of P is called a *supporting hyperplane*. Any polytope can be expressed as the intersection of finitely many closed half-spaces $H_{u,a}^+$ associated to supporting hyperplanes. That is, any polytope $P \subset M_{\mathbb{R}}$ can be written as

$$P = H_{u_1,a_1}^+ \cap \cdots \cap H_{u_k,a_k}^+ = \{m \in M_{\mathbb{R}} \mid \langle u_i, m \rangle + a_i \geq 0, i = 1, \dots, k\} \qquad \text{(D.1.1)}$$

for some $u_1, \dots, u_k \in N_{\mathbb{R}}$, $a_1, \dots, a_k \in \mathbb{R}$. We collect the vectors $u_1, \dots, u_k$ in a matrix $F = [u_1 \ \cdots \ u_k] \in \mathbb{R}^{n \times k}$ (we identify $N_{\mathbb{R}}$ with $\mathbb{R}^n$) and the numbers $a_1, \dots, a_k$ in a vector $a \in \mathbb{R}^k$ to use the short notation

$$P = \{m \in M_{\mathbb{R}} \mid \langle u_i, m \rangle + a_i \geq 0, i = 1, \dots, k\} = \{m \in M_{\mathbb{R}} \mid F^\top m + a \geq 0\}. \quad \text{(D.1.2)}$$

The representation in equations (D.1.1), (D.1.2) is called a *half-space representation* or *H-representation*[1] of the polytope P. There exist infinitely many different H-representations for any polytope. However, if P is full-dimensional, there exists an

[1] A different representation of a polytope using finitely many data is given by a list of its vertices. This is called a *vertex representation* or *V-representation*. This is important for computational purposes, but H-representations are more important in the context of this thesis.

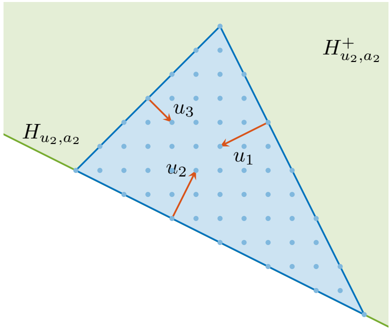


Figure D.1: Illustration of a lattice polytope of dimension 2 and its primitive inward pointing facet normals.

essentially unique, *minimal* H-representation of P, in the sense that it consists of a minimal number k of inequalities where the inequalities are uniquely defined up to multiplication with a nonzero scalar. Suppose that P is full-dimensional. For a supporting hyperplane $H_{u,a}$ corresponding to a facet Q of P, the vector u is uniquely determined up to a nonzero scalar factor. For every facet Q, let u_Q, a_Q be such that $P \subset H^+_{u_Q,a_Q}$, $H_{u_Q,a_Q} \cap P = Q$. The minimal H-representation of P is given by

$$P = \bigcap_{Q \text{ facet of } P} H^+_{u_Q,a_Q}.$$

If P is a full-dimensional lattice polytope, then for any facet $Q \subset P$, u_Q can be chosen in a unique way as the generator of the sublattice

$$\{u \in N \mid \langle u, m\rangle = 0 \text{ for all } m \in Q\}$$

for which $P \in H^+_{u_Q,a_Q}$. This is called the *primitive, inward pointing facet normal* of Q. Geometrically, it is the inward pointing integer vector perpendicular to Q of the smallest length. In the following, by 'the facet normal' associated to Q we mean the primitive, inward pointing facet normal.

Example D.1.1. Figure D.1 shows a full-dimensional polytope in $\mathbb{R}^2$ (a 2-dimensional polytope is also called a *polygon*) together with its interior lattice points and primitive inward pointing facet normals. The matrix F corresponding to the minimal H-representation for this example is given by

$$F = \begin{bmatrix} -2 & 1 & 1 \\ -1 & 2 & -1 \end{bmatrix} = [u_1 \; u_2 \; u_3].$$

The supporting hyperplane H_{u_2,a_2} is also shown in Figure D.1, and its corresponding half-space $H^+_{u_2,a_2}$ (shaded in green) contains the polytope. We note that, strictly speaking, the orange arrows do not belong in the same picture: they live in the dual plane $(\mathbb{R}^2)^\vee$. However, the figure may give some geometric intuition. $\triangle$

We will need to define a few operations on polytopes. For any polytope $P \subset M_{\mathbb{R}}$ and any $\lambda \in \mathbb{R}, \lambda \geq 0$, we define the polytope λP as $\lambda P = \{\lambda p : p \in P\}$. This is called a *dilation* of the polytope P and all dilations are obtained by restricting scalar multiplication in $M_{\mathbb{R}}$ to P. Somewhat less familiar is the binary operation of 'adding polytopes' together.

Definition D.1.3 (Minkowski sum)**.** Let P and Q be polytopes in $M_{\mathbb{R}}$. The Minkowski sum of P and Q is

$$P + Q = \{p + q : p \in P, q \in Q\} \subset M_{\mathbb{R}}.$$

Definition D.1.4. The *n-dimensional volume* of a polytope $P \subset \mathbb{R}^n$ with coordinates $x_1, \ldots, x_n$ on $\mathbb{R}^n$ is defined as

$$\mathrm{Vol}_n(P) = \int \cdots \int_P 1 \, dx_1 \cdots dx_n.$$

Theorem D.1.1. *Given the collection $P_1, \ldots, P_\ell$ of polytopes in $\mathbb{R}^n$, the function*

$$f(\lambda_1, \ldots, \lambda_\ell) = \mathrm{Vol}_n(\sum_{i=1}^{\ell} \lambda_i P_i)$$

is a homogeneous polynomial of degree n in the λ_i.

Proof. See [CLO06, Chapter 7, §4, Proposition 4.9]. □

In the case where $\ell = n$, one coefficient of the homogeneous polynomial of Theorem D.1.1 is of special interest to us, for reasons that are given in Section 5.1.

Definition D.1.5 (Mixed volume)**.** The n-dimensional *mixed volume* of a collection of n polytopes $P_1, \ldots, P_n$ in $\mathbb{R}^n$, denoted $\mathrm{MV}(P_1, \ldots, P_n)$, is the coefficient of the monomial $\lambda_1 \lambda_2 \cdots \lambda_n$ in $\mathrm{Vol}_n(\sum_{i=1}^{n} \lambda_i P_i)$.

There are several different formulas for the mixed volume $\mathrm{MV}(P_1, \ldots, P_n)$, although not all of them are useful for computational purposes. State of the art implementations use the characterization of the mixed volume as the sum of the volumes of the mixed cells in a mixed subdivision of $P_1 + \cdots + P_n$ [HS95, EC95]. An interesting formula for the case $n = 2$ is given by

$$\mathrm{MV}(P_1, P_2) = \mathrm{Vol}_2(P_1 + P_2) - \mathrm{Vol}_2(P_1) - \mathrm{Vol}_2(P_2). \tag{D.1.3}$$

D.2 Polyhedral cones

For a subset $\mathscr{A} \subset V$ of an $\mathbb{R}$-vector space V, the *cone over V*, denoted by $\mathrm{Cone}(\mathscr{A})$, is the set of finite sums $\sum_{u \in \mathscr{A}} \lambda_u u$ with $\lambda_u \in \mathbb{R}_{\geq 0}$.

Definition D.2.1 (Convex polyhedral cone)**.** A convex polyhedral cone (CPC) in a finite dimensional $\mathbb{R}$-vector space V is a subset of the form

$$\sigma = \text{Cone}(\mathscr{A}) = \left\{ \sum_{u \in \mathscr{A}} \lambda_u u : \lambda_u \in \mathbb{R}_{\geq 0} \right\} \subset V,$$

where $\mathscr{A} \subset V$ is finite. We say that σ is *generated* by $\mathscr{A}$. By definition, $\text{Cone}(\varnothing) = \{0\}$.

The *dimension* of a CPC is the dimension of the smallest affine subspace containing it. Convex polyhedral cones are the only type of cones we work with in this thesis, which is why sometimes we refer to them simply as *cones*. Our cones will live in the vector spaces $N_{\mathbb{R}}$ and $M_{\mathbb{R}}$ related to the lattices N and M as defined in Section D.1. For any CPC $\sigma \subset N_{\mathbb{R}}$, its *dual cone* $\sigma^\vee \subset M_{\mathbb{R}}$ is defined as

$$\sigma^\vee = \{m \in M_{\mathbb{R}} \mid \langle u, m \rangle \geq 0, \forall u \in \sigma\}.$$

One can check that the dual cone is indeed a cone and $(\sigma^\vee)^\vee = \sigma$. As suggested by this notation, 'dual cones' (who are themselves cones) live in $M_{\mathbb{R}}$, i.e. in the context of cones we think of $M_{\mathbb{R}}$ as the dual space. Note that for polytopes, this was the other way around. This convention is motivated by toric geometry (see Appendix E).

Hyperplanes and half-spaces in $N_{\mathbb{R}}$ are defined just like in $M_{\mathbb{R}}$. A point $m \in M_{\mathbb{R}} \backslash \{0\}$ and a scalar $a \in \mathbb{R}$ give a hyperplane

$$H_{m,a} = \{u \in N_{\mathbb{R}} : \langle u, m \rangle + a = 0\}$$

and a closed half-space

$$H^+_{m,a} = \{u \in N_{\mathbb{R}} : \langle u, m \rangle + a \geq 0\}.$$

Just like a polytope, a cone is a finite intersection of finitely many closed half-spaces.

Definition D.2.2 (Faces of a cone)**.** Take $m \in M_{\mathbb{R}} \backslash \{0\}$ and let $\sigma \subset V$ be a CPC, the set $\tau = H_{m,0} \cap \sigma$ is a *face* of σ if $\sigma \subset H^+_{m,0}$. By convention, the cone σ is regarded as a face of itself.

One shows that for a CPC σ, every face of σ is a CPC, an intersection of faces is again a face and a face of a face is a face. *Rays* and *facets* of σ are faces of dimension 1 and codimension 1 in σ respectively.

Definition D.2.3 (Strong convexity)**.** A CPC σ is called *strongly convex* or *pointed* if $\sigma \cap (-\sigma) = \{0\}$.

The fact that we are working with cones in $N_{\mathbb{R}}$ and $M_{\mathbb{R}}$ suggests that we will be mainly interested in cones that interact nicely with the lattices $N \subset N_{\mathbb{R}}$ and $M \subset M_{\mathbb{R}}$. This is indeed the case. *Rational polyhedral cones* are to CPCs what lattice polytopes are to convex polytopes.

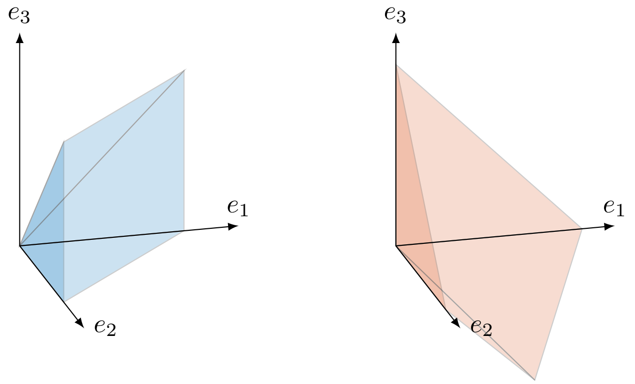


Figure D.2: Left: a rational polyhedral cone σ in $\mathbb{R}^3$. Right: its dual cone $\sigma^\vee$.

Definition D.2.4. A set $\sigma \subset N_\mathbb{R}$ is a *rational polyhedral cone* if $\sigma = \mathrm{Cone}(\mathscr{A})$ for a finite set $\mathscr{A} \subset N$.

A rational polyhedral cone $\sigma^\vee \in M_\mathbb{R}$ gives the subset $\mathsf{S}_\sigma = \sigma^\vee \cap M \subset M$. The set S_σ inherits some algebraic structure from the lattice: it is closed under the (associative and commutative) binary operation '+' and it contains its identity element 0. In other words, S_σ is a commutative monoid. For any finite subset $\mathscr{A} \subset M$ we get the submonoid

$$\mathbb{N}\mathscr{A} = \{\sum_{m\in\mathscr{A}} c_m m \mid c_m \in \mathbb{N}\} \subset M.$$

A subset $\mathsf{S} \subset M$ is called an *affine semigroup* if it arises in this way, i.e., if $\mathsf{S} = \mathbb{N}\mathscr{A}$ for some finite subset $\mathscr{A} \subset M$.

Lemma D.2.1 (Gordan's Lemma). *If $\sigma \subset N_\mathbb{R}$ is a rational polyhedral cone, then $\mathsf{S}_\sigma = \sigma^\vee \cap M \subset M$ is an affine semigroup.*

Proof. See [CLS11, Proposition 1.2.17]. □

Example D.2.1. In Figure D.2 a rational polyhedral cone σ and its dual are depicted. The cone σ is generated by $\{(1,0,0),(0,1,0),(1,0,1),(0,1,1)\}$. The dual cone is generated by $\{(1,0,0),(0,1,0),(0,0,1),(1,1,-1)\}$. Each element of these finite sets generates a ray. △

Example D.2.2. A full-dimensional polytope $P \subset M_\mathbb{R}$ gives rise to some full dimensional cones in the following way. For each vertex $v_i \in P$, we translate P by adding the point $-v_i$ to obtain the polytope $P_i - v_i$. We denote $\sigma_i^\vee = \mathrm{Cone}(P_i - v_i) \subset M_\mathbb{R}$. If P is a lattice polytope, then all the cones $\sigma_i^\vee$ obtained in this way are rational polyhedral cones. An example for the polygon from Example D.1.1 is shown in Figure D.3 with $v_1 = H_{u_1,a_1} \cap H_{u_3,a_3}, v_2 = H_{u_1,a_1} \cap H_{u_2,a_2}, v_3 = H_{u_2,a_2} \cap H_{u_3,a_3}$. △

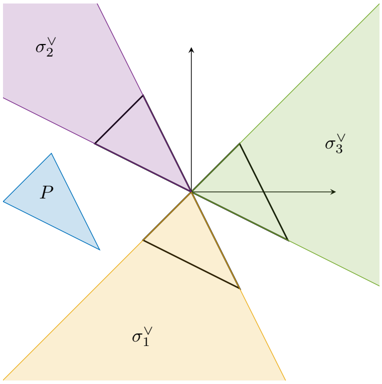


Figure D.3: A translated version of the polytope P from Example D.1.1 and the cones associated to the vertices.

D.3 Fans

Definition D.3.1 (Fan)**.** A *fan* in $N_{\mathbb{R}} \simeq \mathbb{R}^n$ is a finite collection Σ of strongly convex rational polyhedral cones $\sigma \subset N_{\mathbb{R}}$ satisfying

1. for all $\sigma \in \Sigma$, every face $\tau \subset \sigma$ is in Σ.
2. the intersection $\sigma \cap \sigma'$ for any $\sigma, \sigma' \in \Sigma$ is a face of both σ and σ'.

The support $|\Sigma|$ of Σ is defined as $|\Sigma| = \bigcup_{\sigma \in \Sigma} \sigma \subset N_{\mathbb{R}}$ and by $\Sigma(d) \subset \Sigma$ we denote the set of d-dimensional cones of Σ.

