Sparse polynomial systems in optimization

Der Fakultät für Mathematik und Informatik
der Universität Leipzig
eingereichte

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr.rer.nat.)

im Fachgebiet

Mathematik

vorgelegt

von Kemal Rose
geboren am 08.11.1994 in Einbeck

Leipzig, den November 6, 2023

Abstract

Systems of polynomial equations appear both in mathematics, as well as in many applications in the sciences, economics and engineering. Solving these systems is at the heart of computational algebraic geometry, a field which is often associated with symbolic computations based on Gröbner bases. Over the last thirty years, increasing performance and versatility made numerical algebraic geometry emerge as an alternative. It enables us to solve problems which are infeasible with symbolic methods. The focus of this thesis is the rich interplay between algebraic geometry, numerical computation and optimization in various instances.

As a first application of algebraic geometry, we investigate global optimization problems whose objective function and constraints are all described by multivariate polynomials. One of the most important, and also most common, features of real world data is sparsity. We explore the effects of sparsity in global optimization, when exhibited by constraints and objective functions. Exploiting this property can lead to dramatic improvements of computational performance of algorithms.

As a second application of geometry we study a particularly structured class of polynomial programs which stems from the optimization of sequencial decision rules. In the framework of partially observable Markov decision rules, an agent manipulates a system in a sequence of events. It selects an action at every time step, which in turn influences the state of the system at the next time step, and depending on the state it receives an instantaneous reward. Optimizing the long term reward has a long-standing history in computer science, economics and statistics. The ability to incorporate nondeterministic effects makes the framework particularly well suited for real world applications. We initiate a novel, geometric perspective on the underlying optimization problem and explore algorithmic consequences.

As a third application of geometry we present the usage of tropical geometry in order to numerically compute defining equations of unirational varieties from their parametrization. Tropical geometry is an emerging field in mathematics at the boundary of discrete geometry and algebraic geometry. The tropicalization of a variety is a polyhedral complex which encodes geometric information of the variety. Tropical implicitization means computing the tropicalization of a unirational variety from its parametrization. In the case of a hypersurface, this amounts to finding the Newton polytope of the implicit equation, without computing its coefficients. We use this as a preprocessing step for numerical computation.

Contrary to the above uses of geometry in application, we also employ numerical computation in pure mathematics. When relying on numerical methods, problems can be solved that are infeasible with symbolic methods, but the computational results lack a certificate for correctness. This often hinders the application of numerical computation with the purpose of proving mathematical theorems. With this in mind, we develop interval arithmetic as a practical tool for certification in numerical algebraic geometry.

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den November 6, 2023

. . . . . .
(Kemal Rose)

Authorship

Parts of this thesis resolve around results which arose from collaboration with other researchers.

Chapter 2 is based on the articles [LMR23], and [RMR23], which are joint work with Leonid Monin and Julia Lindberg. Section 2.3, Section 2.5.2 and Section 2.6 are my contribution, also the proof of Proposition 2.7.2 and Lemma 2.7.3 and the beginning of subsection 2.7.1.

Chapter 3 is based on [BRT23], which was written in collaboration with Paul Breiding and Sascha Timme. The article is published in the journal ACM Transactions on Mathematical Software. All results arose through equal contribution of all three authors.

Section 4.8, and in particular Theorem 4.8.1 are my contributions. All other sections in Chapter 4 are based on the article [DGLM$^+$24], which arose in collaboration with Marina Garotte, Guido Montúfar, Johannnes Müller and Mareike Dressler and which is published in the Journal of Symbolic Computation. While most parts arose through equal contribution of all authors, I contributed the proof of Theorem 4.5.4.

Chapter 5 is based on the article [RST23], which will be a chapter in a book on the software package `Oscar`. My coauthors are Bernd Sturmfels and Simon Telen. The article arose through equal contribution of all authors, the algorithmic implementations are my contribution.

Aknowledgements

I want to express my sincere gratitude towards all the people who supported and accompanied me during my journey over the last three years at MPI. This is directed first and foremost at both my advisors Bernd and Simon for their guidance, support and patience. Your expertise and mentorship have been invaluable in shaping the direction of my research.

Next, I would like to thank all of my colleagues with whom I shared this exciting time of mathematical and non mathematical growth, and also many Rewe salads. You made the MPI a stimulating academic environment and there are many memories for which I am grateful. I would like to take the opportunity to thank my collaborators for sharing many moments of excitement, mathematical understanding, frustration and success. Your enthusiasm for research has enriched my life. Also thanks to the staff at MPI, whose labour and dedication makes all of this possible, and without whom the MPI would not exist.

To my family for their unconditional support, their confidence and warmth, and to Eva, who does not know how much purpose and joy she has been giving me. Finally, I would like to thank my flatmates for making Leipzig a place where I feel at home.

Thank you all for being an integral part of this chapter of my life.

Contents

Chapter 1

Introduction

Polynomial programs are optimization problems in which polynomial objective functions are minimized subject to polynomial constraints:

$$\min_{x\in\mathbb{R}^n} \; f_0(x) \quad \text{subject to} \quad f_1(x) = \cdots = f_m(x) = 0. \tag{POP}$$

The problem (POP) is also called a polynomial optimization problem. Such problems have broad modelling power, and have found applications in various science and engineering problems, such as signal processing, material sciences, combinatorial optimization, power systems engineering and more [TR01, PRW95, MH19]. Much work in the past 20 years was targeted at studying convex relaxations of (POP), based on the Lasserre Hierarchy and semidefinite optimization. A related line of work has been to study the critical points of (POP). Inspired by recent increases in performance of numerical polynomial system solving, in this work we focus on solving critical point equations. The core of this thesis is the interplay between algebraic geometry and numerical computation, and the study of the geometry of polynomial programs.

The field of computational algebraic geometry has traditionally been associated with symbolic computations based on Gröbner bases. Over the last thirty years the novel computational paradigm of numerical algebraic geometry emerged as an alternative, incorporating techniques from numerical analysis. Based on algorithmic frameworks such as homotopy continuation it solves problems that are infeasible with symbolic methods. While symbolic algorithms reveal the algebraic properties of polynomial systems, numerical methods are predominantly geometric in nature, and compute numerical approximations of solutions.

We now describe in broad strokes a framework that combines intersection theory and numerical computation to solve various polynomial programs with a certificate for correctness. The emerging questions will tie together, and motivate, many of the results in the following chapters. They further touch on fundamental relations between algebraic geometry and numerical computation which are of independent interest. The idea is to use Lagrange multipliers and numerical algebraic geometry to find all critical points of, and therefore globally solve, the problem (POP). We consider the *Lagrange system* $\mathbf{L}_F = (f_1, \ldots, f_m, \ell_1, \ldots, \ell_n)$ of (POP), where

$$\ell_j = \frac{\partial}{\partial x_j}\left(f_0 - \sum_{i=1}^{m} \lambda_i f_i\right).$$

In order to find an optimizer of (POP) we need to *certifiably* find *all* solutions of $\mathbf{L}_F$, since missing any would render our results useless for the purpose of global optimization. Hence, we need to

address fundamental problems of numerical computation. When employing numerical algorithms to generate solutions, we have no guarantees a priori that we did not miss any solutions, or even that our numerical approximations are correct and distinct. It is desirable to know how many solutions there are to find:

Question 1. How many smooth critical points does (POP) have?

Intersection theory furnishes tools for assessing the algebraic complexity of (POP). In some instances, we can either compute the exact number, or tight upper bounds for the number of solutions to $\mathbf{L}_F$. Provided we know the answer to Question 1, we are left with the following question, which also makes sense for arbitrary square polynomial systems $\mathbf{L}_F$.

Question 2. How to numerically certify that all numerically approximated solutions of $\mathbf{L}_F$ represent true and distinct solutions?

Question 2 is contrary to the belief by some algebraists that numerical algorithms can only provide the floating point approximation of a solution, but they cannot in general certify that such a solution is unique, or provide guarantees. We emphasize that, using certification procedures, numerical computation can be used to *prove* lower bounds to the number of solutions by addressing Question 2. Further, should the number of certified solutions agree with the total number of critical points, then we have a proof that all critical points were found. In that sense, addressing both questions above lets us solve (POP).

The algorithmic framework of Homotopy Continuation can be used for finding all solutions of $\mathbf{L}_F$ by tracking solutions from an 'easy' system of polynomial equations $G(x)$ (called the *start system*) to $\mathbf{L}_F$. This is done by constructing a *homotopy*,

$$H(t;x) : [0,1] \times \mathbb{C}^n \longrightarrow \mathbb{C}^n,$$

with $H(0;x) = G(x)$ and $H(1;x) = \mathbf{L}_F(x)$. We call a homotopy H *sufficient* if, by solving the ODE initial value problems $\frac{\partial H}{\partial t} + \frac{\partial H}{\partial x}\dot{x} = 0$ with initial values $\{x \,:\, G(x) = 0\}$, all isolated solutions of $F(x) = 0$ can be obtained. A practical question when employing this method is:

Question 3. How to efficiently compute a sufficient homotopy for $\mathbf{L}_F$, such that the start system G has a minimal number of solutions?

We now describe the contents of this thesis in more detail. At the beginning of each chapter we summarize our main results and give an overview of the state of the art. Each chapter contains concrete examples. Many of the results in this thesis coming from algebraic geometry are accompanied by software that provides concrete algorithmic solutions. Most code is made accessible at MathRepo: https://mathrepo.mis.mpg.de/#.

Chapter 2 In Chapter 2 we study a broad class of polynomial optimization problems whose constraints and objective functions exhibit sparsity patterns. Based on toric intersection theory we give two formulas for the number of critical points of generic members, one as a mixed volume and one as an intersection product based on Porteus' formula. This addresses Question 1. As a corollary, under the same sparsity assumptions, we obtain a convex geometric interpretation of polar degrees, a classical invariant of algebraic varieties, as well as Euclidean distance degrees. Furthermore, we prove BKK generality of Lagrange systems $\mathbf{L}_F$ in many instances.

Finally, we demonstrate how Question 3 can be addressed, based on our previous results from this chapter. In the case where we have a linear objective and a single polynomial constraint

we explicitly construct a polyhedral homotopy algorithm for solving the Lagrange system $\mathbf{L}_F$. A bottleneck in computing a start system for the Lagrange system $\mathbf{L}_F$ is solving the *tropicalization* of $\mathbf{L}_F$. Our algorithm relies on an explicit description of the tropical solution set of $\mathbf{L}_F$. The superiority over traditional homotopy continuation algorithms is demonstrated experimentally.

Chapter 3 Hauenstein and Sottile remark in [HS12] that while numerical methods "routinely and reliably solve systems of polynomial equations with dozens of variables having thousands of solutions", they have the shortcoming that "the output is not certified" and that "this restricts their use in some applications, including those in pure mathematics". In Chapter 3 we combine interval arithmetic and Krawczyk's method with numerical algebraic geometry to rigorously certify solutions to square systems of polynomial equations. We present an extremely fast and easy-to-use implementation of a certification method in `HomotopyContinuation.jl`. This implementation makes the certification of solutions often a matter of seconds and not hours or days. The function `certify` takes as input a *square polynomial system* F, an approximation of a complex zero $x \in \mathbb{C}^n$ and returns a small box around x, provably containing a unique solution of $F = 0$. Our method can be used to *prove* hard lower bounds on the number of (real/positive) zeros of a polynomial system. This method effectively addresses Question 2. Within the last two years it has paved the way for applications of numerical methods in various instances [BRST23, BFS21, KPR$^+$21, BPS21, BHIM22, Ear21, Mar21, Wei21, LAR21, Stu21, BT21, ABF$^+$23, BKK20, SY21, ST21].

Chapter 4 So far, in Chapter 2 the only structure that we leveraged in the study of polynomial programs is sparsity. In Chapter 4 we construct a particular family of highly structured polynomial programs that are motivated by application. Partially observable Markov decision processes (POMDPs) offer a model for sequential decision-making under state uncertainty and model various real-world sequencial decision processes that are based on partial information. Such processes include the optimal control of robots, machine maintenance, search problems, and inventory problems [Whi88, Bel66]. Difficulties that appear in from real world applications, such as noisy sensing and imperfect control, are naturally incorporated in the framework of POMDPs. The optimization of the expected long-term reward is known to be NP-hard in general [VLB12].

In this chapter we open up a new, exciting geometric perspective on the optimization of POMDPs in the case of deterministic observations. By solving an implicitization problem for the reward function, we recast this problem as a polynomial program with a linear objective function. The feasible set of this problem is the positive part of the *state-aggregation variety*, a linear section of a join of Segre varieties. We conduct experiments in which we solve the KKT equations or the Lagrange equations over different boundary components of the feasible set. In Section 4.8 we give a satisfactory answer to Question 1 in this setting, by computing the polar degrees of state aggregation varieties. This is a considerable improvement over the bounds from Chapter 2, which apply more generally. Our results open up many interesting questions. This includes, for example, showing objective value exactness of the first order Lasserre relaxation of the quadratically constrained optimization problem.

Chapter 5 In the previous chapters we investigated families of polynomial equations with a focus on understanding the behaviour of general members. This chapter investigates the discriminant locus that comprises the members that are not general. More precisely, we investigate implicitization problems from a tropical view point and provide a software package based on `Oscar.jl` that predicts Newton polytopes of implicit equations. It solves challenging instances, and can be used

for classical implicitization as well. In particular, it computes A-discriminants. We also develop implicitization in higher codimension via Chow forms, and we pose several open questions.

Chapter 2

The algebraic degree of sparse polynomial optimization

One of the most important and also most common features of real world data is sparsity. Exploiting this property can lead to dramatic improvements of computational performance, affecting both execution time and memory usage of algorithms. Due to its ubiquity, it is very desirable to leverage sparsity in general purpose methods. In this chapter we study a broad class of polynomial optimization problems whose constraints and objective functions exhibit sparsity patterns. We give two characterizations of the number of critical points of these problems, one as a mixed volume and one as an intersection product on a toric variety. As a corollary, we obtain a convex geometric interpretation of polar degrees, a classical invariant of algebraic varieties, as well as Euclidean distance degrees. Furthermore, we prove BKK generality of Lagrange systems in many instances. Finally, we demonstrate how our theoretical results can be made effective by developing a polyhedral homotopy algorithm for solving Lagrange systems in restricted cases.

2.1 Introduction

We again consider the polynomial optimization (POP) from the introduction of this Thesis:

$$\min_{x\in\mathbb{R}^n} \; f_0(x) \quad \text{subject to} \quad f_1(x) = 0, \ldots, f_m(x) = 0$$

When the first order optimality conditions hold, there are finitely many complex critical points to (POP). For a specified objective function f_0 and for fixed constraints $f_1, \ldots, f_m$ we abbreviate $\mathbf{F} = (f_0, \ldots, f_m)$. The number of complex critical points of $\mathbf{F}$ is called the *algebraic degree* of $\mathbf{F}$. While (POP) is a real optimization problem, one considers the number of complex critical points since for polynomials $\mathbf{F}$ with fixed monomial support, the algebraic degree is generically[1] constant.

When the first order optimality conditions of (POP) hold, then the coordinates of the optimal solution of (POP) are algebraic functions of the coefficients of $\mathbf{F}$. The algebraic degree of $\mathbf{F}$ has an additional interpretation as the degree of these algebraic functions. Observe also that the algebraic degree gives an upper bound on the number of real critical points of (POP). This gives a bound on the number of local optima, where local optimization methods can get caught.

A formula for the algebraic degree when $\mathbf{F}$ consists of generic polynomials with full monomial support was given in [NR09]. This was then specialized for many classes of convex polynomial

[1] By generic, we mean generic with respect to the Zariski topology. See Remark 2.2.1 for a detailed explanation.

optimization problems in [GvBR09] and [NRS10]. When the objective function is the Euclidean distance function, i.e. $f_0 = \|x-u\|_2^2$ for a point $u \in \mathbb{R}^n$, the number of critical points to (POP) for general u is called the *ED degree* of $(f_1, \ldots, f_m)$. The study of ED degrees began with [DHO+14] and initial bounds on the ED degree of a variety were given in [DHO+16]. Other work has found the ED degree for real algebraic groups [BD15], Fermat hypersurfaces [Lee17], orthogonally invariant matrices [DLOT17], smooth complex projective varieties [AH18], the multiview variety [MRW20a], when $m = 1$ [BSW21] and when the data u and polynomials $f_1, \ldots, f_m$ are not general [MRW20b].

A related problem is maximum likelihood estimation which considers the objective function $f_0 = x_1^{u_1} \cdots x_n^{u_n}$. The number of complex critical points is called the *ML degree* of $(f_1, \ldots, f_m)$. Relationships between ML degrees and Euler characteristics as well as the ML degree of various statistical models have been studied in [CHKS06, HKS05, Huh13, ABB+19, DM21, MMW21].

Inspired by recent results on the ED and ML degrees of sparse polynomial systems [BSW21, LNRW23], we study the algebraic degree of (POP) when each $f_i \in \mathbb{R}[x_1, \ldots, x_n]$ is assumed to be a *sparse polynomial* (see Section 2.2.1 below), with generic coefficients. Given an optimization problem of the form (POP) where $\mathbf{F}$ is a general list of sparse polynomial equations, define the *Lagrangian* of $\mathbf{F}$ to be

$$\Phi_F(\lambda, x) := f_0 - \sum_{i=1}^{m} \lambda_i f_i.$$

We consider the *Lagrange system* of $\mathbf{F}$, namely $\mathbf{L}_\mathrm{F} = (f_1, \ldots, f_m, \ell_1, \ldots, \ell_n)$, where

$$\ell_j = \frac{\partial}{\partial x_j}\left(f_0 - \sum_{i=1}^{m} \lambda_i f_i\right).$$

Analogous to the *algebraic degree of polynomial optimization* from [NR09], we generalize the common term to the *algebraic degree of sparse polynomial optimization*. It is the number of critical points:

$$\#\mathcal{V}(\mathbf{L}_\mathrm{F}) = \#\left\{(x, \lambda) \in \mathbb{C}^n \times \mathbb{C}^m : \quad 0 = f_1 = \cdots = f_m = \ell_1 = \cdots = \ell_n\right\} \tag{2.1}$$

of f_0 restricted to $\mathcal{V}(f_1, \ldots, f_m)$ where each $f_i \in \mathbb{R}[x_1, \ldots, x_n]$ is a *sparse* polynomial.

There exist classical results in algebraic geometry bounding the number of isolated solutions to a square polynomial system. A result of Bézout says that $\#\mathcal{V}(\mathbf{L}_\mathrm{F})$ is bounded above by the product of the degrees of the polynomials in $\mathbf{L}_\mathrm{F}$. If $\deg(f_i) = d_i$ and $\deg(\ell_j) = h_j$, $0 \le i \le m, 1 \le j \le n$, Bézout 's bound reduces to $d_1 \cdots d_m \cdot h_1 \cdots h_n$ where $h_j \le \max_{i \in [m]}\{d_0 - 1, d_i\}$. The work of Nie and Ranestad refined this bound considerably and showed that

$$\#\mathcal{V}(\mathbf{L}_\mathrm{F}) \le d_1 \cdots d_m \cdot D_{n-m}(d_0 - 1, \ldots, d_m - 1)$$

where $D_r(n_1, \ldots, n_k) = \sum_{i_1 + \cdots + i_k = r} n_1^{i_1} \cdots n_k^{i_k}$ is the symmetric sum of products [NR09]. While this bound is generically tight for dense polynomial systems, the following example shows that it can be quite bad (even worse than Bézout's bound) for sparse polynomial systems.

Example 2.1.1. Consider the following optimization problem:

$$\min_{x \in \mathbb{R}^n} \; c^T x \quad \text{subject to} \quad f = \alpha_1 x_1^3 + \sum_{j=2}^{n-1} \alpha_j x_j^2 + \alpha_n x_n = 1. \tag{2.2}$$

where $c, \alpha \in \mathbb{R}^n$ are generic parameters. The corresponding Lagrange system is given by $\mathbf{L}_\mathrm{F} = (\ell_1, \ldots, \ell_n, f)$ where

$$\ell_1 = c_1 - 3\lambda\alpha_1 x_1^2, \quad , \ell_n = c_n - \alpha_n \lambda, \quad \ell_j = c_j - 2\lambda\alpha_j x_j, \; 2 \le j \le n-1.$$

The Bézout bound tells us that $\#\mathcal{V}(\mathbf{L}_{\mathrm{F}})) \leq 3 \cdot 2^{n-2}$ which is better than the Nie-Ranestad bound which gives $\#\mathcal{V}(\mathbf{L}_{\mathrm{F}})) \leq 3 \cdot D_{n-1}(0,2) = 3 \cdot 2^{n-1}$. In this case, the sparsity of the Lagrange equations allows one to solve the system by hand one variable at a time. One can see that for generic values of c and α, the number of critical points equals $\#\mathcal{V}(\mathbf{L}_{\mathrm{F}}) = 2$.

Motivated by the previous example we proceed to prove stronger bounds for the algebraic degree of sparse polynomial optimization programs in this chapter. We focus on a version of Question 1 from the Introduction of the Thesis: How many critical points does (POP) have for sparse $\mathbf{F}$? The motivation for Question 1 is that if we know how many critical points (POP) has, and we find them all, then we can globally solve (POP). Currently, the only way to *provably* find all smooth critical points is to find all complex solutions of $\mathbf{L}_{\mathrm{F}}$.

The field of computational algebraic geometry has traditionally been associated with symbolic computations based on Gröbner bases. Recent developments in numerical frameworks, such as homotopy continuation [BHSW], provide algorithms that are able to solve problems intractable with symbolic methods. Moreover, numerical algorithms can not only provide the floating point approximation of a solution, but also certify that a given approximation represents a unique solution, and provide guarantees [BRT23, Rum99, Lee19], as we will see in Chapter 3. Therefore, numerical computation can be used to prove lower bounds to the number of solutions. However, to guarantee that there are no other solutions, one needs an upper bound to $\#\mathcal{V}(\mathbf{L}_{\mathrm{F}})$, which can be obtained using intersection theory.

Such an intersection theoretic bound for sparse polynomial systems was given by the celebrated Bernstein-Kouchnirenko-Khovanskii (BKK) theorem. The BKK theorem relates the number of $\mathbb{C}^*$ zeros to a system of polynomial equations to the mixed volume of the corresponding Newton polytopes (see Section 2.2.1). While their bound is generically tight, we note that the coefficients of the system $\mathbf{L}_{\mathrm{F}}$ are linearly dependent, so a priori it is not clear that the system $\mathbf{L}_{\mathrm{F}}$ has the expected number of solutions. This inspires the following question:

Question 4. Does the number of solutions of the Lagrange system of (POP) agree with the BKK bound?

An affirmative answer to Question 4 would show that *polyhedral homotopy* algorithms are optimal for finding all complex critical points to (POP) in the sense that for every solution to $\mathbf{L}_{\mathrm{F}} = 0$, exactly one *homotopy path* is tracked. For more details on polyhedral homotopy continuation see [HS95a]. Furthermore, understanding BKK exactess of non generic polynomial systems is of increasing interest in the applied algebraic geometry community.

Contribution

In this chapter we contribute several results based on intersection theory that determine the number of critical points of generic, sparse polynomial programs. First, we show in Theorem 2.3.7 that the answer to Question 4 is positive for a wide class of sparse polynomial programs having strongly admissible monomial support (see Definition 2.3.4). In particular, our results show that the bound is tight for Example 2.1.1.

As a corollary, we generalize the result in [BSW21] in this case and show that the ED degree of a variety with strongly admissible support is equal to the BKK bound of its corresponding Lagrange system (Corollary 2.4.1). We also prove analogous results for (the sum of) polar degrees (Corollary 2.4.5), giving the first convex geometric interpretation of the algebraic invariant. Further, in Corollary 2.3.8 we show that algebraic degrees of generic sparse polynomial programs are

determined by the Newton polytopes of $\mathbf{F}$. Corollary 2.3.8 has also algorithmic implications which were studied in [RMR23].

For a larger family of sparse polynomial programs, in Theorem 2.3.11 we provide a different formula for the corresponding algebraic degrees. Our main tool here is Porteous' formula, which computes the fundamental class of the degeneracy locus of a morphism between two vector bundles as a polynomial in their Chern classes. Using Porteous' formula, Theorem 2.3.11 expresses the algebraic degree in terms of the intersection theory of a certain toric compactification of $(\mathbb{C}^*)^n$. The formula for the algebraic degree in Theorem 2.3.11 can be expressed as a (non-necessarily positive) linear combination of mixed volumes. However, the explicit connection to the mixed volume of the Lagrange system is still mysterious.

Finally, in Section 2.7 we demonstrate how Question 3 from the Introduction of the thesis can be solved effectively, based on our previous results. In the case where we have a linear objective function and a single constraint we explicitly construct a specialized polyhedral homotopy algorithm for solving the Lagrange system. This is the content of Theorem 2.7.4. We further present numerical results which show that our algorithm outperforms standard polyhedral homotopy solvers.

2.2 Preliminaries and notation

2.2.1 Sparse polynomials and polyhedral geometry

A *sparse polynomial* $f \in \mathbb{C}[x_1, \ldots, x_n]$ is defined by its monomial support and its coefficients. Specifically, for a finite subset $\mathcal{A} \subset \mathbb{N}^n = \mathbb{Z}^n_{\geq 0}$, we write

$$f = \sum_{\alpha \in \mathcal{A}} c_\alpha x^\alpha$$

where $x^\alpha := x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ and $c_\alpha \in \mathbb{C}$. For a sparse polynomial f, we associate to it a polytope called the *Newton polytope* of f which is defined as the convex hull of its exponent vectors. It is denoted $\mathrm{Newt}(f) = \mathrm{Conv}\{\alpha \ : \ \alpha \in \mathcal{A}\}$. A *sparse polynomial system* $\mathbf{F} = (f_0, \ldots, f_m)$ is then defined by a tuple $\mathcal{A} = (\mathcal{A}_0, \ldots, \mathcal{A}_m)$ where $f_i = \sum_{\alpha \in \mathcal{A}_i} c_{\alpha,i} x^\alpha \in \mathbb{C}[x_1, \ldots, x_n]$.

Remark 2.2.1. In this chapter we consider *generic* sparse polynomial systems. A statement holds for a *generic* sparse polynomial system if it holds for all systems $\mathbf{F}$ where for each $i = 0, \ldots, m$ the coefficients $\{c_{\alpha,i}\}_{\alpha \in \mathcal{A}_i}$ of f_i lie in some nonempty Zariski open subset of the space $\mathbb{C}^{\mathcal{A}_0} \times \cdots \times \mathbb{C}^{\mathcal{A}_m}$ of coefficients. This means that the non-generic behavior occurs on a set of measure zero in the space $\mathbb{C}^{\mathcal{A}_0} \times \cdots \times \mathbb{C}^{\mathcal{A}_m}$.

Given polytopes $P_1, \ldots, P_n \subset \mathbb{R}^n$, the *mixed volume* of $P_1, \ldots, P_n$ is the coefficient in front of the monomial $\lambda_1 \cdots \lambda_n$ of the polynomial

$$\mathrm{Vol}_n(\lambda_1 P_1 + \ldots + \lambda_n P_n)$$

where $P+Q = \{p+q \ : \ p \in P, \ q \in Q\}$ is the Minkowski sum and Vol_n is the standard n-dimensional Euclidean volume.

In a series of celebrated results [Ber75, Kou76, Kho78] the connection between the number of solutions over $\mathbb{C}^* := \mathbb{C}\backslash\{0\}$ to a system of sparse polynomial equations and the underlying convex geometry of the polynomials was made.

Theorem 2.2.2 (BKK Bound [Ber75, Kou76, Kho78]). *Let* $\mathbf{F} = (f_1, \ldots, f_n) \subset \mathbb{C}[x_1, \ldots, x_n]$ *be a sparse polynomial system with* η *solutions in* $(\mathbb{C}^*)^n$*, counted with multiplicity, and let* $P_i = \text{Newt}(f_i)$*. Then*

$$\eta \leq \text{MVol}(P_1, \ldots, P_n).$$

Moreover, if the coefficients of $\mathbf{F}$ *are general then* $\eta = \text{MVol}(P_1, \ldots, P_n)$*.*

If for a sparse polynomial system $\mathbf{F}$ the BKK bound holds with equality, we say $\mathbf{F}$ is *BKK general*. Bernstein gave explicit degeneracy conditions under which the above inequality is tight by considering the initial systems of $\mathbf{F}$ [Ber75].

Given a polytope $P \subseteq \mathbb{R}^n$ and a vector $w \in \mathbb{Z}^n \backslash \{0\}$, let P_w denote the *face exposed* by w. Specifically,

$$P_w = \{v \in P \ : \ \langle v, w \rangle \leq \langle y, w \rangle \ \forall y \in P\}.$$

For a sparse polynomial f we call

$$\text{init}_w(f) = \sum_{\alpha \in (\text{Newt}(f))_w} c_\alpha x^\alpha$$

the *initial polynomial* of f with respect to w. For a sparse polynomial system $\mathbf{F}$, we denote $\text{init}_w(\mathbf{F}) = (\text{init}_w(f_1), \ldots, \text{init}_w(f_n))$.

Theorem 2.2.3 (Theorem 2, [Ber75]). *Let* $\mathbf{F} = (f_1, \ldots, f_n) \subset \mathbb{C}[x_1, \ldots, x_n]$ *be a sparse polynomial system with* η *isolated* $\mathbb{C}^*$ *solutions counted with multiplicity and let* $P_i = \text{Newt}(f_i)$*. All* $\mathbb{C}^*$ *solutions of* $F(x) = 0$ *are isolated and* $\eta = \text{MVol}(P_1, \ldots, P_n)$ *if and only if for every* $w \in \mathbb{Z}^n \backslash \{0\}$*,* $\text{init}_w(\mathbf{F})$ *has no* $\mathbb{C}^*$ *solutions.*

Theorem 2.2.2 and Theorem 2.2.3 demonstrate the intimate connection between solutions to systems of polynomial equations and polyhedral geometry. In the remainder of this section we define a few more objects that are helpful when using this connection.

Given $\mathcal{A} = (\mathcal{A}_0, \ldots, \mathcal{A}_m) \in \mathbb{N}^n$ we define the *Cayley polytope* of $\mathcal{A}$ as

$$\text{Cay}(\mathcal{A}) = \text{Conv}\left(\{(x, e_i) \ : x \in \mathcal{A}_i, i = 0, \ldots, m\}\right) \subset \mathbb{R}^{n+m}$$

where e_i is the ith standard basis vector of $\mathbb{R}^m$ and e_0 is the vector of all zeroes. Similarly, for a sparse polynomial system $\mathbf{F}$, with support $\mathcal{A} = (\mathcal{A}_0, \ldots, \mathcal{A}_m)$ we define $\text{Cay}(\mathbf{F}) = \text{Cay}(\mathcal{A})$.

For a face of a convex polytope, $F \subset P \subset \mathbb{R}^n$, the *normal cone* of F is the set of linear functionals which achieve their minimum on F, i.e.

$$\sigma(F) = \{c \in \mathbb{R}^n \ : \ \langle c, x \rangle \leq \langle c, y \rangle, \ \forall x \in F, \ y \in P\}.$$

The normal cones of each face of P form a fan, denoted $\sigma(P) \subset \mathbb{R}^n$.

Finally, throughout the rest of this chapter we will consider the operation of taking the "partial derivative" of a polytope which we define as follows. Let $P \subset \mathbb{R}^n_{\geq 0}$ be a polytope contained in the positive orthant. Then

$$\partial_j P = (P - e_j) \cap \mathbb{R}^n_{\geq 0} = \{\alpha - e_j \ : \ \alpha \in P, \ \alpha_j \geq 1\},$$

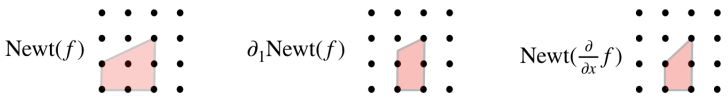
where e_j is the $j-$th standard basis vector of $\mathbb{R}^n$.

Of course, the definition of $\partial_j P$ is motivated by the partial differentiation operation of a polynomial $f \in \mathbb{R}[x_1, \ldots, x_n]$ with $\text{Newt}(f) = P$. Indeed, one always has $\text{Newt}(\frac{\partial}{\partial x_j} f) \subset \partial_j \text{Newt}(f)$. However, the inclusion $\text{Newt}(\frac{\partial}{\partial x_j} f) \subseteq \partial_j \text{Newt}(f)$ can be strict. In general, even if P is integral, the polytope $\partial_i P$ does not have to be integral:

Example 2.2.4. Let f be the bivariate polynomial

$$f = 1 + x + y + xy + x^2 + x^2y + x^2y^2.$$

The three polytopes $\mathrm{Newt}(f), \partial_1\mathrm{Newt}(f)$ and $\mathrm{Newt}(\frac{\partial}{\partial x}f)$ are displayed below.



For a polynomial $f = \sum_{\alpha\in\mathcal{A}} c_\alpha x^\alpha$ having full monomial support (i.e. $c_\alpha \neq 0$ for any $\alpha \in \mathrm{Newt}(f) \cap \mathbb{N}^n$) the two constructions are connected in the following way:

$$\mathrm{Newt}\left(\frac{\partial}{\partial x_j}f\right) = \mathrm{Conv}\{\partial_j\mathrm{Newt}(f) \cap \mathbb{N}^n\}.$$

In particular, if f is a degree d polynomial with all monomials of degree $\leq d$ having non-zero coefficients, then

$$\mathrm{Newt}\left(\frac{\partial}{\partial x_j}f\right) = \partial_j\mathrm{Newt}(f).$$

2.2.2 Toric varieties

Theorem 2.2.2 can be seen as an intersection theory question on toric varieties. A *toric variety* X is an irreducible variety such that $(\mathbb{C}^*)^n$ is a Zariski open subset of X and the action of $(\mathbb{C}^*)^n$ on itself extends to an action of $(\mathbb{C}^*)^n$ on X. We can also associate normal toric varieties to polyhedral fans.

Let $\sigma \in \mathbb{R}^n$ be a rational polyhedral cone which does not contain any vector subspace and denote

$$S_\sigma = \sigma^\vee \cap \mathbb{Z}^n$$

where $\sigma^\vee = \{y \in \mathbb{R}^n \ : \ \langle y, x\rangle \geq 0 \ \forall x \in \sigma\}$ is the dual cone of σ. Then the *affine toric variety* associated to σ is

$$V_\sigma = \mathrm{Spec}(\mathbb{C}[S_\sigma])$$

where $\mathbb{C}[S_\sigma]$ is the semigroup algebra associated to S_σ.

Given a polyhedral fan Σ we have a collection of affine toric varieties indexed by cones in Σ, denoted $\{V_\sigma \ : \ \sigma \in \Sigma\}$. This collection of toric varieties can be 'glued' together to create the toric variety X_Σ as follows. Given $\sigma, \tau \in \Sigma$, then $\rho = \sigma \cap \tau \in \Sigma$ is a face of both σ and τ. This induces the inclusion $V_\rho \subset V_\sigma$ and $V_\rho \subset V_\tau$. We then glue V_σ and V_τ by identifying of the common open subset V_ρ. For a more complete treatment of toric varieties, see [CLS11].

In this chapter we will work with the *total coordinate ring* or *Cox ring* of a toric variety which is a generalization of homogeneous coordinate ring of projective space introduced in [Cox95]. First, let us denote by $\Sigma(1)$ the set of all rays of Σ, where by abuse of notation we often do not distinguish between rays ρ and their primitive ray generators. The Cox ring of X_Σ is

$$S = \mathbb{C}[x_\rho \ : \ \rho \in \Sigma(1)].$$

To every ray ρ we associate the corresponding torus invariant Weyl divisor D_ρ. Note that every torus invariant Weyl divisor D on X is a free linear combination $D = \sum_{\rho\in\Sigma(1)} a_\rho D_\rho$. Then the global sections of the associated sheaf $\mathcal{O}_X(D)$ are spanned by monomials:

$$H^0(\mathcal{O}_X(D), X) = \langle X^m \mid \ m \in \mathbb{Z}^n,\ \langle m, \rho\rangle \geq -a_\rho \rangle\,.$$

Given a global section $f = \sum_{m\in\mathbb{Z}^n} c_m X^m$ of $\mathcal{O}_X(D)$, we define the homogenization $\widetilde{f} \in S$ of f:

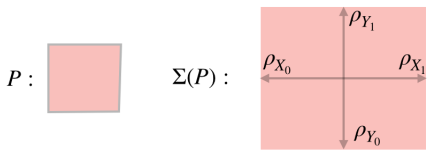
$$\widetilde{f} = \prod_{\rho\in\Sigma(1)} x_\rho^{a_\rho} f(z_1, \ldots, z_n). \tag{2.3}$$

Here the variables z_i are defined by $z_i = \prod x_\rho^{\rho_i}$. Expanding equation (2.3) reads

$$\widetilde{f} = \sum_{m\in\mathbb{Z}^n} c_m \prod_{\rho\in\Sigma(1)} x_\rho^{\langle m,\rho\rangle + a_\rho}.$$

Note that a Laurent polynomial $f \in \mathbb{C}[X_1^{\pm}, \ldots, X_n^{\pm}]$ can be a section of the sheaf $\mathcal{O}_X(D)$ for a certain choice of D, and the homogenization $\widetilde{f}$ depends on the choice of D.

Example 2.2.5. Consider the toric variety $\mathbb{P}^1 \times \mathbb{P}^1$, associated to the dual fan $\Sigma(P)$ of the square P. The four generators X_0, X_1, Y_0, Y_1 of the Cox ring S are in bijection to the four rays.



Homogenizing the bivariate polynomial $f = 1 + x + y + xy$ yields the bihomogeneous polynomial $\widetilde{f} = X_0Y_0 + X_1Y_0 + X_0Y_1 + X_1Y_1$.

2.2.3 Chern and Segre classes of vector bundles

The main ingredient of the intersection theoretic formulas for the algebraic degree of polynomial optimization problems given in [NR09] is Porteous' formula. Porteous' formula computes the expected cohomology class of the degeneracy locus of maps of vector bundles. Loosely speaking, vector bundles are families of vector spaces that are parameterized by another space and cohomology classes are algebraic invariants of topological spaces. In this chapter all vector spaces will be parameterized by algebraic varieties and the vector spaces will all have the same dimension, called the *rank* of the vector bundle. To formulate Porteous formula one needs to use Chern and Segre classes, which are well-studied characteristic classes of vector bundles. Here we list some main properties of these classes. For more detailed introduction we refer to [EH16].

For a vector bundle $\mathcal{E}$ of rank r on a variety X of dimension d and for any $i = 0, \ldots, d$, we denote its ith Chern class by $c_i(\mathcal{E})$. One has $c_0(\mathcal{E}) = 1$ and $c_i(\mathcal{E}) = 0$ for any $i > r$. We will denote by $c(\mathcal{E})$ the *total Chern class* of $\mathcal{E}$, that is

$$c(\mathcal{E}) = c_0(\mathcal{E}) + \cdots + c_{\max(d,r)}(\mathcal{E}).$$

The crucial property of total Chern classes, known as Whitney's formula, is that they are multiplicative under taking direct sums with line bundles:

$$c(\mathcal{E} \oplus F) = c(\mathcal{E}) \cdot c(\mathcal{F}).$$

In what follows we will mostly work with vector bundles coming as a direct sum of line bundles $E = \mathcal{L}_1 \oplus \ldots \oplus \mathcal{L}_n$. By applying Whitney's formula to such vector bundles we obtain a convenient formula for their total Chern class:

$$c(\mathcal{E}) = \prod_{i=1}^{n}(1 + c_1(\mathcal{L}_i)). \text{ These are graded pieces: } c_k(\mathcal{E}) = \sum_{I \in \binom{[n]}{k}} \prod_{i \in I} c_1(\mathcal{L}_i).$$

Finally, let us recall the definition of Segre classes. Note that, the total Chern class $c(\mathcal{E})$ is an invertible element in the cohomology ring of X as its 0-th degree part is equal to 1. Using this one defines a total Segre class $s(\mathcal{E})$ of a vector bundle $\mathcal{E}$ on X to be the inverse of the total Chern class of $\mathcal{E}$:

$$s(\mathcal{E}) = \frac{1}{c(\mathcal{E})}.$$

Individual Segre classes $s_i(\mathcal{E})$ are defined as homogeneous components of the total Segre class. Note that unlike Chern classes, one could have non-trivial Segre class $s_i(\mathcal{E})$ even for $i > r$.

2.3 Statement of the main result

In this section we give an overview of the main results of this chapter and defer the proofs of Theorem 2.3.6 and Theorem 2.3.11 to Section 2.6. The results in this chapter will be proven under certain assumptions on the monomial support of $\mathbf{F}$, which we define in the following. While extensive numerical experiments suggest that both Theorem 2.3.6 and Theorem 2.3.7 are true without these assumptions, we demonstrate in Example 2.3.12 that Theorem 2.3.11 may fail if we drop them.

Although a more detailed discussion of how the assumptions on the monomial support come into play is given at the beginning of Section 2.6, we already say a few explanatory words. The notion of an *admissible* point configuration guarantees that all considered toric varieties contain a distinct copy of affine space $\mathbb{C}^n$, which we use in the proof of Proposition 2.6.2 below. It guarantees the non-vanishing of the gradient of f_0 on the boundary of a toric compactification.

Definition 2.3.1. We call a point configuration $\mathcal{A} = (\mathcal{A}_0, \ldots, \mathcal{A}_m) \in \mathbb{N}^n$ *admissible* if:

1. the positive orthant $\mathbb{R}^n_{\geq 0}$ is a cone in the normal fan of the polytope $\mathrm{Conv}(\mathcal{A}_i)$ for each $i = 0, \ldots, m$, and
2. $\mathrm{Conv}(\mathcal{A}_i)$ meets every coordinate hyperplane of $\mathbb{R}^n$ for each $i = 0, \ldots, m$.

When passing to a toric compactification, for various technical reasons we need to ensure that the constraints $f_1, \ldots, f_m$ define a variety with a smooth closure. This is guaranteed by the following notion of an *appropriate* toric variety and used in the proof of Proposition 2.6.1 below.

Definition 2.3.2. Let X be a proper normal toric variety with underlying polyhedral fan Σ in $\mathbb{R}^n$ and $\mathcal{A} = (\mathcal{A}_0, \ldots, \mathcal{A}_m)$ an admissible point configuration. We call X *appropriate* for $\mathcal{A}$ if the following three properties hold:

1. for each $i = 0, \dots, m$ the normal fan $\Sigma(\mathrm{Conv}(\mathcal{A}_i))$ is refined by Σ ,

2. the fan Σ contains the positive orthant $\mathbb{R}^n_{\geq 0}$ as a cone,

3. for generic functions f_i with monomial support $\mathcal{A}_i$, the closure of $\mathcal{V}(f_1, \dots, f_m)$ in X is disjoint from the singular locus of X.

Remark 2.3.3. Note that for every admissible point configuration $\mathcal{A}$ there exists a smooth appropriate toric variety X. To construct X consider the normal fan Σ' of the Minkowski sum $\mathrm{Conv}(\mathcal{A}_0) + \cdots + \mathrm{Conv}(\mathcal{A}_m)$. A resolution of singularities can be performed by subdividing each singular cone of Σ', resulting in a smooth, complete polyhedral fan Σ which contains the positive orthant. For more details on toric resolution of singularities consider Chapter 11 of [CLS11].

In the proof of Theorem 2.3.6 and Theorem 2.3.11 we consider a natural choice for the toric compactification X, given by the coarsest refinement of all normal fans of the newton polytopes $\mathrm{Newt}(f_0), \dots, \mathrm{Newt}(f_m)$. It is desirable that this compactification is appropriate. This leads us to the following definition:

Definition 2.3.4. We call a point configuration $\mathcal{A} = (\mathcal{A}_0, \dots, \mathcal{A}_m) \in \mathbb{N}^n$ *strongly admissible* if it is admissible and if the variety

$$X = X(\Sigma(\mathrm{Conv}(\mathcal{A}_0) + \cdots + \mathrm{Conv}(\mathcal{A}_m)))$$

is appropriate for $\mathcal{A}$. Here X is the toric variety associated to the common refinement of the normal fans $\Sigma(\mathrm{Conv}(\mathcal{A}_i))$.

We give examples for the above definitions.

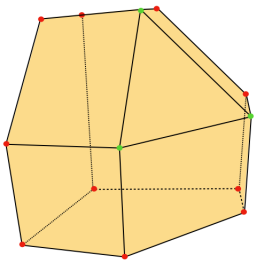
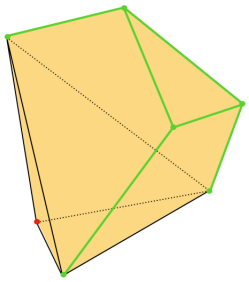
Example 2.3.5. Let $\mathcal{S} \subseteq \mathbb{R}^3$ denote the tetrahedron $\mathrm{Conv}(0, e_1, e_2, e_3)$, let $\mathcal{C}$ denote the three-dimensional cube $\mathrm{Conv}(0, e_1, e_2, e_3, e_1{+}e_2, e_1{+}e_3, e_2{+}e_3, e_1{+}e_2{+}e_3)$ and let $\mathcal{B}$ denote the bipyramid $\mathrm{Conv}(0, e_1, e_2, e_3, e_1 + e_2 + e_3)$.

- The integer points of $\mathcal{S}$ and the points of $-\mathcal{S} + (1, 1, 1)$ do not form an admissible point configuration since the normal fan of $-\mathcal{S}$ does not contain the positive orthant as a cone.

- The integer points of $\mathcal{S}$ and $\mathcal{C}$ form a strongly admissible point configuration. This is because the singular locus of the toric variety defined by $\mathcal{S} + \mathcal{C}$ is zero-dimensional.

- The tetrahedron $\mathcal{S}$ and the bipyramid $\mathcal{B}$ define an admissible, but not strongly admissible, point configuration. In particular, the toric variety defined by $\mathcal{S} + \mathcal{B}$ is not appropriate for the considered point configuration. To see this consider the three-dimensional toric variety defined by $\mathcal{S} + \mathcal{B}$. The torus orbit defined by the cone $\sigma = \mathbb{R}_+(-1, -1, 1) + \mathbb{R}_+(-1, -1, -1)$, dual to the face $\mathrm{Conv}((2, 1, 1), (1, 2, 1))$ of $S + B$, is contained in the singular locus. Further, said orbit intersects V, since the face $\mathrm{Conv}((1, 0, 0), (0, 1, 0))$ of $\mathcal{S}$ revealed by σ is not a vertex.

The polytopes $\mathcal{S} + \mathcal{B}$ and $\mathcal{S} + \mathcal{C}$ are displayed below, with the faces that define singular torus orbits coloured in green.

$\mathscr{S} + \mathscr{C} =$  $\mathscr{S} + \mathscr{B} =$ 

The next two results express the number of solutions of $\mathbf{L}_{\mathrm{F}}$ as mixed volumes.

Theorem 2.3.6. *Let $\mathbf{F} = (f_0, \ldots, f_m)$ be a generic sparse system of polynomials in $\mathbb{C}[x_1, \ldots, x_n]$ with strongly admissible support. Then the algebraic degree of sparse polynomial optimization of (POP) is equal to*

$$\mathrm{MV}\left(\mathrm{Newt}(f_1), \ldots, \mathrm{Newt}(f_m), \partial_1 \mathrm{Newt}(\Phi_F), \ldots, \partial_n \mathrm{Newt}(\Phi_F)\right). \tag{2.4}$$

Here Φ_F denotes the Lagrangian $\Phi_F(\lambda, x) := f_0 - \sum_{i=1}^m \lambda_i f_i$, as above. Note that the polytopes $\partial_j \mathrm{Newt}(\Phi_F)$ might be strictly larger than the Newton polytopes $\mathrm{Newt}(\ell_j)$ of the partial differentials $\ell_j = \frac{\partial}{\partial x_j}(f_0 - \sum_{i=1}^m \lambda_i f_i)$. While the Newton polytopes of $\ell_1, \ldots, \ell_n$ do depend on the exact monomial support of $f_0, \ldots, f_m$, the mixed volume (2.4) does only depend on the convex hulls of $\mathcal{A}_0, \ldots, \mathcal{A}_m$. In particular, by Theorem 2.3.6, also the number of critical points only depends on the Newton polytopes of $f_0, \ldots, f_m$, and not their exact monomial support.

It is natural to ask whether the BKK bound of $\mathbf{L}_{\mathrm{F}}$ does depend on the exact monomial support of $f_0, \ldots, f_m$. The following theorem shows that, although we might have a strict inclusion of polytopes $\mathrm{Newt}(\ell_j) \subseteq \partial_j \mathrm{Newt}(\Phi_F)$, the BKK bound of $\mathbf{L}_{\mathrm{F}}$ is equal to the mixed volume (2.4).

Theorem 2.3.7. *Under the assumptions of Theorem 2.3.6 the Lagrange system $\mathbf{L}_{\mathrm{F}}$ is BKK general and all critical points are smooth. The number of solutions is the mixed volume*

$$\mathrm{MV}\left(\mathrm{Newt}(f_1), \ldots, \mathrm{Newt}(f_m), \mathrm{Newt}(\ell_1), \ldots, \mathrm{Newt}(\ell_n)\right). \tag{2.5}$$

If $\mathbf{F}$ is not generic then (2.5) is an upper bound for the number of isolated, smooth critical points of f_0 restricted to $\mathcal{V}(f_1, \ldots, f_m)$.

