SpringerWienNewYork

Texts and Monographs in Symbolic Computation

A Series of the Research Institute for Symbolic Computation, Johannes Kepler University, Linz, Austria

For further volumes:
http://www.springer.com/series/3073

Lorenzo Robbiano · John Abbott
Editors

Approximate Commutative Algebra

SpringerWienNewYork

Editors

Prof. Lorenzo Robbiano
Università di Genova
Dipartimento di Matematica
Via Dodecaneso, 35
16146 Genova
Italy
robbiano@dima.unige.it

Dr. John Abbott
Università di Genova
Dipartimento di Matematica
Via Dodecaneso, 35
16146 Genova
Italy
abbott@dima.unige.it

SpringerWienNewYork is a part of Springer Science + Business Media
springer.at

Printed on acid-free paper
SPIN 12540785

With 11 Figures

Library of Congress Control Number: 2009934122

ISSN 0943-853X
ISBN 978-3-211-99313-2 e-ISBN 978-3-211-99314-9
DOI 10.1007/978-3-211-99314-9
SpringerWienNewYork

Foreword

What exactly is *Approximate Commutative Algebra*? Where precisely can the approximateness arise? Don't think that it just means

$$xy = 0.9999\,yx$$

and be aware there are certainly some important places where approximation and vagueness are definitely not allowed: *e.g.* in the theorems!

The name ApCoA is an acronym for "Approximate Commutative Algebra". It has received some criticism for its self-contradictory nature: algebra is exact, so it cannot be approximate — but it is for this very same reason that we like it! Our explicit goal is precisely that of building a bridge between the approximate data of the real world and the exact structures of commutative algebra. We believe that the nine papers contained in this volume give an excellent insight into this emerging field of research, and will contribute to the building of this important bridge.

The original stimulus for this book was the first ApCoA workshop hosted in February 2006 by the Radon Institute of Computational and Applied Mathematics (RICAM) of the Austrian Academy of Science and the Research Institute for Symbolic Computation (RISC) of the Johannes Kepler University in Linz, Austria. As interest spread and many new ideas and results sprang up, it quickly became clear that a second ApCoA workshop was warranted. This second workshop was part of the RISC Summer 2008 event, and was again co-organized by RICAM. Most of the articles in this book grew out of the presentations given at this second workshop.

Preface

We have gathered together in this volume nine articles offering highly varied points of view as to what *Approximate Commutative Algebra* (ApCoA) comprises. These diverse perspectives furnish an accessible overview of the current state of research in this burgeoning area. We believe that bringing together these surveys creates a single reference point which will be of benefit both to existing practitioners who wish to expand their horizons, and also to new researchers aspiring to enter this exciting and rapidly developing field. The presentations are intended also to appeal to the interested onlooker who wants to stay informed about recent developments in the field.

The contributions to this book come from active university researchers with a keen interest in ApCoA. Some of them have extensive experience in the field, while others are relative newcomers bringing with them new tools and techniques. The survey articles by their very nature can only scratch the surface, but each one comes with its own bibliography for those who desire to delve more deeply into the numerous topics discussed.

To help the reader orient himself, the paragraphs below summarise the scope of each of the contributed articles. Read and enjoy!

Kreuzer, Poulisse, Robbiano
From Oil Fields to Hilbert Schemes

New techniques for dealing with problems of numerical stability in computations involving multivariate polynomials allow a new approach to real world problems. Using a modelling problem for oil field production optimization as a motivation, the paper presents several recent developments involving border bases of polynomial ideals. To get a deeper understanding for the algebra underlying this approximate world, recent advances concerning border basis and Gröbner basis schemes are discussed. For the reader it will be a long, tortuous, sometimes dangerous, yet hopefully fascinating journey from oil fields to Hilbert schemes.

Bates, Hauenstein, Peterson, Sommese

Numerical Decomposition of the Rank-Deficiency Set of a Matrix of Multivariate Polynomials

Let A be a matrix whose entries are algebraic functions defined on a reduced quasi-projective algebraic set X, *e.g.* multivariate polynomials defined on $X := \mathbb{C}^N$. The sets $S_k(A)$, consisting of $x \in X$ where the rank of the matrix function $A(x)$ is at most k, arise in a variety of contexts: for example, in the description of both the singular locus of an algebraic set and its fine structure; in the description of the degeneracy locus of maps between algebraic sets; and in the computation of the irreducible decomposition of the support of coherent algebraic sheaves, *e.g.* supports of finite modules over polynomial rings. The article presents a numerical algorithm to compute the sets $S_k(A)$ efficiently.

Wu, Reid, Golubitsky

Towards Geometric Completion of Differential Systems by Points

Numerical Algebraic Geometry represents the irreducible components of algebraic varieties over $\mathbb{C}$ by certain points on their components. Such *witness points* are efficiently approximated by Numerical Homotopy Continuation methods, as the intersection of random linear varieties with the components. The paper outlines challenges and progress for extending such ideas to systems of differential polynomials, where prolongation (differentiation) of the equations is required to yield existence criteria for their formal (power series) solutions.

Scott, Reid, Wu, Zhi

Geometric Involutive Bases and Applications to Approximate Commutative Algebra

This article serves to give an introduction to some classical results on Involutive Bases for polynomial systems. Further, it surveys recent developments, including a modification of the above: geometric projected involutive bases, for the treatment of approximate systems, and their application to ideal membership testing and Gröbner basis computation.

Zeng

Regularization and Matrix Computation in Numerical Polynomial Algebra

Numerical polynomial algebra emerges as a growing field of study in recent years with a broad spectrum of applications and many robust algorithms. Among the challenges faced when solving polynomial algebra problems with floating-point arithmetic, the most frequently encountered difficulties include the removal of ill-posedness and the handling of large matrices. This survey develops regularization principles that reformulate the algebraic problems for their well-posed approximate solutions, derives matrix computations arising in numerical polynomial algebra, as well as a subspace strategy that substantially improves the computational efficiency by reducing the matrix sizes. These strategies have been successfully applied to numerical polynomial algebra problems such as GCD, factorization, elimination and determination of multiplicity structure.

Shekhtman

Ideal Interpolation: Translation to and from Algebraic Geometry

This paper discusses four themes that surfaced in multivariate interpolation and which seem to have analogues in algebraic geometry. The hope is that mixing these two areas together will benefit both. In Approximation Theory (AT) the limits of Lagrange projectors correspond to components of the Hilbert scheme of points in Algebraic Geometry (AG). Likewise, error formulas in (AT) may correspond to ideal representations in (AG), and so on.

Riccomagno, Wynn

An Introduction to Regression and Errors in Variables from an Algebraic Viewpoint

There is a need to make a closer connection between classical response surface methods and their experimental design aspects, including optimal design, and algebraic statistics, based on computational algebraic geometry of ideals of points. This is a programme which was initiated by Pistone and Wynn (Biometrika, 1996) and is expanding rapidly. Particular attention is paid to the problem of errors in variables which can be taken as a statistical version of the ApCoA research programme.

Stetter

ApCoA = Embedding Commutative Algebra into Analysis: (my view of computational algebra over $\mathbb{C}$)

This paper deals with the philosophical problem of understanding what ApCoA should mean and, most importantly, what it should do. The main position is that ApCoA comprises consideration of problems of Commutative Algebra over the complex or real numbers, admission of some data of limited accuracy, and use of floating-point arithmetic for the computation of numerical results. In the presence of empirical data, *i.e.* with nearly all computational problems arising from real world applications, the analytic viewpoint is indispensable. The spread of the data may include singular or degenerate situations which would be overlooked if the neighbourhood of a specified problem were neglected.

Kaltofen

Exact Certification in Global Polynomial Optimization Via Rationalizing Sums-Of-Squares

Errors in the coefficients due to floating point round-off or through physical measurement can render exact symbolic algorithms unusable. Hybrid symbolic-numeric algorithms compute minimal deformations of those coefficients that yield non-trivial results, *e.g.* polynomial factorizations or sparse interpolants. The question is: *are the computed approximations the globally nearest to the input?* This paper presents a new alternative to numerical optimization, namely the exact validation via symbolic methods of the global minimality of our deformations.

Acknowledgements Most of the papers in this book grew out of the two ApCoA workshops held at Linz and Hagenberg in Austria. The first workshop was conducted during the *Special Semester on Gröbner Bases* (which lasted from 1st February to 31st July 2006) organized by RICAM (Radon Institute for Computational and Applied Mathematics) of the Austrian Academy of Sciences and RISC (Research Institute for Symbolic Computation) of the Johannes Kepler University, Linz, Austria, under the direction of Professor Bruno Buchberger. The special semester consisted mainly of a series of workshops on various aspects of the theory of Gröbner bases and on important applications of the method.

The second workshop formed part of RISC Summer 2008 event held in the unique and special atmosphere of the Castle of Hagenberg (RISC, Austria). Again RICAM helped organize the meeting. The summer event brought together several important conferences relating to all aspects of symbolic computation — a programme into which ApCoA fitted perfectly.

Contents

Chapter 1
From Oil Fields to Hilbert Schemes

Martin Kreuzer, Hennie Poulisse, and Lorenzo Robbiano

Abstract New techniques for dealing with problems of numerical stability in computations involving multivariate polynomials allow a new approach to real world problems. Using a modelling problem for the optimization of oil production as a motivation, we present several recent developments involving border bases of polynomial ideals. After recalling the foundations of border basis theory in the exact case, we present a number of approximate techniques such as the eigenvalue method for polynomial system solving, the AVI algorithm for computing approximate border bases, and the SOI algorithm for computing stable order ideals. To get a deeper understanding for the algebra underlying this *approximate world*, we present recent advances concerning border basis and Gröbner basis schemes. They are open subschemes of Hilbert schemes and parametrize flat families of border bases and Gröbner bases. For the reader it will be a long, tortuous, sometimes dangerous, and hopefully fascinating journey from oil fields to Hilbert schemes.

Key words: oil field, polynomial system solving, eigenvalue method, Buchberger-Möller algorithm, border basis, approximate algorithm, border basis scheme, Gröbner basis scheme, Hilbert scheme

Martin Kreuzer
Fakultät für Informatik und Mathematik, Universität Passau, D-94030 Passau, Germany,
e-mail: `kreuzer@uni-passau.de`

Hennie Poulisse
Shell Int. Exploration and Production, Exploratory Research, Kessler Park 1, NL-2288 GD Rijswijk, The Netherlands, e-mail: `hennie.poulisse@shell.com`

Lorenzo Robbiano
Dipartimento di Matematica, Università di Genova, Via Dodecaneso 35, I-16146 Genova, Italy,
e-mail: `robbiano@dima.unige.it`

L. Robbiano, J. Abbott (eds.), *Approximate Commutative Algebra*, Texts and Monographs in Symbolic Computation, DOI 10.1007/978-3-211-99314-9_1, © Springer-Verlag/Wien 2009

Contents

Introduction

> *Why did the chicken cross the road?*
> *To boldly go where no chicken has gone before.*
> (James Tiberius Kirk)

A Bridge Between Two Worlds. Oil fields and Hilbert schemes are connected to very different types of ingredients for algorithmic and algebraic manipulation: continuous and discrete data. This apparent dichotomy occurs already in a single polynomial over the real number field. It consists of a discrete part, the support, and a continuous part, the set of its coefficients. The support is well understood and the source of a large amount of literature in classical algebra. On the other hand, if the coefficients are not exact real numbers but *approximate data*, the very notion of a polynomial and all algebraic structures classically derived from it (such as ideals, free resolutions, Hilbert functions, etc.) tend to acquire a blurred meaning.

An easy example is the following. Consider three distinct non-aligned points in the affine plane over the reals. First of all, if the coordinates are not exact, it is not even clear what we mean by “non-aligned”; a better description might be “far from aligned”. The vanishing ideal of the three points is generated by three quadratic polynomials. However, if we change some of the coefficients of these polynomials by a small amount, almost surely we get the unit ideal, since the first two conics still intersect in four points, but the third will almost certainly miss all of them.

How can we cope with this situation? And why should we? The first, easy answer is that approximate coefficients are virtually inevitable when we deal with *real world problems*. In this paper we concentrate on a specific problem where vectors with approximate components encode measurements of physical quantities taken in an **oil field**. Based on actual industrial problems in the field of oil production, we want to popularize the idea that good models of many physical phenomena can be constructed using a *bottom-up* process. The heart of this method is to derive mathematical models by interpolating measured values on a finite set of points. This task can be solved if we know the vanishing ideal of the point set and a suitable vector space basis of its coordinate ring.

This leads us to the next question. Given a zero-dimensional ideal I in a polynomial ring over the reals, if we assume that the coefficients of the generating polynomials are inexact, is it still an ideal? What is the best way of describing this situation? The fact that Gröbner bases are not suitable for computations with inexact data has long been well-known to numerical analysts (see [30]). This is due to the rigid structure imposed by term orderings. Other objects, called *border bases*, behave better. They have emerged as good candidates to complement, and in many cases substitute for, Gröbner bases (see [17], [21], [22], [26], [29]). But possibly the most important breakthrough is the recent discovery of a link between border bases and **Hilbert schemes**. We believe that it may provide a solid mathematical foundation for this new emerging field which tries to combine approximate methods from numerical analysis with exact methods from commutative algebra and algebraic geometry.

You got to be careful if you don't know where you're going
because you might not get there.
(Yogi Berra)

Our Itinerary. In the first part of the introduction we have already suggested the existence of an unexpected bridge between oil fields and Hilbert schemes. Let us now be more specific about the content of the paper and indicate how it tries to build that bridge. Section 1 provides an introduction to one of the main problems arising in oil fields, namely the control of the production. Since we assume that our typical reader is not an expert geologist, we provide some background about the physical nature of an oil reservoir, illustrate the main production problem, and describe a new mathematical approach to solve it. We call it "new", since in our opinion it is very different from the standard view on how to use mathematical models in such a context.

Border bases, the main technical tool we use later, are described in Section 2. This material is mainly taken from [21], Section 6.4 and [17]. We describe the definition and the main properties of border bases and compare them to Gröbner bases using suitable examples. Several important results about border bases are described, in particular their characterization via the commutativity of the formal multiplication matrices due to B. Mourrain (see [26]). A brief excursion is taken into the realm of syzygies, their relation to the border web, and their importance in another fundamental characterization of border bases based on the work of H. Stetter (see [30]).

A useful aspect of border basis theory is that we try to specify a "nice" vector space basis of the quotient ring $\mathbb{R}[x_1,\dots,x_n]/I$. This sort of basis plays a fundamental role in the problem of solving polynomial systems. Notwithstanding the fact that solving polynomial systems is not a main topic in our presentation, we decided to use Section 3 to give a description of a technique which comes from numerical analysis and uses linear algebra methods, in particular eigenvalues and eigenvectors (see [4], [5], and [9]). The importance of a special kind of matrices, called non-derogatory matrices, is illustrated by Example 1.3.9 and also used in [19] in the context of border basis theory.

Sections 4 and 5 are the computational heart of the paper. They describe two somehow complementary algorithmic approaches to the problem of computing the "approximate vanishing ideal" of a finite set of approximate (empirical) points and a basis of the corresponding quotient ring. In particular, the first part of Section 4 deals with the AVI algorithm and is based on the presentation in [14]. The AVI algorithm makes extensive use of the singular value decomposition (SVD) described in Subsection 4.A and of the stable reduced row echelon form explained in Subsection 4.B. Its main outputs are an order ideal of monomials $\mathcal{O}$ and an approximate $\mathcal{O}$-border basis, a concept introduced in Subsection 4.C. The AVI algorithm is then applied in Subsection 4.D to the concrete construction of polynomial models describing the production of a two-zone oil well.

Section 5 deals with the SOI algorithm which treats the following problem: given a finite set of points $\mathbb{X}$ whose coordinates are given with limited precision, find, if there exists one, an order ideal $\mathcal{O}$ such that the residue classes of its elements form a stable basis of the quotient ring $P/\mathcal{I}(\mathbb{X})$ where $P = \mathbb{R}[x_1,\dots,x_n]$ and $\mathcal{I}(\mathbb{X})$ is the vanishing ideal of $\mathbb{X}$. Here stable means that the residue classes of the elements in $\mathcal{O}$ form a basis of the quotient ring for every small perturbation of the set $\mathbb{X}$. This section summarizes the results of [2]. In Subsection 5.B we describe several easy, but illustrative examples and compare the behaviour of the SOI and the AVI algorithm in these cases. The topic studied in Sections 4 and 5 is an active area of research, and several further approaches have been suggested (see for instance [10] and [25]).

Having done all the *dirty* work (oil fields are not places to be dressed formally), it is time to leave the sedimentary rocks and to look at the problems concerning approximate data from a more general perspective. Polynomials with empirical coefficients can be viewed as *families of polynomials*. So, the next question is whether we can describe families of polynomial ideals algebraically. The answer is yes! The possibility of parametrizing families of schemes by one big scheme is a remarkable feature of algebraic geometry. Hilbert schemes are the most widely known instance of this phenomenon, and consequently they have been studied thoroughly. Moreover, the Hilbert scheme of all zero-dimensional ideals in P of colength s can be covered by affine open subschemes which parametrize all subschemes $\mathrm{Spec}(P/I)$ of the affine space $\mathbb{A}^n_K$ with the property that P/I has a fixed vector space basis. It is interesting to note that the construction of such subschemes is performed using border bases (see for instance [15], [16], and [24]). Also Gröbner bases can be used, since they provide tools for constructing suitable stratifications of Hilbert schemes.

Section 6 is devoted to the explanation of these ideas. Its main sources are the two papers [22] and [28]. In Subsection 6.A we start with an informal explanation of two examples (see Examples 1.6.1 and 1.6.2) which are very easy but nevertheless suitable to illustrate the topic. Then we move to Subsection 6.B where we introduce border basis schemes and their associated border basis families. We show the difficulties of generalizing one of the fundamental tools of Gröbner basis theory to the border basis setting, namely the flat deformation to the leading term ideal. Indeed, the problem is only partially solved and still open in general. The final part of the subsection contains Example 1.6.14 where explicit defining equations are given for one particular border basis scheme, and the connection to the approximate border bases of Section 4 is made.

The final Subsection 6.C is devoted to Gröbner basis schemes and summarizes the presentation in [28]. It is shown that Gröbner basis schemes and their associated universal families can be viewed as weighted projective schemes (see Theorem 1.6.19), a fact that constitutes a remarkable difference between Gröbner and border basis schemes. A comparison between the two types of schemes is given by Theorem 1.6.20 and Corollary 1.6.21, and their equality is examined in Proposition 1.6.24. Throughout the section we highlight the connection between border basis schemes, Gröbner basis schemes, and Hilbert schemes.

At that point the journey from oil fields to Hilbert schemes is over. To get you started with this itinerary, let us point out that, unless specifically stated otherwise, our notation follows the two books [20] and [21]. The algorithms we discuss have been implemented in the computer algebra system CoCoA(see [8]) and in the ApCoCoA library (see [3]).

1.1 A Problem Arising in Industrial Mathematics

Are oil fields commutative?
Are they infinite?
What is their characteristic?
Are they stable?
What are their bases?
(from "The Book of Mathematical Geology")

1.1.A. Oil Fields, Gas Fields and Drilling Wells. Research in relation to oil reservoirs faces many times the same kind of difficulty: the true physical state of an intact, working reservoir cannot be observed. Neither in an experiment of thought, for instance a simulation, nor in a physical experiment using a piece of source rock in a laboratory, the reservoir circumstances can be imitated exactly. This means that the physical laws, *i.e.* the relations between the physical quantities, are not known under actual reservoir circumstances.

To shed some additional light upon this problem, let us have a brief look at oil field formation and exploitation. The uppermost crust of the earth in oil and gas-containing areas is composed of sedimentary rock layers. Since the densities of oil and gas are smaller than the density of water, buoyancy forces them to flow upward through small pores in the reservoir rock. When they encounter a *trap*, *e.g.* a dome or an anticline, they are stopped and concentrated according to their density: the gas is on top and forms the free *gas cap*, the oil goes in the middle, and the (salt) water is at the bottom. To complete the trap, a *caprock*, that is a seal which does not allow fluids to flow through it, must overlie the reservoir rock.

Early drillings had some success because many subsurface traps were leaking. Only by the early 1900s it became known that traps could be located by mapping the rock layers and drilling an exploration well to find a new reservoir. If commercial amounts of oil and gas turn out to be present, a long piece of steel pipe (called the *production tubing*) is lowered into the bore hole and connected to the production facilities.

In a gas well, gas flows to the surface by itself. There exist some oil wells, early in the development of an oil field, in which the oil has enough pressure to flow up the surface. Most oil wells, however, do not have enough pressure and a method called *artificial lift* may then be used. This means that gas is injected into the production tubing of the well. The injected gas mixes with the oil and makes it lighter, thereby reducing the back pressure of the reservoir. On the surface the fluids are transported through long pieces of tubing to a large vessel called *separator* where the three physical phases – oil, water and gas – are separated.

During the exploitation of a reservoir, the pressure of the fluid still in the reservoir drops. This decrease of the reservoir pressure over time is depicted by the *decline curve*. The shape of the decline curve and the total volume of fluid that can be produced from a reservoir (which is called the *ultimate recovery*) depend on the *reservoir drive*, the natural energy that pushes the oil or the gas through the subsurface and into the inflow region of the well. The ultimate recovery of gas from a gas reservoir is often about 80% of the gas in the reservoir. Oil reservoirs are far more variable and less efficient: on average, the ultimate recovery is only 30%. This leaves 70% of the oil remaining in the pressure depleted reservoir which cannot be produced anymore.

Thus, on the most abstract level, the problem we want to address is how to increase the ultimate recovery of an oil reservoir.

1.1.B. Production from Multi-Zone Wells. A well may produce from different parts, called *pockets* or *zones*, of an oil reservoir. The total production of such a well consists of contributions from the different zones. The separate contributions can be controlled by valves, called the *down-hole valves*, which determine the production volume flowing into the well tubing at the locations of the different zones. For such a *multi-zone well*, there may be interactions between the zones in the reservoir. Most certainly, the different contributions will interact with each other when they meet in the common production tubing of the multi-zone well. This situation is called *commingled* production.

In this paper we consider a multi-zone well consisting of **two** producing and interacting zones. Like in a single oil well, the common production flows to the bulk separator where the different phases are separated and the production rates of the separated phases are measured. Besides the phase productions, measurements like pressures, temperatures and injected "lift-gas" are collected; down-hole valves positions are also recorded. A typical set of production variables for a such multi-zone well is:

1. the opening of the valve through which the oil from the first zone is entering the multi-zone well; the opening of the valve is measured in percentages: 0% means that the valve is closed; 100% means that the valve is completely open;
2. the opening of the valve through which the oil from the second zone is entering the multi-zone well;

3. the pressure difference over the down-hole valve of the second zone which is a measure for the inflow from the reservoir into the well at the valve position; if the valve is closed we assume this value to be zero;
4. the pressure difference over the down-hole valve of the first zone when the valve in that zone is open; if the valve is closed we assume this value to be zero;
5. the volume of gas produced simultaneously with the oil;
6. the pressure difference between the inflow locations in the production tubing;
7. the pressure difference which drives the oil through the transportation tubing.

One might be tempted to think that the total oil production of a multi-zone well is the sum of the productions of each zone when producing separately. This is in any case the current state of the art, where the total production is regressed against the separate productions, that is the total production is written as a linear combination of the separate productions. The coefficients in this linear sum are called *reconciliation factors*. The oil produced by one of the zones may push back the oil which tries to flow into the well at the other zone. Likewise, the gas which is produced simultaneously with the oil may have stimulating or inhibiting effects on the inflow of the oil with respect to the situation of single zone productions. With reference to the remarks above, this behavior does not sound very linear. Indeed, in Section 4.D we will use our algebraic approach in a two-zone well example to demonstrate that the total production is not a linear combination of the separate productions. We believe that the reason of the (usually) low ultimate recovery of a multi-zone well is due to the fact that the interactions among the different producing zones are unknown.

This leads us to a first concretization of the problem we want to study: find a model for the total production of an oil well which takes the interactions into account and describes the behavior correctly on longer time scales.

1.1.C. Algebraization of the Production Problem. Before plunging into the creation of an algebraic setting for the described production problem, let us spend a few words on why we believe that approximate computational algebra is an appropriate method to deal with it.

The available data correspond to a finite set of points $\mathbb{X}$ in $\mathbb{R}^n$. Their coordinates are *noisy* measurements of physical quantities associated with the well: pressures, oil and gas production, valve positions, etc. These points represent the behavior of the well under various production conditions. The combination of the contribution of the individual zones to the total production is a *sum* which has to be corrected by taking into account the effect of the interactions. As in many other situations (for instance, in statistics), the interactions are related to *products* of the collected data series. Many of the known physical laws and model equations are of a polynomial nature. And even if they are not, some elementary insights into the system (*e.g.* that the result depends exponentially on a certain data series) allow us to prepare the data series appropriately (*e.g.* by computing their logarithms). Consequently, the starting point for us is the polynomial ring $P = \mathbb{R}[x_1,\dots,x_n]$.

In the following we will deal with the case of a two-zone well. The production situation is depicted schematically in Figure 1.1. The notation ΔP refers to pressure differences.

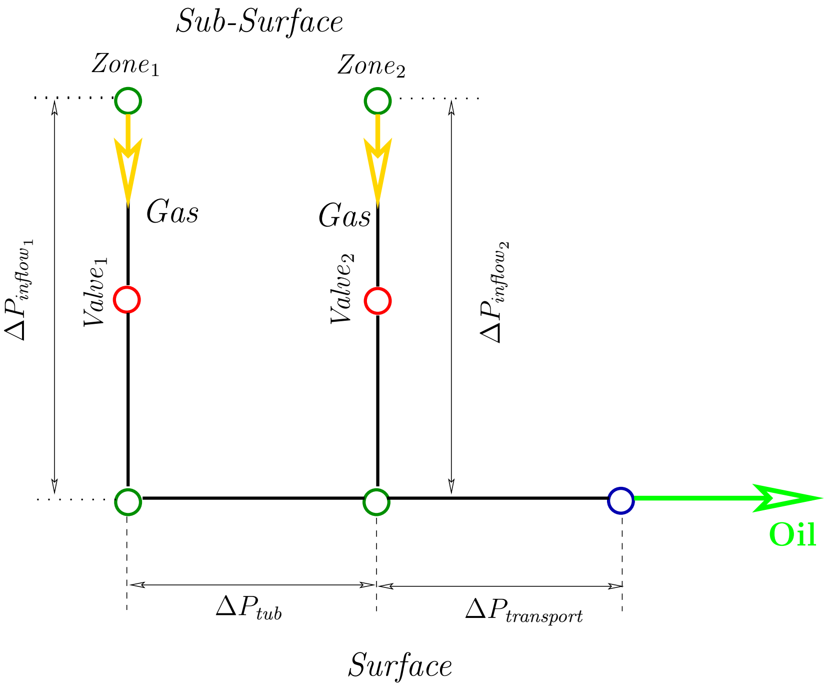


Fig. 1.1 Schematic representation of a two-zone well.

The valves indicated in this figure are used to influence the inflow of the fluids at the two locations into the production tubing of the well. If a valve is closed, there is no inflow from the reservoir at the location of the valve. If the valve is open, the inflow depends on the valve opening and the interactions with the fluids which enter the well through the other inflow opening. In particular, a valve in open position does not imply that there is inflow from the reservoir into the well at its location.

Next we try to formulate the problems associated with this production system more explicitly. Notice that the reservoir is a very special physical system in that it is not possible to check “how it works” using a computer simulation experiment or a physical model laboratory experiment. Traditional modelling techniques assume that equations which describe the flow of the fluids through the reservoir are available. Their limited success is in our view due to the fact that there is no proper representation of the interactions occurring in the production situation. Without these, actions taken to influence the production may have devastating consequences in that the “wrong” effects are stimulated. It is fair to state that the existing low ultimate recovery rates are to a large extent caused by the fact that the interactions in production units have not been acknowledged properly.

As a starting point, let us formulate the production problem in intuitive rather than in precise mathematical terms.

Problem 1. Assume that no *a priori* model is available to describe the production of the two-zone well of Figure 1.1 in terms of measurable physical quantities which determine the production. Find an algebraic model of the production in terms of the determining, measurable physical quantities which specifically models the interactions occurring in this production unit.

Now let us phrase this problem using the polynomial ring $P = \mathbb{R}[x_1,\dots,x_n]$. The first step is to associate the indeterminates x_i with physical quantities in the production problem in the sense that when the indeterminate x_i is evaluated at the points of $\mathbb{X}$, the evaluations are the measurements of the physical quantity associated to x_i. In the sequel we use $n = 5$ and the following associations, where the physical quantities are the ones referenced in Figure 1.1.

x_1	:	$\Delta P_{\mathit{inflow}_1}$
x_2	:	$\Delta P_{\mathit{inflow}_2}$
x_3	:	*Gas production*
x_4	:	ΔP_{tub}
x_5	:	$\Delta P_{\mathit{transport}}$

Table 1.1 Physical interpretation of the indeterminates.

Note that we have not listed an indeterminate associated to the oil production. The explanation for this is that the physical quantities listed in the above table may all be interpreted as *driving forces* for the oil production. For the pressure differences ΔP this is clear. But it holds also for the gas production. When a large amount of gas is produced in the deeper parts of the reservoir, it disperses in the fluid mixture, makes it lighter, and in this way stimulates oil production through this lifting process. Thus the physical quantities listed in the above table may all be viewed as the *causing* quantities, or *inputs*, and the oil production is their *effect*, or *output*. So, basically we make the following crucial assumption.

Assumption. *There exists a causal relationship between the production and the driving forces. Using suitable inputs, this causal relationship is of polynomial nature.*

Denoting the production by f, the algebraic translation of the causal relationship assumption is $f \in \mathbb{R}[x_1,\dots,x_5]$ where the indeterminates x_i are labeled as in the above table. That is, the production is not associated with an indeterminate, but with a polynomial, and the production measurements are the evaluations of this polynomial over the set $\mathbb{X}$. Hence the statement of Problem 1 can be reformulated as follows.

Problem 2. Find the polynomial $f \in \mathbb{R}[x_1,\dots,x_5]$, using only the evaluations $\mathbb{X}$ of the quantities x_i and the evaluations of f!