The set $\Sigma(1)$ is the set of *rays* of Σ. The *primitive ray generator* of a ray $\rho \in \Sigma(1)$ is the generator of the monoid $\rho \cap N$ (i.e. it is the 'smallest' nonzero integer vector contained in the ray). The most important fans for our purpose are those arising as the *normal fan* of a lattice polytope. Consider a minimal H-representation

$$P = \{m \in M_{\mathbb{R}} \mid \langle u_i, m\rangle + a_i \geq 0, i = 1, \dots, k\}$$

of a full-dimensional lattice polytope $P \subset M_{\mathbb{R}}$, where u_i is the primitive, inward pointing facet normal of the facet Q_i (see Section D.1 for a definition). We have seen a way of obtaining cones from P in Example D.2.2. For a vertex $v \in P$, we define $\sigma_v^\vee = \operatorname{Cone}(P - v) = \operatorname{Cone}(\{m - v \mid m \in P\}) \subset M_{\mathbb{R}}$ and $\sigma_v = (\sigma_v^\vee)^\vee$. Every face of $\sigma_v^\vee$ corresponds to a face of P containing v, and in particular all facets of $\sigma_v^\vee$ correspond to facets of P containing v. Hence

$$\sigma_v^\vee = \{m \in M_{\mathbb{R}} \mid \langle u_i, m\rangle \geq 0, \text{ for all } i \text{ such that } \langle u_i, v\rangle + a_i = 0\}.$$

This is exactly the definition of the dual cone of a cone generated by the $\{u_i \mid \langle u_i, v\rangle + a_i = 0\}$, so

$$\sigma_v = \operatorname{Cone}(\{u_i \mid \langle u_i, v\rangle + a_i = 0\}) = \operatorname{Cone}(\{u_i \mid v \in Q_i\}).$$

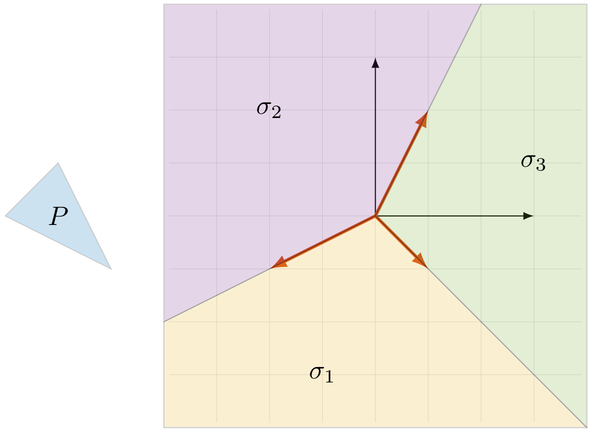


Figure D.4: The normal fan Σ_P of P from Example D.1.1. The primitive ray generators are drawn in orange, the color of the dimension 2 cones of Σ_P corresponds to the color of their duals in Figure D.3.

We generalize this construction for higher dimensional faces $Q \subset P$ by setting

$$\sigma_Q = \operatorname{Cone}(\{u_i \mid Q \subset Q_i\}).$$

The set of cones that we obtain in this way has some nice properties. For example, for any face $Q \subset P$ we have $\dim Q + \dim \sigma_Q = n$. This means that for a vertex v, σ_v is an n-dimensional cone. Also, one can prove that the cones σ_v corresponding to the vertices of P cover the whole vector space:

$$N_{\mathbb{R}} = \bigcup_{v \text{ vertex of } P} \sigma_v = \bigcup_{Q \text{ face of } P} \sigma_Q.$$

See for instance [CLS11, Proposition 2.3.8].

Theorem D.3.1. *Let $P \subset M_{\mathbb{R}}$ be a full dimensional lattice polytope. Then $\{\sigma_Q \mid Q \text{ is a face of } P\}$ is a fan.*

Proof. See [CLS11, Theorem 2.3.2]. □

The collection $\Sigma_P = \{\sigma_Q \mid Q \text{ face of } P\}$ is called the *normal fan* of P. The support of Σ_P is $|\Sigma_P| = N_{\mathbb{R}}$. Fans in $N_{\mathbb{R}}$ whose support is $N_{\mathbb{R}}$ are called *complete*.

Example D.3.1. An illustration of a normal fan for the polytope from Example D.1.1 can be found in Figure D.4. Note that in Figure D.4 the cones are drawn in $N_{\mathbb{R}} \simeq \mathbb{R}^2$, whereas in Figure D.3, the picture is in the dual space $M_{\mathbb{R}} \simeq \mathbb{R}^2$. The primitive ray generators of $\Sigma_P(1)$ are exactly the inward pointing facet normals from Example D.1.1 and Σ_P is complete. △

Appendix E

Toric geometry

This appendix summarizes some basic results from toric geometry to support the material presented in Chapter 5. Our motivation for studying toric varieties is the fact that they are natural solution spaces for systems of polynomial equations coming from polyhedral families. The toric varieties we are mostly interested in are *complete, normal toric varieties*. The structure of such a variety X is completely encoded by a complete fan Σ. The cones in Σ correspond to the *affine toric varieties* which form an open cover of X. In Section E.1 we discuss affine toric varieties, which are the fundamental building blocks of abstract toric varieties. Section E.2 discusses projective toric varieties and their connection with polytopes and their normal fans. For more details, a great first introduction and a modern treatment of toric geometry, the reader is referred to [CLS11]. Alternatively, the books [OM78, Ful93] are standard, more classical references. These notes are strongly based on an exam paper the author wrote for a course on algebraic geometry at KU Leuven taught by Nero Budur.

E.1 Affine toric varieties

Perhaps the most basic example of an affine toric variety is the algebraic torus $(\mathbb{C}^*)^n$. This variety has the extra structure of an abelian group under element-wise multiplication:

$$(t_1, \ldots, t_n) \cdot (u_1, \ldots, u_n) = (t_1u_1, \ldots, t_nu_n).$$

By *a torus* T we mean an affine variety isomorphic to $(\mathbb{C}^*)^n$, where the isomorphism respects this group structure: it is an isomorphism of varieties which is also an isomorphism of groups. A *character* of a torus T is a group homomorphism $\chi : T \to \mathbb{C}^*$. A tuple of integers $m = (m_1, \ldots, m_n) \in \mathbb{Z}^n$ gives a character $\chi^m : (\mathbb{C}^*)^n \to \mathbb{C}^*$ defined by $\chi^m(t_1, \ldots, t_n) = t_1^{m_1} \cdots t_n^{m_n}$. One shows that all possible characters of $(\mathbb{C}^*)^n$ arise in this way [Hum12, Section 16.2, Lemma A-B], so the characters of $(\mathbb{C}^*)^n$ form a group isomorphic to $\mathbb{Z}^n$. For any torus T, the characters form a free abelian group of finite rank

$$M = \operatorname{Hom}_{\mathbb{Z}}(T, \mathbb{C}^*).$$

Such a group is called a *lattice*, which is why M is sometimes referred to as the *character lattice*. Every $m \in M$ gives a character χ^m. The rank of M is equal to the dimension of T as a variety.

Example E.1.1. For $T = (\mathbb{C}^*)^n$, the group of characters $M \simeq \mathbb{Z}^n$ can be thought of as the Laurent monomials in n variables. An element $m \in M$ corresponds to the character of evaluating the Laurent monomial t^m. Therefore, for an arbitrary torus T the isomorphism $T \simeq (\mathbb{C}^*)^n$ induces an isomorphism $M \simeq \mathbb{Z}^n$ that turns characters into Laurent monomials. $\triangle$

Another important group associated to a torus T is the $\mathbb{Z}$-dual N of M:

$$N = \operatorname{Hom}_{\mathbb{Z}}(M, \mathbb{Z}) = \operatorname{Hom}_{\mathbb{Z}}(\mathbb{C}^*, T).$$

This is the group of *one-parameter subgroups* or *cocharacters* of T. By definition, a one-parameter subgroup or cocharacter is a group homomorphism $\lambda : \mathbb{C}^* \to T$. An integer tuple $u = (u_1, \ldots, u_n) \in \mathbb{Z}^n$, gives a cocharacter $\lambda_u : \mathbb{C}^* \to (\mathbb{C}^*)^n$ with

$$\lambda^u(t) = (t^{u_1}, \ldots, t^{u_n}).$$

All cocharacters of $(\mathbb{C}^*)^n$ arise in this way, which establishes $N \simeq \mathbb{Z}^n$. As for any torus T we have $T \simeq (\mathbb{C}^*)^n$ for some n, the (co-)character lattices M and N can be thought of as two (dual) copies of $\mathbb{Z}^n$.

Definition E.1.1 (Affine toric variety)**.** An *affine toric variety* is an irreducible affine variety Y containing a torus $T \simeq (\mathbb{C}^*)^n$ as a Zariski open subset such that the action of T on itself extends to an action $T \times Y \to Y$ of T on Y, given by a morphism.

Example E.1.2. The torus $(\mathbb{C}^*)^n$ itself is obviously an affine toric variety. The same holds for $\mathbb{C}^n$. Indeed, $\mathbb{C}^n$ is irreducible, $(\mathbb{C}^*)^n = \mathbb{C}^n \backslash V_{\mathbb{C}^n}(x_1 \cdots x_n)$ and the action of $(\mathbb{C}^*)^n$ on itself extends to an action $(\mathbb{C}^*)^n \times \mathbb{C}^n \longrightarrow \mathbb{C}^n$ on $\mathbb{C}^n$ by

$$(t_1, \ldots, t_n) \times (x_1, \ldots, x_n) \longrightarrow (t_1 x_1, \ldots, t_n x_n).$$

$\triangle$

Example E.1.3. This is Example 1.1.5 in [CLS11]. Consider the variety $Y = \{xy - zw = 0\} \subset \mathbb{C}^4$. This is an affine toric variety with torus

$$Y \cap (\mathbb{C}^*)^4 = \{(t_1, t_2, t_3, t_1 t_2 t_3^{-1}) : t_i \in \mathbb{C}^*\} \simeq (\mathbb{C}^*)^3. \tag{E.1.1}$$

The torus action extends to an action on Y by

$$(t_1, t_2, t_3) \times (x, y, z, w) \longrightarrow (t_1 x, t_2 y, t_3 z, t_1 t_2 t_3^{-1} w).$$

$\triangle$

We introduce three ways of constructing affine toric varieties: from a set of lattice points, from a toric ideal or from affine semigroups (and cones).
Let $\mathscr{A} = \{m_1, \ldots, m_s\} \subset M$ be a finite subset of the character lattice $M \simeq \mathbb{Z}^n$ of $(\mathbb{C}^*)^n$. Consider the map

$$\phi_{\mathscr{A}} : (\mathbb{C}^*)^n \longrightarrow \mathbb{C}^s \quad \text{given by} \quad \phi_{\mathscr{A}}(t) = (\chi^{m_1}(t), \ldots, \chi^{m_s}(t)).$$

We define $Y_{\mathscr{A}} = \overline{\operatorname{im}\phi_{\mathscr{A}}}$, where $\bar{\cdot}$ is the Zariski closure in $\mathbb{C}^s$.

Proposition E.1.1. *Given the finite subset $\mathscr{A} \subset M$. Let $\mathbb{Z}\mathscr{A}$ be the sublattice generated by $\mathscr{A}$. Then $Y_{\mathscr{A}}$ is an affine toric variety whose torus has character lattice $\mathbb{Z}\mathscr{A} = \{\sum_{m\in\mathscr{A}} c_m m \mid c_m \in \mathbb{Z}$ for all $m \in \mathscr{A}\}$.*

Proof. See [CLS11, Proposition 1.1.8]. □

Example E.1.4. Consider again the affine variety Y of Example E.1.3 with torus $T = (\mathbb{C}^*)^3$. This is the toric variety $Y_{\mathscr{A}}$ defined by

$$\mathscr{A} = \{(1,0,0),(0,1,0),(0,0,1),(1,1,-1)\} \subset \mathbb{Z}^3.$$

Indeed, $\phi_{\mathscr{A}}$ gives the isomorphism of tori in (E.1.1). △

Note that in particular, Proposition E.1.1 implies $\dim(Y_{\mathscr{A}}) = \operatorname{rank}(\mathbb{Z}\mathscr{A})$, so it is equal to n if and only if the elements of $\mathscr{A}$ form an $\mathbb{R}$-basis for $M_{\mathbb{R}} = \mathbb{R}^n$. Equivalently, if we stack the elements of $\mathscr{A}$ into an $n \times s$ matrix $A = \begin{bmatrix} m_1 & \cdots & m_s \end{bmatrix}$, then $\dim(Y_{\mathscr{A}}) = \operatorname{rank}(A)$. Let $\hat{\phi}_{\mathscr{A}} : \mathbb{Z}^s \longrightarrow M$ be the $\mathbb{Z}$-map represented by the matrix A and define $L = \ker \hat{\phi}_{\mathscr{A}}$. Let $e_1, \ldots, e_s$ be the standard basis of $\mathbb{Z}^s$. For $\ell = (\ell_1, \ldots, \ell_s) \in L$, define

$$\ell_+ = \sum_{\ell_i > 0} \ell_i e_i, \qquad \ell_- = -\sum_{\ell_i < 0} \ell_i e_i.$$

The binomial $x^{\ell_+} - x^{\ell_-} \in \mathbb{C}[x_1, \ldots, x_s]$ vanishes on $Y_{\mathscr{A}} \subset \mathbb{C}^s$ by construction (see below). Doing this for all $\ell \in L$ gives an ideal $I_{\mathscr{A}}$.

Proposition E.1.2. *The vanishing ideal $I(Y_{\mathscr{A}})$ of the affine toric variety $Y_{\mathscr{A}}$ is*

$$I_{\mathscr{A}} = \langle x^{\ell_+} - x^{\ell_-} : \ell \in L \rangle.$$

Proof. For $\ell \in L$, let $f_\ell = x^{\ell_+} - x^{\ell_-}$. For any $t \in (\mathbb{C}^*)^n$,

$$f_\ell(\phi_{\mathscr{A}}(t)) = t^{\sum_{\ell_i>0} \ell_i m_i} - t^{-\sum_{\ell_i<0} \ell_i m_i} = 0,$$

since $\sum_{\ell_i>0} \ell_i m_i = -\sum_{\ell_i<0} \ell_i m_i$ by $\ell \in L$. Hence every element of $I_{\mathscr{A}}$ vanishes on $\operatorname{im}\phi_{\mathscr{A}}$ and the inclusion $I_{\mathscr{A}} \subset I(Y_{\mathscr{A}})$ follows immediately. The opposite inclusion is proved by contradiction, see [CLS11, Proposition 1.1.9]. □

Example E.1.5. Consider once more the affine toric variety Y from Example E.1.3. From Example E.1.4 we know that the matrix A defining $\hat{\phi}_{\mathscr{A}}$ is

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

and L is spanned by $\begin{bmatrix} 1 & 1 & -1 & -1 \end{bmatrix}^\top$. Taking coordinates x, y, z, w on $\mathbb{C}^4$, this gives $I_{\mathscr{A}} = \langle xy - zw \rangle$. The generator is exactly the defining equation given in Example E.1.3. △

An ideal of the form $\langle x^{\ell_+} - x^{\ell_-} : \ell \in L \rangle$ for any sublattice $L \subset \mathbb{Z}^s$ is called a *lattice ideal*. A prime lattice ideal is a *toric ideal*. It can be shown [CLS11, Proposition 1.1.11] that the set of toric ideals is the set of prime ideals generated by binomials (note that only one inclusion is obvious). It turns out that *all* affine toric varieties are the zero locus of a toric ideal, hence every affine toric variety is cut out by binomial equations.