Proof. For every $j = 1, \ldots, n$ we have the inclusion $\mathrm{Newt}(\ell_j) \subseteq \partial_j \mathrm{Newt}(\Phi_F)$ of polytopes, showing the inequality (2.5)$\leq$(2.4). On the other hand, by Theorem 2.3.6, the BKK bound (2.5) of $\mathbf{L}_{\mathrm{F}}$ constitutes an upper bound to (2.1) and we obtain

$$(2.1) \leq (2.5) \leq (2.4) = (2.1).$$

By Lemma 2.6.4, all critical points are smooth. □

We obtain the following corollary from Theorem 2.3.6 and Theorem 2.3.7.

Corollary 2.3.8. *Under the assumptions of Theorem 2.3.6 the algebraic degree of the sparse polynomial optimization problem (POP) and the mixed volume (2.5) depend only on the convex hulls of $\mathrm{Newt}(f_i)$ for $i = 0, \ldots, m$.*

Remark 2.3.9. Corollary 2.3.8 has algorithmic consequences if one wishes to numerically find all critical points to (POP) (as opposed to counting them). We leverage this at the end of this chapter and efficiently compute polyhedral start systems for $\mathbf{L}_{\mathrm{F}}$ by imposing maximal sparsity.

Example 2.3.10. Recall the optimization problem (2.2) in Example 2.1.1. Theorem 2.3.7 shows that the algebraic degree of this problem is equal to the mixed volume of its corresponding Lagrange system. A property of mixed volumes is that if $P_1, \dots, P_n \subset \mathbb{R}^n$ and $Q_1, \dots, Q_m \subset \mathbb{R}^{n+m}$, then

$$\mathrm{MVol}(P_1, \dots, P_n, Q_1, \dots, Q_m) = \mathrm{MVol}(P_1, \dots, P_n) \cdot \mathrm{MVol}(\pi(Q_1), \dots, \pi(Q_m)),$$

where $\pi : \mathbb{R}^{n+m} \to \mathbb{R}^m$ is the projection onto the last m coordinates. Observe that $\mathrm{Newt}(\ell_1), \dots, \mathrm{Newt}(\ell_n) \subset \mathbb{R}^{n+1}$ have nth coordinate zero. Therefore,

$$\mathrm{MVol}(\mathrm{Newt}(\ell_1), \dots, \mathrm{Newt}(\ell_n), \mathrm{Newt}(f)) = \mathrm{MVol}(\mathrm{Newt}(\ell_1), \dots, \mathrm{Newt}(\ell_n)) \cdot \mathrm{MVol}(\pi_n(\mathrm{Newt}(f))$$

where $\pi_n : \mathbb{R}^{n+1} \to \mathbb{R}$ is the projection onto the nth coordinate.

Since $\mathrm{Newt}(\ell_j) = \mathrm{Conv}(0, \alpha_j)$ for $j \in [n]$ and some $\alpha_j \in \mathbb{R}^n$, we can compute $\mathrm{MVol}(\mathrm{Newt}(\ell_1), \dots, \mathrm{Newt}(\ell_n)) = \det(M)$ where M is the matrix with jth column equal to α_j. In our case this amounts to computing

$$\det\left(\begin{bmatrix} 2 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & & \\ 0 & 0 & \ddots & 0 \\ 1 & 1 & \dots & 1 \end{bmatrix}\right) = 2.$$

Finally, observe that $\pi_n(\mathrm{Newt}(f)) = [0, 1]$ so it has (mixed) volume one. This gives a geometric proof that the optimization degree of (2.2) is 2, agreeing with the result we computed in Example 2.1.1.

Our next result, Theorem 2.3.11, characterizes the number of solutions to $\mathbf{L}_{\mathrm{F}}$ under slightly weaker assumptions compared to Theorem 2.3.6 and Theorem 2.3.7, since the monomial support $\mathcal{A}$ needs only be admissible instead of strongly admissible. We obtain a description not as a mixed volume but as a more general product in the Chow ring of a toric variety.

To formulate Theorem 2.3.11, we first need some notation. We refer to Section 2.2 and references therein for a brief introduction to the objects we use. Let $\mathcal{A} = (\mathcal{A}_0, \dots, \mathcal{A}_m)$ be an admissible point configuration (Definition 2.3.1) and let X be a smooth toric variety given by the fan Σ which is appropriate for $\mathcal{A}$ (Definition 2.3.2). As usual, the convex hull of each point configuration $\mathcal{A}_i$ defines a line bundle $\mathcal{L}_{\mathcal{A}_i}$.

Further, since we assume that the fan Σ contains the positive orthant as one of its cones, we know that Σ containes the rays generated by the standard basis vectors $e_1, \dots, e_n$ of $\mathbb{R}^n$. We denote by $D_{e_1}, \dots, D_{e_n}$ the corresponding torus-invariant divisors on X and by $\mathcal{O}_X(D_{e_1}), \dots, \mathcal{O}_X(D_{e_n})$ the corresponding line bundles. For further reading on toric line bundles consider chapter 6 in [CLS11].

Theorem 2.3.11. *Let $\mathbf{F} = (f_0, \dots, f_m)$ be a generic sparse system of polynomials in $\mathbb{C}[x_1, \dots, x_n]$ with admissible support $\mathcal{A}$ and let X be as above. Then the algebraic degree of sparse polynomial optimization (POP) is finite and equal to the degree of the following cycle class:*

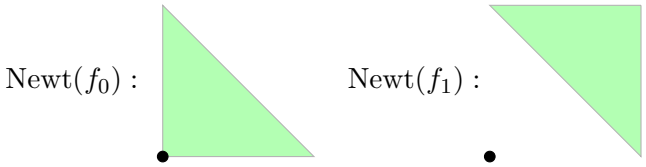
$$c_1(\mathcal{L}_{\mathcal{A}_1}) \dots c_1(\mathcal{L}_{\mathcal{A}_n}) \left(\mathrm{s}(\mathcal{E})\, \mathrm{c}(\mathcal{F})\right)_{n-m}, \tag{2.6}$$

where $\mathcal{E} = \mathcal{L}_{\mathcal{A}_0}^{-1} \oplus \dots \oplus \mathcal{L}_{\mathcal{A}_m}^{-1}$ and $\mathcal{F} = \mathcal{O}_X(-D_{e_1}) \oplus \dots \oplus \mathcal{O}_X(-D_{e_n})$. Moreover, if $\mathbf{F}$ is not generic then (2.6) is an upper bound to the number of isolated, smooth critical points of f_0 restricted to $\mathcal{V}(f_1, \dots, f_m)$.

The purpose of the following example is to demonstrate that the assumptions for Theorem 2.3.11 are necessary.

Example 2.3.12. Consider the following objective f_0 and constraint f_1.

$$f_0 = 7 + 11x - 13y - 19xy - 2x^2 - 5y^2$$
$$f_1 = -5xy + 29xy^2 - 17x^2y + 61x^2y^2 + x^2 - 3y^2.$$



Note that the inner normal fan of Newt(f_1) does not contain the positive orthant $\mathbb{R}^2_{\geq 0}$. In fact, evaluating equation (2.6) amounts to the number 12, while the actual number of isolated solutions to $\mathbf{L}_F$ in the torus are 10. In particular, Theorem 2.3.11 does not hold true. On the other hand, the BKK bound of the corresponding Lagrange system equals 10, so Theorem 2.3.6 and Theorem 2.3.7 hold true. This discrepancy is not due to the specific choice of coefficients for f_0 and f_1. We note that, although the assumptions on Theorem 2.3.6 and Theorem 2.3.7 are stronger than the ones on Theorem 2.3.11, we believe that they hold in greater generality.

We give a rough sketch of how to evaluate (2.6). We denote the line bundles

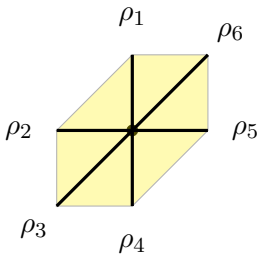
$$c_1\left(\mathcal{L}_{\mathcal{A}_1}\right) = [2D_2 + 4D_3 + 2D_4 - 2D_6], \text{ and } c_1\left(\mathcal{L}_{\mathcal{A}_2}\right) = [2D_2 + 2D_3 + 2D_4],$$

and the vector bundles

$$\mathcal{E} = \mathcal{L}_{\mathcal{A}_0}^{-1} \oplus \mathcal{L}_{\mathcal{A}_1}^{-1} \text{ and } \mathcal{F} = \mathcal{O}_X(D_1)^{-1} \oplus \mathcal{O}_X(D_5)^{-1}.$$

Here, X is the smooth toric variety defined by the complete fan Σ with ray generators

$$\rho_1 = (0,1),\ \rho_2 = (1,1),\ \rho_3 = (1,0),\ \rho_4 = (0,-1),\ \rho_5 = (-1,-1),\ \rho_6 = (-1,0).$$



We have

$$c_1(\mathcal{F}) = [-D_1 - D_5] \text{ and } s_1(\mathcal{E}) = \frac{-1}{c_1(\mathcal{E})} = c_1\left(\mathcal{L}_{\mathcal{A}_0}\right) + c_1\left(\mathcal{L}_{\mathcal{A}_1}\right) = [4D_2 + 6D_3 + 4D_4 - 2D_6].$$

Finally, direct computation shows

$$c_1\left(\mathcal{L}_{\mathcal{A}_1}\right) \cdot \left(\mathrm{s}(\mathcal{E})\,\mathrm{c}(\mathcal{F})\right)_1 = c_1\left(\mathcal{L}_{\mathcal{A}_1}\right) \cdot \left(c_1(\mathcal{F}) + s_1(\mathcal{E})\right) = 12.$$

A short explanation is in order concerning the discrepancy between the number 10, of actual critical points, and the number 12, coming from our cohomological computation. This difference comes from an intersection at the toric boundary, which occurs even for a generic choice of coefficients for f_0 and f_1. Consider the homogenization $\left(\widetilde{\frac{\partial f_0}{\partial x}}, \widetilde{\frac{\partial f_0}{\partial y}}\right)$ of the gradient of f_0 in the Cox ring. Each entry of this gradient is divisible by the variable X_6, associated to the ray ρ_6. There are two points on the toric divisor $D_6 = \mathcal{V}(X_6)$ where $\widetilde{f_1}$ vanishes. These make up for the difference. In conclusion, although the affine equations $\left(f_1, \frac{\partial f_0}{\partial x}\frac{\partial f_1}{\partial y} - \frac{\partial f_0}{\partial y}\frac{\partial f_1}{\partial x}\right)$ generate the ideal for the set of critical points, their homogenizations $\left(\widetilde{f_1}, \widetilde{\frac{\partial f_0}{\partial x}}\widetilde{\frac{\partial f_1}{\partial y}} - \widetilde{\frac{\partial f_0}{\partial y}}\widetilde{\frac{\partial f_1}{\partial x}}\right)$ do not form generators of the homogenized ideal.

This example brings this section to and end, and we conclude with some implications to polar and ED degrees in the next section.

2.4 Sparse ED, polar and sectional degrees

In this section, we discuss important corollaries of Theorem 2.3.6 and Theorem 2.3.7 which relate Euclidean distance optimization, polar degrees and sectional degrees to mixed volumes.

We consider polynomial optimization problems where the objective function is $f_0 = \|x-u\|_2^2$ for a generic point $u \in \mathbb{R}^n$. Let $(f_1, \ldots, f_m)$ be a general sparse polynomial system, and $u \in \mathbb{R}^n$ a general point. The *ED degree* of $(f_1, \ldots, f_m)$ is the number of complex critical points of the optimization problem:

$$\min_{x\in\mathbb{R}^n} \|x-u\|_2^2 \quad \text{subject to} \quad f_1(x) = \ldots = f_m(x) = 0. \tag{ED}$$

Equivalently, it is the number of complex critical points to the corresponding Lagrange system of $\mathbf{F} = (f_0, f_1, \ldots, f_m)$, namely $\mathbf{L}_\mathbf{F} = (\ell_1, \ldots, \ell_n, f_1, \ldots, f_m)$. This brings us to the main result of this section, which relates ED degrees and mixed volumes.

Corollary 2.4.1 (Euclidean distance objective function)**.** *If $f_0 = \|x-u\|_2^2$ is the squared Euclidean distance function for a generic point u of $\mathbb{R}^n$ and $(f_1, \ldots, f_m)$ is a general sparse polynomial system such that the support of $\mathbf{F} = (f_0, \ldots, f_m)$ is strongly admissible, then the mixed volume and degree of the Lagrange system $\mathbf{L}_\mathbf{F}$ are equal.*

Proof. First, consider the *weighted Euclidean distance function* $f_C = \|Cx-u\|_2^2$ where C is an $n \times n$ diagonal matrix with general entries and call $F_C = (f_C, f_1, \ldots, f_m)$. Theorem 2.3.6 implies that the degree and mixed volume of $\mathbf{L}_{F_C}$ are equal. We call this value η.

Observe that the variety of $\mathbf{L}_{F_c}$ is in bijection with the critical points of

$$\min_{x\in\mathbb{R}^n} \|x-u\|_2^2 \quad \text{subject to} \quad f_1(C^{-1}x) = \ldots = f_m(C^{-1}x) = 0. \tag{2.7}$$

This gives that there are η critical points to (2.7). Notice that the monomial support of $(f_1(C^{-1}(x)), \ldots, f_m(C^{-1}(x))$ is the same as that of $(f_1(x), \ldots, f_m(x))$, since C is diagonal. Therefore, the degree of $\mathbf{L}_F$ is equal to its mixed volume. □

In addition, we recall that in [DHO+14] a relationship between ED degrees and polar degrees was established. Let $X \subset \mathbb{P}^{n-1}$ be a projective variety and Y its dual. For a smooth point $x \in X$, denote T_xX as the tangent space of X at x. Denote the conormal variety of X as

$$\mathcal{N}_X = \overline{\{(x,y) \in \mathbb{P}^n \times \mathbb{P}^n \ : \ y \in Y \backslash Y_{\text{sing}},\ x \perp T_yX\}}.$$

Theorem 2.4.2 ([DHO+14, Theorem 5.4]). *If $\mathcal{N}_X$ does not intersect the diagonal $\Delta(\mathbb{P}^{n-1})$, then*

$$ED\ Degree(X) = \delta_0(X) + \delta_1(X) + \ldots + \delta_{n-1}(X)$$

where $\delta_i(X)$ is the ith polar degree of X.

Remark 2.4.3. For a variety $X \subset \mathbb{P}^n$, the ith polar degree $\delta_i(X)$ equals $|\mathcal{N}_X \cap (L_1 \times L_2)|$ where L_1 is a generic linear space of dimension $n+1-i$ and L_2 is a generic linear space of dimension i. The variety $\mathcal{N}_X$ has dimension $n-1$ so intersecting with the linear spaces L_1, L_2 amount to intersecting it with a variety of dimension $n+1$. This ensures that the intersection $\mathcal{N}_X \cap (L_1 \times L_2)$ is finite.

While the assumption of Theorem 2.4.2 that requires X to be an irreducible affine cone does not typicallyhold in our situation, we remark that by considering a variety $X \subset \mathbb{C}^n$ defined by polynomial equations with strongly admissible support, we can simply consider the *projective closure* of X, which we denote $\overline{X}$. Under sufficient generality conditions, the projective closure of X is defined by homogenizing the defining equations of X with respect to a new variable x_0. While the ED degree of X in general is not equal to that of $\overline{X}$ [DHO+16, Example 6.6], given that certain varieties intersect transversally, they are equal. Let $H_\infty = \mathbb{P}^n \backslash \mathbb{C}^n = \mathcal{V}(x_0)$ denote the hyperplane at infinity. Denote $X_\infty = \overline{X} \cap H_\infty$ and $Q_\infty = \{x_0^2 + \ldots + x_n^2 = 0\}$.

Theorem 2.4.4 (Theorem 6.11 [DHO+16]). *Let $X \subset \mathbb{C}^n$ be an irreducible, affine variety and $\overline{X} \subset \mathbb{P}^n$ its projective closure. Assume that the intersections $X_\infty = \overline{X} \cap H_\infty$ and $X_\infty \cap Q_\infty$ are both transversal. Then the ED degree of X is equal to the sum $\sum_{i=0}^n \delta_i(\overline{X})$, where $\delta_i(\overline{X})$ is the i-th polar degree of $\overline{X}$.*

As a consequence of Corollary 2.4.1 and Theorem 2.4.4 we are able to establish a relationship between polar degrees and mixed volumes. To our knowledge this is the first time a connection between convex geometry and polar degrees has been made.

Corollary 2.4.5. *Let $X \subset \mathbb{C}^n$ be an affine variety defined by polynomials $(f_1, \ldots, f_m)$ with strongly admissible support and let $\overline{X} \subset \mathbb{P}^n$ be its projective closure. Assume that the intersections $X_\infty = \overline{X} \cap H_\infty$ and $X_\infty \cap Q_\infty$ are both transversal. Let $\mathbf{L}_F = (\ell_1, \ldots, \ell_n, f_1, \ldots, f_m)$ be the Lagrange system of $\mathbf{F} = (f_0, \ldots, f_m)$ corresponding to the Euclidean distance optimization problem (ED). Then*

$$\mathrm{MVol}(\mathrm{Newt}(\ell_1), \ldots, \mathrm{Newt}(\ell_n), \mathrm{Newt}(f_1), \ldots, \mathrm{Newt}(f_m)) = \delta_0(\overline{X}) + \ldots + \delta_{n-1}(\overline{X})$$

where $\delta_i(\overline{X})$ is the ith polar degree of $\overline{X}$.

Example 2.4.6. Consider the Euclidean distance optimization problem

$$\min_{x \in \mathbb{R}^2} \left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\|_2^2 \quad \text{subject to} \quad 4x_1^2 + 2x_2^2 - x_1x_2 = 1$$

The Lagrange system of this optimization problem is

$$\mathbf{L}_F = (2(x_1 - 1) - \lambda(8x_1 - x_2), 2(x_2 - 1) - \lambda(-x_1 + 4x_2), 4x_1^2 + 2x_2^2 - x_1x_2 - 1).$$

The mixed volume of this system is four and there are indeed four complex solutions (x_1, x_2, λ), two of which are real:

$$[0.3864, 0.5557, -0.4840],$$
$$[-0.3050, -0.6418, 1.4516],$$
$$[-0.1407 - 1.6318\mathbf{i}, 2.3431 - 0.5950\mathbf{i}, 0.2904 - 0.1023\mathbf{i}],$$
$$[-0.1407 + 1.6318\mathbf{i}, 2.3431 + 0.5950\mathbf{i}, 0.2904 + 0.1023\mathbf{i}].$$

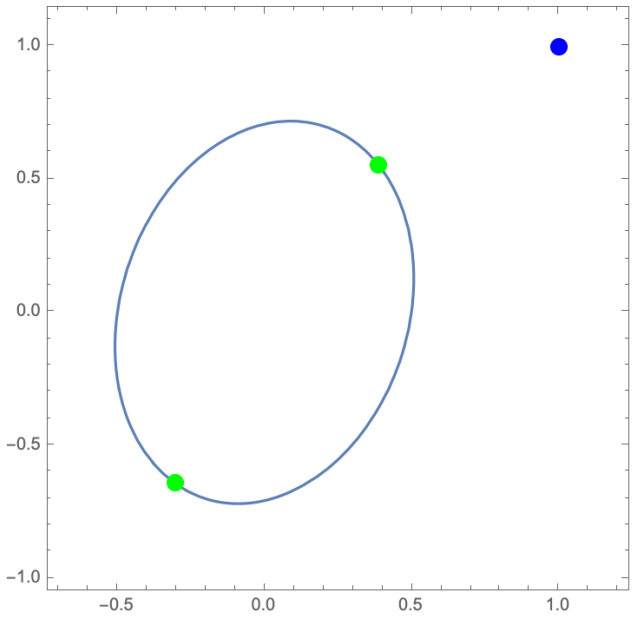


Figure 2.1: The ellipse $4x_1^2 + 2x_2^2 - x_1x_2 = 1$ along with the critical points (green) of the Euclidean distance problem from the point $(1, 1)$ (blue).

By Theorem 2.3.7 we know that the number of complex critical points to this optimization problem will be $\mathrm{MVol}(P_1, P_2, P_3)$ where $P_1 = \mathrm{Conv}\{(1,0,0),(1,0,1),(0,1,1)\}$, $P_2 = \mathrm{Conv}\{(0,1,0),(1,0,1),(0,1,1)\}$ and $P_3 = \mathrm{Conv}\{(2,0,0),(0,2,0),(1,1,0),(0,0,0)\}$.

Now we consider the projective closure of our variety which is defined by $\overline{X} = \mathcal{V}(4x_1^2 + 2x_2^2 - x_1x_2 - x_0^2) \subset \mathbb{P}^2$. The conormal variety of $\overline{X}$, namely $\mathcal{N}_{\overline{X}} \subset \mathbb{P}^2 \times (\mathbb{P}^2)^*$, is defined as the zero set of the following six polynomials:

$$-x_0^2 + 4x_1^2 - x_1x_2 + 2x_2^2,\ x_1y_1 - 4x_2y_1 + 8x_1y_2 - x_2y_2,\ 31y_0^2 - 8y_1^2 - 4y_1y_2 - 16y_2^2,$$
$$31x_2y_0 + 2x_0y_1 + 16x_0y_2,\ 31x_1y_0 + 8x_0y_1 + 2x_0y_2,\ x_0y_0 + 4x_2y_1 - 8x_1y_2 + 2x_2y_2.$$

The i- th polar degree $\delta_i(\overline{X})$ is given by the number of intersection points of $\mathcal{N}_{\overline{X}} \cap (L_1 \times L_2)$ where L_1 is a generic linear space of dimension $n+1-i$ and L_2 is a generic linear space of dimension i. In this case we have $\delta_0(\overline{X}) = 2$ and $\delta_1(\overline{X}) = 2$ and we see that the ED degree of $4x_1^2 + 2x_2^2 - x_1x_2 - 1$ equals the sum of the polar degrees of its projective closure as expected.

Finally, we conclude this section by making a final connection to *sectional degrees* as recently studied in [MRWW23]. Given an affine variety $X \subset \mathbb{C}^n$, the *ith sectional degree* of X, denoted

$s_i(X)$, is defined as the algebraic degree of the optimization problem

$$\min_{x \in \mathbb{R}^n} \langle u, x \rangle \quad \text{subject to} \quad x \in X \cap H_1 \cap H_2 \cap \cdots \cap H_i \tag{SO}$$

where $u \in \mathbb{R}^n$ is a generic linear function and $H_1, \ldots, H_i$ are generic affine linear hyperplanes. As an immediate consequence of Theorem 2.3.7 we have a convex algebraic interpretation of $s_i(X)$.

Corollary 2.4.7. *Let $X \subset \mathbb{C}^n$ be an affine variety defined by generic polynomials $(f_1, \ldots, f_m)$ with strongly admissible support. Let $\mathbf{L}_{\mathrm{F}} = (\ell_1, \ldots, \ell_n, f_1, \ldots, f_m)$ be the Lagrange system of $\mathbf{F} = (\langle u, x \rangle, f_1, \ldots, f_m)$ corresponding to the sectional optimization problem (SO). Then*

$$\mathrm{MVol}(\mathrm{Newt}(\ell_1), \ldots, \mathrm{Newt}(\ell_n), \mathrm{Newt}(f_1), \ldots, \mathrm{Newt}(f_m)) = s_i(X)$$

where $s_i(X)$ is the ith sectional degree of X.

Furthermore, by [MRWW23, Corollary 6.8] we have that if $X \subset \mathbb{C}^n$ is an affine variety with projective closure $\overline{X} \subset \mathbb{P}^n$ such that H_∞ is not contained in the dual of $\overline{X}$, then $s_i(X) = \delta_i(\overline{X})$ for all $0 \leq i \leq \dim(X)$. Given a polynomial system $\mathbf{F} = (f_1, \ldots, f_m)$ we use the notation $\mathrm{MVol}(\mathbf{F}) = \mathrm{MVol}(\mathrm{Newt}(f_1), \ldots, \mathrm{Newt}(f_m))$.

Corollary 2.4.8. *Let $X \subset \mathbb{C}^n$ be an affine variety defined by polynomials $(f_1, \ldots, f_m)$ with strongly admissible support. For generic $u \in \mathbb{R}^n$, let $\mathbf{L}_{\mathrm{F}}$ be the Lagrange system of $\mathbf{F} = (\|x - u\|_2^2, f_1 \ldots, f_m)$ corresponding to the Euclidean distance optimization problem (ED) and $\mathbf{L}_{F_i}$ the Lagrange system of $\mathbf{F}_i = (\langle u, x \rangle, f_1, \ldots, f_m)$ corresponding to the ith sectional optimization problem (SO). Assume that H_∞ is not contained in the dual variety of $\overline{X}$. Then*

$$\mathrm{MVol}(\mathbf{L}_{\mathrm{F}}) = \sum_{i=0}^{n-1} \mathrm{MVol}(\mathbf{L}_i).$$

Observe that by the results in [MRWW23], we can think of sectional degrees as the affine analogue of polar degrees. With this in mind and the aforementioned results, we have the following conjecture.

Conjecture 2.4.9. *Let $X \subseteq \mathbb{C}^n$ be an irreducible, affine variety and $\overline{X} \subset \mathbb{P}^n$ its projective closure. If $\overline{X}$ intersects Q_∞ transversely then the ED degree of X is equal to $s_0(X) + \ldots + s_{n-1}(X)$.*

To provide one piece of evidence for Conjecture 2.4.9 we give an example where Conjecture 2.4.9 is true but Theorem 2.4.4 gives a strict upper bound on the ED degree.

Example 2.4.10. Consider the affine variety $X = \mathcal{V}(x_1^2 - x_2) \subset \mathbb{R}^2$. We can directly compute the ED degree of X to be three. In this case, X has two sectional degrees: $s_0(X) = 1$ and $s_1(X) = 2$. It is then clear that Conjecture 2.4.9 holds in this case.

Conversely, we can consider the polar degrees of $\overline{X} = \mathcal{V}(x_1^2 - x_2 x_0)$. Here, the conormal variety of $\overline{X}$ is defined as the common zero set of:

$$x_2 y_2 - x_0 y_0,\ y_1^2 - 4 y_2 y_0,\ x_0 y_1 + 2 x_1 y_2,$$
$$x_2 y_1 + 2 x_1 y_0,\ x_1 y_1 + 2 x_0 y_0,\ x_1^2 - x_2 x_0.$$

With this, one can directly compute that $\delta_0(\overline{X}) = 2$ and $\delta_1(\overline{X}) = 2$. This provides an example where the sum of the polar degrees of $\overline{X}$ is a strict upper bound on the ED degree of X but the sum of the sectional degrees is exact.

2.5 Homogeneous equations for critical points

In this section we define homogeneous critical point equations for the optimization problem (POP). We give two different sets of critical point equations for (POP). On the one hand, in [NR09] critical points are characterized as an intersection of the vanishing locus of homogeneous equations $\{\widetilde{f}_1 = \cdots = \widetilde{f}_m = 0\}$ with a projective determinantal variety W. We generalize this approach by replacing projective space with an appropriate toric variety X. On the other hand we homogenise the Lagrange equations $\mathbf{L}_{\mathrm{F}} = (f_1, \ldots, f_m, \ell_1, \ldots, \ell_n)$ in the Cox ring of a toric variety, $\mathbb{P}(\mathcal{E})$, which we introduce now. We show both approaches define the desired critical point equations in Lemma 2.5.8 with (2.9) concerning the former approach and (2.10) the latter.

2.5.1 Toric projective bundles

We now describe the toric structure on the projectivization of a direct sum of line bundles on toric varieties. Let X be a complete toric variety given by a fan Σ and let $\mathcal{E} = \mathcal{L}_0 \oplus \ldots \oplus \mathcal{L}_m$ be a fully decomposible vector bundle on X. In this subsection we will describe the fan of the total space of the projectivization $\mathbb{P}(\mathcal{E})$. We will start with a lemma:

Lemma 2.5.1. *Let X be a toric variety and let $\mathcal{E} = \mathcal{L}_0 \oplus \ldots \oplus \mathcal{L}_m$ be a vector bundle which is a direct sum of line bundles. The total spaces of $\mathcal{E}$ and $\mathbb{P}(\mathcal{E})$ have the structure of fibered toric varieties. That is, the natural projection to X is a torus equivariant morphism.*

Proof. For every line bundle $\mathcal{L}_i$, there exist torus invariant divisor D_i such that $\mathcal{L}_i = \mathcal{O}(D_i)$. Therefore, each line bundle $\mathcal{L}_i$ on X could be equipped with an equivariant structure, i.e. the action of T on the total space of $\mathcal{L}_i$, which makes the projection map equivariant.

Now, fixing an equivariant structure on each of line bundles $\mathcal{L}_i$, we obtain a T-action on the total space of $\mathcal{E}$. Finally we extend the T-action on $\mathcal{E}$ to the action of $T \times (\mathbb{C}^*)^{m+1}$ by making the second component act fiberwise in a natural way. This action is faithful and has open-dense orbit in the total space of $\mathcal{E}$.

Moreover, the action of $T \times (\mathbb{C}^*)^{m+1}$ on $\mathcal{E}$ descends to an action on $\mathbb{P}(\mathcal{E})$. The latter action has a one-dimensional kernel given by the diagonal subtorus in $(\mathbb{C}^*)^{m+1}$. Hence $\mathbb{P}(\mathcal{E})$ has the structure of a toric variety with respect to the factor torus

$$T \times \left((\mathbb{C}^*)^{m+1} / \mathbb{C}^* \cdot (1, \ldots, 1)\right). \qquad \square$$

Remark 2.5.2. Note that the divisor D_i is defined up to addition of the principal divisor $div(u)$ of character $u \in \mathbb{Z}^n$ or equivalently, any two equivariant structures on $\mathcal{L}_i$ differ by the action of the character of T. However, the toric variety structures defined by different choices of D_i are isomorphic (as toric varieties).

We conclude by describing the defining fan of the projectivized total space $\mathbb{P}(\mathcal{E})$, when the defining line bundles of $\mathcal{E}$ are torus equivariant. More precisely, we denote for $0 \leq j \leq m$ by D_j a torus invariant divisor such that $\mathcal{L}_i = \mathcal{O}(D_i)$. Each divisor D_j defines a conewise-linear function

$$\psi_j \colon \mathbb{R}^n = |\Sigma| \to \mathbb{R}.$$

Let $\Psi = (\psi_0, \ldots, \psi_m) \colon |\Sigma| \to \mathbb{R}^{m+1}$ be the corresponding piecewise linear map.

Let us denote by $\widetilde{\Sigma} \subset \mathbb{R}^n \times \mathbb{R}^{m+1}$ a fan obtained as the graph of the function Ψ. That is, $\widetilde{\Sigma}$ consists of cones $\widetilde{\sigma}$ where

$$\widetilde{\sigma} = \{(y, \Psi(y)) \,|\, y \in \sigma\}, \text{ for } \sigma \in \Sigma.$$

We now abuse notation and denote by $\mathbb{R}^n_{\geq 0}$ the fan whose cones are the boundary components of the positive orthant in $\mathbb{R}^n$. Then the fan defining the total space of $\mathcal{E}$ consists of cones

$$\widetilde{\sigma} + \tau \text{ for } \sigma \in \Sigma, \tau \in \mathbb{R}^n_{\geq 0} \times \{0\}.$$

Similarly, the fan F defining the total space of $\mathcal{E}$ with the zero section removed is given by

$$\widetilde{\sigma} + \tau \text{ for } \sigma \in \Sigma, \tau \in \partial\mathbb{R}^n_{\geq 0} \times \{0\},$$

where $\partial\mathbb{R}^n_{\geq 0} = \mathbb{R}^n_{\geq 0} \setminus \{\mathbb{R}^n_{>0}\}$ is the fan consisting of all cones in $\mathbb{R}^n_{\geq 0}$, except for the one of dimension n. Finally, let $\mathcal{S}_0 \subset \mathcal{E}$ be the image of the zero section of $\mathcal{E}$. Clearly $\mathcal{S}_0$ is a torus invariant subset of $\mathcal{E}$ and thus the natural projection of $\mathcal{E} \setminus \mathcal{S}_0 \to \mathbb{P}(\mathcal{E})$ is a toric morphism. On the level of fans, consider the projection

$$\tau \colon \mathbb{R}^n \times \mathbb{R}^{m+1} \longrightarrow \mathbb{R}^n \times \left(\mathbb{R}^{m+1}/\mathbb{R} \cdot (1, \ldots, 1)\right) \cong \mathbb{R}^n \times \mathbb{R}^m.$$

The underlying fan of the projectivization $\mathbb{P}(\mathcal{E})$, is the image of F under τ.

Now let S be the Cox ring of X, and let $S_{\mathcal{E}}$ be the Cox ring of $\mathbb{P}(\mathcal{E})$. By the above discussion, each ray of $\widetilde{\Sigma}$ is either of the form $\widetilde{\rho}$, where ρ is a ray of Σ, or of the form $\{0\} \times e_i$, $i = 1, \ldots, m+1$. This splits the generators of $S_{\mathcal{E}}$ over $\mathbb{C}$ into two groups. The first group of generators $x_{\widetilde{\rho}}$ is bijective to the generators x_ρ of S. We denote members of the second group by $\lambda_i = x_{\{0\} \times e_{i+1}}$ and obtain the following proposition.

Proposition 2.5.3. *The Cox ring $S_{\mathcal{E}}$ is isomorphic to the free S-algebra $S[\lambda_0, \ldots, \lambda_m]$.*

Remark 2.5.4. In the following we are often in the situation that f is a global section of a torus invariant line bundle $\mathcal{O}_X(D)$ on X. Now $\widetilde{f}$ denotes an element of $S \subseteq S_{\mathcal{E}}$. At the same time f can be identified with a section of the bundle $\pi^*\mathcal{O}_X(D)$ on $\mathbb{P}(\mathcal{E})$, where $\pi \colon \mathbb{P}(\mathcal{E}) \longrightarrow X$ is the natural projection. When homogenising, this gives rise to another element $\widetilde{f} \in S_{\mathcal{E}}$. Direct computation shows that there is no need for disambiguation, since both expressions are equal.

2.5.2 Constructing critical point equations in Cox rings

We start by fixing some notation and definitions: For the rest of this section let $\mathbf{F} = (f_0, \ldots, f_m)$ be a generic sparse system of polynomials in $\mathbb{C}[X_1, \ldots, X_n]$ with admissible support $\mathcal{A} = (\mathcal{A}_0, \ldots, \mathcal{A}_m)$. Furthermore X denotes a toric variety that is appropriate for $\mathcal{A}$, with fan Σ.

Remark 2.5.5. Note that, since Σ contains the positive orthant $\mathbb{R}^n_{\geq 0}$ as a cone, there is a distinct copy of the affine space $\mathbb{C}^n$ contained in X. For clarity, in this section, we denote the variables in the coordinate ring $\mathbb{C}[X_1, \ldots, X_n]$ of $\mathbb{C}^n$ in capital letters, while the generators of the Cox ring S of X are in lower case. By slight abuse of notation, we will denote the element x_{e_j} in S by x_j for each $j = 1 \ldots, n$.

For every $i = 0, \ldots, m$ let $\mathcal{L}_i = \mathcal{O}(-D_{f_i})$ denote the dual line bundle associated to D_{f_i}. Here D_{f_i} is the torus invariant Weyl divisor on X, corresponding to the Newton polytope $\mathrm{Newt}(f_i)$:

$$D_{f_i} = \sum_{\rho \in \Sigma(1)} a_{\rho,i} D_\rho, \quad \text{where} \quad a_{\rho,i} = -\min\{\langle m, \rho \rangle \ : \ m \in \mathrm{Newt}(f_i)\}. \tag{2.8}$$

We denote by $\mathcal{E}$ the vector bundle $\mathcal{E} = \mathcal{L}_0 \oplus \cdots \oplus \mathcal{L}_m$ with projectivized total space

$$\mathbb{P}(\mathcal{E}) = \{(x, [\lambda]) \mid \ x \in X, \ [\lambda] \in \mathbb{P}\left(\mathcal{E}(x)\right)\}.$$

The rest of this section is devoted to giving two different, but related, systems of homogeneous critical point equations for (POP), one in the Cox ring S of X, and one in the Cox ring $S_{\mathcal{E}}$ of $\mathbb{P}(\mathcal{E})$. On the one hand, critical points are characterized by the vanishing of the Lagrange system $\mathbf{L}_{\mathrm{F}}$. It describes the intersection of the incidence variety

$$Z^\circ := \{(x,[\lambda]) \in \mathbb{C}^n \times \mathbb{P}^n : \ (\nabla f_0 \mid \cdots \mid \nabla f_m)\,\lambda = \underline{0}\}$$

with the vanishing locus of $f_1,\ldots,f_m$. On the other hand, critical points are characterized by the Jacobian $(\nabla f_0 \mid \cdots \mid \nabla f_m)$ dropping rank. They form the intersection $V^\circ \cap W^\circ$, where $V^\circ := \mathcal{V}(f_1,\ldots,f_m) \subseteq \mathbb{C}^n$, and W° is the determinantal variety

$$W^\circ := \{x \in \mathbb{C}^n \ : \ \operatorname{rank}(\nabla f_0 \mid \cdots \mid \nabla f_m) \leq m\}.$$

We proceed by giving homogeneous equations for V°, W° and Z° in S and $S_{\mathcal{E}}$ respectively. Every polynomial f_i is a global section of the line bundle $\mathcal{O}_X(D_{f_i})$, and its homogeneous form can be written as

$$\widetilde{f}_i = \sum_{m \in \mathrm{Newt}(f_i)\cap\mathbb{Z}^n} c_{m,i} \prod_{\rho\in\Sigma(1)} x_\rho^{\langle m,\rho\rangle + a_{\rho,i}}.$$

Here we homogenize f_i as in (2.3) in Section 2.2.2. In particular, $\widetilde{f}_i$ is defined by our choice of line bundle $\mathcal{O}_X(D_{f_i})$.

We denote by V the closure of $V^\circ = \mathcal{V}(f_1,\ldots,f_m)$ in X. Observe that by genericity of $\mathbf{F}$, V is equal to the vanishing locus of the homogeneous equations $V = \mathcal{V}\left(\widetilde{f}_1,\ldots\widetilde{f}_m\right)$.

From now on M denotes a homogeneous version of the Jacobian matrix:

$$M = \left(\widetilde{\nabla}\widetilde{f}_0 \mid \cdots \mid \widetilde{\nabla}\widetilde{f}_m\right).$$

Here $\widetilde{\nabla}$ denotes the vector $(\frac{\partial}{\partial x_1},\ldots,\frac{\partial}{\partial x_n})^T$. We use the notation $\widetilde{\nabla}$ instead of ∇ to indicate that we differentiate with respect to coordinates in the Cox ring. So M has columns $\left(\frac{\partial}{\partial x_1}\widetilde{f}_i,\ldots,\frac{\partial}{\partial x_n}\widetilde{f}_i\right)^T$. We define

$$W := \{x \in X \ : \ \operatorname{rank} M(x) \leq m\}$$

to be the vanishing locus of the maximal minors of M, and furthermore we let

$$Z := \{(x,[\lambda]) \in \mathbb{P}(\mathcal{E}) \ : \ M(x)\lambda = 0\}$$

be the associated incidence variety, contained in the projectivized total space $\mathbb{P}(\mathcal{E})$.

The rest of this Section is devoted to proving Lemma 2.5.8. It shows that the homogeneous critical point equations agree with the affine ones when restricted to affine space $\mathbb{C}^n$.

We need an observation about differentiating homogeneous polynomials. Let D be a torus invariant Weyl divisor on X (or on $\mathbb{P}(\mathcal{E})$), and f a global section of $\mathcal{O}_X(D)$. Observe that for $j = 1,\ldots,n$ the Newton polytope of the differential $X_j\frac{\partial}{\partial X_j}f$ is contained in the rational polytope $e_j + \partial_j\mathrm{Newt}(f)$, and in particular $\frac{\partial}{\partial X_j}f$ is a global section of the sheaf $\frac{1}{X_j}\mathcal{O}_X(D - D_{e_j})$ (or of the sheaf $\frac{1}{X_j}\mathcal{O}_{\mathbb{P}(\mathcal{E})}(D - D_{\widetilde{e}_j})$). Direct computation shows the following proposition.

Proposition 2.5.6. *Homogenization and differentiation commute:* $\widetilde{\frac{\partial}{\partial X_j}f} = \frac{\partial}{\partial x_j}\widetilde{f}$.

We denote by $\widetilde{\Phi}_F$ the homogenization of the Lagrangian Φ, living in the Cox ring $S_{\mathcal{E}}$. This makes sense since $\mathbb{P}(\mathcal{E})$ defines a global section of the sheaf $\mathcal{O}_{\mathbb{P}(\mathcal{E})}(D_{\Phi_F})$, associated to the Cayley polytope $\mathrm{Newt}(\Phi_F) = \mathrm{Cay}(\mathrm{Newt}(f_0), \ldots, \mathrm{Newt}(f_m))$. By the above Proposition, each of the defining equations

$$0 = (M(x)\lambda)_j = \lambda_0 \frac{\partial}{\partial x_j}\widetilde{f}_0 + \cdots + \lambda_m \frac{\partial}{\partial x_j}\widetilde{f}_m = \frac{\partial}{\partial x_j}\widetilde{\Phi}_F(\lambda, x)$$

of Z is equal to the homogenization $\widetilde{\ell_j}$ of $\ell_j = \widetilde{\frac{\partial}{\partial X_j}\Phi_F}(\lambda, x)$. Here $\widetilde{\ell_j}$ is considered a global section of the sheaf $\mathcal{O}_{\mathbb{P}(\mathcal{E})}(D_{\Phi_F} - D_{\widetilde{e}_j})$.

We denote by $\widetilde{\mathbf{L}}_F = (\widetilde{f}_1, \ldots, \widetilde{f}_m, \widetilde{\ell}_1, \ldots, \widetilde{\ell}_n)$ the homogenized Lagrange system. On the one hand, we observed above that Z is equal to the vanishing locus of $\widetilde{\ell}_1, \ldots, \widetilde{\ell}_n$. On the other hand, the vanishing locus of $\widetilde{f}_1, \ldots, \widetilde{f}_m$ in $\mathbb{P}(\mathcal{E})$ is the preimage $\pi^{-1}(V)$ of the vanishing locus V of $\widetilde{f}_1, \ldots, \widetilde{f}_m$ in X. We obtain the following Proposition.

Proposition 2.5.7. *The vanishing locus of $\widetilde{\mathbf{L}}_F$ in $\mathbb{P}(\mathcal{E})$ is the intersection $Z \cap \pi^{-1}(V)$.*

The following lemma shows that the homogeneous critical point equations introduced in this chapter restrict, on $\mathbb{C}^n$, to the expected affine critical point equations.

Lemma 2.5.8. *The following three equalities hold:*

$$V \cap \mathbb{C}^n = V^\circ, \qquad W \cap \mathbb{C}^n = W^\circ \tag{2.9}$$
$$Z \cap \pi^{-1}(\mathbb{C}^n) = Z^\circ \tag{2.10}$$

where the intersection in (2.9) is on X and the intersection in (2.10) is on $\mathbb{P}(\mathcal{E})$.

Proof. The first of the equalities is clear, by the definition of V as the closure of V°. To see the second equality, we prove that the entries of M are homogenizations of the entries of the Jacobian $(\nabla f_0, \ldots, \nabla f_m)$. This is a direct consequence of Proposition 2.5.6, since for every $i = 0, \ldots, m$ and $j = 1, \ldots, n$ it holds $\widetilde{\frac{\partial}{\partial X_j} f_i} = \frac{\partial}{\partial x_j}\widetilde{f}_i$. The third equality is analogous, since homogenising the defining equations $\ell_1, \ldots, \ell_n$ of Z° yields the defining equations $\widetilde{\ell}_1, \ldots, \widetilde{\ell}_n$ of Z. □

We close this section with the following generalization of the Euler equation.

Proposition 2.5.9. *Let $D = \sum_{\rho \in \Sigma(1)} a_\rho D_\rho$ be a torus invariant Weyl divisor, f a global section of $\mathcal{O}_X(D)$ and $\tau \in \Sigma(1)$ a ray. Then the generalized Euler equation*

$$-x_\tau \frac{\partial}{\partial x_\tau}\widetilde{f} + \tau_1 x_1 \frac{\partial}{\partial x_1}\widetilde{f} + \cdots + \tau_n x_n \frac{\partial}{\partial x_n}\widetilde{f} = -a_\tau \widetilde{f} \tag{2.11}$$

holds for the homogenization $\widetilde{f} \in S$ of f.

Proof. Equation (2.3) reads $\widetilde{f} = \sum_{m\in\mathbb{Z}^n} c_m \prod_{\rho\in\Sigma(1)} x_\rho^{\langle m,\rho\rangle+a_\rho}$ and we have

$$\begin{aligned}
&- x_\tau \frac{\partial}{\partial x_\tau}\widetilde{f} + \tau_1 x_1 \frac{\partial}{\partial x_1}\widetilde{f} + \cdots + \tau_n x_n \frac{\partial}{\partial x_n}\widetilde{f} \\
&= \sum_{m\in\mathbb{Z}^n} c_m \left(-x_\tau \frac{\partial}{\partial x_\tau} + \tau_1 x_1 \frac{\partial}{\partial x_1} + \cdots + x_n \frac{\partial}{\partial x_n}\right) \prod_{\rho\in\Sigma(1)} x_\rho^{\langle m,\rho\rangle+a_\rho} \\
&= \sum_{m\in\mathbb{Z}^n} c_m(-\langle m,\tau\rangle - a_\tau + m_1 + a_{e_1} + \cdots + m_n + a_{e_n}) \prod_{\rho\in\Sigma(1)} x_\rho^{\langle m,\rho\rangle+a_\rho} \\
&= \sum_{m\in\mathbb{Z}^n} c_m(-a_\tau) \prod_{\rho\in\Sigma(1)} x_\rho^{\langle m,\rho\rangle+a_\rho} = -a_\tau \widetilde{f}.
\end{aligned}$$

□

2.6 Computing the number of critical points

In this section we finally prove our main results, Theorem 2.3.6, Theorem 2.3.7 and Theorem 2.3.11, relying on the results from Section 2.5. There we characterized critical points of (POP) in two ways. On the one hand, as an intersection $V \cap W$ in X. This is done in equation (2.9) of Lemma 2.5.8. On the other hand by means of homogenized Lagrange equations $\widetilde{\mathbf{L}}_F$ in the Cox ring of $\mathbb{P}(\mathcal{E})$. This is done in in equation (2.10) of Lemma 2.5.8. In this section we show that all intersections are transversal and happen in $\mathbb{C}^n$. This characterises the number of critical points as products of cohomology classes. In the case of Theorem 2.3.6 this product is a mixed volume. The proof of Theorem 2.3.11 rests on a characterization of $[W]$ as a Porteous' class. The assumptions of our main theorems will be used in the following places: we will use the assumption that $\mathcal{A}$ be admissible in the proof of Proposition 2.6.2 below. For the proof of Proposition 2.6.1 below we need that the closure $\mathcal{V}$ of the constraint locus $\mathcal{V}(f_1,\ldots,f_m)$ is smooth. This is guaranteed by stronger assumption that $\mathcal{A}$ is strongly admissible in the proof of Theorem 2.3.6 and Theorem 2.3.7. For Theorem 2.3.11 we assume X to be smooth in order to employ Porteus' formula.

2.6.1 Preliminary results

We start by proving some technical statements that are needed for the desired transversality results. For the rest of the section we again fix the assumptions from Section 2.5.2, and furthermore assume that V does not intersect the singular locus of X. This is in practice guaranteed since we always assume that X is appropriate for the monomial support $\mathcal{A}$ of F. Now V is the vanishing locus of generic sections of basepoint free line bundles on the smooth locus of X. It follows from Bertini's Theorem, that V is also smooth, which is the motivation for Definition 2.3.4 and Definition 2.3.2.

For the next proposition we need the following notion. Similar to the projection $\mathbb{C}^{n+1}\backslash\{0\} \to \mathbb{P}^n$, there exists the open subset $U_\Sigma \subseteq \mathbb{C}^{\Sigma(1)}$ with a projection $\tau : U_\Sigma \to X$. For a subvariety V of X, we define the cone $C(Y)$ over Y to be the closure of the preimage $\tau^{-1}(Y)$ in $\mathbb{C}^{\Sigma(1)}$. The intersection $C(Y) \cap U_\Sigma$ forms a principal bundle over Y. In particular, $C(Y) \cap U_\Sigma$ is smooth if Y is smooth.