The information registered in the set $\mathbb{X}$ refers to the situation where at most one of the valves is closed. The only possible inflows from the reservoir into the production tubing of the two-zone well are at the location of Zone 1, or of Zone 2, or both. Moreover, in all three situations data have been collected at different valve openings. Furthermore, in order for the data in $\mathbb{X}$ to deserve the qualification *driving forces*, some pre-processing has been applied: with reference to Figure 1.1, if $valve_1$ is closed, it may very well be that the pressure difference ΔP_{inflow_1} is not zero, but it does not have the meaning of a driving force over the valve opening because there is no flow over the valve. Hence in the data set $\mathbb{X}$, we set ΔP_{inflow_1} to zero for this situation. Of course, we do the same for $valve_2$ with respect to ΔP_{inflow_2}. Finally, if the valve associated with the deepest zone $valve_1$ is closed, there is no transport of fluids in the lowest part of the production tubing of the well. That is, for ΔP_{tub} really to have the significance of a driving force, it is set to zero if $valve_1$ is closed.

Notice also that all data are based on measurements, *i.e.* they may contain measurement errors. Consequently, we can only expect that the desired polynomial f vanishes *approximately* at the points of $\mathbb{X}$. In Section 4 we will return to this instance of the production problem and solve it with the methods we are going to present.

1.2 Border Bases

Ideally, inside the border
there is order.
(Three anonymous authors)

1.2.A. Motivation and Definition. The problems considered in the previous section lead us to study zero-dimensional ideals in $P = K[x_1,\dots,x_n]$ where K is a field. The two most common ways to describe such an ideal I are by either providing a special system of generators (for instance, a Gröbner basis) of I or by finding a vector space basis $\mathcal{O}$ of P/I and the matrices of the multiplications by the indeterminates with respect to $\mathcal{O}$. One possibility to follow the second approach is to use $\mathcal{O} = \mathbb{T}^n \setminus \mathrm{LT}_\sigma(I)$, the complement of a leading term ideal of I. By Macaulay's Basis Theorem, such a set $\mathcal{O}$ is a K-basis of P/I. Are there other suitable sets $\mathcal{O}$?

A natural choice is to look for sets of terms. We need to fix how a term b_j in the *border* $\partial\mathcal{O} = (x_1\mathcal{O} \cup \cdots \cup x_n\mathcal{O}) \setminus \mathcal{O}$ of $\mathcal{O}$ is rewritten as a linear combination of the terms in $\mathcal{O}$. Thus, for every $b_j \in \partial\mathcal{O}$, a polynomial of the form

$$g_j = b_j - \sum_{i=1}^{\mu} c_{ij} t_i$$

with $c_{ij} \in K$ and $t_i \in \mathcal{O}$ should be contained in I. Moreover, we would not like that $x_k g_j \in I$. Hence we want $x_k b_j \notin \mathcal{O}$. Therefore the set $\mathbb{T}^n \setminus \mathcal{O}$ should be a monoideal. Consequently, $\mathcal{O}$ should be an *order ideal*, that is it should be closed under forming divisors. Let us formulate precise definitions.

Definition 1.2.1. Let $\mathcal{O}$ be a finite set of terms in $\mathbb{T}^n$.

a) The set $\mathcal{O}$ is called an **order ideal** if $t \in \mathcal{O}$ and $t' \mid t$ implies $t' \in \mathcal{O}$.
b) Let $\mathcal{O}$ be an order ideal. The set $\partial\mathcal{O} = (x_1\mathcal{O} \cup \cdots \cup x_n\mathcal{O}) \setminus \mathcal{O}$ is called the **border** of $\mathcal{O}$.
c) Let $\mathcal{O} = \{t_1, \dots, t_\mu\}$ be an order ideal and $\partial\mathcal{O} = \{b_1, \dots, b_\nu\}$ its border. A set of polynomials $\{g_1, \dots, g_\nu\} \subset I$ of the form

$$g_j = b_j - \sum_{i=1}^{\mu} c_{ij} t_i$$

with $c_{ij} \in K$ and $t_i \in \mathcal{O}$ is called an $\mathcal{O}$**-border prebasis** of I.
d) An $\mathcal{O}$-border prebasis of I is called an $\mathcal{O}$**-border basis** of I if the residue classes of the terms in $\mathcal{O}$ are a K-vector space basis of P/I.

The following example will be used frequently throughout this paper.

Example 1.2.2. In the ring $P = \mathbb{R}[x,y]$, consider the ideal $I = (f_1, f_2)$ where

$$\begin{aligned} f_1 &= \tfrac{1}{4}x^2 + y^2 - 1 \\ f_2 &= x^2 + \tfrac{1}{4}y^2 - 1 \end{aligned}$$

The zero set of I in $\mathbb{A}^2(\mathbb{R})$ consists of the four points $\mathbb{X} = \{(\pm\sqrt{0.8}, \pm\sqrt{0.8})\}$. This setting is illustrated in Figure 1.2.

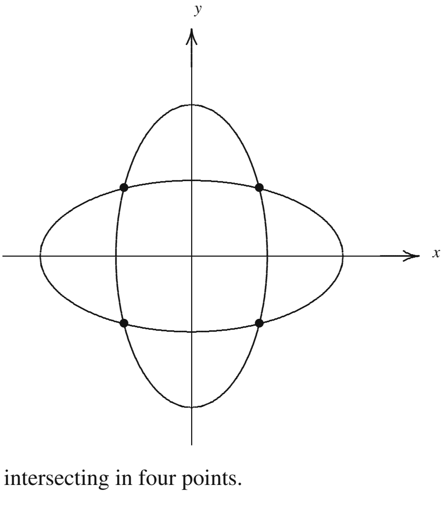


Fig. 1.2 Two ellipses intersecting in four points.

We use $\sigma = \texttt{DegRevLex}$ and compute $\mathrm{LT}_\sigma(I) = (x^2, y^2)$. Thus the order ideal $\mathcal{O} = \{1, x, y, xy\}$ represents a basis of P/I. Its border is $\partial\mathcal{O} = \{x^2, x^2y, xy^2, y^2\}$. The following figure illustrates the order ideal $\mathcal{O}$ and its border.

An $\mathcal{O}$-border basis of I is given by $G = \{g_1, g_2, g_3, g_4\}$ where

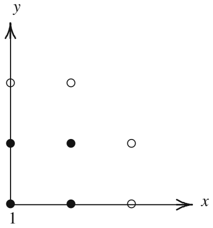


Fig. 1.3 An order ideal and its border.

$$\begin{aligned} g_1 &= x^2 - 0.8 \\ g_2 &= x^2y - 0.8y \\ g_3 &= xy^2 - 0.8x \\ g_4 &= y^2 - 0.8 \end{aligned}$$

Let us see what happens if we disturb this example slightly.

Example 1.2.3. Again we use $P = \mathbb{R}[x,y]$, but now we consider $\tilde{I} = (\tilde{f}_1, \tilde{f}_2)$ where

$$\begin{aligned} \tilde{f}_1 &= 0.25x^2 + y^2 + 0.01xy - 1 \\ \tilde{f}_2 &= x^2 + 0.25y^2 + 0.01xy - 1 \end{aligned}$$

Its zero set consists of four perturbed points $\widetilde{\mathbb{X}}$ close to those in $\mathbb{X}$, as illustrated in Figure 1.4.

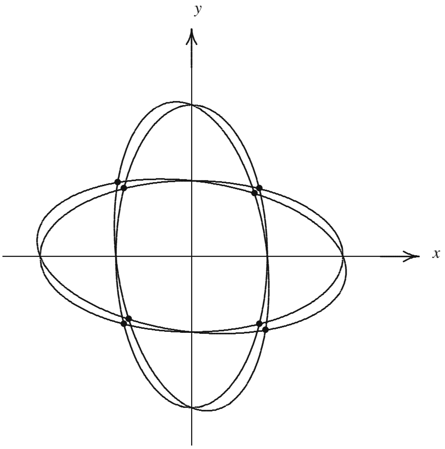


Fig. 1.4 Two slightly moved ellipses and their points of intersection.

The ideal $\tilde{I} = (\tilde{f}_1, \tilde{f}_2)$ has the reduced σ-Gröbner basis

$$\{x^2 - y^2, xy + 125y^2 - 100, y^3 - \tfrac{25}{3906}x + \tfrac{3125}{3906}y\}$$

Moreover, we have $\mathrm{LT}_\sigma(\tilde{I}) = (x^2, xy, y^3)$ and $\mathbb{T}^2 \setminus \mathrm{LT}_\sigma\{\tilde{I}\} = \{1, x, y, y^2\}$.

A *small* change in the coefficients of f_1 and f_2 has led to a *big* change in the Gröbner basis of $(\tilde{f}_1, \tilde{f}_2)$ and in the associated vector space basis of $\mathbb{R}[x,y]/(\tilde{f}_1, \tilde{f}_2)$, although the zeros of the ideal have not changed much. Numerical analysts call this kind of unstable behavior a *representation singularity*.

However, also the ideal $\tilde{I}$ has a a border basis with respect to $\mathcal{O} = \{1,x,y,xy\}$. Recall that the border of $\mathcal{O}$ is $\partial\mathcal{O} = \{x^2, x^2y, xy^2, y^2\}$.

The $\mathcal{O}$-border basis of $\tilde{I}$ is $\tilde{G} = \{\tilde{g}_1, \tilde{g}_2, \tilde{g}_3, \tilde{g}_4\}$ where

$$\begin{aligned}
\tilde{g}_1 &= x^2 + 0.008xy - 0.8\\
\tilde{g}_2 &= x^2y + \tfrac{25}{3906}x - \tfrac{3125}{3906}y\\
\tilde{g}_3 &= xy^2 - \tfrac{3125}{3906}x + \tfrac{25}{3906}y\\
\tilde{g}_4 &= y^2 + 0.008xy - 0.8\}
\end{aligned}$$

When we vary the coefficients of xy in the two generators from zero to 0.01, we can see that one border bases changes continuously into the other. Thus the border basis behaves numerically stable under small perturbations of the coefficient of xy.

1.2.B. Characterizations of border bases. In the sequel, we use the following notation: let $\mathcal{O} = \{t_1, \ldots, t_\mu\}$ be an order ideal in $\mathbb{T}^n$, let $\partial\mathcal{O} = \{b_1, \ldots, b_\nu\}$ be the border of $\mathcal{O}$, let $G = \{g_1, \ldots, g_\nu\}$ be an $\mathcal{O}$-border prebasis, where $g_j = b_j - \sum_{i=1}^{\mu} c_{ij}t_i$ with $c_{ij} \in K$, and let $I = (g_1, \ldots, g_\nu)$ be the ideal generated by G.

The following remark collects some elementary properties of border bases.

Remark 1.2.4. Let $J \subseteq P$ be a zero-dimensional ideal.

a) The ideal J need not have an $\mathcal{O}$-border basis, even if its colength is μ. But if it does, its $\mathcal{O}$-border basis is uniquely determined.
b) If $\mathcal{O}$ is of the form $\mathbb{T}^n \setminus \mathrm{LT}_\sigma(J)$ for some term ordering σ, then J has an $\mathcal{O}$-border basis. It contains the reduced σ-Gröbner basis of J.
c) There exists a **Division Algorithm** for border prebases (see [21], 6.4.11).

The following characterizations of border bases can be shown in analogy to the corresponding results for Gröbner bases (see [21], 6.4.23 and 6.4.28). For a term $t \in \mathbb{T}^n$, its **$\mathcal{O}$-index** $\mathrm{ind}_{\mathcal{O}}(t)$ is the smallest natural number k such that $t = t_1t_2$ with $t_1 \in \mathcal{O}$ and $t_2 \in \mathbb{T}^n_k$.

Proposition 1.2.5. *In the above setting, the set G is an $\mathcal{O}$-border basis of I if and only if one of the following equivalent conditions is satisfied.*

a) For every $f \in I \setminus \{0\}$, there are $f_1, \ldots, f_\nu \in P$ such that $f = f_1g_1 + \cdots + f_\nu g_\nu$ and $\deg(f_i) \le \mathrm{ind}_{\mathcal{O}}(f) - 1$ whenever $f_ig_i \neq 0$.
b) For every $f \in I \setminus \{0\}$, there are $f_1, \ldots, f_\nu \in P$ such that $f = f_1g_1 + \cdots + f_\nu g_\nu$ and $\max\{\deg(f_i) \mid i \in \{1, \ldots, \nu\}, f_ig_i \neq 0\} = \mathrm{ind}_{\mathcal{O}}(f) - 1$.

Proposition 1.2.6. *In the above setting, the set G is an $\mathcal{O}$-border basis of I if and only if the rewrite relation $\xrightarrow{G}$ associated to G is confluent.*

As we mentioned above, the vector space basis $\mathcal{O}$ of P/I can be used to describe the K-algebra structure of P/I via the multiplication matrices of the multiplication maps by the indeterminates. In addition, these multiplication maps can be used to characterize border bases, as the next theorem shows.

Definition 1.2.7. For $r \in \{1,\dots,n\}$, we define the r-th **formal multiplication matrix** $\mathcal{A}_r$ as follows:

Multiply $t_i \in \mathcal{O}$ by x_r. If $x_r t_i = b_j$ is in the border of $\mathcal{O}$, rewrite it using the prebasis polynomial $g_j = b_j - \sum_{k=1}^{\mu} c_{kj} t_k$ and put $(c_{1j},\dots,c_{\mu j})$ into the i-th column of $\mathcal{A}_r$. But if $x_r t_i = t_j$ then put the j-th unit vector into the i-th column of $\mathcal{A}_r$.

Clearly, if G is a border basis and $\mathcal{A}_1,\dots,\mathcal{A}_n$ are the actual multiplication matrices, they commute because P/I is a commutative ring. Surprisingly, the converse holds, too.

Theorem 1.2.8. *(Mourrain [26])*
The set G is the $\mathcal{O}$-border basis of I if and only if the formal multiplication matrices commute, i.e. iff

$$\mathcal{A}_i \mathcal{A}_j = \mathcal{A}_j \mathcal{A}_i \qquad \text{for } 1 \le i < j \le n.$$

For a detailed proof, see [21], 6.4.30. Let us check this result in a concrete case.

Example 1.2.9. In Example 1.2.2 the multiplication matrices are given by

$$\mathcal{A}_x = \begin{pmatrix} 0 & 0.8 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.8 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad \mathcal{A}_y = \begin{pmatrix} 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0.8 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

To check this, we let $t_1 = 1$, $t_2 = x$, $t_3 = y$, and $t_4 = xy$. Then we note that for instance $xt_1 = t_2$ means that the first column of $\mathcal{A}_x$ is $(0,1,0,0)$. If we compute $xt_2 = x^2$, we have to use the coefficients of the corresponding border prebasis polynomial g_1 and put $(0.8,0,0,0)$ into the second column of $\mathcal{A}_1$, etc.

1.2.C. Neighbours and their syzygies. Our next goal is to generalize the Buchberger Criterion for Gröbner bases (see [20], 2.5.3) to the border basis setting. The Buchberger criterion is based on the notion of *lifting syzygies*. Given an order ideal $\mathcal{O} = \{t_1,\dots,t_\mu\}$ and its border $\{b_1,\dots,b_\nu\}$, it is well-known that the syzygy module

$$\mathrm{Syz}_P(b_1,\dots,b_\nu) = \{(f_1,\dots,f_\nu \in P^\nu \mid f_1 b_1 + \cdots + f_\nu b_\nu = 0\}$$

is generated by the **fundamental syzygies**

$$\sigma_{ij} = (\mathrm{lcm}(b_i,b_j)/b_i)\, e_i - (\mathrm{lcm}(b_i,b_j)/b_j)\, e_j$$

with $1 \leq i < j \leq \nu$. However, this system of generators is not minimal and a much smaller subset suffices to generate the same module. The following terminology will be useful to describe such as subset.

Definition 1.2.10. Let $b_i, b_j \in \partial\mathcal{O}$ be two distinct border terms.

a) The border terms b_i and b_j are called **next-door neighbours** if $b_i = x_k b_j$ for some $k \in \{1,\dots,n\}$.
b) The border terms b_i and b_j are called **across-the-street neighbours** if there are $k,\ell \in \{1,\dots,n\}$ such that $x_k b_i = x_\ell b_j$.
c) The border terms b_i and b_j are called **neighbours** if they are next-door neighbours or across-the-street neighbours.
d) The graph whose vertices are the border terms and whose edges are given by the neighbour relation is called the **border web** of $\mathcal{O}$.

Example 1.2.11. For instance, in Example 1.2.2 the border is $\partial\mathcal{O} = \{b_1, b_2, b_3, b_4\}$ with $b_1 = x^2$, $b_2 = x^2 y$, $b_3 = xy^2$, and $b_4 = y^2$. Here we have two next-door neighbour pairs (b_2, b_1), (b_3, b_4) and one across-the-street neighbour pair (b_2, b_3).

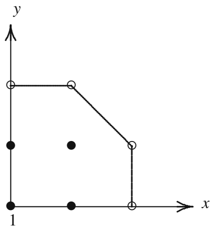


Fig. 1.5 A simple border web.

Proposition 1.2.12. *The border web is connected.*

For a proof, see [17], Prop. 19. Based on the concept of neighbours, we now restrict fundamental syzygies to neighbour pairs.

Definition 1.2.13. Let $\mathcal{O}$ be an order ideal with border $\partial\mathcal{O} = \{b_1, \dots, b_\nu\}$.

a) For next-door neighbours b_i, b_j, *i.e.* for $b_i = x_k b_j$, the fundamental syzygy σ_{ij} has the form $\tau_{ij} = e_i - x_k e_j$ and is called a **next-door neighbour syzygy**.
b) For across-the-street neighbours b_i, b_j, *i.e.* for $x_k b_i = x_\ell b_j$, the fundamental syzygy σ_{ij} has the form $\upsilon_{ij} = x_k e_i - x_\ell e_j$ and is called an **across-the-street neighbour syzygy**.
c) The set of all **neighbour syzygies** is the set of all next-door or across-the street neighbour syzygies.

In [17], Prop. 21, the following result is shown.

Proposition 1.2.14. *The set of neighbour syzygies generates the module of* **border syzygies** $\mathrm{Syz}_P(b_1,\dots,b_\nu)$.

Example 1.2.15. For instance, let us compute the border syzygies for the order ideal $\mathcal{O}=\{1,x,y,xy\}$. We have $\partial\mathcal{O}=\{b_1,b_2,b_3,b_4\}$ with $b_1=x^2$, $b_2=x^2y$, $b_3=xy^2$, and $b_4=y^2$, and the neighbour pairs (b_1,b_2), (b_2,b_3), (b_3,b_4). Therefore the border syzygy module $\mathrm{Syz}_P(b_1,b_2,b_3,b_4)$ is generated by the following three neighbour syzygies:

$$\begin{aligned} e_2-ye_1 &= (-y,\,1,\,0,\,0)\\ ye_2-xe_3 &= (0,\,y,\,-x,\,0)\\ e_4-xe_3 &= (0,\,0,\,-x,\,1) \end{aligned}$$

In order to transfer the Buchberger Criterion from Gröbner to border bases, it suffices to lift neighbour syzygies.

Definition 1.2.16. Let $g_i,g_j\in G$ be two distinct border prebasis polynomials. Then the polynomial

$$S_{ij}=(\mathrm{lcm}(b_i,b_j)/b_i)\cdot g_i-(\mathrm{lcm}(b_i,b_j)/b_j)\cdot g_j$$

is called the **S-polynomial** of g_i and g_j.

Remark 1.2.17. Let $g_i,g_j\in G$.

a) If (b_i,b_j) are next-door neighbours with $b_j=x_kb_i$ then the S-polynomial S_{ij} is of the form $S_{ij}=g_j-x_kg_i$.
b) If (b_i,b_j) are across-the-street neighbours with $x_kb_i=x_\ell b_j$ then S_{ij} is of the form $S_{ij}=x_kg_i-x_\ell b_j$.

In both cases we see that the support of S_{ij} is contained in $\mathcal{O}\cup\partial\mathcal{O}$. Hence there exists constants $a_i\in K$ such that the support of

$$\mathrm{NR}_{\mathcal{O},G}(S_{ij})=S_{ij}-\sum_{m=1}^{\mu}a_m g_m\in I$$

is contained in $\mathcal{O}$. If G is a border basis, this implies $\mathrm{NR}_{\mathcal{O},G}(S_{ij})=0$. We shall say that the syzygy $e_j-x_ke_i-\sum_{m=1}^{\mu}a_me_m$ resp. $x_k\,e_i-x_\ell e_j-\sum_{m=1}^{\mu}a_me_m$ is a **lifting** of the neighbour syzygy $e_j-x_ke_i$ resp. $x_k\,e_i-x_\ell e_j$.

Theorem 1.2.18. *(Stetter [30])*
An $\mathcal{O}$-border prebasis G is an $\mathcal{O}$-border basis if and only if the neighbour syzygies lift, i.e. if and only if we have

$$\mathrm{NR}_{\mathcal{O},G}(S_{ij})=0$$

for all (i,j) such that (b_i,b_j) is a pair of neighbours.

The proof of this theorem is pretty involved. Let us briefly describe the idea. The vanishing conditions for the normal remainders of the S-polynomials entail certain equalities which have to be satisfied by the coefficients c_{ij} of the border prebasis polynomials. Using a (rather nasty) case-by-case argument, one checks that these are the same equalities that one gets from the conditions that the formal multiplication matrices have to commute. A detailed version of this proof is contained in [21], Section 6.4.

Example 1.2.19. Let us look at these conditions for $\mathcal{O}=\{1,x,y,xy\}$. An $\mathcal{O}$-border prebasis $G=\{g_1,g_2,g_3,g_4\}$ is of the form

$$\begin{aligned}
g_1 &= x^2 - c_{11}\cdot 1 - c_{21}\,x - c_{31}\,y - c_{41}\,xy\\
g_2 &= x^2y - c_{12}\cdot 1 - c_{22}\,x - c_{32}\,y - c_{42}\,xy\\
g_3 &= xy^2 - c_{13}\cdot 1 - c_{23}\,x - c_{33}\,y - c_{43}\,xy\\
g_4 &= y^2 - c_{14}\cdot 1 - c_{24}\,x - c_{34}\,y - c_{44}\,xy
\end{aligned}$$

The S-polynomials of its neighbour syzygies are

$$\begin{aligned}
S_{21} &= g_2 - yg_1\\
&= -c_{12} - c_{22}x + (c_{11}-c_{32})y + (c_{21}-c_{42})xy + c_{31}y^2 + c_{41}xy^2\\
S_{23} &= yg_2 - xg_3\\
&= c_{13}x - c_{22}y + (c_{33}-c_{22})xy + c_{23}x^2 + c_{43}x^2y - c_{42}xy^2 - c_{32}y^2\\
S_{34} &= g_3 - xg_4\\
&= -c_{13} + (c_{14}-c_{23})x - c_{33}y + (c_{34}-c_{43})xy + c_{24}x^2 + c_{44}x^2y
\end{aligned}$$

Their normal remainders with respect to G are

$$\begin{aligned}
\mathrm{NR}_{\mathcal{O},G}(S_{21}) &= (-c_{12}+c_{31}c_{14}+c_{41}c_{13}) + (-c_{22}+c_{31}c_{24}+c_{41}c_{23})x\\
&\quad +(c_{11}-c_{32}+c_{31}c_{34}+c_{41}c_{33})y + (c_{21}-c_{42}+c_{31}c_{44}+c_{41}c_{43})xy\\
\mathrm{NR}_{\mathcal{O},G}(S_{23}) &= (c_{11}c_{23}+c_{12}c_{43}-c_{42}c_{13}-c_{32}c_{14}) + (c_{21}c_{23}+c_{22}c_{43}\\
&\quad -c_{42}c_{23}-c_{32}c_{24}+c_{13})x + (-c_{12}+c_{31}c_{23}+c_{32}c_{43}-c_{42}c_{33}-c_{32}c_{34})y\\
&\quad +(c_{33}-c_{22}+c_{41}c_{23}-c_{32}c_{44})xy\\
\mathrm{NR}_{\mathcal{O},G}(S_{34}) &= (-c_{13}+c_{11}c_{24}+c_{12}c_{44}) + (c_{14}-c_{23}+c_{21}c_{24}+c_{23}c_{44})x\\
&\quad +(-c_{33}+c_{31}c_{24}+c_{32}c_{44})y + (c_{34}-c_{43}+c_{41}c_{24}+c_{42}c_{44})xy
\end{aligned}$$

Here G is a border basis if and only if these 12 coefficients are zero. In Example 1.6.14 we shall examine the scheme defined by these vanishing conditions.

1.3 The Eigenvalue Method for Solving Polynomial Systems

When working toward the solution of a problem,
it always helps if you know the answer.
(Rule of Accuracy)

As said in the introduction, this paper deals mainly with the problem of reconstructing polynomial equations from data. The *opposite problem* of solving polynomial systems is also well-known since it plays a key role in many applications. Rather than trying to discuss this problem in its full generality, we will now have a look at a nice method which deserves to be more widely known in the commutative algebra community. While the so called `Lex`-method is amply described in the literature (see for instance [20], Section 3.7), we are going to describe an idea on how to use classical methods in linear algebra to solve polynomial systems. The pioneering work was done in [4] and [5], and a nice introduction can be found in [9].

In the following we let K be a field and $P = K[x_1,\dots,x_n]$. Let the polynomial system be defined by $f_1,\dots,f_s \in P$. Then we let $I = (f_1,\dots,f_s)$ and $A = P/I$. We assume that the ideal I is zero-dimensional, so that A is a finite dimensional K-vector space.

Definition 1.3.1. Given an element f in P, we define a K-linear map $m_f : A \longrightarrow A$ by $m_f(g) = fg \mod I$ and call it the **multiplication map** defined by f. We also consider the induced K-linear map on the dual spaces $m_f^* : A^* \longrightarrow A^*$ defined by $m_f^*(\varphi) = \varphi \circ m_f$.

If we know a vector space basis of A, we can represent a multiplication map by its matrix with respect to this basis. Let us have a look at a concrete case.

Example 1.3.2. Let $P = \mathbb{R}[x]$, let $f = x^2+1$, and I be the principal ideal generated by $x^3-x^2+x-1 = (x-1)(x^2+1)$. The residue classes of the terms in $\{1,x,x^2\}$ form a vector space basis of P/I. Using the two relations $x^3+x \equiv x^2+1 \mod I$ and $x^4+x^2 \equiv x^2+1 \mod I$, we see that the matrix which represents m_f with respect to this basis is

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

The next theorem provides an important link between $\mathscr{Z}(I)$, the set of *zeros* of I over the algebraic closure $\overline{K}$ of K, and the eigenvalues of multiplication maps. Let $\overline{P} = \overline{K}[x_1,\dots,x_n]$ and $\overline{A} = \overline{P}/I\overline{P}$. An element $\lambda \in \overline{K}$ is called a $\overline{K}$**-eigenvalue** of a multiplication map $m_f : A \longrightarrow A$ if it is a zero of the **characteristic polynomial** of m_f, *i.e.* if the $\overline{K}$-linear map $\varphi_{f,\lambda} : \overline{A} \longrightarrow \overline{A}$ defined by $\bar{g} \mapsto \bar{f}\bar{g} - \lambda\,\bar{g}$ is not invertible.

Theorem 1.3.3. *Let I be a zero-dimensional ideal in P, let $f \in P$, and let $\lambda \in \overline{K}$, Then the following conditions are equivalent.*

a) The element λ is a $\overline{K}$-eigenvalue of m_f.

b) There exists a point $p \in \mathscr{Z}(I)$ such that $\lambda = f(p)$.

In this equivalence, if $p \in \mathscr{Z}_K(I)$, we have $\lambda \in K$.

Proof. Let us first prove $a) \implies b)$. If λ does not coincide with any of the values of f at the points $p \in \mathscr{Z}(I)$, the ideal $J = I\overline{P} + (f - \lambda) \subseteq \overline{P}$ satisfies $\mathscr{Z}(J) = \emptyset$. Thus the Weak Nullstellensatz (see [20], Corollary 2.6.14) yields $1 \in J$. Therefore there exist $g \in \overline{P}$ and $h \in I\overline{P}$ such that $1 = g(f - \lambda) + h$. Consequently, we have $1 \equiv g(f - \lambda) \mod I\overline{P}$, so that $\varphi_{f,\lambda}$ is invertible with inverse -$m_{\bar{g}}$. Therefore λ is not a $\overline{K}$-eigenvalue of m_f.

Now let us prove the implication $b) \implies a)$. If λ is not a $\overline{K}$-eigenvalue of m_f then $\varphi_{f,\lambda}$ is an invertible map. In particular, it is surjective, and thus there exists $g \in \overline{P}$ such that $g\,(f - \lambda) \equiv 1 \mod I\overline{P}$. Clearly, this implies that there cannot exist a point $p \in \mathscr{Z}(I)$ such that $f(p) - \lambda = 0$.

The additional claim follows from $\lambda = f(p)$. □

In the setting of Example 1.3.2, the eigenvalues of m_f and the zeros of I are related as follows.

Example 1.3.4. As in Example 1.3.2, we let $I = (x^3 - x^2 + x - 1) \subseteq P = \mathbb{R}[x]$ and $f = x^2 + 1$. Since m_f is singular, the element $\lambda_1 = 0$ is an eigenvalue. And indeed, we have $\mathscr{Z}(I) = \{1, i, -i\}$ and $f(i) = f(-i) = 0$. For the other eigenvalue $\lambda_2 = 2$, we have $f(1) = 2$. Notice that here we have $\lambda_1 \in \mathbb{R}$, but the corresponding zeros of I are not real numbers.

The above theorem can be used in several ways to compute the solutions of a system of polynomial equations. One method is based on the following observation.

Corollary 1.3.5. *Let $i \in \{1, \ldots, n\}$. The i^{th} coordinates of the points of $\mathscr{Z}(I)$ are the $\overline{K}$-eigenvalues of the multiplication map m_{x_i}.*

Proof. This follows immediately from the theorem, since $x_i(p)$ is exactly the i^{th} coordinate of a point $p \in \overline{K}^n$. □

Hence we can determine $\mathscr{Z}(I)$ in the following way. Fix a tuple of polynomials $E = (t_1, \ldots, t_\mu)$ whose residue classes form a K-basis of A. For $f \in P$, we let $fE = (ft_1, \ldots, ft_\mu)$ and describe the multiplication map $m_f : A \longrightarrow A$ by the matrix M_{fE}^E whose the j^{th} column $(a_{1j}, \ldots, a_{\mu j})^{\text{tr}}$ is given by

$$f\,t_j \equiv a_{1j}t_1 + \cdots + a_{\mu j}t_\mu \mod I$$

A more compact way of expressing this fact is the formula

$$fE \equiv E \cdot M_{fE}^E \mod I \tag{$*$}$$

For the tuple E, we usually choose an order ideal of terms (see Definition 1.2.1). In particular, we shall assume that we have $t_1 = 1$.