We have looked at affine toric varieties as the closure of the image of a Laurent monomial map and as the zero locus of a toric ideal. Our third construction will exploit the connection between the coordinate rings of affine toric varieties and semigroup algebras. We have encountered affine semigroups before in Section D.2. We recall the definition.

Definition E.1.2 (Affine semigroup)**.** An *affine semigroup* is a set S with an associative binary operation '+' and identity element 0 such that:

1. '+' is commutative,
2. S is finitely generated: there is a finite set $\mathscr{A}$ such that $\mathbb{N}\mathscr{A} = \{\sum_{m \in \mathscr{A}} a_m m \mid a_m \in \mathbb{N}\} \subset \mathsf{S}$,
3. the semigroup can be embedded in a lattice M.

For our purpose, S is embedded into the character lattice M of some torus, so $\mathsf{S} \subset M$ and S is generated by a finite set $\mathscr{A}$ of characters.

Definition E.1.3 (Semigroup algebra)**.** Given an affine semigroup $\mathsf{S} \subset M$, the *semigroup algebra* $\mathbb{C}[\mathsf{S}]$ over S is the $\mathbb{C}$-vector space with basis S and multiplication induced by the semigroup structure of S. That is,

$$\mathbb{C}[\mathsf{S}] = \left\{ \sum_{m \in \mathsf{S}} c_m \chi^m : c_m \in \mathbb{C},\ c_m \neq 0 \text{ for finitely many } m \right\}$$

and multiplication is defined by $\chi^m \cdot \chi^{m'} = \chi^{m+m'}$.

Note that if $\mathscr{A} = \{m_1, \ldots, m_s\}$ generates S, then $\mathbb{C}[\mathsf{S}] = \mathbb{C}[\chi^{m_1}, \ldots, \chi^{m_s}]$. The reader can think of M as $\mathbb{Z}^n$ and of the χ^m as Laurent monomials.

Proposition E.1.3. *Let $S \subset M$ be an affine semigroup generated by $\mathscr{A} = \{m_1, \ldots, m_s\}$. Then*

1. $\mathbb{C}[\mathsf{S}]$ *is an integral domain and finitely generated as a $\mathbb{C}$-algebra.*
2. $\mathbb{C}[\mathsf{S}] \simeq \mathbb{C}[Y_{\mathscr{A}}]$ *where $\mathbb{C}[Y_{\mathscr{A}}]$ is the coordinate ring of $Y_{\mathscr{A}}$, hence $Y_{\mathscr{A}} = \mathrm{MaxSpec}(\mathbb{C}[\mathsf{S}])$.*
3. *The character lattice of the torus $T_{Y_{\mathscr{A}}}$ of $Y_{\mathscr{A}}$ is $\mathbb{Z}\mathsf{S}$.*

Proof. Since $\mathsf{S} \subset M$ we have $\mathbb{C}[\mathsf{S}] \subset \mathbb{C}[M]$ and $\mathbb{C}[M]$ is the coordinate ring of the torus $(\mathbb{C}^*)^n$ with character lattice M. Since $(\mathbb{C}^*)^n$ is irreducible, $\mathbb{C}[M]$ is an integral domain and so is $\mathbb{C}[\mathsf{S}]$. The algebra $\mathbb{C}[\mathsf{S}]$ is finitely generated because $\mathbb{C}[\mathsf{S}] = \mathbb{C}[\chi^{m_1}, \ldots, \chi^{m_s}]$. For the second statement, consider the $\mathbb{C}$-algebra homomorphism $(\phi_{\mathscr{A}})^* : \mathbb{C}[x_1, \ldots, x_s] \to \mathbb{C}[M]$ defined by $x_i \mapsto \chi^{m_i} \in \mathbb{C}[M]$ (note that this is the pullback of the Laurent monomial map $\phi_{\mathscr{A}}$). We have $\ker(\phi_{\mathscr{A}})^* = I_{\mathscr{A}}$ and the image $\mathrm{im}(\phi_{\mathscr{A}})^*$ is $\mathbb{C}[\chi^{m_1}, \ldots, \chi^{m_s}] = \mathbb{C}[\mathsf{S}]$. Therefore

$$\begin{aligned}\mathbb{C}[Y_{\mathscr{A}}] &= \mathbb{C}[x_1, \ldots, x_s]/I_{\mathscr{A}} \\ &= \mathbb{C}[x_1, \ldots, x_s]/\ker(\phi_{\mathscr{A}})^* \simeq \mathrm{im}(\phi_{\mathscr{A}})^* = \mathbb{C}[\mathsf{S}].\end{aligned}$$

The third statement follows from $\mathbb{Z}\mathsf{S} = \mathbb{Z}(\mathbb{N}\mathscr{A}) = \mathbb{Z}\mathscr{A}$ and Proposition E.1.1. □

A nice fact is that all affine toric varieties arise from the three equivalent constructions introduced above. The following is Theorem 1.1.17 in [CLS11].

Theorem E.1.1. *Let Y be an affine variety. The following are equivalent:*

1. *Y is an affine toric variety,*
2. *$Y = Y_{\mathscr{A}}$ for a finite set $\mathscr{A}$ in a lattice,*
3. *Y is the variety of a toric ideal,*
4. *$Y = \mathrm{MaxSpec}(\mathbb{C}[\mathsf{S}])$ for an affine semigroup S.*

The interpretation of $Y_{\mathscr{A}}$ as $\mathrm{MaxSpec}(\mathbb{C}[\mathsf{S}])$ for an affine semigroup S leads to an interesting relation with rational polyhedral cones (see Section D.2).

Proposition E.1.4. *Let $\sigma \subset N_{\mathbb{R}} \simeq \mathbb{R}^n$ be a rational polyhedral cone and let $\mathsf{S}_\sigma = \sigma^\vee \cap M$. Then $U_\sigma = \mathrm{MaxSpec}(\mathbb{C}[\mathsf{S}_\sigma])$ is an affine toric variety.*

Proof. The theorem follows immediately from Lemma D.2.1 and Proposition E.1.3. □

One can show that $\dim(U_\sigma) = n$ if and only if σ is strongly convex [CLS11, Proposition 1.2.18]. The reason why the affine toric variety U_σ is defined by the affine semigroup $\sigma^\vee \cap M$ rather than $\sigma \cap N$ will become clear later. A nice property of affine toric varieties of the form U_σ where σ is strongly convex is that they are *normal*. The reason that this is a desirable property for a variety is that it allows to develop a nice theory of *divisors*. The definition of normality is quite technical, but we will include it for completeness. We will see shortly that 'normality' is easy to describe for affine toric varieties. An integral domain R is called *integrally closed* if it is integrally closed in its field of fractions. This means that for any monic polynomial $f \in R[x]$ and $x^* \in K(R)$, $f(x^*) = 0$ implies $x^* \in R \subset K(R)$.

Definition E.1.4 (Normal varieties)**.** An irreducible affine variety Y is normal if its coordinate ring $\mathbb{C}[Y]$ is integrally closed.

For semigroup algebras the property of being integrally closed corresponds to the more geometric notion of the semigroup being *saturated*. Intuitively speaking, this means that the semigroup has 'no holes' in its ambient lattice.

Definition E.1.5 (Saturated semigroup)**.** An affine semigroup $\mathsf{S} \subset M$ is said to be *saturated* in M if for all $k \in \mathbb{N}\backslash\{0\}$ and $m \in M$, $km \in \mathsf{S}$ implies $m \in \mathsf{S}$.

Theorem E.1.2. *Let Y be an affine toric variety with torus T. Let M and N be the character and cocharacter lattice of T. The following are equivalent:*

1. *Y is normal,*
2. *$Y = \mathrm{MaxSpec}(\mathbb{C}[\mathsf{S}])$ where $\mathsf{S} \subset M$ is a saturated affine semigroup,*
3. *$V = U_\sigma$ where $\sigma \subset N_\mathbb{R}$ is a strongly convex rational polyhedral cone.*

Proof. This is Theorem 1.3.5 in [CLS11]. □

Example E.1.6. Let $\mathsf{S} = \mathbb{N}\{2,3\} \subset \mathbb{Z}$, such that $\mathbb{C}[\mathsf{S}] = \mathbb{C}[t^2, t^3]$. The associated toric variety is $Y = Y_{\{2,3\}} = V_{\mathbb{C}^2}(x^3 - y^2)$. This variety is not normal, since $\mathsf{S} = \{0, 2, 3, 4, \ldots\}$ is not saturated in the character lattice $M = \mathbb{Z}\{2,3\}$ of Y. Its coordinate ring $\mathbb{C}[\mathsf{S}]$ is not integrally closed, since $t \in K(\mathbb{C}[\mathsf{S}]) \setminus \mathbb{C}[\mathsf{S}]$ is a root of the monic polynomial $x^2 - t^2 \in \mathbb{C}[\mathsf{S}][x]$. △

Example E.1.7. Let $\mathsf{S} = \mathbb{N}\{2\} \subset \mathbb{Z}$, such that $\mathbb{C}[\mathsf{S}] = \mathbb{C}[t^2] \simeq \mathbb{C}[t]$, and $Y = \mathrm{MaxSpec}(\mathbb{C}[\mathsf{S}]) \simeq \mathbb{C}$ is normal. This does not contradict Theorem E.1.2, since the torus T of Y has character lattice $\mathbb{Z}\mathsf{S} = 2\mathbb{Z} \subset \mathbb{Z}$, in which S is saturated. △

Example E.1.8. The variety Y from Example E.1.3 is the variety $Y_\mathscr{A}$ as shown in Example E.1.4 and it is the variety U_σ for the convex polyhedral cone σ from Example D.2.1. Therefore $Y = \mathrm{MaxSpec}(\mathbb{C}[\mathsf{S}_\sigma])$ with S_σ the affine semigroup generated by $\mathscr{A}$. The affine semigroup S_σ is saturated, hence Y is a normal affine variety. △

E.2 Projective toric varieties and polytopes

Like in the affine case, our first description of projective toric variety will be based on monomial maps. Next, we show how such a projective toric variety is covered by open subsets isomorphic to affine toric varieties. Finally, we define the toric variety of a polytope through its normal fan.

The inclusion $(\mathbb{C}^*)^n \to \mathbb{P}^n$ given by

$$(t_1, \ldots, t_n) \mapsto (1 : t_1 : \cdots : t_n)$$

shows that $(\mathbb{C}^*)^n$ is a dense open subset of $\mathbb{P}^n$. We will denote the (isomorphic) image of $(\mathbb{C}^*)^n$ by $T_{\mathbb{P}^n} \subset \mathbb{P}^n$. Moreover, the action of $(\mathbb{C}^*)^n$ on itself extends to an action of $(\mathbb{C}^*)^n$ on $\mathbb{P}^n$, which is given by a morphism. These will be the requirements we impose on a projective variety for it to be *toric*, generalizing Definition E.1.1.

Definition E.2.1 (Projective toric variety)**.** A *projective toric variety* is an irreducible projective variety X containing a torus $T \simeq (\mathbb{C}^*)^n$ as a Zariski open subset such that the action of T on itself extends to an action $T \times X \to X$ of T on X, given by a morphism.

Let M be the character lattice of $(\mathbb{C}^*)^n$ and consider a finite set $\mathscr{A} = \{m_0, \ldots, m_s\} \subset M$. We consider the map

$$\pi : (\mathbb{C}^*)^{s+1} \to \mathbb{P}^s \quad \text{given by} \quad (t_0, \ldots, t_s) \mapsto (t_0 : \cdots : t_s).$$

The set $\mathscr{A} \subset M$ gives an affine toric variety $Y_{\mathscr{A}} = \overline{\operatorname{im} \phi_{\mathscr{A}}} \subset \mathbb{C}^{s+1}$ as before, and $\operatorname{im} \phi_{\mathscr{A}} \subset (\mathbb{C}^*)^{s+1}$. Composing the map $\phi_{\mathscr{A}}$ with π, we get a map

$$(\mathbb{C}^*)^n \xrightarrow{\phi_{\mathscr{A}}} (\mathbb{C}^*)^{s+1} \xrightarrow{\pi} \mathbb{P}^s.$$

We define $X_{\mathscr{A}} = \overline{\operatorname{im}(\pi \circ \phi_{\mathscr{A}})} \subset \mathbb{P}^s$.

Theorem E.2.1. *The projective variety $X_{\mathscr{A}}$ is a projective toric variety whose dimension is equal to the dimension of the smallest affine subspace of $M_{\mathbb{R}}$ containing $\mathscr{A}$. Its torus has character lattice*

$$\mathbb{Z}'\mathscr{A} = \left\{ \sum_{i=0}^{s} a_i m_i \mid a_i \in \mathbb{Z}, \sum_{i=1}^{s} a_i = 0 \right\}.$$

Proof. See Propositions 2.1.2 and 2.1.6 in [CLS11]. □

Example E.2.1 (Segre embedding)**.** Let $M = \mathbb{Z}^2$ and

$$\mathscr{A} = \{(0,0), (1,0), (0,1), (1,1)\} \subset M.$$

Let x, y be coordinates on $(\mathbb{C}^*)^2$ and let u, s, v, t be homogeneous coordinates on $\mathbb{P}^3$. The map $\pi \circ \phi_{\mathscr{A}} : (\mathbb{C}^*)^2 \to \mathbb{P}^3$ is given by

$$(x, y) \to (1 : x : y : xy).$$

The closure of the image in $\mathbb{P}^3$ can be shown to be $X_{\mathscr{A}} = \{ut - sv = 0\} \subset \mathbb{P}^3$. The projective variety $X_{\mathscr{A}}$ is the image of the Segre embedding of $\mathbb{P}^1 \times \mathbb{P}^1$ in $\mathbb{P}^3$ and hence $X_{\mathscr{A}} \simeq \mathbb{P}^1 \times \mathbb{P}^1$. △

Theorem E.2.1 shows that the dimension of the projective toric variety is not determined by the dimension of the *linear* $\mathbb{R}$-span of the lattice points in M, but rather by the *affine* $\mathbb{R}$-span. This is illustrated by our standard example.

Example E.2.2. For the variety Y from Example E.1.3, we have that $Y = Y_{\mathscr{A}} = V_{\mathbb{C}^4}(xy - zw)$ with $\mathscr{A} = \{(1,0,0),(0,1,0),(0,0,1),(1,1,-1)\}$. Note that $Y_{\mathscr{A}}$ is closed under the action $\mathbb{C}^* \times \mathbb{C}^4 \to \mathbb{C}^4$ of 'scalar multiplication' by $\mathbb{C}^*$. The map π is the projection $(x,y,z,w) \mapsto (x:y:z:w)$ along the orbits of this action. The projective toric variety $X_{\mathscr{A}} \subset \mathbb{P}^3$ is given by the equation $xy - zw = 0$, which was already homogeneous. In this example, $Y_{\mathscr{A}}$ is of one dimension higher than $X_{\mathscr{A}}$ (π maps lines in $\operatorname{im}\phi_{\mathscr{A}}$ to points in $X_{\mathscr{A}}$, so it takes away one dimension) and it is the affine cone over $X_{\mathscr{A}}$. $\triangle$

Example E.2.2 is an illustration of the following result, which is Proposition 2.1.4 in [CLS11]. It uses the notation $H_{u,a}$ for a hyperplane in $M_{\mathbb{R}}$ defined by $u \in N_{\mathbb{R}}$ and $a \in \mathbb{R}$ (see Section D.1).