Proposition 2.6.1. *The matrices* $M = \left(\widetilde{\nabla}\widetilde{f}_0,\ldots,\widetilde{\nabla}\widetilde{f}_m\right)$ *and* $\left(\widetilde{\nabla}\widetilde{f}_1,\ldots,\widetilde{\nabla}\widetilde{f}_m\right)$ *have full ranks* $m+1$ *and* m *everywhere on* $\mathcal{V}\left(\widetilde{f}_0,\ldots,\widetilde{f}_m\right)$ *and* $\mathcal{V}\left(\widetilde{f}_1,\ldots,\widetilde{f}_m\right)$ *respectively.*

Proof. The proof for the second matrix is analogous, so we only present the proof for

$$M = \left(\widetilde{\nabla}\widetilde{f}_0,\ldots,\widetilde{\nabla}\widetilde{f}_m\right).$$

Let $x \in \mathcal{V}\left(\widetilde{f}_0, \ldots, \widetilde{f}_m\right)$ be arbitrary and $\sigma \in \Sigma$ the unique cone such that x is contained in the torus orbit $O(\sigma)$. Let $\widetilde{M}$ denote the matrix with rows

$$\left(\frac{\partial}{\partial x_\rho}\widetilde{f}_0, \ldots, \frac{\partial}{\partial x_\rho}\widetilde{f}_m\right) \tag{2.12}$$

for each ray ρ in $\Sigma(1)$. The left kernel of $\widetilde{M}$ is the tangent space of the cone $C\left(\mathcal{V}\left(\widetilde{f}_0, \ldots, \widetilde{f}_m\right)\right)$ in $\mathbb{C}^{\Sigma(1)}$. The Jacobian $\widetilde{M}_\sigma$ of the cone over the variety $O(\sigma) \cap \mathcal{V}\left(\widetilde{f}_0, \ldots, \widetilde{f}_m\right)$ is a submatrix of $\widetilde{M}$. Its rows correspond to those rays ρ that are not contained in σ. By our assumption at the beginning of this section, V is disjoint from the singular locus of X, and we can apply Bertini's Theorem to show that $\mathcal{V}\left(\widetilde{f}_0, \ldots, \widetilde{f}_m\right)$ is a smooth variety. Furthermore, the intersection $O(\sigma) \cap \mathcal{V}\left(\widetilde{f}_0, \ldots, \widetilde{f}_m\right)$ is transversal by [Kho78], so $\widetilde{M}_\sigma$ is of full rank $m+1$ at x. We now finish the proof by showing that the row span of $\widetilde{M}_\sigma$ is contained in the row span of M. Let ρ be any ray that is not contained in σ. To show that the corresponding row (2.12) of $\widetilde{M}_\sigma$ is contained in the row span of M, we apply Proposition 2.5.9 to all functions $\widetilde{f}_0, \ldots, \widetilde{f}_m$. The right side of equation (2.11) vanishes, and we obtain

$$\begin{pmatrix} \rho_1 x_1 \\ \vdots \\ \rho_n x_n \end{pmatrix}^T M = x_\rho \begin{pmatrix} \frac{\partial}{\partial x_\rho}\widetilde{f}_0 \\ \vdots \\ \frac{\partial}{\partial x_\rho}\widetilde{f}_m \end{pmatrix}^T . \qquad \square$$

Proposition 2.6.2. *The gradient* $\widetilde{\nabla}\widetilde{f}_0 = \left(\frac{\partial}{\partial x_j}\widetilde{f}_0\right)_{j=1,\ldots,n}$ *does not vanish on any torus orbit.*

Proof. Towards a contradiction we assume that there exists a cone $\sigma \in \Sigma$ such that for every $j = 1, \ldots, n$ the polynomial $\frac{\partial}{\partial x_j}\widetilde{f}_0$ vanishes on the associated torus orbit $O(\sigma)$ of X. We denote by $\frac{\partial}{\partial x_j}\widetilde{f}_0\Big|_{O(\sigma)}$ the restriction of $\frac{\partial}{\partial x_j}\widetilde{f}_0$ to the cone over $O(\sigma)$. It is obtained by substituting all variables x_ρ with zero, where ρ is contained in σ.

Now consider the face $\mathrm{Newt}(f_0)^\sigma$ of $\mathrm{Newt}(f_0)$ exposed by σ. For every lattice point m of $\mathrm{Newt}(f_0)^\sigma$ the monomial

$$\frac{\partial}{\partial x_j} \prod_{\rho \in \Sigma(1)} x_\rho^{\langle m, \rho \rangle + a_{\rho,0}}, \quad a_{\rho,0} = -\min\{\langle m, \rho \rangle \ : \ m \in \mathrm{Newt}(f_0)\}$$

of $\frac{\partial}{\partial x_j}\widetilde{f}_0$ only vanishes on $O(\sigma)$ if $m_j = 0$. In particular, the face $\mathrm{Newt}(f_0)^\sigma$ can only contain the single element $\underline{0}$. By assumption 2.3.2 on X, $\underline{0}$ is a smooth vertex of $\mathrm{Newt}(f_0)$, and dual to the cone $\mathbb{R}^n_{\geq 0}$. Since σ reveals the vertex $\underline{0}$, it has to intersect the interior of the positive orthant $\mathbb{R}^n_{\geq 0}$ and in fact both cones are equal. This leaves us with the case where the torus orbit is $\{\underline{0}\}$. But the gradient $\widetilde{\nabla}\widetilde{f}_0$ does not vanish uniformly at $\underline{0}$. $\square$

Let again V and W denote the varieties from Subsection 2.5.2.

Proposition 2.6.3. *The variety* $V \cap W$ *is of dimension* 0.

Proof. Towards a contradiction we assume that there exists a torus orbit $O(\sigma)$, and a curve C such that C is contained in the intersection $W \cap V \cap O(\sigma)$. We denote by $\widetilde{f}_0\Big|_{O(\sigma)}$ the restriction of $\widetilde{f}_0$

to $O(\sigma)$. It is obtained by substituting all variables x_ρ with zero, where ρ is contained in σ. We now distinguish two cases: either $\widetilde{f}_0\Big|_{O(\sigma)}$ vanishes somewhere on $O(\sigma)$, or it is a scalar multiple of a monomial. In the first case $\widetilde{f}_0\Big|_{O(\sigma)}$ vanishes on C by genericity. In particular, the matrix M drops rank somewhere on the vanishing locus $\mathcal{V}\left(\widetilde{f}_0,\dots\widetilde{f}_m\right)$, contradicting Proposition 2.6.1.

In the second case we now derive a contradiction from Proposition 2.6.1 by showing that the matrix $\left(\widetilde{\nabla}\widetilde{f}_1,\dots,\widetilde{\nabla},\widetilde{f}_m\right)$ drops rank somewhere on C. Suppose $\widetilde{f}_0\Big|_{O(\sigma)}$ is a monomial. Then each restriction $\frac{\partial}{\partial x_j}\widetilde{f}_0\Big|_{O(\sigma)}$ is either a monomial or zero, and by Proposition 2.6.2 there is an index $l=1,\dots,n$ such that $\frac{\partial}{\partial x_l}\widetilde{f}_0\Big|_{O(\sigma)}$ is not zero. Without loss of generality we assume l to be 1. Consider the following matrix, M^*, obtained by subtracting for each $j=2,\dots,n$ from the $j-$th row of M the multiple

$$\frac{\frac{\partial}{\partial x_j}\widetilde{f}_0}{\frac{\partial}{\partial x_1}\widetilde{f}_0}\left(\frac{\partial}{\partial x_1}\widetilde{f}_0,\dots,\frac{\partial}{\partial x_1}\widetilde{f}_m\right)$$

of the first row, eliminating the first entry in the process:

$$M^*=\begin{bmatrix}\frac{\partial}{\partial x_1}\widetilde{f}_0 & \frac{\partial}{\partial x_1}\widetilde{f}_1\cdots\frac{\partial}{\partial x_1}\widetilde{f}_m\\ 0 & \\ \vdots & A\\ 0 & \end{bmatrix}.$$

Since M drops rank everywhere on C and $\frac{\partial}{\partial x_1}\widetilde{f}_0$ is not identically zero, A also drops rank on C. Let $\mu=(\mu_1,\dots,\mu_n)^T$ be a vector of rational functions on $O(\sigma)$ satisfying

$$A\mu=0$$

everywhere on C. Since the expression $\mu_1\frac{\partial}{\partial x_1}\widetilde{f}_1+\cdots+\mu_n\frac{\partial}{\partial x_n}\widetilde{f}_1$ is not a monomial on $O(\sigma)$, it vanishes at some point x in C by genericity. This shows that $\mu(x)$ is in the right kernel of $\left(\widetilde{\nabla}\widetilde{f}_1(x),\dots,\widetilde{\nabla}\widetilde{f}_m(x)\right)$, finishing the proof. □

Lemma 2.6.4. *The intersection $Z\cap\mathcal{V}(\widetilde{f}_1,\dots,\widetilde{f}_m)$ is transversal and contained in the big torus $(\mathbb{C}^*)^{n+m}$ in the toric variety $\mathbb{P}(\mathcal{E})$.*

Proof. The image of $Z\cap\mathcal{V}(\widetilde{f}_1,\dots,\widetilde{f}_m)$ under the natural projection $\pi:\mathbb{P}(\mathcal{E})\longrightarrow X$ is $W\cap\mathcal{V}(\widetilde{f}_1,\dots,\widetilde{f}_m)$, which by Proposition 2.6.3 is finite. In fact, we prove below that π bijectively identifies both sets. In particular, the n defining equations of Z, given by $M(x)\lambda=0$, form a complete intersection when restricted to $\mathcal{V}(\widetilde{f}_1,\dots,\widetilde{f}_m)$.

To inductively apply Bertini to the equations $(M(x)\lambda)_j=0$ we now show that, for varying coefficients of f_0, $Z\cap\mathcal{V}(\widetilde{f}_1,\dots,\widetilde{f}_m)$ defines a basepoint free family of varieties on the vanishing locus $\mathcal{V}(\widetilde{f}_1,\dots,\widetilde{f}_m)$ in $\mathbb{P}(\mathcal{E})$. To do this, we fix any element x of V and show that Z does not have a fixed point in the fiber $\pi^{-1}(x)$. By Proposition 2.6.1 the last m columns $\left(\widetilde{\nabla}\widetilde{f}_1,\dots,\widetilde{\nabla}\widetilde{f}_m\right)$ of M are linearly independent. In particular, varying the first column $\widetilde{\nabla}\widetilde{f}_0$ changes the unique solution $[\lambda]$ to $M(x)\lambda=0$. It now suffices to see that the gradient $\widetilde{\nabla}\widetilde{f}_0$ does not vanish uniformly at x, which by Proposition 2.6.2 is true for generic coefficients of f_0. To finally see that $Z\cap\mathcal{V}(\widetilde{f}_1,\dots,\widetilde{f}_m)$ is

contained in the big torus orbit, we apply the same Bertini type argument to show transversality of the intersection $O \cap \mathcal{V}\left(\widetilde{f}_1, \ldots, \widetilde{f}_m\right) \cap Z$. Here O denotes any torus orbit on $\mathbb{P}(\mathcal{E})$. For dimensional reasons only for the big torus this intersection can be non empty. □

2.6.2 The proof of Theorem 2.3.6

The idea behind the proof of Theorem 2.3.6 is to study the system of homogenized Lagrange equations $\widetilde{\mathbf{L}}_F = (\widetilde{f}_1, \ldots, \widetilde{f}_m, \widetilde{\ell}_1, \ldots, \widetilde{\ell}_n)$. We will show that it comprises global sections of $\mathbb{Q}$-Cartier divisors that intersect transversally and away from infinity. This expresses the number of solutions as a product of Chern classes, which is a mixed volume.

For the rest of this subsection we impose the assumptions of Theorem 2.3.6. Let Σ be the normal fan of the Minkowski sum of the polytopes $\mathrm{Newt}(f_0), \ldots, \mathrm{Newt}(f_n)$, and let X be the associated normal toric variety. Then the assumptions from the beginning of Subsection 2.5.2 are fulfilled, since X is appropriate for $\mathcal{A}$, and V does not intersect the singular locus of X.

Let again Φ_F denote the Lagrangian $\Phi_F(\lambda, x) = f_0 - \sum_{i=1}^m \lambda_i f_i$, and the partial differentials $\frac{\partial}{\partial x_j}\left(f_0 - \sum_{i=1}^m \lambda_i f_i\right)$ of Φ_F are denoted by ℓ_j. For each $j = 1, \ldots, n$, we defined the homogenization $\widetilde{\ell}_j$ of ℓ_j as a section of the divisor $D_{\Phi_F} - D_{\widetilde{e}_j}$ on $\mathbb{P}(\mathcal{E})$. We first prove that, up to isomorphy, this divisor is associated to the rational polytope $\partial_j \mathrm{Newt}(\Phi_F)$.

Lemma 2.6.5. *For all $j = 1, \ldots, n$, the divisor $D_{\Phi_F} - D_{\widetilde{e}_j}$ on $\mathbb{P}(\mathcal{E})$ is linearly equivalent to the divisor associated to the rational polytope $\partial_j \mathrm{Newt}(\Phi_F)$.*

Proof. We now prove that the divisor $D_{\Phi_F} - D_{\widetilde{e}_j}$ is associated to the polytope $e_j + \partial_j \mathrm{Newt}(\Phi_F)$. Note that $e_j + \partial_j \mathrm{Newt}(\Phi_F)$ is the intersection of $\mathrm{Newt}(\Phi_F)$ with the affine halfspace $\{x_j \geq 1\}$. We have to prove that the support function of $\mathrm{Newt}(\Phi_F) \cap \{x_j \geq 1\}$ takes the same value on all rays of $\Sigma_{\mathcal{E}}$, except for $\widetilde{e_j}$, where it differs by one. Let v be any element of $\mathbb{R}^{n+m}$. The value

$$-\min\{\langle w, v\rangle \ : \ w \in \mathrm{Newt}(\Phi_F)\}$$

of the support function of $\mathrm{Newt}(\Phi_F)$ on v can only differ if the face $\mathrm{Newt}(\Phi_F)^v$ is contained in the facet

$$\mathrm{Newt}(\Phi_F)^{\widetilde{e}_j} = \mathrm{Newt}(\Phi_F) \cap \{x_j = 0\}.$$

Note that a face of the form $\mathrm{Newt}(\Phi_F)^{\{0\}\times e_i}$ is equal to the Cayley polytope $\mathrm{Cay}(\mathcal{A}_0, \ldots, \mathcal{A}_{i-1}, \mathcal{A}_{i+1}, \ldots, \mathcal{A}_m)$, where we omit one of the constraints. In particular, it is always a facet, so we may restrict to rays of the form $\widetilde{\rho}$. Let now $\widetilde{\rho}$ be a ray such that $\mathrm{Newt}(\Phi_F)^{\widetilde{\rho}}$ is contained in $\mathrm{Newt}(\Phi_F)^{\widetilde{e}_j}$. By Proposition 2.6.6 below we have

$$\begin{aligned}
&\mathrm{Cay}\left(\mathrm{Newt}(f_0)^\rho, \ldots, \mathrm{Newt}(f_m)^\rho\right)\\
=&\,\mathrm{Newt}(\Phi_F)^{\widetilde{\rho}}\\
\subseteq&\,\mathrm{Newt}(\Phi_F)^{\widetilde{e_j}}\\
=&\,\mathrm{Cay}\left(\mathrm{Newt}(f_0)^{e_j}, \ldots, \mathrm{Newt}(f_m)^{e_j}\right),
\end{aligned}$$

implying $\mathrm{Newt}(f_i)^\rho \subseteq \mathrm{Newt}(f_i)^{e_j}$ for all $i = 0, \ldots, m$. We obtain

$$\left(\sum_{i=0}^m \mathrm{Newt}(f_i)\right)^\rho = \sum_{i=0}^m \mathrm{Newt}(f_i)^\rho \subseteq \sum_{i=0}^m \mathrm{Newt}(f_i)^{e_j} = \left(\sum_{i=0}^m \mathrm{Newt}(f_i)\right)^{e_j},$$

which is an inclusion of facets of the Minkowski sum $\sum_{i=0}^m \mathrm{Newt}(f_i)$, showing $\rho = e_j$. □

Proposition 2.6.6. *Let σ be a cone in the normal fan $\Sigma(P_1 + \cdots + P_n)$. Then the face of $\mathrm{Cay}(P_1, \ldots, P_n)$ exposed by $\widetilde{\sigma}$ is equal to the Cayley polytope of the faces $P_1^\sigma, \ldots, P_n^\sigma$:*

$$\mathrm{Cay}(P_1, \ldots, P_n)^{\widetilde{\sigma}} = \mathrm{Cay}(P_1^\sigma, \ldots, P_n^\sigma). \tag{2.13}$$

Proof. This can be done by direct computation. A different argument relies on Proposition 2.6.7 below. For this, denote by X_σ the closure of the torus orbit of X_Σ corresponding to $\sigma \in \Sigma$. By Proposition 2.6.7, the equation (2.13) is equivalent to the equality

$$\mathcal{O}_{\mathbb{P}(\mathcal{E})}(1)\big|_{X_\sigma} = \mathcal{O}_{\mathbb{P}(\mathcal{E}|_{X_\sigma})}(1). \qquad \square$$

Proposition 2.6.7. *Let X be a toric variety and $\mathcal{E} = \bigoplus \mathcal{L}_{P_i}$ be a direct sum of line bundles on X. Then the relative $\mathcal{O}(1)$ bundle of $\mathbb{P}(\mathcal{E})$ is represented by the Cayley polytope: $\mathrm{Cay}(P_1, \ldots, P_m)$*

Proof. The space of sections $H^0(\mathbb{P}(\mathcal{E}), \mathcal{O}(1))$ is canonically isomorphic to $H^0(X, \mathcal{E}) = \bigoplus_i H^0(X, \mathcal{L}_{P_i})$. Moreover, $H^0(X, \mathcal{L}_{P_i})$ is the weight space of $H^0(X, \mathcal{E})$ with respect to $(\mathbb{C}^*)^m$ torus acting fiberwise on $\mathbb{P}(\mathcal{E})$ corresponding to i-th basis vector of $\mathbb{Z}^m$. Since the weights of the base torus acting on $H^0(X, \mathcal{L}_{P_i})$ is given by the lattice points of P_i, we obtain the result. $\square$

Before proving Theorem 2.3.6 we need to show a statement about the intersection of $\mathbb{Q}$-Cartier divisors.

Lemma 2.6.8 (Generic intersection of $\mathbb{Q}$-cartier divisors)**.** *Let X be a normal, proper variety of dimension n with Weyl divisors $D_1, \ldots, D_n$, and let k be an integer such that $\mathcal{O}(kD_i)$ is a line-bundle for each $i = 1, \ldots, n$. Let $\widetilde{f}_i$ be a global section of $\mathcal{O}(D_i)$ for $i = 1, \ldots, n$ such that $\mathcal{V}(\widetilde{f}_1, \ldots, \widetilde{f}_n)$ is a zero dimensional smooth scheme contained in the smooth locus of X. Then*

$$k^n \# \mathcal{V}(\widetilde{f}_1, \ldots, \widetilde{f}_n) = c_1(\mathcal{O}(kD_1)) \cdots c_1(\mathcal{O}(kD_n)).$$

Proof. The length of the zero-dimensional scheme $\mathcal{V}((\widetilde{f}_1)^k, \ldots, (\widetilde{f}_n)^k)$ is equal to the product $c_1(\mathcal{O}(kD_1)) \cdots c_1(\mathcal{O}(kD_n))$ of Chern classes. On the other hand, since $\widetilde{f}_1, \ldots, \widetilde{f}_n$ intersect transversally, each isolated point of $\mathcal{V}((\widetilde{f}_1)^k, \ldots, (\widetilde{f}_n)^k)$ is isomorphic to the scheme $\operatorname{Spec} \mathbb{C}[x_1, \ldots, x_n]/\langle x_1^k, \ldots, x_n^k \rangle$. In particular we have

$$k^n \cdot \mathrm{length}(\mathcal{V}(\widetilde{f}_1, \ldots, \widetilde{f}_m)) = \mathrm{length}(\mathcal{V}((\widetilde{f}_1)^k, \ldots, (\widetilde{f}_n)^k)),$$

finishing the proof. $\square$

Proof of Theorem 2.3.6. The vanishing locus of the homogenized system of Lagrange equations $\widetilde{\mathbf{L}}_F = (\widetilde{f}_1, \ldots, \widetilde{f}_m, \widetilde{\ell}_1, \ldots, \widetilde{\ell}_n)$ is the intersection of Z with the vanishing locus of $\widetilde{f}_1, \ldots, \widetilde{f}_m$ and by Lemma 2.6.4 this intersection is a smooth zero-dimensional variety, contained in the big torus. By Lemma 2.5.8, the algebraic degree of sparse polynomial optimization is equal to its cardinality. According to Lemma 2.6.5, the system $\widetilde{\mathbf{L}}_F$ comprises global sections of $\mathbb{Q}$-Cartier divisors, associated to the respective, rational, polytopes $\mathrm{Newt}(f_1), \ldots, \mathrm{Newt}(f_m), \partial_1 \mathrm{Newt}(\Phi_F), \ldots, \partial_n \mathrm{Newt}(\Phi_F)$. As a consequence of Lemma 2.6.8, and using multilinearity of the mixed volume, we can express the number of solutions to $\widetilde{\mathbf{L}}_F$ as the mixed volume (2.4) of these polytopes. $\square$

2.6.3 The proof of Theorem 2.3.11

In this section we study the intersection of the determinantal variety W with the vanishing locus V of $\widetilde{f}_1, \dots, \widetilde{f}_m$ in X. The proof of Theorem 2.3.11 rests on a proof of transversality, and a characterisation of the cohomology class $[W]$ as a Porteous' class.

We start by recalling Porteous' formula, also called the Giambelli–Thom–Porteous formula. For more details refer to chapter 14 in [Ful98] and chapter 12 in [EH16]. The following statement is a special case of Theorem 12.4 in [EH16].

Theorem 2.6.9. *(Porteous' formula) Let $\varphi : \mathcal{E} \longrightarrow \mathcal{F}$ be a morphism of vector bundles of rank $m+1 \leq n$ on a smooth proper variety X of dimension n. We denote by W the (possibly non reduced) degeneracy locus of φ, supported on the set*

$$|W| = \{x \in X \ : \ \varphi_x : \mathcal{E}(x) \longrightarrow \mathcal{F}(x) \text{ is not injective}\}.$$

If W is pure of codimension $n-m$ then the cohomology class of W is the $n-m$ graded part

$$[W] = (\mathrm{s}(\mathcal{E})\,\mathrm{c}(\mathcal{F}))_{n-m}$$

of the product of the total Segre class $\mathrm{s}(\mathcal{E})$ and the total Chern class $\mathrm{c}(\mathcal{F})$.

We need the following, modified version of Porteous' formula which only requires W to be pure dimensional after restricting to a subvariety V of X.

Corollary 2.6.10. *Under the assumptions of Theorem 2.6.9, let V be an irreducible closed subvariety of X of codimension k, which intersects W transversally and let the intersection $V \cap W$ be pure of codimension $n-m+k$. Then the cohomology class $[V \cap W]$ is given by*

$$[V \cap W] = [V] \cdot (\mathrm{s}(\mathcal{E})\,\mathrm{c}(\mathcal{F}))_{n-m}\,.$$

Proof. Note that $V \cap W$ is the degeneracy locus of the restriction $\varphi|_V$. By applying Porteous formula to $\phi|_V : \mathcal{E}|_V \to \mathcal{F}|_V$ we get

$$[V \cap W] = (\mathrm{s}(\mathcal{E}|_V)\,\mathrm{c}(\mathcal{F}|_V))_{n-m+k}\,.$$

Lastly we notice that $(\mathrm{s}(\mathcal{E}|_V)\,\mathrm{c}(\mathcal{F}|_V))_{n-m+k} = [V] \cdot (\mathrm{s}(\mathcal{E})\,\mathrm{c}(\mathcal{F}))_{n-m}$ by naturality of characteristic classes. □

Lemma 2.6.11. *Under the assumptions of Theorem 2.3.11 the intersection $V \cap W$ is transversal and contained in the torus $(\mathbb{C}^*)^n$.*

Proof. Under the assumptions of Theorem 2.3.11, the assumptions from Section 2.5.2 are satisfied. The inclusion in the big torus follows from Lemma 2.6.4. We now show that transversality of the intersection $V \cap W$ follows from transversality of the intersection of Z with $\mathcal{V}(\widetilde{f}_1, \dots, \widetilde{f}_m)$. Let $\pi : \mathbb{P}(\mathcal{E}) \longrightarrow X$ denote the natural projection and let $z = (x, [\lambda])$ be any element of $\mathbb{P}(\mathcal{E})$. If Z intersects $\pi^{-1}(V)$ transversally at z, then for the tangent spaces $T_{Z,z}$ and $T_{\pi^{-1}(V),z}$ at z it holds

$$T_{Z,z} + T_{\pi^{-1}(V),z} = T_{\mathbb{P}(\mathcal{E}),z}. \tag{2.14}$$

To see that W and V intersect transversally at x we show $T_{W,x} + T_{V,x} = T_{X,x}$. We apply the differential $d\pi$ to both sides of (2.14) and note that we have the inclusions

$$d\pi(T_{Z,z}) \subseteq T_{W,x},\ d\pi(T_{\pi^{-1}(V),z}) \subseteq T_{V,x},\ d\pi(T_{\mathbb{P}(\mathcal{E}),z}) = T_{X,x}.$$ □

Proof of Theorem 2.3.11. By Lemma 2.5.8 the algebraic degree of sparse polynomial optimization is the cardinality of $V \cap W \cap \mathbb{C}^n$. By Lemma 2.6.11, the scheme theoretic intersection $V \cap W$ is a smooth variety of dimension zero, contained in the big torus. We now finish the proof by verifying the assumptions of Corollary 2.6.10.

Let D_{f_i} be the Weyl divisors introduced in equation (2.8). By the assumptions of Theorem 2.3.11, X is smooth. In particular, all divisors considered in this proof are Cartier. The variety W is defined to be the degeneracy locus of the matrix M, whose entries are global sections of the bundle $\mathcal{O}_X(D_{f_i} - D_{e_j})$. The transpose of M defines a morphism $\varphi : \mathcal{E} \to \mathcal{F}$ of vector bundles, where

$$\mathcal{E} = \mathcal{O}_X(-D_{f_0}) \oplus \cdots \oplus \mathcal{O}_X(-D_{f_m}), \text{ and } \mathcal{F} = \mathcal{O}_X(-D_{e_1}) \oplus \cdots \oplus \mathcal{O}_X(-D_{e_n}).$$

Now W is the degeneracy locus of φ, further $V \cap W$ is pure of dimension zero. Finally, $\mathcal{L}_{\mathcal{A}_i} = \mathcal{O}_X(D_{f_i})$ which finishes the proof. □

2.7 A polyhedral homotopy algorithm for computing critical points

In this final section we demonstrate how Question 3 from the introduction of this thesis can be addressed, based on our previous results. In the case where we have a linear objective and a single constraint we explicitly construct a polyhedral homotopy algorithm for solving the Lagrange system $\mathbf{L}_F$. A bottleneck in computing a start system for the Lagrange system $\mathbf{L}_F$ is solving the *tropicalization* of $\mathbf{L}_F$. Our algorithm relies on an explicit description of the tropical solution set of $\mathbf{L}_F$. The superiority over traditional homotopy continuation algorithms is demonstrated experimentally. Correctness of our algorithm follows from the previous intersection theoretic computations. We start with a reminder on polyhedral homotopy methods.

2.7.1 A discussion of polyhedral homotopy continuation

Homotopy continuation algorithms are a class of numerical algorithms used for finding all isolated solutions to a square system of polynomial equations. Specifically, suppose you have a square system of polynomial equations

$$F(x) = \{f_1(x), \ldots, f_n(x)\} = 0$$

where $f_i \in \mathbb{R}[x_1, \ldots, x_n]$ and the number of complex solutions to $F(x) = 0$ is finite. Homotopy continuation works by tracking solutions from an 'easy' system of polynomial equations (called the *start system*) to the desired one (called the *target system*). This is done by constructing a *homotopy*,

$$H(t; x) : [0, 1] \times \mathbb{C}^n \longrightarrow \mathbb{C}^n,$$

such that

1. $H(0; x) = G(x)$ and $H(1; x) = F(x)$,
2. the solutions to $G(x) = 0$ are isolated and easy to find
3. H has no singularities along the path $t \in [0, 1)$ and
4. H is sufficient for F.

Here we call a homotopy H *sufficient* for $F = H(1;x)$ if, by solving the ODE initial value problems $\frac{\partial H}{\partial t} + \frac{\partial H}{\partial x}\dot{x} = 0$ with initial values $\{x \ : \ G(x) = 0\}$, all isolated solutions of $F(x) = 0$ can be obtained.

One example of a homotopy, known as a *straight-line homotopy*, is defined as a convex combination of the start and target systems:

$$H(t;x) = \gamma(1-t)G(x) + tF(x)$$

where $\gamma \in \mathbb{C}$ is a generic constant. Choosing generic γ ensures $H(x;t)$ is non-singular for $t \in [0,1)$. Path tracking is typically done using standard predictor-corrector methods. For more information, see [BHSW13, Stu02]. The main question when employing homotopy continuation techniques is how to select such an 'easy' start system. If the target system roughly achieves the *Bézout bound* then a *total degree* start system is suitable. An example of this is

$$G(x) = \{x_1^{d_1} - 1, \ldots, x_n^{d_n} - 1\}$$

where $\deg(f_i) = d_i$.

Often in applications, the target system is defined by sparse polynomial equations. In this case, the Bézout bound can be a strict upper bound on the total number of complex solutions so using a total degree start system leads to wasted computation. The *BKK bound*, gives an upper bound on the number of complex solutions in the torus to a sparse polynomial system. If the BKK bound is much less than the Bézout bound, a *polyhedral* start system is a more economic choice, compared to a total degree start system. The downside of polyhedral homotopy is that the start system is more difficult to construct. This is not surprising since computing the mixed volume is #P hard [Kha93]. Still, there is an algorithm that computes this start system [HS95a]. We briefly outline the idea behind polyhedral homotopy here but give [HS95a] as a more complete reference.

Recall that $F = \{f_1, \ldots, f_n\}$, where $f_i = \sum_{\alpha \in \mathcal{A}_i} c_\alpha x^\alpha \in \mathbb{C}[x_1, \ldots, x_n]$. For each monomial, $\alpha \in \mathcal{A}_i$, we consider a *lifting*, $w(\alpha)$, and the corresponding lifted system $F^w(x,t) = (f_1^w(x,t), \ldots, f_n^w(x,t))$ where

$$f_i(x,t) = \sum_{\alpha \in \mathcal{A}_i} c_\alpha x^\alpha t^{w(\alpha)}. \tag{2.15}$$

Solutions to $F^w(x,t) = 0$ are algebraic functions in the parameter t. Such solutions can be written as

$$x(t) = (x_1(t), \ldots, x_n(t)).$$

In a neighborhood of $t = 0$, each solution can be written as $x(t) = (x_1(t), \ldots, x_n(t))$ where

$$x_i(t) = y_i t^{u_i} \ + \quad \text{higher order terms in } t$$

where $y_i \neq 0$ is a constant and $u_i \in \mathbb{Q}$. Substituting this into (2.15) we have

$$f_i(x,t) = c_\alpha y^\alpha t^{u^T\alpha + w(\alpha)} \ + \quad \text{higher order terms in } t.$$

By [HS95a, Lemma 3.1] We wish to find u such that

$$\min_{u \in \mathbb{R}^n} \{u^T\alpha + w(\alpha)\}$$

is achieved twice. For each solution u, the vector $(u,1)$ is an inner normal to one of the lower facets of the *Cayley polytope* of F. Further more, each such solution, u, then induces a binomial

polynomial system $\mathcal{B}_u$ which can be solved using Smith normal forms as well as a homotopy to track solutions from $\mathcal{B}_u(x) = 0$ to $F(x) = 0$. The sum of the number of solutions to $B_u(x) = 0$ for each solution u is equal to the BKK bound of $F(x)$. Therefore, if the coefficients of F are generic with respect to its monomial support, then polyhedral homotopy will track one homotopy path for each solution to $F(x) = 0$. We illustrate this on a small example.

Example 2.7.1. Consider the system of one polynomial equation in one unknown

$$f(x) = x^3 - x^2 + 2x - 1 = 0.$$

We wish to solve this polynomial system using homotopy continuation and a polyhedral start system. To do this we consider a lifted system of f which we obtain by weighting each monomial of f by some power of t:

$$f_t = t^{\omega_3}x^3 - t^{\omega_2}x^2 + 2t^{\omega_1}x - t^{\omega_0}.$$

Now suppose we choose weighting $(\omega_0, \omega_1, \omega_2, \omega_3) = (0, 3, 1, 2)$ so

$$f_t = t^2x^3 - tx^2 + 2xt^3 - t^0.$$

A figure of this lifting is given in Figure 2.3. Solutions to $f_t = 0$ lie in the field of Puiseux series of t and are of the form

$$x(t) = \hat{x}t^a + \text{ higher order terms in } t$$

where $a \in \mathbb{Q}$ and $\hat{x} \in \mathbb{C}^*$. For $x(t)$ to be a root of f_t, the lowest terms in t must cancel out. Substituting in $x(t) = \hat{x}t^a$ into f_t, we have

$$f_t(x(t)) = \hat{x}^3t^{3a+2} - \hat{x}^2t^{2a+1} + 2\hat{x}t^{a+3} - t^0. \tag{2.16}$$

To have cancellation of the lowest terms, we must have the minimum exponent in t achieved twice. In other words,

$$\min_a\{3a + 2, 2a + 1, a + 3, 0\}. \tag{2.17}$$

must be achieved twice. There are six options:

1. $3a + 2 = 2a + 1 < a + 3, 0$
2. $3a + 2 = a + 3 < 2a + 1, 0$
3. $3a + 2 = 0 < 2a + 1, a + 3$
4. $2a + 1 = a + 3 < 3a + 2, 0$
5. $2a + 1 = 0 < 3a + 2, a + 3$
6. $a + 3 = 0 < 3a + 2, 2a + 1$

The only feasible solutions are the first and fifth where $a = -1$ and $a = -\frac{1}{2}$, respectively. For the first case, we substitute $a = -1$ into (2.16) giving

$$\hat{x}^3t^{-1} - \hat{x}^2t^{-1} + 2\hat{x}t^2 - 1$$

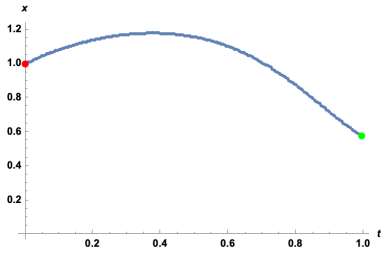


Figure 2.2: The homotopy $h_1(\hat{x}, t)$ from Example 2.7.1. The red point is the starting point induced by the binomial system $\hat{x}^3 - \hat{x}^2 = 0$ while the green point is the target solution, namely a zero of $f(x) = 0$.

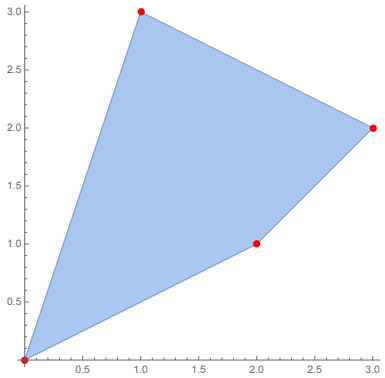


Figure 2.3: The polyhedral lift from Example 2.7.1

Multiplying through by t, we get

$$h_1(\hat{x}, t) = \hat{x}^3 - \hat{x}^2 + 2\hat{x}t^3 - t.$$

When $t = 0$ we have $h_1(\hat{x}, 0) = \hat{x}^3 - \hat{x}^2$ which has a unique $\mathbb{C}^*$ solution, $\hat{x} = 1$.

Similarly, we consider when $a = -\frac{1}{2}$ and substitute this value of a into (2.16) to get

$$h_2(\hat{x}, t) = \hat{x}^3 t^{\frac{1}{2}} - \hat{x}^2 + 2\hat{x}^{\frac{7}{2}} - 1.$$

When $t = 0$ we have $h_2(\hat{x}, 0) = -\hat{x}^2 - 1$ which has two $\mathbb{C}^*$ solutions, $\hat{x} = \pm\sqrt{-1}$. Therefore, to find all three solutions to $f(x) = 0$, we track the solution $\hat{x} = 1$ using the homotopy $h_1(\hat{x}, t)$ from $t = 0$ to $t = 1$ and the solutions $\hat{x} = \pm\sqrt{-1}$ using the homotopy $h_2(\hat{x}, t)$ from $t = 0$ to $t = 1$. A graphical depiction of the homotopy h_1 is shown in Figure 2.2.

Finally, one can observe in Figure 2.3 that the lifted polytope of $\text{Newt}(f)$ has two lower facets, $\mathcal{F}_1 = \text{Conv}\{(0, 0), (2, 1)\}$ and $\mathcal{F}_2 = \text{Conv}\{(2, 1), (3, 2)\}$. $\mathcal{F}_1$ has inner normal given by $(-\frac{1}{2}, 1)$ while $\mathcal{F}_2$ has inner normal given by $(-1, 1)$. These are precisely the solutions to (2.17).

The main bottleneck with employing polyhedral homotopy algorithms is finding the binomial start systems and corresponding homotopies. Example 2.7.1 shows how finding these start systems is equivalent to solving a tropical system for a fixed lifting. The main contribution of this section is to find these binomial start systems for polynomial systems arising as the Lagrange systems of polynomial optimization programs.

2.7.2 A linear optimization algorithm given a single constraint

We consider (POP) when $m = 1$ and $\deg(f_0) = 1$. Specifically, we consider a polynomial optimization problem of the form

$$\min_{x\in\mathbb{R}^n} u^T x \quad \text{s.t.} \quad f(x) = 0 \tag{2.18}$$

where $u \in \mathbb{R}^n$ and $f(x)$ is a general degree $d \geq 2$ polynomial. We wish to design a homotopy algorithm to find all critical points to (2.18). We first consider the Lagrange system $\mathbf{L}_{u,f} = \{\ell_1, \ldots, \ell_n, f\}$ of (2.18) where

$$\ell_i = u_i - \lambda \frac{\partial}{\partial x_i} f(x). \tag{2.19}$$

If f is a generic degree d polynomial and $u \in \mathbb{R}^n$ is generic, then by Theorem 2.3.7, the number of critical points to (2.18) is the same as that of

$$\min_{x\in\mathbb{R}^n} u^T x \quad \text{s.t.} \quad \hat{f}(x) = 0 \tag{2.20}$$

where $\hat{f} = \sum_{i=1}^n c_i x_i^d$ and c_i is generic for $i \in [n]$. The Lagrange system of (2.20) is $\mathbf{L}_{u,\hat{f}} = \{\hat{\ell}_1, \ldots, \hat{\ell}_n, \hat{f}\}$ where for $i \in [n]$

$$\hat{\ell}_i = u_i - d\lambda c_i x_i^{d-1} \tag{2.21}$$

Observe that by Theorem 2.3.7, not only are the algebraic degrees of (u, f) and $(u, \hat{f})$ the same, but the BKK bound of $\mathbf{L}_{u,\hat{f}}$ is the same as that of $\mathbf{L}_{u,f}$.

The Lagrange system $\mathbf{L}_{u,\hat{f}}$ is sparser than $\mathbf{L}_{u,f}$ and in fact a binomial start system G for $\mathbf{L}_{u,\hat{f}}$ can be constructed efficiently. The following lemma shows that this is desirable since start systems for $\mathbf{L}_{u,\hat{f}}$ are start systems for $\mathbf{L}_{u,f}$ as well. We first need an observation about the existence of straight-line homotopies.

Proposition 2.7.2. *Let $F(x;p) : \mathbb{C}^n \times \mathbb{C}^k \longrightarrow \mathbb{C}^n$ denote a family of polynomial systems $F(x;p)$ that depends polynomially on parameters $p \in \mathbb{C}^k$ and $F(x;p_1)$ a fixed member of that family. Then there is a nonempty set $U \subseteq \mathbb{C}^k$, open and dense in the Euclidean topology, such that for every parameter p_0 in U the straight-line homotopy*

$$H(t;x) = tF(x;p_1) + (1-t)F(x;p_0)$$

is sufficient for $F(x;p_1)$.

Proof. By the *Parameter Continuation Theorem* by Morgan and Sommese [SW05] there exists a proper algebraic subvariety $\Sigma \subset \mathbb{C}^k$ with the following property: Let $\rho : [0,1] \to \mathbb{C}^k$ be any smooth path and $H(t,x) = F(x, \rho(t))$ the corresponding homotopy. If

$$\rho([0,1)) \cap \Sigma = \emptyset,$$

then as $t \to 1$, the limits of the solution paths $x(t)$ satisfying $H(x(t),t) = 0$ include all the isolated solutions to $F(x;\rho(1)) = 0$. In particular, $H(t,x)$ is a sufficient homotopy.

From now on we identify the complex affine space $\mathbb{C}^k$ with real affine space $\mathbb{R}^{2k}$ and denote by $\overline{\Sigma}$ the closure of Σ in real projective space $\mathbb{P}^{2k+1}_{\mathbb{R}}$. Consider the projection $\pi : \mathbb{P}^{2k+1}_{\mathbb{R}} \to \mathbb{P}^{2k}_{\mathbb{R}}$ away from the point p_1. Since the codimension of $\overline{\Sigma}$, considered as a manifold, is at least two, the image

$\pi\left(\overline{\Sigma}\right)$ has codimension at least one in $\mathbb{P}^{2k}_{\mathbb{R}}$. In particular, the image $\pi(p_0)$ of a generic element p_0 is not contained in $\pi\left(\overline{\Sigma}\right)$. Since the image $\rho([0,1))$ of the straight path

$$\rho(t) = (1-t)p_1 + tp_0$$

between p_0 and p_1 is is contained in the fiber $\pi^{-1}(\pi(p_0))$, it does not intersect Σ. Consequently the to ρ associated straight-line homotopy is sufficient. □

Let $BKK(\mathbf{L}_{u,\hat{f}})$ denote the BKK bound of $\mathbf{L}_{u,\hat{f}}$.

Lemma 2.7.3. *Let G be a zero dimensional square system of polynomials with exactly $BKK(\mathbf{L}_{u,\hat{f}})$ solutions. There is a sufficient homotopy connecting G to $\mathbf{L}_{u,f}$.*

Proof. Let $F(x;c)$ denote the family of polynomials with monomial support contained in the support of $\mathbf{L}_{u,f}$. In particular, the coefficient vector c has one entry for each monomial of each polynomial of $\mathbf{L}_{u,f}$. We denote by $F(x;c_0)$ a generic member of this family.

The desired homotopy will be constructed explicitly as a composition. We start by connecting $F(x;c_0)$ to both $\mathbf{L}_{u,f}$ and G with a straight-line homotopy, which by Proposition 2.7.2 is a sufficient homotopy in both cases. We denote the straight-line homotopy from $F(x;c_0)$ to G by H. It now suffices to prove that H does not merge any solutions of $F(x;c_0)$, allowing us to define the inverted homotopy H^* by setting for t in $(0,1)$ $H^*(t,x) = H(1-t,x)$ and $H^*(0,x) = G(x)$. Since tracking the roots of $F(x;c_0)$ to the roots of G along the sufficient homotopy H defines a surjective map, it is enough to prove that $F(x;c_0)$ and G have the same number of solutions.

By the results of Bernstein and Kushnirenko [Ber75, Kou76], the number of solutions of $F(x;c_0)$ is equal to the BKK bound of $\mathbf{L}_{u,f}$. By Theorem 2.3.7, the polynomial system $\mathbf{L}_{u,f}$ achieves this bound. Furthermore, as we noted at the beginning of Section 2.7.2, $\mathbf{L}_{u,f}$ and $\mathbf{L}_{u,\hat{f}}$ have the same number of solutions:

$$\#\{\mathbf{L}_{u,\hat{f}}\} = \#\{\mathbf{L}_{u,f} = 0\} = BKK(\mathbf{L}_{u,\hat{f}}). \tag{2.22}$$

At the same time the number of solutions to G is equal to the BKK bound of $\mathbf{L}_{u,\hat{f}}$, which is upper bounded by the BKK bound of $\mathbf{L}_{u,f}$ by inclusion on Newton poytopes:

$$\#\{\mathbf{L}_{u,\hat{f}}\} \leq \#\{G = 0\} \leq BKK(\mathbf{L}_{u,\hat{f}}). \tag{2.23}$$

Together, inequalities (2.22) and (2.23) imply that $\mathbf{L}_{u,\hat{f}}$ and G have the same root count. □

We now give the main result of this section.

Theorem 2.7.4. *For any $d \geq 2$ consider the Lagrange system of (2.18). Then for generic u and f there are $d(d-1)^{n-1}$ complex solutions to the corresponding Lagrange system. Moreover, all of these solutions can be found via the homotopy*

$$H(x,\lambda;t) = (1-t)B(x,\lambda) + t\gamma\mathbf{L}_{u,f}(x,\lambda)$$

where

$$B(x,\lambda) = \begin{cases} u_1 - d\lambda c_1 x_1^{d-1} = 0 \\ \vdots \\ u_n - d\lambda c_n x_n^{d-1} = 0 \\ c_0 + c_1 x_1^d = 0, \end{cases} \tag{2.24}$$

$\gamma \in \mathbb{C}$ is a generic constant and $\mathbf{L}_{u,f}(x,\lambda)$ is the Lagrange system of (2.18).

Proof. In order to design a polyhedral homotopy algorithm as described in [HS95a], in the following we construct a binomial start system B of $\mathbf{L}_{u,\hat{f}}$ by solving a tropical system. By the proof of Lemma 2.7.3 we then obtain a homotopy from B to $\mathbf{L}_{u,f}$. Note that, by genericity of f, this homotopy can be chosen to be a straight-line homotopy.

By Lemma 2.7.3 it suffices to design a polyhedral homotopy algorithm as described in [HS95a] for $\mathbf{L}_{u,\hat{f}}$. In order to define this algorithm, we need to first find a binomial start system of $\mathbf{L}_{u,\hat{f}}$ which can be done by solving a tropical system.

Let a_i be the tropical variable corresponding to x_i and b the tropical variable corresponding to λ. Then for a given lifting $\omega \in \mathbb{R}^{3n+1}$, the corresponding tropical system that we want to solve is

$$\begin{aligned} &\min_{a\in\mathbb{Q}^n, b\in\mathbb{Q}} \{\omega_{1,1}, (d-1)a_1 + b + \omega_{1,2}\} \\ &\vdots \\ &\min_{a\in\mathbb{Q}^n, b\in\mathbb{Q}} \{\omega_{n,1}, (d-1)a_n + b + \omega_{n,2}\} \\ &\min_{a\in\mathbb{Q}^n, b\in\mathbb{Q}} \{\omega_{n+1,1}, da_1 + \omega_{n+1,2}, \ldots, da_n + \omega_{n+1,n+1}\} \end{aligned} \tag{2.25}$$

We consider a specific lifting that induces a unique solution to (2.25), giving a homotopy from one binomial start system to the desired target system (2.21). With the particular lifting

$$\omega_{ij} = \begin{cases} 0 & \text{if} \quad 1 \le i \le n+1, \ j = 1 \\ 1-d & \text{if} \quad 1 \le i \le n, \ j = 2 \\ -d & \text{if} \quad (i,j) = (n+1, 2) \\ 1-d & \text{else} \end{cases} \tag{2.26}$$

This gives the following tropical system:

$$\begin{aligned} &\min_{a\in\mathbb{Q}^n, b\in\mathbb{Q}} \{0, (d-1)a_1 + b + 1 - d\} \\ &\vdots \\ &\min_{a\in\mathbb{Q}^n, b\in\mathbb{Q}} \{0, (d-1)a_n + b + 1 - d\} \\ &\min_{a\in\mathbb{Q}^n, b\in\mathbb{Q}} \{0, da_1 - d, da_2 + 1 - d, \ldots, da_n + 1 - d\} \end{aligned} \tag{2.27}$$

We claim there is a unique solution to (2.27) given by $a_i = 1$ for $i \in [n]$ and $b = 0$.

First, observe that the first n equations of (2.27) force $(d-1)a_i + b + 1 - d = 0$ for $i \in [n]$. This gives $a_i = \frac{d-1-b}{d-1}$. Substituting this into the final equation and simplifying we have that

$$\min_{a\in\mathbb{Q}^n, b\in\mathbb{Q}} \{0, \frac{bd}{1-d}, \frac{bd}{1-d} + 1, \ldots, \frac{bd}{1-d} + 1\}$$

must have minimum attained twice. It is then clear that the only solution is $b = 0$ where the minimum is achieved at the first two terms. Back substituting then gives that $a_i = \frac{d-1}{d-1} = 1$ for $i \in [n]$. The binomial start system $\mathcal{B}(x, \lambda)$ defined in (2.24) then follows immediately from the solution to this tropical system. □

Observe that Bézout's Theorem gives an upper bound that (2.21) has at most d^{n+1} solutions but we see that the binomial system (2.24) has $d(d-1)^{n-1}$ solutions. This gives another proof of

n	20	30	40	50	60	70	80	90
Polyhedral	0.14	0.51	1.01	2.30	4.49	NA	NA	NA
H	0.07	0.20	0.35	0.87	1.65	2.54	3.78	6.45

Table 2.1: Average time (sec) to find all critical points to (2.18) when $d = 2$ using standard polyhedral homotopy versus the homotopy, H, outlined in Theorem 2.7.4.

n	6	7	8	9	10	11	12
Polyhedral	0.29	0.93	3.06	9.79	27.42	88.37	556.92
H	0.21	0.68	2.29	7.35	20.35	70.02	395.64

Table 2.2: Average time (sec) to find all critical points to (2.18) when $d = 3$ using standard polyhedral homotopy versus the homotopy, H, outlined in Theorem 2.7.4.

the bound given in [NR09] for hypersurfaces and highlights the benefit of using a polyhedral start system over a total degree start system.

Finally, we wish to remark that the homotopy defined in Theorem 2.7.4 will work for finding all smooth critical points for the optimization of a linear function over any hypersurface, f, so long as $\text{Newt}(f)$ is contained in $\text{Conv}\{0, de_1, \ldots, de_n\}$. When $\text{Newt}(f)$ is a strict subset of $\text{Conv}\{0, de_1, \ldots, de_n\}$, then the algebraic degree of f can be less than $d(d-1)^{n-1}$. This homotopy may lead to wasted computation in tracking divergent paths.

2.7.3 Numerical results

We implement the homotopy in Theorem 2.7.4 with start system (2.24) using the path tracking function in `HomotopyContinuation.jl`. We compare our implementation of the homotopy outlined in Theorem 2.7.4 against the polyhedral one in `HomotopyContinuation.jl` and give the time it takes to run each homotopy algorithm in Table 2.1, Table 2.2 and Table 2.3. The computations are all run using a 2018 Macbook Pro with 2.3 GHz Quad-Core Intel Core i5.