If the ideal I contains a linear polynomial, we can reduce the problem of computing $\mathscr{Z}(I)$ to a problem for an ideal in a polynomial ring having fewer indeterminates. Thus we shall now assume that I contains no linear polynomial. Consequently, we suppose that the indeterminates are in E, specifically that we have $t_2 = x_1$, ..., $t_{n+1} = x_n$.

One method of finding $\mathscr{Z}(I)$ is to compute the $\overline{K}$-eigenvalues $\lambda_{i1},\dots,\lambda_{i\mu}$ of $M^E_{x_iE}$ for $i = 1,\dots,n$ and then to check for all points $(\lambda_{1j_1},\dots,\lambda_{nj_n})$ such that $j_1,\dots,j_n \in \{1,\dots,\mu\}$ whether they are zeros of I. Clearly, this approach has several disadvantages:

1. Usually, the $\overline{K}$-eigenvalues of the multiplication matrices $M^E_{x_iE}$ can only be determined approximately.
2. The set of candidate points is a grid which is typically much larger than the set $\mathscr{Z}(I)$.

A better approach uses the next theorem. For a K-linear map $\varphi : A \longrightarrow A$, we let $\bar{\varphi} = \varphi \otimes_K \overline{K} : \overline{A} \longrightarrow \overline{A}$. Given a $\overline{K}$-eigenvalue $\lambda \in \overline{K}$ of φ, the $\overline{K}$-vector space $\ker(\bar{\varphi} - \lambda\, \mathrm{id}_{\overline{A}})$ is called the corresponding **$\overline{K}$-eigenspace** and its non-zero vectors are called the corresponding **$\overline{K}$-eigenvectors**. For the matrices representing φ, we use a similar terminology.

Theorem 1.3.6. *In the above setting, let $f \in P$, let $p \in \mathscr{Z}(I)$, and let $E = (t_1,\dots,t_\mu)$ be a tuple of polynomials whose residue classes form a K-basis of A. Then the vector $E(p)^{\mathrm{tr}} = (t_1(p),\dots,t_\mu(p))^{\mathrm{tr}}$ is a $\overline{K}$-eigenvector of $(M^E_{fE})^{\mathrm{tr}}$ corresponding to the $\overline{K}$-eigenvalue $f(p)$.*

Proof. When we evaluate both sides of the above formula $(*)$ at p, we get the equality $f(p)\,E(p) = E(p)\,M^E_{fE}$. Transposing both sides yields

$$f(p)(E(p))^{\mathrm{tr}} = (M^E_{fE})^{\mathrm{tr}}\, E(p)^{\mathrm{tr}}$$

and this is precisely the claim. □

Note that the matrix $(M^E_{fE})^{\mathrm{tr}}$ represents the linear map m^*_f (see Definition 1.3.1). To make good use of this theorem, we need the following notion.

Definition 1.3.7. A matrix $M \in \mathrm{Mat}_\mu(K)$ is called **$\overline{K}$-non-derogatory** if it has the property that all its $\overline{K}$-eigenspaces are 1-dimensional.

It is a well-known result in Linear Algebra that this condition is equivalent to requiring that the Jordan canonical form of M over $\overline{K}$ has one Jordan block per eigenvalue, or to the condition that the minimal polynomial and the characteristic polynomial of M agree. Using the preceding theorem and a non-derogatory multiplication matrix, we can solve a zero-dimensional polynomial system as follows.

Corollary 1.3.8. *Let $E = (t_1,\dots,t_\mu)$ be a tuple of polynomials whose residue classes form a K-basis of A, let $1, x_1,\dots,x_n$ be the first $n+1$-entries of E, and let $f \in P$ be such that the matrix $(M^E_{fE})^{\mathrm{tr}}$ is $\overline{K}$-non-derogatory. Let $V_1,\dots,V_r$ be* *the $\overline{K}$-eigenspaces of this matrix. For $j = 1,\dots,r$, choose a basis vector v_j of V_j of the form $v_j = (1, a_{2j},\dots,a_{\mu j})$ with $a_{ij} \in \overline{K}$. Then $\mathscr{Z}(I)$ consists of the points $p_j = (a_{2j},\dots,a_{n+1\,j})$ such that $j \in \{1,\dots,r\}$.*

Proof. Let $j \in \{1,\dots,r\}$. By Theorem 1.3.6, the vector $E(p_j) = (t_1(p_j),\dots,t_\mu(p_j))$ is a $\overline{K}$-eigenvector of $(M_{fE}^E)^{\mathrm{tr}}$ corresponding to the $\overline{K}$-eigenvalue $f(p_j)$. Hence it is a non-zero vector in V_j. Since V_j is 1-dimensional and $t_1(p_j) = 1$ equals the first component of v_j, we have the equality $E(p_j) = v_j$. Now the observation that $E(p_j) = (1, x_1(p_j),\dots,x_n(p_j),\dots)$ finishes the proof. □

The technique given in this corollary addresses the second problem stated above: no exponentially large set of candidate points has to be examined. However, we note that the first problem still persists. For instance, if $K = \mathbb{Q}$, instead of the $\overline{\mathbb{Q}}$-eigenvalues of $(M_{fE}^E)^{\mathrm{tr}}$ we can usually only compute approximate eigenvalues. Hence the corresponding $\overline{\mathbb{Q}}$-eigenspaces are not computable as true kernels of linear maps. However, in the next section we will introduce *approximate kernels* of linear maps which take care of this task.

Let us end this section with an example which illustrates the methods described above.

Example 1.3.9. Let I be the ideal in $P = \mathbb{R}[x,y]$ generated by the set of polynomials $\{x^2 + 4/3xy + 1/3y^2 - 7/3x - 5/3y + 4/3,\ y^3 + 10/3xy + 7/3y^2 - 4/3x - 20/3y + 4/3,\ xy^2 - 7/3xy - 7/3y^2 - 2/3x + 11/3y + 2/3\}$. It is easy to check that this set is a Gröbner basis of I with respect to $\sigma = \mathtt{DegRevLex}$. Hence $E = \{1, x, y, xy, y^2\}$ is an order ideal of terms whose residue classes form a K-basis of P/I. By computing the normal forms $\mathrm{NF}_{\sigma,I}(x^2)$, $\mathrm{NF}_{\sigma,I}(x^2y)$, $\mathrm{NF}_{\sigma,I}(xy^2, I)$ and $\mathrm{NF}_{\sigma,I}(y^3)$, we get the multiplication matrices

$$M_{xE}^E = \begin{pmatrix} 0 & -4/3 & 0 & 4/3 & -2/3 \\ 1 & 7/3 & 0 & -4/3 & 2/3 \\ 0 & 5/3 & 0 & 4/3 & -11/3 \\ 0 & -4/3 & 1 & 1/3 & 7/3 \\ 0 & -1/3 & 0 & -2/3 & 7/3 \end{pmatrix} \text{ and } M_{yE}^E = \begin{pmatrix} 0 & 0 & 0 & -2/3 & -4/3 \\ 0 & 0 & 0 & 2/3 & 4/3 \\ 1 & 0 & 0 & -11/3 & 20/3 \\ 0 & 1 & 0 & 7/3 & -10/3 \\ 0 & 0 & 1 & 7/3 & -7/3 \end{pmatrix}$$

First, let us follow the method of Corollary 1.3.5. The characteristic polynomial of M_{xE}^E is $(x+1)(x-1)^2(x-2)^2$ and the characteristic polynomial of M_{yE}^E is given by $x(x-1)(x+1)(x-2)(x+2)$. If we check the 15 candidate points, we find that five of them, namely $(1,0)$, $(1,1)$, $(2,-1)$, $(-1,2)$, and $(2,-2)$ form the set of zeros of I.

Now we apply the method of Corollary 1.3.8. The characteristic polynomial of $(M_{xE}^E)^{\mathrm{tr}}$ is the same as that of M_{xE}^E. It is easy to check (for instance, using CoCoA) that the dimension of the eigenspace corresponding to the eigenvalue 1 is 2. Therefore the matrix $(M_{xE}^E)^{\mathrm{tr}}$ is derogatory and cannot be used for the proposed method.

On the other hand, the characteristic polynomial of the matrix $(M_{yE}^E)^{\mathrm{tr}}$ is given by $x(x-1)(x+1)(x-2)(x+2)$. Consequently, this matrix is non-derogatory. We compute basis vectors for its eigenspaces and norm them to have first component 1.

The result is $v_1 = (1,1,0,0,0)$, $v_2 = (1,1,1,1,1)$, $v_3 = (1,2,-1,-2,1)$, $v_4 = (1,-1,2,-2,4)$, and $v_5 = (1,2,-2,-4,4)$. We get $\mathscr{Z}(I) = \{(1,0), (1,1), (2,-1), (-1,2), (2,-2)\}$, as before.

1.4 Approximate Vanishing Ideals

Two is not equal to three;
not even for large values of two.
(Grabel's Law)

It is time to enter the *real world*. When dealing with industrial applications, we do not always have exact data available. Thus our computations have to be based on measured values with intrinsic errors. How can we perform symbolic computation in this world? Let us start to discuss this question in a first relevant case. Then, based on our answer, we shall present an actual industrial example. We want to deal with the following situation. Let $\mathbb{X} = \{p_1,\dots,p_s\}$ be a set of s points in $\mathbb{R}^n$. These points are meant to represent measured values. In the computer, they will be stored as tuples of floating point numbers.

If $\mathbb{X}$ was an exact set of points, we could compute its **vanishing ideal**

$$I(\mathbb{X}) = \{f \in \mathbb{R}[x_1,\dots,x_n] \mid f(p_1) = \cdots = f(p_s) = 0\}$$

However, in the presented setting, it is well-known that this leads to a numerically unstable and virtually meaningless result. Instead, we are looking for a reasonable definition of an **approximate vanishing ideal** of $\mathbb{X}$. To this end, we have to overcome a number of impediments. First of all, we need a **threshold number** $\varepsilon \in \mathbb{R}_+$. We say that a polynomial $f \in \mathbb{R}[x_1,\dots,x_n]$ **vanishes ε-approximately** at $\mathbb{X}$ if $|f(p_i)| < \varepsilon$ for $i = 1,\dots,s$. This definition entails several problems.

1. The polynomials which vanish ε-approximately at $\mathbb{X}$ do not form an ideal!
2. All polynomials with very small coefficients vanish ε-approximately at $\mathbb{X}$!

To address the second problem, we introduce a topology on the polynomial ring $P = \mathbb{R}[x_1,\dots,x_n]$.

Definition 1.4.1. Let $f = a_1t_1 + \cdots + a_st_s \in P$, where $a_1,\dots,a_s \in \mathbb{R} \setminus \{0\}$ and $t_1,\dots,t_s \in \mathbb{T}^n$. Then the number $\|f\| = \|(a_1,\dots,a_s)\| = \sqrt{a_1^2 + \cdots + a_s^2}$ is called the **(Euclidean) norm** of f.

Clearly, this definition turns P into a normed vector space. A polynomial $f \in P$ with $\|f\| = 1$ will be called **unitary**. Now it is reasonable to consider the condition that polynomials $f \in P$ with $\|f\| = 1$ vanish ε-approximately at $\mathbb{X}$, and we can try the following definition.

Definition 1.4.2. An ideal $I \subseteq P$ is called an ε**-approximate vanishing ideal** of $\mathbb{X}$ if there exists a system of generators $\{f_1,\dots,f_r\}$ of I such that $\|f_i\| = 1$ and f_i vanishes ε-approximately at $\mathbb{X}$ for $i = 1,\dots,r$.

In itself, this definition is certainly still too loose. For instance, it is clear that the unit ideal is always an ε-approximate vanishing ideal. Nevertheless, we shall see below that we arrive at a very usable definition if we impose additional structure on the generators. Before we move to this topic, though, we need two additional ingredients.

1.4.A. The Singular Value Decomposition (SVD). In approximate computation, we frequently have to decide whether something is zero or not. The following theorem and its corollary can be used to determine the vectors which are approximately in the kernel of a linear map of $\mathbb{R}$-vector spaces.

Theorem 1.4.3 (The Singular Value Decomposition).
Let $\mathscr{A} \in \mathrm{Mat}_{m,n}(\mathbb{R})$.

1. *There are orthogonal matrices $\mathscr{U} \in \mathrm{Mat}_{m,m}(\mathbb{R})$ and $\mathscr{V} \in \mathrm{Mat}_{n,n}(\mathbb{R})$ and a matrix $\mathscr{S} \in \mathrm{Mat}_{m,n}(\mathbb{R})$ of the form $\mathscr{S} = \begin{pmatrix} \mathscr{D} & 0 \\ 0 & 0 \end{pmatrix}$ such that*
$$\mathscr{A} = \mathscr{U} \cdot \mathscr{S} \cdot \mathscr{V}^{\mathrm{tr}} = \mathscr{U} \cdot \begin{pmatrix} \mathscr{D} & 0 \\ 0 & 0 \end{pmatrix} \cdot \mathscr{V}^{\mathrm{tr}}$$
where $\mathscr{D} = \mathrm{diag}(s_1,\dots,s_r)$ is a diagonal matrix.
2. *In this decomposition, it is possible to achieve $s_1 \geq s_2 \geq \cdots \geq s_r > 0$. The numbers $s_1,\dots,s_r$ depend only on $\mathscr{A}$ and are called the* **singular values** *of $\mathscr{A}$.*
3. *The number r is the rank of $\mathscr{A}$.*
4. *The matrices $\mathscr{U}$ and $\mathscr{V}$ have the following interpretation:*
$$\begin{aligned} \textit{first } r \textit{ columns of } \mathscr{U} &\equiv \textit{ONB of the column space of } \mathscr{A} \\ \textit{last } m-r \textit{ columns of } \mathscr{U} &\equiv \textit{ONB of the kernel of } \mathscr{A}^{\mathrm{tr}} \\ \textit{first } r \textit{ columns of } \mathscr{V} &\equiv \textit{ONB of the row space of } \mathscr{A} \\ &\equiv \textit{ONB of the column space of } \mathscr{A}^{\mathrm{tr}} \\ \textit{last } n-r \textit{ columns of } \mathscr{V} &\equiv \textit{ONB of the kernel of } \mathscr{A} \end{aligned}$$
Here ONB is an abbreviation for "orthonormal basis".

For a proof, see for instance [12], Sections 2.5.3 and 2.6.1. The SVD of a real matrix allows us to define and compute its **approximate kernel**.

Corollary 1.4.4. *Let $\mathscr{A} \in \mathrm{Mat}_{m,n}(\mathbb{R})$, and let $\varepsilon > 0$ be given. Choose $k \in \{1,\dots,r\}$ such that $s_k > \varepsilon \geq s_{k+1}$, form the matrix $\widetilde{\mathscr{S}}$ by setting $s_{k+1} = \cdots = s_r = 0$ in $\mathscr{S}$, and let $\widetilde{\mathscr{A}} = \mathscr{U}\,\widetilde{\mathscr{S}}\,\mathscr{V}^{\mathrm{tr}}$.*

1. *We have $\min\{\|\mathscr{A} - \mathscr{B}\| : \mathrm{rank}(\mathscr{B}) \leq k\} = \|\mathscr{A} - \widetilde{\mathscr{A}}\| = s_{k+1}$. (Here $\|\cdots\|$ denotes the 2-operator norm of a matrix.)*

2. *The vector subspace* $\operatorname{apker}(\mathscr{A},\varepsilon) = \ker(\widetilde{\mathscr{A}})$ *is the largest dimensional kernel of a matrix whose Euclidean distance from* $\mathscr{A}$ *is at most* ε. *It will be called the* **ε-approximate kernel** *of* $\mathscr{A}$.
3. *The last* $n-k$ *columns* $v_{k+1},\dots,v_n$ *of* $\mathscr{V}$ *are an ONB of* $\operatorname{apker}(\mathscr{A},\varepsilon)$. *They satisfy* $\|\mathscr{A}v_i\| < \varepsilon$.

Proof. See [12], Section 2.5.4 and the theorem. To prove the third claim, observe that $\|\mathscr{A}v_i\| = \|(\mathscr{A}-\widetilde{\mathscr{A}})v_i\| \le \|\mathscr{A}-\widetilde{\mathscr{A}}\| < \varepsilon$. □

1.4.B. The Stable Reduced Row Echelon Form. Our next task is to find the leading terms contained in a vector space of polynomials. Again we are of course interested in leading terms of unitary polynomials for which the leading coefficient is not smaller than a given threshold number.

Let $V \subset P$ be a finite dimensional vector space of polynomials. Given a term ordering σ and a basis $B = \{f_1,\dots,f_r\}$ of V, We can identify V with a real matrix as follows.

Definition 1.4.5. Let $S = \operatorname{Supp}(f_1) \cup \cdots \cup \operatorname{Supp}(f_r)$, and write $S = \{t_1,\dots,t_s\}$ where the terms $t_i \in \mathbb{T}^n$ are ordered such that $t_1 \ge_\sigma t_2 \ge_\sigma \cdots \ge_\sigma t_s$. Clearly, the support of every polynomial of V is contained in S. For $i = 1,\dots,r$, we write $f_i = c_{i1}t_1 + \cdots + c_{is}t_s$ with $c_{ij} \in \mathbb{R}$. Then the matrix $M_{\sigma,B} = (c_{ij}) \in \operatorname{Mat}_{r,s}(\mathbb{R})$ is called the **Macaulay matrix** of V with respect to σ and B.

In other words, the columns of $M_{\sigma,B}$ are indexed by the terms in S and the rows correspond to the coefficients of the basis polynomials f_i. If we use Gaussian elimination to bring $M_{\sigma,B}$ into row echelon form, the first non-zero entries of each row will indicate the leading term of the corresponding polynomial. Hence the pivot columns will correspond precisely to the set $\operatorname{LT}_\sigma(V)$ of all leading terms of polynomials in V.

To imitate this in the approximate world, we should perform the Gaussian elimination in a numerically stable way. However, we cannot use complete pivoting, since the order of the rows is fixed by the term ordering. The following adaptation of the QR-decomposition uses partial pivoting and provides the "best" leading terms available under the given circumstances.

Proposition 1.4.6 (Stabilized Reduced Row Echelon Form).
Let $A \in \operatorname{Mat}_{m,n}(\mathbb{R})$ *and* $\tau > 0$ *be given. Let* $a_1,\dots,a_n$ *be the columns of* A. *Consider the following instructions.*

(1) Let $\lambda_1 = \|a_1\|$. *If* $\lambda_1 < \tau$, *we let* $R = (0,\dots,0) \in \operatorname{Mat}_{m,1}(\mathbb{R})$. *Otherwise, we let* $Q = ((1/\lambda_1)a_1) \in \operatorname{Mat}_{m,1}(\mathbb{R})$ *and* $R = (\lambda_1,0,\dots,0) \in \operatorname{Mat}_{m,1}(\mathbb{R})$.

(2) For $i = 2,\dots,n$, *compute* $q_i = a_i - \sum_{j=1}^{i-1}\langle a_i,q_j\rangle q_j$ *and* $\lambda_i = \|q_i\|$. *If* $\lambda_i < \tau$, *append a zero column to* R. *Otherwise, append the column* $(1/\lambda_i)q_i$ *to* Q *and the column* $(\lambda_i\langle a_1,q_1\rangle,\dots,\lambda_i\langle a_{i-1},q_{i-1}\rangle,\lambda_i,0,\dots,0)$ *to* R.

(3) Starting with the last row and working upwards, use the first non-zero entry of each row of R *to clean out the non-zero entries above it.*

(4) For $i = 1,\dots,m$, compute the norm ρ_i of the i-th row of R. If $\rho_i < \tau$, set this row to zero. Otherwise, divide this row by ρ_i. Then return the matrix R.

This is an algorithm which computes a matrix R in reduced row echelon form. The row space of R is contained in the row space of the matrix $\overline{A}$ which is obtained from A by setting the columns whose norm is less than τ to zero. Here the pivot elements of R are not 1, but its rows are unitary vectors.

Furthermore, if the rows of A are unitary and mutually orthogonal, the row vectors of R differ by less than $\tau m \sqrt{n}$ from unitary vectors in the row space of A.

The proof of this proposition in contained in [14], Section 3.

1.4.C. The AVI-Algorithm. Finally we are ready to combine all ingredients and produce an algorithm which computes a "good" system of generators of an approximate vanishing ideal of $\mathbb{X}$. By "good" we mean the following.

Definition 1.4.7. Let $\mathcal{O} = \{t_1,\dots,t_\mu\}$ be an order ideal of terms in $\mathbb{T}^n$, denote its border by $\partial\mathcal{O} = \{b_1,\dots,b_\nu\}$, and let $G = \{g_1,\dots,g_\nu\}$ be an $\mathcal{O}$-border prebasis of the ideal $I = (g_1,\dots,g_\nu)$ in P. Recall that this means that g_j is of the form $g_j = b_j - \sum_{i=1}^{\mu} c_{ij} t_i$ with $c_{ij} \in \mathbb{R}$.

For every pair (i,j) such that b_i, b_j are neighbours in $\partial\mathcal{O}$, we compute the normal remainder $S'_{ij} = \mathrm{NR}_{\mathcal{O},G}(S_{ij})$ of the S-polynomial of g_i and g_j with respect to G. We say that G is an **ε-approximate border basis** of the ideal $I = (G)$ if we have $\|S'_{ij}\| < \varepsilon$ for all such pairs (i,j).

Given a finite set of points $\mathbb{X} = \{p_1,\dots,p_s\}$ in $\mathbb{R}^n$, the first thing one should do in every approximate computation is to normalize the data, *i.e.* to transform $\mathbb{X}$ such that $\mathbb{X} \subset [-1,1]^n$. Then the following algorithm computes an approximate border basis of an approximate vanishing ideal of $\mathbb{X}$.

Theorem 1.4.8 (The Approximate Vanishing Ideal Algorithm (AVI-Algorithm)).

Let $\mathbb{X} = \{p_1,\dots,p_s\} \subset [-1,1]^n \subset \mathbb{R}^n$, let $P = \mathbb{R}[x_1,\dots,x_n]$, let $\mathrm{eval}_{\mathbb{X}} : P \longrightarrow \mathbb{R}^s$ be the associated evaluation map $\mathrm{eval}_{\mathbb{X}}(f) = (f(p_1),\dots,f(p_s))$, and let $\varepsilon > \tau > 0$ be small positive numbers. Moreover, let σ be a degree compatible term ordering. Consider the following sequence of instructions.

A1 *Start with lists $G = \emptyset$, $\mathcal{O} = [1]$, a matrix $\mathcal{M} = (1,\dots,1)^{\mathrm{tr}} \in \mathrm{Mat}_{s,1}(\mathbb{R})$, and $d = 0$.*

A2 *Increase d by one and let L be the list of all terms of degree d in $\partial\mathcal{O}$, ordered decreasingly w.r.t. σ. If $L = \emptyset$, return the pair $(G,\mathcal{O})$ and stop. Otherwise, let $L = (t_1,\dots,t_\ell)$.*

A3 *Let m be the number of columns of $\mathcal{M}$. Form the matrix*

$$\mathcal{A} = (\mathrm{eval}_{\mathbb{X}}(t_1),\dots,\mathrm{eval}_{\mathbb{X}}(t_\ell),\mathcal{M}) \in \mathrm{Mat}_{s,\ell+m}(\mathbb{R}).$$

Using its SVD, calculate a matrix $\mathcal{B}$ whose column vectors are an ONB of the approximate kernel $\mathrm{apker}(\mathcal{A},\varepsilon)$.

A4 *Using Proposition 1.4.6, compute the stabilized reduced row echelon form of* $\mathscr{B}^{\mathrm{tr}}$ *with respect to the given* τ. *The result is a matrix* $\mathscr{C} = (c_{ij}) \in \mathrm{Mat}_{k,\ell+m}(\mathbb{R})$ *such that* $c_{ij} = 0$ *for* $j < \nu(i)$. *Here* $\nu(i)$ *denotes the column index of the pivot element in the* i^{th} *row of* $\mathscr{C}$.

A5 *For all* $j \in \{1,\dots,\ell\}$ *such that there exists a* $i \in \{1,\dots,k\}$ *with* $\nu(i) = j$ *(i.e. for the column indices of the pivot elements), append the polynomial*

$$c_{ij}t_j + \sum_{j'=j+1}^{\ell} c_{ij'}t_{j'} + \sum_{j'=\ell+1}^{\ell+m} c_{ij'}u_{j'}$$

to the list G, *where* $u_{j'}$ *is the* $(j'-\ell)^{\mathrm{th}}$ *element of* $\mathscr{O}$.

A6 *For all* $j = \ell, \ell-1, \dots, 1$ *such that the* j^{th} *column of* $\mathscr{C}$ *contains no pivot element, append the term* t_j *as a new first element to* $\mathscr{O}$ *and append the column* $\mathrm{eval}_{\mathbb{X}}(t_j)$ *as a new first column to* $\mathscr{M}$.

A7 *Using the SVD of* $\mathscr{M}$, *calculate a matrix* $\mathscr{B}$ *whose column vectors are an ONB of* $\mathrm{apker}(\mathscr{M}, \varepsilon)$.

A8 *Repeat steps* **A4** – **A7** *until* $\mathscr{B}$ *is empty. Then continue with step* **A2**.

This is an algorithm which computes a pair $(G, \mathscr{O})$ *of sets* $G = \{g_1, \dots, g_\nu\}$ *and* $\mathscr{O} = \{t_1, \dots, t_\mu\}$ *with the following properties:*

a) The set G *consists of unitary polynomials which generate a* δ*-approximate vanishing ideal of* $\mathbb{X}$, *where* $\delta = \varepsilon\sqrt{\nu} + \tau\nu(\mu+\nu)$.

b) The set $\mathscr{O} = \{t_1, \dots, t_\mu\}$ *contains an order ideal of terms such that there is no unitary polynomial in* $\langle \mathscr{O} \rangle_K$ *which vanishes* ε*-approximately on* $\mathbb{X}$.

c) The set $\widetilde{G} = \{(1/\mathrm{LC}_\sigma(g))\,g \mid g \in G\}$ *is an* $\mathscr{O}$*-border prebasis.*

d) Let γ *denote the smallest absolute value of the border term coefficient of one of the polynomials* g_i. *Then the set* $\widetilde{G}$ *is an* η*-approximate border basis for* $\eta = 2\delta + 2\nu\delta^2/\gamma\varepsilon + 2\nu\delta\sqrt{s}/\varepsilon$.

For a proof, see [14], Section 3. Let us add some remarks on the performance of this algorithm.

1. The AVI-Algorithm follows in principle the method of the Buchberger-Möller Algorithm for computing the exact vanishing ideal of $\mathbb{X}$. However, we are not processing one term at a time, but all terms of a given degree simultaneously, in order to filter out “almost relations” among the evaluation vectors using the SVD. Of course, if these sets of terms are too large, we can partition them into smaller chunks to speed up the SVD calculation.
2. The stated bounds for δ and η are rather crude. Using a more refined analysis, they could be improved significantly. In practical examples, the behavior of the computed approximate border bases is much better than predicted by these bounds.
3. By changing the construction of the list L in step **A2** appropriately, the AVI-Algorithm can be used to compute an “approximate Gröbner basis” of an approximate vanishing ideal of $\mathbb{X}$. More precisely, the list L should be defined as all terms in $\mathbb{T}^n_d$ which are not contained in $\langle \mathrm{LT}_\sigma(G) \rangle$. Unfortunately, there is

no guarantee that the computed polynomials are close to an actual Gröbner basis. The computation of the normal remainders of the S-polynomials requires a number of reductions steps which can be very large. Therefore no bound for the size of the evaluation vectors of these normal remainders can be given. In many practical examples, however, the Gröbner basis version works fine.

4. The AVI-Algorithm can also be combined with a threshold control in order to obtain a smoother evaluation behaviour of the computed border prebasis. Details can be found in [14], Section 3.

What is an approximate $\mathcal{O}$-border basis good for? In the next subsection we shall see an actual industrial application. Moreover, given a further order ideal $\mathcal{O}'$ of the same cardinality, we can compute an approximate $\mathcal{O}'$-border basis using the technique of [18], Prop. 5. (In general, this will come at the expense of a partial loss of the quality of approximation.) Finally, we can compute a "close-by" exact $\mathcal{O}$-border basis having coefficients in $\mathbb{Q}$ via the *rational recovery* technique in [19], and this exact border basis can be used as input for "higher" algebraic operations such as the computation of syzygy modules.

1.4.D. An Application of the AVI Algorithm. Let us now return to Problems 1 and 2 discussed in Section 1. Our idea is to construct the desired polynomial $f \in P = \mathbb{R}[x_1,\dots,x_5]$ using the AVI algorithm 1.4.8. Finding f means *explaining* the production as a linear combination of linearly independent "data series" which, in turn, depend on the evaluations of the indeterminates x_i. This implies that linear dependencies among those input combinations have to be removed first, *i.e.* that we have to pass to a suitable quotient modulo certain relations among the indeterminates. In the context of the AVI algorithm, where we are dealing with large uncertainties in the set of points $\mathbb{X}$, we need to consider *approximate* relations among the variables.

In summary, we are specifically looking for a polynomial $f \in P = \mathbb{R}[x_1,\dots,x_5]$ of the form

$$f \;=\; \sum_{i=1}^{\mu} c_i t_i \,+\, g$$

where $g \in P$ is contained in an ε-approximate vanishing ideal of $\mathbb{X}$, where we have $c_i \in \mathbb{R}$, and where $\mathcal{O} = \{t_1,\dots,t_\mu\}$ is an order ideal of monomials whose evaluation vectors at the points of $\mathbb{X}$ are almost linearly independent. The evaluation vector of f should represent the production data used for the modelling experiment.

Why do we expect that such a representation exists? Observe that the order ideal $\mathcal{O} = \{t_1,\dots,t_\mu\}$ is the one calculated by the AVI algorithm. Its evaluation vectors $\{\mathrm{eval}_{\mathbb{X}}(t_1),\dots,\mathrm{eval}_{\mathbb{X}}(t_\mu)\}$ span approximately the vector space of all possible evaluation vectors of terms at $\mathbb{X}$. Moreover, this agrees with the assumption that we tried to motivate in Section 1. Our method to compute f is to take its evaluation vector $\mathrm{eval}_{\mathbb{X}}(f)$, the measured production, and to project it to the linear span of the evaluation vectors $\mathrm{eval}_{\mathbb{X}}(t_i)$.