Proposition E.2.1. *Let $Y_{\mathscr{A}}$, $X_{\mathscr{A}}$, $I_{\mathscr{A}}$ be the affine toric variety, projective toric variety and toric ideal defined by the finite subset $\mathscr{A} = \{m_0, \dots, m_s\} \subset M$. Let I_L be as in Proposition E.1.2 and let $S = \mathbb{C}[x_0, \dots, x_s]$ be the homogeneous coordinate ring of $\mathbb{P}^s$. The following are equivalent:*

1. *$Y_{\mathscr{A}}$ is the affine cone over $X_{\mathscr{A}}$,*
2. *$I_L = I_S(X_{\mathscr{A}})$,*
3. *I_L is homogeneous,*
4. *$\mathscr{A} \subset H_{u,a} \subset M_{\mathbb{R}}$ for some $u \in N$ and $a \in \mathbb{N}_{>0}$.*

Proposition E.2.1 can be used to obtain the ideal $I_S(X_{\mathscr{A}})$ even if $\mathscr{A}$ is not contained in a hyperplane in $M_{\mathbb{R}}$. The trick is to replace $\mathscr{A}$ by $\mathscr{A} \times \{1\} = \{(m_0, 1), \dots, (m_s, 1)\} \subset M \times \mathbb{Z}$. Observe that $X_{\mathscr{A}} = X_{\mathscr{A} \times \{1\}}$ and $\mathscr{A} \times \{1\} \subset H_{u,1}$ with $u = (0, \dots, 0, 1)$.

Let $x_0, \dots, x_s$ be homogeneous coordinates on $\mathbb{P}^s$ and define $U_i = \mathbb{P}^s \setminus V(x_i), i = 0, \dots, s$ as the usual affine charts of $\mathbb{P}^s$. It is clear that $T_{\mathbb{P}^s} \subset U_i$ for all i. Let $T_{X_{\mathscr{A}}}$ denote the torus of $X_{\mathscr{A}}$. We have

$$T_{X_{\mathscr{A}}} = X_{\mathscr{A}} \cap T_{\mathbb{P}^s} \subset X_{\mathscr{A}} \cap U_i.$$

Since $X_{\mathscr{A}}$ is the Zariski closure of $T_{X_{\mathscr{A}}}$ in $\mathbb{P}^s$, $X_{\mathscr{A}} \cap U_i$ is the Zariski closure of $T_{X_{\mathscr{A}}} \cap U_i = T_{X_{\mathscr{A}}}$ in $U_i \simeq \mathbb{C}^s$ and hence $X_{\mathscr{A}} \cap U_i$ is an affine toric variety.

Proposition E.2.2. *Let $X_{\mathscr{A}} \subset \mathbb{P}^s$ be defined as above by $\mathscr{A} = \{m_0, \dots, m_s\} \subset M$. The affine piece $X_{\mathscr{A}} \cap U_i$ is isomorphic to the affine toric variety $Y_{\mathscr{A}_i} = \operatorname{MaxSpec}(\mathbb{C}[\mathsf{S}_i])$ with $\mathscr{A}_i = \mathscr{A} - m_i = \{m_0 - m_i, \dots, m_{i-1} - m_i, m_{i+1} - m_i, \dots, m_s - m_i\}$ and $\mathsf{S}_i = \mathbb{N}\mathscr{A}_i$.*

Proof. The isomorphism $U_i \simeq \mathbb{C}^s$ is given by

$$(a_0 : \cdots : a_s) \xrightarrow{\phi_i} \left(\frac{a_0}{a_i}, \dots, \frac{a_{i-1}}{a_i}, \frac{a_{i+1}}{a_i}, \dots, \frac{a_s}{a_i} \right),$$

see Subsection 2.2.5. Now, we can apply ϕ_i to $X_{\mathscr{A}} \cap U_i$. Combining this with the map $\pi \circ \phi_{\mathscr{A}}$ we obtain that $X_{\mathscr{A}} \cap U_i$ is isomorphic to the closure of the image of the map $(\mathbb{C}^*)^n \longrightarrow \mathbb{C}^s$ given by

$$t \mapsto (\chi^{m_0 - m_i}(t), \ldots, \chi^{m_{i-1}-m_i}(t), \chi^{m_{i+1}-m_i}(t), \ldots \chi^{m_s - m_i}(t)),$$

which is by definition equal to the affine toric variety $Y_{\mathscr{A}_i}$. The equality $Y_{\mathscr{A}_i} = \operatorname{MaxSpec}(\mathbb{C}[S_i])$ follows from Proposition E.1.3. $\square$

Since the isomorphism $X_{\mathscr{A}} \cap U_i \simeq Y_{\mathscr{A}_i} = \operatorname{MaxSpec}(\mathbb{C}[\mathsf{S}_i])$ induces an isomorphism of coordinate rings $\mathbb{C}[\mathsf{S}_i] \to \mathbb{C}[X_{\mathscr{A}} \cap U_i]$ which sends $\chi^{m_j - m_i}$ to $\frac{x_j}{x_i} + I_{\mathscr{A}_i}$, we get that $X_{\mathscr{A}} \cap U_i \cap U_j = X_{\mathscr{A}} \cap U_j \cap U_i$ is isomorphic to

$$(Y_{\mathscr{A}_i})_{\chi^{m_j - m_i}} = \operatorname{MaxSpec}(\mathbb{C}[\mathsf{S}_i]_{\chi^{m_j-m_i}}) \simeq \operatorname{MaxSpec}(\mathbb{C}[\mathsf{S}_j]_{\chi^{m_i-m_j}}) = (Y_{\mathscr{A}_j})_{\chi^{m_i-m_j}}.$$

Since $\mathbb{P}^s = \bigcup_{i=0}^s U_i$ we have $X_{\mathscr{A}} = \bigcup_{i=0}^s X \cap U_i$. It turns out that some of the affine pieces in this decomposition may be redundant.

Proposition E.2.3. *Given* $\mathscr{A} = \{m_0, \ldots, m_s\} \subset M$, *let* $P = \operatorname{Conv}(\mathscr{A}) \subset M_{\mathbb{R}}$ *and define* $\mathscr{T} = \{j \in \{0, \ldots, s\} : m_j \text{ is a vertex of } P\}$. *Then*

$$X_{\mathscr{A}} = \bigcup_{j \in \mathscr{T}} X_{\mathscr{A}} \cap U_j.$$

Proof. We give a sketch of the proof, more details can be found in [CLS11, Proposition 2.1.9]. The key observation is that when m_i is not a vertex of P, then there is $m_j, j \in \mathscr{T}$ such that both $m_i - m_j \in \mathsf{S}_i$ and $m_j - m_i \in \mathsf{S}_i$. This means that $\chi^{m_j - m_i}$ is invertible in $\mathbb{C}[\mathsf{S}_i]$ and thus $\mathbb{C}[\mathsf{S}_i]_{\chi^{m_j-m_i}} = \mathbb{C}[\mathsf{S}_i]$. It follows that $X_{\mathscr{A}} \cap U_i \cap U_j \simeq Y_{\mathscr{A}_i} \cap Y_{\mathscr{A}_j} = \operatorname{MaxSpec}(\mathbb{C}[\mathsf{S}_i]_{\chi^{m_j-m_i}}) = \operatorname{MaxSpec}(\mathbb{C}[\mathsf{S}_i]) = Y_{\mathscr{A}_i} \simeq X_{\mathscr{A}} \cap U_i$. This implies $X_{\mathscr{A}} \cap U_i \subset U_j$. $\square$

Proposition E.2.3 illustrates how polytopes pop up naturally in describing projective toric varieties. Given a full-dimensional lattice polytope $P \subset M_{\mathbb{R}} \simeq \mathbb{R}^n$ where $M = \mathbb{Z}^n$, we can associate a projective toric variety to it by constructing $X_{P \cap M}$ via the monomial map $\phi_{P \cap M}$. In some cases, however, when the polytope P contains 'too few' lattice points, this construction leads to a toric variety which is not *normal*, which is a property we need our toric variety to have for some of the purposes in this thesis.

Definition E.2.2. A variety X with affine open cover $X = \cup_{i \in \mathscr{T}} U_i$ is *normal* if each of the affine varieties U_i is normal.

From this definition we see that if the affine semigroup generated by $P \cap M - m_i$ for some vertex $m_i \in P$ has 'holes' in M, the variety X is not normal. We will avoid this by enlarging the polytope P until it has 'enough' lattice points. Let us make this precise.

Definition E.2.3. A lattice polytope $P \subset M_{\mathbb{R}}$ is called *very ample* if for every vertex $m \in P$, the semigroup $\mathsf{S}_{P,m} = \mathbb{N}(P \cap M - m)$ generated by the set $P \cap M - m$ is saturated in M.

Proposition E.2.4. *If $P \subset M_{\mathbb{R}} \simeq \mathbb{R}^n$ is a full dimensional lattice polytope, then if $n \geq 2$, ℓP is very ample for all $\ell \geq n-1$.*

Proof. See [EW91]. □

An immediate corollary of Proposition E.2.4 is that every lattice polygon in $\mathbb{R}^2$ is very ample in $\mathbb{Z}^2$. We are now ready to define the toric variety of a polytope.

Definition E.2.4. Let $P \subset M_{\mathbb{R}} \simeq \mathbb{R}^n$ be a full dimensional lattice polytope. The *toric variety of* P is $X_P = X_{(\ell P) \cap M}$ where ℓ is a positive integer such that ℓP is very ample.

By Proposition E.2.4 we know that such an ℓ for which ℓP is normal always exists. For Definition E.2.4 to make sense, if there are two integers ℓ and ℓ' such that ℓP and $\ell' P$ are very ample, $X_{(\ell P) \cap M}$ and $X_{(\ell' P) \cap M}$ must be the same variety. We will show that they are. They are just embedded in a different projective space. With this definition, the affine pieces of the projective toric variety X_P of a polytope P correspond to strongly convex rational polyhedral cones in $\mathbb{R}^n$.

Theorem E.2.2. *Let X_P be the toric variety of a full-dimensional polytope $P \subset M_{\mathbb{R}} \simeq \mathbb{R}^n$. For each vertex $m_i \in P \cap M$, let $\mathscr{A}_i = (\ell P) \cap M - \ell m_i$ for any $\ell \in \mathbb{N}$ such that ℓP is very ample. Then*

$$X_{P \cap M} \cap U_i = U_{\sigma_i} = \operatorname{MaxSpec}(\mathbb{C}[\sigma_i^\vee \cap M]) \simeq Y_{\mathscr{A}_i}$$

where σ_i is the strongly convex rational polyhedral cone dual to $\operatorname{Cone}(P \cap M - m_i) \subset M_{\mathbb{R}}$. *The dimension of σ_i is n.*

Proof. The theorem follows from the previous discussion and the fact that the semigroup $\mathbb{N}\mathscr{A}$ is saturated in the lattice M if and only if $\mathbb{N}\mathscr{A} = \operatorname{Cone}(\mathscr{A}) \cap M$. For details we refer to [CLS11, §2.3]. □

Using Proposition E.1.3 one shows that for P very ample, the character lattice of the affine piece U_{σ_i} is $\mathbb{Z}\mathsf{S}_i = \mathbb{Z}(\sigma_i^\vee \cap M) = M$, so its torus is $(\mathbb{C}^*)^n$. Then $(\mathbb{C}^*)^n \subset U_{\sigma_i} = X_{P \cap M} \cap U_i \subset X_P$ shows that $(\mathbb{C}^*)^n$ is the torus of X_P.

The affine varieties $Y_{\mathscr{A}_i}$ from Theorem E.2.2 are thought of as toric subvarieties of some affine space. That is, we think of them as *embedded* affine toric varieties via some monomial map. They are isomorphic to the affine varieties $\{U_{\sigma_i}\}_{i \in \mathscr{T}}$ in the affine open covering

$$X_P = \bigcup_{i \in \mathscr{T}} U_{\sigma_i}.$$

The cones $\sigma_i, i \in \mathscr{T}$ are exactly the maximal cones in the normal fan Σ_P of P (see Section D.3). We will now show how the normal fan Σ_P encodes the gluing data that is used to glue X_P from the affine varieties $\{Y_{\mathscr{A}_i}\}_{i\in\mathscr{T}}$.

Let us take a closer look at the smaller open subsets $U_{\sigma_i} \cap U_{\sigma_j} \subset X_P$, $i, j \in \mathscr{T}$. These are again affine and their algebras are $\mathbb{C}[\mathsf{S}_i]_{\chi^{m_j - m_i}}$. This is again a normal semigroup algebra $\mathbb{C}[\mathsf{S}_{ij}]$, but $\tau_{ij}^\vee = \mathrm{Cone}(\mathsf{S}_{ij})$ is no longer pointed: $(m_j - m_i) \in \mathsf{S}_{ij} \cap (-\mathsf{S}_{ij})$. However, $\tau_{ij}^\vee$ has dimension n, since it contains $\sigma_i^\vee$ and $\sigma_j^\vee$, which means that the dual τ_{ij} is pointed and of dimension $< n$ [CLS11, Proposition 1.2.12]. The fact that $\tau_{ij}^\vee$ contains $\sigma_i^\vee$ and $\sigma_j^\vee$ also implies that τ_{ij} is contained in the intersection $\sigma_i \cap \sigma_j$. In fact, the other inclusion also holds.

Proposition E.2.5. *Let m_i, m_j be vertices of a full dimensional lattice polytope P and let $U_{\sigma_i}, U_{\sigma_j}$ be the corresponding affine open subsets of X_P. We have that*

$$U_{\sigma_i} \cap U_{\sigma_j} = \mathrm{MaxSpec}(\mathbb{C}[\tau_{ij}^\vee \cap M]) = U_{\tau_{ij}}$$

where $\tau_{ij} = \sigma_i \cap \sigma_j$ is the cone in the normal fan Σ_P of P corresponding to the smallest face of P containing both m_i and m_j.