In all cases, our homotopy algorithm is much faster than the standard off the shelf software. When the hypersurface is of degree two, there are only two complex critical points. Despite this, standard polyhedral homotopy was unable to compute a start system when $n \geq 70$. In contrast, our specialized algorithm was able to find both critical points in a few seconds. We note that in this case, the Bézout bound of the corresponding polynomial system is 2^{n+1} where n is the number of variables. When $n = 70$, the Bézout bound is $\approx 2.36 \times 10^{21}$, so it is unreasonable to expect that a total degree homotopy would work in this case.

Similarly, in Table 2.2 and Table 2.3 we see that when the degree of the hypersurface is three or four, our algorithm increasingly outperforms the state-of-the-art polyhedral homotopy software as the number of variables increases.

n	3	4	5	6	7	8	9
Polyhedral	0.03	0.17	1.16	7.04	40.11	228.48	1225.78
H	0.03	0.15	0.83	5.15	34.79	181.11	1027.64

Table 2.3: Average time (sec) to find all critical points to (2.18) when $d = 4$ using standard polyhedral homotopy versus the homotopy, H, outlined in Theorem 2.7.4.

2.8 Conclusion

In this chapter we studied polynomial programs that exhibit sparsity patterns with the tools of toric geometry. Our focus was their algebraic complexity which we measure in terms of the number of their critical points. We quantified this number in many instances in Theorem 2.3.6, Theorem 2.3.7 and Theorem 2.3.11. We further demonstrated in some special cases that these results can be made effective by implementing a polyhedral homotopy algorithm that efficiently solves Lagrange systems, based on Theorem 2.7.4.

Chapter 3

Certifying zeros of polynomial systems using interval arithmetic

Numerical algebraic geometry has been emerging as an alternative to symbolic computation methods with increasing performance and also versatility. Although many problems can be solved that are infeasible with symbolic methods, the computation results lack a certificate for correctness. This drawback keeps researchers from using numerical computation in proofs and pure mathematics. This chapter develops interval arithmetic as a practical tool for certification in numerical algebraic geometry. We present a built-in function `certify` in the software `HomotopyContinuation.jl`. It proves the correctness of an isolated nonsingular solution to a square system of polynomial equations, resting on Krawczyk's method. We demonstrate that it dramatically outperforms earlier approaches to certification. We see this contribution as a powerful new tool in numerical algebraic geometry.

3.1 Introduction

Systems of polynomial equations appear in many areas of mathematics, as well as in many applications in the sciences and engineering. In physics and chemistry the geometry of molecules is often modelled with algebraic constraints on the distance or the angle between atoms. In kinematics the relation between robot joints is defined by polynomial equations. In systems biology the steady-state equations for many bio-chemical reaction networks are algebraic equations. A central task in all those applications is computing the isolated zeros of a system of polynomials.

The study of zeros of polynomial systems is at the heart of algebraic geometry. The field of *computational algebraic geometry* is often associated with symbolic computations based on Gröbner bases. But over the last thirty years *numerical algebraic geometry* (NAG) [SW05] emerged as an alternative; enabling us to solve problems infeasible with symbolic methods. An important algorithmic framework in NAG is *numerical homotopy continuation*. Several implementations of this are available: `Bertini` [BHSW], `Hom4PS-3` [CLL14], `HomotopyContinuation.jl` [BT18], `NAG4M2` [Ley11] and `PHCpack` [Ver99].

Hauenstein and Sottile remark in [HS12] that while all of these softwares "routinely and reliably solve systems of polynomial equations with dozens of variables having thousands of solutions", they have the shortcoming that "the output is not certified" and that "this restricts their use in some applications, including those in pure mathematics". To remedy this, Hauenstein and Sottile developed the software `alphaCertified` [HS12]. It can rigorously certify that Newton's method,

starting at a given numerical approximation, converges quadratically to a true zero by using Smale's α-theory [Sma86]. Hauenstein and Sottile's contribution to numerical algebraic geometry was a milestone. Yet, `alphaCertified` produces rigorous certificates using expensive rational arithmetic. This turns the big advantage of numerical computations, namely that they are fast, upside-down and makes certification of large problems prohibitively expensive. Thus, up to this point, the majority of researchers in applied algebraic geometry were kept from using numerical methods, because certification was too expensive and because without certification numerical methods can't be used for proofs.

We give researchers a new powerful tool in numerical algebraic geometry. Our implementation is integrated in `HomotopyContinuation.jl` [BT18], so that in principle we can certify *all* zeros of a system of polynomial equations (see Section 3.1.2 below for more details). With a fast implementation certification becomes the default and is not just an option and enables the extensive use of numerical methods for rigorous proofs. This is underlined by at least 15 research works [BRST23, BFS21, KPR+21, BPS21, BHIM22, Ear21, Mar21, Wei21, LAR21, Stu21, BT21, ABF+23, BKK20, SY21, ST21] that have used our implementation in the last two years.

3.1.1 Contribution

Our contribution to the field of computational and applied algebraic geometry is an extremely fast and easy-to-use implementation of a certification method. This implementation outperforms `alphaCertified` by several orders of magnitude. It makes the certification of solutions often a matter of seconds and not hours or days. This leap in performance can turn certification in numerical algebraic geometry into default and not just an option.

Starting from version 2.1, `HomotopyContinuation.jl` has a function `certify`[1]. The function `certify` takes as input a *square polynomial system* F and a *numerical approximation of a complex zero* $x \in \mathbb{C}^n$ (or a list of zeros). If the output says "certified", then this is a rigorous proof that a solution of $F = 0$ is near x. If the output says "not certified", then this does not necessarily mean that there is no zero near x, just that the method couldn't find one. Figure 3.2 shows an example of `certify`.

We combine interval arithmetic and Krawczyk's method with numerical algebraic geometry to rigorously certify solutions to square systems of polynomial equations. In technical terms, our implementation returns *strong interval approximate zeros*. We introduce this notion in Definition 3.3.8 below. The strong interval approximate zero consists of a box in $\mathbb{C}^n$, which contains a unique true zero of the polynomial system. If the input is a list of zeros, the routine returns a list of distinct strong interval approximate zeros. Therefore, our method can be used to *prove* hard lower bounds on the number of zeros of a polynomial system. Combined with theoretical upper bounds this can constitute rigorous mathematical proofs on the number of zeros of such systems. We explain this in more detail in the next subsection. In addition, if the given polynomial system is real, we give a certificate whether the certified zero is a real zero (the approximate zero does not need to be real for this). The returned boxes may also be used to check whether a real zero is positive real. Therefore, our method can also be used to prove lower bounds on the number of real and positive real zeros of a polynomial system.

It is also possible to give a square system of rational functions as input to our implementation. Although this chapter is mostly formulated in terms of polynomial systems, Krawczyk's method also applies to square systems of rational functions (in fact, to all analytic functions $\mathbb{R}^n \to \mathbb{R}^n$; see

[1] The technical documentation is available at
`https://www.juliahomotopycontinuation.org/HomotopyContinuation.jl/stable/certification`

Section 3.3). Consequently, all statements about using our implementation for proofs are equally valid for square systems of rational functions.

3.1.2 Certifying all zeros

Our implementation is integrated in `HomotopyContinuation.jl` [BT18]. This is a software for numerically solving systems of polynomials equations via homotopy continuation. The basic idea was already explained in Chapter 2, but we repeat it for the reader: suppose that $F(x) = (f_1(x), \dots, f_n(x))$ is a system of polynomials in n variables $x = (x_1, \dots, x_n)$. To compute the solutions of the systems of equation $F(x) = 0$ one takes another system of polynomials $G(x) = (g_1(x), \dots, g_n(x))$, called *start system*, for which the zeros are simple to compute. Then, F and G are joined with a path in the vector space of systems of polynomials. This path defines a homotopy $H(x,t) : \mathbb{C}^n \times \mathbb{C} \to \mathbb{C}$, such that $H(x,1) = G(x)$ and $H(x,0) = F(x)$. The zeros of G are continued towards the zeros of F by solving the ODE initial value problems $\frac{\partial H}{\partial t} + \frac{\partial H}{\partial x}\dot{x}(t) = 0$, where $x(1)$ ranges over the zeros of G. For more details see the textbook [SW05].

In the last paragraph there is nothing special about polynomials. This approach works for any analytic functions F and G. However, in the case of systems of polynomial equations we can choose G such that we compute *all* zeros of F. This follows from the *Parameter Continuation Theorem* by Morgan and Sommese [MS89]: suppose that $F(x) = F(x; p_0)$ is a point in family of polynomial systems $F(x;p)$ that depends polynomially on parameters $p \in \mathbb{C}^k$. The Parameter Continuation Theorem says that there exists a proper algebraic subvariety $\Sigma \subset \mathbb{C}^k$ with the following property. Let $\gamma(t) \colon [0,1] \to \mathbb{C}^k$ with $\gamma(0) = p_0$ be a continuous path and $H(x,t)$ the corresponding homotopy. If $\gamma((0,1]) \cap \Sigma = \emptyset$, then as $t \to 0$, the limits of the solution paths $x(t)$ with $H(x(t),t) = 0$ include all the isolated solutions to $F(x;p_0) = 0$. This includes both regular solutions and solutions with multiplicity greater than one. Consequently, every parameter outside Σ provides a suitable start system for F.

The *Parameter Continuation Theorem* implies the existence of start systems, such that we can compute all zeros of F, but it does not tell us how to set up these start systems, nor how to compute their zeros. In fact, different choices of families of parametrized systems lead to different start systems and thus different homotopy methods. In `HomotopyContinuation.jl` [BT18] one can choose between two well-established strategies for choosing start systems: the so-called totaldegree start system and the polyhedral start system [HS95a].

Coming back to interval arithmetic we see that the zeros computed by polynomial homotopy continuation can be used as input for certification, so that we can *certify all solutions* of a system of polynomial equations. There is one subtlety, though. Although the Parameter Continuation Theorem asserts that in principle we can find all solutions, since homotopy continuation involves numerical computations we can't rule out the possibiliy that some computations of solutions paths fail. Still, combining certification with homotopy continuation always gives *lower bounds* on the number of zeros. This can be exploited in situations, where we know upper bounds. An example of such a scenario in enumerative geometry is discussed in Section 3.5.1 below, where we certify 3264 real zeros of a system of polynomials that is known to have at most 3264 complex zeros. Another example from optimization is discussed in [BSW21, Section 3.3]. Here, certifying all zeros of a system of polynomial equations helps to rigorously compute the minimal Euclidean distance of a point to an algebraic hypersurface. In Chapter 2 we generalize the results from [BSW21] in several ways. The constraint set is now allowed to be a complete intersection of higher codimension than one, and the objective function needs not be the euclidean distance. In Chapter 4 below we will study similar bounds in the context of optimizing decision rules and provide bounds to the number

of critical points of the associated polynomial program.

3.1.3 Comparison to previous works

There are other implementations of certification methods using Krawczyk's method and interval arithmetic, e.g., the commercial `MATLAB` package `INTLAB` [Rum99], the `Macaulay2` package `NumericalCertification` [Lee19], and the `Julia` package `IntervalRootFinding.jl` [BS]. The theory of Krawczyk's method and interval arithmetic are explained, for instance, in [Rum83].

Unlike `INTLAB`, the source code of our implementation is freely available and can be verified by anyone. Additionally, `INTLAB` doesn't support the use of arbitrary precision interval arithmetic which limits its capability to certify badly conditioned solutions. `NumericalCertification`, as of version 1.0, takes as input not the numerical approximation of a complex zero $x \in \mathbb{C}^n$, but instead a box I in $\mathbb{C}^n$. Then, `NumericalCertification` attempts to certify that interval I is a strong interval approximate. The process of going from a numerical approximation x to a good candidate interval I needs particular care, as illustrated in Section 3.4. `INTLAB` and `NumericalCertification` also both require manual work to obtain a list of all distinct strong interval approximate zeros. The package `IntervalRootFinding.jl` finds all zeros of a multivariate function inside a given box in $\mathbb{R}^n$, whereas our implementation works in $\mathbb{C}^n$ and additionally certifies reality of zeros; see Section 3.3.2.

3.1.4 Outline

The rest of this chapter is organized as follows. In the next two sections we give a short introduction to interval arithmetic and explain the Krawczyk method. Section 3.4 focuses on implementation details. In Section 3.5 we demonstrate features of our implementation using two examples. In Section 3.3.2 we discuss how to certify reality of zeros, and for completeness, we give a proof of Krawczyk's method in Section 3.3.

3.2 Interval arithmetic

Before we discuss our implementation, let us briefly introduce the basics of interval arithmetic.

Since the 1950s researchers [Moo66, Sun58] have worked on interval arithmetic, which allows certified computations while still using floating point arithmetic. We briefly introduce the concepts from interval arithmetic which are relevant for our chapter.

3.2.1 Real interval arithmetic

Real interval arithmetic concerns computing with compact real intervals. Following [May17] we denote the set of all compact real intervals by

$$\mathbb{IR} := \{[a, b] \mid a, b \in \mathbb{R}, a \leq b\}.$$

For $X, Y \in \mathbb{IR}$ and the binary operation $\circ \in \{+, -, \cdot, /\}$ we define

$$X \circ Y = \{x \circ y \,|\, x \in X, y \in Y\}, \tag{3.1}$$

where we assume $0 \notin Y$ in the case of division. The interval arithmetic version of these binary operations, as well as other standard arithmetic operations, have explicit formulas. See, e.g., [May17, Sec. 2.6] for more details.

3.2.2 Complex interval arithmetic

We define the set of *rectangular complex intervals* as

$$\mathbb{IC} := \{X + iY \mid X, Y \in \mathbb{IR}\},$$

where $X + iY = \{x + iy \mid x \in X, y \in Y\}$ and $i = \sqrt{-1}$. Following [May17, Ch. 9] we define the algebraic operations for $I = X + iY, J = W + iZ \in \mathbb{IC}$ in terms of operations on the real intervals from (3.1):

$$\begin{aligned} I + J &:= (X + W) + i(Y + Z), & I \cdot J &:= (X \cdot W - Y \cdot Z) + i(X \cdot Z + Y \cdot W) \\ I - J &:= (X - W) + i(Y - Z), & \frac{I}{J} &:= \frac{X \cdot W + Y \cdot Z}{W \cdot W + Z \cdot Z} + i\frac{Y \cdot W - X \cdot Z}{W \cdot W + Z \cdot Z}. \end{aligned} \tag{3.2}$$

It is necessary to use (3.1) instead of complex arithmetic for the definition of algebraic operations in $\mathbb{IC}$. The following example from [May17] demonstrates this. Consider the intervals $I = [1, 2]+i[0, 0]$ and $J = [1, 1] + i[1, 1]$. Then, $\{x \cdot y | x \in I, y \in J\} = \{t(1 + i) \mid 1 \leq t \leq 2\}$ is not a rectangular complex interval, while $I \cdot J = [1, 2] + i[1, 2]$ is.

The algebraic structure of $\mathbb{IC}$ is given by following theorem; see, e.g., [May17, Theorem 9.1.4].

Theorem 3.2.1. *The following holds.*

1. $(\mathbb{IC}, +)$ *is a commutative semigroup with neutral element.*
2. $(\mathbb{IC}, +, \cdot)$ *has no zero divisors.*

Furthermore, if $I, J, K, L \in \mathbb{IC}$, *then*

3. $I \cdot (J + K) \subseteq I \cdot J + I \cdot K$, *but equality does not hold in general.*
4. $I \subseteq J, K \subseteq L$, *then* $I \circ K \subseteq J \circ L$ *for* $\circ \in \{+, -, \cdot, /\}$.

Working with interval arithmetic is challenging because of the third item from the previous theorem: distributivity does not hold in $\mathbb{IC}$. As a consequence, in $\mathbb{IC}$ the evaluation of polynomials depends on the exact order of the evaluation steps. Therefore, the evaluation of polynomial maps $F : \mathbb{IC}^n \to \mathbb{IC}$ is only well-defined if F is defined by a straight-line program, and not just by a list of coefficients. Figure 3.1 demonstrates this issue in an example. See, e.g., [BCS13, Sec. 4.1] for an introduction to straight-line programs.

Figure 3.1: The picture shows two straight-line programs for evaluating the polynomial $f(x, y, z) = (x + y)z$. Let $I = ([-1, 0], [1, 1], [0, 1])^T$. Then, the program on the left evaluated at I yields $f(I) = ([-1, 0] + [1, 1])[0, 1] = [0, 1]$, while the program on the right yields $f(I) = [-1, 0][0, 1] + [1, 1][0, 1] = [-1, 1]$.

Arithmetic in $\mathbb{IC}^n$ is defined in the expected way. If $I = (I_1, \ldots, I_n), J = (J_1, \ldots, J_n) \in \mathbb{IC}^n$,

$$I + J = (I_1 + J_1, \ldots, I_n + J_n).$$

Scalar multiplication for $I \in \mathbb{IC}$ and $J \in \mathbb{IC}^n$ is defined as $I \cdot J = (I \cdot J_1, \ldots, I \cdot J_n)$. The product of an interval matrix $A = (A_{i,j}) \in \mathbb{IC}^{n \times n}$ and an interval vector $I \in \mathbb{IC}^n$ is

$$A \cdot I := I_1 \cdot \begin{bmatrix} A_{1,1} \\ \vdots \\ A_{n,1} \end{bmatrix} + \cdots + I_n \cdot \begin{bmatrix} A_{1,n} \\ \vdots \\ A_{n,n} \end{bmatrix}. \tag{3.3}$$

Similar to the one-dimensional case $(\mathbb{IC}^n, +)$ is a commutative semigroup with neutral element.

3.3 Certifying zeros with interval arithmetic

In 1969 Krawczyk [Kra69] developed an interval arithmetic version of Newton's method. Later in 1977 Moore [Moo77] recognized that Krawczyk's method can be used to certify the existence and uniqueness of a solution to a system of nonlinear equations. Interval arithmetic and interval Newton's method are a prominent tool in many areas of applied mathematics; e.g., in chemical engineering [GS05], thermodynamics [GD05] and robotics [KSS15]. See also the overview in [Rum10].

The results in this section are stated for general functions. For a practical implementation it is however necessary to compute interval enclosures (see Definition 3.3.1). We discuss our approach in the context of polynomial systems in Section 3.4.1 below. A generalization in this spirit is discussed in [BLL19].

3.3.1 Krawczyk's method

In this section we recall Krawczyk's method. First, we need three definitions.

Definition 3.3.1 (Interval enclosure)**.** Let $F : \mathbb{C}^n \to \mathbb{C}^n$. A map $\square F : \mathbb{IC}^n \to \mathbb{IC}^n$ is an interval enclosure of F if for every $I \in \mathbb{IC}^n$ we have $\{F(x) \mid x \in I\} \subseteq \square F(I)$.

In the rest of this chapter we use the notation $\square F$ to denote the interval enclosure of F. Also, we do not distinguish between a point $x \in \mathbb{C}^n$ and the complex interval $[\mathrm{Re}(x), \mathrm{Re}(x)] + i[\mathrm{Im}(x), \mathrm{Im}(x)]$ defined by x. We simply use the symbol "x" for both terms so that $\square F(x)$ is well-defined.

Definition 3.3.2 (Interval matrix norm)**.** Let $A \in \mathbb{IC}^{n \times n}$. We define the operator norm of A as $\|A\|_\infty := \max_{B \in A} \max_{v \in \mathbb{C}^n} \frac{\|Bv\|_\infty}{\|v\|_\infty}$, where $\|(v_1, \ldots, v_n)\|_\infty = \max_{1 \le i \le n} |v_i|$ is the infinity norm in $\mathbb{C}^n$.

Next we introduce an interval version of the Newton operator, the *Krawczyk operator* [Kra69].

Definition 3.3.3. Let $F : \mathbb{C}^n \to \mathbb{C}^n$ be differentiable, and $\mathrm{J}F$ be its Jacobian matrix seen as a function $\mathbb{C}^n \to \mathbb{C}^{n \times n}$. Let $\square F$ be an interval enclosure of F and $\square \mathrm{J}F$ be an interval enclosure of $\mathrm{J}F$. Furthermore, let $I \in \mathbb{IC}^n$ and $x \in \mathbb{C}^n$ and let $Y \in \mathbb{C}^{n \times n}$ be an invertible matrix. We define the Krawczyk operator

$$K_{x,Y}(I) := x - Y \cdot \square F(x) + (\mathbf{1}_n - Y \cdot \square \mathrm{J}F(I))(I - x).$$

Here, $\mathbf{1}_n$ is the $n \times n$-identity matrix.

Remark 3.3.4. In the literature, $K_{x,Y}(I)$ is often defined using $F(x)$ and not $\square F(x)$. Here, we use this definition, because in practice it is usually not feasible to evaluate $F(x)$ exactly. Instead, $F(x)$ is replaced by an interval enclosure.

Remark 3.3.5. The second part of Theorem 3.3.6 motivates to find a matrix $Y \in \mathbb{C}^{n\times n}$ such that $||1_n - Y \cdot \square \mathrm{J}F(I)||_\infty$ is minimized. A good choice is an approximation of the inverse of $\mathrm{J}F(x)$.

We are now ready to state the theorem behind Krawczyk's method. The first proof for real interval arithmetic is due to Moore [Moo77]. A proof for complex data is at least known since the work by Rump [Rum83]. Note that all the data in the theorem can be computed using interval arithmetic.

Theorem 3.3.6. *Let $F : \mathbb{C}^n \to \mathbb{C}^n$ be differentiable and $I \in \mathbb{IC}^n$. Let $x \in I$ and $Y \in \mathbb{C}^{n\times n}$ be an invertible complex $n \times n$ matrix. The following holds:*

1. *If $K_{x,Y}(I) \subset I$, there is a zero of F in I.*
2. *If additionally $\sqrt{2}\,\|\mathbf{1}_n - Y\square \mathrm{J}F(I)\|_\infty < 1$, then F has exactly one zero in I.*

Remark 3.3.7. One can prove a version of this theorem without $\sqrt{2}$ in the second item. We state the version above, because our implementation uses the $\sqrt{2}$ factor. In most cases, it does not make much of a difference whether one has this additional factor or not. In our implementation, the infinity norm $\|\mathbf{1}_n - Y\square \mathrm{J}F(I)\|_\infty$ is evaluated using interval arithmetic.

To simplify our language when talking about intervals $I \in \mathbb{IC}^n$ satisfying Theorem 3.3.6 we introduce the following definitions.

Definition 3.3.8. Let $F : \mathbb{C}^n \to \mathbb{C}^n$ be differentiable and $I \in \mathbb{IC}^n$. Let $K_{x,Y}(I)$ be the associated Krawczyk operator (see Definition 3.3.3). If there exists an invertible matrix $Y \in \mathbb{C}^{n\times n}$, such that $K_{x,Y}(I) \subset I$, we say that I is an *interval approximate zero* F. We call I a *strong interval approximate zero* of F if in addition $\sqrt{2}\|\mathbf{1}_n - Y\square \mathrm{J}F(I)\|_\infty < 1$.

Remark 3.3.9. The name "strong interval approximate zero" is not common in the field of interval arithmetic. We introduce it as a reference to the work of Shub and Smale and the software `alphaCertified` [HS12] that inspired our work. Shub and Smale coined the name *strong approximate zero* for points in the radius of quadratic convergence of Newton's method.

Definition 3.3.10. If I is an interval approximate zero, then, by Theorem 3.3.6, I contains a zero of F. We call such a zero an *associated zero* of I. If I is a strong interval approximate zero then there is a unique associated zero and we refer to it as *the* associated zero of I.

The notion of strong interval approximate zero is stronger than the definition suggests at first sight. We not only certify that a unique zero of F exists inside I, but even that we can approximate this zero with arbitrary precision. This is shown in the next proposition.

Proposition 3.3.11. *Let I be a strong interval approximate zero of F and let $x^* \in I$ be the unique zero of F inside I. Let $x \in I$ be any point in I. We define $x_0 := x$ and for all $i \geq 1$ we define the iterates $x_i := x_{i-1} - Y\,F(x_{i-1})$, where $Y \in \mathbb{C}^{n\times n}$ is the matrix from Definition 3.3.8. Then, the sequence $(x_i)_{i\geq 0}$ converges (at least linearly) to x^*.*

Proof. Define $G_Y(x) = x - YF(x)$. Let $(x_i)_{i\geq 0}$ be the sequence defined in the statement of the proposition. By assumption, $x_0 \in I$ and $x_{i+1} = G_Y(x_i)$, $i \geq 0$. Let $I \in \mathbb{IC}^n$ be an interval vector and $x \in I$. Then,

$$G_Y(I) \subset K_{x,Y}(I).$$

(see, e.g., [BLL19, Lemma 2].). This implies, using an induction argument, that $x_i \in I$ for all i. Furthermore, if x^* denotes the unique zero of F in I, we have

$$x^* - x_{i+1} = G_Y(x^*) - G_Y(x_i)$$

We also have (see, e.g., [BLL19, Lemma 2].)

$$G_Y(x^*) - G_Y(x_i) \in (\mathbf{1}_n - Y \cdot \square \mathrm{J}F(I))\mathrm{Re}(x^* - x_i) + (\mathbf{1}_n - Y \cdot \square \mathrm{J}F(I))i\mathrm{Im}(x^* - x_i).$$

(Note that we can't apply the distributivity law because of Theorem 3.2.1 3.). Applying norms and using submultiplicativity yields

$$\|x^* - x_{i+1}\|_\infty \leq \|(\mathbf{1}_n - Y \cdot \square \mathrm{J}F(I))\|_\infty \, (\|\mathrm{Re}(x^* - x_i)\|_\infty + \|\mathrm{Im}(x^* - x_i)\|_\infty).$$

Since $\|\mathrm{Re}(x^* - x_i)\|_\infty + \|\mathrm{Im}(x^* - x_i)\|_\infty \leq \sqrt{2}\|x^* - x_i\|_\infty$ it holds

$$\|x^* - x_{i+1}\|_\infty \leq \sqrt{2}\|\mathbf{1}_n - Y \cdot \square \mathrm{J}F(I)\|_\infty \|x^* - x_i\|_\infty. \tag{3.4}$$

We get $\|x^* - x_{i+1}\|_\infty < c \cdot \|x^* - x_i\|_\infty$ with a constant $c := \sqrt{2}\|\mathbf{1}_n - Y \cdot \square \mathrm{J}F(I)\|_\infty < 1$ by Theorem 3.3.6. This shows linear convergence. □

3.3.2 Certifying reality

For many applications only the real zeros of a polynomial system are of interest. Since numerical homotopy continuation computes in $\mathbb{C}^n$, it is important to have a rigorous method to determine whether a zero is real.

Recall from Definition 3.3.8 the notion of *strong interval approximate zero*.

Lemma 3.3.12. *Let $F : \mathbb{C}^n \to \mathbb{C}^n$ be a real square system of polynomials (or rational functions) and $I \in \mathbb{IC}^n$ a strong interval approximate zero of F. Then there exists $x \in I$ and $Y \in \mathbb{C}^{n\times n}$ satisfying $K_{x,Y}(I) \subset I$ and $\sqrt{2}\,\|\mathbf{1}_n - Y\square \mathrm{J}F(I)\|_\infty < 1$. If additionally $\{\bar{z} \,|\, z \in K_{x,Y}(I)\} \subset I$, the associated zero of I is real.*

Proof. Theorem 3.3.6 implies that F has a unique zero $s \in K_{x,Y}(I) \subset I$. The element-wise complex conjugate $\bar{s}$ is also a zero of F. If we have that $\bar{s} \in \{\bar{z} \,|\, z \in K_{x,Y}(I)\} \subset I$, then $\bar{s} = s$, since otherwise $\bar{s}$ and s would be two distinct zeros of F in I. This contradicts the uniqueness result from Theorem 3.3.6, finishing the proof. □

For a wide range of applications positive real zeros are of particular interest.

Corollary 3.3.13. *Let $F : \mathbb{C}^n \to \mathbb{C}^n$ be a real square system of polynomials and $I \in \mathbb{IC}^n$ a strong interval approximate zero of F satisfying the conditions of Lemma 3.3.12. If $\mathrm{Re}(I) > 0$ then the associated zero of I is real and positive.*

If the reality test in Lemma 3.3.12 fails for a strong interval approximate zero $I \in \mathbb{C}^n$ then this does not necessarily mean that the associated zero of I is not real. A sufficient condition that I is not real is that there is a coordinate such that the imaginary part of it does not contain zero.

Lemma 3.3.14. *Let $F(x)$ be a square system of polynomials or rational functions and let $I \in \mathbb{IC}^n$ be a strong interval approximate zero of F. If there exists $k \in \{1, \ldots, n\}$ such that $0 \notin \mathrm{Im}(I_k)$ then the associated zero of I is not real.*

Proof. The associated zero x of I is contained in I. Since $0 \notin \operatorname{Im}(I_k)$ it follows $x_k \notin \mathbb{R}$, $x \notin \mathbb{R}^n$. □

Now assume that the certification routine produced a list $\mathcal{I}$ of m distinct strong interval approximate zeros for a given system F, and that m also agrees with the theoretical upper bound on the number of isolated, nonsingular zeros of F. If we apply Lemma 3.3.12 to $I_k \in \mathcal{I}$, then we obtain only a *lower bound*, say r, on the number of real zeros of F. However, combined with Lemma 3.3.14 we can also obtain an *upper bound* of the number of real zeros. If these two bounds agree we obtain a certificate that, among the associated zeros of the intervals in $\mathcal{I}$, there are *exactly* r real zeros. An application of this is, e.g., the study of the distribution of the number of real solutions of the power flow equations [LZBL20].

3.4 Implementation details

In this section we describe details of our implementation of Krawcyzk's method.

The certification routine takes as input a square polynomial system $F : \mathbb{C}^n \to \mathbb{C}^n$ and a finite list $X \subset \mathbb{C}^n$ of (suspected) approximations of isolated nonsingular zeros of F. It is also possible to provide a square system of rational functions as input, but in the following we focus on polynomial systems for simplicity. Our implementation returns a list of strong interval approximate zeros $\mathcal{I} = \{I_1, \ldots, I_m\}$ in $\mathbb{IC}^n$, such that no two intervals I_k and I_ℓ, $k \neq \ell$, overlap. If two strong interval approximate zeros don't overlap then this implies that their associated zeros are distinct. Additionally, if F is a real polynomial system then for each $I_k \in \mathcal{I}$ it is determined whether its associated zero is real. In this chapter we take as an input for out certification routine approximations of all isolated nonsingular solutions $X \subset \mathbb{C}^n$ of F, as computed by numerical homotopy continuation methods as discussed in Section 3.1.2. This is a prototypical application. We do however want to emphasize that our method can be applied to any set of approximate solutions. In [BT21] for example the latter are numerical eigenvalues.

3.4.1 Interval enclosures for polynomial systems

The fact that distributivity does not hold in $\mathbb{IC}$ makes it necessary for us to define the polynomial system $F : \mathbb{C}^n \to \mathbb{C}^n$, and its interval enclosure $\square F$, by a straight-line program, and not just by a list of coefficients. The overestimation of the interval enclosure $\square F$ increases with the size of the straight-line program. Therefore, it is good to express F and its enclosure $\square F$ by the smallest straight-line program possible. To achieve this, `HomotopyContinuation.jl` automatically applies a multivariate version of Horner's rule to reduce the number of operations necessary to evaluate F and $\square F$.

Remark 3.4.1. Our implementation of interval enclosures can also be used to prove that a polynomial map $F : \mathbb{C}^n \to \mathbb{C}^m$ with real coefficients, evaluated at a real point $p \in \mathbb{R}^n$, is positive. To verify this, one takes an interval $I \in \mathbb{IC}^n$ of the form $I = J + i[0,0]^{\times n}$ such that $p \in J$. If $\square F$ is an interval enclosure of F, and if $\square F(I) \subset \mathbb{R}^m_{>0} + i[0,0]^{\times m}$, then this is a proof that $F(p) \in \mathbb{R}^m_{>0}$.

3.4.2 Machine interval arithmetic

In the next subsection we give a method to construct a candidate $I \in \mathbb{IC}^n$ for a strong interval approximate zero. Before, we need to study *machine interval arithmetic*; the realization of interval arithmetic with finite precision floating point arithmetic. We assume the standard model of floating point arithmetic [Hig02, Section 2.3], where the result of a floating point operation is accurate

up to relative unit roundoff u: $\mathrm{fl}(x \circ y) = (x \circ y)(1+\delta)$, where $|\delta| \leq u$ and $\circ \in \{+,-,*,/\}$. For instance, following the IEE-754 standard, the unit roundoff in double precision arithmetic is $u = 2^{-53} \approx 2.2 \cdot 10^{-16}$. The key property in the context of interval arithmetic is that each result of a floating point operation can be rounded outwards, such that the resulting *interval* contains the true (exact) result; see, e.g., [May17, Section 3.2]. Therefore, given $X, Y \in \mathbb{IC}$ the result of $X \circ Y$, $\circ \in \{+,-,*,/\}$, is $\mathrm{fl}(X \circ Y) := \{(x \circ y)(1+\delta) \,|\, |\delta| \leq u, x \in X, y \in Y\}$ in machine arithmetic. This interval contains $X \circ Y$. It is *larger*. Additionally, for a given $x \in \mathbb{IC}$, all intervals of the form $\{x + (|\mathrm{Re}(x_j)| + i|\mathrm{Im}(x_j)|)\delta \,|\, |\delta| \leq \mu\}$ with $0 < \mu \leq u$ are indistinguishable when working with precision u.

Consequently it is possible that the Krawczyk operator $K_{x,Y}$, see Definition 3.3.3, is a contraction for the interval I, but that machine arithmetic can't verify this, because $\mathrm{fl}(X \circ Y)$ is larger than $X \circ Y$. In such a case, the unit roundoff u needs to be sufficiently decreased. For this reason our implementation uses machine interval arithmetic based on double precision arithmetic as well as, if necessary, the arbitrary precision interval arithmetic implemented in `Arb` [Joh17]. For instance, we could not certify all solutions in the example in Section 3.5.1 below using only 64-bit arithmetic, because the zeros are too ill-posed.

3.4.3 Determining strong interval approximate zeros

In a first step, the certification routine attempts to produce for a given $x_0 \in X$ a strong interval approximate zero $I \in \mathbb{IC}^n$. Recall that for $I \in \mathbb{IC}^n$ to be a strong interval approximate zero we need by Theorem 3.3.6 to have a point $x \in I$, and a matrix $Y \in \mathbb{C}^{n \times n}$ such that $K_{x,Y}(I) \subset I$, and $\sqrt{2}\,\|\mathbf{1}_n - Y \square \mathrm{J}F(I)\|_\infty < 1$.

Given a point $x_0 \in X$ and a unit roundoff u, the point x_0 is refined using Newton's method to maximal accuracy. Let this refined point be x. Here, we assume that x_0 is already in the region of quadratic convergence of Newton's method. Next, the point x needs to be inflated to an interval I with $x \in I$. This process is called ε-inflation in the literature [May17, Sec. 4.3]. However, choosing the correct I is a hard problem: if I is too small or too large, then the Krawcyzk operator is not a contraction.

In spite of these difficulties, we found that the following heuristic to determine I works very well. First, we compute $JF(x)^{-1}$ in floating arithmetic, which yields a matrix Y. Then, we set

$$I := (x_j \pm |(Y \cdot \square F(x))_j| u^{-\frac{1}{4}})_{j=1,\ldots,n},$$

where u is the unit roundoff. The motivation behind this choice is as follows: If we assume x to be in the region of quadratic convergence of Newton's method, it follows from the Newton-Kantorovich theorem that $||\mathrm{J}F(x)^{-1}F(x)||_\infty$ is a good estimate of the distance between x and the convergence limit x^*. This distance is approximated by $(Y \cdot \square F(x))_j$ for $1 \leq j \leq n$. The factor $u^{-\frac{1}{4}}$ accounts for the overestimation by machine interval arithmetic. Here is how we arrived at this factor: The best relative accuracy we can expect to get for the j-th entry of x^* is about $|(x^*)_j| \cdot u$, so that $||I - x^*||_\infty$ needs to be larger than $(|(x^*)_j| u)^{-1/2}$ for quadratic convergence. On the other hand, we need to have an ε-inflation of at least $|(Y \cdot \square F(x))_j|$ so that the inflated interval contains $(x^*)_j$. In the typical case we have $|(Y \cdot \square F(x))_j| > 1$, i.e., $|(Y \cdot \square F(x))_j| > |(Y \cdot \square F(x))_j|^{\frac{1}{2}}$. All of this motivates us to use $|(Y \cdot \square F(x))_j| u^{-\frac{1}{2}}$ as the inflation constant. However, to account for hidden constant factors we need to increase this estimate. We found that replacing $u^{-\frac{1}{2}}$ by $u^{-\frac{1}{4}}$ produces a good estimate that works well in all the examples we tested.

Finally, if I doesn't satisfy the conditions in Theorem 3.3.6, then the procedure is repeated with a smaller unit roundoff u. This repeats until either a minimal unit roundoff is reached or the certification is successful.

3.4.4 Producing distinct intervals

Assume now that the steps in Section 3.4.3 have been performed for all $x \in X$. We obtain a list of strong interval approximate zeros $I_1, \ldots, I_r \in \mathbb{IC}^n$. In a final step we want to select a subset $M \subset \{1, \ldots, r\}$ such that for all $k, j \in M$, $k \neq j$, the intervals I_k and I_j do not overlap. If two strong interval approximate zeros do not overlap then it is guaranteed that they have distinct associated zeros. A simple approach to determine M is to compare all intervals pairwise. However, this approach requires us to perform $\binom{r}{2}$ interval vector comparisons. For larger problems this becomes prohibitively expensive: in the example in Section 3.5.2 the number of necessary comparisons is already larger than $4 \cdot 10^9$.

Instead, we employ the following improved scheme to determine all non-overlapping intervals. First, we pick a random point $q \in \mathbb{C}^n$ and compute in interval arithmetic for each I_k, $k \in M$, the squared Euclidean distance $d_k \in \mathbb{IR}$ between I_k and q. Due to the guarantees of interval arithmetic we have that d_k and d_ℓ overlap if I_k and I_ℓ overlap (but the converse it not necessarily true). Next, we check for all overlapping intervals $d_k, d_\ell \in \{d_k \in \mathbb{IR} \,|\, k = 1, \ldots, r\}$, whether I_k and I_ℓ overlap, and if so, we group them accordingly. This allows us to construct the set M by selecting those intervals which don't overlap with any other and by picking one representative of each cluster of overlapping intervals. The worst case complexity of this procedure still requires $O(r^2)$ operations, but in the common case where no or only a small number of intervals overlap $O(r \log r)$ operations are sufficient.

3.5 Applications

In this section we showcase example applications of our certification method. The first example is from enumerative geometry and demonstrates how our method can be used for rigorous proofs. The second example is an application from kinematics, which shows that our implementation can deal with large problems and that our strategy for producing distinct intervals from Section 3.4.4 is indispensable. This is underlined by the fact that with our computation we improve a result from the literature. Both examples emphasize the speed compared to the symbolic approaches, and they rely on the option to modulate the precision thanks to our usage of `Arb` [Joh17].

All reported timings were obtained on an desktop computer with a 3.4 GHz processor running `Julia` 1.5.2 and `HomotopyContinuation.jl` version 2.2.2.

3.5.1 3264 real conics

We demonstrate how certification methods in numerical algebraic geometry allow to prove theorems in algebraic geometry. This example furthermore reveals the superior speed of our implementation compared to `alphaCertified`.

In [BST20] `alphaCertified` was used to prove that a certain arrangement of five conics in the plane had 3264 real conics, which were simultaneously tangent to each of the five given conics. Such an arrangement is called *totally real*. It was known before that such arrangements exist [RTV97], but an explicit instance was not known. The fact that `alphaCertified` provides a proof for a totally real instance highlights the relevance of certification software in algebraic geometry.

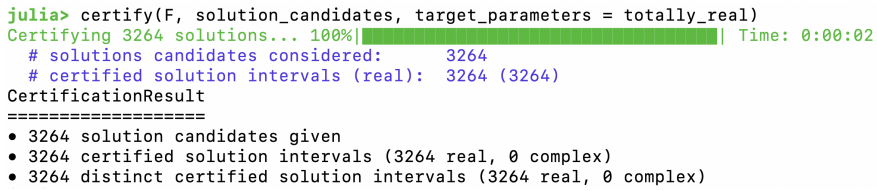


Figure 3.2: Screenshot from a `Julia` session, where we certify the 3264 real conics for the totally real arrangement from [BST20]. Here, `F` is the system of polynomials from (12) in [BST20]. The screenshot also demonstrates the simple syntax of our implementation.

The strategy for the computation is this. The zeros of the system (12) in [BST20] give the coordinates of the 3264 conics which are tangent to five given conics. We compute the zeros for the coordinates of the specific instance in [BST20, Figure 2] using `HomotopyContinuation.jl`. This is a numerical computation. Therefore, it is inexact and cannot be used in a proof. Next, we take the inexact numerical zeros as starting points for our certification method. If our implementation outputs that it has found a real certified zero, then this is an exact result and hence it is a proof that the zero is real. This way we can prove that indeed all the 3264 conics for the instance in [BST20, Figure 2] are real. See also the proof of [BST20, Proposition 1] for a more detailed discussion.

The certification with `alphaCertified` took us *more than 36 hours*. In contrast, our implementation certifies the reality and distinctness of the 3264 conics in *less than three seconds*.

3.5.2 Numerical Synthesis of Six-Bar Linkages

Now we demonstrate that the certification routine can cope with large problems. With our computation we improve a result from the literature.

We consider the kinematic synthesis of six-bar linkages that use eight prescribed accuracy points as described in [PM14]. In this chapter, the authors derive the synthesis equations for six-bar linkages of the Watt II, Stephenson II, and Stephenson III type. Additionally, in [PM14, Eq. (35)] they construct a system of 22 polynomials in 22 unknowns and 224 parameters, that can be used as a start system in a parameter homotopy to solve the synthesis equations of all three considered six-bar linkage types.

The number of non-singular zeros of this generalized start system is reported as 92,736. It was computed using `Bertini` and a multi-homogeneous start system. To certify the reported count, we solved the generalized start system using the monodromy method [DHJ+18] implementation in `HomotopyContinuation.jl`. In our computation we obtained 92,752 non-singular zeros for a generic choice of the 224 parameters. These are sixteen *more* than reported in [PM14]. We certified this count using our certification routine and obtained 92,752 distinct strong interval approximate zeros. Therefore, we have a certificate that the generalized system has in general (at least) 92,752 non-singular solution. This establishes that the result in [PM14] undercounts the true number of solutions. The certification needed only 38.34 seconds which underlines the scalability of the certification routine. Notice that the naive method for comparing intervals in Section 3.4.4 gives 4.301.420.376 pairs to check. This underlines the need for having an efficient algorithm for comparing pairs.

3.6 Conclusion

We reported on a novel implementation of Krawczyk's method. Based on interval arithmetic it is able to numerically certify isolated solutions to quadratic polynomial systems. It has already been employed by researchers in various instances, enabling new applications of numerical computation both within and outside of mathematics. Due to its speed, this certification method has now been made an automatic computation step after numerically solving polynomial systems in the software package `HomotopyContinuation.jl`.

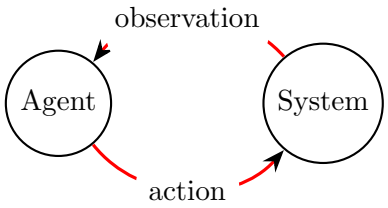
Chapter 4

Algebraic methods in decision processes

Solving sequential decision problems has a long-standing history in computer science, economics, mathematics, and statistics. A sequential decision problem is particularly challenging if only partial information about the true state of the system is available to the acting agent. In this chapter we study partially observable Markov decision processes and the optimization of their long-term reward. We contribute a new geometric formulation of this optimization problem, and show that it is equivalent to optimizing a linear objective subject to quadratic constraints. The feasible set of this problem is the positive part of a join of Segre varieties subject to linear constraints. We conduct experiments in which we solve the KKT equations or the Lagrange equations over different boundary components of the feasible set, and compare the result to other constrained optimization methods. Finally, we quantify the algebraic complexity of the optimization problem in many instances by computing polar degrees of the Zariski closure of the feasible set.

4.1 Introduction

Partially observable Markov decision processes (POMDPs) offer a mathematical framework for sequential decision-making under uncertainty. Their ability to incorporate nondeterministic effects of actions, and partial observability, makes them particularly well suited for modelling real world problems, but also highly complex. In the framework of POMDPs, an agent manipulates a system in a sequence of events. It selects an action at every time step and receives an instantaneous reward depending on the selected action and the current state of the system, which in turn influences the state at the next time step. However, the agent selects its actions based on observations that might not fully reveal the underlying state.



We study stochastic action selection mechanisms that do not depend on the prior history of observations but only on the current observation, known as memoryless policies. A common measure for the performance of a policy is the expectation of the instantaneous rewards, accumulated over time, and discounted into the future. We will refer to this measure simply as the reward function. Identifying a policy that maximizes the reward is challenging, since it is a nonconcave function that can exhibit non global strict local optima [BR19]. Indeed it has been shown that this optimization problem is NP-hard in general [VLB12]. A common approach is local optimization, such as policy gradient optimization [SMSM99, AYA18], an approach that has no global optimality guarantees. Whereas global optimality guarantees for gradient methods in fully observable systems (where the observation fully identifies the underlying state) have been given in [BR19], for general POMDPs such guarantees do not exist.

In this chapter we study POMDPs geometrically and express reward optimization as a linear program with polynomial constraints. That is, we are concerned with the optimization of a linear function over a nonconvex semialgebraic set. Our work builds on [MM22a]. We focus on the case of deterministic observations, and prove that the polynomial constraints define a join of Segre varieties (Theorem 4.5.4). By investigating the geometry of the semialgebraic set, we determine upper bounds on the number of (complex) critical points of the reward optimization problem, i.e., its algebraic degree (Theorem 4.6.3). We then provide a computational method that solves the optimization problem by computing the critical points via the Karush-Kuhn-Tucker conditions, whereby we identify ways to reduce the combinatorial complexity of the problem by focusing on relevant boundary components (Theorem 4.6.1, [MR17]). This approach is implemented using numerical algebra methods from Chapter 3 that automatically certify the correctness of the results. We further employ a convex relaxation of the polynomial problem to certify the global optimality of the results. Moreover, we observe that in specific instances this numerical algebraic approach leads to superior results than two commonly used optimization methods. Afterwards we compare the number of critical points obtained in numerical experiments with our theoretical bounds. These bounds to the algebraic complexity of the optimization problem are rather coarse and take into account only the (low) degrees of the objective and constraint. A more satisfactory description of the algebraic degree of the reward optimization problem of POMDPs will be given based on a study of the polar degrees of state aggregation varieties with Theorem 4.8.5 in Section 4.8. This addresses Question 1 from the Introduction of this Thesis for reward optimization. We compare Theorem 4.8.5 to the previous bounds from this chapter, namely Theorem 4.6.3, and the bounds from Chapter 2, namely Theorem 2.3.7 in Example 4.8.3 and Example 4.8.4. There where we also demonstrate that the bound from Theorem 4.6.3 are tight in some, but not all cases.

The chapter is organized as follows. The following section, Section 4.2 discusses previous work. Afterwards there is a short Prelude which we use to anticipate the geometric results from this chapter. In Section 4.4, we introduce partially observable Markov decision processes and related notation. In Section 4.5, we define the state aggregation variety and identify the feasible set and its defining (in)equalities for the reward optimization problem in POMDPs with its positive part. This is followed by Section 4.6, where we provide an upper bound on the number of complex critical points for the optimization problem we are considering. In Section 4.7, we use the description of reward maximization as a constrained polynomial optimization problem to solve the critical equations numerically. Finally, in Section 4.8, we compute polar degrees of state-aggregation varieties.

4.2 Previous work

Various approaches have been suggested to study the geometry underlying the optimization problem. A classic line of works has established that in the fully observable case, the optimization problem is equivalent to a linear program over a polytope of feasible state-action frequencies [Der70, Kal94]. These studies have been complemented by the characterization of the set of feasible value functions of a Markov decision process as a finite union of polytopes [DTLR+19, WKZ+22, WDL22]. However, for partially observable systems the geometry of the reward optimization problem is more complex. The problem can be formulated as a quadratically constrained linear program with the policy and the value function as search variables [ABZ06]. More recently, the set of feasible state-action frequencies was described as a union of convex sets in [MGZA15] and as a semialgebraic set in [MM22a], which also provided a method for computing the polynomial constraints. The possible advantages of taking this constrained optimization perspective in state-action space were recently studied in [MM22b] using interior point methods.

Related approaches have been proposed in other settings as well. In continuous time and space, a convex relaxation of linear quadratic control problems based on state-action frequencies has been proposed and studied in [LHPT08]. In [Ney03] the graphs of different stochastic games are described as semialgebraic sets, where (generalized) Nash equilibria, including a convex relaxation for their computation, have been studied with algebraic tools in [NT21, PS22].

4.3 Prelude

Before we formally define Markov decision processes we anticipate the geometric objects and the main geometric ideas of the following sections. This avoids heavy notation in the hope to make them accessible to the more geometrically inclined readers. We start off with an example of a POMDP that will guide us through the rest of the section.