The results of trying this method using actual industrial data are shown in the following figure. The values of the physical quantities associated to $x_1,\dots,x_5$ were available at 7400 time stamps. The first 6000 data points were used for the modelling experiment, and the computed polynomial f was evaluated at the remaining 1400 data points in the validation phase. The physical interpretation of the indeterminates is according to Table 1.1.

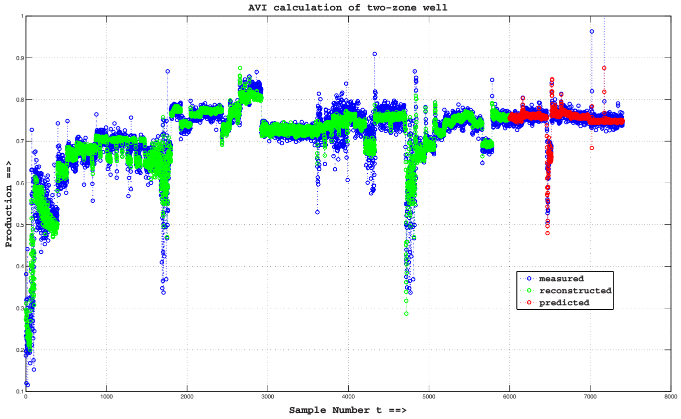


Fig. 1.6 Result of an AVI application.

Figure 1.6 shows that our model f does an excellent job: the comparison of the predicted values for the production with the measured values shows that the model passes the validation test unambiguously. The spikes shown in the figure result from instrumentation sensor failures.

Which choices and *a priori* information went into this computation? The term ordering we used is `DegRevLex`. The significance of this choice will be discussed below. For the threshold value ε, we used $\varepsilon = 0.1$. A suitable value for ε cannot be inferred from inspecting the measured data. As a rule-of-thumb, we choose it according to the size of the relative measurement errors, but we do not know a mathematical argument to determine a judicious choice of this parameter. In more intuitive terms, the value for ε is related to the level of detail we are looking at the physical problem we are investigating. In loose terms, when choosing relatively large values for ε, we are only following the large trends in the data, whereas when choosing smaller values for ε, we are zooming in on the “local” variations in the data.

Now we address one of the most remarkable features of the AVI algorithm, namely that it extracts structural information from numerical, measured data. This is unlike virtually any other method which is used in applications where a model structure in whatever form has to be provided as *input* for the algorithm which is used. Using the AVI algorithm, the model structure is *output*. Specifying a model structure up front means that a prescription is imposed on the data how the physical system under investigation works. But specifically in the case of an oil reservoir one cannot know how it works. To emphasize this crucial point we have summarized this unique feature of the AVI algorithm in Figure 1.7.

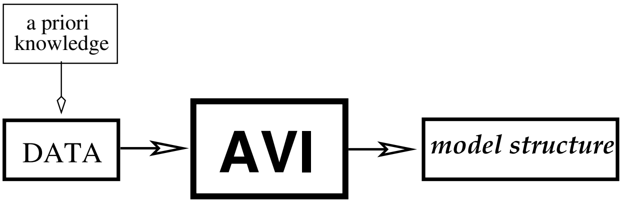


Fig. 1.7 Model construction using the AVI algorithm.

The motivation we have given in Section 1 for our problem statements indicates that a good numerical approximation and prediction of production values is not enough to deal completely with the production problem. In itself, the computed model does not give information about the all-determining interactions occurring in a production unit. For that we need to inspect the structure of the model for the production in terms of the driving inputs. In other words, we have to study the structure of the polynomial f. A first representation is

$$\begin{aligned} f = & -1.97x_3^2 - 0.18x_1x_4 - 0.30x_2x_4 + 2.37x_3x_4 - 0.16x_4^2 - 0.36x_1x_5 \\ & +0.40x_2x_5 - 3.03x_3x_5 - 1.19x_4x_5 + 0.32x_5^2 + 0.34x_1 - 0.09x_2 \\ & +4.03x_3 + 0.94x_4 + 0.68x_5 - 0.36 \end{aligned}$$

Already at first glance we notice the dominant presence of x_5. As given in Table 1.1, this indeterminate is related to the transport of the fluids through the tubing at the surface. This comes rather unexpected, indeed almost as an unpleasant surprise. For we have stated repeatedly the importance of the sub-surface for this production problem, and also that the notorious interactions are taking place in the neighbourhood of the inflow from the reservoir into the production tubing of the well. But now it seems an indeterminate related to the surface, far away from the reservoir, is a key element in all this? Well, this is the answer we instructed the algorithm to find! Recall that the chosen term ordering is `DegRevLex`. Hence x_5 is the indeterminate which is most unlikely to be a leading term of one of the polynomials in the approximate border basis. In other words, it is the indeterminate which is most likely to occur in many of the terms of $\mathcal{O}$, and our method amounts to the attempt to explain the data predominantly in terms of x_5.

Rather than continuing the attempt to reveal the significance of the above structure of f, we should therefore hasten to repair our “physical inconsistency”. To do this, we have two options: we can either switch to a different term ordering or we can change the physical interpretation of the indeterminates. To ease the comparison of the two models we get, we opt for the second method. The following physical interpretation of the indeterminates acknowledges the “physical hierarchy” of the system. We consider the polynomial ring $\mathbb{R}[y_1,\dots,y_5]$ and let

y_1	: $\Delta P_{transport}$
y_2	: ΔP_{tub}
y_3	: *Gas production*
y_4	: ΔP_{inflow_2}
y_5	: ΔP_{inflow_1}

Table 1.2 New physical interpretation of the indeterminates.

For the role played by these indeterminates in the two-zone well we refer to Figure 1.1. We repeat the calculation using the AVI algorithm with $\varepsilon = 0.1$ and term ordering `DegRevLex`. The result is a polynomial g of the form

$$\begin{aligned} g = & -5.35y_3y_5^2 - 0.73y_4y_5^2 - 0.21y_5^3 + 2.37y_2y_3 - 7.32y_3^2 - 0.88y_1y_4 - 0.15y_2y_4 \\ & +0.34y_3y_4 - 0.55y_4^2 - 2.20y_1y_5 - 0.35y_2y_5 + 3.85y_3y_5 + 0.67y_4y_5 + 0.61y_5^2 \\ & +0.62y_1 - 0.26y_2 + 2.69y_3 + 0.98y_4 + 1.63y_5 - 0.12 \end{aligned}$$

To judge the quality of this new model, we consider the differences of the evaluations of f and g at the points of $\mathbb{X}$. We obtain the following Figure 1.8 which shows that $f-g$ vanishes approximately at $\mathbb{X}$, apart from some spikes caused by faults in the data due to instrumentation failures. Thus, from the numerical point of view, the polynomial g is as good a model of the system as f.

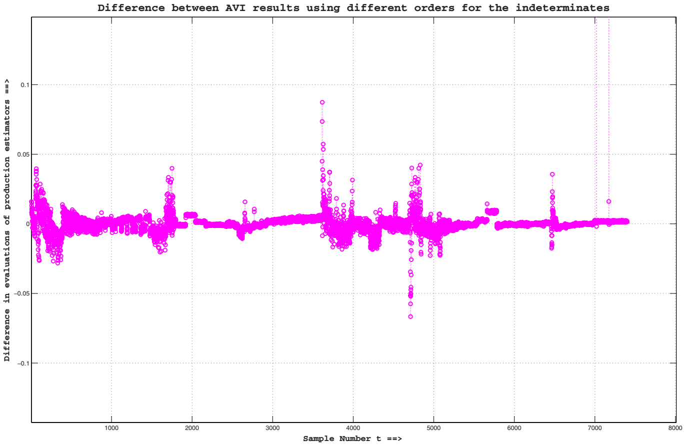


Fig. 1.8 Differences of the evaluations of two models.

Notice that also in g, the "last" indeterminate y_5 plays a dominant role. However, this time there is no physical inconsistency associated with this fact. Quite to the contrary, the terms in the support of the model, and in particular the factors we put in parenthesis, have physical interpretations revealing the flow mechanisms inside the well. Although a detailed discussion of these interpretations would exceed the scope of this paper, it should be mentioned here that our findings have been assessed positively in discussions with production engineers and have been confirmed by dedicated field experiments.

There is, however, one aspect in this vein which warrants to be mentioned here. Recall the brief discussion in Section 1.B of the commonly accepted procedure to express the total production as a linear combination of the separate productions. The terms y_4y_5 and $y_4y_5^2$ in the above polynomial g indicate that g cannot be written as a linear combination of the separate productions which correspond to y_4 and y_5. Clearly, the inappropriate decomposition of the total production resulting from the traditional procedures may have substantial consequences for the production strategy used in the exploitation of the reservoir.

To wrap up the discussion, we note that the information about the two-zone well captured in the data set $\mathbb{X}$ has been coded in merely *20* functions, namely the terms in $\mathscr{O}$. Using suitable linear combinations of these terms, we can find excellent *estimators* of the oil production of the two-zone well under different production conditions. It should be stressed that such a physical interpretation can usually not be given for linear combinations of terms contained in the expression of g, nor to the individual monomials for that matter; in particular, their evaluations over $\mathbb{X}$ may be negative or exceed physically meaningful bounds. In this sense, the terms in $\mathscr{O}$ should be considered as purely *mathematical* states of the system into which the production of the well can be decomposed. The structure of this decomposition reveals deep insights into the production system which are only available via the described modelling procedure based on the AVI algorithm.

1.5 Stable Order Ideals

There is nothing so stable as change.
(Bob Dylan)

1.5.A. The SOI Algorithm. In this subsection we consider the following setting. Let $\mathbb{X} \subset \mathbb{R}^n$ be a finite set of points whose coordinates are known only with limited precision, and let

$$\mathscr{I}(\mathbb{X}) = \{f \in \mathbb{R}[x_1,\dots,x_n] \mid f(p) = 0 \text{ for all } p \in \mathbb{X}\}$$

be its vanishing ideal. Our goal is to compare the different residue class rings $P/\mathscr{I}(\widetilde{\mathbb{X}})$ where $P = \mathbb{R}[x_1,\dots,x_n]$ and $\widetilde{\mathbb{X}}$ is an admissible perturbation of $\mathbb{X}$, *i.e.* a set made up of points differing by less than the data uncertainty from the corresponding points of $\mathbb{X}$.

Given two distinct admissible perturbations $\widetilde{\mathbb{X}}_1$ and $\widetilde{\mathbb{X}}_2$ of $\mathbb{X}$, it can happen that their affine coordinate rings $P/\mathscr{I}(\widetilde{\mathbb{X}}_1)$ and $P/\mathscr{I}(\widetilde{\mathbb{X}}_2)$ as well as their vanishing ideals $\mathscr{I}(\widetilde{\mathbb{X}}_1)$ and $\mathscr{I}(\widetilde{\mathbb{X}}_2)$ have very different bases – this is a well known phenomenon in Gröbner basis theory. When dealing with a set $\mathbb{X}$ of empirical points, a notion of "numerically stable" basis of the quotient ring $P/\mathscr{I}(\mathbb{X})$ is necessary. A basis $\mathscr{O} \subseteq \mathbb{T}^n$ is stable if its residue classes form a vector space basis of $P/\mathscr{I}(\widetilde{\mathbb{X}})$ for any admissible perturbation $\widetilde{\mathbb{X}}$ of the empirical set $\mathbb{X}$. Furthermore, a stable order ideal $\mathscr{O}$ provides a common characterization of the ideals $\mathscr{I}(\mathbb{X})$ and $\mathscr{I}(\widetilde{\mathbb{X}})$ by means of their $\mathscr{O}$-border bases.

One way of dealing with the negative effects of data uncertainty is to replace elements of $\mathbb{X}$ which differ from each other by less than the data accuracy with a single representative point. This "preprocessing", using for instance the algorithms described in [1], may reduce the computational complexity but also loose information contained in the data. In general, it is not sufficient to eliminate the instabilities of exact bases of $P/\mathscr{I}(\mathbb{X})$. However, if we are given a finite set $\mathbb{X}$ of s well-separated empirical points, we can use the **S**table **O**rder **I**deal (SOI) Algorithm presented in this subsection. It computes a stable order ideal $\mathscr{O}$, and if $\mathscr{O}$ contains enough elements to form a basis of $P/\mathscr{I}(\mathbb{X})$, the corresponding stable border basis is also computed.

The following definition formalizes some concepts defined "empirically" in [30].

Definition 1.5.1. Let $p = (c_1, \dots, c_n)$ be a point in $\mathbb{R}^n$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n) \in (\mathbb{R}_+)^n$.

a) The pair (p, ε) is called an **empirical point** in $\mathbb{R}^n$. We shall denote it also by p^ε. The point p is called the **specific value** and ε is called the **tolerance** of p^ε.

b) A point $\widetilde{p} = (\widetilde{c}_1, \dots, \widetilde{c}_n) \in \mathbb{R}^n$ is called an **admissible perturbation** of p if

$$\| ((\widetilde{c}_1 - c_1)/\varepsilon_1, \dots, (\widetilde{c}_n - c_n)/\varepsilon_n) \| \leq 1$$

c) Let $\mathbb{X}^\varepsilon = \{p_1^\varepsilon, \dots, p_s^\varepsilon\}$ be a set of empirical points which share the same tolerance ε, and let $\mathbb{X} = \{p_1, \dots, p_s\}$ be its specific value. A set of points $\widetilde{\mathbb{X}} = \{\tilde{p}_1, \dots, \tilde{p}_s\}$ is called an **admissible perturbation** of $\mathbb{X}$ if each point $\tilde{p}_i$ is an admissible perturbation of p_i.

d) Let a set $\mathbb{X}^\varepsilon = \{p_1^\varepsilon, \dots, p_s^\varepsilon\}$ of empirical points be given with specific values $p_i = (c_{i1}, \dots, c_{in})$. We introduce ns **error indeterminates**

$$\mathbf{e} = (e_{11}, \dots, e_{s1}, e_{12}, \dots, e_{s2}, \dots, e_{1n}, \dots, e_{sn})$$

Then the set $\widetilde{\mathbb{X}}(\mathbf{e}) = \{\widehat{p}_1, \dots, \widehat{p}_s\}$ where $\widehat{p}_k = (c_{k1} + e_{k1}, \dots, c_{kn} + e_{kn})$ is called the **generic perturbation** of $\mathbb{X}$.

Obviously, an admissible perturbation of $\mathbb{X}$ is obtained from the generic perturbation by substituting values $\tilde{e}_{ij}$ for the error indeterminates such that we have $\|(\tilde{e}_{i1}/\varepsilon_1, \dots, \tilde{e}_{in}/\varepsilon_n)\| \leq 1$.

Next we define the notion of stability for order ideals.

Definition 1.5.2. Let $\mathscr{O}$ be an order ideal of $\mathbb{T}^n$. The set $\mathscr{O}$ is called **stable** w.r.t. $\mathbb{X}^\varepsilon$ if the evaluation matrix $\mathrm{eval}_{\widetilde{\mathbb{X}}}(\mathscr{O})$ has full rank for each admissible perturbation $\widetilde{\mathbb{X}}$ of $\mathbb{X}$. Furthermore, if $\mathscr{O}$ is also a basis of $P/\mathscr{I}(\mathbb{X})$, it is called a **stable quotient basis** of $\mathscr{I}(\mathbb{X})$.

Given any finite set of points $\mathbb{X}$, any quotient basis $\mathscr{O}$ for $\mathscr{I}(\mathbb{X})$ is stable w.r.t. $\mathbb{X}^\delta$ for a sufficiently small value of the tolerance δ. This is equivalent to saying that $\mathscr{O}$ has a "region of stability" w.r.t. $\mathbb{X}$ and follows from the next theorem.

Theorem 1.5.3. *Let $\mathbb{X}$ be a finite set of points in $\mathbb{R}^n$ and $\mathscr{O}$ a quotient basis for $\mathscr{I}(\mathbb{X})$. Then there exists a tolerance $\delta = (\delta_1, \dots, \delta_n)$, with $\delta_i > 0$, such that $\mathscr{O}$ is stable w.r.t. $\mathbb{X}^\delta$.*

Proof. Let $\mathrm{eval}_{\mathbb{X}}(\mathscr{O})$ be the evaluation matrix of $\mathscr{O}$ at the points of $\mathbb{X}$. Its entries depend continuously on the points in $\mathbb{X}$. By hypothesis, the set $\mathscr{O}$ is a quotient basis for $\mathscr{I}(\mathbb{X})$. It follows that $\mathrm{eval}_{\mathbb{X}}(\mathscr{O})$ is invertible. Recalling that the determinant is a polynomial function in the matrix entries and noting that the entries of $\mathrm{eval}_{\mathbb{X}}(\mathscr{O})$ are polynomials in the points' coordinates, we can conclude that there exists a tolerance $\delta = (\delta_1, \dots, \delta_n) \in (\mathbb{R}_+)^n$ such that $\det(\mathrm{eval}_{\widetilde{\mathbb{X}}}(\mathscr{O})) \neq 0$ for any perturbation $\widetilde{\mathbb{X}}$ of $\mathbb{X}$ in $\mathbb{X}^\delta$. □

Nevertheless, since the tolerance ε of the empirical points in $\mathbb{X}^\varepsilon$ is given *a priori* by the measurements, Theorem 1.5.3 does not provide a direct answer to the problem of stability. If the given tolerance ε for the points in $\mathbb{X}$ allows us to leave the "region of stability" of a chosen quotient basis $\mathscr{O}$, then $\mathscr{O}$ will not be stable w.r.t. $\mathbb{X}^\varepsilon$. Such a situation is shown in the following example.

Example 1.5.4. Let $\mathbb{X} = \{(0,2),\ (1,2.1),\ (2,1.9)\} \subseteq \mathbb{R}^2$ be a set of specified values. The order ideal $\mathscr{O} = \{1, y, y^2\}$ is a basis of $P/\mathscr{I}(\mathbb{X})$. Given the generic perturbation

$$\widetilde{\mathbb{X}}(\mathbf{e}) = \{(0+e_{11}, 2+e_{12}),\ (1+e_{21}, 2.1+e_{22}),\ (2+e_{31}, 1.9+e_{32})\}$$

the evaluation matrix of $\mathscr{O}$ at $\widetilde{\mathbb{X}}(\mathbf{e})$ is the Vandermonde matrix

$$\mathrm{eval}_{\widetilde{\mathbb{X}}(\mathbf{e})}(\mathscr{O}) = \begin{pmatrix} 1 & 2+e_{12} & (2+e_{12})^2 \\ 1 & 2.1+e_{22} & (2.1+e_{22})^2 \\ 1 & 1.9+e_{32} & (1.9+e_{32})^2 \end{pmatrix}$$

Since the matrix $\mathrm{eval}_{\widetilde{\mathbb{X}}}(\mathscr{O})$ is invertible if and only if the values $2+\tilde{e}_{12}$, $2.1+\tilde{e}_{22}$ and $1.9+\tilde{e}_{32}$ are pairwise distinct, we have that $\mathscr{O}$ is stable w.r.t. $\mathbb{X}^\varepsilon$ if the tolerance $\varepsilon = (\varepsilon_1, \varepsilon_2)$ satisfies $|\varepsilon_2| < 0.1$. *Vice versa,* if we consider $\mathbb{X}^{(\delta_1,\delta_2)}$, where $\delta_2 > 0.1$, there exists the admissible perturbation $\widetilde{\mathbb{X}} = \{(0,2),\ (1,2),\ (2,2)\}$ whose evaluation matrix $\mathrm{eval}_{\widetilde{\mathbb{X}}}(\mathscr{O})$ is singular. So, the order ideal $\mathscr{O}$ is not stable w.r.t. $\mathbb{X}^{(\delta_1,\delta_2)}$ since its "region of stability" is too small w.r.t. the given tolerance δ.

Intuitively, a border basis G of the vanishing ideal $\mathscr{I}(\mathbb{X})$ is considered to be structurally stable if, for each admissible perturbation $\widetilde{\mathbb{X}}$ of $\mathbb{X}$, it is possible to produce a border basis $\widetilde{G}$ of $\mathscr{I}(\widetilde{\mathbb{X}})$ only by means of a slight and continuous variation of the coefficients of the polynomials of G. This situation arises when G and $\widetilde{G}$ are founded on the same stable quotient basis $\mathscr{O}$, as shown in the following theorem (for a proof see [1]).

Theorem 1.5.5. *Let $\mathbb{X}^\varepsilon$ be a set of s distinct empirical points and $\mathscr{O} = \{t_1,\dots,t_s\}$ a quotient basis for $\mathscr{I}(\mathbb{X})$ which is stable w.r.t. $\mathbb{X}^\varepsilon$. Then, for each admissible perturbation $\widetilde{\mathbb{X}}$ of $\mathbb{X}^\varepsilon$, the vanishing ideal $\mathscr{I}(\widetilde{\mathbb{X}})$ has an $\mathscr{O}$-border basis $\widetilde{G}$. Furthermore, if $\partial\mathscr{O} = \{b_1,\dots,b_\nu\}$ is the border of $\mathscr{O}$ then $\widetilde{G}$ consists of ν polynomials of the form*

$$g_j = b_j - \sum_{i=1}^{s} \alpha_{ij} t_i \qquad \textit{for } j \in \{1,\dots,\nu\}$$

where the coefficients $a_{ij} \in \mathbb{R}$ satisfy the linear systems

$$\mathrm{eval}_{\widetilde{\mathbb{X}}}(b_j) = \sum_{i=1}^{s} \alpha_{ij}\, \mathrm{eval}_{\widetilde{\mathbb{X}}}(t_i)$$

Note that the coefficients α_{ij} of each polynomial $g_j \in \widetilde{G}$ are just the components of the solution α_j of the linear system $\mathrm{eval}_{\widetilde{\mathbb{X}}}(\mathscr{O})\, \alpha_j = \mathrm{eval}_{\widetilde{\mathbb{X}}}(b_j)$. It follows that the coefficients α_{ij} are continuous functions of the coordinates of the points of $\widetilde{\mathbb{X}}$. Since the order ideal $\mathscr{O}$ is stable w.r.t. $\mathbb{X}^\varepsilon$, they undergo only continuous variations as $\widetilde{\mathbb{X}}$ changes. Now the definition of stable border bases follows naturally.

Definition 1.5.6. Let $\mathbb{X}^\varepsilon$ be a finite set of distinct empirical points, and let $\mathscr{O}$ be a quotient basis for the vanishing ideal $\mathscr{I}(\mathbb{X})$. If $\mathscr{O}$ is stable w.r.t. $\mathbb{X}^\varepsilon$ then the $\mathscr{O}$-border basis G of $\mathscr{I}(\mathbb{X})$ is said to be **stable** w.r.t. the set $\mathbb{X}^\varepsilon$.

Given $\mathbb{X}$ and a stable quotient basis $\mathscr{O}$, it is possible to obtain a stable $\mathscr{O}$-border basis of $\mathscr{I}(\mathbb{X})$ by simple linear algebra computations. The SOI Algorithm addresses the problem of finding a stable quotient basis as follows. As in the Buchberger-Möller algorithm [6], the order ideal $\mathscr{O}$ is built stepwise: initially $\mathscr{O}$ comprises just the term 1; then at each iteration, a new term t is considered. If the evaluation matrix $\mathrm{eval}_{\widetilde{\mathbb{X}}}(\mathscr{O} \cup \{t\})$ has full rank for all admissible perturbations $\widetilde{\mathbb{X}}$ then t is added to $\mathscr{O}$; otherwise t is added to the corner set of the order ideal.

The rank condition is equivalent to checking whether $\rho(\widetilde{\mathbb{X}})$, the component of the evaluation vector $\mathrm{eval}_{\widetilde{\mathbb{X}}}(t)$ orthogonal to the column space of the matrix $\mathrm{eval}_{\widetilde{\mathbb{X}}}(\mathscr{O})$, vanishes for any admissible $\widetilde{\mathbb{X}}$. In the following theorem this check is greatly simplified by restricting it to first order error terms, as our interest is essentially focused on small perturbations $\widetilde{\mathbb{X}}$ of $\mathbb{X}$. In practice, the SOI algorithm solves an underdetermined system to test whether the first order approximation of $\rho(\widetilde{\mathbb{X}})$ vanishes for some admissible set $\widetilde{\mathbb{X}}$.

Theorem 1.5.7. (The Stable Order Ideal Algorithm (SOI))
Let $\mathbb{X}^{\varepsilon} = \{p_1^{\varepsilon}, \dots, p_s^{\varepsilon}\}$ be a finite set of well-separated empirical points having a common tolerance $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$. Let σ be a term ordering on $\mathbb{T}^n$ and $\gamma \geq 0$. Consider the following sequence of instructions.

S1 Start with the lists $\mathscr{O} = [1]$, $L = [x_1, \dots, x_n]$, the empty list $C = [\,]$, the matrix $M_0 \in \mathrm{Mat}_{s \times 1}(\mathbb{R})$ with all entries equal to 1, and $M_1 \in \mathrm{Mat}_{s \times 1}(\mathbb{R})$ with all entries equal to 0.

S2 If $L = [\,]$ return the set $\mathscr{O}$ and stop. Otherwise, let $t = \min_{\sigma}(L)$ and delete it from L.

S3 Let v_0 and v_1 be the homogeneous components of degrees 0 and 1 of the evaluation vector $v = \mathrm{eval}_{\widetilde{\mathbb{X}}(\mathbf{e})}(t)$. Compute the vectors

$$\begin{aligned} \rho_0 &= v_0 - M_0 \alpha_0 \\ \rho_1 &= v_1 - M_0 \alpha_1 - M_1 \alpha_0 \end{aligned}$$

where

$$\begin{aligned} \alpha_0 &= (M_0^{\mathrm{tr}} M_0)^{-1} M_0^{\mathrm{tr}} v_0 \\ \alpha_1 &= (M_0^{\mathrm{tr}} M_0)^{-1} (M_0^{\mathrm{tr}} v_1 + M_1^{\mathrm{tr}} v_0 - M_0^{\mathrm{tr}} M_1 \alpha_0 - M_1^{\mathrm{tr}} M_0 \alpha_0). \end{aligned}$$

S4 Let $C_t \in \mathrm{Mat}_{s,sn}(\mathbb{R})$ be such that $\rho_1 = C_t \mathbf{e}$. Let k be the maximum integer such that the matrix $\widehat{C}_t$, formed by selecting the first k rows of C_t, has minimum singular value $\widehat{\sigma}_k$ greater than $\|\varepsilon\|$. Let $\widehat{\rho}_0$ be the vector comprising the first k elements of ρ_0, and let $\widehat{C}_t^{\dagger}$ be the pseudo-inverse of $\widehat{C}_t$. Compute $\widehat{\delta} = -\widehat{C}_t^{\dagger} \widehat{\rho}_0$, which is the minimal 2-norm solution of the underdetermined system $\widehat{C}_t \widehat{\delta} = -\widehat{\rho}_0$.

S5 If $\|\hat{\delta}\| > (1+\gamma)\sqrt{s}\|\varepsilon\|$, then adjoin the vector v_0 as a new column of M_0 and the vector v_1 as a new column of M_1. Append the power product t to $\mathscr{O}$, and add to L those elements of $\{x_1 t, \dots, x_n t\}$ which are not multiples of an element of L or C. Continue with step S2.

S6 Otherwise append t to the list C, and remove from L all multiples of t. Continue with step S2.

This is an algorithm which returns an order ideal $\mathscr{O} \subset \mathbb{T}^n$. If for every admissible perturbation $\widetilde{\mathbb{X}}$ the value γ satisfies $\|\rho_{2+}(\widetilde{\mathbb{X}})\| \leq \gamma\sqrt{s}\|\varepsilon\|^2$, where $\rho_{2+}(\widetilde{\mathbb{X}})$ is the evaluation at $\widetilde{\mathbb{X}}$ of the component of $\rho(\widetilde{\mathbb{X}}(\mathbf{e}))$ of degree greater than 1, then $\mathscr{O}$ is an order ideal which is stable w.r.t. the empirical set $\mathbb{X}^{\varepsilon}$. In particular, when $\#\mathscr{O} = s$, the ideal $\mathscr{I}(\mathbb{X})$ has a corresponding stable border basis w.r.t. $\mathbb{X}^{\varepsilon}$.

To implement the SOI Algorithm a value of γ has to be chosen even if an estimate of $\|\rho_{2+}(\widetilde{\mathbb{X}})\|$ is unknown. Since we consider small perturbations $\widetilde{\mathbb{X}}$ of the set $\mathbb{X}$, in most cases $\rho_0 + \rho_1(\widetilde{\mathbb{X}})$ is a good linear approximation of $\rho(\widetilde{\mathbb{X}})$. For this reason $\|\rho_{2+}(\widetilde{\mathbb{X}})\|$ is small and a value of $\gamma \ll 1$ can be chosen to obtain a set $\mathscr{O}$ which is stable w.r.t. $\mathbb{X}^{\varepsilon}$. On the other hand, if ρ is not well approximated by its homogeneous components of degrees 0 and 1 the strategy of the SOI algorithm loses its meaning, since it is based on a first order analysis.

1.5.B. Comparison of the SOI and AVI Algorithms. Now we present some numerical examples to show the effectiveness of the SOI and AVI algorithms. The first two examples show how the algorithms detect simple geometrical configurations almost satisfied by the given set $\mathbb{X}$.