Example E.2.3. To illustrate Proposition E.2.5, consider the polytope shown in Figure E.1 and its two vertices m_i and m_j. The associated semigroups are represented at the bottom of the figure by blue dots. Localizing the semigroup algebras $\mathbb{C}[\mathsf{S}_i]$ and $\mathbb{C}[\mathsf{S}_j]$ at $\chi^{m_j - m_i}$ and $\chi^{m_i - m_j}$ respectively, we obtain the algebras over the semigroups formed by the union of the blue and the orange dots. The figure shows that $\mathbb{C}[\mathsf{S}_i]_{\chi^{m_j - m_i}} = \mathbb{C}[\mathsf{S}_j]_{\chi^{m_i - m_j}}$. The resulting semigroup is the intersection of the closed halfspace $\tau_{ij}^\vee$ with the lattice, where $\tau_{ij} = \sigma_i \cap \sigma_j$, see Figure E.2. △

To describe the gluing in the notation of Section 2.3, for $i, j \in \mathscr{T}$ we set

$$Y_i = Y_{\mathscr{A}_i}, \quad \text{and} \quad Y_{ij} = (Y_{\mathscr{A}_i})_{\chi^{m_j - m_i}}$$

and $\phi_{ij} : Y_{ij} \to Y_{ji}$ is given by $\mathbb{C}[\mathsf{S}_i]_{\chi^{m_j - m_i}} = \mathbb{C}[\mathsf{S}_j]_{\chi^{m_i - m_j}}$. One can check that this data satisfies conditions 1-3 from Section 2.3. We obtain the abstract variety

$$X_P = \bigsqcup_{i \in \mathscr{T}} Y_i \Big/ \sim$$

which is isomorphic to $X_{\ell P \cap M}$ for each $\ell \in \mathbb{N}$ such that ℓP is very ample. We illustrate this construction with some examples.

Example E.2.4 (The gluing of $\mathbb{P}^1$ revisited)**.** Consider the polytope $[0, 1] \subset \mathbb{R}$. Its normal fan is supported on the real line $\mathbb{R}$ with cones $(-\infty, 0], \{0\}, [0, \infty)$. The maximal cones are $\sigma_1 = (-\infty, 0]$ and $\sigma_0 = [0, \infty)$ and they correspond to the vertices $m_1 = 1$ and $m_0 = 0$ respectively. These cones are self-dual, and their algebras are $\mathbb{C}[\mathsf{S}_1] = \mathbb{C}[\sigma_1^\vee \cap \mathbb{Z}] = \mathbb{C}[-\mathbb{N}] = \mathbb{C}[u]$ and $\mathbb{C}[\mathsf{S}_0] = \mathbb{C}[\mathbb{N}] = \mathbb{C}[t]$. We see that the affine varieties corresponding to the vertices of P are two copies of $\mathbb{C}$. By setting

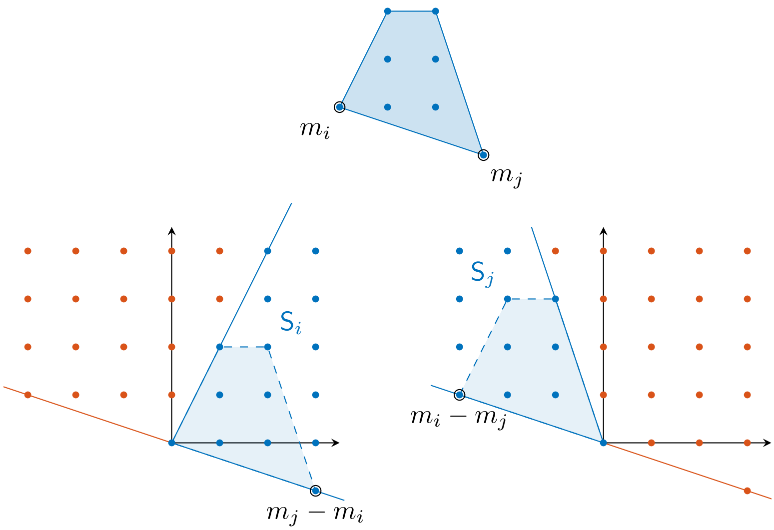


Figure E.1: Polytope and semigroups from Example E.2.3.

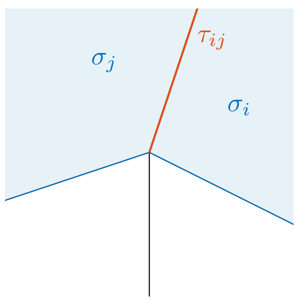


Figure E.2: Normal fan of the polytope in Figure E.1 with relevant cones highlighted.

$\mathbb{C}[-\mathbb{N}] = \mathbb{C}[u]$, we identify the variable u with the character $\chi^{-1} = \chi^{0-1} = \chi^{m_0-m_1}$. Hence $\mathbb{C}[\mathsf{S}_1]_{\chi^{m_0-m_1}} = \mathbb{C}[u]_u$ and analogously we find $\mathbb{C}[\mathsf{S}_0]_{\chi^{m_1-m_0}} = \mathbb{C}[t]_t$. The isomorphism $\mathbb{C}[u]_u \to \mathbb{C}[t]_t$ is given by $u/1 \mapsto 1/t$, and therefore $\phi_{01} : \mathbb{C}^* \to \mathbb{C}^*$ is given by $\phi_{01}(t) = t^{-1}$. We conclude that the two copies of $\mathbb{C}$ are glued together in a way identical to Example 2.3.1, and thus $X_{[0,1]} = \mathbb{P}^1$. More generally, the toric variety of the n-dimensional elementary simplex Δ_n is $X_{\Delta_n} = \mathbb{P}^n$. This makes sense because Δ_n is very ample and the image of $\phi_{\Delta_n \cap \mathbb{Z}^n} : (\mathbb{C}^*)^n \to \mathbb{P}^n$ is dense in $\mathbb{P}^n$. For a dilation $\ell\Delta_n$ of the elementary simplex, the closure of the image of $\phi_{\ell\Delta_n \cap \mathbb{Z}^n}$ is the ℓ-th *Veronese embedding* of $\mathbb{P}^n$, which shows that we obtain the same abstract variety, embedded in a different projective space. △

Example E.2.5 ($\mathbb{P}^1 \times \mathbb{P}^1$). For the polytope $P = [0,1]^2 \subset \mathbb{R}^2$, we know that $X_P = X_{\mathscr{A}}$

from Example E.2.1 and we saw that $X_P = \mathbb{P}^1 \times \mathbb{P}^1$. In particular, $X_{\mathscr{A}}$ from Example E.2.1 is a Segre embedding of $\mathbb{P}^1 \times \mathbb{P}^1$. To see from the gluing construction above that $X_P \simeq \mathbb{P}^1 \times \mathbb{P}^1$, denote the vertices of P by

$$m_1 = (0,0), \quad m_2 = (1,0), \quad m_3 = (1,1), \quad m_4 = (0,1).$$

These vertices give 4 affine toric varieties $Y_1, \dots, Y_4$ where Y_i corresponds to m_i, each of which is a copy of $\mathbb{C}^2$. We use the identification

$$\begin{aligned} Y_1 &= \operatorname{MaxSpec}(\mathbb{C}[\chi^{(1,0)}, \chi^{(0,1)}]), & Y_2 &= \operatorname{MaxSpec}(\mathbb{C}[\chi^{(-1,0)}, \chi^{(0,1)}]), \\ Y_3 &= \operatorname{MaxSpec}(\mathbb{C}[\chi^{(-1,0)}, \chi^{(0,-1)}]), & Y_4 &= \operatorname{MaxSpec}(\mathbb{C}[\chi^{(1,0)}, \chi^{(0,-1)}]). \end{aligned}$$

The isomorphisms ϕ_{ij} for $i = 1$ are given by

$$\begin{aligned} \phi_{11}(t_1,t_2) &= (t_1,t_2), & \phi_{12}(t_1,t_2) &= (t_1^{-1},t_2), \\ \phi_{13}(t_1,t_2) &= (t_1^{-1},t_2^{-1}), & \phi_{14}(t_1,t_2) &= (t_1,t_2^{-1}). \end{aligned}$$

Note that the overlap of $U_{\sigma_1} \simeq Y_1$ and U_{σ_3} is $(\mathbb{C}^*)^2$ (the cones intersect in a single point, the origin, whose toric variety is $(\mathbb{C}^*)^2$), so the map $\phi_{13} : (\mathbb{C}^*)^2 \to (\mathbb{C}^*)^2$ is well-defined on this overlap. The intersection of σ_1 with the cones σ_2 and σ_4 are rays, which can be seen from the polytope by the fact that m_1 and m_2/m_4 are connected by an edge. Now let $(x_0 : x_1, y_0 : y_1)$ be homogeneous coordinates on $\mathbb{P}^1 \times \mathbb{P}^1$ and for $0 \le i,j \le 1$ let

$$U_{ij} = \{(x_0 : x_1, y_0 : y_1) \in \mathbb{P}^1 \times \mathbb{P}^1 \mid x_i \neq 0 \text{ and } y_j \neq 0\}.$$

The open subsets $U_1 = U_{00}, U_2 = U_{10}, U_3 = U_{11}, U_4 = U_{01}$ cover $\mathbb{P}^1 \times \mathbb{P}^1$. We identify these open subsets with $Y_1, \dots Y_4$ by

$$\begin{aligned} h_1 : U_1 \to Y_1 &\quad \text{where} \quad (x_0 : x_1, y_0 : y_1) \mapsto (x_1/x_0, y_1/y_0), \\ h_2 : U_2 \to Y_2 &\quad \text{where} \quad (x_0 : x_1, y_0 : y_1) \mapsto (x_0/x_1, y_1/y_0), \\ h_3 : U_3 \to Y_3 &\quad \text{where} \quad (x_0 : x_1, y_0 : y_1) \mapsto (x_0/x_1, y_0/y_1), \\ h_4 : U_4 \to Y_4 &\quad \text{where} \quad (x_0 : x_1, y_0 : y_1) \mapsto (x_1/x_0, y_0/y_1). \end{aligned}$$

This gives an isomorphism $\mathbb{P}^1 \times \mathbb{P}^1 \to \bigsqcup_{i=1}^4 Y_i/\sim$ given by

$$p \mapsto [(h_i(p), Y_i)] \quad \text{for any } i \text{ such that } p \in U_i,$$

where $[\cdot]$ denotes the equivalence class in $X_P = \bigsqcup_{i=1}^4 Y_i/\sim$. △

The statement that the construction presented here does not depend on which very ample dilate ℓP of P we consider can be generalized. In fact, the construction *only depends on the fan* Σ_P. Different polytopes P may have the same normal fan, and for that they do not have to be dilated versions of each other (consider for instance a square and a rectangle in $\mathbb{R}^2$). For this reason, the variety X_P is sometimes denoted X_{Σ_P}. In fact, any fan $\Sigma \in N_{\mathbb{R}}$ gives a normal toric variety X_Σ. We will not discuss this in full generality. The fans we will encounter are complete and they come from a polytope.

Since the toric variety X_P always contains the torus $(\mathbb{C}^*)^n$ as a dense open subset (this is the intersection of all the open subsets $U_{\sigma_i}, i \in \mathscr{T}$ corresponding to the cone $\{0\}$ in Σ_P), we can think of X_P as '$(\mathbb{C}^*)^n$ plus its *boundary*'. The way this boundary looks like is completely encoded by the polytope P, and by its normal fan Σ_P. The dense torus $(\mathbb{C}^*)^n \subset X_P$ is an orbit of the action of $(\mathbb{C}^*)^n$ on X_P. The following nice result shows that X_P can be decomposed as a disjoint union of torus orbits, each of which corresponds to a face of P or, equivalently, to a cone in Σ_P.

Theorem E.2.3 (The orbit-(cone/face) correspondence)**.** *Let $X_P = X_{\Sigma_P}$ be the toric variety of a full-dimensional polytope $P \subset M_{\mathbb{R}} \simeq \mathbb{R}^n$. The following statements hold.*

1. *There is a one-to-one correspondence between faces $Q \subset P$, cones $\sigma \in \Sigma_P$ and $(\mathbb{C}^*)^n$-orbits in X_P. For a cone $\sigma \in \Sigma_P$, we denote the corresponding $(\mathbb{C}^*)^n$-orbit by $O(\sigma) \subset X_P$.*
2. *For each $\sigma \in \Sigma_P$, $\dim O(\sigma) = n - \dim \sigma$.*
3. *For each $\sigma \in \Sigma_P$, the affine open subset $U_\sigma \subset X_P$ can be written as*

$$U_\sigma = \bigcup_{\tau \textit{ face of } \sigma} O(\tau)$$

   and the closure $\overline{O(\sigma)}$ in X_P with respect to both the classical and the Zariski topology is

$$\overline{O(\sigma)} = \bigcup_{\sigma \textit{ face of } \tau} O(\tau).$$

Proof. This is Theorem 3.2.6 in [CLS11]. □

Bibliography

[ABB$^+$99] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999.

[AG12] E. L. Allgower and K. Georg. *Numerical continuation methods: an introduction*, volume 13. Springer Science & Business Media, 2012.

[AGH$^+$17] B. Assarf, E. Gawrilow, K. Herr, M. Joswig, B. Lorenz, A. Paffenholz, and T. Rehn. Computing convex hulls and counting integer points with `polymake`. *Math. Program. Comput.*, 9(1):1–38, 2017.

[AL94] W. W. Adams and P. Loustaunau. *An introduction to Gröbner bases*. Number 3. American Mathematical Soc., 1994.

[AL11] M. F. Anjos and J. B. Lasserre. *Handbook on semidefinite, conic and polynomial optimization*, volume 166. Springer Science & Business Media, 2011.

[AM69] M. F. Atiyah and I. G. Macdonald. Introduction to commutative algebra, 1969.

[AMVW15] J. L. Aurentz, T. Mach, R. Vandebril, and D. S. Watkins. Fast and backward stable computation of roots of polynomials. *SIAM Journal on Matrix Analysis and Applications*, 36(3):942–973, 2015.

[AS88] W. Auzinger and H. J. Stetter. An elimination algorithm for the computation of all zeros of a system of multivariate polynomial equations. In *Numerical Mathematics Singapore 1988*, pages 11–30. Springer, 1988.

[Atk72] F. V. Atkinson. *Multiparameter eigenvalue problems*. Academic Press New York, 1972.

[Aud12] M. Audin. *The topology of torus actions on symplectic manifolds*, volume 93. Birkhäuser, 2012.

[Bat13] K. Batselier. *A numerical linear algebra framework for solving problems with multivariate polynomials*. PhD thesis, KU Leuven, 2013.

[BBV19] C. Beltrán, P. Breiding, and N. Vannieuwenhoven. Pencil-based algorithms for tensor rank decomposition are not stable. *SIAM Journal on Matrix Analysis and Applications*, 40(2):739–773, 2019.

[BC94] V. V. Batyrev and D. A. Cox. On the Hodge structure of projective hypersurfaces in toric varieties. *Duke Math. J.*, 75(2):293–338, 08 1994.

[BC13] P. Bürgisser and F. Cucker. *Condition: The geometry of numerical algorithms*, volume 349. Springer Science & Business Media, 2013.

[BCDM+06] L. Bos, M. Caliari, S. De Marchi, M. Vianello, and Y. Xu. Bivariate Lagrange interpolation at the Padua points: the generating curve approach. *Journal of Approximation Theory*, 143(1):15–25, 2006.

[BCMT10] J. Brachat, P. Comon, B. Mourrain, and E. Tsigaridas. Symmetric tensor decomposition. *Linear Algebra and its Applications*, 433(11-12):1851–1872, 2010.

[BCR13] J. Bochnak, M. Coste, and M.-F. Roy. *Real algebraic geometry*, volume 36. Springer Science & Business Media, 2013.

[BCSS12] L. Blum, F. Cucker, M. Shub, and S. Smale. *Complexity and real computation*. Springer Science & Business Media, 2012.

[BDDM14] K. Batselier, P. Dreesen, and B. De Moor. A fast recursive orthogonalization scheme for the Macaulay matrix. *Journal of Computational and Applied Mathematics*, 267:20–32, 2014.