Example 4.3.1. We imagine a parent interacting with their baby in a sequence of events. The baby could be in any of three states $s_1 = \texttt{happy}$, $s_2 = \texttt{neutral}$, $s_3 = \texttt{unhappy}$. The parent could observe that the baby is either $o_1 = \texttt{not crying}$ or $o_2 = \texttt{crying}$, where state s_1 and s_2 lead to observation o_1 and state s_3 leads to observation o_2. This encodes an uncertainty of the parent when assessing the needs of the baby. At each point in time the parent has two available actions to perform. It can $a_1 = \texttt{don't feed}$ and $a_2 = \texttt{feed}$ the baby. We consider the (deterministic) transition graph depicted in Figure 4.1, with associated matrix

$$\alpha = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \end{array} \begin{array}{c} \begin{array}{cccccc} s_1,a_1 & s_1,a_2 & s_2,a_1 & s_2,a_2 & s_3,a_1 & s_3,a_2 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \end{pmatrix} \end{array}. \tag{4.1}$$

If the baby is `neutral` then the action `feed` causes a transition to `happy`, similarly, if the baby is `unhappy`, then feeding it transitions to `neutral`. But if the baby is already `happy` then the same action causes a transition to `unhappy`, modelling overindulgence. This illustrates how partial observability of the state can make it difficult for the agent to choose the "right" action.

According to framework of POMDPs, the parent has to choose their actions at each time step, without memory, based only on its observation. We study the chain of events that emerges when at each time step the parent does not feed the baby with probability $0 \leq \pi_1 \leq 1$ if it is `not crying`,

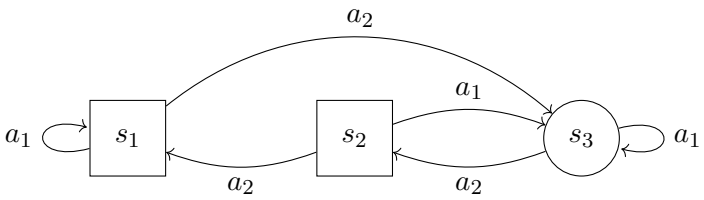


Figure 4.1: Transition graph of Example 4.3.1; states s_1, s_2 lead to observation o_1, and s_3 leads to observation o_2.

and decides to not feed the baby with probability $0 \leq \pi_2 \leq 1$ if it is `crying` for fixed values π_1, π_2. The pair $\pi = (\pi_1, \pi_2) \in [0,1]^2$ is called the policy of the agent.

In this thesis we measure the performance of a policy by capturing how the agent typically interacts with the system. More concretely, for each pair of state and action (s, a) we are interested in the expected value $\Phi(\pi)_{s,a} \coloneqq \lim_{N\to\infty} \frac{1}{N+1}\, \mathbb{E}\left[\sum_{t=0}^{N} \mathbb{P}(s_t = s, a_t = a)\right]$. Here we denote by $\mathbb{P}(s_t = s, a_t = a)$ the probability that at time step t the agent chooses action a in state s. The quantity $\Phi(\pi)_{s,a}$ records how frequent this event occurs. In our example the map $\Phi : [0,1]^2 \longrightarrow \mathbb{R}^{3\times 2}$ is given by:

$$\Phi(\pi)^T = \frac{\begin{pmatrix} \pi_1 - \pi_1\pi_2 + \pi_1\pi_2^2 - \pi_1^2 & \pi_1 - \pi_1\pi_2 + \pi_1\pi_2^2 - \pi_1^2 & \pi_2 - \pi_1\pi_2 \\ 1 - 2\pi_1 - \pi_2 + 2\pi_1\pi_2 - \pi_1\pi_2^2 + \pi_1^2 & 1 - 2\pi_1 - \pi_2 + 2\pi_1\pi_2 - \pi_1\pi_2^2 + \pi_1^2 & 1 - \pi_1 - \pi_2 + \pi_1\pi_2 \end{pmatrix}}{(3 - 3\pi_1 - 2\pi_2 + 2\pi_2\pi_1)}.$$

Remark 4.3.2. This definition of $\Phi(\pi)$ is the limit of the state action frequency $\Phi(\pi)$ from equation (4.5) below, for $\gamma \to 1$. We use this limit instead of the actual value of $\Phi(\pi)$, since it is very similar and simplifies the following equations. For now we call $\Phi(\pi)$ the state action frequency of π.

In this chapter we reward an agent at each time step, the reward depending only on the state of the system and the action of the agent. Given a policy π, we study the expected total reward and show that it is linear in the state action frequency $\Phi(\pi)$. Thus, we are interested in classifying the family $\Phi([0,1]^2)$ of feasible state action frequencies. We will see in Theorem 4.5.4 below that the entries of

$$\Phi(\pi)^T = \begin{array}{c} \\ a_1 \\ a_2 \end{array} \begin{array}{c} \begin{array}{ccc} s_1 & s_2 & s_3 \end{array} \\ \begin{pmatrix} \eta_{1,1} & \eta_{2,1} & \eta_{3,1} \\ \eta_{1,2} & \eta_{2,2} & \eta_{3,2} \end{pmatrix} \end{array}$$

exhibit algebraic relations coming from exactly two sources. On the one hand, the first two columns of $\Phi(\pi)^T$ are linearly dependent. This relation is a consequence of the states s_1 and s_2 being indistinguishable to the parent, which implies that the probability of choosing action a_1 in state s_1 is equal to the probability of choosing the action a_1 in state s_2, independent of the policy of the agent. We obtain the toric hypersurface $\mathcal{X} = \{0 = \eta_{1,1}\eta_{2,2} - \eta_{2,1}\eta_{1,2}\}$. On the other hand, the transition matrix α gives rise to a stationarity property of the state-action frequencies for every state s, encoded by the expression $l_s = \sum_a \eta_{sa} - \sum_{s',a'} \eta_{s'a'}\alpha_{s|s',a'}$. Here the quantity $\sum_a \eta_{sa}$ encodes how frequent the agent is in state s, while $\sum_{s',a'} \eta_{s'a'}\alpha_{s|s',a'}$ encodes how frequent the agent transitions from an arbitrary state to the state s.

We can now explicitly write down all polynomial relations that are satisfied by η. The image $\psi([0,1]^2)$ is the positive part of the intersection $\mathcal{X} \cap \mathcal{L}$, where $\mathcal{X}$ is the toric hypersurface from before,

and the affine linear space $\mathcal{L} = \{0 = l_{s_1} = l_{s_2} = l_{s_3} = 1 - \sum_{i,j} \eta_{i,j}\}$ is defined by polynomials, i.e.

$$l_{s_1} = \eta_{1,2} - \eta_{2,2},\ l_{s_2} = \eta_{2,1} + \eta_{2,2} - \eta_{3,2},\ l_{s_3} = -\eta_{1,2} - \eta_{2,1} + \eta_{3,2}.$$

We note that both $\mathcal{L}$ and $\mathcal{X}$ are symmetric under exchanging the states s_1 and s_2. It may not be immediate from the transition graph displayed in Figure 4.1, but ultimately the child will be in state s_1 equally as often as in state s_2, regardless of the choice of policy.

When computing an optimal policy, we have to decide which interactions between parent and child we want to penalize or reward. We might want to maximize the time that the baby is `happy`, leading to the linear optimization problem with reward function $\eta_{1,1} + \eta_{1,2}$ on the feasible set $\psi([0,1]^2)$. The number of critical points on the relative interior $\psi((0,1)^2)$ of the feasible set is bounded above by the number of complex critical points on the variety $\mathcal{X} \cap \mathcal{L}$. For a generic choice of α this bound is given by Theorem 4.8.1 in terms of the polar degree $2 = \delta_4(\overline{\mathcal{X}})$, and similar bounds can be obtained for each boundary component of $\psi([0,1]^2)$. In this example our specific choice of α is not general and we encounter an infinite family of singular critical points. In fact, every policy $(\pi_1, \pi_2) \in [0,1] \times \{0\}$ is optimal with objective value $\frac{1}{3}$.

Notation: For a finite set $\mathcal{S}$ we denote the free linear space over $\mathcal{S}$ by $\mathbb{R}^{\mathcal{S}} = \{f \colon \mathcal{S} \to \mathbb{R}\}$ and the simplex of probability distributions over $\mathcal{S}$ as $\Delta_{\mathcal{S}} = \{\mu \in \mathbb{R}^{\mathcal{S}} : \sum_s \mu_s = 1 \text{ and } \mu \geq 0\}$. For $s \in \mathcal{S}$ we denote the Dirac measure at s with $\delta_s \in \Delta_{\mathcal{S}}$ which corresponds to a unit vector. The *conditional probability polytope* consisting of all column-stochastic matrices[1] in $\mathbb{R}^{\mathcal{O} \times \mathcal{S}}$ is the product $\Delta_{\mathcal{O}}^{\mathcal{S}} = \Delta_{\mathcal{O}} \times \cdots \times \Delta_{\mathcal{O}}$. We call the elements of this set *conditional probability distributions* or *Markov kernels* from $\mathcal{S}$ to $\mathcal{O}$. Given a Markov kernel $Q \in \Delta_{\mathcal{O}}^{\mathcal{S}}$, the conditional probability $Q(o|s)$ is the entry Q_{os}. Note that a composition of Markov kernels is matrix multiplication. For a probability distribution $p \in \Delta_{\mathcal{S}}$ and a Markov kernel $Q \in \Delta_{\mathcal{O}}^{\mathcal{S}}$ we denote their composition into a joint probability distribution by $p * Q \in \Delta_{\mathcal{S} \times \mathcal{O}}$ and define it as $p * Q = \operatorname{diag}(p) Q^T$, that is, with entries $(p * Q)(s, o) \coloneqq p(s) Q(o|s)$. For a subset $A \subseteq \mathcal{S}$ we denote the complement $\mathcal{S} \setminus A$ of A in $\mathcal{S}$ by A^c.

4.4 Partially observable Markov decision processes

Partially observable Markov decision processes provide a powerful model to describe sequential decision making problems with state uncertainty.

Definition 4.4.1. A finite *partially observable Markov decision process* or shortly *POMDP* is a tuple $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \alpha, \beta, r)$, where $\mathcal{S}, \mathcal{O}$, and $\mathcal{A}$ are finite sets called the *state*, *observation*, and *action space* respectively and $\alpha \in \Delta_{\mathcal{S}}^{\mathcal{S} \times \mathcal{A}}$ and $\beta \in \Delta_{\mathcal{O}}^{\mathcal{S}}$ are Markov kernels, which we call the *transition* and *observation kernel* respectively. Furthermore, we consider an *instantaneous reward vector* $r \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$. We denote the cardinalities of $\mathcal{S}, \mathcal{A}$, and $\mathcal{O}$ by $n_{\mathcal{S}}, n_{\mathcal{A}}$, and $n_{\mathcal{O}}$.

From a modeling perspective, $\alpha(s'|s,a)$ is the probability of transitioning from state s to state s' upon taking action a, and $\beta(o|s)$ is the probability of making the observation o if the system is in state s. The entry r_{sa} corresponds to an instantaneous reward received upon selecting action a in state s.

[1] We choose to work with column-stochastic rather than row-stochastic matrices to have $Q_{os} = Q(o|s)$, which makes composition of two Markov kernels $Q_1 \circ Q_2$ equivalent to matrix multiplication $Q_1 Q_2$.

Remark 4.4.2. Example 4.3.1 is represented by the POMDP $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \alpha, \beta, r)$ with state space $\mathcal{S} = \{s_1, s_2, s_3\}$, observation space $\mathcal{O} = \{o_1, o_2\}$, and action space $\mathcal{A} = \{a_1, a_2\}$. The (deterministic) transitions kernel α is given by the incidence matrix of the directed graph represented in Figure 4.1, which is the column stochastic matrix presented in equation (4.1). The observation kernel is

$$\beta = \begin{array}{c} \\ o_1 \\ o_2 \end{array} \begin{array}{c} \begin{array}{ccc} s_1 & s_2 & s_3 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array} \in \Delta_{\mathcal{O}}^{\mathcal{S}}.$$

A *(memoryless stochastic) policy* is a column-stochastic matrix $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$, mapping from the set of observations to the set of actions. The entry $\pi(a|o)$ is the probability with which action $a \in \mathcal{A}$ is selected given the observation $o \in \mathcal{O}$. A policy can be interpreted as a randomized decision rule that encodes which action should be taken, based on the current observation. Every observation based policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ defines a state policy $\tau = \pi \circ \beta \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ according to

$$\tau(a|s) = \sum_{o \in \mathcal{O}} \pi(a|o)\beta(o|s). \tag{4.2}$$

A state policy $\tau \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ defines transition kernels $P_\tau \in \Delta_{\mathcal{S}\times\mathcal{A}}^{\mathcal{S}\times\mathcal{A}}$ and $p_\tau \in \Delta_{\mathcal{S}}^{\mathcal{S}}$ with entries

$$P_\tau(s', a'|s, a) \coloneqq \alpha(s'|s, a)\tau(a'|s') \quad \text{and} \quad p_\tau(s'|s) \coloneqq \sum_{a \in \mathcal{A}} \alpha(s'|s, a)\tau(a|s), \tag{4.3}$$

which we call the *state-action transition kernel* and the *state transition kernel* associated with τ and α. Given an initial distribution, the state-action transition kernel $P_\tau = P_{\pi\circ\beta}$ defines a Markov process on the state-action space $\mathcal{S} \times \mathcal{A}$.

One is particularly interested in the probability that the Markov process assigns to any given state-action pair, averaged over time, whereby it is convenient to discount events at larger times t by weighting them by $(1-\gamma)\gamma^t$ for a *discount factor* $\gamma \in (0, 1)$. Given an initial state distribution $\mu \in \Delta_{\mathcal{S}}$ and a discount factor $\gamma \in (0, 1)$, one thus defines the *(discounted) state-action frequency* associated with policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ as the following element of $\Delta_{\mathcal{S}\times\mathcal{A}}$:

$$\eta^\pi \coloneqq (1-\gamma) \sum_{t \geq 0} \gamma^t P_{\pi\circ\beta}^t (\mu * (\pi \circ \beta)) = (1-\gamma)(I - \gamma P_{\pi\circ\beta})^{-1}(\mu * (\pi \circ \beta)), \tag{4.4}$$

where I is the identity matrix; see [Der70, Kal94]. We denote the parametrization of η^π by

$$\begin{array}{rccl} \Phi\colon & \Delta_{\mathcal{A}}^{\mathcal{O}} & \to & \Delta_{\mathcal{S}\times\mathcal{A}} \\ f\colon & \pi & \mapsto & \eta^\pi = (1-\gamma)(I - \gamma P_{\pi\circ\beta})^{-1}(\mu * (\pi \circ \beta)). \end{array} \tag{4.5}$$

Elementary calculations show that for any given η^π the conditional probabilities of actions given states satisfy $\eta^\pi(a|s) = (\pi \circ \beta)(a|s)$. We denote the state-marginal of η^π by $\rho_s^\pi = \sum_{a \in \mathcal{A}} \eta_{sa}^\pi$ and refer to it as the *state frequency*. By the definition of conditional probability distributions it holds that

$$\eta_{sa}^\pi = \eta^\pi(a|s)\rho_s^\pi = (\pi \circ \beta)(a|s)\rho_s^\pi. \tag{4.6}$$

Finally, as a measure for the performance of policies, we introduce the *reward function*[2]:

$$R(\pi) \coloneqq \sum_{s \in \mathcal{S}, a \in \mathcal{A}} r_{sa}\Phi(\pi)_{sa} = \langle r, \Phi(\pi) \rangle_{\mathcal{S}\times\mathcal{A}}. \tag{4.7}$$

[2] More precisely, this is the infinite-horizon expected discounted reward function.

The reward function R is a widely used criterion to evaluate the performance of a policy. It is equal to the expected value $\mathbb{E}\left[(1-\gamma)\sum_{t\geq 0}\gamma^t r(s_t,a_t)\right]$ of the (discounted) accumulated instantaneous rewards along state-action trajectories distributed according to the Markov process with transition kernel $P_{\pi\circ\beta}$ and initial state-action distribution $\mu * (\pi\circ\beta)$. We refer to standard textbooks for an in-depth discussion [How60, Der70, Put14].

We consider the following **reward maximization problem**, which is the standard problem in (discounted) Markov decision processes:

$$\text{maximize } R(\pi) \quad \text{subject to } \pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}. \tag{4.8}$$

In this work, we focus on **deterministic observations** $\beta \in \Delta_{\mathcal{O}}^{\mathcal{S}} \cap \{0,1\}^{\mathcal{O}\times\mathcal{S}}$, where we can identify the observation kernel with a deterministic mapping $g_\beta\colon \mathcal{S}\to\mathcal{O}$. We denote the fibers of g_β by $S_o \coloneqq \{s\in\mathcal{S} : g_\beta(s)=o\}$ and their cardinality by $d_o = |S_o|$. Note that the fibers S_o are a disjoint partition of the states $\mathcal{S}$ and hence $(d_o)_{o\in\mathcal{O}}$ is a partition of $n_{\mathcal{S}}$, i.e., $\sum_{o\in\mathcal{O}} d_o = n_{\mathcal{S}}$. This special type of partial observability is known in the literature as **state-aggregation**.

Example 4.4.3. Consider again Example 4.3.1 from above. The policies of the agent are encoded as stochastic matrices

$$\pi = \begin{array}{c} \\ a_1 \\ a_2 \end{array}\!\!\begin{array}{c} \begin{array}{cc} o_1 & o_2 \end{array} \\ \begin{pmatrix} \pi_{a_1o_1} & \pi_{a_1o_2} \\ \pi_{a_2o_1} & \pi_{a_2o_2} \end{pmatrix} \end{array} \in \Delta_{\mathcal{A}}^{\mathcal{O}}.$$

For the given observation kernel β, the state policies (4.2) take the form

$$\tau = \pi\circ\beta = \begin{array}{c} \\ a_1 \\ a_2 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} s_1 & s_2 & s_3 \end{array} \\ \begin{pmatrix} \pi_{a_1o_1} & \pi_{a_1o_1} & \pi_{a_1o_2} \\ \pi_{a_2o_1} & \pi_{a_2o_1} & \pi_{a_2o_2} \end{pmatrix} \end{array} \in \Delta_{\mathcal{A}}^{\mathcal{S}}.$$

The state-action transition kernel (4.3) associated to α and τ is given by

$$P_\tau = \begin{array}{c} \\ s_1,a_1 \\ s_1,a_2 \\ s_2,a_1 \\ s_2,a_2 \\ s_3,a_1 \\ s_3,a_2 \end{array}\!\!\begin{array}{c} \begin{array}{cccccc} s_1,a_1 & s_1,a_2 & s_2,a_1 & s_2,a_2 & s_3,a_1 & s_3,a_2 \end{array} \\ \begin{pmatrix} \pi_{a_1o_1} & 0 & 0 & \pi_{a_1o_1} & 0 & 0 \\ \pi_{a_2o_1} & 0 & 0 & \pi_{a_2o_1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \pi_{a_1o_1} \\ 0 & 0 & 0 & 0 & 0 & \pi_{a_2o_1} \\ 0 & \pi_{a_1o_2} & \pi_{a_1o_2} & 0 & \pi_{a_1o_2} & 0 \\ 0 & \pi_{a_2o_2} & \pi_{a_2o_2} & 0 & \pi_{a_2o_2} & 0 \end{pmatrix} \end{array} \in \Delta_{\mathcal{S}\times\mathcal{A}}^{\mathcal{S}\times\mathcal{A}},$$

and the state transition kernel is given by

$$p_\tau = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} s_1 & s_2 & s_3 \end{array} \\ \begin{pmatrix} \pi_{a_1o_1} & \pi_{a_2o_1} & 0 \\ 0 & 0 & \pi_{a_2o_2} \\ \pi_{a_2o_1} & \pi_{a_1o_1} & \pi_{a_1o_2} \end{pmatrix} \end{array} \in \Delta_{\mathcal{S}}^{\mathcal{S}}.$$

Further, we consider a uniform initial distribution $\mu\in\Delta_{\mathcal{S}}$ and a discount factor $\gamma = 1/2$. Finally, let us assume the instantaneous reward vector is $r(s,a) = \delta_{s_1 s}$, which corresponds to a reward of $+1$ obtained in state s_1. Combining the Neumann series with Cramer's rule (see [MM22a]) one

sees that the reward function R is a rational function with the explicit expression $R(\pi) = \frac{f(\pi)}{2g(\pi)} - \frac{1}{2}$, where f and g are determinantal polynomials given by

$$\begin{aligned} f(\pi) &= 24\det(I - \gamma p_{\pi\circ\beta} + \mu\delta_{s_1 s}^T) \\ &= \pi_{a_1o_1}^2\pi_{a_2o_2} - 2\pi_{a_1o_1}\pi_{a_2o_1}\pi_{a_1o_2} - 2\pi_{a_2o_1}^2\pi_{a_1o_2} - \pi_{a_2o_1}^2\pi_{a_2o_2} + 4\pi_{a_1o_1}\pi_{a_2o_1} \\ &+ 2\pi_{a_1o_1}\pi_{a_1o_2} - 6\pi_{a_1o_1}\pi_{a_2o_2} + 4\pi_{a_2o_1}^2 - 4\pi_{a_2o_1}\pi_{a_1o_2} - 4\pi_{a_1o_1} + 8\pi_{a_2o_1} - 12\pi_{a_1o_2} + 24 \end{aligned} \tag{4.9}$$

and

$$\begin{aligned} g(\pi) &= 24\det(I - \gamma p_{\pi\circ\beta}) \\ &= 3\pi_{a_1o_1}^2\pi_{a_2o_2} - 3\pi_{a_2o_1}^2\pi_{a_2o_2} + 6\pi_{a_1o_1}\pi_{a_1o_2} - 6\pi_{a_1o_1}\pi_{a_2o_2} - 12\pi_{a_1o_1} - 12\pi_{a_1o_2} + 24. \end{aligned} \tag{4.10}$$

The reward function is to be optimized over the observation policy, that is, we have

$$\text{maximize } R(\pi) \quad \text{subject to } \begin{cases} \pi_{oa} \geq 0 & \text{for all } o \in \mathcal{O}, a \in \mathcal{A}, \\ \sum_{a\in\mathcal{A}} \pi_{oa} = 1 & \text{for all } o \in \mathcal{O}. \end{cases} \tag{4.11}$$

4.5 The geometry of reward optimization

In this section, we discuss the formulation of the reward optimization problem as a polynomially constrained linear program from [MM22a]. For deterministic observations we provide a new description of the feasible state-action frequencies as the intersection of a product of varieties of rank-one matrices, an affine space, and the simplex (see Theorem 4.5.4).

Clearly, optimizing $R(\pi) = \langle r, \Phi(\pi)\rangle$ over $\Delta_{\mathcal{A}}^{\mathcal{O}}$ is equivalent to the **reward maximization problem in the state-action space**:

$$\text{maximize } \langle r, \eta\rangle \quad \text{subject to } \eta \in \Phi(\Delta_{\mathcal{A}}^{\mathcal{O}}). \tag{4.12}$$

By definition, the feasible set $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$ is a subset of the probability simplex $\Delta_{\mathcal{S}\times\mathcal{A}}$. Cramer's rule implies that the parametrization Φ is a rational map and hence, by the Tarski-Seidenberg theorem, the range $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$ is semialgebraic. Next we discuss the solution of the **implicitization problem** for the parametric set $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$, i.e., a representation of this set as the solution set to a list of polynomial (in)equalities.

The mapping Φ can be seen as a composition $\Psi \circ f_\beta$ of a linear and non-linear map, illustrated in Figure 4.2 for the POMDP of our running example (Examples 4.3.1, 4.4.3, and 4.5.7), with

$$f_\beta\colon \begin{array}{ccc} \Delta_{\mathcal{A}}^{\mathcal{O}} & \longrightarrow & \Delta_{\mathcal{A}}^{\mathcal{S}} \\ \pi & \longmapsto & \tau = \pi\circ\beta \end{array} \quad \text{and} \quad \Psi\colon \begin{array}{ccc} \Delta_{\mathcal{A}}^{\mathcal{S}} & \longrightarrow & \Delta_{\mathcal{S}\times\mathcal{A}} \\ \tau & \longmapsto & \eta = (1-\gamma)(I - \gamma P_\tau)^{-1}(\mu * \tau). \end{array}$$

We recall the following classic result.

Proposition 4.5.1 (The state-action polytope of Markov decision processes, [Der70])**.** *The image $\Psi(\Delta_{\mathcal{A}}^{\mathcal{S}})$ is a polytope given by $\Psi(\Delta_{\mathcal{A}}^{\mathcal{S}}) = \mathcal{L} \cap \mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}} \subseteq \Delta_{\mathcal{S}\times\mathcal{A}}$, where*

$$\mathcal{L} = \left\{\eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} :\ \ell_s(\eta) = 0 \text{ for all } s \in \mathcal{S}\right\}, \tag{4.13}$$

and $\ell_s(\eta) \coloneqq \sum_a \eta_{sa} - \gamma\sum_{s',a'} \eta_{s'a'}\alpha(s|s',a') - (1-\gamma)\mu_s$.

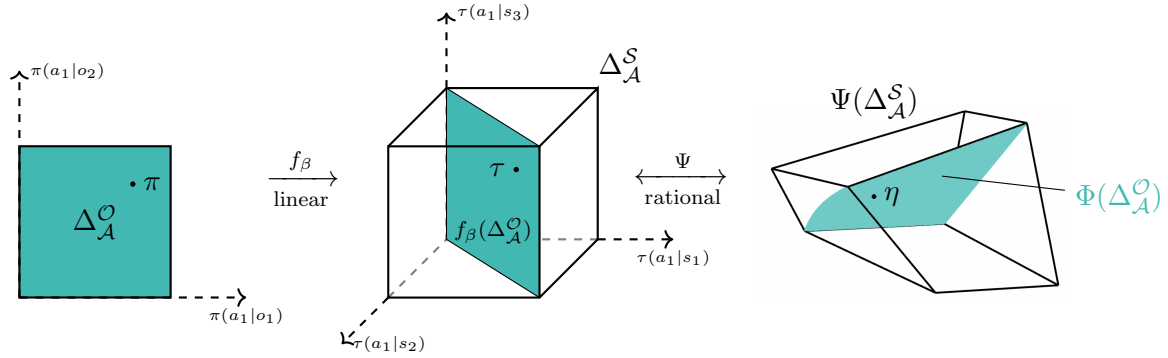


Figure 4.2: Shown is the policy polytope $\Delta_{\mathcal{A}}^{\mathcal{O}}$ (left), the effective policy polytope $f_\beta(\Delta_{\mathcal{A}}^{\mathcal{O}})$ within the state policy polytope $\Delta_{\mathcal{A}}^{\mathcal{S}}$ (middle), and the set of feasible state-action frequencies $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$ within the state-action polytope $\Psi(\Delta_{\mathcal{A}}^{\mathcal{S}})$ (right). This figure illustrates concretely our running example (Examples 4.3.1, 4.4.3, and 4.5.7); $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$ is the nonlinear solution set of the constraints given in equation (4.17) and $\Psi(\Delta_{\mathcal{A}}^{\mathcal{S}})$ is a three-dimensional polytope that is combinatorially equivalent to the cube $\Delta_{\mathcal{A}}^{\mathcal{S}}$.

In particular, the set of state-action frequencies $\Psi(\Delta_{\mathcal{A}}^{\mathcal{S}})$ of a fully observable Markov decision process forms a polytope, referred to as the *state-action polytope*. The constraints encoded in $\mathcal{L}$ describe a generalized stationarity property of the state-action frequencies, recovering stationarity in the limit where the discount factor is $\gamma = 1$. In order to relate the space of state policies to state-action frequencies, we make the following assumption.

Assumption 4.5.2 (Positivity)**.** *For every $s \in \mathcal{S}$ and $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$, we assume that $\sum_a \eta_{sa} > 0$.*

This assumption is satisfied, for example, if the system is ergodic or if the initial distribution $\mu \in \Delta_{\mathcal{S}}$ has full support, i.e., has only strictly positive entries. This can be seen by interpreting $\sum_a \eta_{sa}$ as a weighted average of the time spent in state s when following the policy π. An important consequence of this assumption is that the state policies τ and the state-action frequencies η are in one-to-one correspondence, whereby the state policies can easily be computed from the state-action frequencies by conditioning.

Proposition 4.5.3 ([MM22a])**.** *Under Assumption 4.5.2, the mapping $\Psi\colon \Delta_{\mathcal{A}}^{\mathcal{S}} \to \Psi(\Delta_{\mathcal{A}}^{\mathcal{S}})$ is rational and bijective with rational inverse given by conditioning*

$$\begin{aligned} \Gamma\colon \Psi(\Delta_{\mathcal{A}}^{\mathcal{S}}) &\longrightarrow \Delta_{\mathcal{A}}^{\mathcal{S}} \\ \eta &\longmapsto \tau, \quad \text{where } \tau_{as} = \frac{\eta_{sa}}{\sum_{a'} \eta_{sa'}}. \end{aligned}$$

The mapping Ψ extends to a rational bijection with rational inverse between open dense subsets of the affine span $\operatorname{affine}(\Delta_{\mathcal{A}}^{\mathcal{S}}) = \{\tau \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \sum_a \tau_{as} = 1 \text{ for } s \in \mathcal{S}\}$ *of $\Delta_{\mathcal{A}}^{\mathcal{S}}$ and $\mathcal{L}$ given in (4.13).*

The function Ψ is defined everywhere on $\Delta_{\mathcal{A}}^{\mathcal{S}}$ and bijectively identifies the defining inequalities of the polytope $f_\beta(\Delta_{\mathcal{A}}^{\mathcal{O}})$ within $\Delta_{\mathcal{A}}^{\mathcal{S}}$ with the defining inequalities of $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$ within $\Psi(\Delta_{\mathcal{A}}^{\mathcal{S}})$ via the pullback along Γ. This relates the geometry of $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$ and $f_\beta(\Delta_{\mathcal{A}}^{\mathcal{O}})$. The defining inequalities of $f_\beta(\Delta_{\mathcal{A}}^{\mathcal{O}})$ can be computed algorithmically, see e.g., [JKM04]. As we demonstrate in what follows, for deterministic observations the defining inequalities can be given in closed form. In particular, we show the following:

Theorem 4.5.4 (Feasible state-action frequencies)**.** *Let Assumption 4.5.2 hold. For deterministic observation β, the set of feasible state-action frequencies $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$ is the intersection $\mathcal{L} \cap \mathcal{X} \cap \Delta_{\mathcal{S}\times\mathcal{A}}$ of the linear space $\mathcal{L}$ defined in (4.13), the product of real determinantal varieties*

$$\mathcal{X} = \Big\{\eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \ \eta_{sa}\eta_{s'a'} - \eta_{sa'}\eta_{s'a} = 0 \ \forall a, a' \in \mathcal{A} \text{ and } s, s' \in \mathcal{S} \text{ with } g_\beta(s) = g_\beta(s')\Big\}, \tag{4.14}$$

and the probability simplex $\Delta_{\mathcal{S}\times\mathcal{A}}$. In particular, the only inequalities are of the form $\eta \geq 0$.

We call $\mathcal{L} \cap \mathcal{X}$ the *state aggregation variety*. Note that $\mathcal{X}$ is determined by the condition that for every observation o the $d_o \times n_{\mathcal{A}}$ submatrix $(\eta_{sa})_{s\in S_o,\, a\in\mathcal{A}}$ of η, consisting of all entries η_{sa} with $\beta(s) = o$, has rank one. In particular, the projective variety associated to $\mathcal{X}$ is a join of Segre varieties.

Proof of Theorem 4.5.4. Proposition 4.5.1 provides a description of the polytope $\Psi(\Delta_{\mathcal{A}}^{\mathcal{S}})$ as the intersection $\mathcal{L}\cap\Delta_{\mathcal{S}\times\mathcal{A}}$ so we are left with finding the defining equations and inequalities for $\Psi(f_\beta(\Delta_{\mathcal{A}}^{\mathcal{O}}))$ in $\Psi(\Delta_{\mathcal{A}}^{\mathcal{S}})$. To do this, observe that the polytope $f_\beta(\Delta_{\mathcal{A}}^{\mathcal{O}})$ consists of those elements $\tau \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ satisfying the linear equations $\tau_{as} - \tau_{as'}$ for all $a \in \mathcal{A}$, $s, s' \in S$ such that $g_\beta(s) = g_\beta(s')$. In particular, the description of $f_\beta(\Delta_{\mathcal{A}}^{\mathcal{O}})$ in $\Delta_{\mathcal{A}}^{\mathcal{S}}$ does not entail any inequalities. In other words, the only requirement is that all columns of τ indexed by states with equal observations coincide. Fix an action $a_o \in \mathcal{A}$ and a state $s_o \in S_o$ for each observation $o \in \mathcal{O}$. Then the non-redundant defining equalities of $f_\beta(\Delta_{\mathcal{A}}^{\mathcal{O}})$ are given by

$$l_{sa}^o(\tau) \coloneqq \tau_{as} - \tau_{as_o} = 0,$$

for all observations $o \in \mathcal{O}$, actions $a \in \mathcal{A} \setminus \{a_o\}$, and states in the fiber $s \in S_o \setminus \{s_o\}$. These equations determine the range of f_β as a function $\mathbb{R}^{\mathcal{A}\times\mathcal{O}} \to \mathbb{R}^{\mathcal{A}\times\mathcal{S}}$ (corresponding to the set $\mathcal{U}$ in [MM22a, Theorem 12]). Using the positivity Assumption 4.5.2 we can apply the pullback Γ^* of the conditioning map Γ to the linear functions l_{sa}^o and obtain the rational equations

$$(\Gamma^* l_{sa}^o)(\eta) = l_{sa}^o(\Gamma(\eta)) = \eta_{sa}\left(\sum_{a'\in\mathcal{A}} \eta_{sa'}\right)^{-1} - \eta_{s_oa}\left(\sum_{a'\in\mathcal{A}} \eta_{s_oa'}\right)^{-1} = 0, \tag{4.15}$$

which we rephrase as the vanishing of the polynomials

$$p_{sa}^o(\eta) \coloneqq \eta_{sa}\sum_{a'\in\mathcal{A}} \eta_{s_oa'} - \eta_{s_oa}\sum_{a'\in\mathcal{A}} \eta_{sa'} = \sum_{a'\in\mathcal{A}\setminus\{a\}} (\eta_{sa}\eta_{s_oa'} - \eta_{s_oa}\eta_{sa'}). \tag{4.16}$$

These are defining polynomial equations of $\Psi(f_\beta(\Delta_{\mathcal{A}}^{\mathcal{O}}))$ in $\Psi(\Delta_{\mathcal{A}}^{\mathcal{S}})$. Let now $\mathcal{W}$ be the variety determined by the equations (4.15). It remains to show $\mathcal{L}\cap\mathcal{X}\cap\Delta_{\mathcal{S}\times\mathcal{A}} = \mathcal{L}\cap\mathcal{W}\cap\Delta_{\mathcal{S}\times\mathcal{A}}$. Since p_{sa}^o is a linear combination of 2×2 minors, we have the inclusion $\mathcal{X} \subseteq \mathcal{W}$. On the other hand, equation (4.15) implies the linear dependence of the two vectors

$$(\eta_{sa})_a,\ (\eta_{s_oa})_a \in \mathbb{R}^{\mathcal{A}}$$

for every observation o and state $s \in S_o$. Consequently, every 2×2 minor in the definition of $\mathcal{X}$ vanishes on $\mathcal{L}\cap\mathcal{W}\cap\Delta_{\mathcal{S}\times\mathcal{A}}$. This shows the desired inclusion

$$\mathcal{L}\cap\mathcal{W}\cap\Delta_{\mathcal{S}\times\mathcal{A}} \subseteq \mathcal{L}\cap\mathcal{X}\cap\Delta_{\mathcal{S}\times\mathcal{A}},$$

which finishes the proof. □

Hence by Theorem 4.5.4, in the case of a deterministic observation kernel, the set $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$ is described semi-algebraically by linear inequalities and equalities which are either linear or 2×2 principal minors. This is in contrast to the case of general observation kernels, where nonlinear defining inequalities appear and the polynomial constraints might be of higher degree (see [MM22a, Theorem 16]). Since all defining equalities of $\mathcal{X}$ are binomial, it is a toric variety. The following monomial parametrization of $\mathcal{X}$ can be inferred from the discussion of the family of state-frequencies and equation (4.6):

$$\begin{aligned} \mathbb{R}^{\mathcal{S}} \times \mathbb{R}^{\mathcal{A}\times\mathcal{O}} &\longrightarrow \mathcal{X} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \\ (\rho, \pi) &\longmapsto \eta, \quad \text{where } \eta(s,a) = \pi(a|g_\beta(s))\rho(s). \end{aligned}$$

The subsequent characterization of the set of feasible state-action frequencies with fewer equations will be useful later.

Corollary 4.5.5 (Alternative characterization of feasible state-action frequencies)**.** *For a deterministic observation β, fix an arbitrary action $a_o \in \mathcal{A}$ and an arbitrary state $s_o \in S_o$ for every $o \in \mathcal{O}$. The set of feasible state-action frequencies $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$ can be described as the intersection $\mathcal{L} \cap \mathcal{Y} \cap \Delta_{\mathcal{S}\times\mathcal{A}}$, where*

$$\mathcal{Y} = \left\{\eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \ p_{sa}^{o}(\eta) = 0 \text{ for all } o \in \mathcal{O}, a \in \mathcal{A} \setminus \{a_o\}, s \in S_o \setminus \{s_o\}\right\}.$$

The polynomials p_{sa}^{o} are given in (4.16), and $\mathcal{Y}$ is a complete intersection of their hypersurfaces.

Proof. This follows directly from the proof of Theorem 4.5.4. □

Remark 4.5.6. We clarify that the description of $\mathcal{Y}$ as a complete intersection in general does not hold for the state aggregation variety $\mathcal{X} \cap \mathcal{L}$. The variety $\mathcal{Y}$ might have additional components, possibly of higher dimension than the state aggreggation variety. By Theorem 4.5.4 these are disjoint from the feasible set $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$. The description of $\mathcal{Y}$ as a complete intersection will be exploited in the next sections where we employ numerical methods to compute critical points by solving Lagrange and KKT equations.

Example 4.5.7. We continue Example 4.3.1 and 4.4.3 from above. The defining equalities of $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$, described in (4.13) and Theorem 4.5.4, take the form

$$\begin{aligned} \mathcal{L} = \{\eta \in \mathbb{R}^{3\times 2} : 3\eta_{s_1a_1} + 6\eta_{s_1a_2} - 3\eta_{s_2a_2} - 1 = 0,\ 6\eta_{s_2a_1} + 6\eta_{s_2a_2} - 3\eta_{s_3a_2} - 1 = 0, \\ 3\eta_{s_3a_1} + 6\eta_{s_3a_2} - 3\eta_{s_1a_2} - 3\eta_{s_2a_1} - 1 = 0\} \end{aligned}$$

and

$$\mathcal{X} = \left\{\eta \in \mathbb{R}^{3\times 2} : \ \eta_{s_1a_1}\eta_{s_2a_2} - \eta_{s_1a_2}\eta_{s_2a_1} = 0\right\},$$

and the defining inequalities are $\eta_{sa} \geq 0$. Thus, the reward maximization problem (4.12) is

$$\text{maximize } \eta_{s_1a_1} + \eta_{s_1a_2} \quad \text{subject to} \begin{cases} 3\eta_{s_1a_1} + 6\eta_{s_1a_2} - 3\eta_{s_2a_2} - 1 = 0 \\ 6\eta_{s_2a_1} + 6\eta_{s_2a_2} - 3\eta_{s_3a_2} - 1 = 0 \\ 3\eta_{s_3a_1} + 6\eta_{s_3a_2} - 3\eta_{s_1a_2} - 3\eta_{s_2a_1} - 1 = 0 \\ \eta_{s_1a_1}\eta_{s_2a_2} - \eta_{s_1a_2}\eta_{s_2a_1} = 0 \\ \eta_{s_1a_1}, \eta_{s_1a_2}, \eta_{s_2a_1}, \eta_{s_2a_2}, \eta_{s_3a_1}, \eta_{s_3a_2} \geq 0. \end{cases} \tag{4.17}$$

The feasible set of this optimization problem is shown on the right in Figure 4.2. Comparing this to the optimization problem over the policy polytope (4.11) with objective function (4.9), now the constraints are more complex and nonlinear but the objective is linear.

4.6 Combinatorial and algebraic complexity of the problem

In this section, we study the number of critical points of the reward optimization problem in the case of deterministic observations. A similar approach was pursued in [MM22a] for the case of invertible observation matrix β, in which case there are linear equations and polynomial inequalities.

The description of $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$ obtained in Corollary 4.5.5 allows to reformulate the reward maximization problem (4.12) as the following constrained polynomial optimization problem:

$$\text{maximize } \langle r, \eta \rangle \quad \text{subject to } \begin{cases} \ell_s(\eta) = 0 & \text{for } s \in \mathcal{S}, \\ p^o_{sa}(\eta) = 0 & \text{for } o \in \mathcal{O}, a \in \mathcal{A} \setminus \{a_o\}, s \in S_o \setminus \{s_o\}, \\ \eta_{sa} \geq 0 & \text{for } s \in \mathcal{S}, a \in \mathcal{A}, \end{cases} \tag{4.18}$$

where the linear constraints ℓ_s are given in Proposition 4.5.1, the polynomial constraints $p^o_{sa}(\eta)$ are provided in (4.16) taking a fixed action $a_o \in \mathcal{A}$ and a fixed state $s_o \in S_o$ for each observation $o \in \mathcal{O}$, and the inequality constraints simply ensure the entries of η being nonnegative. Observe that problem (4.18) is in fact a quadratically constrained linear program.

We bound the number of critical points individually for each boundary component of the feasible set. A boundary component consists of all feasible points for which a given subset of the inequality constraints are active. The boundary components of the feasible set $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$ are in one-to-one correspondence with the faces of $\Delta_{\mathcal{A}}^{\mathcal{O}}$ according to

$$\left\{\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}} : \ \pi(a|o) = 0 \ \forall a \in A_o, o \in \mathcal{O}\right\} \longleftrightarrow \left\{\eta \in \Phi(\Delta_{\mathcal{A}}^{\mathcal{O}}) : \ \eta_{sa} = 0 \text{ for } a \in A_{g_\beta(s)}\right\}, \tag{4.19}$$

where A_o is a proper subset of $\mathcal{A}$ for every $o \in \mathcal{O}$, and $g_\beta(s)$ is the observation associated with state s. In particular, there is a boundary component associated to each tuple $(A_o)_{o\in\mathcal{O}}$ with $A_o \subsetneq \mathcal{A}$, $o \in \mathcal{O}$.

We point out the following result, which allows us to ignore high-dimensional boundary components when searching for a maximizer of the reward. Recall that for an observation $o \in \mathcal{O}$, the cardinalities of the fibers of g_β are denoted by $d_o = |S_o|$.

Theorem 4.6.1 (Existence of maximizers in low dimensional faces, [MR17])**.** *There exist $A_o \subsetneq \mathcal{A}$ with $|A_o^c| \leq d_o$, $o \in \mathcal{O}$, such that the set B described in (4.20) contains a (globally optimal) solution of the problem (4.18).*

Remark 4.6.2. One approach to solving (4.18) is to solve the critical equations over every boundary component and then selecting the critical point with the highest objective value. According to Theorem 4.6.1 there is a lower-dimensional boundary component that contains a global maximizer. This implies that, instead of considering the critical points in all $(2^{n_\mathcal{A}}-1)^{n_\mathcal{O}}$ boundary components, it is enough to consider those in the boundary components with $A_o \subsetneq \mathcal{A}$ satisfying $|A_o| \geq n_\mathcal{A} - d_o$. This reduces the number of boundary components that need to be checked to

$$\prod_{o\in\mathcal{O}} \left(\sum_{k_o=\max(n_\mathcal{A}-d_o,0)}^{n_\mathcal{A}-1} \binom{n_\mathcal{A}}{k_o} \right),$$

which we call *relevant* boundary components. Note that this number only depends on the number of actions $n_\mathcal{A}$ and d_o (the cardinality of the fibers of g_β).

Bounds via algebraic degrees of polynomial optimization

With the description of the boundary components of the feasible set at hand, we can deduce upper bounds on the number of critical points over each of them based on the degrees of the defining equations and the degree of the objective function.

Theorem 4.6.3 (Bound on the algebraic degree)**.** *Consider a POMDP with deterministic observations. Fix $A_o \subsetneq \mathcal{A}$ for every $o \in \mathcal{O}$ and set $n := n_{\mathcal{S}} n_{\mathcal{A}} - n_{\mathcal{S}} - \sum_o d_o |A_o|$ and $m := \sum_o (d_o - 1)(|A_o^c| - 1)$, where we assume n is not zero. Then the number of isolated critical points of the linear function $\eta \mapsto \langle r, \eta \rangle$ over*

$$B = \{\eta \in \mathcal{L} \cap \mathcal{X} : \ \eta_{sa} = 0 \text{ for } a \in A_{g_\beta(s)}\} \tag{4.20}$$

is upper bounded by $2^m \binom{n-1}{m-1}$.

Proof. We now write the Zariski closure of B as a complete intersection of m quadratic equations on an $n-$dimensional affine-linear space. Recall from Corollary 4.5.5 that $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$ is defined in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ as an intersection of $n_{\mathcal{S}}$ linear equations, $\sum_o (d_o - 1)(n_{\mathcal{A}} - 1)$ quadratic equations of the form (4.16), and the linear inequalities $\eta \geq 0$. We start by showing that the family of linear equations $\ell_s(\eta) = 0$, $s \in \mathcal{S}$ and $\eta_{sa} = 0$, $a \in A_o$, $s \in S_o$, $o \in \mathcal{O}$ is linearly independent for any choice of $A_o \subsetneq \mathcal{A}$, $o \in \mathcal{O}$. For this we first note that the linear equations $\ell_s(\eta) = 0$, $s \in \mathcal{S}$ define the space $\mathcal{L} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ of dimension $\dim(\mathcal{L}) = \dim(\text{affine}(\Delta_{\mathcal{A}}^{\mathcal{S}})) = n_{\mathcal{S}}(n_{\mathcal{A}} - 1)$ (see Proposition 4.5.3), which implies their linear independence. It now suffices to see that the restrictions of η_{sa} to $\mathcal{L}$ are linearly independent. Note that the pullback of the equations $\eta_{sa} = 0$ restricted to $\mathcal{L}$ along the birational map Ψ are the equations $\tau_{as} = 0, a \in A_o, s \in S_o$, which are linearly independent on affine$(\Delta_{\mathcal{A}}^{\mathcal{S}})$.

On the set B given in (4.20) there are $\sum_o d_o |A_o|$ active linear inequalities with $A_o \subsetneq \mathcal{A}$ for each $o \in \mathcal{O}$, and hence B is contained in an affine space of dimension

$$n = n_{\mathcal{S}} n_{\mathcal{A}} - n_{\mathcal{S}} - \sum_o d_o |A_o|.$$

Further, given these linear equations, the quadratic equations

$$p_{sa}^o(\eta) = \eta_{sa} \sum_{a' \in \mathcal{A}} \eta_{s_o a'} - \eta_{s_o a} \sum_{a' \in \mathcal{A}} \eta_{sa'} = 0$$

are redundant for all $a \in A_o$, $s \in S_o$. By choosing $a_o \in A_o^c$ in Corollary 4.5.5 for every $o \in \mathcal{O}$ there remain $n_{\mathcal{A}} - |A_o| - 1$ non-redundant quadratic equalities for every $s \in S_o \setminus \{s_o\}$. Therefore, we get $m = \sum_o (d_o - 1)(|A_o^c| - 1)$ non-redundant quadratic equalities. By Theorem 2.2 and Corollary 2.5 in [NR09] the algebraic degree for the optimization of the linear function $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ over an n-dimensional affine space subject to m non-redundant quadratic constraints is upper bounded by $2^m \binom{n-1}{m-1}$. □

Remark 4.6.4. We will observe in Example 4.8.3 below that the bounds from Theorem 4.6.3 are weaker than the bounds obtained in Chapter 2. They are however a lot easier to evaluate, and based on a different method we will be able to obtain sharp bounds in some instances later, with Theorem 4.8.1.

With Theorem 4.6.3 we can provide upper bounds for the number of critical points of the optimization problem (4.18). Indeed, the number of critical points over the interior

$$\{\eta \in \mathcal{L} \cap \mathcal{X} : \ \eta_{sa} = 0 \text{ for all } a \in A_{g_\beta(s)}, \eta_{sa} > 0 \text{ otherwise}\} \tag{4.21}$$

of a boundary component is clearly upper bounded by the number of critical points over B defined in (4.20). This bound over the individual boundary components can be summed to obtain an upper bound on the number of critical points of the polynomial optimization problem (4.18) (see also [NR09]). Note that the Zariski closure of the interior of a boundary component defined in (4.21) is contained in B but might be a strict subset. In fact, by Remark 4.5.6, Theorem 4.6.3 can fail to give tight bounds on the number of (complex) critical points in those instances where B is not a complete intersection.

Remark 4.6.5 (Tighter bounds via polar degrees)**.** A more satisfactory approach is to compute polar degrees of the state aggregation variety $\mathcal{X} \cap \mathcal{L}$. This approach yields tighter bounds, as demonstrated in the special case of a blind controller with two actions, i.e., a system with one observation and two actions in [MM22a]. The authors obtain an upper bound linear in $n_{\mathcal{S}}$ compared to the exponential upper bound of $n_{\mathcal{S}} \cdot 2^{n_{\mathcal{S}}-1} + 2$ that follows from Theorem 4.6.3. We address this approach in Section 4.8 in full generality, and give a description of the number of critical points in Theorem 4.8.5 under some generality assumption. We compare the bounds obtained from polar degrees to the previous results from this chapter in Example 4.8.4 below.

4.6.1 Evaluation of the bounds

In Table 4.1 we present the upper bounds on the number of critical points for problems of different sizes. We compare the bound on the total number of critical points obtained by iterating Theorem 4.6.3 over all boundary components and the one iterating only over the relevant components described in Theorem 4.6.1. In addition, we report the total and relevant number of boundary components discussed in Remark 4.6.2. Both the number of boundary components and the upper bound on the number of critical points, depend on $n_{\mathcal{S}}, n_{\mathcal{A}}$, and the tuple $(d_o)_{o\in\mathcal{O}}$. The two extreme cases for the tuple $(d_o)_{o\in\mathcal{O}}$, namely $(n_{\mathcal{S}})$ and $(1,\ldots,1)$, correspond to a *blind controller*, i.e., all states map to the same observation, and the *fully observable case*, i.e., states and observations are in one-to-one correspondence, respectively. The bounds are independent of the specific α, so long as Assumption 4.5.2 is satisfied.