Example 1.5.8. (Four Almost Aligned Points)
Let $\mathbb{X}^{\varepsilon}$ be a set of empirical points with the specified values

$$\mathbb{X} = \{(0, 0.01),\ (0.34, 0.32),\ (0.65, 0.68),\ (0.99, 1)\} \subseteq \mathbb{R}^2$$

and the tolerance $\varepsilon = (0.03, 0.03)$.

a) The SOI algorithm computes the quotient basis $\mathcal{O} = \{1, y, y^2, y^3\}$ which is stable w.r.t. $\mathbb{X}^{\varepsilon}$. Hence we can compute the stable border basis G founded on it and get

$$G = \begin{cases} x \;-\; 0.654y^3 + 1.013y^2 - 1.362y + 0.014 \\ xy \;-\; 0.303y^3 - 0.552y^2 - 0.137y + 0.001 \\ xy^2 \;-\; 1.16y^3 + 0.238y^2 - 0.068y + 0.001 \\ xy^3 \;-\; 2.094y^3 + 1.368y^2 - 0.266y + 0.002 \\ y^4 \;-\; 2.01y^3 - 1.238y^2 - 0.23y + 0.002 \end{cases}$$

The algorithm also yields the almost vanishing polynomial $f = x - 0.984y$. This polynomial highlights the fact that $\mathbb{X}$ contains "almost aligned" points. Since the quotient basis $\mathcal{O}$ is stable w.r.t. $\mathbb{X}^{\varepsilon}$, we can conclude that there exists a small perturbation $\widetilde{\mathbb{X}}$ of $\mathbb{X}$ containing aligned points and for which the associated evaluation matrix $\mathrm{eval}_{\widetilde{\mathbb{X}}}(\mathcal{O})$ is invertible. Notice that this fact is not easily discernible from the computed border basis.
A further interesting polynomial is obtained by taking the difference of f and the first border basis polynomial. The resulting polynomial $h = 0.654y^3 - 1.013y^2 + 0.378y - 0.014$ has small values at the points of $\mathbb{X}$. This is not a contradiction to the "almost linear independence" of $\mathcal{O}$ in the sense of [2], since there is no admissible perturbation of $\mathbb{X}$ for which h vanishes. The correct interpretation is that there is an almost $\mathcal{O}$-border prebasis close to the computed $\mathcal{O}$-border basis which is not an approximate $\mathcal{O}$-border basis.

b) A completely different result is obtained by applying the Buchberger-Möller algorithm to the set $\mathbb{X}$. We use the same term ordering σ and obtain the following σ-Gröbner basis H of $\mathscr{I}(\mathbb{X})$:

$$\begin{aligned} &x^2 \;-\; 5525/5324y^2 - 30456/33275x + 103649/106480y - 6409/665500 \\ &xy \;-\; 1358/1331y^2 - 15391/33275x + 32811/66550y - 8033/1663750 \\ &y^3 \;-\; 205967/133100y^2 - 1271124/831875x + 1384811/665500y \\ &\quad - 429556/20796875 \end{aligned}$$

The associated quotient basis is $\mathcal{O}_\sigma(\mathcal{I}(\mathbb{X})) = \mathbb{T}^2 \setminus \mathrm{LT}_\sigma\{\mathcal{I}(\mathbb{X})\} = \{1, y, x, y^2\}$. We observe that $\mathcal{O}_\sigma(\mathcal{I}(\mathbb{X}))$ is not stable because the matrix $\mathrm{eval}_{\widetilde{\mathbb{X}}}(\mathcal{O}_\sigma(\mathcal{I}(\mathbb{X})))$ is singular for some admissible perturbations of $\mathbb{X}$. In particular, the information that the points of $\mathbb{X}$ are "almost aligned" is not at all evident from H.

c) Finally, we apply the AVI algorithm to $\mathbb{X}$. If we use $\varepsilon = 0.05$, we obtain the quotient basis $\mathcal{O} = \{1, y, y^2\}$ and as the approximate $\mathcal{O}$-border basis

$$\begin{aligned} x &- 0.984y \\ xy &- 1.013y^2 + 0.03y - 0.004 \\ xy^2 &- 1.568y^2 + 0.614y - 0.026 \\ y^3 &- 1.556y^2 + 0.588y - 0.023 \end{aligned}$$

Notice that the first and last polynomial generate the vanishing ideal of the set of *three* points $\mathbb{X}' = \{(0.044, 0.044), (0.522, 0.529), (0.967, 0.982)\}$. Thus the "almost alignment" of $\mathbb{X}$ was correctly detected, and the algorithm found a cubic curve passing close to all four points.

If instead we apply the AVI algorithm with $\varepsilon = 0.03$, which is approximately the size of the data inaccuracies inherent in $\mathbb{X}$, we get $\mathcal{O} = \{1, y, y^2, y^3\}$ and the approximate border basis

$$\begin{aligned} x &- 0.984y \\ xy &- -1.013y^2 + 0.03y - 0.004 \\ xy^2 &- 1.16y^3 + 0.237y^2 - 0.068y \\ xy^3 &- 2.094y^3 + 1.367y^2 - 0.265y + 0.002 \\ y^4 &- 2.01y^3 + 1.237y^2 - 0.229y + 0.002 \end{aligned}$$

Here the ideal generated by the first and last polynomial corresponds to four perfectly aligned points very close to the points of $\mathbb{X}$.

In the following example we show the behavior of the SOI and AVI algorithms when applied to two sets of points with similar geometrical configuration but with different cardinality.

Example 1.5.9. (Points Close to a Circle)
Let $\mathbb{X}_8$ and $\mathbb{X}_{16}$ be sets of points created by perturbing slightly the coordinates of 8 and 16 points lying on the unit circle $x^2 + y^2 - 1 = 0$.

a) First we apply the SOI algorithm with tolerance $\varepsilon = (0.01, 0.01)$. The following table summarizes the results. The first two columns contain the name of the processed set and the value of its cardinality. The column labeled with "Corners" refers to the set of corners of the stable order ideal computed by the algorithm.

Note that the sets of corners of the stable quotient bases computed by the SOI algorithm always contain the power product x^2. This means that there is a numerical linear dependence among the empirical vectors associated to the power products $\{1, y, x, y^2, xy, x^2\}$ and that some useful information on the geometrical configuration of the points could be found.

Input	$\#\mathbb{X}_i$	$\#\mathcal{O}$ (SOI)	Corners (SOI)	$\#\mathcal{O}$ (AVI)	Corners (AVI)
$\mathbb{X}_8$	8	8	$\{x^2,xy^3,y^5\}$	8	$\{x^2,xy^3,y^5\}$
$\mathbb{X}_{16}$	16	16	$\{x^2,xy^7,y^9\}$	11	$\{x^2,xy^5,y^6\}$

Table 1.3 SOI and AVI on sets of points close to a circle

If we enlarge the tolerances, however, already for $\varepsilon = (0.04, 0.04)$ the SOI algorithm finds no stable border basis for X_{16} anymore.

b) Now we apply the AVI algorithm. Since the points are near the unit circle, no normalization is necessary. We use for instance $\varepsilon = 0.06$.
For $\mathbb{X}_8$, we obtain the same order ideal $\mathcal{O} = \{1,x,y,xy,y^2,xy^2,y^3,y^4\}$ as the SOI algorithm, and an approximate $\mathcal{O}$-border basis containing $\{0.57x^2 + 0.57y^2 - 0.57, 0.89xy^3 + 0.01y^3 - 0.44xy - 0.01y, 0.53y^5 - 0.79y^3 + 0.26y\}$. This shows that the circle close to the points has been detected.
Using $\mathbb{X}_{16}$, we find the order ideal $\mathcal{O} = \{1,x,y,xy,y^2,xy^2,y^3,xy^3,y^4\}$ and an approximate border basis which contains $0.57x^2 + 0.58y^2 - 0.57$. Again the close-by unit circle has been detected, but there are also three sextics passing close to the original 16 points. Unlike with the SOI algorithm, we find an approximate vanishing ideal of a smaller number (namely 11 instead of 16) of points here.

Our next example shows that the term ordering σ used in the SOI Algorithm is only an implementation detail. In general, any strategy that chooses the power product t such that $\mathcal{O} \cup \{t\}$ is always an order ideal can be applied. The example illustrates the case where σ can lead to an $\mathcal{O}$-border basis which does not contain the τ-Gröbner basis of $\mathcal{I}(\mathbb{X})$ for any term ordering τ. Similarly, the AVI algorithm can be modified in such a way that the same property holds.

Example 1.5.10. (A Quotient Basis Not of Gröbner Type)
Let $\mathbb{X}^\varepsilon$ be a set of distinct empirical points having

$$\mathbb{X} = \{(1,1),\ (0.82,-1),\ (-0.82,0.82),\ (-1,-0.82)\}$$

as the set of specified values and $\varepsilon = (0.1, 0.1)$ as the tolerance.

a) Applying the SOI algorithm to $\mathbb{X}^\varepsilon$, we get the quotient basis $\mathcal{O} = \{1,x,y,xy\}$ which is stable with respect to $\mathbb{X}^\varepsilon$. Let τ be any term ordering on $\mathbb{T}^n$ and $\mathcal{O}_\tau(\mathcal{I}(\mathbb{X})) = \mathbb{T}^n \backslash \mathrm{LT}_\tau\{\mathcal{I}(\mathbb{X})\}$ the quotient basis associated to τ. We note that we have $\mathcal{O} \neq \mathcal{O}_\tau(\mathcal{I}(\mathbb{X}))$ here. In fact, according to τ, either $x^2 <_\tau xy$ or $y^2 <_\tau xy$. Furthermore, at least one of the two evaluation vectors $\mathrm{eval}_{\mathbb{X}}(x^2)$, $\mathrm{eval}_{\mathbb{X}}(y^2)$ is linearly independent of $\{\mathrm{eval}_{\mathbb{X}}(1),\ \mathrm{eval}_{\mathbb{X}}(x),\ \mathrm{eval}_{\mathbb{X}}(y)\}$ so that one of x^2 or y^2 must belong to $\mathcal{O}_\tau(\mathcal{I}(\mathbb{X}))$. We conclude that the $\mathcal{O}$-border basis of $\mathcal{I}(\mathbb{X})$ does not contain any Gröbner basis of $\mathcal{I}(\mathbb{X})$.

b) Next we start from the set $\mathbb{X}$ and use the AVI algorithm with $\varepsilon = 0.1$ and $\varepsilon' = 0.01$. The result is the order ideal $\mathcal{O} = \{1,x,y,xy\}$ and the (unitary) approximate border basis $G = \{0.76x^2 - 0.15xy - 0.62,\, 0.76y^2 - 0.13x - 0.63,\, 0.76x^2y - 0.12x - 0.62y,\, 0.76xy^2 - 0.63x - 0.11\}$.

Our final example illustrates that the order ideal $\mathscr{O}$ can have cardinality less than $s = \#\mathbb{X}$ both for the SOI and the AVI algorithm, but for different reasons. In the case of the SOI algorithm this happens when the tolerance on the points is, in some sense, too large. With a fixed set of specified values, the SOI algorithm may produce different results for different values of ε, some of which do not span all of $P/\mathscr{I}(\mathbb{X})$. For the AVI algorithm, the computed order ideal may satisfy $\#\mathscr{O} < s$ even for small ε. The reason is that the algorithm may detect many low degree polynomials vanishing ε-approximately at the given point, and those polynomials may generate a zero-dimensional ideal of lower codimension.

Example 1.5.11. (Five Points Close to Two Conics and a Cubic)
Let $\mathbb{X} = \{(0,1), (0.2,0.4), (0.28,0.28), (0.4,0.2), (1,0)\} \subset \mathbb{R}^2$.

a) First we apply the SOI algorithm to the set of well-separated empirical points $\mathbb{X}^\varepsilon$ with specified values $\mathbb{X}$ and tolerance $\varepsilon = (0.02, 0.02)$. We find the stable order ideal $\mathscr{O} = \{1, y, x, y^2\}$. However, this is not a quotient basis, so we cannot obtain the corresponding stable border basis. This is due to the fact that the points of $\mathbb{X}$ lie close to the hyperbola $xy + 0.17x + 0.14y - 0.17 = 0$, the ellipse $(x - 0.95)^2 + 0.87(y - 1)^2 - 0.9 = 0$ and the cubic defined by the equation $y^3 - 1.8y^2 + 0.23x - 1.03y - 0.23 = 0$. So, if the tolerance ε is too big, they "almost satisfy" all of them.
Observe how the problem does not arise if we use a smaller tolerance, *e.g.* $\delta = (0.01, 0.01)$. Applying SOI to $\mathbb{X}^\delta$, we obtain the stable quotient basis $\mathscr{O}' = \{1, y, x, xy, y^2\}$ and its corresponding border basis

$$G' = \begin{cases} x^2 + 3.83xy + y^2 - 1.23x - 1.23y + 0.23 \\ y^3 - 0.07xy - 1.8y^2 + 0.22x + 1.02y - 0.22 \\ xy^2 - 0.07xy + 0.2y^2 - 0.05x - 0.25y + 0.05 \\ x^2y - 0.84xy - 0.2y^2 + 0.19y \end{cases}$$

b) Next we use the AVI algorithm. Choosing $\varepsilon = 0.06$, we get $\mathscr{O} = \{1, x, y\}$ and the (unitary) approximate border basis

$$G = \begin{cases} 0.52x^2 - 0.77x - 0.25y + 0.25 \\ 0.94xy + 0.18x + 0.18y - 0.18 \\ 0.51y^2 - 0.26x - 0.77y + 0.26 \end{cases}$$

The same result is produced for any $0.06 \le \varepsilon \le 0.25$. The set G approximates a system of generators of the vanishing ideal $\mathscr{I}(\widetilde{\mathbb{X}})$ of the point set $\widetilde{\mathbb{X}} = \{(0, 0.98), (0.28, 0.29), (0.98, 0)\}$. Notice that $\widetilde{\mathbb{X}}$ is approximately contained in all three conics.
A smaller choice of ε, for instance $\varepsilon = 0.01$, leads to $\mathscr{O}' = \{1, x, y, y^2\}$ and

$$G' = \begin{cases} 0.3x^2 + 0.3y^2 - 0.6x - 0.6y + 0.3 \\ 0.94xy + 0.18x + 0.18y - 0.18 \\ 0.95xy^2 + 0.19y^2 - 0.03x - 0.22y + 0.03 \\ 0.42y^3 - 0.77y^2 + 0.1x + 0.44y - 0.1 \end{cases}$$

The set G' approximates a system of generators of $\mathscr{I}(\widetilde{\mathbb{X}}')$ for $\widetilde{\mathbb{X}}' = \{(0,0.99), (0.21,0.37), (0.37,0.21), (0.99,0)\}$. Thus even this small choice of ε leads to a decrease in the codimension of the corresponding $\mathscr{I}(\mathbb{X})$.

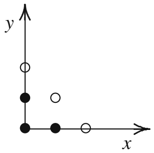
1.6 Border Basis and Gröbner Basis Schemes

Without geometry,
life is pointless.
(Sam Wormley)

Let $\mathscr{O} = \{t_1, \dots, t_\mu\}$ be an order ideal in $\mathbb{T}^n$. In this section we define a *moduli space*, called the border basis scheme, for *all* zero-dimensional ideals which have an $\mathscr{O}$-border basis. Then we define another space, called the Gröbner basis scheme, and explore their main properties, their connection to problems concerning approximate data, and their connection to Hilbert schemes of zero-dimensional schemes.

1.6.A. Two Basic Examples. Before starting with the technical details, we introduce two basic examples which will help us to understand the general theory.

Example 1.6.1. (Three Non-Collinear Points)
In this example we want to represent *all* zero-dimensional subschemes of $\mathbb{A}^2_K$ which share the property that the residue classes of the elements in $\mathscr{O} = \{1, x, y\}$ form a K-vector space basis of their coordinate ring. Another way of saying this is that we want to represent all ideals I in $P = K[x,y]$ such that the residue classes of the elements in $\mathscr{O}$ form a K-basis of P/I.



In this picture the elements of $\mathscr{O} = \{1, x, y\}$ are represented by bullets. Knowing that their residue classes form a K-vector space basis of P/I implies, in particular, that the elements represented by circles, *i.e.* x^2, xy, y^2 can be expressed modulo I as linear combinations of the elements in $\mathscr{O}$. In other words, the ideal I has to contain three polynomials of the form $g_1 = x^2 - c_{11} - c_{21}x - c_{31}y$, $g_2 = xy - c_{12} - c_{22}x - c_{32}y$, and $g_3 = y^2 - c_{13} - c_{23}x - c_{33}y$ for suitable values of the coefficients $c_{ij} \in K$.

But, of course, this is not the end of the discussion. For instance, the unit ideal contains such polynomials, but $\{1,x,y\}$ is not a basis modulo it. To achieve this property, we observe that $\{1,x,y\}$ is an order ideal of monomials and that the complementary monomial ideal is generated by $\{x^2,xy,y^2\}$. If σ is a degree-compatible term ordering, for instance $\sigma = \texttt{DegRevLex}$, we have $\mathrm{LT}_\sigma(g_1) = x^2$, $\mathrm{LT}_\sigma(g_2) = xy$, and $\mathrm{LT}_\sigma(g_3) = y^2$, independent of the values of the coefficients c_{ij}.

Macaulay's Basis Theorem (see [20], Theorem 1.5.7) implies that we have $\dim_K(P/I) = \dim_K(P/\mathrm{LT}_\sigma(I))$, and we want that this number is three. On the other hand, we have $\dim_K(P/(x^2,xy,y^2)) = 3$. Hence we want that $\mathrm{LT}_\sigma(I) = (x^2,xy,y^2)$. In other words, we want to impose that $\{g_1,g_2,g_3\}$ is a σ-Gröbner basis of I. Due to the particular shape of the equations involved, this requirement is equivalent to imposing that $\{g_1,g_2,g_3\}$ is the reduced σ-Gröbner basis of I.

How do we do that? There are two non-trivial fundamental syzygies of the tuple of terms (x^2,xy,y^2), namely $(-y,x,0)$ and $(0,-y,x)$. First we consider the S-polynomial $-yg_1 + xg_2 = c_{11}y + c_{21}xy + c_{31}y^2 - c_{12}x - c_{22}x^2 - c_{32}xy$. Using g_1,g_2,g_3, it can be rewritten as

$$(c_{21}c_{12} + c_{31}c_{13} - c_{22}c_{11} - c_{32}c_{12}) + (-c_{12} + c_{31}c_{23} - c_{32}c_{22})x \\ +(c_{11} + c_{21}c_{32} + c_{31}c_{33} - c_{22}c_{31} - c_{32}^2)y$$

Second, we consider the S-polynomial $-yg_2 + xg_3 = c_{12}y + c_{22}xy + c_{32}y^2 - c_{13}x - c_{23}x^2 - c_{33}xy$. Using g_1,g_2,g_3, it can be rewritten as

$$(c_{22}c_{12} + c_{32}c_{13} - c_{23}c_{11} - c_{33}c_{12}) + (-c_{13} + c_{22}^2 + c_{32}c_{23} - c_{23}c_{21} - c_{33}c_{22})x \\ +(c_{12} + c_{22}c_{32} - c_{23}c_{31})y$$

Imposing that $\{g_1,g_2,g_3\}$ is the reduced σ-Gröbner basis of I is therefore equivalent to imposing that the following expressions are all zero:

$$\begin{aligned}
F_1 &= c_{21}c_{12} + c_{31}c_{13} - c_{22}c_{11} - c_{32}c_{12} \\
F_2 &= -c_{12} + c_{31}c_{23} - c_{32}c_{22} \\
F_3 &= c_{11} + c_{21}c_{32} + c_{31}c_{33} - c_{22}c_{31} - c_{32}^2 \\
F_4 &= c_{22}c_{12} + c_{32}c_{13} - c_{23}c_{11} - c_{33}c_{12} \\
F_5 &= -c_{13} + c_{22}^2 + c_{32}c_{23} - c_{23}c_{21} - c_{33}c_{22} \\
F_6 &= c_{12} + c_{22}c_{32} - c_{23}c_{31}
\end{aligned}$$

Let J be the ideal of $K[c_{11},\dots,c_{33}]$ generated by $\{F_1,F_2,F_3,F_4,F_5,F_6\}$. We note the equality $F_6 = -F_2 = c_{12} + c_{22}c_{32} - c_{23}c_{31}$ and check with CoCoA that $J = (F_2,F_3,F_5)$. By mapping c_{11} to $-c_{21}c_{32} - c_{31}c_{33} + c_{22}c_{31} + c_{32}^2$, c_{12} to $c_{31}c_{23} - c_{32}c_{22}$, and c_{13} to $c_{22}^2 + c_{32}c_{23} - c_{23}c_{21} - c_{33}c_{22}$, we define an isomorphism

$$K[c_{11},\dots,c_{33}]/J \xrightarrow{\sim} K[c_{21},c_{31},c_{22},c_{32},c_{23},c_{33}]$$

The conclusion is that *all* zero-dimensional subschemes of $\mathbb{A}^2_K$ which have the property that the residue classes of the elements in $\{1,x,y\}$ form a K-vector space basis of their coordinate ring are parametrized by an affine space $\mathbb{A}^6_K$. Their vanishing ideals are generated by polynomials $\{g_1,g_2,g_3\}$ where

$$\begin{aligned} g_1 &= x^2-(-c_{21}c_{32}-c_{31}c_{33}+c_{22}c_{31}+c_{32}^2)-c_{21}x-c_{31}y\\ g_2 &= xy-(c_{31}c_{23}-c_{32}c_{22})-c_{22}x-c_{32}y\\ g_3 &= y^2-(c_{22}^2+c_{32}c_{23}-c_{23}c_{21}-c_{33}c_{22})-c_{23}x-c_{33}y \end{aligned}$$

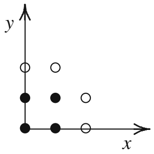
and the parameters which show up in these three polynomials can vary freely. Notice that this family contains only one monomial ideal, namely (x^2,xy,y^2).

To summarize this discussion, we note that we found the parametrizing scheme $\mathbb{A}^6_K$ by imposing that certain fundamental syzygies lift properly. This process is far from canonical. Nevertheless, it can be shown that the output is independent of the choices made (see [28], Proposition 3.5).

Furthermore, we observe that the dimension of the parameter space is six. How can we explain this number? We could argue as follows. Among the ideals represented by the family, there are the vanishing ideals of three non-collinear points. (For three collinear points, the set $\{1,x,y\}$ would not be linearly independent modulo their vanishing ideal.) Clearly, to represent three points in the affine plane one needs six independent coordinates. But we have to be careful: this argument does not work in general! Indeed, we cannot exclude *a priori* the existence of a component of the parameter space of higher dimension. In other words, we do not know *a priori* whether *degenerate schemes* which have $\{1,x,y\}$ as a basis of their coordinate ring can be represented as *limits* of sets of three distinct points. It turns out that this is true in our example, but not for more complicated order ideals $\mathcal{O}$. A similar counter-intuitive situation arises in automatic theorem proving (see for instance [21], Section 6.7).

The next interesting case is the order ideal $\mathcal{O}=\{1,x,y,xy\}$ in $\mathbb{T}^2$.

Example 1.6.2. (Four Points)
As in the example before, we would like to parametrize *all* zero-dimensional subschemes of $\mathbb{A}^2_K$ such that the residue classes of the elements in $\mathcal{O}=\{1,x,y,xy\}$ form a K-vector space basis of their coordinate ring.



Let us try to argue as in the preceding example. The complement of the set $\mathcal{O}$ is the monomial ideal generated by $\{x^2,y^2\}$. Thus we want $\mathrm{LT}_\sigma(I)=(x^2,y^2)$. However, at this point we encounter a serious problem: no matter which term ordering we choose, it is not possible that both x^2 and y^2 are bigger than xy. The best we can do is to pick a degree-compatible term ordering, say $\sigma=\mathtt{DegRevLex}$, and use two polynomials of the form $g_1 = x^2 - c_{11} - c_{21}x - c_{31}y - c_{41}xy$ and $g_2 = y^2 - c_{12} - c_{22}x - c_{32}y$. Then, for every choice of the parameters $c_{ij} \in K$, the set $\{g_1, g_2\}$ is the reduced σ-Gröbner basis of an ideal I such that the residue classes of the elements in $\{1, x, y, xy\}$ form a K-vector space basis of P/I.

Here we have seven free parameters. But four points in the affine plane need eight parameters to describe them completely. This shows that our Gröbner basis approach is not sufficient. A better way to proceed is to consider the border of the given order ideal. Its elements are marked by circles in the above picture. We represent every element in the border as a generic linear combination of $\{1, x, y, xy\}$ and impose that the given generic border prebasis is a border basis. In this way we obtain a set of equations which define an 8-dimensional moduli scheme (see Example 1.6.14).

1.6.B. Border Basis Schemes. The above examples indicate that border bases are well suited for describing families of affine subschemes of $\mathbb{A}^n$ whose coordinate rings have a given K-basis. In fact, they do the job better than Gröbner bases. It is time to provide the precise definitions and technical details necessary for the theoretical foundation of these observations.

Definition 1.6.3. Let $\mathcal{O} = \{t_1, \dots, t_\mu\}$ be an order ideal, let $\partial\mathcal{O} = \{b_1, \dots, b_\nu\}$, and let $\{c_{ij} \mid 1 \le i \le \mu,\ 1 \le j \le \nu\}$ be a set of new indeterminates.

a) The **generic $\mathcal{O}$-border prebasis** is the set of polynomials $G = \{g_1, \dots, g_\nu\}$ in $K[x_1, \dots, x_n, c_{11}, \dots, c_{\mu\nu}]$ given by

$$g_j = b_j - \sum_{i=1}^{\mu} c_{ij} t_i \qquad \text{for } j = 1, \dots, \nu$$

b) For $k = 1, \dots, n$, let $\mathcal{A}_k \in \mathrm{Mat}_\mu(K[c_{ij}])$ be the k^{th} formal multiplication matrix associated to G (cf. [21], Def. 6.4.29). It is also called the k^{th} **generic multiplication matrix** with respect to $\mathcal{O}$.

c) The affine scheme $\mathbb{B}_\mathcal{O} \subseteq \mathbb{A}^{\mu\nu}$ defined by the ideal $I(\mathbb{B}_\mathcal{O})$ which is generated by the entries of the matrices $\mathcal{A}_k\mathcal{A}_\ell - \mathcal{A}_\ell\mathcal{A}_k$ with $1 \le k < \ell \le n$ is called the **$\mathcal{O}$-border basis scheme**.

d) The coordinate ring $K[c_{11}, \dots, c_{\mu\nu}]/I(\mathbb{B}_\mathcal{O})$ of the scheme $\mathbb{B}_\mathcal{O}$ will be denoted by $B_\mathcal{O}$.

We observe that, by definition, the ideal $I(\mathbb{B}_\mathcal{O})$ is generated by polynomials of degree two. By [21], Thm. 6.4.30, a point $(\alpha_{ij}) \in K^{\mu\nu}$ yields a border basis $\sigma(G)$ when we apply the substitution $\sigma(c_{ij}) = \alpha_{ij}$ to G if and only if $\sigma(\mathcal{A}_k)\,\sigma(\mathcal{A}_\ell) = \sigma(\mathcal{A}_\ell)\,\sigma(\mathcal{A}_k)$ for $1 \le k < \ell \le n$. Therefore the K-rational points of $\mathbb{B}_\mathcal{O}$ are in 1–1 correspondence with the $\mathcal{O}$-border bases of zero-dimensional ideals in P, and thus with all zero-dimensional ideals having an $\mathcal{O}$-border basis.

In the following remark, we collect some basic properties of border basis schemes.

Remark 1.6.4. Let $\mathcal{O}$ be an order ideal in $\mathbb{T}^n$, and let $\mathbb{B}_\mathcal{O}$ be the $\mathcal{O}$-border basis scheme.

a) There is an irreducible component of $\mathbb{B}_{\mathcal{O}}$ of dimension $n\mu$ which is the closure of the set of radical ideals having an $\mathcal{O}$-border basis.
b) There is an example by Iarrobino (see [23] and [22], Example 5.6) which exhibits a border basis scheme having an irreducible component whose dimension is higher than $n\mu$.
c) For every term ordering σ, there is a subset of $\mathbb{B}_{\mathcal{O}}$ which parametrizes all ideals I such that $\mathcal{O} = \mathcal{O}_\sigma(I)$. These subsets have turned out to be useful for studying the Hilbert scheme parametrizing subschemes of $\mathbb{A}^n$ of length μ (see for instance [7] and [27]).
d) In the case $n = 2$ more precise information about $\mathbb{B}_{\mathcal{O}}$ is available: for instance, it is known that $\mathbb{B}_{\mathcal{O}}$ is reduced, irreducible and smooth of dimension 2μ (see [13], [15] and [24], Ch. 18).

Our next remark clarifies the relation between border basis schemes and Hilbert schemes.

Remark 1.6.5. For an order ideal $\mathcal{O}$ in $\mathbb{T}^n$, the border basis scheme $\mathbb{B}_{\mathcal{O}}$ can be embedded as an open affine subscheme of the Hilbert scheme parametrizing subschemes of $\mathbb{A}^n$ of length μ (see [24], Section 18.4). This can be seen as follows.

Let $I_{\mathcal{O}}$ be the monomial ideal generated by the complement of $\mathcal{O}$. The Hilbert polynomial of $P/I_{\mathcal{O}}$ is the constant polynomial μ. Among all schemes having this Hilbert polynomial there are the schemes for which $\mathcal{O}$ yields a basis of their coordinate ring. This condition defines a Zariski open subset.

As usual, a moduli space such as the border basis scheme comes together with a universal flat family. In the present setting it is defined as follows.

Definition 1.6.6. Let $G = \{g_1,\dots,g_\nu\} \subset K[x_1,\dots,x_n,c_{11},\dots,c_{\mu\nu}]$ with $g_j = b_j - \sum_{i=1}^{\mu} c_{ij}t_i$ for $j = 1,\dots,\nu$ be the generic $\mathcal{O}$-border prebasis. We will denote the ring $K[x_1,\dots,x_n,c_{11},\dots,c_{\mu\nu}]/(I(\mathbb{B}_{\mathcal{O}}) + (g_1,\dots,g_\nu))$ by $U_{\mathcal{O}}$. Then the natural homomorphism of K-algebras

$$\Phi :\ B_{\mathcal{O}} \longrightarrow U_{\mathcal{O}} \cong B_{\mathcal{O}}[x_1,\dots,x_n]/(g_1,\dots,g_\nu)$$

is called the **universal $\mathcal{O}$-border basis family**.

What are the fibers of this family? It is easy to understand that they are the quotient rings P/I for which I is a zero-dimensional ideal which has an $\mathcal{O}$-border basis. The special fiber, *i.e.* the fiber corresponding to $(0,\dots,0)$, is the ring $P/(\partial\mathcal{O})$. It is the only fiber in the family which is defined by a monomial ideal. A remarkable result is the following.

Theorem 1.6.7. (The Universal Border Basis Family)
Let $\Phi : B_{\mathcal{O}} \longrightarrow U_{\mathcal{O}}$ be the universal $\mathcal{O}$-border basis family. Then the residue classes of the elements of $\mathcal{O}$ are a $B_{\mathcal{O}}$-module basis of $U_{\mathcal{O}}$. In particular, the map Φ is a flat homomorphism.

Proof. See [11] or [16]. For an elementary proof see [22], Theorem 3.4. □

Let us have a look at the first consequences of this fundamental result. A rational curve on the $\mathcal{O}$-border basis scheme corresponds to a surjective K-algebra homomorphism $\Psi : B_{\mathcal{O}} \longrightarrow K[z]$ of the corresponding affine coordinate rings. If we restrict the universal family of $\mathcal{O}$-border bases to this rational curve, we obtain the following flat deformation of border bases.