[Bea68] A. F. Beardon. On the location of poles of Padé approximants. *Journal of Mathematical Analysis and Applications*, 21(3):469–474, 1968.

[Ben19] M. R. Bender. *Algorithms for sparse polynomial systems: Gröbner basis and resultants*. PhD thesis, Sorbonne Université, 2019.

[Ber75] D. N. Bernstein. The number of roots of a system of equations. *Functional Analysis and its applications*, 9(3):183–185, 1975.

[Béz79] E. Bézout. *Théorie générale des équations algébriques*. Ph.-D. Pierres, 1779.

[BFT18] M. R. Bender, J.-C. Faugère, and E. Tsigaridas. Towards mixed Gröbner basis algorithms: The multihomogeneous and sparse case. In *Proceedings of the 2018 ACM International Symposium on Symbolic and Algebraic Computation*, pages 71–78, 2018.

[BFT19] M. R. Bender, J.-C. Faugère, and E. Tsigaridas. Gröbner basis over semigroup algebras: Algorithms and applications for sparse polynomial systems. In *Proceedings of the 2019 on International Symposium on Symbolic and Algebraic Computation*, pages 42–49, 2019.

[BG65] P. Businger and G. H. Golub. Linear least squares solutions by Householder transformations. *Numerische Mathematik*, 7(3):269–276, 1965.

[BHSW08] D. J. Bates, J. D. Hauenstein, A. J. Sommese, and C. W. Wampler. Adaptive multiprecision path tracking. *SIAM J. Numer. Anal.*, 46(2):722–746, 2008.

[Bie55] L. Bieberbach. *Analytische Fortsetzung*, volume 3 of *Ergebnisse der Mathematik und Ihrer Grenzgebiete*. Springer-Verlag, 1955.

[BJA07] M. Byröd, K. Josephson, and K. Aström. Improving numerical accuracy of Gröbner basis polynomial equation solvers. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[BJA08] M. Byröd, K. Josephson, and K. Aström. A column-pivoting based strategy for monomial ordering in numerical Gröbner basis calculations. In *European Conference on Computer Vision*, pages 130–143. Springer, 2008.

[BJGM96] G. A. Baker Jr and P. Graves-Morris. *Padé Approximants*, volume 59. Cambridge University Press, 1996.

[BKM05] L. Busé, H. Khalil, and B. Mourrain. Resultant-based methods for plane curves intersection problems. In *International Workshop on Computer Algebra in Scientific Computing*, pages 75–92. Springer, 2005.

[BL13] C. Beltrán and A. Leykin. Robust certified numerical homotopy tracking. *Foundations of Computational Mathematics*, 13(2):253–295, 2013.

[BM15] B. Beckermann and A. C. Matos. Algebraic properties of robust Padé approximants. *Journal of Approximation Theory*, 190:91–115, 2015.

[Bre20] P. Breiding. An algebraic geometry perspective on topological data analysis. *arXiv preprint arXiv:2001.02098*, 2020.

[BS87] D. Bayer and M. Stillman. A criterion for detecting m-regularity. *Inventiones mathematicae*, 87(1):1–11, 1987.

[BSHW13] D. J. Bates, A. J. Sommese, J. D. Hauenstein, and C. W. Wampler. *Numerically solving polynomial systems with Bertini*. SIAM, 2013.

[BST19] P. Breiding, B. Sturmfels, and S. Timme. 3264 conics in a second. *Notices of the American Mathematical Society*, 2019.

[BT18] P. Breiding and S. Timme. Homotopycontinuation. jl: A package for homotopy continuation in Julia. In *International Congress on Mathematical Software*, pages 458–465. Springer, 2018.

[BT20a] M. R. Bender and S. Telen. Toric eigenvalue methods for solving sparse polynomial systems. *arXiv preprint arXiv:2006.10654*, 2020.

[BT20b] A. Bernardi and D. Taufer. Waring, tangential and cactus decompositions. *Journal de Mathématiques Pures et Appliquées*, 2020.

[Buc70] B. Buchberger. Ein algorithmisches kriterium für die lösbarkeit eines algebraischen gleichungssystems. *Aequationes mathematicae*, 4(3):374–383, 1970.

[Buc06] B. Buchberger. Bruno buchbergers PhD thesis 1965: An algorithm for finding the basis elements of the residue class ring of a zero dimensional polynomial ideal. *Journal of symbolic computation*, 41(3-4):475–511, 2006.

[Bul06] A. Bultheel. *Inleiding tot de numerieke wiskunde.* Acco, 2006.

[Bus01] L. Busé. Residual resultant over the projective plane and the implicitization problem. In *Proceedings of the 2001 international symposium on Symbolic and algebraic computation*, pages 48–55, 2001.

[BV18a] N. Bliss and J. Verschelde. The method of Gauss–Newton to compute power series solutions of polynomial homotopies. *Linear Algebra and its Applications*, 542:569–588, 2018.

[BV18b] P. Breiding and N. Vannieuwenhoven. The condition number of join decompositions. *SIAM Journal on Matrix Analysis and Applications*, 39(1):287–309, 2018.

[BvDD+17] A. Boralevi, J. van Doornmalen, J. Draisma, M. E. Hochstenbach, and B. Plestenjak. Uniform determinantal representations. *SIAM journal on applied algebra and geometry*, 1(1):415–441, 2017.

[Cay64] A. Cayley. Nouvelles recherches sur l'élimination et la théorie des courbes. *Journal für die reine und angewandte Mathematik*, 63:34–39, 1864.

[CCC+05] E. Cattani, D. A. Cox, G. Chèze, A. Dickenstein, M. Elkadi, I. Z. Emiris, A. Galligo, A. Kehrein, M. Kreuzer, and B. Mourrain. *Solving polynomial equations: foundations, algorithms, and applications (algorithms and computation in mathematics).* Springer-Verlag, 2005.

[CDSS09] G. Craciun, A. Dickenstein, A. Shiu, and B. Sturmfels. Toric dynamical systems. *Journal of Symbolic Computation*, 44(11):1551–1565, 2009.

[CE93] J. Canny and I. Z. Emiris. An efficient algorithm for the sparse mixed resultant. In *International Symposium on Applied Algebra, Algebraic Algorithms, and Error-Correcting Codes*, pages 89–104. Springer, 1993.

[CGPS08] G. Craciun, L. D. García-Puente, and F. Sottile. Some geometrical aspects of control points for toric patches. In *International Conference on Mathematical Methods for Curves and Surfaces*, pages 111–135. Springer, 2008.

[CGT97] R. M. Corless, P. M. Gianni, and B. M. Trager. A reordered Schur factorization method for zero-dimensional polynomial systems with multiple roots. In *Proceedings of the 1997 international symposium on Symbolic and algebraic computation*, pages 133–140, 1997.

[CH92] T. F. Chan and P. C. Hansen. Some applications of the rank revealing QR factorization. *SIAM Journal on Scientific and Statistical Computing*, 13(3):727–741, 1992.

[CI94] S. Chandrasekaran and I. C. Ipsen. On rank-revealing factorisations. *SIAM Journal on Matrix Analysis and Applications*, 15(2):592–622, 1994.

[CLO06] D. A. Cox, J. B. Little, and D. O'Shea. *Using algebraic geometry*, volume 185 of *Graduate Texts in Mathematics*. Springer Science & Business Media, 2006.

[CLO13] D. A. Cox, J. B. Little, and D. O'Shea. *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer Science & Business Media, 2013.

[CLS11] D. A. Cox, J. B. Little, and H. K. Schenck. *Toric varieties*. American Mathematical Soc., 2011.

[CMDL+15] A. Cichocki, D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and H. A. Phan. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE signal processing magazine*, 32(2):145–163, 2015.

[Com02] P. Comon. Tensor decompositions. *Mathematics in Signal Processing V*, pages 1–24, 2002.

[Cox95] D. A. Cox. The homogeneous coordinate ring of a toric variety. *Journal of Algebraic Geometry*, 4:17–50, 1995.

[Cox20a] D. A. Cox. *Applications of Polynomial Systems*. CBMS Regional Conference Series in Mathematics. Conference Board of the Mathematical Sciences, 2020.

[Cox20b] D. A. Cox. Stickelberger and the eigenvalue theorem. *arXiv preprint arXiv:2007.12573*, 2020.

[Cut18] S. D. Cutkosky. *Introduction to Algebraic Geometry*, volume 188. American Mathematical Soc., 2018.

[D'A02] C. D'Andrea. Macaulay style formulas for sparse resultants. *Transactions of the Americal Mathematical Society*, 354(7):2595–2619, 2002.

[DB04] C. De Boor. Ideal interpolation. *Approximation Theory XI: Gatlinburg*, pages 59–91, 2004.

[DBDM12] P. Dreesen, K. Batselier, and B. De Moor. Back to the roots: Polynomial system solving, linear algebra, systems theory. *IFAC Proceedings Volumes*, 45(16):1203–1208, 2012.

[DHJ$^+$19] T. Duff, C. Hill, A. Jensen, K. Lee, A. Leykin, and J. Sommars. Solving polynomial systems via homotopy continuation and monodromy. *IMA Journal of Numerical Analysis*, 39(3):1421–1446, 2019.

[DHO$^+$16] J. Draisma, E. Horobeţ, G. Ottaviani, B. Sturmfels, and R. R. Thomas. The Euclidean distance degree of an algebraic variety. *Foundations of computational mathematics*, 16(1):99–149, 2016.

[Dic16] A. Dickenstein. Biochemical reaction networks: An invitation for algebraic geometers. In *Mathematical Congress of the Americas*, volume 656, pages 65–83. Contemp. Math, 2016.

[Die57] P. Dienes. *The Taylor series: an introduction to the theory of functions of a complex variable*. Dover New York, 1957.

[DL06] L. De Lathauwer. A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM journal on Matrix Analysis and Applications*, 28(3):642–666, 2006.

[dM02] R. de Montessus. Sur les fractions continues algébriques. *Bulletin de la Société Mathématique de France*, 30:28–36, 1902.

[Dre13] P. Dreesen. *Back to the roots: polynomial system solving using linear algebra*. PhD thesis, KU Leuven, 2013.

[DS15] C. D'Andrea and M. Sombra. A Poisson formula for the sparse resultant. *Proceedings of the London Mathematical Society*, 110(4):932–964, 2015.

[DSL08] V. De Silva and L.-H. Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1084–1127, 2008.

[DYY16] B. Dong, B. Yu, and Y. Yu. A homotopy method for finding all solutions of a multiparameter eigenvalue problem. *SIAM Journal on Matrix Analysis and Applications*, 37(2):550–571, 2016.

[DZ05] B. H Dayton and Z. Zeng. Computing the multiplicity structure in solving polynomial systems. In *Proceedings of the 2005 international symposium on Symbolic and algebraic computation*, pages 116–123, 2005.

[EC93] I. Z. E. and J. Canny. A practical method for the sparse resultant. In *Proceedings of the 1993 international symposium on Symbolic and algebraic computation*, pages 183–192, 1993.

[EC95] I. Z. Emiris and J. F. Canny. Efficient incremental algorithms for the sparse resultant and the mixed volume. *Journal of Symbolic Computation*, 20(2):117–149, 1995.

[EdW19] A. A. Ergür and T. de Wolff. A polyhedral homotopy algorithm for real zeros. *arXiv preprint arXiv:1910.01957*, 2019.

[EGSS01] D. Eisenbud, D. R. Grayson, M. Stillman, and B. Sturmfels. *Computations in algebraic geometry with Macaulay 2*, volume 8. Springer Science & Business Media, 2001.

[EH06] D. Eisenbud and J. Harris. *The geometry of schemes*, volume 197. Springer Science & Business Media, 2006.

[EH16] D. Eisenbud and J. Harris. *3264 and all that: A second course in algebraic geometry.* Cambridge University Press, 2016.

[Eis13] D. Eisenbud. *Commutative Algebra: with a view toward algebraic geometry*, volume 150. Springer Science & Business Media, 2013.

[EM99a] I. Z. Emiris and B. Mourrain. Computer algebra methods for studying and computing molecular conformations. *Algorithmica*, 25(2-3):372–402, 1999.

[EM99b] I. Z. Emiris and B. Mourrain. Matrices in elimination theory. *Journal of Symbolic Computation*, 28(1-2):3–44, 1999.

[EM07] M. Elkadi and B. Mourrain. *Introduction à la résolution des systèmes polynomiaux*, volume 59. Springer, 2007.

[Emi96] I. Z. Emiris. On the complexity of sparse elimination. *Journal of Complexity*, 12(2):134–166, 1996.

[EW91] G. Ewald and U. Wessels. On the ampleness of invertible sheaves in complete projective toric varieties. *Results in Mathematics*, 19(3-4):275–278, 1991.

[EY36] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[Fab96] E. Fabry. Sur les points singuliers d'une fonction donnée par son développement en série et l'impossibilité du prolongement analytique dans des cas très généraux. In *Annales scientifiques de l'École Normale Supérieure*, volume 13, pages 367–399, 1896.

[Fau99] J.-C. Faugère. A new efficient algorithm for computing Gröbner bases (F4). *Journal of pure and applied algebra*, 139(1-3):61–88, 1999.

[Fau02] J.-C. Faugère. A new efficient algorithm for computing Gröbner bases without reduction to zero (F5). In *Proceedings of the 2002 international symposium on Symbolic and algebraic computation*, pages 75–83, 2002.

[Fau10] J.-C. Faugère. FGb: a library for computing Gröbner bases. In *International Congress on Mathematical Software*, pages 84–87. Springer, 2010.

[Ful93] W. Fulton. *Introduction to toric varieties*. Number 131. Princeton University Press, 1993.

[Gat90] K. Gatermann. Symbolic solution of polynomial equation systems with symmetry. In Sh. Watanabe and M. Nagata, editors, *Proceedings of ISSAC-90 (Tokyo, Japan, August 20–24, 1990)*, pages 112–119. ACM, 1990.

[GE96] M. Gu and S. C. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

[GGT13] P. Gonnet, S. Güttel, and L. N. Trefethen. Robust Padé approximation via SVD. *SIAM review*, 55(1):101–117, 2013.

[GKZ94] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky. *Discriminants, Resultants, and Multidimensional Determinants*. Mathematics (Birkhäuser). Springer, 1994.

[Gon81] A. A. Gončar. Poles of rows of the Padé table and meromorphic continuation of functions. *Matematicheskii Sbornik*, 157(4):590–613, 1981.

[Grü13] B. Grünbaum. *Convex polytopes*, volume 221. Springer Science & Business Media, 2013.

[GS] D. R. Grayson and M. E. Stillman. Macaulay2, a software system for research in algebraic geometry. Available at `http://www.math.uiuc.edu/Macaulay2/`.

[GS04] J. J. Gervais and H. Sadiky. A continuation method based on a high order predictor and an adaptive steplength control. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik: Applied Mathematics and Mechanics*, 84(8):551–563, 2004.

[GT09] S. Graillat and P. Trébuchet. A new algorithm for computing certified numerical approximations of the roots of a zero-dimensional system. In *Proceedings of the 2009 international symposium on Symbolic and algebraic computation*, pages 167–174, 2009.