In these examples, we observe that restricting to the relevant boundary components significantly reduces the upper bound. This is reflected in the last two columns in Table 4.1. The difference is most notable when the fibers of g_β have a small cardinality, i.e., only few states lead to the same observation. In the fully observable case, the relevant boundary components correspond to the vertices of $\Delta_{\mathcal{A}}^{\mathcal{O}}$. This is consistent with the fact that in the fully observable case the feasible set $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$ is a polytope [Der70] and hence the optimization problem (4.18) is a linear program, for which the solutions are attained at the vertices. On the other hand, in the case of a blind controller (with a single observation o), all boundary components are relevant since $d_o = n_{\mathcal{S}}$.

4.7 Numerical methods for the optimization of decision rules

In POMDPs, reward optimization over the set of memoryless stochastic policies (4.8) is known to be hard in theory (NP-hard [VLB12]) and also difficult in practice as the reward function R is nonconvex and has sub-optimal strict local optima [PLT11, BR19]. In this section, we discuss how the geometric description of reward optimization facilitates computational approaches based on

$n_{\mathcal{S}}$	$n_{\mathcal{A}}$	partitions of $n_{\mathcal{S}}$: $(d_o)_{o\in\mathcal{O}}$	Number of boundary components		Bound on number of critical points	
			total	relevant	total	relevant
3	2	(3)	3	3	10	10
		(2, 1)	9	6	10	8
		(1, 1, 1)	27	8	8	8
4	3	(4)	7	7	1419	1419
		(3, 1)	49	21	2237	561
		(2, 2)	49	36	1265	153
		(2, 1, 1)	343	54	1189	81
		(1, 1, 1, 1)	2401	81	81	81
5	3	(5)	7	7	9411	9411
		(4, 1)	49	21	23745	4257
		(3, 2)	49	42	13431	4371
		(3, 1, 1)	343	63	24363	1683
		(2, 2, 1)	343	108	12159	459
		(2, 1, 1, 1)	2401	162	9195	243
		(1, 1, 1, 1, 1)	16807	243	243	243

Table 4.1: Listed are the number of boundary components and the upper bound on the number of critical points from Theorem 4.6.3 both over all boundary components and over the subset of relevant boundary components from Theorem 4.6.1 for problems of different size.

numerical algebra. We derive polynomial systems for the critical points, globally from the Karush-Kuhn-Tucker (KKT) conditions, and separately for each boundary component of $\Phi(\Delta_{\mathcal{A}}^{\mathcal{O}})$ from the Lagrangian criterion. For different choices of $n_{\mathcal{S}}$, $n_{\mathcal{A}}$, and generic data (i.e., generic α, μ, and r), we compute the complex and real solutions of the KKT and Lagrangian systems, and we compare the number of solutions with the theoretical upper bounds established in Section 4.6. Finally, we compare these approaches with other popular methods from constrained optimization: the interior point solver `Ipopt` and convex relaxations via the moment-SOS approach.

4.7.1 Critical equations and computation

The KKT critical point equations A standard approach for constrained optimization problems are the KKT conditions [KT51], which provide necessary conditions of stationary points under certain regularity conditions; see, e.g., [Aba67, Ber97, BSS06]. If both the constraints and objective function are polynomial, the KKT conditions form a polynomial system, which can be solved using various numerical algebraic methods.

Applied to our problem, the KKT conditions reduce to the following polynomial system in $\eta \in \mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}}$ with multipliers $\lambda \in \mathbb{R}^{\mathcal{S}}, \nu_{sa}^o \in \mathbb{R}, \kappa \in \mathbb{R}_{\geq 0}^{\mathcal{S}\times\mathcal{A}}$:

$$\begin{aligned}
\text{Primal feasibility:} \quad & \ell_s(\eta) = 0 \text{ for } s \in \mathcal{S}, \\
& p_{sa}^o(\eta) = 0 \text{ for } o \in \mathcal{O}, a \in \mathcal{A}\setminus\{a_o\}, s \in S_o \setminus \{s_o\}, \\
\text{Complementary slackness:} \quad & \kappa_{s_o a}\eta_{s_o a} = 0 \text{ for all } s_o, a, \\
\text{Stationarity:} \quad & r + \sum_s \lambda_s \nabla \ell_s(\eta) + \sum_{o,s,a} \nu_{sa}^o \nabla p_{sa}^o(\eta) + \kappa = 0,
\end{aligned} \tag{4.22}$$

where $a_o \in \mathcal{A}$ and $s_o \in S_o$ for every $o \in \mathcal{O}$ are fixed arbitrarily. Here we have included the primal feasibility $\eta_{sa} \geq 0$ for $s \in \mathcal{S}, a \in \mathcal{A}$ and the dual feasibility $\kappa_{sa} \geq 0$ for $s \in \mathcal{S}, a \in \mathcal{A}$ in the definition

of the search space for η and κ.

The number of linear constraints ℓ_s is $n_{\mathcal{S}}$, while the number of polynomial constraints p^o_{sa} is $(n_{\mathcal{A}}-1)\sum_{o\in\mathcal{O}}(d_o-1) = (n_{\mathcal{A}}-1)(n_{\mathcal{S}}-n_{\mathcal{O}})$. Due to the symmetry of the effective policies, there are only $n_{\mathcal{O}}n_{\mathcal{A}}$ inequalities, $\eta_{s_oa} \geq 0$ for each $a \in \mathcal{A}, o \in \mathcal{O}$. Hence the dimension of the square KKT system (4.22) is

$$n_{\mathcal{S}}n_{\mathcal{A}} + n_{\mathcal{S}} + (n_{\mathcal{A}}-1)(n_{\mathcal{S}}-n_{\mathcal{O}}) + n_{\mathcal{O}}n_{\mathcal{A}} = 2n_{\mathcal{S}}n_{\mathcal{A}} + n_{\mathcal{O}}.$$

In this setting, we can verify that the linear independence constraint qualification is satisfied. Given an element η^* in the feasible set $\Phi(\Delta^{\mathcal{O}}_{\mathcal{A}})$, it suffices to verify the linear independence of the gradients of the active inequality constraint functions and the equality constraint functions at η^*. Notice that under the pullback along the birational morphism $\Gamma = \Psi^{-1}$ the equality constraints in (4.16) are identified with the affine-linear functions l^o_{sa} defined in the proof of Theorem 4.5.4. Checking the linear independence of their gradients can be done by counting the dimension of the faces.

The Lagrange critical point equations over boundary components Alternatively to solving the KKT system, one can compute the critical equations given by the Lagrange criterion over every boundary component individually. If there are no inequality constraints, the KKT equations specialize to the Lagrange multiplier equations. Consider a boundary component B in (4.20) for a choice of $A_o \subsetneq \mathcal{A}$ for every $o \in \mathcal{O}$, and consider the optimization problem over B. This amounts to setting $\eta(s,a) = 0$ for $a \in A_o$ whenever $g_\beta(s) = o, o \in \mathcal{O}$, which reduces optimization to a subspace of $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. We denote the new primal variables by $\hat{\eta}$. Similarly, we denote the restriction of ℓ_s and p^o_{sa} to this space by $\hat{\ell}_s$ and $\hat{p}^o_{sa}$ and the projection of r onto this space (i.e., the vector obtained by dropping the indices which are set to zero in η) by $\hat{r}$. In the lower dimensional variables $\hat{\eta}$ for a given B the Lagrange system becomes

$$\begin{aligned} \text{Feasibility:} \quad & \hat{\ell}_s(\hat{\eta}) = 0 \text{ for } s \in \mathcal{S}, \\ & \hat{p}^o_{sa}(\hat{\eta}) = 0 \text{ for } o \in \mathcal{O}, a \in \mathcal{A}\setminus\{a_o\}, s \in S_o \setminus \{s_o\}, \\ \text{Stationarity:} \quad & \hat{r} + \sum_s \lambda_s \nabla\hat{\ell}_s(\hat{\eta}) + \sum_{o,s,a} \nu^o_{sa}\nabla\hat{p}^o_{sa}(\hat{\eta}) = 0, \end{aligned} \tag{4.23}$$

where $a_o \in A^c_o$ and $s_o \in S_o$ are fixed arbitrarily for every $o \in \mathcal{O}$. The dimension of the primal variable $\hat{\eta}$ is $n_{\mathcal{S}}n_{\mathcal{A}} - \sum_o d_o|A_o|$, the dimension of the Lagrange multipliers λ is $n_{\mathcal{S}}$ and of ν^o_{sa} is $\sum_o(d_o-1)(|A^c_o|-1)$ (see also the proof of Theorem 4.6.3). Overall, the Lagrange system (4.23) is a square polynomial system of dimension

$$2n_{\mathcal{S}}n_{\mathcal{A}} - (n_{\mathcal{A}}-1)n_{\mathcal{O}} - \sum_o (2d_o-1)|A_o|.$$

Remark 4.7.1 (Lagrange vs KKT system)**.** It is easy to see that every real solution of the KKT system satisfying the primal and dual inequality constraints $\eta \geq 0, \kappa \geq 0$ is a solution of the Lagrange system over a boundary component, namely the boundary component defined by the zeros of η; see Figure 4.3 for an illustrated example of this situation. When solving the KKT system (4.22), usually one solves the system of equations without the nonnegativity conditions $\eta \geq 0$ and $\kappa \geq 0$ and then selects the nonnegative solutions. Note that every solution of the Lagrange system over a boundary component appears as the solution of the KKT system without the nonnegativity constraints. Hence, solving the KKT system gives at least as many solutions as solving the Lagrange system over every boundary component.

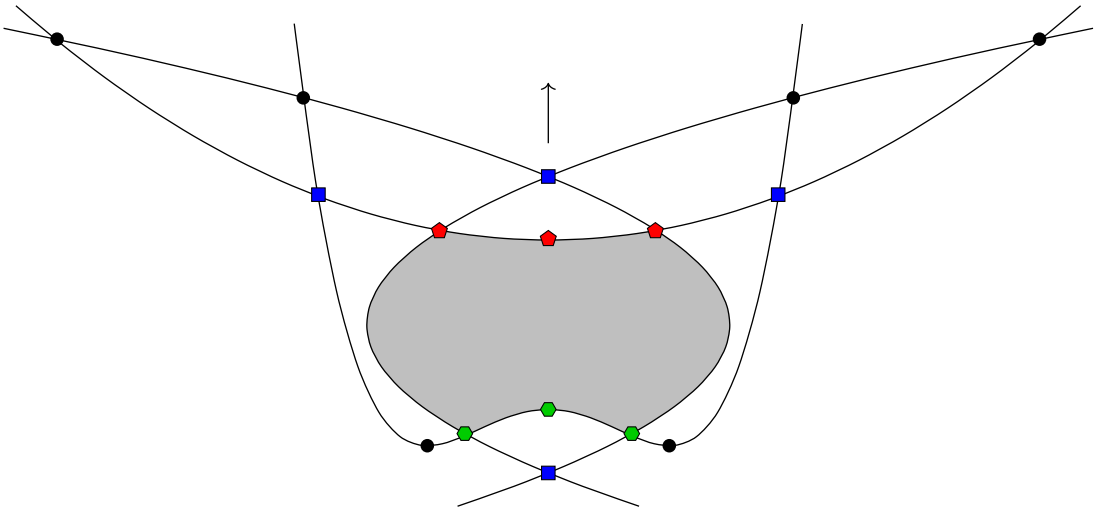


Figure 4.3: Schematic illustration of the feasible region (gray) and objective gradient (arrow) of a polynomially constrained linear program showing (i) the solutions of the KKT system checking only primal ($\eta \geq 0$) inequality constraints (red pentagons and green hexagons); and checking primal ($\eta \geq 0$) and dual ($\kappa \geq 0$) inequality constraints (red pentagons), (ii) the solutions of the KKT system without checking inequalities (all points), (iii) the positive ($\eta \geq 0$) solutions of the Lagrange systems over all boundary components (red pentagons and green hexagons), (iv) all solutions of the Lagrange systems over all boundary components (red pentagons, green hexagons, and black circles).

4.7.2 Experiments

Description of the experiments We test our computational approach on random POMDPs of different sizes. To this end, we first specify the number of states $n_{\mathcal{S}}$, the number of actions $n_{\mathcal{A}}$, and the number of states aggregated in each observation $(d_o)_{o\in\mathcal{O}}$ with $\sum_o d_o = n_{\mathcal{S}}$. For each specification of these values, we generate 20 random problems as follows. We sample the initial state distribution μ and the transition probabilities $\alpha(\cdot|s,a)$, $(s,a) \in \mathcal{S}\times\mathcal{A}$ from a symmetric Dirichlet distribution, and sample the instantaneous reward vector $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ from a standard Gaussian distribution. We use the same random data for both approaches, KKT and Lagrange over boundary components.

Computation The optimization problem (4.18) can be solved using several methods:

- First, we use the numerical algebra package `HomotopyContinuation.jl` [BT18] to solve the KKT system (4.22) and the Lagrange system (4.23) of each boundary component. This automatically certifies the results, meaning that for every returned solution, a unique true solution is guaranteed in a small neighborhood. From the returned solutions to the critical equations, we then just need to select the real ones that satisfy the primal inequality constraints $\eta_{s,a} \geq 0$, and among them the one that has the maximum objective value.

- Alternatively, we solve a convex relaxation of the polynomial optimization problem (4.18). Namely, we relax the problem to a semidefinite program (SDP) via the moment-SOS-approach that is implemented in the freeware `GloptiPoly3` [HLL09], and solve the SDPs using the numerical solver `Mosek`; see [DA21] for details. We note that `GloptiPoly3` builds upon a hierarchy of moment/SOS programs (also called Lasserre hierarchy), which allows to approximate the optimal value arbitrarily close, and can be used to test optimality and extract global optimizers [HL05, Nie11]. We use this key feature to check if our methods reach global optimality.

- We may also solve the constrained optimization problem (4.18) using the interior point solver `Ipopt` [WB06], which is a local optimization method for large-scale nonlinear optimization, an approach recently pursued in [MM22b].

$n_{\mathcal{S}}$	$n_{\mathcal{A}}$	$(d_o)_{o\in\mathcal{O}}$	KKT			Lagrange (all)			Lagrange (relevant)		
			complex	real	positive	complex	real	positive	complex	real	positive
3	2	(3)	6±0	4.4±1.2	2.1±0.3	6±0	4.4±1.2	2.1±0.3	6±0	4.4±1.2	2.1±0.3
		(2,1)	12±0	10.1±1.9	4.25±0.44	10±0	8.2±1.9	4.25±0.44	8±0	6.7±1.6	4.25±0.44
		(1,1,1)	20±0	20±0	8±0	8±0	8±0	8±0	8±0	8±0	8±0
4	3	(4)	45±0	17.1±4.3	4.3±1.3	45±0	17.1±4.3	4.3±1.3	45±0	17.1±4.3	4.3±1.3
		(3,1)	150±0	79±11	11±1.9	129±0	68.7±9.7	11±1.9	81±0	41.6±8.5	10.9±1.8
		(2,2)	281.6±0.75	154±16	13.9±4.7	263±0	136±16	13.9±4.7	153±0	89±10	13.65±4.3
		(2,1,1)	381.2±0.7	292±23	31.5±4.3	216±0	168±16	31.5±4.3	81±0	68±11	30.9±4.0
		(1,1,1,1)	495±0	495±0	81±0	81±0	81±0	81±0	81±0	81±0	81±0
5	3	(5)	71±0	21.4±6	3.7±0.98	71±0	21.4±6	3.7±0.98	71±0	21.4±6	3.7±0.98
		(3,2)	637.95±0.76	219±28	12.60±2.9	626±0	213±29	12.6±2.9	477±0	171±24	12.6±2.9
		(4,1)	269.85±0.49	99±20	11.9±3.3	234±0	87±18	11.9±3.3	144±0	52±13	11.55±2.6
		(3,1,1)	881.95±0.22	436±68	36±10	558±0	285±47	36±10	243±0	117±20	35.3±9.2
		(2,2,1)	1717.3±2.5	890±49	35.6±5.3	1260±0	624±56	36.5±7.1	459±0	244±25	35.7±6.6
		(2,1,1,1)	2269.9±3.9	1712±142	89±12	810±0	624±74	89.3±12.3	243±0	195±37	88.1±9.5
		(1,1,1,1,1)	3002.9±0.31	3002.9±0.3	243±0	243±0	243±0	243±0	243±0	243±0	243±0

Table 4.2: Mean and standard deviation of the number of solutions of the KKT system (4.22), the Lagrange system (4.23) over all boundary components, and the Lagrange system over the relevant boundary components, for 20 random POMDPs with the indicated number of states $n_{\mathcal{S}}$, actions $n_{\mathcal{A}}$, and state-aggregation partition $(d_o)_{o\in\mathcal{O}}$. In our setting, positive solutions are feasible solutions.

Discussion of the experimental results In this section, we discuss the experimental results on the number of solutions obtained by solving the KKT and Lagrange systems introduced above. In Table 4.2, we report the average and standard deviation of the number of complex, real, and positive solutions returned in each case. Note that in our setting, positive solutions (i.e., solutions satisfying $\eta \geq 0$) are (primal) feasible solutions. We also compare these methods' performance and computational times with convex relaxations and interior point methods.

Following the discussion in Remark 4.7.1, we start by comparing the number of solutions of the KKT and the Lagrange systems. In Table 4.2 we see that the KKT system has at least as many complex solutions as the Lagrange systems over all boundary components. This is consistent with our previous discussion, since, as we have pointed out, any solution of the Lagrange system over a boundary component is a solution of KKT. Moreover, KKT and Lagrange over all boundary components generally have the same number of positive solutions (see Remark 4.7.1 and Figure 4.3).

The difference between the number of complex, real, and positive solutions is also worth noting. Table 4.2 reveals a drop between the number of complex solutions and the number of real and positive solutions of the three types of systems. However, we find an exception to this in the Lagrange system for fully observable systems ($d_o = (1, \ldots, 1)$), where the number of complex, real, and positive solutions coincide. Indeed, in this case all boundary components are affine spaces, so only the zero-dimensional boundary components have a solution, and these correspond precisely to the $n_{\mathcal{A}}^{n_{\mathcal{S}}}$ vertices of the feasible set.

We also observe that the number of complex solutions has a much smaller variance than the number of real or positive ones. This is expected, since choosing the coefficients of polynomial systems randomly gives the same number of complex solutions with probability one. In fact, the number of complex solutions for the Lagrange system has no variance across the different random parameters. Still, we see a small variance in the number of complex KKT solutions, which we attribute to numerical instability: this can prevent the software package `HomotopyContinuation.jl` from finding all solutions to the KKT system. In contrast to the complex case, the variance on the number of real and positive solutions is not due to numerical errors. This is a typical phenomenon in polynomial system solving and is one of the possible limitations of classic algebraic methods when one wants to estimate the number of real solutions of a system.

In the following we compare the experimental results presented in Table 4.2 with the theoretical upper bounds shown in Table 4.1 and highlight two particular facts. First notice that in most cases the theoretical bound is significantly larger than the number of solutions of the Lagrange system. Moreover, this gap becomes particularly pronounced for problems where the fibers of g_β are large. This indicates that there is a discrepancy between the theoretical bounds and the algebraic degree of the optimization problem. Indeed, our bounds are based on the theory for generic polynomials. Hence, we do not expect that they provide a tight estimate of the algebraic degree for the particular polynomials we are dealing with. Here we also observe a particular behavior in the case of fully observable systems where the number of critical points of the Lagrange systems agrees with our bounds. On the other hand, we see that in some cases the number of solutions of KKT is larger than the bound, which agrees with our discussion on solutions of KKT and Lagrange systems in Remark 4.7.1.

In addition to analyzing the number of solutions of the KKT and Lagrange systems, we are interested in comparing the different solution methods for the optimization problem. Therefore, we compare the optimal solution found by solving these systems with `HomotopyContinuation.jl` with the one found by `Ipopt` and `GloptiPoly3`. Although `HomotopyContinuation.jl` is not guaranteed to find all solutions to the KKT and Lagrange systems, we observe that this approach yields a reward that is at least as high as the one obtained by the interior point method `Ipopt` and, in a few instances, strictly higher. In fact, solving the optimization problem with `GloptiPoly3` returns a certificate for the optimality of the result, which in all computed instances coincides with the optimal value obtained by solving the KKT and Lagrange systems with `HomotopyContinuation.jl`. That is, `GloptiPoly3` offers numerical evidence that they always provide globally optimal solutions. In all computed instances using `GloptiPoly3`, the optimal value of the optimization problem was already attained at the first-order relaxation of the Lasserre hierarchy [Las01]. We conjecture that objective value exactness for the first order relaxation of (4.18) holds with high probability for generic input data. Since the size of the SDP depends very sensitively on the order of the relaxation, this conjecture would remedy one of the major drawbacks of the SDP relaxation method. In more detail, the t-th order relaxation for both, the moment and the SOS relaxation of a polynomial optimization problem, can be computed via an SDP of size $\binom{n+t}{t}$, where n is the number of variables of the involved polynomials.

As described in [Nie11], finite convergence of the Lasserre hierarchy, i.e., convergence after finitely many relaxation steps, is closely related to certifying the flat truncation property. In fact, finite convergence holds generically [Nie14]. However, studying exactness properties of the SOS and the moment relaxation is still an ongoing topic of current research, see e.g. [BM22].

Finally, in Table 4.3 we report the computation times of the different approaches. The KKT and Lagrange systems as well as Ipopt were computed on a server with a 2x 32-Core AMD Epyc 7601 at

2.2 GHz and 1024 GB RAM, whereas the SDP relaxation was computed on a Intel(R) Core(TM) i7-8550U CPU with 4 cores at 1.8 GHz and 16GB RAM. Solving the Lagrange equations only over the relevant boundary components was up to two orders of magnitude faster than solving them over all boundary components. The improvements are more pronounced when g_β has small fibers in which we can exclude more faces by means of Theorem 4.6.1; see also Table 4.1. The computation times for the solution of the KKT system are in the same order to magnitude as the computation time of the solution of the Lagrange systems over all boundary components. KKT is slightly faster when g_β has small fibers and slightly slower when g_β has large fibers. The SDP approach is several orders of magnitude faster compared to the solution of the KKT and Lagrange systems with the gap becoming more pronounced for problems of increasing size. The interior point method Ipopt is again several orders of magnitude faster. Ipopt and SDP return one candidate solution, whereas homotopy continuation attempts to return all critical points. Note however that in contrast to the SDP relaxation the interior point method only guarantees locally optimal solutions. In our experiments we consistently observed that Ipopt yields less accurate solutions and sometimes converges to suboptimal points. Indeed, the maximum euclidean difference of the reward obtained by Ipopt and SDP is 9.68×10^{-2}, whereas the maximum difference between either of KKT and Lagrange methods and SDP is 2.98×10^{-7}.

	Partitions of $n_\mathcal{S}$	Ipopt	SDP	KKT	Lagrange (all)	Lagrange (relevant)
$n_\mathcal{S}=3, n_\mathcal{A}=2$	(3)	0.01	0.213	1.575	0.046	1.175
	(2,1)	0.009	0.168	1.551	3.563	2.757
	(1,1,1)	0.006	0.171	0.114	0.119	0.03
$n_\mathcal{S}=4, n_\mathcal{A}=3$	(4)	0.011	1.167	19.885	7.642	10.407
	(3,1)	0.01	1.114	76.071	43.759	22.17
	(2,2)	0.011	1.278	173.644	114.208	48.52
	(2,1,1)	0.009	1.292	79.775	191.394	27.004
	(1,1,1,1)	0.007	1.184	13.82	32.637	0.693
$n_\mathcal{S}=5, n_\mathcal{A}=3$	(5)	0.011	7.394	62.321	31.257	31.501
	(3,2)	0.01	6.338	1768.722	509.877	259.054
	(4,1)	0.011	7.256	307.524	163.88	69.5
	(3,1,1)	0.01	6.608	895.701	704.813	91.901
	(2,2,1)	0.011	6.078	2831.482	2175.098	313.557
	(2,1,1,1)	0.009	6.22	899.981	2058.912	188.536
	(1,1,1,1,1)	0.006	5.159	172.621	319.165	3.667

Table 4.3: Average run times for the different approaches reported in seconds. KKT and Lagrange are computed with homotopy continuation.

4.8 Polar degrees of state aggregation varieties

In this section, we review the reward optimization problem (4.12) from the perspective of algebraic geometry. The main result of this section, Theorem 4.8.1, bounds the algebraic complexity of this optimization problem. Contrary to Theorem 4.6.3, this bound is tight under some genericity conditions. We compare the results to each other and to the results from Chapter 2 in Example 4.8.4.

We start by fixing a POMDP with deterministic observations and remind the reader of the notation. Let $\mathcal{S}, \mathcal{A}, \mathcal{O}$ denote the sets of states and actions and let the map $\beta : \mathcal{S} \longrightarrow \mathcal{O}$ denote the

(deterministic) observation kernel and let $\alpha \in \mathbb{R}^{A\times S\times A}$ denote the transition kernel. We denote $0 < \gamma \leq 1$ the discount factor and $\mu \in \Delta_S$ the starting distribution. Let $\mathcal{X}$ be the associated determinantal variety from equation (4.14):

$$\mathcal{X} = \left\{\eta \in \mathbb{C}^{\mathcal{S}\times\mathcal{A}} : \ \eta_{sa}\eta_{s'a'} - \eta_{sa'}\eta_{s'a} = 0 \ \forall a, a' \in \mathcal{A} \text{ and } s, s' \in \mathcal{S} \text{ with } g_\beta(s) = g_\beta(s')\right\}$$

and let further $\mathcal{L}$ denote the affine linear space

$$\mathcal{L} = \left\{\eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \ \ell_s(\eta) = 0 \text{ for all } s \in \mathcal{S}\right\}, \ \ell_s(\eta) := \sum_a \eta_{sa} - \gamma \sum_{s',a'} \eta_{s'a'}\alpha(s|s',a') - (1-\gamma)\mu_s$$

from Equation (4.13). In view of Theorem 4.5.4 our aim is to study the following optimization problem:

$$\text{maximize } \langle r, \eta\rangle \quad \text{subject to } \eta \in \mathcal{X} \cap \mathcal{L} \cap \mathbb{R}^{S\times A}_{\geq 0}.$$

Analogous to the discussion from Section 4.7, in this section we study the number of critical points individually for each boundary component of the feasible set. As we observed at the beginning of Section 4.6, every boundary component B is the intersection of the feasible set with an orthant

$$\text{Orth}(B) = \{\eta : \quad \eta_{sa} = 0 \text{ for } a \in A_{g_\beta(s)}\}, \tag{4.24}$$

where for every observation o, $A_o \subsetneq \mathcal{A}$ is a proper subset of the set of actions. The boundary component B constitutes the state action frequencies of those policies π that, given an observation o, *never* choose any action from the set A_o.

Under the assumption that the complex linear space $\mathcal{L} \cap \text{Orth}(B)$ is in general position relative to $\mathcal{X}$ we can attach an algebraic degree to the optimization problem. This is done by our main result in this section, namely Theorem 4.8.5. In order to state it, we need the following notation.

Let m_1, m_2, r be positive integers. We set $N = m_1m_2 - 1$, $d = m_1 + m_2 - 2$, and define

$$\gamma_r(m_1, m_2) = \sum_{k=0}^{d-N+1+r} (-1)^k \binom{d+1-k}{N-r} (N-k)! \left(\sum_{i+j=k} \frac{\binom{m_1}{i}}{(m_1-1-i)!} \cdot \frac{\binom{m_2}{j}}{(m_2-1-j)!}\right).$$

Further, we define the homogeneous polynomial

$$H(m_1, m_2) = \gamma_1(m_1, m_2)s^N t^1 + \cdots + \gamma_N(m_1, m_2)s^1 t^N \in \mathbb{Z}[s, t]. \tag{4.25}$$

Let $c = \left(\sum_{o\in O} \#A_o\right) + \#S$ denote the codimension of the linear space $\mathcal{L} \cap \text{Orth}(B)$.

Theorem 4.8.1. *Let $r \in \mathbb{R}^{\mathcal{A}\times\mathcal{S}}$ be a generic reward vector. The number of isolated critical points of the linear function $\eta \mapsto \langle r, \eta\rangle$ on the relative interior of the boundary component B is upper bounded by the coefficient of the monomial $s^{\#\mathcal{S}\times\#\mathcal{A}-c}t^c$ in the polynomial*

$$\mathcal{H} = \prod_{o\in\mathcal{O}} H(\#\beta^{-1}(o), \#\mathcal{A}) \in \mathbb{Z}[s, t].$$

The number of isolated complex critical points of the linear function $\eta \mapsto \langle r, \eta\rangle$ on the complex variety $\mathcal{X} \cap \mathcal{L} \cap \text{Orth}(B)$ is also bounded above by this number, and equal to it if we replace $\mathcal{L} \cap \text{Orth}(B)$ with a generic affine linear space of codimension c.

Remark 4.8.2. It is natural to ask whether for a general choice of the matrix M_α the bound from Theorem 4.8.1 on the number of complex critical points on the variety $\mathcal{X} \cap \mathcal{L} \cap \mathrm{Orth}(B)$ is tight. We encountered examples that exhibit both behaviours, depending on the combinatorics of the observation kernel β. This is subject to future research.

In the following two examples we demonstrate that there are instances for which the main result of this section, Theorem 4.8.1, is tight, and instances for which it fails to be tight. In both cases we consider a ,,general" transition kernel $\alpha \in \Delta_{\mathcal{A}}^{\mathcal{S}\times\mathcal{A}}$ and a generic reward vector $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. We further compare the bounds from Chapter 2, namely Theorem 2.3.7, to Theorem 4.8.1 and Theorem 4.6.3. We start by demonstrating that the bound from Theorem 4.8.1 needs not be tight.

Example 4.8.3. Consider the following POMDP describing a blind controller: We fix 4 states s_1, s_2, s_3, s_4, one observations and 3 actions a_1, a_2, a_3. Then the variety $\mathcal{X}$ is equal to the affine variety $M_{4\times 3}$ of rank one 4×3 matrices. It defines a projective variety of dimension 5 in $\mathbb{P}^{11}$. The polynomial $\mathcal{H}$ from the statement of Theorem 4.8.1 describes its polar class and evaluates to the following expression:

$$\mathcal{H} = 0s^{11}t^1 + 6s^10t^2 + 16s^9t^3 + 27s^8t^4 + 24s^7t^5 + 10s^6t^6 \in \mathbb{Z}[s,t].$$

We identify $\alpha \in \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$ with the product of the following three column stochastic matrices.

$$\alpha(\cdot \mid (\cdot, a_1)) = \frac{1}{7}\begin{pmatrix} 0 & 1 & 3 & 2 \\ 2 & 3 & 0 & 1 \\ 3 & 1 & 1 & 1 \\ 2 & 2 & 3 & 3 \end{pmatrix}, \ \alpha(\cdot \mid (\cdot, a_2)) = \frac{1}{7}\begin{pmatrix} 3 & 0 & 1 & 1 \\ 1 & 1 & 2 & 3 \\ 2 & 4 & 4 & 1 \\ 1 & 2 & 0 & 2 \end{pmatrix}, \ \alpha(\cdot \mid (\cdot, a_3)) = \frac{1}{7}\begin{pmatrix} 2 & 0 & 4 & 0 \\ 1 & 3 & 0 & 2 \\ 0 & 0 & 2 & 3 \\ 4 & 4 & 1 & 2 \end{pmatrix}.$$

We set $\gamma = 1$, and the affine linear space $\mathcal{L}$ is defined by the vanishing of the following expressions:

$$\begin{aligned}
l_{s_1} &= 7\eta_{1,1} + 4\eta_{1,2} + 5\eta_{1,3} - \eta_{2,1} - 3\eta_{3,1} - \eta_{3,2} - 4\eta_{3,3} - 2\eta_{4,1} - \eta_{4,2} \\
l_{s_2} &= -2\eta_{1,1} - \eta_{1,2} - \eta_{1,3} + 4\eta_{2,1} + 6\eta_{2,2} + 4\eta_{2,3} - 2\eta_{3,2} - \eta_{4,1} - 3\eta_{4,2} - 2\eta_{4,3} \\
l_{s_3} &= -3\eta_{1,1} - 2\eta_{1,2} - \eta_{2,1} - 4\eta_{2,2} + 6\eta_{3,1} + 3\eta_{3,2} + 5\eta_{3,3} - \eta_{4,1} - \eta_{4,2} - 3\eta_{4,3} \\
l_{s_4} &= -2\eta_{1,1} - \eta_{1,2} - 4\eta_{1,3} - 2\eta_{2,1} - 2\eta_{2,2} - 4\eta_{2,3} - 3\eta_{3,1} - \eta_{3,3} + 4\eta_{4,1} + 5\eta_{4,2} + 5\eta_{4,3} \\
&\eta_{1,1} + \eta_{1,2} + \eta_{1,3} + \eta_{2,1} + \eta_{2,2} + \eta_{2,3} + \eta_{3,1} + \eta_{3,2} + \eta_{3,3} + \eta_{4,1} + \eta_{4,2} + \eta_{4,3} - 1.
\end{aligned}$$

Direct computation confirms that the number of critical points of a generic linear objective function on the state aggregation variety $\mathcal{X} \cap \mathcal{L}$ is 24. This is different from the coefficient 27 of the monomial s^8t^4 of $\mathcal{H}$, whence Theorem 4.8.1 does not give a sharp bound on the number of complex critical points. This discrepancy is not explained with our particular choice of α, but can be observed for every generic choice.

In the proof of Theorem 4.6.3 we express the positive part of $\mathcal{X} \cap \mathcal{L}$ as a complete intersection of the following 6 quadratic equations with the 8-dimensional affine linear space $\mathcal{L}$:

$$\begin{aligned}
&-\eta_{1,1}(\eta_{2,1} + \eta_{2,2} + \eta_{2,3}) + \eta_{2,1}(\eta_{1,1} + \eta_{1,2} + \eta_{1,3}) \\
&-\eta_{1,2}(\eta_{2,1} + \eta_{2,2} + \eta_{2,3}) + \eta_{2,2}(\eta_{1,1} + \eta_{1,2} + \eta_{1,3}) \\
&-\eta_{1,1}(\eta_{3,1} + \eta_{3,2} + \eta_{3,3}) + \eta_{3,1}(\eta_{1,1} + \eta_{1,2} + \eta_{1,3}) \\
&-\eta_{1,2}(\eta_{3,1} + \eta_{3,2} + \eta_{3,3}) + \eta_{3,2}(\eta_{1,1} + \eta_{1,2} + \eta_{1,3}) \\
&-\eta_{1,1}(\eta_{4,1} + \eta_{4,2} + \eta_{4,3}) + \eta_{4,1}(\eta_{1,1} + \eta_{1,2} + \eta_{1,3}) \\
&-\eta_{1,2}(\eta_{4,1} + \eta_{4,2} + \eta_{4,3}) + \eta_{4,2}(\eta_{1,1} + \eta_{1,2} + \eta_{1,3}).
\end{aligned}$$

This complete intersection is the union of the state aggregation variety $\mathcal{X} \cap \mathcal{L}$ with a linear space of dimension 5, so in this particular instance no new critical points are introduced away from $\mathcal{X} \cap \mathcal{L}$. Theorem 4.6.3 gives a bound to the number of complex critical points of a generic linear objective by invoking the results from [NR09], which evaluates to $2^6\binom{7}{5} = 5376$. In the proof of Theorem 4.6.3 we can replace the algebraic degree of polynomial optimization with the sparse analogue introduced in Chapter 2, in particular with the BKK bound of the corresponding Lagrange system. This evaluates to the number 252.

Example 4.8.4. Consider the following POMDP: We fix 6 states $s_1, s_2, s_3, s_4, s_5, s_6$, two observations o_1, o_2 and 2 actions a_1, a_2. The observation kernel $\beta : S \longrightarrow O$ maps s_1, s_2 and s_3 to o_1 and maps s_4, s_5 and s_6 to o_2. Then the variety $\mathcal{X}$ is the product $M_{3,2} \times M_{3,2}$, where $M_{3,2}$ denotes the affine variety of rank one 3×2 matrices. It defines a projective variety of dimension 7 in $\mathbb{P}^{11}$.

$$\mathcal{X} = \left\{\eta \in \mathbb{R}^{4\times 3} : \operatorname{rank}\begin{pmatrix}\eta_{1,1} & \eta_{1,2}\\ \eta_{2,1} & \eta_{2,2}\\ \eta_{3,1} & \eta_{3,2}\end{pmatrix} = \operatorname{rank}\begin{pmatrix}\eta_{4,1} & \eta_{4,2}\\ \eta_{5,1} & \eta_{5,2}\\ \eta_{6,1} & \eta_{6,2}\end{pmatrix} = 1\right\}.$$

For some choice of α we obtain the following defining equations of $\mathcal{L}$:

$$\begin{aligned}
l_{s_1} =& 2/3\eta_{1,1} + 2\eta_{2,1} + 3/4\eta_{3,1} - \eta_{3,2} - \eta_{4,1} + 2/3\eta_{4,2} + -2/5\eta_{5,1} + \eta_{5,2} - 3\eta_{6,1} + 1/5\eta_{6,2}\\
l_{s_2} =& -2/3\eta_{1,1} - \eta_{2,1} - \eta_{2,2} + -1/2\eta_{3,1} + 2/3\eta_{3,2} - \eta_{4,1} + -1/3\eta_{4,2} + 2/5\eta_{5,1} - \eta_{5,2} + 2\eta_{6,1} + -2/5\eta_{6,2}\\
l_{s_3} =& -2/3\eta_{1,1} + \eta_{1,2} + \eta_{2,1} - 2\eta_{2,2} + 5/4\eta_{3,1} + \eta_{3,2} + 3\eta_{4,1} + -1/3\eta_{4,2} + -3/5\eta_{5,1} + 2\eta_{5,2} - 2\eta_{6,1}\\
l_{s_4} =& -1/3\eta_{1,1} + -1/3\eta_{1,2} - 3\eta_{2,2} + -1/2\eta_{3,1} - \eta_{4,1} + 5/3\eta_{4,2} - \eta_{5,2} + 3\eta_{6,1} + -3/5\eta_{6,2}\\
l_{s_5} =& 2/3\eta_{1,1} + -1/3\eta_{1,2} + \eta_{2,1} + 3\eta_{2,2} + -1/2\eta_{3,1} + 1/3\eta_{3,2} + 2\eta_{4,1} - \eta_{4,2} + \eta_{5,1} + 2\eta_{6,1}\\
l_{s_6} =& 1/3\eta_{1,1} + -1/3\eta_{1,2} - 3\eta_{2,1} + 3\eta_{2,2} + -1/2\eta_{3,1} - \eta_{3,2} - 2\eta_{4,1} + -2/3\eta_{4,2} + -2/5\eta_{5,1} - \eta_{5,2} - 2\eta_{6,1} + 4/5\eta_{6,2}\\
& - 1 + \eta_{1,1} + \eta_{1,2} + \eta_{2,1} + \eta_{2,2} + \eta_{3,1} + \eta_{3,2} + \eta_{4,1} + \eta_{4,2} + \eta_{5,1} + \eta_{5,2} + \eta_{6,1} + \eta_{6,2}.
\end{aligned}$$

Direct computation shows that a general linear objective function has 34 critical points on the state aggregation variety. Contrary to Example 4.8.3 this is equal to the bound from Theorem 4.8.1: The polar class $\mathcal{H}$ of χ is represented by the following polynomial:

$$\mathcal{H} = 0s^1 0t^2 + 0s^9t^3 + 9s^8t^4 + 24^7t^5 + 34s^6t^6 + 24s^5t^7 + 9s^4t^8 \in \mathbb{Z}[s,t].$$

Evaluating the bound from Theorem 4.6.3 yields $2^4\binom{5}{3} = 160$, while the KKT bound to the Lagrange system evaluates to the slightly sharper bound 128. In this instance, the complete intersection described in the proof of Theorem 4.6.3 comprises three additional components in dimension 7, of respective degrees 1, 3 and 3. For a generic linear objective function these contribute an additional six critical points that do not live on the variety $\mathcal{X} \cap \mathcal{L}$.

The rest of this chapter is devoted to the proof of Theorem 4.8.1. Our approach is to compute the polar degrees of $\mathcal{X}$ by proving the following Theorem:

Theorem 4.8.5. *The polar class of the projective closure $\overline{\mathcal{X}}$ of $\mathcal{X}$ is represented by the polynomial*

$$t\prod_{o\in\mathcal{O}} H(\#\beta^{-1}(o), \#\mathcal{A}) \in \mathbb{Z}[s,t].$$

Although we already encountered polar degrees in this thesis, we will now remind the reader of all necessary definitions and relate them to linear optimization, showing that Theorem 4.8.1 is an easy corollary of Theorem 4.8.5. The rest of this section is then devoted to a proof of Theorem 4.8.5.

For the rest of the section $X \subseteq \mathbb{P}^n$, $Y \subseteq \mathbb{P}^m$ are projective varieties. In this section we denote

$$\pi_1 : \mathbb{P}^{m+n+1} \dashrightarrow \mathbb{P}^n, \text{ and } \pi_2 : \mathbb{P}^{m+n+1} \dashrightarrow \mathbb{P}^m$$

the rational projections on the first $n+1$ and on the last $m+1$ coordinates respectively. We now remind the reader of the definition of conormal varieties.

Definition 4.8.6. We define the projective conormal variety $\mathcal{N}_X$ of X to be the closure

$$\mathcal{N}_X = \overline{\{(x,u) \in \mathbb{P}^n \times (\mathbb{P}^n)^\vee : \quad x \in X_{\text{sm}} \text{ and } u \in T^\perp_{X,x}\}},$$

of all pairs (x,u), where x is a smooth point and u a hyperplane, tangent to X at x.

Definition 4.8.7. Let

$$[\mathcal{N}_X] = \delta_1(X)s^n t^1 + \cdots + \delta_n(X)s^1 t^n$$

denote the class of $\mathcal{N}_X$ in the cohomology ring $H^*(\mathbb{P}^n \times \mathbb{P}^n) = \mathbb{Z}[s,t]/\langle s^{n+1}, t^{n+1}\rangle$. We call $[\mathcal{N}_X]$ the polar class of X, and for each $i = 1, \ldots, n$ we call $\delta_i(X)$ the i-the polar degree of X.

Remark 4.8.8. Since $\mathcal{N}_X$ is a variety of codimension $n+1$ the above definition makes sense. The polar degrees satisfy $\delta_i(X) = \#(\mathcal{N}_X \cap L \times L')$, where $L \subseteq \mathbb{P}^n$ and $L' \subseteq (\mathbb{P}^n)^\vee$ are generic linear subspaces of dimension $n+1-i$ and i respectively. In particular, it holds $\delta_i(X) = 0$ for all i with $i > \dim(X)+1$. For more on polar degrees we point the reader to [Pie78]. We remark that in our notation, the indices of polar degrees start from 1, while in some sources they start from zero.

The reason we are interested in polar degrees is their relation to linear optimization.

Definition 4.8.9. Let $X \subseteq \mathbb{C}^n$ be a an affine variety. For each $i = 0, \ldots, n-1$ we define the i-th sectional linear optimization degree $c_i(X)$ of X as follows: Let L be a general affine linear space of codimension i. We define $c_i(X)$ to be the number of critical points of a generic linear functional over the intersection $X_{\text{sm}} \cap L$ of the smooth locus with L.

The following is a consequence of Corollary 6.3 in [MRWW23]:

Corollary 4.8.10. *Let $X \subseteq \mathbb{P}^n$ be the projective closure of the affine cone over a projective variety living in $\mathbb{P}^{n-1}$, and let X be of dimension d. Then for each $i = 0, \ldots, d$ it holds $\delta_{i+1}(\overline{X}) = c_i(X)$.*

Theorem 4.8.1 now follows directly from Theorem 4.8.5.

Proof of Theorem 4.8.1. The number of isolated complex critical points of the linear function $\eta \mapsto \langle r, \eta\rangle$ over the complex variety $\mathcal{X} \cap \mathcal{L} \cap \text{Orth}(B)$ is upper bounded by, and generically equal to the sectional linear optimization degree $c_k(\mathcal{X})$ of $\mathcal{X}$. By Corollary 4.8.10 the latter is equal to the polar degree $\delta_{k+1}(\overline{\mathcal{X}})$, where k is the codimension of $\mathcal{L} \cap \text{Orth}(B)$. This is the content of Theorem 4.8.5. □

The rest of this section is devoted to the proof of Theorem 4.8.5. It rests on a characterization of polar degrees of joins of varieties as convolution products in Theorem 4.8.15, and on a description of of the projective closure $\overline{\mathcal{X}}$ of $\mathcal{X}$ as a join of Segre varieties in Proposition 4.8.12. We start by defining the join.

Definition 4.8.11. Let $X \subseteq \mathbb{P}^n$, $Y \subseteq \mathbb{P}^m$ be projective varieties. We define their join to be

$$\mathcal{J}(X,Y) = \overline{\{(x_0 : \cdots : x_n : y_0 : \cdots y_m) \in \mathbb{P}^{n+m+1} : \quad (x_0 : \cdots : x_n) \in X,\ (y_0 : \cdots y_m) \in Y\}}.$$

Equivalently, $\mathcal{J}(X,Y)$ is the closure of the intersection of preimages $\pi_1^{-1}(X) \cap \pi_2^{-1}(Y)$.

The relation to POMDPs is the following. Let for any natural numbers m_1 and m_2, $N = m_1 m_2 - 1$ the variety $\mathcal{M}_{m_1,m_2}, \subseteq \mathbb{P}^N$, denote the projective Segre variety of rank one $m_1 \times m_2$ matrices of dimension $d = m_1 + m_2 - 2$, and let $M_{m_1,m_2} \subseteq \mathbb{C}^{m_1 m_2}$ denote its respective affine cone. Then we have the following theorem:

Proposition 4.8.12. *The projective closure $\overline{\mathcal{X}}$ of $\mathcal{X}$ is the common join of the point $\mathbb{P}^0$, and of all of the Segre varieties $\mathcal{M}_{\beta^{-1}(o),\mathcal{A}}$, where o is an element in the set of observations $\mathcal{O}$.*

Proof. As elaborated below the statement of Theorem 4.5.4, the affine variety $\mathcal{X}$ is equal to the product $\mathcal{X} = \prod_{o\in\mathcal{O}} M_{\beta^{-1}(o),\mathcal{A}}$. In particular, the affine cone over the projective closure $\mathcal{X}$ equals the product $\prod_{o\in\mathcal{O}} M_{\beta^{-1}(o),\mathcal{A}} \times \mathbb{C}$. When projectivizing we obtain the desired join of varieties. □

We need an analogue of the join for bihomogeneus varieties:

Definition 4.8.13. By a slight abuse of notation we denote by $\mathcal{J}(\mathcal{N}_X, \mathbb{P}^m)$ the variety

$$\overline{\{((x:y),(u:v)) \in \mathbb{P}^{n+m+1} \times (\mathbb{P}^{n+m+1})^\vee : \quad (x,u) \in \mathcal{N}_X\}}.$$

Equivalently,$\mathcal{J}(\mathcal{N}_X, \mathbb{P}^m)$ is the closure of the preimage $(\pi_1 \times \pi_1)^{-1}(\mathcal{N}_X)$, and we define $\mathcal{J}(\mathbb{P}^n, \mathcal{N}_Y)$ analogously as the closure $\overline{(\pi_2 \times \pi_2)^{-1}(\mathcal{N}_Y)}$:

$$\overline{\{((x:y),(u:v)) \in \mathbb{P}^{n+m+1} \times (\mathbb{P}^{n+m+1})^\vee : \quad (y,v) \in \mathcal{N}_Y\}}.$$

Proposition 4.8.14. *The cohomology class of the variety $\mathcal{J}(\mathcal{N}_X, \mathbb{P}^m)$ is represented by*

$$[\mathcal{J}(\mathcal{N}_X, \mathbb{P}^m)] = \delta_1(X) s^n t^1 + \cdots + \delta_n(X) s^1 t^n,$$

which is also the polynomial that represents $[\mathcal{N}_X]$. And analogously

$$[\mathcal{J}(\mathbb{P}^n, \mathcal{N}_Y)] = \delta_1(Y) s^m t^1 + \cdots + \delta_n(Y) s^1 t^m.$$

Proof. By symmetry we only treat the first case. Let $W \subseteq \mathbb{P}^{n+m+1}$, $W' \subseteq (\mathbb{P}^{n+m+1})^\vee$ be generic linear spaces of dimension $n+1-i$ and dimension i respectively for $i = 1, \ldots, n$. We denote by $L = \pi(W)$ and $L' = \pi(W')$ the projections onto the first $n+1$ variables, having the same dimensions $n+1-i$ and i respectively. By genericity of W and W' we have

$$\delta_i(X) = \#\mathcal{N}_X \cap L \times L' = \#\mathcal{J}(\mathcal{N}_X, \mathbb{P}^m) \cap W \times W'.$$ □

With the aim of computing the polar class of $\overline{\mathcal{X}}$, we now investigate polar classes of joins.

Theorem 4.8.15. *The polar class of the join $\mathcal{J}(X,Y)$ is represented by the product*

$$[\mathcal{N}_{\mathcal{J}(X,Y)}] = \left(\delta_1(X) s^n t^1 + \cdots + \delta_n(X) s^1 t^n\right) \left(\delta_1(Y) s^m t^1 + \cdots + \delta_n(Y) s^1 t^m\ \right)$$

of the polynomials that represent the polar classes of and X and Y respectively.