Corollary 1.6.8. *Let z be a new indeterminate, and let $\Psi : B_{\mathcal{O}} \longrightarrow K[z]$ be a surjective homomorphism of K-algebras. By applying the base change Ψ to the universal family Φ, we get a homomorphism of $K[z]$-algebras*

$$\Phi_{K[z]} = \Phi \otimes_{B_{\mathcal{O}}} K[z] : \; K[z] \longrightarrow U_{\mathcal{O}} \otimes_{B_{\mathcal{O}}} K[z]$$

Then the residue classes of the elements of $\mathcal{O}$ form a $K[z]$-module basis of the right-hand side. In particular, the map $\Phi_{K[z]}$ defines a flat family.

As explained in [22], this corollary can be used to construct flat deformations over $K[z]$ of border bases. Suppose the maximal ideal $\Psi^{-1}(z-1)$ corresponds to a given $\mathcal{O}$-border basis and the maximal ideal $\Psi^{-1}(z)$ is the ideal $(c_{11},\dots,c_{\mu\nu})$ which corresponds to the border term ideal $(b_1,\dots,b_\nu)$. In other words, suppose that there exists a rational curve which connects the given point to the point $(0,\dots,0)$. Then the map $\Phi_{K[z]}$ defines a flat family over $K[z]$ whose generic fiber P/I is defined by the ideal I generated by the given $\mathcal{O}$-border basis and whose special fiber $P/(b_1,\dots,b_\nu)$ is defined by the border term ideal.

In the next part of this subsection we try to construct explicit flat deformations to the border term ideal. The idea is to imitate the method used in Gröbner basis theory, namely the technique of homogenization. The first step is to deform to a suitable degree form ideal.

Lemma 1.6.9. *Let P be graded by a matrix $W \in \mathrm{Mat}_{m,n}(\mathbb{Z})$, let $\mathcal{O}$ be an order ideal in $\mathbb{T}^n$, and let $I \subset P$ be a homogeneous ideal which has an $\mathcal{O}$-border basis. Then this $\mathcal{O}$-border basis of I consists of homogeneous polynomials.*

Proof. See [22], Lemma 2.3. □

As for the idea to deform a border basis of I to a homogeneous border basis of the degree form ideal $\mathrm{DF}_W(I)$, we have the following result.

Theorem 1.6.10. (Deformation to the Degree Form Ideal)
Let $W = (w_1,\dots,w_n) \in \mathrm{Mat}_{1,n}(\mathbb{N}_+)$ be a row of positive integers, let P be graded by W, and let $I \subset P$ be a zero-dimensional ideal. Then the following conditions are equivalent.

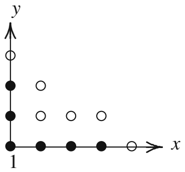
a) The ideal I has an $\mathcal{O}$-border basis, say $G = \{g_1,\dots,g_\nu\}$, and we have $b_j \in \mathrm{Supp}(\mathrm{DF}_W(g_j))$ for $j = 1,\dots,\nu$.
b) The degree form ideal $\mathrm{DF}_W(I)$ has an $\mathcal{O}$-border basis.

If these conditions are satisfied, the $\mathcal{O}$-border basis of $\mathrm{DF}_W(I)$ is $\mathrm{DF}_W(G) = \{\mathrm{DF}_W(g_1),\dots,\mathrm{DF}_W(g_s)\}$ and there is a flat family $K[x_0] \longrightarrow \overline{P}/I$hom whose general fiber is isomorphic to P/I, where $I = (g_1,\dots,g_\nu)$, and whose special fiber is isomorphic to $P/\mathrm{DF}_W(I)$, where $\mathrm{DF}_W(I) = (\mathrm{DF}_W(g_1),\dots,\mathrm{DF}_W(g_\nu))$.

Proof. See [22], Theorem 2.4. □

Let us look at an example for the application of this theorem.

Example 1.6.11. Consider the ideal $I = (-2x^2+xy-y^2-1,\, 8y^3+10x+9y)$ in the polynomial ring $P = \mathbb{Q}[x,y]$. The degree form ideal of I with respect to the standard grading, *i.e.* the grading defined by $W = (1\ 1)$, is $\mathrm{DF}_W(I) = (-2x^2+xy-y^2,\, y^3)$. We want to use the order ideal $\mathcal{O} = \{1,x,x^2,x^3,y,y^2\}$ whose border is given by $\partial\mathcal{O} = \{xy,y^3,xy^2,x^2y,x^3y,x^4\}$.



It is easy to check that $\mathrm{DF}_W(I)$ has an $\mathcal{O}$-border basis, namely $H = \{h_1,\dots,h_6\}$ with $h_1 = xy-2x^2-y^2$, $h_2 = y^3$, $h_3 = xy^2+4x^3$, $h_4 = x^2y+2x^3$, $h_5 = x^3y$, and $h_6 = x^4$. Therefore the theorem says that I has an $\mathcal{O}$-border basis $G = \{g_1,\dots,g_6\}$, and that $h_i = \mathrm{DF}_W(g_i)$ for $i = 1,\dots,6$. Indeed, if we compute this border basis we find that it is given by $g_1 = xy-2x^2-y^2-1$, $g_2 = y^3+\frac{5}{4}x+\frac{9}{8}y$, $g_3 = xy^2+4x^3+\frac{3}{4}x-\frac{1}{8}y$, $g_4 = x^2y+2x^3-\frac{1}{4}x-\frac{1}{8}y$, $g_5 = x^3y-\frac{1}{2}x^2-\frac{1}{8}y^2-\frac{3}{32}$, and $g_6 = x^4-\frac{1}{64}$.

An easy modification of this example shows that the implication "a) $\Longrightarrow$ b)" in the theorem is not true without the hypothesis $b_j \in \mathrm{Supp}(\mathrm{DF}_W(g_j))$. This observation inspires the following definition.

Definition 1.6.12. Let P be graded by a matrix $W \in \mathrm{Mat}_{1,n}(\mathbb{N}_+)$. The order ideal $\mathcal{O}$ is said to have a **maxdeg$_W$ border** if $\deg_W(b_j) \ge \deg_W(t_i)$ for $i = 1,\dots,\mu$ and $j = 1,\dots,\nu$. In other words, no term in $\mathcal{O}$ is allowed to have a degree larger than any term in the border.

Using this notion, we can combine the deformation given by the theorem with a second deformation from the degree form ideal to the border term ideal by using the following result.

Theorem 1.6.13. (Homogeneous Maxdeg Border Bases)
Suppose that the order ideal $\mathcal{O}$ has a maxdeg$_W$ *border. Let $I \subset P$ be a homogeneous ideal which has an $\mathcal{O}$-border basis $G = \{g_1,\dots,g_\nu\}$. Then there exists a flat family $K[z] \longrightarrow K[z][x_1,\dots,x_n]/J$ such that $\mathcal{O}$ is a $K[z]$-basis of the right-hand side, such that $J|_{z\mapsto 1} \cong I$, and such that $J|_{z\mapsto 0} \cong (b_1,\dots,b_\nu)$. In fact, the ideal J may be defined by writing $g_j = b_j - \sum_{i=1}^{\mu} c_{ij}t_i$ and replacing $c_{ij} \in K$ by $c_{ij}z \in K[z]$ for all i,j.*

Proof. See [22], Theorem 5.3. □

To get a good grasp of these deformations, we look at one particular border basis scheme in detail, namely the one corresponding to Example 1.6.2.

Example 1.6.14. Consider the case $n=2$ and $\mathscr{O}=\{1,x,y,xy\}$. The border of $\mathscr{O}$ is $\partial\mathscr{O}=\{y^2,x^2,xy^2,x^2y\}$, so that in our terminology we have $\mu=4$, $\nu=4$, $t_1=1$, $t_2=x$, $t_3=y$, $t_4=xy$, $b_1=y^2$, $b_2=x^2$, $b_3=xy^2$, and $b_4=x^2y$.

The generic multiplication matrices are

$$\mathscr{A}_x=\begin{pmatrix} 0 & c_{12} & 0 & c_{14}\\ 1 & c_{22} & 0 & c_{24}\\ 0 & c_{32} & 0 & c_{34}\\ 0 & c_{42} & 1 & c_{44}\end{pmatrix} \quad\text{and}\quad \mathscr{A}_y=\begin{pmatrix} 0 & 0 & c_{11} & c_{13}\\ 0 & 0 & c_{21} & c_{23}\\ 1 & 0 & c_{31} & c_{33}\\ 0 & 1 & c_{41} & c_{43}\end{pmatrix}$$

When we compute the ideal generated by the entries of $\mathscr{A}_x\mathscr{A}_y-\mathscr{A}_y\mathscr{A}_x$ and simplify its system of generators, we see that the ideal $I(\mathbb{B}_{\mathscr{O}})$ is generated by

$$\begin{aligned}\{&c_{23}c_{41}c_{42}-c_{21}c_{42}c_{43}+c_{21}c_{44}+c_{11}-c_{23},\ -c_{21}c_{32}-c_{34}c_{41}+c_{33},\\ &c_{34}c_{41}c_{42}-c_{32}c_{41}c_{44}+c_{32}c_{43}+c_{12}-c_{34},\ -c_{21}c_{32}-c_{23}c_{42}+c_{24},\\ &-c_{23}c_{32}c_{41}+c_{21}c_{32}c_{43}-c_{21}c_{34}+c_{13},\ c_{21}c_{42}+c_{41}c_{44}+c_{31}-c_{43},\\ &-c_{21}c_{34}c_{42}+c_{21}c_{32}c_{44}-c_{23}c_{32}+c_{14},\ c_{32}c_{41}+c_{42}c_{43}+c_{22}-c_{44}\}\end{aligned}$$

Thus there are eight free indeterminates, namely c_{21}, c_{23}, c_{32}, c_{34}, c_{41}, c_{42}, c_{43}, and c_{44}, while the remaining indeterminates depend on the free ones by the polynomial expressions above. From this we conclude that the border basis scheme $\mathbb{B}_{\mathscr{O}}$ is an *affine cell* of the corresponding Hilbert scheme, *i.e.* an open subset which is isomorphic to an affine space.

Its coordinate ring is explicitly represented by the isomorphism

$$B_{\mathscr{O}} \xrightarrow{\sim} K[c_{21},c_{23},c_{32},c_{34},c_{41},c_{42},c_{43},c_{44}]$$

given by

$$\begin{aligned} c_{11} &\longmapsto -c_{23}c_{41}c_{42}+c_{21}c_{42}c_{43}-c_{21}c_{44}+c_{23}\\ c_{12} &\longmapsto -c_{34}c_{41}c_{42}+c_{32}c_{41}c_{44}-c_{32}c_{43}+c_{34}\\ c_{13} &\longmapsto c_{23}c_{32}c_{41}-c_{21}c_{32}c_{43}+c_{21}c_{34}\\ c_{14} &\longmapsto c_{21}c_{34}c_{42}-c_{21}c_{32}c_{44}+c_{23}c_{32}\\ c_{22} &\longmapsto -c_{32}c_{41}-c_{42}c_{43}+c_{44}\\ c_{24} &\longmapsto c_{21}c_{32}+c_{23}c_{42}\\ c_{31} &\longmapsto -c_{21}c_{42}-c_{41}c_{44}+c_{43}\\ c_{33} &\longmapsto c_{21}c_{32}+c_{34}c_{41}\end{aligned}$$

Hence we have $U_{\mathscr{O}}\cong K[x,y,c_{21},c_{23},c_{32},c_{34},c_{41},c_{42},c_{43},c_{44}]/(\widetilde{g}_1,\widetilde{g}_2,\widetilde{g}_3,\widetilde{g}_4)$ where

$$\begin{aligned}
\widetilde{g}_1 &= y^2-(-c_{23}c_{41}c_{42}+c_{21}c_{42}c_{43}-c_{21}c_{44}+c_{23})\\
&\quad -c_{21}x-(-c_{21}c_{42}-c_{41}c_{44}+c_{43})y-c_{41}xy,\\
\widetilde{g}_2 &= x^2-(-c_{34}c_{41}c_{42}+c_{32}c_{41}c_{44}-c_{32}c_{43}+c_{34})\\
&\quad -(-c_{32}c_{41}-c_{42}c_{43}+c_{44})x-c_{32}y-c_{42}xy,\\
\widetilde{g}_3 &= xy^2-(c_{23}c_{32}c_{41}-c_{21}c_{32}c_{43}+c_{21}c_{34})\\
&\quad -c_{23}x-(c_{21}c_{32}+c_{34}c_{41})y-c_{43}xy,\\
\widetilde{g}_4 &= x^2y-(c_{21}c_{34}c_{42}-c_{21}c_{32}c_{44}+c_{23}c_{32})\\
&\quad -(c_{21}c_{32}+c_{23}c_{42})x-c_{34}y-c_{44}xy,
\end{aligned}$$

The ideal $(\widetilde{g}_1,\widetilde{g}_2,\widetilde{g}_3,\widetilde{g}_4)$ is the defining ideal of the family of all subschemes of length four of the affine plane which have the property that their coordinate ring admits $\mathcal{O}$ as a vector space basis. Since the border basis scheme is isomorphic to an affine space in this case, we can connect every point to the point corresponding to (x^2,y^2) by a rational curve. Therefore every ideal in the family can be deformed by a flat deformation to the monomial ideal (x^2,y^2). Algebraically, it suffices to substitute each free indeterminate c_{ij} with zc_{ij} where z is a new indeterminate. This yields the K-algebra homomorphism

$$\Phi_{K[z]} : K[z] \longrightarrow K[x,y,z,c_{21},c_{23},c_{32},c_{34},c_{41},c_{42},c_{43},c_{44}]/(\overline{g}_1,\overline{g}_2,\overline{g}_3,\overline{g}_4)$$

where

$$\begin{aligned}
\overline{g}_1 &= y^2-(-z^3c_{23}c_{41}c_{42}+z^3c_{21}c_{42}c_{43}-z^2c_{21}c_{44}+zc_{23})\\
&\quad -zc_{21}x-(-z^2c_{21}c_{42}-z^2c_{41}c_{44}+zc_{43})y-zc_{41}xy,\\
\overline{g}_2 &= x^2-(-z^3c_{34}c_{41}c_{42}+z^3c_{32}c_{41}c_{44}-z^2c_{32}c_{43}+zc_{34})\\
&\quad -(-z^2c_{32}c_{41}-z^2c_{42}c_{43}+zc_{44})x-zc_{32}y-zc_{42}xy,\\
\overline{g}_3 &= xy^2-(z^3c_{23}c_{32}c_{41}-z^3c_{21}c_{32}c_{43}+z^2c_{21}c_{34})\\
&\quad -zc_{23}x-(z^2c_{21}c_{32}+z^2c_{34}c_{41})y-zc_{43}xy,\\
\overline{g}_4 &= x^2y-(z^3c_{21}c_{34}c_{42}-z^3c_{21}c_{32}c_{44}+z^2c_{23}c_{32})\\
&\quad -(z^2c_{21}c_{32}+z^2c_{23}c_{42})x-zc_{34}y-zc_{44}xy,
\end{aligned}$$

By Corollary 1.6.8, the homomorphism $\Phi_{K[z]}$ is flat. For every point on the border basis scheme, it induces a flat deformation from the corresponding coordinate ring P/I to $P/(\partial\mathcal{O})$ where the border term ideal is $(\partial\mathcal{O})=(y^2,x^2,xy^2,x^2y)=(x^2,y^2)$.

Finally, we want to draw the connection between border basis schemes and the approximate border bases defined in Section 4.

Remark 1.6.15. Let $\mathcal{O}=\{t_1,\dots,t_\mu\}$ be an order ideal in $\mathbb{T}^n$ and $\partial\mathcal{O}=\{b_1,\dots,b_\nu\}$ its border.

a) An approximate $\mathcal{O}$-border basis $G=\{g_1,\dots,g_\nu\}$ with $g_j = b_j - \sum_{i=1}^{\mu} a_{ij}t_i$, as defined in Section 4, yields a point $(a_{ij}) \in \mathbb{R}^{\mu\nu}$ which is *close* to the $\mathcal{O}$-border basis scheme. In fact, in Definition 1.4.7 we required that the normal remainders of the S-polynomials of neighbour syzygies are small. This implies that the coefficients of these normal remainders are small, and those coefficients are precisely the evaluations of the defining equations of $\mathbb{B}_{\mathcal{O}}$ at (a_{ij}).
b) The AVI algorithm computes an approximate $\mathcal{O}$-border basis. As mentioned above, this corresponds to a point p close to the border basis scheme. Therefore it is natural to ask how one can find an exact $\mathcal{O}$-border basis defined over $\mathbb{Q}$, *i.e.* a rational point on $\mathbb{B}_{\mathcal{O}}$ which is close to p. This problem, called the *rational recovery problem*, will be addressed in [19].

1.6.C. Gröbner Basis Schemes. In the first subsection we tried to use the shape of a Gröbner basis in order to parametrize families of zero-dimensional ideals, but we encountered difficulties. Then we saw that border bases are more suited for this purpose. Now we return to the Gröbner basis approach and try to put it in relation to the border basis technique. To this end, we plan to define $(\mathcal{O},\sigma)$-Gröbner basis schemes.

Before we start the discussion, some extra bits of notation are required. Given an order ideal $\mathcal{O}=\{t_1,\dots,t_\mu\}$, the set of minimal generators of the monoideal $\mathbb{T}^n \setminus \mathcal{O}$ (which are also called the **corners** of $\mathcal{O}$) is denoted by $c\mathcal{O}$, and we let η be the cardinality of $c\mathcal{O}$. Since $c\mathcal{O} \subseteq \partial\mathcal{O}$, it follows that $\eta \le \nu$. Without loss of generality, we label the elements in $\partial\mathcal{O}$ so that $c\mathcal{O}=\{b_1,\dots,b_\eta\}$.

Next we let σ be a term ordering on $\mathbb{T}^n$. Recall that, for an ideal I in the polynomial ring P, we denote the order ideal $\mathbb{T}^n \setminus \mathrm{LT}_\sigma(I)$ by $\mathcal{O}_\sigma(I)$. Moreover, we denote by $S_{\mathcal{O},\sigma}$ the set $\{c_{ij} \in \{c_{11},\dots,c_{\mu\nu}\} \mid b_j >_\sigma t_i\}$, by $L_{\mathcal{O},\sigma}$ the ideal generated by $\{c_{11},\dots,c_{\mu\nu}\} \setminus S_{\mathcal{O},\sigma}$ in $K[c_{11},\dots,c_{\mu\nu}]$, by $S_{c\mathcal{O},\sigma}$ the intersection $S_{\mathcal{O},\sigma} \cap \{c_{11},\dots,c_{\mu\eta}\}$, and by $L_{c\mathcal{O},\sigma}$ the ideal generated by $\{c_{11},\dots,c_{\mu\eta}\} \setminus S_{c\mathcal{O},\sigma}$ in $K[c_{11},\dots,c_{\mu\eta}]$. Furthermore, we denote the cardinality of $S_{c\mathcal{O},\sigma}$ by $s(c\mathcal{O},\sigma)$.

Definition 1.6.16. For $j=1,\dots,\nu$, we define a polynomial g_j^* by

$$g_j^* \;=\; b_j - \sum_{\{i \,\mid\, b_j >_\sigma t_i\}} c_{ij}t_i \;=\; b_j - \sum_{c_{ij} \in S_{\mathcal{O},\sigma} \cap \{c_{1j},\dots,c_{\mu j}\}} c_{ij}t_i$$

a) The set of polynomials $\{g_1^*,\dots,g_\eta^*\}$ is called the **generic** $(\mathcal{O},\sigma)$**-Gröbner prebasis**.
b) The ideal $\big(L_{\mathcal{O},\sigma} + I(\mathbb{B}_{\mathcal{O}})\big) \cap K[S_{c\mathcal{O},\sigma}]$ of $K[S_{c\mathcal{O},\sigma}]$ defines an affine subscheme of $\mathbb{A}^{s(c\mathcal{O},\sigma)}$ which will be denoted by $\mathbb{G}_{\mathcal{O},\sigma}$ and called the $(\mathcal{O},\sigma)$**-Gröbner basis scheme**. Its defining ideal $\big(L_{\mathcal{O},\sigma} + I(\mathbb{B}_{\mathcal{O}})\big) \cap K[S_{c\mathcal{O},\sigma}]$ will be denoted by $I(\mathbb{G}_{\mathcal{O},\sigma})$ and its coordinate ring $K[S_{c\mathcal{O},\sigma}]/I(\mathbb{G}_{\mathcal{O},\sigma})$ by $G_{\mathcal{O},\sigma}$.

Notice that the polynomial g_j^* is obtained from g_j by setting all indeterminates in $L_{\mathcal{O},\sigma} \cap \{c_{1j},\dots,c_{\mu j}\}$ to zero.

What is the relation between Gröbner basis schemes and border basis schemes? Well, by now it should be clear that a Gröbner basis scheme is a closed subscheme of the corresponding border basis scheme.

Example 1.6.17. Let us examine the inclusion $c\mathcal{O} \subseteq \partial\mathcal{O}$. If $\mathcal{O} = \{1,x,y,xy\}$ then $c\mathcal{O} = \{x^2,y^2\}$ while $\partial\mathcal{O} = \{x^2,y^2,x^2y,xy^2\}$, so that $c\mathcal{O} \subset \partial\mathcal{O}$. On the other hand, if $\mathcal{O} = \{1,x,y\}$ then $c\mathcal{O} = \partial\mathcal{O} = \{x^2,xy,y^2\}$.

Returning to $\mathcal{O} = \{1,x,y,xy\}$ we have $t_1 = 1$, $t_2 = x$, $t_3 = y$, $t_4 = xy$, $b_1 = x^2$, $b_2 = y^2$, $b_3 = x^2y$, $b_4 = xy^2$. Let $\sigma = \texttt{DegRevLex}$, so that $x >_\sigma y$. Then we get $L_{\mathcal{O},\sigma} = L_{c\mathcal{O},\sigma} = (c_{42})$, $g_1^* = g_1$, $g_2^* = y^2 - (c_{12} + c_{22}x + c_{32}y)$, $g_3^* = g_3$, and $g_4^* = g_4$.

Having introduced the Gröbner basis scheme, we define a naturally associated universal family. We recall that $K[x_1,\dots,x_n,c_{11},\dots,c_{\mu\nu}]/\big(I(\mathbb{B}_\mathcal{O}) + (g_1,\dots,g_\nu)\big)$ was denoted by $U_\mathcal{O}$ in Definition 1.6.6, and the natural homomorphism of K-algebras $\Phi:\ B_\mathcal{O} \longrightarrow U_\mathcal{O}$ was called the universal $\mathcal{O}$-border basis family.

Definition 1.6.18. The ring $K[x_1,\dots,x_n,S_{c\mathcal{O},\sigma}]/\big(I(\mathbb{G}_{\mathcal{O},\sigma}) + (g_1^*,\dots,g_\eta^*)\big)$ will be denoted by $U_{\mathcal{O},\sigma}$.

a) The natural homomorphism of K-algebras $\Psi:\ G_{\mathcal{O},\sigma} \longrightarrow U_{\mathcal{O},\sigma}$ is called the **universal $(\mathcal{O},\sigma)$-Gröbner basis family**.
b) The induced homomorphism of K-algebras $B_\mathcal{O}/\overline{L}_{\mathcal{O},\sigma} \longrightarrow U_\mathcal{O}/\overline{L}_{\mathcal{O},\sigma}$ will be denoted by $\overline{\Phi}$.

The next result shows than Gröbner basis schemes have a very nice property which is not shared by some border basis schemes. To help the reader, we simply write $\mathbf{x}$ for $x_1,\dots,x_n$ and $\mathbf{c}$ for $c_{11},\dots c_{\mu\nu}$.

Theorem 1.6.19. *There exists a system W of positive weights on the elements of $S_{c\mathcal{O},\sigma}$, a system $\overline{W}$ of positive weights on the elements of $S_{\mathcal{O},\sigma}$, and a system V of positive weights on $\mathbf{x}$ such that the following conditions hold true.*

a) The system $\overline{W}$ is an extension of the system W.
b) The ideal $I(\mathbb{G}_{\mathcal{O},\sigma})$ in $K[S_{c\mathcal{O},\sigma}]$ is W-homogeneous.
c) The ideal $I(\mathbb{G}_{\mathcal{O},\sigma}) + (g_1^*,\dots,g_\eta^*)$ in $K[\mathbf{x},S_{c\mathcal{O},\sigma}]$ is (V,W)-homogeneous.
d) The image of $I(\mathbb{B}_\mathcal{O})$ in $K[S_{\mathcal{O},\sigma}] \cong K[\mathbf{c}]/L_{\mathcal{O},\sigma}$ is $\overline{W}$-homogeneous.
e) The image of $I(\mathbb{B}_\mathcal{O}) + (g_1^*,\dots,g_\nu^*)$ in the ring $K[\mathbf{x},S_{\mathcal{O},\sigma}] \cong K[\mathbf{x},\mathbf{c}]/L_{\mathcal{O},\sigma}$ is $(V,\overline{W})$-homogeneous.

Proof. See [28], Theorem 2.8.

In other words, this theorem says that a Gröbner basis scheme has an intrinsic graded structure. It follows that it is isomorphic to an affine space if and only if the point corresponding to the unique monomial ideal is smooth (see [28], Corollary 3.7). Moreover, Gröbner basis schemes are connected. The analogous result for border basis schemes is not known. (A partial result is that if $\mathcal{O}$ has a $\max\deg_W$ border, then $\mathbb{B}_\mathcal{O}$ is connected. This follows by combining Theorem 1.6.10 with [22], Theorem 5.3.)

Our next goal is to make the connection between Gröbner and border basis schemes more explicit. We recall the equality $I(\mathbb{G}_{\mathcal{O}}) = \big(L_{\mathcal{O},\sigma} + I(\mathbb{B}_{\mathcal{O}})\big) \cap K[S_{c\mathcal{O},\sigma}]$ which yields the homomorphism φ below. A further homomorphism ϑ is obtained as follows: let $\Theta : K[\mathbf{x}, S_{c\mathcal{O},\sigma}] \longrightarrow K[\mathbf{x},\mathbf{c}]$ be the natural inclusion of polynomial rings. Then clearly $I(\mathbb{G}_{\mathcal{O},\sigma}) + (g_1^*,\dots,g_\eta^*) \subseteq \Theta^{-1}\big(L_{\mathcal{O},\sigma} + I(\mathbb{B}_{\mathcal{O}}) + (g_1,\dots g_\nu)\big)$.

Now we consider the following commutative diagram of canonical homomorphisms.

$$\begin{tikzcd}
G_{\mathcal{O},\sigma} \arrow[r, "\varphi"] \arrow[d, "\Psi"] & B_{\mathcal{O}}/\overline{L}_{\mathcal{O},\sigma} \arrow[d, "\overline{\Phi}"] \\
U_{\mathcal{O},\sigma} \arrow[r, "\vartheta"] & U_{\mathcal{O}}/\overline{L}_{\mathcal{O},\sigma}
\end{tikzcd}$$

Using explicit representations, this diagram has the following form.

$$\begin{tikzcd}
K[S_{c\mathcal{O},\sigma}]/I(\mathbb{G}_{\mathcal{O},\sigma}) \arrow[r, "\varphi"] \arrow[d, "\Psi"] & K[\mathbf{c}]/\big(L_{\mathcal{O},\sigma} + I(\mathbb{B}_{\mathcal{O}})\big) \arrow[d, "\overline{\Phi}"] \\
K[\mathbf{x}, S_{c\mathcal{O},\sigma}]/\big(I(\mathbb{G}_{\mathcal{O},\sigma}) + (g_1^*,\dots,g_\eta^*)\big) \arrow[r, "\vartheta"] & K[\mathbf{x},\mathbf{c}]/\big(L_{\mathcal{O},\sigma} + I(\mathbb{B}_{\mathcal{O}}) + (g_1,\dots g_\nu)\big)
\end{tikzcd}$$

At this point we are ready for the following fundamental results about Gröbner basis schemes.

Theorem 1.6.20. (Gröbner Basis and Border Basis Schemes)
Let $\mathcal{O} = \{t_1,\dots,t_\mu\}$ be an order ideal of monomials, and let σ be a term ordering on $\mathbb{T}^n$.

a) The classes of the elements in $\mathcal{O}$ form a $B_{\mathcal{O}}/\overline{L}_{\mathcal{O},\sigma}$-module basis of $U_{\mathcal{O}}/\overline{L}_{\mathcal{O},\sigma}$.
b) The classes of the elements in $\mathcal{O}$ form a $G_{\mathcal{O},\sigma}$-module basis of $U_{\mathcal{O},\sigma}$.
c) We have $I(\mathbb{G}_{\mathcal{O},\sigma}) + (g_1^*,\dots,g_\eta^*) = \vartheta^{-1}\big(L_{\mathcal{O},\sigma} + I(\mathbb{B}_{\mathcal{O}}) + (g_1,\dots g_\nu)\big)$.
d) The maps φ and ϑ in the above diagram are isomorphisms.

Proof. See [28], Theorem 2.9.

Corollary 1.6.21. *Let $\mathcal{O} = \{t_1,\dots,t_\mu\}$ be an order ideal of monomials in P and let σ be a term ordering on $\mathbb{T}^n$.*

a) The affine scheme $\mathbb{G}_{\mathcal{O},\sigma}$ parametrizes all zero-dimensional ideals I in P for which $\mathcal{O} = \mathcal{O}_\sigma(I)$.
b) The fibers over the K-rational points of the universal $(\mathcal{O},\sigma)$ Gröbner family $\Psi : G_{\mathcal{O},\sigma} \longrightarrow U_{\mathcal{O},\sigma}$ are the quotient rings P/I for which I is a zero-dimensional ideal with the property that $\mathcal{O} = \mathcal{O}_\sigma(I)$. Moreover, the reduced σ-Gröbner basis of I is obtained by specializing the $(\mathcal{O},\sigma)$-Gröbner prebasis $\{g_1^*,\dots,g_\eta^*\}$ to the corresponding maximal linear ideal.

Proof. See [28]. Corollary 2.11.

Finally, we reformulate these results in the language of algebraic geometry.