[GT17] S. Güttel and F. Tisseur. The nonlinear eigenvalue problem. *Acta Numerica*, 26:1–94, 2017.

[GV08] Y. Guan and J. Verschelde. PHClab: a MATLAB/Octave interface to PHCpack. In *Software for Algebraic Geometry*, pages 15–32. Springer, 2008.

[GVL12] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU press, 2012.

[Har77] R. Hartshorne. *Algebraic geometry.* Springer-Verlag, New York, 1977. Graduate Texts in Mathematics, No. 52.

[Har13] J. Harris. *Algebraic geometry: a first course*, volume 133. Springer Science & Business Media, 2013.

[Hig02] N. J. Higham. *Accuracy and stability of numerical algorithms*, volume 80. SIAM, 2002.

[HLJ16] J. D. Hauenstein and A. C. Liddell Jr. Certified predictor–corrector tracking for newton homotopies. *Journal of Symbolic Computation*, 74:239–254, 2016.

[HMP19] M. E. Hochstenbach, C. Mehl, and B. Plestenjak. Solving singular generalized eigenvalue problems by a rank-completing perturbation. *SIAM Journal on Matrix Analysis and Applications*, 40(3):1022–1046, 2019.

[HOOS19] J. D. Hauenstein, L. Oeding, G. Ottaviani, and A. J. Sommese. Homotopy techniques for tensor decomposition and perfect identifiability. *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 2019(753):1–22, 2019.

[HP92] Y. P. Hong and C.-T. Pan. Rank-revealing QR factorizations and the singular value decomposition. *Mathematics of Computation*, 58(197):213–232, 1992.

[HS95] B. Huber and B. Sturmfels. A polyhedral method for solving sparse polynomial systems. *Mathematics of computation*, 64(212):1541–1555, 1995.

[HS97] B. Huber and B. Sturmfels. Bernstein's theorem in affine space. *Discrete & Computational Geometry*, 17(2):137–141, 1997.

[HS12] J. D. Hauenstein and F. Sottile. Algorithm 921: alphacertified: certifying solutions to polynomial systems. *ACM Transactions on Mathematical Software (TOMS)*, 38(4):1–20, 2012.

[HSS98] B. Huber, F. Sottile, and B. Sturmfels. Numerical schubert calculus. *Journal of Symbolic Computation*, 26(6):767–788, 1998.

[Hum12] J. E. Humphreys. *Linear algebraic groups*, volume 21. Springer Science & Business Media, 2012.

[HV98] B. Huber and J. Verschelde. Polyhedral end games for polynomial continuation. *Numerical Algorithms*, 18(1):91–108, 1998.

[HZ03] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision.* Cambridge university press, 2003.

[IA13] O. L. Ibryaeva and V. M. Adukov. An algorithm for computing a Padé approximant with minimal degree denominator. *Journal of Computational and Applied Mathematics*, 237:529–541, 2013.

[IK99] A. Iarrobino and V. Kanev. *Power sums, Gorenstein algebras, and determinantal loci.* Springer Science & Business Media, 1999.

[Jou91] J.-P. Jouanolou. Le formalisme du résultant. *Advances in Mathematics*, 90(2):117–263, 1991.

[JV05] G. Jónsson and S. Vavasis. Accurate solution of polynomial equations using Macaulay resultant matrices. *Mathematics of computation*, 74(249):221–262, 2005.

[Kah66] W. Kahan. Numerical linear algebra. *Canadian Mathematical Bulletin*, 9(5):757–801, 1966.

[Kat90] S. Katsura. Spin glass problem by the method of integral equation of the effective field. In M.D. Coutinho-Filho and S.M. Resende, editors, *New Trends in Magnetism*, pages 110–121. World Scientific, 1990.

[Kat94] S. Katsura. Users posing problems to PoSSO. In the PoSSO Newsletter, no. 2, edited by L. Gonzelez-Vega and T. Recio., 1994.

[KB09] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[Kho77] A. G. Khovanskii. Newton polytopes and toric varieties. *Functional Anal. Appl*, 11:289–298, 1977.

[Kho92] A. G. Khovanskii. Newton polyhedron, Hilbert polynomial, and sums of finite sets. *Functional Analysis and Its Applications*, 26(4):276–281, 1992.

[KK05] A. Kehrein and M. Kreuzer. Characterizations of border bases. *Journal of Pure and Applied Algebra*, 196(2-3):251–270, 2005.

[KK06] A. Kehrein and M. Kreuzer. Computing border bases. *Journal of Pure and Applied Algebra*, 205(2):279–295, 2006.

[KKR05] A. Kehrein, M. Kreuzer, and L. Robbiano. An algebraists view on border bases. In *Solving polynomial equations*, pages 169–202. Springer, 2005.

[KL18] Y.-C. Kuo and T.-L. Lee. Computing the unique candecomp/parafac decomposition of unbalanced tensors by homotopy method. *Linear Algebra and its Applications*, 556:238–264, 2018.

[KLT20] M. Kaluba, B. Lorenz, and S. Timme. Polymake. jl: A new interface to polymake. *arXiv preprint arXiv:2003.11381*, 2020.

[KP07] Z. Kukelova and T. Pajdla. A minimal solution to the autocalibration of radial distortion. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2007.

[Kuk13] Z. Kukelova. Algebraic methods in computer vision. 2013. Ph. D. thesis.

[Kul20] A. Kulkarni. Solving p-adic polynomial systems via iterative eigenvector algorithms. *Linear and Multilinear Algebra*, pages 1–22, 2020.

[Kus76a] A. G. Kushnirenko. Newton polytopes and the Bézout theorem. *Functional analysis and its applications*, 10(3):233–235, 1976.

[Kus76b] A. G. Kushnirenko. Polyèdres de Newton et nombres de Milnor. *Inventiones mathematicae*, 32(1):1–31, 1976.

[KX94] R. B. Kearfott and Z. Xing. An interval step control for continuation methods. *SIAM Journal on Numerical Analysis*, 31(3):892–914, 1994.

[Lan02] S. Lang. *Algebra*, volume 211. Springer-Verlag, 2002.

[Li97] T.-Y. Li. Numerical solution of multivariate polynomial systems by homotopy continuation methods. *Acta numerica*, 6:399–436, 1997.

[Lip76] J. D. Lipson. Newton's method: a great algebraic algorithm. In *Proceedings of the third ACM symposium on Symbolic and algebraic computation*, pages 260–270, 1976.

[Lju86] L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, Inc., USA, 1986.

[LLM$^+$13] J.-B. Lasserre, M. Laurent, B. Mourrain, P. Rostalski, and P. Trébuchet. Moment matrices, border bases and real radical computation. *Journal of Symbolic Computation*, 51:63–85, 2013.

[LLT08] T.-L. Lee, T.-Y. Li, and C.-H. Tsai. HOM4PS-2.0: a software package for solving polynomial systems by the polyhedral homotopy continuation method. *Computing*, 83(2-3):109, 2008.

[LT09] T. Y. Li and C. H. Tsai. HOM4PS-2.0para: Parallelization of HOM4PS-2.0 for solving polynomial systems. *Parallel Computing*, 35(4):226–238, 2009.

[Mac02] F. S. Macaulay. Some formulae in elimination. *Proceedings of the London Mathematical Society*, 1(1):3–27, 1902.

[Mac94] F. S. Macaulay. *The algebraic theory of modular systems*, volume 19. Cambridge University Press, 1994.

[Map18] Maplesoft. Maple, a division of waterloo maple inc. *Waterloo, Ontario*, 2018.

[Mas80] J. C. Mason. Near-best multivariate approximation by Fourier series, Chebyshev series and Chebyshev interpolation. *Journal of Approximation Theory*, 28(4):349–358, 1980.

[Mas16] C. Massri. Solving a sparse system using linear algebra. *Journal of Symbolic Computation*, 73:157–174, 2016.

[MAT17] MATLAB. *version 9.2.0.556344 (R2017a).* The MathWorks Inc., Natick, Massachusetts, 2017.

[Mau80] J. Maurer. Puiseux expansion for space curves. *Manuscripta Mathematica*, 32(1-2):91–100, 1980.

[MM12] William D. Mac M. A method for determining the solutions of a system of analytic functions in the neighborhood of a branch point. *Mathematische Annalen*, 72(2):180–202, 1912.

[MMM91] M. G. Marinari, M. Möller, and T. Mora. Gröbner bases of ideals given by dual bases. In *Proceedings of the 1991 international symposium on Symbolic and algebraic computation*, pages 55–63, 1991.

[MMM93] M. G. Marinari, M. Möller, and T. Mora. Gröbner bases of ideals defined by functionals with an application to ideals of projective points. *Applicable Algebra in Engineering, Communication and Computing*, 4(2):103–145, 1993.

[Möl93] M. Möller. Systems of algebraic equations solved by means of endomorphisms. In *International Symposium on Applied Algebra, Algebraic Algorithms, and Error-Correcting Codes*, pages 43–56. Springer, 1993.

[Mor09] A. Morgan. *Solving polynomial systems using continuation for engineering and scientific problems*, volume 57 of *Classics in Applied Mathematics*. SIAM, 2009.

[Mou99] B. Mourrain. A new criterion for normal form algorithms. In *International Symposium on Applied Algebra, Algebraic Algorithms, and Error-Correcting Codes*, pages 430–442. Springer, 1999.

[Mou07] B. Mourrain. Pythagores dilemma, symbolic-numeric computation, and the border basis method. In *Symbolic-Numeric Computation*, pages 223–243. Springer, 2007.

[MP09] B. Mourrain and J. P. Pavone. Subdivision methods for solving polynomial equations. *Journal of Symbolic Computation*, 44(3):292–306, 2009.

[MS87] A. Morgan and A. J. Sommese. Computing all solutions to polynomial systems using homotopy continuation. *Applied Mathematics and Computation*, 24(2):115–138, 1987.

[MS95] M. Möller and H. J. Stetter. Multivariate polynomial equations with multiple zeros solved by matrix eigenproblems. *Numerische Mathematik*, 70(3):311–329, 1995.

[MS00] M. Möller and T. Sauer. H-bases for polynomial interpolation and system solving. *Advances in Computational Mathematics*, 12(4):335–362, 2000.

[MS03] D. Maclagan and G. Smith. Multigraded Castelnuovo-Mumford regularity. *Journal fur die Reine und Angewandte Mathematik*, 05 2003.

[MSW90] A. P. Morgan, A. J. Sommese, and C. W. Wampler. Computing singular solutions to nonlinear analytic systems. *Numerische Mathematik*, 58(1):669–684, 1990.

[MSW92a] A. P. Morgan, A. J. Sommese, and C. W. Wampler. Computing singular solutions to polynomial systems. *Advances in Applied Mathematics*, 13(3):305–327, 1992.

[MSW92b] A. P. Morgan, A. J. Sommese, and C. W. Wampler. A power series method for computing singular solutions to nonlinear analytic systems. *Numerische Mathematik*, 63(1):391–409, 1992.

[MT00] B. Mourrain and P. Trébuchet. Solving projective complete intersection faster. In *Proceedings of the 2000 international symposium on Symbolic and algebraic computation*, pages 234–241, 2000.

[MT01] M. Möller and R. Tenberg. Multivariate polynomial system solving using intersections of eigenspaces. *Journal of symbolic computation*, 32(5):513–531, 2001.

[MT05] B. Mourrain and P. Trébuchet. Generalized normal forms and polynomial system solving. In *Proceedings of the 2005 international symposium on Symbolic and algebraic computation*, pages 253–260, 2005.

[MT08] B. Mourrain and P. Trébuchet. Stable normal forms for polynomial system solving. *Theoretical Computer Science*, 409(2):229–240, 2008.

[MTVB19] B. Mourrain, S. Telen, and M. Van Barel. Truncated normal forms for solving polynomial systems: Generalized and efficient algorithms. *Journal of Symbolic Computation*, 2019.

[Mum96] D. Mumford. The red book of varieties and schemes. *Lecture notes in mathematics*, 1358:14–01, 1996.

[MVD15] N. Mastronardi and P. Van Dooren. Revisiting the stability of computing the roots of a quadratic polynomial. *Electronic Transactions on Numerical Analysis*, 44:73–82, 2015.

[Nis04] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, 26(6):756–770, 2004.

[NNT15] Y. Nakatsukasa, V. Noferini, and A. Townsend. Computing the common zeros of two bivariate functions via Bézout resultants. *Numerische Mathematik*, 129(1):181–209, 2015.

[Noo89] V. W. Noonburg. A neural network modeled by an adaptive Lotka-Volterra system. *SIAM J. Appl. Math.*, 49(6):1779–1792, 1989.

[NST18] Y. Nakatsukasa, O. Sète, and L. N. Trefethen. The AAA algorithm for rational approximation. *SIAM Journal on Scientific Computing*, 40(3):A1494–A1522, 2018.

[NT16] V. Noferini and A. Townsend. Numerical instability of resultant methods for multidimensional rootfinding. *SIAM Journal on Numerical Analysis*, 54(2):719–743, 2016.

[Oda89] T. Oda. Convex bodies and algebraic geometry: an introduction to toric varieties. *Bull. Amer. Math. Soc*, 21:360–364, 1989.

[OM78] T. Oda and K. Miyake. *Lectures on torus embeddings and applications*, volume 58. Tata Institute of Fundamental Research, 1978.

[PH16] B. Plestenjak and M. E. Hochstenbach. Roots of bivariate polynomial systems via determinantal representations. *SIAM Journal on Scientific Computing*, 38(2):A765–A788, 2016.

[Poi02] S.-D. Poisson. Mémoire sur l'élimination dans les équations algébriques. *Journal de l'ecole polytechnique*, 4(11):199, 1802.

[PS93] P. Pedersen and B. Sturmfels. Product formulas for resultants and Chow forms. *Mathematische Zeitschrift*, 214(1):377–396, 1993.

[PS96] P. Pedersen and B. Sturmfels. Mixed monomial bases. In *Algorithms in algebraic geometry and applications*, pages 307–316. Springer, 1996.

[PSS19] M. Panizzut, E. C. Sertöz, and B. Sturmfels. An octanomial model for cubic surfaces. *arXiv preprint arXiv:1908.06106*, 2019.

[PU99] F. Pauer and A. Unterkircher. Gröbner bases for ideals in Laurent polynomial rings and their application to systems of difference equations. *Applicable Algebra in Engineering, Communication and Computing*, 9(4):271–291, 1999.

[PV10] K. Piret and J. Verschelde. Sweeping algebraic curves for singular solutions. *Journal of computational and applied mathematics*, 234(4):1228–1237, 2010.

[Rei95] M. Reid. *Undergraduate commutative algebra*, volume 29. Cambridge University Press Cambridge, 1995.

[RLY18] J. I. Rodriguez, L.-H. Lim, and Y. You. Fiber product homotopy method for multiparameter eigenvalue problems. *arXiv preprint arXiv:1806.10578*, 2018.

[Roj99] J. M. Rojas. Toric intersection theory for affine root counting. *Journal of Pure and Applied algebra*, 136(1):67–100, 1999.

[Rot10] J. J. Rotman. *Advanced modern algebra*, volume 114. American Mathematical Soc., 2010.

[RW96] J. M. Rojas and X. Wang. Counting affine roots of polynomial systems via pointed newton polytopes. *Journal of Complexity*, 12(2):116–133, 1996.