Proof. The idea is to split the defining equations of $\mathcal{N}_{\mathcal{J}(X,Y)}$ into two sets of polynomials in disjoint families of variables. We now show that the conormal variety $\mathcal{N}_{\mathcal{J}(X,Y)}$ is the intersection of the varieties $\mathcal{J}(\mathcal{N}_X, \mathbb{P}^m)$ and $\mathcal{J}(\mathbb{P}^n, \mathcal{N}_Y)$. Furthermore, at a generic point of $\mathcal{N}_{\mathcal{J}(X,Y)}$ this intersection is transversal. In particular, the polar class of $\mathcal{J}(X,Y)$ is $[\mathcal{J}(\mathcal{N}_X, \mathbb{P}^m)] \cdot [\mathcal{J}(\mathbb{P}^n, \mathcal{N}_Y)]$. Proposition 4.8.14 then finishes the proof.

Let $U \subseteq \mathbb{P}^{n+m+1} \times (\mathbb{P}^{n+m+1})^\vee$ denote the open subset consisting of all elements with only non zero entries and let $p = ((x_0 : \cdots x_n : y_0 \cdots : y_m), (u_0 : \cdots : u_n : v_0 \cdots v_m))$ be an arbitrary element. We first observe that the desired equality $\mathcal{J}(\mathcal{N}_X, \mathbb{P}^m) \cap \mathcal{J}(\mathbb{P}^n, \mathcal{N}_Y) = \mathcal{N}_{\mathcal{J}(X,Y)}$ holds on U.

$$\begin{aligned} p \in \mathcal{N}_{\mathcal{J}(X,Y)} &\iff (x:y) \in \mathcal{J}(X,Y) \text{ and } (u:v) \in T^\perp_{\mathcal{J}(X,Y),(x:y)} \\ &\iff x \in X,\ y \in Y,\ x \in T^\perp_{X,x},\ y \in T^\perp_{Y,y} \\ &\iff p \in \mathcal{J}(\mathcal{N}_X, \mathbb{P}^m) \text{ and } p \in \mathcal{J}(\mathbb{P}^n, \mathcal{N}_Y). \end{aligned}$$

To see transversality of the intersection, it suffices to observe that the bihomogeneous defining equations of $\mathcal{J}(\mathcal{N}_X, \mathbb{P}^m)$ and $\mathcal{J}(\mathbb{P}^n, \mathcal{N}_Y)$ are in disjoint sets of variables. By the same argument, the intersection $\mathcal{J}(\mathcal{N}_X, \mathbb{P}^m) \cap \mathcal{J}(\mathbb{P}^n, \mathcal{N}_Y)$ is irreducible and hence equal to $\mathcal{N}_{\mathcal{J}(X,Y)}$.

□

We recall the following description of polar classes of Segre varieties:

Corollary 4.8.16 (Corollary 15, [CJM+21])**.** *The polar class of the Segre variety $\mathcal{M}_{m_1,m_2}$ is represented by the polynomial*

$$[\mathcal{N}_{\mathcal{M}_{m_1,m_2}}] = H(m_1, m_2)$$

from equation (4.25). In other words, for $r = 1, \ldots, N, N = m_1 m_2 - 1$, $d = m_1 + m_2 - 2$ it holds

$$\alpha(\cdot)\delta_r(\mathcal{M}_{m_1,m_2}) = \sum_{k=0}^{d-N+1+r} (-1)^k \binom{d+1-k}{N-r} (N-k)! \left(\sum_{i+j=k} \frac{\binom{m_1}{i}}{(m_1-1-i)!} \cdot \frac{\binom{m_2}{j}}{(m_2-1-j)!} \right).$$

We finally prove Theorem 4.8.5, it is now a direct consequence of the previous results:

Proof of Theorem 4.8.5. Combining the description of $\overline{\mathcal{X}}$ as a join of Segre varieties from Proposition 4.8.12 with Theorem 4.8.15 and Corollary 4.8.16 gives the desired result. □

4.9 Conclusion

In this chapter we initiated the study of Markov decision problems from a new, geometric perspective. Our main object of interest was the set of state action frequencies, which we proved to be the positive part of the state aggregation variety. Based on this description, an algebraic optimization of the long term reward is possible by solving KKT and Lagrange equations. We conducted various numerical experiments based on this approach. Finally, by computing polar degrees of the state aggregation variety we are able to characterize the algebraic complexity of long term reward optimization.

Chapter 5

Discriminants and tropical implicitization

Tropical implicitization means computing the tropicalization of a unirational variety from its parametrization. In the case of a hypersurface, this amounts to finding the Newton polytope of the implicit equation, without computing its coefficients. We present a new implementation of this procedure in Oscar.jl. It solves challenging instances, and can be used for classical implicitization as well. We also develop implicitization in higher codimension via Chow forms, and we pose several open questions.

5.1 Introduction

Let $X \subset \mathbb{C}^n$ be a d-dimensional affine variety defined as the closure of the image of a map

$$f : \mathbb{C}^d \dashrightarrow \mathbb{C}^n, \quad t \longmapsto \big(f_1(t), f_2(t), \ldots, f_n(t)\big). \tag{5.1}$$

Here $t = (t_1, \ldots, t_d)$, and $f_1, f_2, \ldots, f_n \in \mathbb{C}(t)$ are rational functions. The problem of *implicitization* asks for the defining polynomial equations of X in the coordinates $x = (x_1, \ldots, x_n)$ on $\mathbb{C}^n$. When $f_1, \ldots, f_n \in \mathbb{Q}(t)$ have rational coefficients, these equations can be computed via *symbolic elimination* [DCO97, Chapter 3, §3]. More precisely, one eliminates $t_1, \ldots, t_d$ from

$$x_1 - f_1(t) \;=\; x_2 - f_2(t) \;=\; \cdots \;=\; x_n - f_n(t) \;=\; 0 \tag{5.2}$$

using Gröbner basis or resultant techniques. Unfortunately, these methods run out of steam for larger instances. This has motivated the question whether we can obtain interesting partial information in cases where computing the ideal of X is out of reach.

Tropical geometry [MS15] replaces an algebraic variety by a polyhedral complex which encodes many of its geometric properties. A commonly used slogan is that this complex serves as a *combinatorial shadow* of the original variety. The *tropicalization* $\mathrm{trop}(X)$ of X is a pure d-dimensional polyhedral fan in $\mathbb{R}^n$, satisfying a balancing condition. The task of *tropical implicitization* [ST08, STY06] is to compute $\mathrm{trop}(X)$ from the data in (5.1). This was the goal in the paper [SY08], which includes demonstrations of an implementation called `TrIm`. Theorem 6.4 in [SY08] suggests the following two-step procedure for performing tropical implicitization:

1. Compute the tropicalization of the graph Γ_f of f, given by the n equations in (5.2). The result is the d-dimensional balanced fan $\mathrm{trop}(\Gamma_f)$ in the product space $\mathbb{R}^d \times \mathbb{R}^n$.

2. Project this fan to $\mathbb{R}^n$ and assign appropriate multiplicities to each cone in the image.

This is illustrated in Figure 5.2. Computing $\mathrm{trop}(\Gamma_f)$ in step 1 can be complicated in general. It involves the computation of a *tropical basis* for the ideal generated by $x_1 - f_1(t), \ldots, x_n - f_n(t)$.

In this chapter, we consider two different assumptions on the map f. Both assumptions circumvent the tropical basis computation, and are relevant in practice. First, in Section 5.2, we assume that the functions f_i are Laurent polynomials which are generic with respect to their Newton polytopes. This is the assumption in [ST08, STY06, SY08]. It reduces step 1 above to computing the stable intersection of n codimension one fans in $\mathbb{R}^d \times \mathbb{R}^n$. Second, in Section 5.3, we assume that f is the composition of a linear map $\lambda : \mathbb{C}^d \to \mathbb{C}^\ell$ followed by a Laurent monomial map $\mu : \mathbb{C}^\ell \dashrightarrow \mathbb{C}^n$. In symbols, we have $f = \mu \circ \lambda$. This allows to compute $\mathrm{trop}(X)$ as the linear projection of a tropical linear space in $\mathbb{R}^\ell$. For details see [MS15, Section 5.5]. An important special case arises from the computation of *tropical A-discriminants* [DFS07].

Tropical implicitization is a first step towards classical implicitization. Let $X = V(F)$ be the hypersurface defined by a polynomial $F \in \mathbb{C}[x]$. Then $\mathrm{trop}(X)$ is the union of the $(n-1)$-dimensional cones in the normal fan of the Newton polytope $\mathcal{N}(F)$, decorated with multiplicities. From $\mathrm{trop}(X)$ we can recover $\mathcal{N}(F)$. The key ingredient is a *vertex oracle* which, for a generic weight vector $w \in \mathbb{R}^n$, returns the vertex v of $\mathcal{N}(F)$ which minimizes the dot product with w on $\mathcal{N}(F)$. The algorithm realizing the oracle is suggested by [DFS07, Theorem 2.2]. We provide an implementation using `Oscar.jl` and use it to recover $\mathcal{N}(F)$ via the algorithm in [Hug06]. Once we have the Newton polytope $\mathcal{N}(F)$, we can find F via (numerical) linear algebra. The task is to compute the unique kernel vector of a matrix constructed via *numerical integration* [CGKW00] or *sampling* [BKSW18, EKKB13]. Sampling is preferred when the f_i have rational coefficients. We can then use the parametrization to find rational points on X, and F can be computed using exact arithmetic over $\mathbb{Q}$. However, the size of the matrix is the number of lattice points in $\mathcal{N}(F)$, and we may have to resort to floating point arithmetic when this number is too large. An alternative is rational reconstruction from linear algebra over finite fields. We discuss these techniques in Section 5.4. We use them to solve instances for which elimination via Gröbner bases does not terminate within reasonable time.

If the f_i are Laurent polynomials which are generic with respect to their Newton polytopes, as in Section 5.2, then $\mathcal{N}(F)$ is a *mixed fiber polytope* [EKP07, EK08, STY06]. Our implementation in `Oscar.jl` for computing $\mathcal{N}(F)$ gives a practical way of computing mixed fiber polytopes.

When $\dim X < n-1$, we present a new way of finding its implicit equations from $\mathrm{trop}(X)$. This is the topic of Section 5.5. The idea is to pass through the *Chow form* $\mathrm{Ch}(X)$ of X [DS95]. The polytope we compute is the *Chow polytope* $\mathcal{C}(X)$, which is a linear projection of the Newton polytope $\mathcal{N}(\mathrm{Ch}(X))$. This computation rests on a result by Fink [Fin13], which describes the (weighted) normal fan of $\mathcal{C}(X)$ in terms of $\mathrm{trop}(X)$. We explain how to recover $\mathrm{Ch}(X)$ from $\mathcal{C}(X)$, using the parameterizing functions and an appropriate ansatz. Defining equations for X are obtained from $\mathrm{Ch}(X)$ in the standard manner [DS95, Proposition 3.1].

The implementation of the algorithms supporting this work have benefited from the flexibility provided by `Oscar.jl`. The possibility to combine polyhedral computations with symbolic linear and nonlinear algebra in the same environment has greatly simplified the task. This feature has been our incentive to revisit tropical implicitization. Throughout the chapter, we include several open problems and computational challenges which we hope will inspire the reader to join this effort. This thesis relies heavily on software and data. These materials are made available at `https://mathrepo.mis.mpg.de/TropicalImplicitization`, in the repository `MathRepo` at MPI-MiS [Fev22].

5.2 Generic tropical implicitization

In this section, we start with n Laurent polynomials in n variables with complex coefficients:

$$f_i \,=\, \sum_{a\in A_i} c_{i,a}\, t^a \,\in\, \mathbb{C}[t_1^{\pm 1},\ldots,t_d^{\pm 1}] \quad \text{for} \quad i \,=\, 1,2,\ldots,n.$$

We use these Laurent polynomials in (5.1). The tuple $f = (f_1,\ldots,f_n)$ gives a map $(\mathbb{C}^*)^d \to (\mathbb{C}^*)^n$. Let $X \subset (\mathbb{C}^*)^n$ be the closure of the image of f. Our first task is to find its tropicalization $\mathrm{trop}(X)$. In Section 5.4, we use $\mathrm{trop}(X)$ for classical implicitization. As a set,

$$\mathrm{trop}(X) \,=\, \{w \in \mathbb{R}^n \,:\, \mathrm{in}_w(I(X)) \text{ does not contain a monomial } \} \,\subset \mathbb{R}^n.$$

Here $I(X) \subset \mathbb{C}[x_1^{\pm 1},\ldots,x_n^{\pm 1}]$ is the vanishing ideal of X, and in_w takes the initial ideal with respect to the weight vector w. It is well known that $\mathrm{trop}(X)$ is the support of a fan Σ of dimension $\dim(X)$. This fan is not unique, but for the purposes of this text we can choose any fan Σ with support $\mathrm{trop}(X)$. Assigning a *multiplicity* m_σ to each top dimensional cone $\sigma \in \Sigma$ in the appropriate way [MS15, Definition 3.4.3], the fan Σ is *balanced* [MS15, Theorem 3.4.14]. We will see that these multiplicities are crucial when using $\mathrm{trop}(X)$ for implicitization.

Classically, the variety $X \subset (\mathbb{C}^*)^n$ is the closure of the projection $\Gamma_f \to (\mathbb{C}^*)^n$ of the graph

$$\Gamma_f \,=\, \{(x,t) \in (\mathbb{C}^*)^n \times (\mathbb{C}^*)^d \,:\, x_1 - f_1(t) = 0,\ldots,x_n - f_n(t) = 0\}$$

onto the n x-coordinates. It turns out this has an easy tropical analog.

Theorem 5.2.1. *Let $X = \overline{\mathrm{im}\, f} \subset (\mathbb{C}^*)^n$. The tropical variety $\mathrm{trop}(X)$ is the image of the projection $\mathrm{trop}(\Gamma_f) \to \mathbb{R}^n$, where $\mathrm{trop}(\Gamma_f) \subset \mathbb{R}^n \times \mathbb{R}^d$ is the tropicalization of the graph of f.*

This is an instance of [STY06, Theorem 2.1]. See also [SY08, Theorem 6.4]. We can thus obtain $\mathrm{trop}(X)$ from $\mathrm{trop}(\Gamma_f)$ via a simple projection. However, Theorem 5.2.1 is only useful in practice when $\mathrm{trop}(\Gamma_f)$ is easy to compute. Our next theorem describes $\mathrm{trop}(\Gamma_f)$ under the assumption that the f_i are *generic* with respect to their Newton polytopes $\mathcal{N}(f_i)$. It uses the following notation. For a polytope $P \subset \mathbb{R}^k$ and a vector $w \in (\mathbb{R}^k)^*$, we write $P_w = \{p \in P \,:\, w\cdot p \le w\cdot q \text{ for all } q \in P\}$. In words, P_w is the face of P supported by w.

Theorem 5.2.2. *Suppose f_i is generic with respect to $\mathcal{N}(f_i)$, and let $P_i = \mathcal{N}(x_i - f_i(t)) \subset \mathbb{R}^n \times \mathbb{R}^d$ for $i = 1,\ldots,n$. The tropical variety $\mathrm{trop}(\Gamma_f)$ is the support of a d-dimensional subfan of the normal fan of $P = P_1 + \cdots + P_n$. It consists of the normal cones σ of P for which the face polytopes $(P_1)_w,\ldots,(P_n)_w$ have positive mixed volume MV_w in the affine lattice of P_w, for each $w \in \mathrm{int}(\sigma)$. Moreover, the multiplicity m_σ of σ in $\mathrm{trop}(\Gamma_f)$ equals MV_w.*

This is [ST08, Theorem 4.3]. We illustrate this theorem for a parametric plane curve.

Example 5.2.3. Consider the parametrization $f = (f_1, f_2) : \mathbb{C}^* \longrightarrow (\mathbb{C}^*)^2$ given by

$$f_1 = 11\,t^2 + 5\,t^3 - t^4 \quad \text{and} \quad f_2 = 11 + 11\,t + 7\,t^8.$$

The image is the plane curve $C = \overline{\mathrm{im}\, f}$ given by the implicit equation $F(x,y) = 0$, with

$$F \,=\, 2401\,x^8 - 1372\,x^6y - 422576\,x^5y + \cdots + y^4 + \cdots + 1247565503668. \tag{5.3}$$

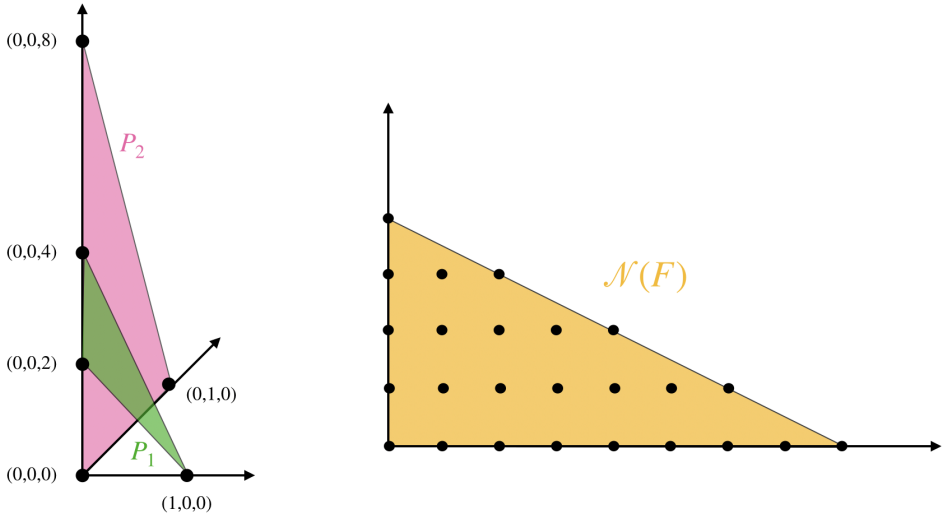


Figure 5.1: Newton polytopes P_1, P_2 and $\mathcal{N}(F)$ from Example 5.2.3.

This has 25 terms, one for each lattice point of $\mathcal{N}(F)$, shown on the right side of Figure 5.1. The Newton polytopes of $x - f_1$ and $y - f_2$ of Γ_f are the triangles seen on the left side.

The tropical curve $\mathrm{trop}(\Gamma_f)$ can be constructed according to Theorem 5.2.2. It is shown in blue on the right of Figure 5.2. The result is a balanced, one-dimensional fan with four rays:

$$\mathrm{trop}(\Gamma_f) \;=\; \mathbb{R}_+ \cdot (1,0,0) \;\cup\; \mathbb{R}_+ \cdot (-4,-8,-1) \;\cup\; \mathbb{R}_+ \cdot (0,1,0) \;\cup\; \mathbb{R}_+ \cdot (2,0,1),$$

with respective multiplicities $2, 1, 8$ and 1. We demonstrate how to obtain these multiplicities.

Consider the primitive ray generator $w = (2,0,1)$, revealing the face polytopes

$$(P_1)_w \;=\; \mathrm{conv}((0,0,2),(1,0,0)) \quad \text{and} \quad (P_2)_w \;=\; \mathrm{conv}((0,0,0),(0,1,0)).$$

The multiplicity of $\mathbb{R}_+ \cdot (2,0,1)$ in $\mathrm{trop}(\Gamma_f)$ can be computed as the mixed volume of the line segments $(P_1)_w$ and $(P_2)_w$ inside the 2-dimensional lattice define by the affine hull of their Minkowski sum. This is the mixed volume MV_w, and we find that it is equal to 1.

Now that we know $\mathrm{trop}(\Gamma_f)$ and its multiplicities (when the Laurent polynomials f_i are generic), and we know that $\mathrm{trop}(X)$ is obtained from its projection, it remains to determine the multiplicities of $\mathrm{trop}(X)$ from those of $\mathrm{trop}(\Gamma_f)$. The answer is given by [ST08, Theorem 1.1], which is the second part of [SY08, Theorem 6.4]. In order to recall the formula, we introduce some more notation. Let Σ_X be a fan in $\mathbb{R}^n$ whose support is $\mathrm{trop}(X)$, and Σ_{Γ_f} a fan in $\mathbb{R}^n \times \mathbb{R}^d$ whose support is $\mathrm{trop}(\Gamma_f)$. Let v be a point in the interior of a top dimensional cone $\sigma_v \in \Sigma_X$. We write $\mathbb{L}_v$ for the linear span of a small open neighborhood of v in $\mathrm{trop}(X)$. Similarly, $w \in \mathrm{int}(\sigma_w)$ for a top dimensional cone $\sigma_w \in \Sigma_{\Gamma_f}$ defines a linear space $\mathbb{L}_w$. If the projection $\Gamma_f \to X$ is generically finite of degree δ, then the multiplicity of $\sigma_v \in \Sigma_X$ is

$$m_{\sigma_v} \;=\; \frac{1}{\delta} \sum_{w \in \pi^{-1}(v)} m_{\sigma_w} \cdot \mathrm{index}\big(\mathbb{L}_v \cap \mathbb{Z}^n : \pi(\mathbb{L}_w \cap \mathbb{Z}^{n+d})\big). \tag{5.4}$$

Here π is the projection $\mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^n$ and the sum is over all points w in the pre-image of v under the map $\pi_{|\mathrm{trop}(\Gamma_f)} : \mathrm{trop}(\Gamma_f) \to \mathrm{trop}(X)$. It is assumed that there are only finitely many such points, and each of them lies in the interior of a top dimensional cone of Σ_{Γ_f}.

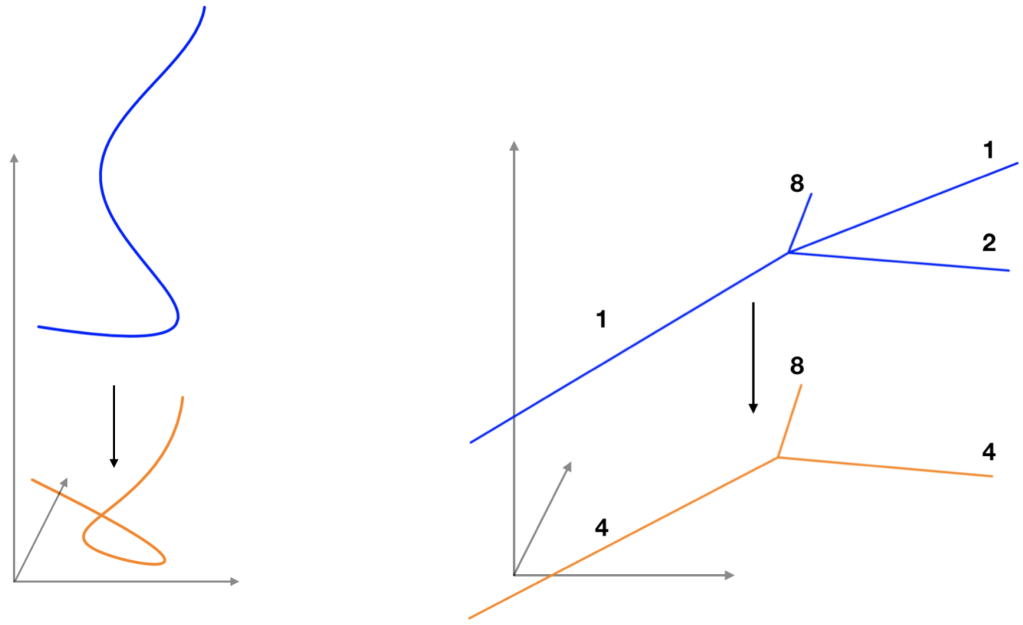


Figure 5.2: Classical (left) and tropical (right) implicitization of a parametric plane curve.

With the choice of weights (5.4), the image fan is balanced. This is a non-trivial fact, derived in a more general setting in [MS15, Lemma 3.6.3]. See also [MS15, Theorem 6.5.16] for a textbook discussion of tropical implicitization in the context of geometric tropicalization.

Example 5.2.4. According to Theorem 5.2.1, the tropical curve $\mathrm{trop}(\Gamma_f)$ projects to $\mathrm{trop}(C)$. This is displayed on the right of Figure 5.2, where $\mathrm{trop}(C)$ is shown in orange as the fan

$$\mathrm{trop}(C) \;=\; \mathbb{R}_+ \cdot (1,0) \;\cup\; \mathbb{R}_+ \cdot (-1,-2) \;\cup\; \mathbb{R}_+ \cdot (0,1).$$

This fan is balanced with ray multiplicities $4, 4, 8$, in that order. We demonstrate the computation of $m_\rho = 4$ for the first ray $\rho = \mathbb{R}_+ \cdot (1,0)$ using (5.4). The ray $\hat{\rho} = \mathbb{R}_+ \cdot (2,0,1)$ of $\mathrm{Trop}(\Gamma_f)$ projects to ρ. Its primitive ray generator $(2,0,1)$ projects to the imprimitive lattice vector $(2,0)$. The contribution of $\hat{\rho} = \mathbb{R}_+ \cdot (2,0,1)$ to the multiplicity m_ρ is the product of two numbers: its intrinsic multiplicity $m_{\hat{\rho}} = 1$, and the lattice index 2. The ray $\mathbb{R}_+ \cdot (1,0,0)$ also projects to ρ, which leads to a total of $m_\rho = 1 \cdot 2 + 2 \cdot 1$. The tropical curve $\mathrm{trop}(C)$ equals the normal fan of the Newton polytope $\mathcal{N}(F)$, displayed on the right of Figure 5.1.

The discussion above leads to Algorithm 1, which makes the results in this section effective. It takes the Newton polytopes $Q_i = \mathcal{N}(f_i)$ as an input, and returns the tropicalization of $X = \overline{\mathrm{im}\, f}$. Here the Laurent polynomials f_i are assumed to be generic with respect to their Newton polytopes Q_i. The output is a set of pairs (m_τ, τ), where $\tau \subset \mathbb{R}^n$ is a cone, and m_τ is a positive integer. The tropical hypersurface $\mathrm{trop}(X)$ is the union of all these cones τ, and the multiplicity of $\mathrm{trop}(X)$ at a generic point x is the sum $\sum_{x \in \tau} m_\tau$.

We warn the reader that, although the union of all cones τ forms the support of a fan, the collection of cones itself is generally not a fan. This representation of a tropical variety is unconventional. However, it is easy to compute and convenient for our algorithmic purposes.

We now explain Algorithm 1. The polytopes $P_1, \ldots, P_n$ in line 3 are the Newton polytopes of the equations $x_1 - f_1, \ldots, x_n - f_n$ of Γ_f. The standard basis $e_1, e_2, \ldots$ of $\mathbb{R}^{n+d}$ is indexed by

Algorithm 1 Generic tropical implicitization

1: **procedure** GETTROPICALCYCLE($Q_1, \dots, Q_n$)
2: **for** $i \in \{1, \dots, n\}$ **do**
3: $P_i \leftarrow \text{conv}(e_i \cup Q_i)$
4: $P \leftarrow P_1 + \cdots + P_n$
5: $\Sigma \leftarrow$ normal fan of P
6: $\text{trop}(X) \leftarrow \emptyset$
7: **for** $\sigma \in \Sigma$ **do**
8: $m_\sigma \leftarrow \text{MV}((P_1)_\sigma, \dots, (P_n)_\sigma)$
9: **if** $m_\sigma > 0$ **then**
10: $\tau \leftarrow \pi(\sigma)$
11: $m_{\text{lattice}} \leftarrow \text{index}(\mathbb{L}_\tau \cap \mathbb{Z}^n : \pi(\mathbb{L}_\sigma \cap \mathbb{Z}^{n+d}))$
12: $\text{trop}(X) \leftarrow \text{trop}(X) \cup \{(m_{\text{lattice}} \cdot m_\sigma, \tau)\}$
13: **return** $\text{trop}(X)$

the variables $x_1, \dots, x_n, t_1, \dots, t_d$ in that order. Following Theorem 5.2.2, Algorithm 1 selects all cones σ in the normal fan of $P_1 + \cdots + P_n$ that contribute to the tropicalization $\text{trop}(\Gamma_f)$. Line 8 computes the mixed volume $m_\sigma = \text{MV}((P_1)_\sigma, \dots, (P_n)_\sigma)$, where $(P_i)_\sigma = (P_i)_w$ for any $w \in \text{int}(\sigma)$. We denote by $\pi(\sigma)$ the projection of $\sigma \subset \mathbb{R}^{n+d}$ to the first n coordinates. The multiplicity with which $\pi(\sigma)$ contributes to $\text{trop}(X)$ is computed in lines 8 and 11. Based on (5.4), it is the product of m_σ with the index of the lattice $\pi(\mathbb{L}_\sigma \cap \mathbb{Z}^{n+d})$ in the lattice $\mathbb{L}_\tau \cap \mathbb{Z}^n$. Here $\mathbb{L}_\sigma$ is the linear span of σ and $\mathbb{L}_\tau = \pi(\mathbb{L}_\sigma)$. We implemented Algorithm 1 in `Julia`.

Example 5.2.5. We show how to apply our `Julia` implementation to Example 5.2.3:

```
using TropicalImplicitization, Oscar
R, (t,) = polynomial_ring(QQ,["t"])
f1 = 11*t^2 + 5*t^3 - 1*t^4
f2 = 11 + 11*t + 7*t^8
Q1 = newton_polytope(f1)
Q2 = newton_polytope(f2)
newton_pols = [Q1, Q2]
cone_list, weight_list = get_tropical_cycle(newton_pols)
```

The lists `cone_list` and `weight_list` returned by our program have four elements each. The first list contains the planar cones

$$\mathbb{R}_+ \cdot (1,0)\,, \quad \mathbb{R}_+ \cdot (1,0)\,, \quad \mathbb{R}_+ \cdot (0,1)\,, \quad \mathbb{R}_+ \cdot (-1,-2),$$

and the second list consists of their respective multiplicities $(2, 2, 8, 4)$. Notice that $\mathbb{R}_+ \cdot (1,0)$ appears twice, and its multiplicity is split up as $4 = 2 + 2$, like in Example 5.2.4.

Problem 5.2.6. Suppose the coefficients of $f_1, \dots, f_n$ lie in a field with a non-trivial valuation, such as the p-adic numbers $\mathbb{Q}_p$ or the Puiseux series $\mathbb{C}\{\{t\}\}$. While the theory of tropical implicitization generalizes nicely to this setting, with balanced fans replaced by balanced polyhedral complexes, useful algorithms and their implementations are yet to be developed.

5.3 A-discriminants

We fix a $d \times n$ integer matrix A of rank d which has the vector $(1, 1, \ldots, 1)$ in its row span. The associated $(d-1)$-dimensional projective toric variety X_A is the closure in $\mathbb{P}^{n-1}$ of the set

$$\{\, (t^{a_1} : t^{a_2} : \cdots : t^{a_n}) \in \mathbb{P}^{n-1} \,:\, t = (t_1, \ldots, t_d) \in (\mathbb{C}^*)^d \,\}. \tag{5.5}$$

Here a_i denotes the ith column of the matrix A. We are interested in the dual variety X_A^*, which parametrizes hyperplanes that are tangent to X_A at some points. Equivalently, X_A^* is the closure in $\mathbb{P}^{n-1}$ of the set of points $x = (x_1 : x_2 : \cdots : x_n)$ such that the hypersurface

$$\{\, t \in (\mathbb{C}^*)^d \,:\, \sum_{i=1}^{n} x_i\, t^{a_i} \,=\, 0 \,\} \tag{5.6}$$

has a singular point. The variety X_A^* is irreducible, and it is usually a hypersurface. The *A-discriminant* Δ_A is the unique (up to scaling) irreducible polynomial vanishing on X_A^*.

In this section we address the following computational problem: given the matrix A, compute its A-discriminant Δ_A. Along the way, we will discover whether X_A^* is not a hypersurface. In this event, we turn to Section 5, and we compute its Chow form instead.

Our algorithm is based on the *Horn uniformization*, which writes X_A^* as the image of a map whose coordinates are products of linear forms. We follow the exposition given in [DFS07]. For additional information, see the book references in [GKZ08, Section 9.3.F] and [MS15, Section 5.5]. Given two vectors u and v in $(\mathbb{C}^*)^n$, we define $u \star v = (u_1 v_1 : u_2 v_2 : \cdots : u_n v_n) \in \mathbb{P}^{n-1}$. If U and V are varieties in $\mathbb{P}^{n-1}$, neither contained in a coordinate hyperplane, then their *Hadamard product* $U \star V$ is the closure of all such points $u \star v$, where $u \in U$ and $v \in V$.

Theorem 5.3.1 (Horn Uniformization)**.** *The dual variety X_A^* is the Hadamard product in $\mathbb{P}^{n-1}$ of the $(d-1)$-dimensional toric variety X_A with an $(n-d-1)$-dimensional linear space:*

$$X_A^* \;=\; X_A \star \operatorname{kernel}(A). \tag{5.7}$$

We illustrate this theorem with several examples. In each of them, we refer to the $(d-1)$-dimensional polytope $Q = \operatorname{conv}(a_1, a_2, \ldots, a_n)$, and we fix an $(n-d) \times n$-matrix B whose rows span the kernel of A. In polytope language, B is a Gale transform of the polytope Q. For (5.7), we introduce unknowns $u = (u_1, \ldots, u_{n-d})$ and we write uB for vectors in kernel(A).

Example 5.3.2 (Determinant)**.** Fix $n = k^2$ and $d = 2k-1$, for some integer $k \geq 2$, and let A represent the linear map that extracts the row sums and column sums of a $k \times k$ matrix. Naively, this matrix has $2k$ rows, but only $2k-1$ of them are linearly independent. Here $Q = \Delta_{k-1} \times \Delta_{k-1}$ is the product of two $(k-1)$-simplices. The toric variety X_A consists of $k \times k$ matrices of rank 1 and X_A^* consists of $k \times k$ matrices of rank $\leq k-1$. We parametrize X_A^* by the Hadamard product of a rank 1 matrix with a matrix whose row and columns are zero. E.g., for $k = 3$, the Horn uniformization writes all singular 3×3 matrices as follows:

$$\begin{bmatrix} t_1 t_4\, u_1 & t_1 t_5 (u_2 - u_1) & t_1 t_6 (-u_2) \\ t_2 t_4 (u_3 - u_1) & t_2 t_5 (u_1 - u_2 - u_3 + u_4) & t_2 t_6 (u_2 - u_4) \\ t_3 t_4 (-u_3) & t_3 t_5 (u_3 - u_4) & t_3 t_6\, u_4 \end{bmatrix}. \tag{5.8}$$

This matrix has $(t_4^{-1}, t_5^{-1}, t_6^{-1})^t$ in its right kernel and $(t_1^{-1}, t_2^{-1}, t_3^{-1})$ in its left kernel. The A-discriminant Δ_A is the determinant of a square matrix, which obviously vanishes on (5.8).

Example 5.3.3 (Resultant)**.** The resultant of a square system of homogeneous polynomials is the A-discriminant where A is the Cayley configuration of the given monomial supports. We examine the Sylvester resultant Δ_A of two binary quadrics ($d = 3, n = 6$). We set

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 2 & 0 & 1 & 2 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \\ 1 & -1 & 0 & -1 & 1 & 0 \end{bmatrix}.$$

This yields the following parametrization for pairs of univariate quadrics with a common zero:

$$\begin{aligned} x_1 + x_2\, z + x_3\, z^2 &= t_1(t_3 u_1\, z - u_1 - u_3)(t_3\, z - 1), \\ x_4 + x_5\, z + x_6\, z^2 &= t_2(t_3 u_2\, z - u_2 - u_3)(t_3\, z - 1). \end{aligned}$$

These Horn uniformizations exist for resultants of polynomials in any number of variables.

Example 5.3.4 (Hyperdeterminant)**.** The hyperdeterminant of a multidimensional tensor vanishes whenever the hypersurface defined by the associated multilinear form is singular. In our notation, this is the A-discriminant Δ_A where the columns of A are the vertices of a product of simplices. As an illustration, we here present the Horn uniformization for the hyperdeterminant of format $2 \times 2 \times 2$. Here $n = 8$ and our configuration is the regular 3-cube:

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \\ 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \end{bmatrix}.$$

These two matrices yield the following map from $\mathbb{C}^8$ into the space of $2 \times 2 \times 2$ tensors

$$\begin{aligned} &x_{000} = t_1(u_1 + u_3 + u_4), \quad x_{001} = t_1 t_4(-u_1 - u_3), \quad x_{010} = t_1 t_3(-u_1 - u_4), \quad x_{011} = t_1 t_3 t_4 u_1, \\ &x_{100} = t_1 t_2(u_2 - u_3 - u_4), \quad x_{101} = t_1 t_2 t_4(u_3 - u_2), \quad x_{110} = t_1 t_2 t_3(u_4 - u_2), \quad x_{111} = t_1 t_2 t_3 t_4 u_2. \end{aligned}$$

Implicitization of this parametrization gives us the hyperdeterminant:

$$\begin{aligned} \Delta_A = {} & x_{000}^2 x_{111}^2 + x_{001}^2 x_{110}^2 + x_{011}^2 x_{100}^2 + x_{010}^2 x_{101}^2 + 4x_{000}x_{011}x_{101}x_{110} + 4x_{001}x_{010}x_{100}x_{111} \\ & - 2x_{000}x_{001}x_{110}x_{111} - 2x_{000}x_{010}x_{101}x_{111} - 2x_{000}x_{011}x_{100}x_{111} \\ & - 2x_{001}x_{010}x_{101}x_{110} - 2x_{001}x_{011}x_{100}x_{110} - 2x_{010}x_{011}x_{100}x_{101}. \end{aligned}$$

We now return to tropical implicitization. Our aim is to compute the tropical variety $\mathrm{trop}(X_A^*)$ directly from A. Here we identify X_A^* with its affine cone in $(\mathbb{C}^*)^n$. If X_A^* has codimension 1 then $\mathrm{trop}(X_A^*)$ is an $(n-1)$-dimensional balanced fan in $\mathbb{R}^n$, with a one-dimensional lineality space. This is the normal fan of the Newton polytope of the A-discriminant Δ_A. We recover the polytope from the fan using Algorithm 3 below; see also [MS15, Remark 3.3.11].

The Horn uniformization of Theorem 5.3.1 gives a convenient way of computing $\mathrm{trop}(X_A^*)$. It is an instance of parametrizations given by monomials in linear forms. These admit an elegant solution to the tropical implicitization problem; see [MS15, Section 5.5]. Let U and V be integer matrices of size $r \times m$ and $s \times r$ respectively. The rows of V are $v_1, \ldots, v_s \in \mathbb{Z}^r$. We denote by λ_U the linear map defined by U, and by μ_V the monomial map specified by V:

$$\lambda_U : (\mathbb{C}^*)^m \dashrightarrow (\mathbb{C}^*)^r \qquad \mu_V : (\mathbb{C}^*)^r \longrightarrow (\mathbb{C}^*)^s$$
$$v \longmapsto U\,v \qquad x \longmapsto (x^{v_1}, \ldots, x^{v_s}).$$

The composition of these maps gives the unirational variety $Y_{U,V} = \overline{\mathrm{im}(\mu_V \circ \lambda_U)}$ in $(\mathbb{C}^*)^s$. Its tropicalization $\mathrm{trop}(Y_{U,V})$ is obtained by tropicalizing the map $\mu_V \circ \lambda_U$. We begin with the tropical linear space $\mathrm{trop}(\mathrm{im}\,\lambda_U)$. This is computed purely combinatorially, as the *Bergman fan* of the matroid of U; see [MS15, Section 4.2]. The monomial map μ_V tropicalizes to the linear map $V : \mathbb{R}^r \to \mathbb{R}^s$. The following result is [DFS07, Theorem 3.1] and [MS15, Theorem 5.5.1].

Theorem 5.3.5. *The tropical variety* $\mathrm{trop}(Y_{U,V})$ *is the image, as a balanced fan via [MS15, Lemma 3.6.3], of the Bergman fan* $\mathrm{trop}(\mathrm{im}\,\lambda_U)$ *under the linear map* $\mathbb{R}^r \to \mathbb{R}^s$ *given by* V.

By Theorem 5.3.1, the affine cone over the A-discriminant in $(\mathbb{C}^*)^n$ is the variety $Y_{U,V}$ with

$$U = \begin{pmatrix} B^t & 0 \\ 0 & I_d \end{pmatrix}, \quad V = \begin{pmatrix} I_n & A^t \end{pmatrix}. \tag{5.9}$$

Here $m = s = n$ and $r = n + d$. This leads to Algorithm 2 for computing $\mathrm{trop}(X_A^*)$.

Algorithm 2 Compute the tropical A-discriminant

1: **procedure** getTropADisc(A)
2: $B \leftarrow$ Gale dual of A
3: $U \leftarrow \begin{pmatrix} B^t & 0 \\ 0 & I_d \end{pmatrix}$
4: $V \leftarrow \begin{pmatrix} I_n & A^t \end{pmatrix}$
5: $M \leftarrow$ matroid of U
6: $\mathrm{trop}(\mathrm{im}\,\lambda_U) \leftarrow$ Bergman fan of M
7: $\mathrm{trop}(X_A^*) \leftarrow \emptyset$
8: **for** $(m_\sigma, \sigma) \in \mathrm{trop}(\mathrm{im}\,\lambda_U)$ **do**
9: $\tau \leftarrow V\sigma$
10: $m_{\text{lattice}} \leftarrow \mathrm{index}(\mathbb{L}_\tau \cap \mathbb{Z}^n : V(\mathbb{L}_\sigma \cap \mathbb{Z}^{n+d}))$
11: $\mathrm{trop}(X_A^*) \leftarrow \mathrm{trop}(X_A^*) \cup \{(m_\sigma \cdot m_{\text{lattice}}, \tau)\}$
12: **return** $\mathrm{trop}(X_A^*)$

The matrix B in line 2 is Gale dual to A. Using the symbolic linear algebra functionality provided by `Oscar.jl`, we find this with the command `nullspace(A)`. Lines 5 and 6 compute the tropicalization $\mathrm{trop}(\mathrm{im}\,\lambda_U)$ of the column span of U. They are based on the `Oscar.jl` commands `Oscar.Polymake.matroid.Matroid(VECTORS = U)` and `Oscar.Polymake.tropical.matroid_fan{min}(matroid)`. From line 8 on, the algorithm computes a projection of the Bergman fan $\mathrm{trop}(\mathrm{im}\,\lambda_U)$. This is analogous to Algorithm 1.

Example 5.3.6. We compute the tropicalized $2 \times 2 \times 2$ hyperdeterminant from Example 5.3.4:

```
A = [1 1 1 1 1 1 1 1; 0 0 0 0 1 1 1 1; 0 0 1 1 0 0 1 1; 0 1 0 1 0 1 0 1]
cone_list, weight_list = get_trop_A_disc(A)
```

The result consists of 32 7-dimensional cones and a list of their multiplicities, constituting the weighted normal fan of the Newton polytope $\mathcal{N}(\Delta_A)$. The following code uses an implementation of Algorithm 3 below. It computes $\mathcal{N}(\Delta_A)$, its lattice points, and its f-vector.

```
Delta = get_polytope_from_cycle(cone_list, weight_list)
f_vec, lattice_pts = f_vector(Delta), lattice_points(Delta)
```

The result is `f_vec` $= (6, 14, 16, 8)$, and `lattice_pts` contains the 12 exponents of Δ_A.

Mixed discriminants [CCD+13] are special cases of A-discriminants. We discuss a non-trivial one.

Example 5.3.7. We revisit [DFS07, Example 5.1]. Here, $d = 4$ and $n = 8$, and we fix the matrix

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 11 & 13 & 17 & 19 \\ 19 & 17 & 13 & 11 & 7 & 5 & 3 & 2 \end{bmatrix}.$$

This represents the following sparse system of two polynomial equations in two variables:

$$x_1 s^2 t^{19} + x_2 s^3 t^{17} + x_3 s^5 t^{13} + x_4 s^7 t^{11} \;=\; x_5 s^{11} t^7 + x_6 s^{13} t^5 + x_7 s^{17} t^3 + x_8 s^{19} t^2 \;=\; 0.$$

The mixed volume of the two Newton polygons equals 39, so we expect 39 common solutions in $(\mathbb{C}^*)^2$. The A-discriminant Δ_A is the condition for two of these solutions to come together. We know from [DFS07, Example 5.1] that Δ_A is a polynomial of degree 126 in $x_1, \ldots, x_8$. Our software computes that $\mathcal{N}(\Delta_A)$ has 43400 lattice points, and f-vector $(45, 92, 63, 16)$.

Problem 5.3.8. Computing the coefficients of Δ_A in Example 5.3.7 amounts to solving a linear system of equations over $\mathbb{Q}$ with 43400 unknowns. This is one of our topics in Section 5.4. Solving that system is hard for at least two reasons. First, systems of this size are beyond the reach of symbolic black box solvers on most personal computers at present. Second, the large condition number and the unbalanced nature of the coefficients of the implicit equation, as in (5.3), hinder the naive use of numerical linear algebra. It is an interesting challenge to develop symbolic or mixed symbolic-numerical techniques for solving such problems.

5.4 Polytope reconstruction and interpolation

Suppose that X is an irreducible hypersurface in $(\mathbb{C}^*)^n$, given by its parametrization (5.1) or (5.7). Using Algorithms 1 and 2, we have computed the tropical variety $\Sigma_X = \mathrm{trop}(X)$. Thus, Σ_X is a balanced fan of dimension $d = n-1$ in $\mathbb{R}^n$, represented by a collection of weighted cones. Our aim in this section is to compute the polynomial $F \in \mathbb{C}[x_1, \ldots, x_n]$ that defines the hypersurface X. We identify X with its closure in $\mathbb{C}^n$. This makes F uniquely defined up to scaling. In particular, the Newton polytope $P = \mathcal{N}(F)$ is uniquely specified.

Before spelling out the details, we summarize our approach. First, we compute the Newton polytope P from the balanced fan Σ_X. This relies on Theorem 5.4.1 below. Second, we find the polynomial F from the parametrization by interpolation. Here we use the ansatz

$$F(x) \;\;=\;\; \sum_{a \,\in\, \mathcal{N}(F) \cap \mathbb{Z}^n} c_a \, x^a, \tag{5.10}$$

and we determine the unknown coefficients c_a by evaluating (5.10) at many points x on X.

We start with computing $P = \mathcal{N}(F)$. The fan Σ_X is dual to the Newton polytope P, namely, it is the $(n-1)$-skeleton of the normal fan of P. Taking into account the multiplicities of all maximal

cones of Σ_X, we can go back and forth between P and Σ_X. Obtaining Σ_X and its multiplicities from P is straightforward. The multiplicity of an $(n-1)$-dimensional cone in Σ_X is the lattice length of the corresponding edge of P. The other direction is more interesting to us: we want to compute P from the output of Algorithm 1. This is discussed in [MS15, Remark 3.3.11]. The main tool is a *vertex oracle*, provided by [DFS07, Theorem 2.2].

Theorem 5.4.1. *Let $X = V(F) \subset \mathbb{C}^n$ be a hypersurface, whose tropicalization $\mathrm{trop}(X) \subset \mathbb{R}^n$ is the support of a fan Σ_X. For a generic weight vector $w \in \mathbb{R}^n$, the vertex $\mathcal{N}(F)_w$ is*

$$\sum_{i=1}^{n} \left(\sum_{\sigma \in \Sigma_X} m_\sigma \cdot \mathrm{IM}(w + \mathbb{R}_+ \cdot e_i, \sigma) \right) \cdot e_i.$$

Here e_i is a standard basis vector, and the inner sum is over all maximal cones of Σ_X.

The intersection multiplicity $\mathrm{IM}(w + \mathbb{R}_+ \cdot e_i, \sigma)$ is the *lattice multiplicity* of the intersection of the ray $\mathbb{R} \cdot e_i$ with the hyperplane $\mathbb{L}_\sigma = \mathbb{R} \cdot \sigma$. This is the absolute value of the determinant of any $n \times n$ matrix whose columns are e_i and a lattice basis for $\mathbb{L}_\sigma \cap \mathbb{Z}^n$. Algorithm 3 implements Theorem 5.4.1. It finds the vertex $\mathcal{N}(F)_w$ from the output of Algorithm 1 or 2. Theorem 5.4.1 and Algorithm 3 are illustrated in Figure 5.3 for the curve in Example 5.2.4.

Algorithm 3 Compute vertex oracle from a tropical hypersurface

```
1: procedure GetVertex(trop(X ∩ (C*)^n), w)
2:     v ← 0
3:     for i ∈ {1, ..., n} do
4:         for (m_σ, σ) ∈ trop(X) do
5:             v ← v + IM(w + R_+ · e_i, σ) · m_σ · e_i
6:     return v
```

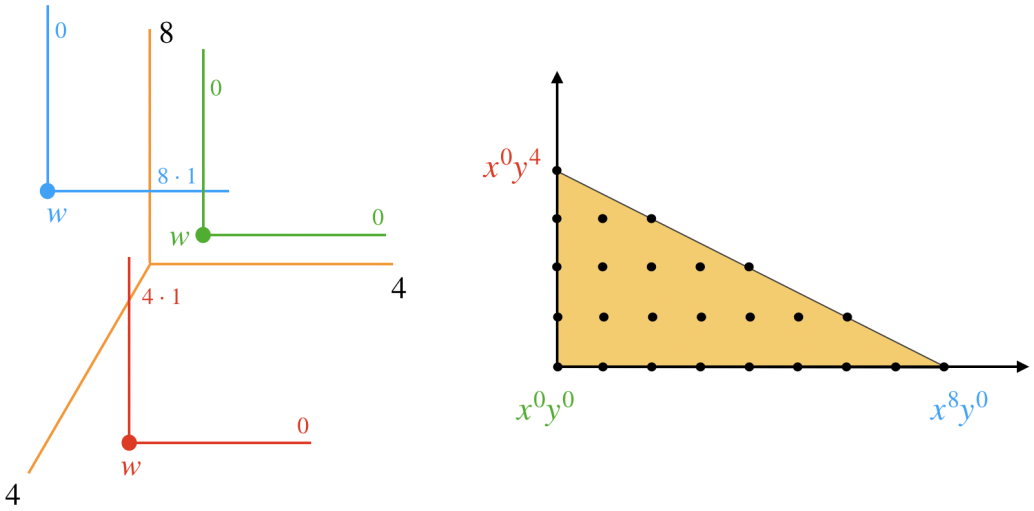


Figure 5.3: Computing vertices of $\mathcal{N}(F)$ by intersecting $\mathrm{trop}(X)$ with $w + \mathbb{R}_{\geq 0} \cdot e_i$.