Remark 1.6.22. There is a commutative diagram

$$\begin{tikzcd}
\mathbb{G}_{\mathcal{O},\sigma} \arrow[r, phantom, "\cong"] & \operatorname{Spec}(B_{\mathcal{O}}/\overline{L}_{\mathcal{O},\sigma}) \\
\operatorname{Spec}(U_{\mathcal{O},\sigma}) \arrow[u, "\pi_{\Psi}"'] \arrow[r, phantom, "\cong"] & \operatorname{Spec}(U_{\mathcal{O}}/\overline{L}_{\mathcal{O},\sigma}) \arrow[u, "\pi_{\overline{\Phi}}"']
\end{tikzcd}$$

of affine schemes, but more can be said. Let W, $\overline{W}$, and V be systems of positive weights, chosen suitably to satisfy Theorem 1.6.19. Then $G_{\mathcal{O},\sigma}$ is a W-graded ring, $B_{\mathcal{O}}$ is a $\overline{W}$-graded ring, $U_{\mathcal{O},\sigma}$ is a (V,W)-graded ring, and $U_{\mathcal{O}}/\overline{L}_{\mathcal{O},\sigma}$ is a $(V,\overline{W})$-graded ring.

Hence we see that the above diagram gives rise to a diagram

$$\begin{tikzcd}
\operatorname{Proj}(G_{\mathcal{O},\sigma}) \arrow[r, phantom, "\cong"] & \operatorname{Proj}(B_{\mathcal{O}}/\overline{L}_{\mathcal{O},\sigma}) \\
\operatorname{Proj}(U_{\mathcal{O},\sigma}) \arrow[u, "\Pi_{\Psi}"'] \arrow[r, phantom, "\cong"] & \operatorname{Proj}(U_{\mathcal{O}}/\overline{L}_{\mathcal{O},\sigma}) \arrow[u, "\Pi_{\overline{\Phi}}"']
\end{tikzcd}$$

of projective schemes such that $\operatorname{Proj}(G_{\mathcal{O},\sigma}) \subset \mathbb{P}(W)$, $\operatorname{Proj}(B_{\mathcal{O}}/\overline{L}_{\mathcal{O},\sigma}) \subset \mathbb{P}(\overline{W})$, $\operatorname{Proj}(U_{\mathcal{O},\sigma}) \subset \mathbb{P}(V,W)$, and $\operatorname{Proj}(U_{\mathcal{O}}/\overline{L}_{\mathcal{O},\sigma}) \subset \mathbb{P}(V,\overline{W})$. The corresponding weighted projective spaces are denoted by $\mathbb{P}(W)$, $\mathbb{P}(\overline{W})$, $\mathbb{P}(V,W)$, and $\mathbb{P}(V,\overline{W})$.

Moreover, let $\mathbf{p} = (a_{ij}) \in \mathbb{G}_{\mathcal{O},\sigma}$ be a rational point, let $I \subset P$ be the corresponding ideal according to Corollary 1.6.21, let $v_i = \deg(x_i)$ in the V-grading, and let $w_{ij} = \deg(c_{ij})$ in the W-grading. Then it is well-known that the substitution $a_{ij} \longrightarrow t^{w_{ij}} a_{ij}$ gives rise to a flat family of ideals whose general fibers are ideals isomorphic to I, and whose special fiber is the monomial ideal $\operatorname{LT}_\sigma(I)$. In the setting of the first diagram, the rational monomial curve which parametrizes this family is a curve in $\mathbb{G}_{\mathcal{O},\sigma}$ which connects the two points representing I and $\operatorname{LT}_\sigma(I)$. In the setting of the second diagram, the rational monomial curve is simply a point in $\operatorname{Proj}(G_{\mathcal{O},\sigma}) \subset \mathbb{P}(W)$, which represents all of these ideals except for the special one.

In [28], Section 3, the relation between the construction of $I(\mathbb{G}_{\mathcal{O}})$ and other constructions described in the literature (see for instance [7] and [27]) is discussed. Our next remark collects the main points.

Remark 1.6.23. Starting with the generic σ-Gröbner prebasis $\{g_1^*,\dots,g_\eta^*\}$, one can construct an affine subscheme of $\mathbb{A}^{s(c\mathcal{O},\sigma)}$ in the following way. As in the Buchberger Algorithm, one reduces the critical pairs of the leading terms of the σ-Gröbner prebasis as much as possible. The reduction stops when a polynomial is obtained which is a linear combination of the elements in $\mathcal{O}$ with coefficients in $K[S_{c\mathcal{O},\sigma}]$. Collecting all coefficients obtained in this way for all the critical pairs, one gets a set of polynomials which generates an ideal J in $K[S_{c\mathcal{O},\sigma}]$. Clearly, every zero of J gives rise to a specialization of the generic σ-Gröbner prebasis which is, by construction, the reduced σ-Gröbner basis of a zero-dimensional ideal I in P for which we have $\mathscr{O} = \mathscr{O}_\sigma(I)$.

However, this procedure is not canonical, since for instance the choice of the critical pairs to be reduced and the order of the reduction steps is not fixed. Based on the construction presented in this subsection, one can show that all possible ideals J define *the same scheme*, namely the one defined in Definition 1.6.16.

Another interesting problem is to look for conditions under which the two schemes $G_{\mathscr{O},\sigma}$ and $B_{\mathscr{O}}$ are isomorphic. Proposition 3.11 of [28] yields a partial answer. Essentially, it says the following.

Proposition 1.6.24. *Let $\mathscr{O}$ be an order ideal and σ a term ordering on $\mathbb{T}^n$, and assume that the order ideal $\mathscr{O}$ is a* σ**-cornercut**, *i.e. that we have $b >_\sigma t$ for every $b \in c\mathscr{O}$ and every $t \in \mathscr{O}$. Then the canonical embedding of $K[S_{\mathscr{O},\sigma}]$ into $K[c_{11},\dots,c_{\mu\nu}]$ induces an isomorphism between $G_{\mathscr{O},\sigma}$ and $B_{\mathscr{O}}$.*

The study of border basis and Gröbner basis schemes is still in its infancy. There are many open questions, for instance whether the converse of the preceding proposition holds, or whether border basis schemes are always connected. Although our journey from oil fields to Hilbert schemes ends here, the topics we discussed offer many possibilities to continue it in various directions.

> *When you have completed 95 percent of your journey,*
> *you are only halfway there.*
> (Japanese proverb)

Acknowledgements The first and third author gratefully acknowledge the financial support of the Algebraic Oil Project provided by the Stichting Shell Research Foundation. All three authors are grateful to J. Abbott, C. Fassino and M.L. Torrente for allowing them to use their paper [2] as a basis for Section 5 above. Further thanks are due to D. Heldt and J. Limbeck for the implementation of preliminary versions of the AVI algorithm and to J. Abbott and M.L. Torrente for the implementation of the SOI algorithm which both aided the understanding of the peculiarities of the approximate world substantially. Finally, the authors thank D. Heldt and M. Popoviciu for useful discussions about the subjects of this paper.

References

1. J. Abbott, C. Fassino, and M. Torrente, Thinning out redundant empirical data, Math. Comput. Sci. **1** (2007), 375–392.
2. J. Abbott, C. Fassino, and M. Torrente, Stable border bases for ideals of points, J. Symb. Comput. (to appear)

3. The ApCoCoA Team, ApCoCoA: Approximate Computations in Commutative Algebra, available at `http://www.apcocoa.org`.
4. W. Auzinger and H. Stetter, An elimination algorithm for the computation of all zeros of a system of multivariate polynomial equations, in: R.G. Agarwal, Y.M. Chow, S.J. Wilson (eds.), Int. Conf. on Numerical Mathematics, Singapore 1988, Birkhäuser ISNM **86**, Basel 1988, 11–30.
5. W. Auzinger and H. Stetter, A study of numerical elimination for the solution of multivariate polynomial systems, Technical Report, TU Wien 1989.
6. B. Buchberger and H. M. Möller, The construction of multivariate polynomials with preassigned zeros, in: J. Calmet (ed.), Proceedings of EUROCAM'82, Lect. Notes in Comp. Sci. **144**, Springer, Heidelberg 1982, 24–31.
7. A. Conca and G. Valla, Canonical Hilbert-Burch matrices for ideals of $k[x,y]$, preprint available at `arXiv:math\0708.3576`.
8. The CoCoA Team, CoCoA: a system for doing Computations in Commutative Algebra, available at `http://cocoa.dima.unige.it`.
9. D. Cox, Solving equations via algebras, in: A. Dickenstein and I. Emiris (eds.), Solving Polynomial Equations, Springer, Berlin 2005.
10. C. Fassino, Almost vanishing polynomials for sets of limited precision points, preprint available at `arXiv:math\0807.3412`.
11. T.S. Gustavsen, D. Laksov and R.M. Skjelnes, An elementary, explicit proof of the existence of Hilbert schemes of points, preprint available at `arXiv:math\0506.161v1`.
12. G.H. Golub and C.F. van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore 1989.
13. M. Haiman, q,t-Catalan numbers and the Hilbert scheme, Discr. Math. **193** (1998), 201–224.
14. D. Heldt, M. Kreuzer, S. Pokutta, and H. Poulisse, Approximate computation of zero-dimensional polynomial ideals, J. Symb. Comput. (to appear)
15. M. Huibregtse, A description of certain affine open schemes that form an open covering of $\mathrm{Hilb}^n_{\mathbb{A}^2_k}$, Pacific J. Math. **204** (2002), 97–143.
16. M. Huibregtse, An elementary construction of the multigraded Hilbert scheme of points, Pacific J. Math. **223** (2006), 269–315.
17. M. Kreuzer and A. Kehrein, Characterizations of border bases, J. Pure Appl. Alg. **196** (2005), 251–270.
18. M. Kreuzer and A. Kehrein, Computing border bases, J. Pure Appl. Alg. **205** (2006), 279–295.
19. M. Kreuzer and H. Poulisse, Computing approximate border bases, in preparation (2008)
20. M. Kreuzer and L. Robbiano, *Computational Commutative Algebra 1*, Springer, Heidelberg 2000.
21. M. Kreuzer and L. Robbiano, *Computational Commutative Algebra 2*, Springer, Heidelberg 2005.
22. M. Kreuzer and L. Robbiano, Deformations of border bases, Coll. Math. **59** (2008), 275–297.
23. A. Iarrobino, Reducibility of the families of 0-dimensional schemes on a variety, Invent. Math. **15** (1972), 72–77.
24. E. Miller and B. Sturmfels, *Combinatorial Commutative Algebra*, Springer, New York 2005.
25. H.M. Möller and T. Sauer, H-bases II: applications to numerical problems, in: A. Cohen, C. Rabut, and L.L. Schumaker (eds.), *Curve and Surface Fitting*, Vanderbilt Univ. Press, Nashville 2000, 1–10.
26. B. Mourrain, A new criterion for normal form algorithms, AAECC Lect. Notes Comp. Sci. **1719** (1999), 430–443.
27. R. Notari and M.L. Spreafico, A stratification of Hilbert schemes by initial ideals and applications, Manus. Math. **101** (2000), 429–448.
28. L. Robbiano, On border basis and Gröbner basis schemes, Coll. Math. **60** (2008), to appear.
29. H. Stetter, "Approximate Commutative Algebra" - an ill-chosen name for an important discipline, ACM Commun. Comp. Alg. **40** (2006), 79–81.
30. H. Stetter, *Numerical Polynomial Algebra*, SIAM, Philadelphia 2004.

Chapter 2
Numerical Decomposition of the Rank-Deficiency Set of a Matrix of Multivariate Polynomials

Daniel J. Bates, Jonathan D. Hauenstein, Chris Peterson, and Andrew J. Sommese

Abstract Let A be a matrix whose entries are algebraic functions defined on a reduced quasi-projective algebraic set X, *e.g.* multivariate polynomials defined on $X := \mathbb{C}^N$. The sets $\mathscr{S}_k(A)$, consisting of $x \in X$ where the rank of the matrix function $A(x)$ is at most k, arise in a variety of contexts: for example, in the description of both the singular locus of an algebraic set and its fine structure; in the description of the degeneracy locus of maps between algebraic sets; and in the computation of the irreducible decomposition of the support of coherent algebraic sheaves, *e.g.* supports of finite modules over polynomial rings. In this article we present a numerical algorithm to compute the sets $\mathscr{S}_k(A)$ efficiently.

Keywords: rank deficiency, matrix of polynomials, homotopy continuation, irreducible components, numerical algebraic geometry, polynomial system, Grassmannians.

Daniel J. Bates
Department of Mathematics, Colorado State University, Fort Collins, CO 80523, e-mail: bates@math.colostate.edu This author was supported by Colorado State University and the Institute for Mathematics and Its Applications (IMA)

Jonathan D. Hauenstein
Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556, e-mail: jhauenst@nd.edu This author was supported by the Duncan Chair of the University of Notre Dame, the University of Notre Dame Center for Applied Mathematics, NSF grant DMS-0410047, and NSF grant DMS-0712910

Chris Peterson
Department of Mathematics, Colorado State University, Fort Collins, CO 80523, e-mail: peterson@math.colostate.edu This author was supported by Colorado State University, NSF grant MSPA-MCS-0434351, AFOSR-FA9550-08-1-0166, and the Institute for Mathematics and Its Applications (IMA)

Andrew J. Sommese
Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556, e-mail: sommese@nd.edu This author was supported by the Duncan Chair of the University of Notre Dame, NSF grant DMS-0410047, and NSF grant DMS-0712910

L. Robbiano, J. Abbott (eds.), *Approximate Commutative Algebra*, Texts and Monographs in Symbolic Computation, DOI 10.1007/978-3-211-99314-9_2, © Springer-Verlag/Wien 2009

AMS Subject Classification. 13CXX, 14Q15, 14M15, 15A54, 65H10, 65H20, 65E05

Introduction

Let A be an $m \times n$ matrix with polynomial entries, *i.e.*

$$A(x) := \begin{bmatrix} p_{1,1}(x) & \cdots & p_{1,n}(x) \\ \vdots & \ddots & \vdots \\ p_{m,1}(x) & \cdots & p_{m,n}(x) \end{bmatrix} \tag{2.1}$$

with each $p_{i,j}(x) \in \mathbb{C}[x_1, x_2, \ldots, x_N]$ and where $x := (x_1, x_2, \ldots, x_N) \in \mathbb{C}^N$.

The main contribution of this article is an efficient numerical algorithm to decompose the algebraic sets

$$\mathscr{S}_k(A) := \left\{x \in \mathbb{C}^N \mid \operatorname{rank} A(x) \leq k\right\} \tag{2.2}$$

for each k from 0 to $\min\{m,n\}$. By computing the sets $\mathscr{S}_k(A)$, we also compute the algebraic subsets of $\mathbb{C}^N$ where the rank of $A(x)$ equals k, *i.e.* $\mathscr{S}_k(A) \setminus \mathscr{S}_{k-1}(A)$. By taking adjoints and relabeling if necessary, we may assume that $m \geq n$. By convention, $\mathscr{S}_{-1}(A) := \emptyset$.

We work in the general setting of finding the irreducible decompositions of sets of the form $\mathscr{S}_k(A_{V(f)})$, where f is a system of polynomials defined on $\mathbb{C}^N$; $V(f)$ denotes the common zeros of f; and $A_{V(f)}$ denotes the restriction of the matrix of polynomials in Eq. 2.1 to $V(f)$. One advantage of this generality is that many related problems may be restated in this context. For example, given a matrix $\widehat{A}(x)$ of homogeneous polynomials on $\mathbb{P}^N$ with degrees of entries compatible with the rank of $\widehat{A}(x)$ being well defined for each $x \in \mathbb{P}^N$, the irreducible components of $\mathscr{S}_k(\widehat{A})$ may be computed by regarding $\widehat{A}$ as a matrix of polynomials on $\mathbb{C}^{N+1}$ with $f(x)$ a single linear equation on $\mathbb{C}^{N+1}$ having general coefficients.

In §2.1, we present background material. Besides reviewing the general setup of Numerical Algebraic Geometry, we highlight several results we will use in the article. In §2.2 we give a description of random coordinate patches on Grassmannians. This is a useful generalization of random coordinate patches for projective space [18], see also [30, §3.7]. The generalization applies more broadly to rational homogeneous manifolds.

The strategy of the algorithm presented in §2.3 is to work with the system

$$\begin{bmatrix} f(x) \\ A(x)\cdot\xi \end{bmatrix} = 0, \tag{2.3}$$

where

$$f(x) := \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix} \tag{2.4}$$

is a system of polynomials on $\mathbb{C}^N$ and where

$$A(x)\cdot\xi \tag{2.5}$$

is a parametrized family of ξ-linear equations with

$$\xi := \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} \in \mathbb{P}^{n-1}.$$

This is a standard construct which has been used multiple times in numerical algebraic geometry, though in cases where much weaker information was sought. For example, systems which include terms of the form $A(x)\cdot\xi = 0$, have occurred for deflation [15] and [30]; and for the degeneracy set of a map from a curve to $\mathbb{C}$, [16] and [2].

Using Eq. 2.5, one can compute the components of the set $\mathscr{S}_{n-1}(A)$ as the images of the irreducible components of the reduced solution set of Eq. 2.3: this is straightforward using the numerical irreducible decomposition available in Bertini [1] or PHC [32]. The computation of the components of the remaining $\mathscr{S}_k(A)$ is more subtle.

A natural approach to computing the structure of the sets $\mathscr{S}_k(A)$ would be to decompose the projection map $V(A(x)\cdot\xi) \to \mathbb{C}^N$ into sets of constant dimension. This can be done using fiber products [31]. However, since the fibers of the map $V(A(x)\cdot\xi) \to \mathbb{C}^N$ are linear subspaces of $\mathbb{C}^N$, it is natural to use Grassmannians to parametrize fibers of a given dimension. This leads to a considerably simpler and more efficient algorithm than the general fiber product approach.

Let $\text{Grass}(a,b)$ be the Grassmannian of a-dimensional vector subspaces of $\mathbb{C}^b$; and let

$$\mathbb{T}(n;k,N) := \mathbb{C}^N \times \text{Grass}(n-k,n),$$

where n denotes the number of columns of $A(x)$. We consider the algebraic subset

$$\mathscr{E}_k(A) := \{(x,y) \in \mathbb{T}(n;k,N) \mid A(x)\cdot y = 0\}.$$

Let $\pi : \mathbb{C}^N \times \text{Grass}(n-k,n) \to \mathbb{C}^N$ be the map obtained by restricting the product projection from $\mathbb{C}^N \times \text{Grass}(n-k,n)$ to $\mathbb{C}^N$. The irreducible components of $\mathscr{S}_k(A)$

that are not irreducible components of $\mathscr{S}_{k-1}(A)$ are precisely the images under π of irreducible components of $\mathscr{E}_k(A)$ on which π is generically one-to-one. Thus, the problem reduces to setting up a polynomial system to compute $\mathscr{E}_k(A)$.

To set up this system we construct a coordinate system on a Zariski open set $U \subset \text{Grass}(n-k,n)$ such that every irreducible component of $\mathscr{S}_k(A)$ that is not an irreducible component of $\mathscr{S}_{k-1}(A)$ is the closure of the image of an irreducible component of $\mathscr{E}_k(A)$ under the product projection π. This construction uses the random coordinate patches described in §2.2 and leads to the system

$$A(x) \cdot B \cdot \begin{bmatrix} I_{n-k} \\ \Xi \end{bmatrix} = 0, \tag{2.6}$$

where B is a generic $n \times n$ unitary matrix; I_{n-k} is the $(n-k) \times (n-k)$ identity matrix; and where

$$\Xi := \begin{bmatrix} \xi_{1,1} & \cdots & \xi_{1,n-k} \\ \vdots & \ddots & \vdots \\ \xi_{k,n-k} & \cdots & \xi_{k,n-k} \end{bmatrix},$$

is an element of $\mathbb{C}^{k \times (n-k)}$. The solution components of the reduced solution set of Eq. 2.6 give the desired decomposition of $\mathscr{E}_k(A)$.

The system

$$\begin{bmatrix} f(x) \\ A(x) \cdot B \cdot \begin{bmatrix} I_{n-k} \\ \Xi \end{bmatrix} \end{bmatrix} = 0 \tag{2.7}$$

allows us to compute the decomposition of $\mathscr{S}_k(A(x)_{V(f)})$.

In §2.4, we discuss several generalizations. For example, we may compute the decomposition of $\mathscr{S}_k(A(x)_X)$, where X is an irreducible component of $V(f)$. We also show how to deal with more general $A(x)$, *e.g.* $A(x)$ having entries that are algebraic functions on algebraic sets, or when $A(x)$ is a map between vector bundles.

In §2.5, we present several applications. For example, if $f(x)$ is a system of polynomials on $\mathbb{C}^N$, then applying the algorithm of §2.3 to the Jacobian

$$Jf(x) := \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \cdots & \frac{\partial f_1}{\partial x_N}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(x) & \cdots & \frac{\partial f_m}{\partial x_N}(x) \end{bmatrix} \tag{2.8}$$

computes the decomposition of the singular set of the solution set $f^{-1}(0)$ of f. Note in this case that $n = N$. We use $f^{-1}(0)$ to denote the solution set of $f(x) = 0$ with its induced scheme-theoretic structure, *i.e.* the scheme-theoretic fiber of $f : \mathbb{C}^N \to \mathbb{C}^m$. If a component Z of $V(f)$ occurs with multiplicity at least two, then Z is contained in the singular set of $f^{-1}(0)$.

In §2.6, we give implementation details and computational results in the context of several specific examples.

In appendix 2.7, we show how to compute the singular set, $\mathrm{Sing}(V(f))$, of the reduced algebraic set $V(f)$, *i.e.* of the solution set of the radical of the ideal generated by $f(x)$. We first recall in §2.7.1 that given an irreducible component Z of a solution set $V(f)$ of a system of polynomials, there is a classical prescription, *e.g.* given in [19], to construct a system of polynomials g with $Z = V(g)$. Then in §2.7.2, a modified version of this construction is combined with the algorithm in §2.3 to give an algorithm to compute the singular set of $V(f)$.

2.1 Background Material

We work over the complex numbers. By an algebraic set we mean a possibly non-reduced quasi-projective algebraic set X. The reduction of X is denoted by X_{red}.

Let f be a system of polynomials

$$f(x) := \begin{bmatrix} f_1(x) \\ \vdots \\ f_{N-k}(x) \end{bmatrix} = 0 \tag{2.9}$$

on $\mathbb{C}^N$. By $f^{-1}(0)$ we denote the solution set of f with its possibly non-reduced structure. By $V(f)$ we denote the reduced algebraic set $f^{-1}(0)_{\mathrm{red}}$.

An algebraic set X is irreducible if it has no embedded components and if the smooth points of its reduction X_{red} are connected. By the dimension of X, we mean the maximum of the dimensions of the connected components of the set of smooth points of X_{red}. We say that an algebraic set is *pure dimensional* if it has no embedded components and if the connected components of the set of smooth points of X_{red} all have the same dimension. Given a function $G : A \to B$ between sets, we denote the restriction of G to a subset $C \subset A$ by G_C.

The main approach of Numerical Algebraic Geometry is to use intersections with generic affine linear spaces to reduce problems about positive dimensional algebraic sets to finding isolated solutions by homotopy continuation.

The use of intersections with affine linear spaces has been a standard tool in algebraic geometry for well over a century, *e.g.* see [3]. Affine slices, *e.g.* lifting fibers, have been used in related ways in symbolic work, *e.g.* [9, 10, 14]. For further discussion of this, see [29, §2.3]. For background and a detailed description of the tools for Numerical Algebraic Geometry, see [30].

2.1.1 Genericity and Randomness

A major source of efficiency in Numerical Algebraic Geometry is the use of randomness. Typically, there is an irreducible algebraic set Q, which serves as a parameter space, and some property $\mathscr{P}$ for an object corresponding to a point in Q. We need to choose a point $q \in Q$ for which $\mathscr{P}$ is true, though it might not hold for all parameter values. We say the property $\mathscr{P}$ holds generically if it is true for a nonempty Zariski open set $\mathscr{U}$ of Q.

For example, the polynomial $f(x,q) := qx-1$ may be regarded as a family of polynomials in the variable x with parameter q in the parameter space $Q := \mathbb{C}$. The property that "$f(x,q) = 0$ has a solution" is true except when $q = 0$. Thus, this property holds generically.

Some algorithms depend on choosing $q \in \mathscr{U}$. We typically do this using a random number generator, and refer to the object depending on the parameter chosen, *e.g.* a coordinate patch, as random or generic, *e.g.* a random coordinate patch. If our random number generator truly determines a random complex number, the probability of choosing a point in Q for which the property and the algorithm fails would be zero, and hence such algorithms are called *probability-one algorithms*. Of course, the numbers available on a computer are finite, but with error checking and use of high precision arithmetic, such algorithms may be designed to work very well. These matters are discussed further in [30].

The parameter spaces Q which we use are usually defined over $\mathbb{C}$, but there are times when we restrict our choice of a random point $q \in Q$ to lie in a special subset. For example, we might have a property $\mathscr{P}$ that holds generically for points in the parameter space $\mathbb{C}^* := \mathbb{C} \setminus \{0\}$, but for reasons of numerical stability, we might prefer to choose q to be of absolute value 1. Since the subset of $S^1 := \{q \in \mathbb{C}^* \mid |q| = 1\}$, for which $\mathscr{P}$ fails, is closed and has Lebesgue measure zero, choosing q randomly from S^1 is justified. A slight generalization of this situation, which occurs in this article, is when the parameter space Q is $\mathrm{GL}(\mathrm{n},\mathbb{C})$, Here, for reasons of numerical stability, we choose $q \in \mathrm{U}(\mathrm{n})$, the unitary group acting on $\mathbb{C}^n$. Since the intersection of $\mathrm{U}(\mathrm{n})$ with any proper algebraic subset X of $\mathrm{GL}(\mathrm{n},\mathbb{C})$, is a closed set of Lebesgue measure zero, choosing q randomly from $\mathrm{U}(\mathrm{n})$ is justified.

Remark 2.1.1. Note that if a complex semisimple group G is our parameter space, as it is in Remark 2.2.2, we could, for similar reasons, choose q randomly in a maximal compact subgroup of G.

2.1.2 The Numerical Irreducible Decomposition

Given a system of polynomials $f(x)$ as in Eq. 2.9, the *irreducible decomposition* of $V(f)$ is the decomposition

$$V(f) = \cup_{i=1}^{\dim V(f)} Z_i = \cup_{i=1}^{\dim V(f)} \cup_{j \in \mathscr{I}_i} Z_{i,j},$$

where Z_i is a pure i-dimensional algebraic set; each set $\mathscr{I}_i$ is finite; and the $Z_{i,j}$ are irreducible algebraic sets with the property that $Z_{i,j} \subset Z_{a,b}$ if and only if $(i,j) = (a,b)$. The *Numerical Irreducible Decomposition* of $V(f)$ is the collection of

1. linear equations $L_1(x), \ldots, L_n(x)$ general with respect to all the $Z_{i,j}$;

2. the sets $W_{i,j}$ consisting of the $\deg Z_{i,j}$ smooth points $V(L_1,\ldots,L_i)\cap Z_{i,j}$ of $Z_{i,j}$ for each non-negative integer $i \leq \dim V(f)$ and each $j \in \mathscr{I}_i$.

The elements of the $W_{i,j}$ are called witness points for $Z_{i,j}$. This decomposition is developed in [23, 24, 25, 26]. See also [30]. The programs Bertini [1] and PHC [27, 32] compute this decomposition. As an algorithm, we have

Algorithm *NumIrredDecomp*

```
Input: A system of polynomials $\{f_1(x),\ldots,f_m(x)\}$ on $\mathbb{C}^N$.
Output: The dimension $d = \dim V(f)$;
    the dimensions $\dim_{i,j}$ and degrees $\deg_{i,j}$ of the irreducible components
            $Z_{i,j}$ of $V(f)$ for $j \in \mathscr{I}_i$ with non-negative integers $i \leq d$;
    linear equations $L_1(x),\ldots,L_n(x)$ general with respect to all the $Z_{i,j}$; and
    witness sets $W_{i,j}$ consisting of the $\deg Z_{i,j}$ smooth points $V(L_1,\ldots,L_i)\cap Z_{i,j}$
            of $Z_{i,j}$ for each non-negative integer $i \leq \dim V(f)$ and each $j \in \mathscr{I}_i$.
```

The algorithm as implemented also outputs auxiliary information needed for further computation, *e.g.* deflated systems, as discussed following Algorithm 1, which is needed to track paths on components whose multiplicity is greater than one.

By varying the linear equations, it is computationally inexpensive to generate additional points on each $Z_{i,j}$.

The membership test from [24, 25] gives a computation of the multiplicity of a point on the reduction of an irreducible algebraic set. Since a smooth point is precisely one of multiplicity one, this gives a criterion for a point to be a smooth point of the reduction of an algebraic set. Since we need this criterion in this article, let us state it as an algorithm.

Algorithm 1. *CheckSmoothness*

```
Input: A system of polynomials $\{f_1(x),\ldots,f_m(x)\}$ on $\mathbb{C}^N$;
and a point $x^*$ on an irreducible component $Z$ of $V(f)$.
Output: The multiplicity $\mu$ of $x^*$ on $Z_{\mathrm{red}}$.
    Compute a set $W$ of witness points $w_1,\ldots,w_{\deg Z_{\mathrm{red}}}$ for linear equations
            $L := \{L_1,\ldots,L_{\dim Z}\}$ general with respect to $Z$.
    Choose a system of linear equations $\widehat{L} := \{\widehat{L}_1,\ldots,\widehat{L}_{\dim Z}\}$ satisfying
            $\widehat{L}_i(x^*) = 0$ and which other than this are general with
            respect to the choices made in previous steps.
    Choose a general complex number $\gamma$ satisfying $|\gamma| = 1$.
    Compute the limits $\widehat{W} := \{\widehat{w}_1,\ldots,\widehat{w}_{\deg Z_{\mathrm{red}}}\}$ of the paths
            $Z \cap V(tL(x) + \gamma(1-t)\widehat{L}(x))$ starting at $W$ and traced
            as $t$ goes from $1$ to $0$.
    Set $\mu$ equal to the number of points of $\widehat{W}$ equal to $x^*$.
```

There are a number of numerical issues that must be dealt with in implementations of this algorithm. If Z is not generically reduced, then tracking must be done using a deflated system. Deflation for isolated points was developed in [15]: see also [8, 14, 20, 21]. For the deflation of irreducible components and the use of it for tracking paths see [30, §13.3.2 and §15.2.2]. Another numerical issue is how to decide equality. These matters are discussed in [30].

We need to compute generic fiber dimensions in the main algorithm of this article; the following suffices.