[SC87] H. Schwetlick and J. Cleve. Higher order predictors and adaptive steplength control in path following algorithms. *SIAM journal on numerical analysis*, 24(6):1382–1393, 1987.

[SDLF+17] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

[Ser55] J.-P. Serre. Faisceaux algébriques cohérents. *Annals of Mathematics*, pages 197–278, 1955.

[SKKT04] K. Smith, L. Kahanpää, P. Kekääinen, and W. Traves. *An invitation to algebraic geometry.* Springer Science & Business Media, 2004.

[Sop05] I. Soprunov. Toric residue and combinatorial degree. *Transactions of the American Mathematical Society*, 357(5):1963–1975, 2005.

[Sot03] F. Sottile. Enumerative real algebraic geometry. In *Algorithmic and quantitative real algebraic geometry*, volume 60, pages 139–180. AMS, 2003.

[Sot11] F. Sottile. *Real solutions to equations from geometry*, volume 57. American Mathematical Soc., 2011.

[Sot17] F. Sottile. Ibadan lectures on toric varieties. *arXiv preprint arXiv:1708.01842*, 2017.

[SR94] I. R. Shafarevich and M. Reid. *Basic algebraic geometry*, volume 2. Springer, 1994.

[SR17] W. R. F. Santos and A. Rittatore. *Actions and invariants of algebraic groups.* CRC press, 2017.

[ŞS16] M. Şahin and I. Soprunov. Multigraded Hilbert functions and toric complete intersection codes. *Journal of Algebra*, 459:446–467, 2016.

[Sta97] H. Stahl. The convergence of Padé approximants to functions with branch points. *Journal of Approximation Theory*, 91(2):139–204, 1997.

[Ste91] G. W. Stewart. Perturbation theory for the singular value decomposition. *SVD and Signal Processing, II: Algorithms, Analysis and Applications*, pages 99–109, 1991.

[Ste97] H. J. Stetter. Stabilization of polynomial systems solving with Gröbner bases. In *Proceedings of the 1997 international symposium on Symbolic and algebraic computation*, pages 117–124, 1997.

[Ste04] H. J. Stetter. *Numerical polynomial algebra*, volume 85. SIAM, 2004.

[Ste06] M. Stewart. Perturbation of the SVD in the presence of small singular values. *Linear algebra and its applications*, 419(1):53–77, 2006.

[Stu94] B. Sturmfels. On the Newton polytope of the resultant. *Journal of Algebraic Combinatorics*, 3(2):207–236, 1994.

[Stu96] B. Sturmfels. *Gröbner bases and convex polytopes*, volume 8. American Mathematical Soc., 1996.

[Stu02] B. Sturmfels. *Solving systems of polynomial equations.* Number 97. American Mathematical Soc., 2002.

[Sue85] S. P. Suetin. On an inverse problem for the m-th row of the Padé table. *Mathematics of the USSR-Sbornik*, 52(1):231, 1985.

[Sue02] S. P. Suetin. Padé approximants and efficient analytic continuation of a power series. *Russian Mathematical Surveys*, pages 43–141, 2002.

[Sul18] S. Sullivant. *Algebraic Statistics*, volume 194. American Mathematical Soc., 2018.

[SVBDL14] L. Sorber, M. Van Barel, and L. De Lathauwer. Numerical solution of bivariate and polyanalytic polynomial systems. *SIAM Journal on Numerical Analysis*, 52(4):1551–1572, 2014.

[SVTW06] J. Sidman, A. Van Tuyl, and H. Wang. Multigraded regularity: coarsenings and resolutions. *Journal of Algebra*, 301(2):703–727, 2006.

[SVW01] A. J. Sommese, J. Verschelde, and C. W. Wampler. Numerical decomposition of the solution sets of polynomial systems into irreducible components. *SIAM Journal on Numerical Analysis*, 38(6):2022–2046, 2001.

[SVW05] A. J. Sommese, J. Verschelde, and C. W. Wampler. Introduction to numerical algebraic geometry. In *Solving polynomial equations*, pages 301–337. Springer, 2005.

[Syl40] J. J. Sylvester. LVII. Note on elimination. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 17(111):379–380, 1840.

[Sze39] G. Szegő. *Orthogonal polynomials*, volume 23. American Mathematical Soc., 1939.

[TBI97] L. N. Trefethen and D. Bau III. *Numerical linear algebra*, volume 50. SIAM, 1997.

[Tel16] S. Telen. Solving systems of polynomial equations. 2016. Master's thesis, available at `https://www.scriptiebank.be/scriptie/2016/het-oplossen-van-stelsels-veeltermvergelijkingen`.

[Tel20] S. Telen. Numerical root finding via Cox rings. *Journal of Pure and Applied Algebra*, 224(9), 2020.

[Tim20] S. Timme. Mixed precision path tracking for polynomial homotopy continuation. *arXiv preprint arXiv:1902.02968*, 2020.

[TMVB18] S. Telen, B. Mourrain, and M. Van Barel. Solving polynomial systems via truncated normal forms. *SIAM Journal on Matrix Analysis and Applications*, 39(3):1421–1447, 2018.

[Tre02] L. N. Trefethen. The SIAM 100-dollar, 100-digit challenge. *SIAM News*, 35(6):2, 2002.

[Tre17] L. N. Trefethen. Multivariate polynomial approximation in the hypercube. *Proceedings of the American Mathematical Society*, 145(11):4837–4844, 2017.

[Tre19] L. N. Trefethen. *Approximation theory and approximation practice*, volume 164. SIAM, 2019.

[Tre20] L. N. Trefethen. Quantifying the ill-conditioning of analytic continuation. *BIT Numerical Mathematics*, pages 1–15, 2020.

[TT13] A. Townsend and L. N. Trefethen. An extension of chebfun to two dimensions. *SIAM Journal on Scientific Computing*, 35(6):C495–C518, 2013.

[TTVB20] S. Telen, S. Timme, and M. Van Barel. Backward error measures for roots of polynomials. *Numerical Algorithms*, pages 1–21, 2020.

[TVB18] S. Telen and M. Van Barel. A stabilized normal form algorithm for generic systems of polynomial equations. *Journal of Computational and Applied Mathematics*, 342:119–132, 2018.

[TVB20] F. Tisseur and M. Van Barel. Min-max elementwise backward error for roots of polynomials and a corresponding backward stable root finder. *arXiv:2001.05281*, 2020.

[TVBV19] S. Telen, M. Van Barel, and J. Verschelde. A robust numerical path tracking algorithm for polynomial homotopy continuation. *arXiv preprint arXiv:1909.04984*, 2019.

[Vak17] R. Vakil. The rising sea: foundations of algebraic geometry. *preprint*, 2017.

[VC94] J. Verschelde and R. Cools. Symmetric homotopy construction. *J. Comput. Appl. Math.*, 50:575–592, 1994.

[vdH15] J. van der Hoeven. Reliable homotopy continuation. Technical report, LIX, Ecole polytechnique, 2015.

[VDS+16] N. Vervliet, O. Debals, L. Sorber, M. Van Barel, and L. De Lathauwer. Tensorlab 3.0, Mar. 2016. Available online at `https://www.tensorlab.net`.

[Ver99] J. Verschelde. Algorithm 795: PHCpack: A general-purpose solver for polynomial systems by homotopy continuation. *ACM Transactions on Mathematical Software (TOMS)*, 25(2):251–276, 1999.

[VLLP79] V. V. Vavilov, G. López Lagomasino, and V. A. Prokhorov. On an inverse problem for the rows of a Padé table. *Matematicheskii Sbornik*, 152(1):117–127, 1979.

[VSDL17a] J. Vanderstukken, A. Stegeman, and L. De Lathauwer. Systems of polynomial equations, higher-order tensor decompositions and multidimensional harmonic retrieval: a unifying framework–Part I: The canonical polyadic decomposition. *Available at* ***ftp://ftp.esat.kuleuven.be/pub/stadius/nvervliet/vanderstukken2017systems1.pdf***, 2017.

[VSDL17b] J. Vanderstukken, A. Stegeman, and L. De Lathauwer. Systems of polynomial equations, higher-order tensor decompositions and multidimensional harmonic retrieval: a unifying framework–Part II: The block term decomposition. *Available at* ***ftp://ftp.esat.kuleuven.be/pub/stadius/nvervliet/vanderstukken2017systems2.pdf***, 2017.

[VVC94] J. Verschelde, P. Verlinden, and R. Cools. Homotopies exploiting Newton polytopes for solving sparse polynomial systems. *SIAM Journal on Numerical Analysis*, 31(3):915–930, 1994.

[Wam93] C. W. Wampler. An efficient start system for multi-homogeneous polynomial continuation. *Numerische Mathematik*, 66(1):517–523, 1993.

[War91] E. Waring. *Meditationes Algebraicae: An English Translation of the Work of Edward Waring; Ed. and Transl. from the Latin by Dennis Weeks*. American Mathematical Society, 1991.

[WS05] C. W. Wampler and A. J. Sommese. *The Numerical solution of systems of polynomials arising in engineering and science*. World Scientific, 2005.

[WS11] C. W. Wampler and A. J. Sommese. Numerical algebraic geometry and algebraic kinematics. *Acta Numerica*, 20:469–567, 2011.

[XBY18] J. Xu, M. Burr, and C. Yap. An approach for certifying homotopy continuation paths: Univariate case. In *Proceedings of the 2018 ACM International Symposium on Symbolic and Algebraic Computation*, pages 399–406, 2018.

[Zen16] Z. Zeng. Sensitivity and computation of a defective eigenvalue. *SIAM Journal on Matrix Analysis and Applications*, 37(2):798–817, 2016.

[ZL14] Z. Zeng and T.-Y. Li. NAClab: A matlab toolbox for numerical algebraic computation. *ACM Communications in Computer Algebra*, 47(3/4):170–173, 2014.

Curriculum

Simon Telen was born on October 27, 1993 in Maaseik, Belgium.

Education

Doctoral researcher in applied mathematics September 2016 - Present
KU Leuven, Department of Computer Science
Supervisor: Marc Van Barel
Supervisory committee: Marc Van Barel, Nick Vannieuwenhoven, Wim Veys

M.Sc. summa cum laude 2014 - 2016
KU Leuven
Mathematical Engineering
Master's thesis title: *Solving Systems of Polynomial Equations*

B.Sc. magna cum laude 2011 - 2014
KU Leuven
Major in Electrical Engineering
Minor in Mathematical Modelling of Living Systems

Awards

Best poster award at the MEGA 2019 Conference June 2019
Universidad Complutense de Madrid, Spain
for our poster 'Robust Numerical Path Tracking for Polynomial Homotopies' with Marc Van Barel and Jan Verschelde

Best poster presentation award at the ISSAC 2018 conference July 2018
City University of New York, USA
for our poster 'Truncated Normal Forms for Solving Polynomial Systems' with Bernard Mourrain and Marc Van Barel

List of publications

Articles in internationally reviewed academic journals

Simon Telen, Sascha Timme and Marc Van Barel. *Backward error measures for roots of polynomials.* Numerical Algorithms (2020):1-21.

Simon Telen. *Numerical root finding via Cox rings.* Journal of Pure and Applied Algebra, 224(9), 2020.

Bernard Mourrain, Simon Telen and Marc Van Barel. *Truncated normal forms for solving polynomial systems: generalized and efficient algorithms.* Journal of Symbolic computation, `https://doi.org/10.1016/j.jsc.2019.10.009`, 2019.

Simon Telen, Bernard Mourrain and Marc van Barel. *Solving polynomial systems via truncated normal forms.* SIAM Journal on Matrix Analysis and Applications, 39(3):1421-1447, 2018.

Simon Telen and Marc Van Barel. *A stabilized normal form algorithm for generic systems of polynomial equations.* Journal of Computational and Applied Mathematics, 342:199-132, 2018.

Articles in review

Simon Telen, Marc Van Barel and Jan Verschelde. *A robust numerical path tracking algorithm for polynomial homotopy continuation.* arXiv:1909.04984, 2019.

Matías R. Bender, Simon Telen. *Toric eigenvalue methods for solving sparse polynomial systems.* arXiv:2006.10654, 2020.

Seminar talks and invited talks

Numerical Root Finding via Cox Rings — January 2020
Forschungsseminar Diskrete Mathematik/Geometrie, FU Berlin, Germany

Truncated Normal Forms — December 2019
Algorithmic Algebra Seminar, TU Berlin, Germany

Numerical Root Finding via Cox Rings — November 2019
Seminar Algebraische Geometrie, FU Berlin, Germany

Robust Numerical Path Tracking in Polynomial Homotopies — April 2019
NUMA seminar, KU Leuven, Belgium

Stabilized Algebraic Methods for Multivariate Polynomial Root Finding — March 2019
Research visit with Tomas Pajdla, CIIRC, Prague

Stabilized Algebraic Methods for Multivariate Polynomial Root Finding December 2018
Research visit with Bernd Sturmfels, MPI Leipzig

Polynomial System Solving through Stabilized Representation of Quotient Algebras April 2018
Research visit with Tyler Jarvis, BYU, Provo

Polynomial System Solving and Numerical Linear Algebra September 2017
Research visit with Bernard Mourrain, INRIA, Sophia-Antipolis

Systems of Polynomial Equations and Numerical Linear Algebra May 2017
NUMA seminar, KU Leuven, Belgium

Talks and posters at international conferences

Solving Polynomial Systems using Cox Rings February 2020
Milestone conference of the thematic Einstein semester 'Algebraic Geometry', Berlin

Numerical Root Finding via Cox Rings (poster) October 2019
Opening conference of the thematic Einstein semester 'Algebraic Geometry', Berlin

Robust Numerical Path Tracking in Polynomial Homotopies July 2019
ICIAM conference, Valencia

Solving Nonlinear Eigenvalue Problems using Contour Integration July 2019
ICIAM conference, Valencia

Numerical Root Finding via Cox Rings July 2019
SIAM AG conference, Bern

Robust Numerical Path Tracking for Polynomial Homotopies (poster) June 2019
MEGA conference, Madrid

Numerical Root Finding via Cox Rings (poster) June 2019
Conference 'Ideals, Varieties and Applications' (celebrating the influence of David Cox)

Truncated Normal Forms for Solving Polynomial Systems (poster) September 2018
ICERM nonlinear algebra semester, workshop 'Core Computational Methods', Providence

Truncated Normal Forms for Solving Polynomial Systems (poster) July 2018
ISSAC conference, New York

Truncated Normal Forms for Solving Polynomial Systems (poster) June 2018
CBMS conference on Applications of Polynomial Systems, Fort Worth

Structured Matrices in Polynomial System Solving May 2018
SIAM ALA conference, Hong Kong

Solving Nonlinear Eigenvalue Problems using Contour Integration May 2018
SIAM ALA conference, Hong Kong

Polynomial System Solving and Numerical Linear Algebra August 2017
SIAM AG conference, Atlanta

Matrices in Polynomial System Solving May 2017
Rencontre en Algèbre Linéaire Numérique Amiens-Calais, Amiens

Solving Systems of Polynomial Equations July 2016
ILAS conference, Leuven

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF COMPUTER SCIENCE
NUMA
Celestijnenlaan 200A box 2402
B-3001 Leuven
simon.telen@kuleuven.be
https://simontelen.webnode.com/