Algorithm 3 can be used to compute *all* vertices of $\mathcal{N}(F)$. A naive approach applies the vertex oracle to many random vectors w. However, it is not clear how many w would be needed, and

which stop criterion to use. A deterministic way of constructing $\mathcal{N}(F)$ using a vertex oracle like Algorithm 3 was proposed by Huggins [Hug06]. Our implementation uses that.

Example 5.4.2. The following `Julia` code computes a polytope from a tropical hypersurface:

```
Delta = get_polytope_from_cycle(cone_list, weight_list)
```

If the variables `cone_list`, `weight_list` are carried over from Example 5.2.5, then `Delta` is the yellow polytope shown in Figure 5.3. For `cone_list`, `weight_list` from Example 5.3.6, the polytope `Delta` is the Newton polytope of the hyperdeteminant Δ_A from Example 5.3.4.

Remark 5.4.3. Under the assumptions of Section 5.2, i.e., the f_i are Laurent polynomials which are generic with respect to their Newton polytopes, $\mathcal{N}(F)$ is a *mixed fiber polytope*. This was discovered independently by several authors [EKP07, EK08, STY06]. For instance, in Figure 5.1, $\mathcal{N}(F)$ is the mixed fiber polytope of P_1 and P_2. Our implementation of Huggins' algorithm [Hug06] combined with Algorithm 3 provides a practical way of computing mixed fiber polytopes. This includes the computation of fiber polytopes and secondary polytopes [SY08, Section 3].

Once the Newton polytope $\mathcal{N}(F)$ of the defining equation $F = 0$ of X is known, we can obtain its coefficients c_a in (5.10) using linear algebra. The set $\mathcal{B} = \mathcal{N}(F) \cap \mathbb{Z}^n$ is a superset of the monomial support $\operatorname{supp}(F)$ of F. It can be computed in `Oscar.jl` via the command `lattice_points`. The interpolation method is most efficient when $\mathcal{B}$ is not much larger than $\operatorname{supp}(F)$, that is, few of the c_a in (5.10) are zero. The inclusion $\mathcal{B} \supseteq \operatorname{supp}(F)$ can be strict:

Example 5.4.4. Consider the map $f : \mathbb{C} \to \mathbb{C}^2$ given by $f_1(t) = a_1 t^4 + a_2 t$ and $f_2(t) = a_3 t^2 + a_4 t$ for generic complex numbers a_1, a_2, a_3, a_4. Here the implicit polynomial $F(x, y)$ equals

$$a_1^2 y^4 - 2a_1 a_3^2 xy^2 + a_3^4 x^2 - 4a_1 a_3 a_4^2 xy + 3a_1 a_2 a_3 a_4 y^2 - a_4(a_1 a_4^3 - a_2 a_3^3)x + a_2(a_1 a_4^3 - a_2 a_3^3)y.$$

Note that the term y^3 does not appear, in spite of it being in $\mathcal{N}(F)$. This shows that some lattice points in a predicted Newton polytope may never appear with nonzero coefficient.

Problem 5.4.5. We propose to refine the observation in Remark 5.4.3 by predicting the monomial support of F from the monomial support of $f_1, \ldots, f_n$. That is, which lattice points in the mixed fiber polytope, other than its vertices, contribute to the implicit equation?

For simplicity, we work with the superset $\mathcal{B} \supseteq \operatorname{supp}(F)$ and allow some coefficients to be zero. We identify a set $\mathcal{P}$ of m points in X, so that the interpolation conditions $F(p) = 0$ for $p \in \mathcal{P}$ uniquely determine F (up to a constant factor). We obtain $\mathcal{P} \subset X$ by sending random points in $\mathbb{C}^d$ through the parametrization (5.1). The unknown coefficients c_a in (5.10) form a vector $c = (c_a)_{a \in \mathcal{B}} \in \mathbb{C}^{\mathcal{B}}$. For each point $p \in \mathbb{C}^n$, let $p^{\mathcal{B}} = (p^a)_{a \in \mathcal{B}}$ be the vector of monomials corresponding to $\mathcal{B}$, evaluated at p. We interpret $p^{\mathcal{B}}$ as an element of the dual vector space $(\mathbb{C}^{\mathcal{B}})^*$. With this set-up, c is the unique vector (up to scaling) satisfying

$$p^{\mathcal{B}} \cdot c = 0 \quad \text{for all } p \in X.$$

If the sample points $\mathcal{P} \subset X$ are sufficiently random and $m \geq |\mathcal{B}| - 1$, this is equivalent to

$$p^{\mathcal{B}} \cdot c = 0, \quad \text{for all } p \in \mathcal{P}.$$

The *Vandermonde matrix* $M(\mathcal{B}, \mathcal{P})$ has the vectors $p^{\mathcal{B}}$ for its rows, where $p \in \mathcal{P}$. It has size $m \times |\mathcal{B}|$ and, by the above discussion, the kernel of $M(\mathcal{B}, \mathcal{P}) : \mathbb{C}^{\mathcal{B}} \to \mathbb{C}^m$ is spanned by c.

Our problem is now reduced to the computation of the one-dimensional kernel of a Vandermonde matrix $M(\mathcal{B}, \mathcal{P})$. Below, we will fix $\mathcal{B}$ and $\mathcal{P}$ and use the simpler notation $M = M(\mathcal{B}, \mathcal{P})$, where there is no danger for confusion. When the parametrizing functions f_i have coefficients in $\mathbb{Q}$, like in the case of A-discriminants in Section 5.4, we can use $\mathcal{P} \subset \mathbb{Q}^n$. In particular, M has rational entries, and its kernel can be computed in exact arithmetic.

Example 5.4.6. We now demonstrate our implementation of the above discussion by computing the implicit equation F from Example 5.2.3. The following code computes the Vandermonde matrix $M(\mathcal{B}, \mathcal{P})$ of size 24 by 25 with rational entries. This is done by plugging 24 random rational numbers into the parametrization (5.1). The functions `f1`, `f2` are taken from Example 5.2.5, and the Newton polytope `Delta` $= \mathcal{N}(F)$ was computed in Example 5.4.2.

```
B = lattice_points(Delta)
n_samples = length(B)-1
P = sample([f1,f2], n_samples)
M_BP = get_vandermonde_matrix(B,P)
coeffs_F = nullspace(M_BP)[2]
```

Up to scaling, `coeffs_F` consists of the 25 coefficients of F. Some are shown in (5.3).

Often, in practical computations, the points $p \in \mathcal{P}$ are approximations of points on X, so the entries of M are finite precision floating point numbers. In that case, the task of computing $\ker M$ is one of *numerical linear algebra*. This is not supported in the current version of `Oscar.jl`. The standard way to proceed using, for instance, the numerical linear algebra functionality in `Julia`, is via the *singular value decomposition* (SVD) of M. Alternatives include *QR factorization with optimal pivoting* and *iterative eigenvalue methods*. We refer to [BKSW18, Section 5] for such numerical considerations and pointers to the relevant literature.

When f is defined over $\mathbb{Q}$, one might still want to use floating point computations for speed. Let c_a be a nonzero entry of a generator c for $\ker M$. The vector $c_a^{-1}c$ has rational entries. Its numerical approximation $\tilde{c}_a^{-1}\tilde{c}$ is contaminated by rounding errors. We approximate the entries of $\tilde{c}_a^{-1}\tilde{c}$ by rational numbers using the built in function `rationalize` in `Julia`. This has an optional input `tol`, so that `rationalize(a,tol = e)` returns a rational number `q` which satisfies $|\mathtt{q} - \mathtt{a}| \leq \mathtt{e}$. A sensible choice for `tol` is $100 \cdot \tilde{c}_a^{-1} \cdot \varepsilon \cdot \sigma_1/\sigma_{|\mathcal{B}|-1}$.

If symbolic computation is preferable to numerical methods, then one might solve the linear equations over various finite fields and recover rational solutions via the Chinese remainder theorem. This can be done in a computer algebra system. Sometimes, one is only interested in a fixed finite field. We illustrate the finite field computation in `Oscar.jl`.

Example 5.4.7. We seek the A-discriminant for a matrix whose entries are large integers:

$$A \quad = \quad \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 11 & 13 \\ 13 & 8 & 5 & 3 & 2 & 1 \end{bmatrix}$$

The following code finds that the Newton polytope of Δ_A over $\mathbb{Q}$ has dimension 3 and f-vector $(12, 18, 8)$. It terminated on a MacBook Pro with a 3,3 GHz Intel Core i5 processor within 120 seconds. The number of lattice points equals 2295. In order to compute the coefficients of the A-discriminant, we must solve a linear system of 2294 equations with large integer coefficients. We solve this over the field with 101 elements instead:

```
A = [1 1 1 1 1 1; 2 3 5 7 11 13; 13 8 5 3 2 1];
cone_list, weight_list = get_trop_A_disc(A);
Delta = get_polytope_from_cycle(cone_list, weight_list);
@time mons, coeffs = compute_A_discriminant(A, Delta, GF(101));
```

For the same computation over the rational numbers, the machine ran out of memory.

We close this section with a combinatorics problem that arises naturally from Remark 5.4.3.

Problem 5.4.8. Let $P_1, \ldots, P_n$ be polytopes in $\mathbb{R}^{n-1}$ having $v_1, \ldots, v_n$ vertices. Give a sharp upper bound in terms of $v_1, \ldots, v_n$ for the number of vertices of their mixed fiber polytope. In other words, prove an Upper Bound Theorem for f-vectors arising in tropical implicitization.

Example 5.4.9. We illustrate Problem 5.4.8 for three triangles ($n = v_1 = v_2 = v_3 = 3$). After many runs for different random configurations, the following example is our current winner:

```
verts1 = [898 -614; -570 817; 892 -594]
verts2 = [-603 -481; -623 -127; -36 732]
verts3 = [-548 -864; -151 873; 800 -861]
(T1,T2,T3) = convex_hull.([verts1, verts2, verts3])
Delta = get_polytope_from_cycle(get_tropical_cycle([T1,T2,T3])...)
f_vec = f_vector(Delta)
```

This code computes a mixed fiber polytope that has 25 vertices, 49 edges and 26 facets. Can you find three triangles in $\mathbb{R}^2$ whose mixed fiber polytope has more than 25 vertices?

5.5 Higher codimension

In this section we address the implicitization problem for varieties X that are not hypersurfaces. The role of the Newton polytope $\mathcal{N}(F)$ of a polynomial F will now be played by the Chow polytope $\mathcal{C}(X)$. We begin by reviewing some definitions from [DS95] and [GKZ08, Chapter 6].

Let X be an irreducible projective variety of dimension d in complex projective space $\mathbb{P}^n$. Suppose we are given the tropical variety $\mathrm{trop}(X)$, a balanced fan of dimension d in $\mathbb{R}^{n+1}/\mathbb{R}\mathbf{1}$. Our goal is to compute the *Chow form* $\mathrm{Ch}(X)$, which is a hypersurface in the Grassmannian $\mathrm{Gr}(n-d-1, \mathbb{P}^n)$. Its points are the linear subspaces of dimension $n-d-1$ whose intersection with X is non-empty. We identify $\mathrm{Ch}(X)$ with its defining polynomial of degree $\deg(X)$ in primal Plücker coordinates $p_{i_0 i_1 \cdots i_d}$, where $1 \leq i_0 < i_1 < \cdots i_d \leq n$. The $p_{i_0 i_1 \cdots i_d}$ are the maximal minors of any $(d+1) \times (n+1)$ matrix whose kernel is the subspace. The Chow form $\mathrm{Ch}(X)$ is only well-defined up to the Plücker relations that vanish on $\mathrm{Gr}(n-d-1, \mathbb{P}^n)$. By [Stu08, Theorem 3.1.7], $\mathrm{Ch}(X)$ is a unique linear combination of standard tableaux. In our computations, we always use that standard representation for Chow forms.

The weight of the Plücker coordinate $p_{i_0 i_1 \cdots i_d}$ is the vector $e_{i_0} + e_{i_1} + \cdots + e_{i_d}$ in $\mathbb{Z}^n$, and the weight of a Plücker monomial is the sum of the weights of its variables, with multiplicity. By definition, the Chow polytope $\mathcal{C}(X)$ is the convex hull of the weights occurring in $\mathrm{Ch}(X)$.

Fink [Fin13] gave a combinatorial recipe for constructing the weighted normal fan of the Chow polytope $\mathcal{C}(X)$ from the tropical variety $\mathrm{trop}(X)$. Let L_{n-d-1} denote the standard tropical linear space of dimension $n-d-1$ in $\mathbb{R}^{n+1}/\mathbb{R}\mathbf{1}$. Its maximal cones are the orthants spanned by $(n-d-1)$-tuples of unit vectors. It is proved in [Fin13, Theorem 4.8] that the weighted normal fan of $\mathcal{C}(X)$

is the *stable sum* of $\mathrm{trop}(X)$ with the negated linear space $-L_{n-d-1}$. The stable sum is a dual operation to the stable intersection. It always produces a balanced fan of expected dimension. Hence $\mathrm{trop}(X) - L_{n-d-1}$ is a balanced fan of codimension 1 in $\mathbb{R}^{n+1}/\mathbb{R}\mathbf{1}$. Fink's result states that this is the *outer* normal fan of $\mathcal{C}(X)$.

We can compute $\mathcal{C}(X)$ from $\mathrm{trop}(X) - L_{n-d-1}$ by the algorithm for building Newton polytopes in Section 4, *up to an integer translation*. Indeed, the normal fan of $\mathcal{C}(X)$ and $\mathcal{C}(X)+\mathbf{t}$ are identical, for any $\mathbf{t} \in \mathbb{Z}^n$. Algorithm 3 finds vertices of $\mathcal{C}(X) + \mathbf{t}$, where $\mathbf{t}$ shifts $\mathcal{C}(X)$ so that it touches each coordinate hyperplane. In previous examples, we had $\mathbf{t} = 0$. Indeed, if F is irreducible, then the polytope $\mathcal{N}(F)$ touches all coordinate hyperplanes. This is not true for the Chow polytope, as illustrated by the example below. Finding the correct $\mathbf{t}$ is an interesting combinatorial problem which we plan to investigate in a future project.

Example 5.5.1 $(d = 1, n = 3)$**.** Let X be the curve in $\mathbb{C}^3$ which is given by the parametrization

$$x_1 = t(t-1)(t+1), \quad x_2 = t^2(t+1), \quad x_3 = t^3(t-1).$$

The tropical curve is determined by the orders of the coordinate functions at all zeros and poles. Hence $\mathrm{trop}(X)$ is the fan with four rays $(1,2,3)$, $(1,1,0)$, $(1,0,1)$ and $(-3,-3,-4)$. We identify X with its projective closure in $\mathbb{P}^3$, obtained by adding an extra coordinate x_0. The tropical line L_1 is spanned by e_0, e_1, e_2, e_3, and we form the sum of $\mathrm{trop}(X)$ with the negated line $-L_1$. This 2-dimensional fan is the normal fan of the Chow polytope $\mathcal{C}(X)$.

We implemented the stable sum using `Oscar.jl`, and obtain this fan as follows.

```
cone_list = positive_hull.([[1, 1, 0], [1, 2, 3], [1,0,1], [-1, -1, -4//3]])
weight_list = ones(Int64, 4)
cone_list, weight_list = get_chow_fan(cone_list, weight_list)
```

The output consists of 16 2-dimensional cones and their multiplicities. A translated version of the Chow polytope is obtained from this output as in the previous section:

```
C_translated = get_polytope_from_cycle(cone_list, weight_list)
```

This is a three-dimensional polytope touching all coordinate hyperplanes. It has vertices

$$(0,2,3,1), (0,3,1,2), (0,4,1,1), (1,0,4,1), (1,2,3,0), (1,3,0,2),$$
$$(1,4,0,1), (1,4,1,0), (2,0,1,3), (2,0,4,0), (2,4,0,0), (3,0,0,3).$$

To identify the shift $\mathbf{t}$, we compare this to the Chow polytope. We obtain $\mathcal{C}(X)$ as the convex hull of the weights of the Plücker monomials in the Chow form $\mathrm{Ch}(X)$ of our curve:

$$\begin{aligned}
&p_{03}^4 - p_{01}^3p_{13} - 3p_{01}^2p_{02}p_{13} - 3p_{01}p_{02}^2p_{13} - p_{02}^3p_{13} + 3p_{01}^2p_{03}p_{13} + 9p_{01}p_{02}p_{03}p_{13} + 6p_{02}^2p_{03}p_{13}\\
&+p_{01}p_{03}^2p_{13} - 5p_{02}p_{03}^2p_{13} + 2p_{01}^2p_{12}p_{13} + p_{01}p_{02}p_{12}p_{13} + 2p_{01}^2p_{13}^2 - 2p_{01}p_{02}p_{13}^2 + 4p_{02}^2p_{13}^2\\
&+p_{01}p_{03}p_{13}^2 - 4p_{01}p_{12}p_{13}^2 - p_{01}^3p_{23} - 3p_{01}^2p_{02}p_{23} - 3p_{01}p_{02}^2p_{23} - p_{02}^3p_{23} + 4p_{01}^2p_{03}p_{23}\\
&+11p_{01}p_{02}p_{03}p_{23} + 7p_{02}^2p_{03}p_{23} - 2p_{01}p_{03}^2p_{23} - 10p_{02}p_{03}^2p_{23} + 2p_{03}^3p_{23} + 2p_{01}^2p_{12}p_{23}\\
&+p_{01}p_{02}p_{12}p_{23} + 9p_{01}^2p_{13}p_{23} - p_{01}p_{02}p_{13}p_{23} + 6p_{02}^2p_{13}p_{23} + 2p_{01}p_{03}p_{13}p_{23} - 2p_{02}p_{03}p_{13}p_{23}\\
&-6p_{01}p_{12}p_{13}p_{23} + 2p_{01}p_{13}^2p_{23} + 9p_{01}^2p_{23}^2 + 2p_{01}p_{02}p_{23}^2 + 2p_{02}^2p_{23}^2 - 4p_{01}p_{03}p_{23}^2 - 2p_{01}p_{12}p_{23}^2.
\end{aligned}$$

We find that $\mathcal{C}(X)$ is the 3-dimensional polytope with the following 12 vertices:

$$(1,2,3,2), (1,3,1,3), (1,4,1,2), (2,0,4,2), (2,2,3,1), (2,3,0,3),$$
$$(2,4,0,2), (2,4,1,1), (3,0,1,4), (3,0,4,1), (3,4,0,1), (4,0,0,4).$$

We conclude that $\mathbf{t} = (-1,0,0,-1)$. For now, we apply this shift manually.

After listing all lattice points in $\mathcal{C}(X)$, we can compute the Chow form $\mathrm{Ch}(X)$ by interpolation. This is done as follows. For each lattice point u in $\mathcal{C}(X)$ we list all standard Plücker monomials of weight u, and form their linear combination with unknown coefficients. Our ansatz is the sum of these $\mathbb{Z}^n$-homogeneous Plücker polynomials, with distinct unknown coefficients. We generate random points on the Chow hypersurface as follows. Pick a random point in X and a random linear space of dimension $n-d-1$ through that point. We read off the Plücker coordinates of that linear space and substitute them into the ansatz. Repeating this process many times gives the desired linear system of equations in the unknown coefficients. Up to scaling, this system has a unique solution, namely the Chow form $\mathrm{Ch}(X)$.

Example 5.5.2. We use this strategy to recover the Chow form $\mathrm{Ch}(X)$ from Example 5.5.1. For each lattice point u in the polytope $\mathcal{C}(X)$, we form the general linear combination of standard Plücker monomials of weight u. For instance, for $u = (2,2,2,2)$ this linear combination is

$$\gamma_{u,1} \cdot p_{01}^2 p_{23}^2 \; + \; \gamma_{u,2} \cdot p_{01}p_{02}p_{13}p_{23} \; + \; \gamma_{u,3} \cdot p_{02}^2 p_{13}^2.$$

Our ansatz for the Chow form $\mathrm{Ch}(X)$ is the sum of these expressions over all $u \in \mathcal{C}(X) \cap \mathbb{Z}^4$.

We sample from the Chow hypersurface by picking random matrices of the form

$$\begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 \\ 1 & t(t-1)(t+1) & t^2(t+1) & t^3(t-1) \end{bmatrix}.$$

The 2×2 minors of this matrix are the dual Plücker coordinates of the corresponding sample point in $\mathrm{Ch}(X) \subset \mathrm{Gr}(1,\mathbb{P}^3)$. We read off its primal Plücker coordinates as follows:

$$\begin{array}{ccc} p_{01} = (t^4-t^3)\alpha_2 + (t^3+t^2)\alpha_3, & p_{02} = (t^3-t^4)\alpha_1 + (t^3-t)\alpha_3, & p_{03} = (t^3+t^2)\alpha_1 + (t-t^3)\alpha_2, \\ p_{12} = (t^4-t^3)\alpha_0 - \alpha_3, & p_{13} = (t^3+t^2)\alpha_0 + \alpha_2, & p_{23} = (t^3-t)\alpha_0 - \alpha_1. \end{array}$$

We substitute many such sample points into the ansatz, and we solve the resulting system of linear equations for the unknown coefficients $\gamma_{u,i}$. The output is the desired Chow form. This yields defining equations for X by setting $p_{ij} = \alpha_i x_j - \alpha_j x_i$ for any $\alpha_0, \ldots, \alpha_3 \in \mathbb{Q}$.

5.6 Conclusion

In this chapter, we discussed an application of tropical geometry to computer algebra, namely implicitization with tropical preprocessing. Guided by many examples it was shown that `Oscar.jl` provides excellent capabilities for performing tropical implicitization in practice. Our implementation in `Oscar` realizes the vision in [SY08] and fulfils the promise made by `TrIm`. In particular, it computes the tropicalization of unirational varieties in many instances. Furthermore, it uses this tropical data in conjunction with numerical interpolation to compute defining equations of A-discriminants. In Section 5 we ventured into a setting where the desired hypersurface is not in an affine or projective space, but inside a Grassmannian. This suggests yet one more problem for future research.

Problem 5.6.1. Many applications lead to interesting subvarieties of Grassmannians. For instance, in computer vision, certain cameras are represented by curves and surfaces in $\mathrm{Gr}(1,\mathbb{P}^3)$. Their tropicalizations lie inside the tropical Grassmannian, and their cohomology classes are computed by Schubert calculus. It would be desirable to develop tropical implicitization in the setting when the ambient spaces are Grassmannians, or even flag varieties.

Bibliography

[Aba67] Jean Abadie. On the Kuhn-Tucker theorem. In *Nonlinear Programming (NATO Summer School, Menton, (1964)*, pages 19–36. North-Holland, Amsterdam, 1967.

[ABF+23] Daniele Agostini, Taylor Brysiewicz, Claudia Fevola, Lukas Kühne, Bernd Sturmfels, Simon Telen, and Thomas Lam. Likelihood degenerations. *Advances in Mathematics*, 414:108863, 02 2023.

[AH18] Paolo Aluffi and Corey Harris. The Euclidean distance degree of smooth complex projective varieties. *Algebra Number Theory*, 12(8):2005–2032, 2018.

[ABZ06] Christpoher Amato, Daniel S. Bernstein and Shlomo Zilberstein. Solving POMDPs using quadratically constrained linear programs. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 341–343, 2006.

[ABB+19] Carlos Améndola, Nathan Bliss, Isaac Burke, Courtney R. Gibbons, Martin Helmer, Serkan Hoşten, Evan D. Nash, Jose Israel Rodriguez, and Daniel Smolkin. The maximum likelihood degree of toric varieties. *J. Symbolic Comput.*, 92:222–242, 2019.

[AYA18] Kamyar Azizzadenesheli, Yisong Yue and Animashree Anandkumar. Policy gradient in partially observable environments: Approximation and convergence, 2018. Preprint, arXiv:1810.07900.

[BD15] Jasmijn A. Baaijens and Jan Draisma. Euclidean distance degrees of real algebraic groups. *Linear Algebra Appl.*, 467:174–187, 2015.

[BM22] Lorenzo Baldi and Bernard Mourrain. Exact moment representation in polynomial optimization, 2022.

[BHSW] Daniel J. Bates, Jonathan D. Hauenstein, Andrew J. Sommese, and Charles W. Wampler. Bertini: Software for numerical algebraic geometry.

[BHSW13] Daniel J. Bates, Jonathan D. Hauenstein, Andrew J. Sommese, and Charles W. Wampler. *Numerically solving polynomial systems with Bertini*, volume 25 of *Software, Environments, and Tools*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2013.

[BSS06] Mokhtar S. Bazaraa, Hanif D. Sherali, and C.M. Shetty. *Nonlinear programming*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition, 2006. Theory and algorithms.

[Bel57] Richard Bellman,. A Markovian decision process. *Journal of mathematics and mechanics*, 6(5):679–684, 1957.

[Bel66] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[BT21] Matías R. Bender and Simon Telen. Yet another eigenvalue algorithm for solving polynomial systems. *arXiv preprint arXiv:2105.08472*, 2021.

[BS] Luis Benet and David P. Sanders. IntervalRootFinding.jl. `https://juliaintervals.github.io/IntervalRootFinding.jl`.

[Ber75] David N. Bernstein. The number of roots of a system of equations. *Funkcional. Anal. i Priložen.*, 9(3):1–4, 1975.

[Ber97] Dimitri P. Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

[BR19] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *Preprint,* arXiv:1906.01786, 2019.

[BPS21] Tobias Boege, Sonja Petrović and Bernd Sturmfels. Marginal independence models. 2021. Preprint, arXiv:2112.10287.

[BHIM22] Paul Breiding, Reuven Hodges, Christian Ikenmeyer and Mateusz Michałek. Equations for gl invariant families of polynomials. *Vietnam Journal of Mathematics*, pages 1–12, 2022.

[BKSW18] Paul Breiding, Sara Kališnik, Bernd Sturmfels and Madeleine A. Weinstein. Learning algebraic varieties from samples. *Revista Matemática Complutense*, 31:545–593, 2018.

[BRST23] Paul Breiding, Felix Rydell, Elima Shehu and Angélica Torres. Line multiview varieties. *SIAM Journal on Applied Algebra and Geometry*, 7(2):470–504, 2023.

[BRT23] Paul Breiding, Kemal Rose and Sascha Timme. Certifying zeros of polynomial systems using interval arithmetic. *ACM Transactions of Mathematical Software*, 49(1):14, 2023.

[BST20] Paul Breiding, Bernd Sturmfels and Sascha Timme. 3264 Conics in a Second. *Notices of the American Mathematical Society*, 67:30–37, 2020.

[BSW21] Paul Breiding, Frank Sottile and JamesWoodcock. Euclidean distance degree and mixed volume. *Foundations of Computational Mathematics*, 09 2021.

[BT18] Paul Breiding and Sascha Timme. HomotopyContinuation.jl: A package for homotopy continuation in julia. In *Mathematical Software – ICMS 2018*, pages 458–465, Cham, 2018. Springer International Publishing.

[BKK20] Taylor Brysiewicz, Khazhgali Kozhasov and Mario Kummer. Nodes on quintic spectrahedra. *arXiv preprint arXiv:2011.13860*, 2020.

[BFS21] Taylor Brysiewicz, Claudia Fevola and Bernd Sturmfels. Tangent quadrics in real 3-space. *Le Matematiche*, 76(2):355–367, 2021.

[BCS13] Peter Bürgisser, Michael Clausen and Mohammad A. Shokrollahi. *Algebraic complexity theory*, volume 315. Springer Science & Business Media, 2013.

[BLL19] Michael Burr, Kisun Lee and Anton Leykin. Effective certification of approximate solutions to systems of equations involving analytic functions. In *Proceedings of the 2019 on International Symposium on Symbolic and Algebraic Computation*, ISSAC '19, pages 267–274, New York, NY, USA, 2019. Association for Computing Machinery.

[CHKS06] Fabrizio Catanese, Serkan Hoşten, Amit Khetan and Bernd Sturmfels. The maximum likelihood degree. *Amer. J. Math.*, 128(3):671–697, 2006.

[CCD+13] Eduardo Cattani, María Angélica Cueto, Alicia Dickenstein, Sandra Di Rocco and Bernd Sturmfels. Mixed discriminants. *Mathematische Zeitschrift*, 274(3-4):761–778, 2013.

[CJM+21] Türkü Özlüm Celik, Asgar Jamneshan, Guido Montúfar, Bernd Sturmfels and Lorenzo Venturello. Wasserstein distance to independence models. *Journal of Symbolic Computation*, 104:855–873, 2021.

[CLL14] Tianran Chen, Tsung-Lin Lee and Tien-Yien Li. Hom4PS-3: A parallel numerical solver for systems of polynomial equations based on polyhedral homotopy continuation methods. In Hoon Hong and Chee Yap, editors, *Mathematical Software – ICMS 2014*, pages 183–190. Springer Berlin Heidelberg, 2014.

[Che68] Hermann Chernoff. Optimal stochastic control. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 221–252, 1968.

[CGKW00] Robert M Corless, Mark W Giesbrecht, Ilias S Kotsireas and Stephen M Watt. Numerical implicitization of parametric hypersurfaces with linear algebra. In *International Conference on Artificial Intelligence and Symbolic Computation*, pages 174–183. Springer, 2000.

[CLS11] David A. Cox, John B. Little and Henry K. Schenck. *Toric varieties*, volume 124 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2011.

[Cox95] David A. Cox. The homogeneous coordinate ring of a toric variety. In *International Conference on Machine Learning*, pages 1486–1495. 2019.

[DTLR+19] Robert Dadashi, Adrien A. Taiga, Nicolas Le Roux, Dale Schuurmans and Marc G. Bellemare The value function polytope in reinforcement learning. In *Algebraic Geometry*, pages 17–50. PMLR, 1995.

[DA21] Joachim Dahl and Erling D. Anderse. A primal-dual interior-point algorithm for nonsymmetric exponential-cone optimization. *Math. Program. 194, 341–370 (2022).*

[DS95] John Dalbec and Bernd Sturmfels. Introduction to Chow forms. In *Invariant Methods in Discrete and Computational Geometry: Proceedings of the Curaçao Conference, 13–17 June, 1994*, pages 37–58. Springer, 1995.

[DM21] Harm Derksen and Visu Makam. Maximum likelihood estimation for matrix normal models via quiver representations. *SIAM J. Appl. Algebra Geom.*, 5(2):338–365, 2021.

[Der70] Cyrus Derman. *Finite state Markovian decision processes*. Academic Press, Inc., USA, 1970.

[DFS07] Alicia Dickenstein, Eva Feichtner and Bernd Sturmfels. Tropical discriminants. *Journal of the American Mathematical Society*, 20(4):1111–1133, 2007.

[DHO+14] Jan Draisma, Emil Horobeţ, Giorgio Ottaviani, Bernd Sturmfels and Rekha Thomas. The Euclidean distance degree. In *SNC 2014—Proceedings of the 2014 Symposium on Symbolic-Numeric Computation*, pages 9–16. ACM, New York, 2014.

[DHO+16] Jan Draisma, Emil Horobeţ, Giorgio Ottaviani, Bernd Sturmfels and Rekha R. Thomas. The Euclidean distance degree of an algebraic variety. *Found. Comput. Math.*, 16(1):99–149, 2016.

[DGLM+24] Mareike Dressler, Marina Garrote-López, Guido Montúfar, Johannes Müller and Kemal Rose. Algebraic optimization of sequential decision problems. *Journal of Symbolic Computation*, 121:102241, 2024.

[DLOT17] Dmitriy Drusvyatskiy, Hon-Leung Lee, Giorgio Ottaviani and Rekha R. Thomas. The Euclidean distance degree of orthogonally invariant matrix varieties. *Israel J. Math.*, 221(1):291–316, 2017.

[DHJ+18] Timothy Duff, Cvetelina Hill Anders Jensen, Kisun Lee, Anton Leykin and Jeff Sommars. Solving polynomial systems via homotopy continuation and monodromy. *IMA Journal of Numerical Analysis*, 39(3):1421–1446, 04 2018.

[Ear21] Nick Early. Planarity in generalized scattering amplitudes: Pk polytope, generalized root systems and worldsheet associahedra. arXiv:2106.07142 (2021).

[EH16] David Eisenbud and Joe Harris. *3264 and all that: A second course in algebraic geometry*. Cambridge University Press, 2016.

[EKKB13] Ioannis Z Emiris, Tatjana Kalinka, Christos Konaxis and Thang Luu Ba. Implicitization of curves and (hyper) surfaces using predicted support. *Theoretical Computer Science*, 479:81–98, 2013.

[EKP07] Ioannis Z Emiris, Christos Konaxis and Leonidas Palios. Computing the Newton polytope of specialized resultants. In *Presented at MEGA (Effective Methods in Algebraic Geometry)*, 2007.

[EK08] Alexander Esterov and Askold Khovanskii. Elimination theory and Newton polytopes. *Functional Analysis and Other Mathematics*, 2(1):45–71, 2008.

[Est07] Alexander Esterov. Determinantal singularities and newton polyhedra. *Proceedings of the Steklov Institute of Mathematics*, 259:16–34, 2007.

[Fev22] Claudia Fevola and Christiane Görgen The mathematical research-data repository MathRepo, Computeralgebra Rundbrief 70 (2022) 16–20.

[Fin13] Alex Fink. Tropical cycles and Chow polytopes. *Beiträge zur Algebra und Geometrie/Contributions to Algebra and Geometry*, 54:13–40, 2013.

[Ful98] William Fulton. *Intersection Theory*. Springer New York, NY, 2 edition, 1998.

[GD05] Hatice Gecegormez and Yasar Demirel. Phase stability analysis using interval Newton method with NRTL model. *Fluid Phase Equilibria*, 237(1-2):48—58, 2005.

[GKZ08] Israel M. Gelfand, Mikhail M. Kapranov and Andrej V. Zelevinsky. *Discriminants, resultants and multidimensional determinants*. Modern Birkhäuser Classics. Birkhäuser Boston, Inc., Boston, MA, 2008. Reprint of the 1994 edition.

[GS05] Balajit Gopalan and Jay-Dean Seader. Application of interval Newton's method to chemical engineering problems. *Reliable Computing*, 1(3):215—223, 2005.

[GvBR09] Hans-Christian Graf von Bothmer and Kristian Ranestad. A general formula for the algebraic degree in semidefinite programming. *Bull. Lond. Math. Soc.*, 41(2):193–197, 2009.

[HS12] Jonathan D. Hauenstein and Frank Sottile. Algorithm 921: alphaCertified: Certifying solutions to polynomial systems. *ACM Trans. Math. Softw.*, 38(4), Aug 2012.

[HL05] Didier Henrion and Jean-Bernard Lasserre. *Detecting global optimality and extracting solutions in GloptiPoly*, pages 293–310. Springer Berlin, Heidelberg, 2005.

[HLL09] Didier Henrion, Jean-Bernard Lasserre and Johan Löfberg Gloptipoly 3: moments, optimization and semidefinite programming. *Optimization Methods & Software*, 24(4-5):761–779, 2009.

[Hig02] Nicholas J. Higham. *Accuracy and stability of numerical algorithms*, volume 80. Siam, 2002.

[HKS05] Serkan Hoşten, Amit Khetan and Bernd Sturmfels. Solving the likelihood equations. *Found. Comput. Math.*, 5(4):389–407, 2005.

[How60] Ronald A. Howard. *Dynamic programming and Markov processes*. John Wiley, 1960.

[HS95a] Birkett Huber and Bernd Sturmfels. A polyhedral method for solving sparse polynomial systems. *Math. Comp.*, 64(212):1541–1555, 1995.

[Hug06] Peter Huggins. iB4e: A software framework for parametrizing specialized LP problems. In *Mathematical Software-ICMS 2006: Second International Congress on Mathematical Software, Castro Urdiales, Spain, September 1-3, 2006. Proceedings 2*, pages 245–247. Springer, 2006.

[Huh13] June Huh. The maximum likelihood degree of a very affine variety. *Compos. Math.*, 149(8):1245–1266, 2013.

[Joh17] Frederik Johansson. Arb: efficient arbitrary-precision midpoint-radius interval arithmetic. *IEEE Transactions on Computers*, 66:1281–1292, 2017.

[JKM04] Colin Jones, Eric Kerrigan and Jan Maciejowski. Equality set projection: A new algorithm for the projection of polytopes in halfspace representation. Technical report, University of Cambridge, Cambridge, 2004.

[Kal94] Lodewijk Kallenberg Survey of linear programming for standard and nonstandard Markovian control problems. Part I: Theory. *Zeitschrift für Operations Research*, 40(1):1–42, 1994.

[KKE21] B Ya Kazarnovskii, Askold Georgievich Khovanskii and Alexander Isaakovich Esterov. Newton polytopes and tropical geometry. *Russian Mathematical Surveys*, 76(1):91, 2021.

[Kha93] Leonid Khachiyan. *Complexity of polytope volume computation*, pages 91–101. Springer Berlin Heidelberg, Berlin, Heidelberg, 1993.

[Kho78] Askold G. Khovanskii. Newton polyhedra and the genus of complete intersections. *Funktsional. Anal. i Prilozhen.*, 12(1):51–61, 1978.

[Kou76] Anatoli G. Kouchnirenko. Polyèdres de Newton et nombres de Milnor. *Invent. Math.*, 32(1):1–31, 1976.

[KPR+21] Kathlén Kohn, Ragni Piene, Kristian Ranestad, Felix Rydell, Boris Shapiro, Rainer Sinn, Miruna-Stefana Sorea and Simon Telen. Adjoints and canonical forms of polypols. *arXiv preprint arXiv:2108.11747*, 2021.

[Kra69] Rudolf Krawczyk. Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken. *Computing*, 4(3):187–201, 1969.

[KT51] Harold W. Kuhn and Albert W. Tucker. Nonlinear programming. In *Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, 1951.

[KSS15] Virendra Kumar, Soumen Sen and Sankar Shome. Inverse kinematics of redundant manipulator using interval newton method. *International Journal of Engineering and Manufacturing*, 2:19—-20, 2015.

[Las01] Jean-Bernard Lasserre. Global optimization with polynomials and the problem of moments. *SIAM J. Optim.*, 11(3):796–817, 2000/01.

[LHPT08] Jean-Bernard Lasserre, Didier Henrion, Christophe Prieur and Emmanuel Trélat. Nonlinear optimal control via occupation measures and lmi-relaxations. *SIAM Journal on Control And Optimization*, 47(4):1643–1666, 2008.

[Lee17] Hwangrae Lee. The Euclidean distance degree of Fermat hypersurfaces. *J. Symbolic Comput.*, 80(part 2):502–510, 2017.

[Lee19] Kisun Lee. Certifying approximate solutions to polynomial systems on Macaulay2. *ACM Communications in Computer Algebra*, 53(2):45–48, 2019.

[Ley11] Anton Leykin. Numerical algebraic geometry for Macaulay2. *The Journal of Software for Algebra and Geometry: Macaulay2*, 3:5–10, 2011.

[LAR21] Julia Lindberg, Carlos Améndola and Jose Israel Rodriguez. Estimating gaussian mixtures using sparse polynomial moment systems. *arXiv preprint arXiv:2106.15675*, 2021.

[LMR23] Julia Lindberg, Leonid Monin and Kemal Rose. The algebraic degree of sparse polynomial optimization. *arXiv preprint arXiv:2308.07765*, 2023.

[RMR23] Julia Lindberg, Leonid Monin and Kemal Rose. A polyhedral homotopy algorithm for computing critical points of polynomial programs. *arXiv preprint arXiv:2302.04117*, 2023.

[LNRW23] Julia Lindberg, Nathan Nicholson, Jose I. Rodriguez and Zinan Wang. The maximum likelihood degree of sparse polynomial systems. *SIAM Journal on Applied Algebra and Geometry*, 7(1):159–171, 2023.

[LZBL20] Julia Lindberg, Alisha Zachariah, Nigel Boston and Bernard C. Lesieutre. The distribution of the number of real solutions to the power flow equations, 2020. *IEEE Transactions on Power Systems 38.2*, (2022): 1058-1068.

[DCO97] John Littles and David A. Cox and Donal O'Shea. Ideals, varieties and algorithms. *Undergraduate Texts in Mathematics*, 1997.

[MS15] Diane Maclagan and Bernd Sturmfels. *Introduction to tropical geometry*, volume 161 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2015.

[Mar21] Ivan A. Martyanov. Solving the Delsarte problem for 4-designs on the sphere. *Chebyshevskii Sb.*, 22:154–165, 2021.

[MRW20a] Laurentiu G. Maxim, Jose I. Rodriguez and Botong Wang. Euclidean distance degree of the multiview variety. *SIAM J. Appl. Algebra Geom.*, 4(1):28–48, 2020.

[MRW20b] Laurentiu G. Maxim, Jose Israel Rodriguez and Botong Wang. Defect of Euclidean distance degree. *Adv. in Appl. Math.*, 121:102101, 22, 2020.

[MRWW23] Laurentiu G. Maxim, Jose Israel Rodriguez, Botong Wang and Lei Wu. Linear optimization on varieties and Chern-Mather classes, 2023.

[May17] Günter Mayer. *Interval analysis*. De Gruyter, Berlin, Boston, 2017.

[MH19] Daniel K. Molzahn and Ian A. Hiskens. A survey of relaxations and approximations of the power flow equations. *Foundations and trends in electric energy systems*, 4(1-2):1–221, 2019.

[MMW21] Mateusz Michałek, Leonid Monin and Jarosław A. Wiśniewski. Maximum likelihood degree, complete quadrics and $\mathbb{C}^*$-action. *SIAM J. Appl. Algebra Geom.*, 5(1):60–85, 2021.

[MGZA15] Guido Montúfar, Keyan Ghazi-Zahedi and Nihat Ay. Geometry and determinism of optimal stationary control in partially observable Markov decision processes, 2015. Preprint, arXiv:1503.07206.

[MR17] Guido Montúfar and Johannes Rauh. Geometry of policy improvement. In *International Conference on Geometric Science of Information*, pages 282–290. Springer, 2017.

[MRA19] Guido Montúfar, Johannes Rauh and Ay, Nihat. Task-agnostic constraining in average reward POMDPs. In *ICLR 2019 Workshop on Task-Agnostic Reinforcement Learning*, 2019.

[Moo66] Ramon E. Moore. *Interval analysis*, volume 4. Prentice-Hall, 1966.

[Moo77] Ramon E. Moore. A test for existence of solutions to nonlinear systems. *SIAM Journal on Numerical Analysis*, 14(4):611–615, 1977.

[MS89] Alexander P. Morgan and Andrew J. Sommese. Coefficient-parameter polynomial continuation. *Applied Mathematics and Computation*, 29(2):123–160, 1989.

[MM22a] Johannes Müller and Guido Montúfar. The geometry of memoryless stochastic policy optimization in infinite-horizon POMDPs. In *International Conference on Learning Representations*, 2022.

[MM22b] Johannes Müller and Guido Montúfar. Solving infinite-horizon POMDPs with memoryless stochastic policies in state-action space. In *5th Multi-disciplinary Conference on Reinforcement Learning and Decision Making*, 2022.

[Ney03] Abraham Neyman. Real algebraic tools in stochastic games. In *Stochastic games and applications*, pages 57–75. Springer, 2003.

[Nie11] Jiawang Nie. Certifying convergence of Lasserre's hierarchy via flat truncation. *Mathematical Programming*, 142:485–510, 2011.

[Nie14] Jiawang Nie. Optimality conditions and finite convergence of lasserre's hierarchy. *Mathematical programming*, 146(1):97–121, 2014.

[NR09] Jiawang Nie and Kristian Ranestad Algebraic degree of polynomial optimization. *SIAM Journal on Optimization*, 20(1):485–502, 2009.

[NRS10] Jiawang Nie, Kristian Ranestad and Bernd Sturmfels. The algebraic degree of semidefinite programming. *Math. Program.*, 122(2, Ser. A):379–405, 2010.

[NT21] Jiawang Nie and Xindong Tang. Convex generalized Nash equilibrium problems and polynomial optimization. *Mathematical Programming*, pages 1–34, 2021.

[Pie78] Ragni Piene. Polar classes of singular varieties. *Annales scientifiques de l'École Normale Supérieure*, 11(2):247–276, 1978.

[PM14] Mark M. Plecnik and John M. McCarthy. Numerical synthesis of six-bar linkages for mechanical computation. *Journal of Mechanisms and Robotics*, 6(3), 06 2014. 031012.

[PRW95] Svata Poljak,Franz Rendl and Henry Wolkowicz. A recipe for semidefinite relaxation for $(0,1)$-quadratic programming. *J. Global Optim.*, 7(1):51–73, 1995.

[PS22] Irem Portakal and Bernd Sturmfels. Geometry of dependency equilibria, 2022. Preprint, arXiv:2201.05506.

[PLT11] Pascal Poupart, Tobias Lang and MarcToussaint. Analyzing and escaping local optima in planning as inference for partially observable domains. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 613–628. Springer, 2011.

[Put14] Martin L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[RTV97] Felice Ronga, Alberto Tognoli and Thierry Vust. The number of conics tangent to five given conics: the real case. *Rev. Mat. Univ. Complut. Madrid*, 10:391–421, 1997.

[RST23] Kemal Rose, Bernd Sturmfels and Simon Telen. Tropical implicitization revisited. *arXiv preprint arXiv:2306.13015*, 2023.

[Rum83] Siegfried M. Rump. Solving algebraic problems with high accuracy. In *Proc. of the Symposium on A New Approach to Scientific Computation*, page 51–120, USA, 1983. Academic Press Professional, Inc.

[Rum99] Siegfried M. Rump. INTLAB - INTerval LABoratory. In *Developments in reliable computing*, pages 77–104. Kluwer Academic Publishers, 1999.

[Rum10] Siegfried M. Rump. Verification methods: Rigorous results using floating-point arithmetic. *Acta Numerica*, 19:287–449, 2010.

[Sma86] Steve Smale. Newton's method estimates from data at one point. In Richard E. Ewing, Kenneth I. Gross and Clyde F. Martin, editors, *The Merging of Disciplines: New Directions in Pure, Applied and Computational Mathematics*, pages 185–196. Springer, 1986.

[SW05] Andrew Sommese and Charles Wampler. *The numerical solution of systems of polynomials arising in engineering and science*. World Scientific, 2005.

[SY21] Frank Sottile and Thomas Yahl. Galois groups in enumerative geometry and applications. *arXiv preprint arXiv:2108.07905*, 2021.

[ST08] Bernd Sturmfels and Jenia Tevelev. Elimination theory for tropical varieties. *Mathematical Research Letters*, 15(3):543–562, 2008.

[STY06] Bernd Sturmfels, Jenia Tevelev and Josephine Yu. The newton polytope of the implicit equation. *Computing Research Repository - CORR*, 7, 07 2006.

[ST21] Bernd Sturmfels and Simon Telen. Likelihood equations and scattering amplitudes. *Algebraic Statistics*, 12(2):167–186, 2021.

[SY08] Bernd Sturmfels and Josephine Yu. Tropical implicitization and mixed fiber polytopes. *Software for algebraic geometry*, pages 111–131, 2008.

[Stu02] Bernd Sturmfels. *Solving systems of polynomial equations*, volume 97 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI, 2002.

[Stu08] Bernd Sturmfels. *Algorithms in invariant theory.* Springer Science & Business Media, 2008.

[Stu21] Bernd Sturmfels. Beyond linear algebra. *arXiv preprint arXiv:2108.09494*, 2021.

[Sun58] Teruo Sunaga. Theory of an interval algebra and its application to numerical analysis. *Research Association of Applied Geometry*, 2:29–46, 1958.

[SMSM99] Richard S Sutton, David McAllester, Satinder Singh and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, volume 12. MIT Press, 1999.

[TR01] Peng Hui Tan and L.K. Rasmussen. The application of semidefinite programming for detection in CDMA. *IEEE Journal on Selected Areas in Communications*, 19(8):1442–1449, 2001.

[Ver99] Jan Verschelde. Algorithm 795: PHCpack: A general-purpose solver for polynomial systems by Homotopy Continuation. *ACM Trans. Math. Softw.*, 25(2):251–276, June 1999.

[VLB12] Nico Vlassis, Michael L. Littman and David Barber. On the computational complexity of stochastic controller optimization in POMDPs. *ACM Transactions on Computation Theory (TOCT)*, 4(4):1–8, 2012.

[WB06] Andres Wächter and Lorenz T. Biegler On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106(1):25–57, 2006.

[WKZ+22] Kaixin Wang, Navdeep Kumar, Kuangqi Zhou, Bryan Hooi, Jiashi Feng and Shie Mannor. The geometry of robust value functions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22727–22751. PMLR, 17–23 Jul 2022.

[Wei21] Madeleine A. Weinstein. *Metric Algebraic Geometry.* University of California, Berkeley, 2021.

[Whi88] Douglas J. White. Further real applications of Markov decision processes. *Interfaces*, 18(5):55–61, 1988.

[WDL22] Yue Wu and Jesús A. De Loera Geometric policy iteration for Markov decision processes. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 2070–2078, New York, NY, USA, 2022. Association for Computing Machinery.