Algorithm *FiberDimension*
Input: A system of polynomials $\{f_1(x),\dots,f_m(x)\}$ on $\mathbb{C}^N$;
a polynomial map $\Phi:\mathbb{C}^N\to\mathbb{C}^s$ with $\Phi(x)=(\Phi_1(x),\dots,\Phi_s(x))$;
and a point $x^*\in V(f)$.
Output: The dimension at x^* of the fiber of $\Phi_{V(f)}$ containing x^*.

This is a simple consequence of having *NumIrredDecomp*. First compute

$$\mathit{NumIrredDecomp}(f_1(x),\dots,f_m(x),\Phi_1(x)-\Phi_1(x^*),\dots,\Phi_s(x)-\Phi_s(x^*)).$$

Now do a membership test using the witness sets from this computation to find which components of the Irreducible Decomposition of

$$V(f_1(x),\dots,f_m(x),\Phi_1(x)-\Phi_1(x^*),\dots,\Phi_s(x)-\Phi_s(x^*))$$

contain x^*. The maximum of the dimensions $\dim_{i,j}$ among these components gives the desired dimension.

2.1.3 Images of Algebraic Sets

What is required for the numerical irreducible decomposition is data that allows us to carry out homotopy continuation. Often this is a system of equations on the Euclidean space which contains the algebraic set, but this is not necessary.

With deflation of a multiple k-dimensional component Z of a system $f(x)=0$ [30, §13.3.2 and §15.2.2], we have a system of the form

$$\mathscr{D}(f,Z):=\begin{bmatrix} f(x)\\ A(x)\cdot\begin{bmatrix}1\\ \xi\end{bmatrix}\\ c\cdot\xi-1\end{bmatrix}=0$$

with $x\in\mathbb{C}^N$; $\xi\in\mathbb{C}^n$; $A(x)$ is a $s\times(n+1)$ matrix of polynomials; and c a $(n+N-s-k)\times n$ matrix of constants. The key property of the system $\mathscr{D}(f,Z)$ is that there is a multiplicity one component Z' of $\mathscr{D}(f,Z)^{-1}(0)$ which maps generically one-to-one onto Z under the product projection $(x,\xi)\to x$. To carry out operations such as tracking a path on Z as a complementary linear space $\mathscr{L}$ moves, it suffices to track the path cut out on Z' as the pullback of $\mathscr{L}$ to $\mathbb{C}^{N+n}$ moves.

Similarly, assume that we have a system of polynomials

$$\mathscr{F}(x_1,\ldots,x_N;\xi_1,\ldots,\xi_n)=0$$

on $\mathbb{C}^{N+n}$. Let $\pi:\mathbb{C}^{n+r}\to\mathbb{C}^r$ denote the product projection. Let Z be an irreducible component of $V(\mathscr{F})$. Given this information, it is straightforward to numerically work with the closure $\overline{\pi(Z)}$ of the image of Z by lifting computations to Z.

This is the special situation that occurs in this article, assuming that each fiber of π_Z is a linear space but with possibly varying dimensions. Using *FiberDimension*, compute the dimension v of the fiber of π_Z containing a witness point of Z. Now choose v general linear equations $L_1,\ldots,L_v$ in the x variables. There is a unique component Z' of $Z\cap V(L_1,\ldots,L_v)$ which maps generically one-to-one onto $\overline{\pi(Z)}$. For numerical continuation this is as good as having the equations for $\overline{\pi(Z)}$.

In several algorithms we will manipulate irreducible components of an algebraic set. Numerically we always use the Numerical Irreducible Decomposition, but with the possibility that the equations are defined on an auxiliary space as above.

2.2 Random Coordinate Patches on Grassmannians

The Grassmannian $\mathrm{Grass}(a,b)$ parametrizes all a-dimensional vector subspaces of $\mathbb{C}^b$. When $a=1$, this is the usual $(b-1)$-dimensional projective space, $\mathbb{P}^{b-1}$. An a-dimensional vector subspace S of $\mathbb{C}^b$ is specified uniquely by a linearly independent vectors $v_1,\ldots,v_a$ in $\mathbb{C}^b$. It is convenient to regard these as forming a $b\times a$ matrix

$$\begin{bmatrix} v_1 & \cdots & v_a \end{bmatrix}. \tag{2.10}$$

Note that if $v'_1,\ldots,v'_a$ is a second basis of S, then there is an invertible $a\times a$ matrix T of complex numbers such that

$$\begin{bmatrix} v'_1 & \cdots & v'_a \end{bmatrix} = \begin{bmatrix} v_1 & \cdots & v_a \end{bmatrix}\cdot T.$$

$\mathrm{Grass}(a,b)$ is an $a(b-a)$-dimensional projective manifold on which the group $\mathrm{GL(b)}$ of invertible $b\times b$ matrices g acts homogeneously under the action

$$\begin{bmatrix} v_1 & \cdots & v_a \end{bmatrix} \to g\cdot\begin{bmatrix} v_1 & \cdots & v_a \end{bmatrix}.$$

More details on Grassmannians may be found in [12, 13].

Random Coordinate Patches

A basic numerical trick, first exploited in [18] to find isolated solutions in $\mathbb{C}^N$ or $\mathbb{P}^N$, is to carry out all computation on a random Euclidean coordinate patch on $\mathbb{P}^N$. The advantage of this trick is that, with probability one, all solutions of the system are now finite, and so for the purposes of computation, points at infinity may be treated as finite (albeit such points are often highly singular). Though no patch can contain a positive dimensional solution component at infinity, a general coordinate patch meets every irreducible component of a solution set in a Zariski open set of the given component. For this reason, this same trick is widely used in Numerical Algebraic Geometry [30].

In this article, we have need of a random patch on a Grassmannian $\mathrm{Grass}(a,b)$ of linear vector subspaces $\mathbb{C}^a \subset \mathbb{C}^b$. A straightforward generalization of the above trick is to embed $\mathrm{Grass}(a,b)$ in $P := \mathbb{P}^{\binom{b}{a}-1}$ and pullback a random patch from P. This patch is complicated to work with because of the number of variables involved and because of the non-linearity of the conditions for a point to be on the patch.

There is a much better way to choose a random patch, which is particularly efficient for numerical computation. We present the approach and justification for choosing the patch in the following paragraphs.

Let B be a $b \times b$ unitary matrix. Then for a coordinate patch we take

$$B \cdot \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \\ \xi_{1,1} & \cdots & \xi_{1,a} \\ \vdots & \ddots & \vdots \\ \xi_{b-a,1} & \cdots & \xi_{b-a,a} \end{bmatrix}. \tag{2.11}$$

We often abbreviate this as

$$B \cdot \begin{bmatrix} I_a \\ \Xi \end{bmatrix}$$

where I_a denotes the $a \times a$ identity matrix.

Theorem 2.2.1. *Let W be an arbitrary algebraic set and let $\mathscr{T}$ denote an algebraic subset of $W \times \mathrm{Grass}(a,b)$. Given a unitary matrix B, let U_B denote the Zariski open set of* $\mathrm{Grass}(a,b)$ *for which*

$$B \cdot \begin{bmatrix} I_a \\ \Xi \end{bmatrix}$$

are coordinates. There is an open dense subset $\mathscr{U}$ of the unitary matrices $U(n)$ such that the Lebesgue measure of $U(n) \setminus \mathscr{U}$ is zero and such that for $B \in \mathscr{U}$, $(W \times U_B) \cap \mathscr{T}$ is a non-empty Zariski open subset of $\mathscr{T}$.

Proof. By the discussion in §2.1.1, it suffices to show this for generic B in the general linear group, $\mathrm{GL(n,\mathbb{C})}$. For the closure of $(W \times U_B) \cap \mathscr{T}$ to not contain a component C of $\mathscr{T}$ is an algebraic condition, *i.e.* a condition picking out an algebraic subset of the General linear group. Let D_C denote this algebraic subset of $\mathrm{GL}(\mathrm{n},\mathbb{C})$. The set, D_C, is a proper subset due to the fact that U_B may be chosen to contain any particular point of $\mathrm{Grass}(a,b)$. Let $\mathscr{C}$ denote the set of components of $\mathscr{T}$. Since $\mathscr{T}$ has finitely many components, any invertible matrix B in the complement of $\cup_{C\in\mathscr{C}} D_C$ will suffice.

Remark 2.2.2. Let $\mathscr{X}$ be a rational homogeneous projective manifold. The Borel-Remmert characterization [5, 22] of such manifolds is that they are the compact Kähler manifolds $\mathscr{X}$ with a complex semisimple group G as biholomorphism group, and such that there is a parabolic subgroup P of G with $\mathscr{X}$ biholomorphic to G/P with its natural complex structure. Thus $\mathrm{Aut}(\mathscr{X})$ possesses a conjugation with fixed points a maximal compact subgroup K with $\dim_{\mathbb{R}} K = \dim_{\mathbb{C}} \mathrm{Aut}(\mathscr{X})$. The analogous notion of a random coordinate patch would be the set gU, where g is a general element of K and U is any dense Bruhat Cell. More details on parabolic subgroups and Bruhat decompositions may be found in [6, §3 and §11]

2.3 Finding Rank-Dropping Sets

Let A be an $m \times n$ matrix with polynomial entries as in Eq. 2.1 and let $f(x)$ denote the system as in Eq. 2.9. By taking adjoints and relabeling if necessary, we may assume without loss of generality that $m \geq n$. Let $\mathscr{S}_k(A) := \left\{x \in \mathbb{C}^N \mid \mathrm{rank}\, A(x) \leq k\right\}$. Since the rank of $A(x)$ can be at most n, we may restrict ourselves to computing $\mathscr{S}_k(A)$ for $k \leq n$.

For each k, $\mathscr{S}_k(A_{V(f)})$ is algebraic since it is the solution set of the system comprised of $f_1(x),\ldots,f_n(x)$ plus the determinants of all $k \times k$ subminors of A. The irreducible components Z of $\mathscr{S}_k(A)$, with the property that $\mathrm{rank}\, A(x^*) = k$ for a general point $x^* \in Z$, are precisely the irreducible components of $\mathscr{S}_k(A)$, which are not components of $\mathscr{S}_{k-1}(A)$.

The sets $S_k(A_{V(f)})$ may theoretically be computed via Gröbner basis techniques by solving each of these systems with software such as CoCoA, Macaulay, or Singular [7, 17, 11]. However, for many applications, the system of determinants of all $k \times k$ subminors of A is impractically large and complex. Such systems consist of $\binom{m}{k}\binom{n}{k}$ equations with degrees considerably larger than those of the entries of A. As a result, this approach will only work when both the size of A and the degrees of the entries of A are relatively small. We follow an alternative approach.

Our starting point is the system

$$\begin{bmatrix} p_{1,1}(x) & \cdots & p_{1,n}(x) \\ \vdots & \ddots & \vdots \\ p_{m,1}(x) & \cdots & p_{m,n}(x) \end{bmatrix} \cdot \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix} = 0 \tag{2.12}$$

where $[\xi_1,\dots,\xi_n]$ is a set of homogeneous coordinates on $\mathbb{P}^{n-1}$. We let $\mathscr{T}(A)$ denote the solution set of Eq. 2.12. Let $\pi : \mathscr{T}(A) \to \mathbb{C}^N$ and $\sigma : \mathscr{T}(A) \to \mathbb{P}^{n-1}$ denote the maps induced by the product projections of $\mathbb{C}^N \times \mathbb{P}^{n-1}$.

We let $\mathscr{T}(A)_y \subset \mathbb{P}^{n-1}$ denote the solution set of the fiber of $\mathscr{T}(A)$ over y regarded as a subset of $\mathbb{P}^{n-1}$, *i.e.* $\mathscr{T}(A)_y = \sigma(\pi^{-1}(y))$. In this setting, we have

$$y \in \mathscr{S}_k(A) \quad \text{if and only if} \quad \dim \mathscr{T}(A)_y \geq n-1-k.$$

While computing the irreducible decomposition of $\mathscr{T}(A)$ provides a great deal of information, it does not allow for the full determination of the sets

$$\mathscr{S}_k(A) = \{y \in \mathbb{C}^N \,|\, \dim \mathscr{T}(A)_y \geq n-1-k\}.$$

One could completely determine these sets by applying the fiber product algorithm as developed in [31]. However, since we want to find fibers of π that are points in $\mathrm{Grass}(n-k,n)$, there is a different approach which has the advantage of being computationally more efficient.

The approach is to consider the system

$$\mathscr{F}_k(f) := \begin{bmatrix} f(x) \\ A(x)\cdot B \cdot \begin{bmatrix} I_{n-k} \\ \Xi \end{bmatrix} \end{bmatrix} = 0 \tag{2.13}$$

where B is a random $n \times n$ unitary matrix; I_{n-k} is the $(n-k)\times(n-k)$ identity matrix; and Ξ is an $k \times (n-k)$ matrix of indeterminates $\xi_{i,j}$. The discussion in §2.2 gives the following result.

Theorem 2.3.1. *Let A be an $m \times n$ matrix with polynomial entries in Eq. 2.1 and let $f(x)$ denote the system as in Eq. 2.9. Assume that $m \geq n$. For a generic B in the $n \times n$ unitary group, and a non-negative integer k between 0 and n, let $\mathscr{F}_k(f)$ denote the system in Eq. 2.13. Let $\mathscr{A}_k$ denote the set of the irreducible components of $\mathscr{S}_k(A_{V(f)})$, which are not irreducible components of $\mathscr{S}_{k-1}(A_{V(f)})$. Let $\mathscr{B}_k$ denote the set made up of the closures of the images under π of the irreducible components Z of $V(\mathscr{F}_k(f))$, such that the generic fiber of the projection from Z to $\mathbb{C}^N$ is zero dimensional, i.e. such that $A(x)$ is of rank k at the image under π of a witness point of Z. The sets in $\mathscr{A}_k$ are the maximal elements under set inclusion of the elements of $\mathscr{B}_k$.*

Note we can use a membership test to determine inclusion relations.

Algorithm 2. *RankDropSet*

Input: A system of polynomials $\{f_1(x),\dots,f_m(x)\}$ on $\mathbb{C}^N$;

a matrix of polynomials $A(x) := \begin{bmatrix} p_{1,1}(x) & \cdots & p_{1,n}(x) \\ \vdots & \ddots & \vdots \\ p_{m,1}(x) & \cdots & p_{m,n}(x) \end{bmatrix}$ on $\mathbb{C}^N$; and

a non-negative integer k.

```
Output: The Numerical Irreducible Decomposition of the set $\mathscr{S}_k(A_{V(f)})$.
    Choose a random unitary matrix $B \in U(n)$.
    Compute NumIrredDecomp of the system in Eq. 2.13.
    Use FiberDimension to pick out the set $\mathscr{Z}$ of irreducible components $Z$
            with fibers having generic fiber dimension zero.
    Output the projection of the components $Z \in \mathscr{Z}$ in $\mathbb{C}^N \times \mathbb{C}^{k(n-k)}$ to
            $\mathbb{C}^N$ under the product projection.
```

In line with the discussion of §2.1.3, we have not explicitly given the details of the standard steps to compute the full witness sets for the components output in the last line of the algorithm. For the convenience of the reader, we give a brief discussion of steps involved.

Fix an a-dimensional component $Z \in \mathscr{Z}$ with fiber dimension zero. Let $\widehat{\mathscr{L}}$ denote the generic a-codimensional affine linear subspace of $\mathbb{C}^{N+k(n-k)}$ with $W_Z := \widehat{\mathscr{L}} \cap Z$ the witness points of Z.

Under the product projection, $\mathbb{C}^N \times \mathbb{C}^{k(n-k)}$, Z has, as image in $\mathbb{C}^N$, a dense constructable subset A of an algebraic subset B of $\mathbb{C}^N$. Choose a generic a-codimensional affine linear subspace $\mathscr{L} \subset \mathbb{C}^N$. A contains a dense Zariski open subset $\mathscr{O}$ of B, see for example [30, Lemma 12.5.8 and 12.5.9]. $\mathscr{L}$ meets $\mathscr{O}$ in $\deg B$ points. The witness points $\mathscr{L} \cap B$ for B may be simply computed from the known witness points W_Z of Z.

To do this, pull back $\mathscr{L}$ to an a-codimensional affine linear subspace $\mathscr{L}'$ of $\mathbb{C}^{N+k(n-k)}$. Using a homotopy deforming $\widehat{\mathscr{L}}$ to $\mathscr{L}'$, we can, starting with the points W_Z find the set of points $W' := \mathscr{L}' \cap Z$. The images in $\mathbb{C}^N$, of the points W', are the witness points $\mathscr{L} \cap B$.

2.4 Generalizations

Theorem 2.3.1 and the corresponding algorithm are stated for a matrix of polynomials on $\mathbb{C}^N$. These results hold much more generally.

Using Open Zariski Subsets

Since we work with witness points, we can replace $\mathbb{C}^N$ with a nontrivial Zariski open set U. Indeed, either

1. U meets a d-dimensional component Z of $V(f) \subset \mathbb{C}^N$ in a nontrivial Zariski open set: or
2. $U \cap Z$ is empty.

In the first case, a generic $N-d$ dimensional linear space that meets Z in $\deg Z$ points will with probability one meet $Z \cap U$ in $\deg Z$ points.

Restriction of A to an Irreducible Set

Let f and A be as in §2.3. Let X be an irreducible component of $V(f)$. It is straightforward to find $\mathscr{S}_k(A_X)$. We have found that each component Z of $\mathscr{S}_k(A_{V(f)})$ is the closure of the image of an irreducible set Z' from $V(f) \times \mathrm{Grass}(n-k,n)$ under the product projection π with general fiber dimension zero. Using [28], we can find the irreducible components Z'' of the sets $Z' \cap [X \times \mathrm{Grass}(k,N)]$. The set $\mathscr{S}_k(A_X)$ is precisely the union of the closures of the images under π of the components Z'' with general fiber dimension under π equal to zero.

Algebraic Functions instead of Polynomials

The results of §2.3 hold for the restriction of $A(x)$, made up of algebraic functions defined on a Zariski open set $\mathscr{U}$ of $\mathbb{C}^N$, to the solution set of a system of polynomials $f(x)=0$ defined on $\mathbb{C}^N$. For example, $A(x)$ is a matrix of rational functions on $\mathbb{C}^N$ and $\mathscr{U}$ is the complement on $\mathbb{C}^N$ of the union of the pole sets of the entries of $A(x)$. By clearing the denominators, we have reduced to a polynomial matrix.

Algebraic Vector Bundle Maps

Let $\mathscr{F}$ and $\mathscr{G}$ be algebraic vector bundles of ranks n and m, respectively, defined on a quasi-projective manifold Y. Let A be an element of $\mathrm{Hom}(\mathscr{F},\mathscr{G})$, *i.e.* an algebraic section of $\mathscr{F}^* \otimes_{\mathbb{C}} \mathscr{G}$. Let F be a section of an algebraic vector bundle $\mathscr{H}$ on Y and let $X \subset Y$ be an algebraic subset of the set of zeroes of F. For each integer $k \geq 0$, the set $\mathscr{S}_k(A)$ of points $y \in Y$ where $\mathrm{rank}(A) \leq k$ is an algebraic subset of Y. The set $\mathscr{S}_k(A_X)$ is also algebraic. By convention, $\mathscr{S}_{-1}(A) = \emptyset$.

We wish to decompose $\mathscr{S}_k(A_X)$ into irreducible components. Since algebraic vector bundles are locally free in the Zariski topology, this general situation may be reduced to finding $\mathscr{S}_k(A_{V(f)}) \cap U$ for a matrix A of polynomials on some Zariski open set U of $\mathbb{C}^N$ and a system of polynomials $f(x)=0$.

The only practical case of this generalization is the case where f is a system of homogeneous polynomials on $\mathbb{P}^N$ and $A(x)$ is a matrix of homogeneous polynomials whose degrees are compatible with the rank drop loci being considered to lie in $\mathbb{P}^N$.

To be explicit, let $\mathscr{O}_{\mathbb{P}^N}(k)$ denote the sheaf of algebraic sections of the k-th power of the hyperplane section bundle on $\mathbb{P}^N$, *i.e.* the rank one locally free coherent algebraic sheaf whose sections are the homogeneous polynomials of degree k. Consider an $\mathscr{O}_{\mathbb{P}^N}$ linear mapping A of the form

$$A : \bigoplus_{i=1}^{m} \mathscr{O}_{\mathbb{P}^N}(a_i) \to \bigoplus_{j=1}^{n} \mathscr{O}_{\mathbb{P}^N}(b_j). \tag{2.14}$$

A is represented by a matrix of homogeneous polynomials with $\deg A_{i,j}(x) = b_j - a_i$. For a matrix of this form, the rank of $A(x)$ is well defined for any given point in projective space. Choosing a generic Euclidean coordinate patch $U \approx \mathbb{C}^N$ on $\mathbb{P}^N$, U meets each irreducible component of each $\mathscr{S}_k(A)$ in a Zariski open dense set.

To set up the equations, regard A as a matrix of homogeneous polynomials on $\mathbb{C}^{N+1}$ and f as a system of homogeneous polynomials on $\mathbb{C}^{N+1}$. Let $c \cdot x - 1$ be a general linear equation on $\mathbb{C}^{N+1}$ with $\mathbb{P}^N \setminus U$ defined by $c \cdot x = 0$. The system for $\mathscr{S}_k(A_{V(f)})$ is

$$\begin{bmatrix} f(x) \\ A(x) \cdot B \cdot \begin{bmatrix} I_{n-k} \\ \Xi \end{bmatrix} \\ c \cdot x - 1 \end{bmatrix} = 0 \tag{2.15}$$

where we regard A as an $m \times n$ matrix of polynomials on $\mathbb{C}^{N+1}$ (thus x is a vector of $N+1$ indeterminates); B is a random $n \times n$ unitary matrix; I_{n-k} is the $(n-k) \times (n-k)$ identity matrix; Ξ is a $k \times (n-k)$ matrix of indeterminates $\xi_{i,j}$; and c is a generic unit vector on $\mathbb{C}^{N+1}$.

2.5 Applications

We present three applications of our algorithm:

1. the numerical irreducible decomposition of the support of a finite module over a ring of algebraic functions, *i.e.* of a coherent algebraic sheaf on an algebraic set;
2. the decomposition into sets where the differential of an algebraic map is of constant rank: one special case of this is the computation of the singular set of an algebraic set; and
3. the singular set of an algebraic set.

In each of the following applications we work over $\mathbb{C}^N$. Generalizations, *e.g.* to quasi-projective manifolds, follow from the ideas outlined in §2.4.

2.5.1 Support of a Module

Let $\mathscr{O}_{\text{alg},U}$ denote the sheaf of algebraic functions on a Zariski open set $U \subset \mathbb{C}^N$. A finitely generated coherent algebraic sheaf $\mathscr{F}$ is the quotient sheaf of an $\mathscr{O}_{\text{alg},U}$-linear map

$$\bigoplus_{i=1}^{n} \mathscr{O}_{\text{alg},U} \to \bigoplus_{i=1}^{m} \mathscr{O}_{\text{alg},U}.$$

Such a map is given by an $m \times n$ matrix of algebraic functions $A(x)$. Entries $A_{i,j}(x)$ of $A(x)$ are rational functions, which must be of the form $\dfrac{p_{i,j}(x)}{q_{i,j}(x)}$ for polynomials $p_{i,j}(x)$ and $q_{i,j}(x)$ with the solution set of $q_{i,j}(x)=0$ contained in $\mathbb{C}^N \setminus U$. Decomposing the support of $\mathscr{F}$ is the same as computing the sets $\mathscr{S}_k(A)$.

2.5.2 Degeneracy Sets of the Differential of a Map

Let X denote the solution set of a system of polynomials $f(x)=0$ (as in Eq. 2.9) defined on $\mathbb{C}^N$. Let $Jf(x)$ denote its Jacobian matrix as in Eq. 2.16. For simplicity, let $\pi_X : X \to \mathbb{C}^M$ be the restriction to $X \subset \mathbb{C}^N$ of a surjective linear projection from $\pi : \mathbb{C}^N \to \mathbb{C}^M$. Let $n := N-M$ and let R denote an $N \times n$ matrix of orthonormal vectors spanning the n-dimensional vector subspace of $\mathbb{C}^N$, which is the fiber of π containing the origin of $\mathbb{C}^N$. If X is irreducible and $\pi(X)$ is dense in $\mathbb{C}^M$, then the degeneracy for the map π_X is the set of points $x^* \in X$ where the rank of the $(m \times n)$-matrix

$$Jf(x^*) \cdot R$$

is less than n.

2.5.3 Singular Sets

The special case in §2.5.2 when $M=0$ is of special interest. For simplicity assume that we are trying to find the singular set of a possibly non-reduced pure k-dimensional algebraic set X defined by a system of $N-k$ polynomials on $\mathbb{C}^N$ as in Eq. 2.9, with Jacobian matrix

$$Jf(x) := \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_{N-k}}{\partial x_1} & \cdots & \frac{\partial f_{N-k}}{\partial x_N} \end{bmatrix}. \tag{2.16}$$

The singular set consists of those points $x^* \in X$ such that

$$\mathrm{rank}_{x^*} Jf(x) < N-k.$$

The results apply immediately to this situation. We codify this for use in the Appendix. For a matrix A, we denote the transpose by A^T.

Algorithm 3. *FindSchemeTheoreticSingularSet*
Input: A system of polynomials $\{f_1(x), \ldots, f_{N-k}(x)\}$ on $\mathbb{C}^N$;
with $V(f)$ pure k-dimensional.
Output: *RankDropSet*$(f, (Jf)^T, N-k-1)$.

Note that it may well turn out that the set produced by the previous algorithm is the singular set of $V(f) = f^{-1}(0)_{\text{red}}$. It is simple to check this. Given a witness point x^* on a component Z of *FindSchemeTheoreticSingularSet*(f), we have that with probability one, $Z \subset \text{Sing}(V(f))$ if either x^* is contained in more than one irreducible component of $V(f)$ or x^* is contained in a single component X of $V(f)$ and *CheckSmoothness*$(f, X, x^*) > 1$.

2.6 Implementation Details and Computational Results

The computational examples discussed here were run on an Opteron 250 processor running Linux using the numerical irreducible decomposition [23] implemented in the Bertini software package [1], which is under development by the first, second and fourth authors and Charles Wampler of GM Research and Development. Bertini v1.0 was used for the following examples.

2.6.1 Singular Set for a Matrix

Consider the matrix

$$A(a,b,c,d,e,f) = \begin{bmatrix} 0 & a & b & c \\ -a & 0 & d & e \\ -b & -d & 0 & f \\ -c & -e & -f & 0 \end{bmatrix}$$

Clearly, $\mathscr{S}_0(A) = \{(0,0,0,0,0,0)\}$. One can compute $\mathscr{S}_1(A) = \mathscr{S}_0(A)$ and

$$\mathscr{S}_2(A) = \mathscr{S}_3(A) = \{(a,b,c,d,e,f) : af + cd - be = 0\}.$$

It should be noted that $\det(A) = (af + cd - be)^2$.

Bertini identified the components numerically for $\mathscr{S}_0(A)$ in $0.03s$, $\mathscr{S}_1(A)$ in 6.47 seconds, $\mathscr{S}_2(A)$ in 5.77 seconds, and $\mathscr{S}_3(A)$ in 0.40 seconds.

2.6.2 Singular Set for a Hessian Matrix

For a given polynomial $g : \mathbb{C}^N \to \mathbb{C}$, consider computing the singular set of its Hessian matrix $H_g(x)$ where

$$H_g(x)_{ij} = \frac{\partial^2 g}{\partial x_i \partial x_j}(x).$$

In particular, consider the polynomial $g(x,y,z) = x^3 + x^2 + 2xy^2 - y^3 + 3yz^2 + z^3$ which has the Hessian matrix

$$H_g(x,y,z) = \begin{bmatrix} 6x+2 & 4y & 0 \\ 4y & 4x-6y & 6z \\ 0 & 6z & 6y+6z \end{bmatrix}.$$

By inspection, $\mathscr{S}_0(H_g) = \emptyset$. One can compute

$$\mathscr{S}_1(H_g) = \left\{(0,0,0), \left(-\frac{1}{3}, 0, -\frac{2}{9}\right), \left(-\frac{1}{3}, 0, 0\right)\right\}$$

and

$$\mathscr{S}_2(H_g) = \{(x,y,z) : \det(H_g(x,y,z)) = 0\}.$$

Bertini identified the components numerically for $\mathscr{S}_0(A)$ in 0.01 seconds, $\mathscr{S}_1(A)$ in 0.30 seconds, $\mathscr{S}_2(A)$ in 0.18 seconds.

2.6.3 Singular Solutions for a Polynomial System

Consider computing the singular solutions of the cyclic-4 system [4] given by

$$f(x_1,x_2,x_3,x_4) = \begin{bmatrix} x_1+x_2+x_3+x_4 \\ x_1x_2+x_2x_3+x_3x_4+x_4x_1 \\ x_1x_2x_3+x_2x_3x_4+x_3x_4x_1+x_4x_1x_2 \\ x_1x_2x_3x_4 - 1 \end{bmatrix}.$$

It is well-known that $V(f) = \{(x_1,x_2,x_3,x_4) : f(x_1,x_2,x_3,x_4) = 0\}$ has two irreducible quadric curve components given by $\{(x_1,x_2,-x_1,-x_2) : x_1x_2 = 1\}$ and $\{(x_1,x_2,-x_1,-x_2) : x_1x_2 = -1\}$. Denoting the Jacobian of f as Jf, for this system $\mathrm{Sing}(V(f)) = V(f) \cap \mathscr{S}_2(Jf)$ is the set of solutions of f with exceptional rank. The polynomial system that defines $\mathrm{Sing}(V(f))$ consists of 12 polynomials in 8 variables. Bertini performed a full numerical irreducible decomposition on this system in 4.45 minutes to discover that $\mathrm{Sing}(V(f))$ consists of 8 isolated points, namely

$$\mathrm{Sing}(V(f)) = \left\{\left(a, \frac{1}{a}, -a, \frac{1}{a}\right), \left(a, -\frac{1}{a}, -a, \frac{1}{a}\right) : a = \pm 1, \pm\sqrt{-1}\right\}.$$

2.7 The Singular Set of the Reduction of an Algebraic Set

Let $f := \{f_1(x), \ldots, f_m(x)\}$ denote a system of polynomials on $\mathbb{C}^N$. In this section we give an algorithm *FindSingularSet*, which starting with the input f, outputs a system of polynomials $\mathscr{I}$ satisfying $\mathrm{Sing}(V(f)) = V(\mathscr{I})$. Combined with

DefiningEquations from §2.7.1, this constructs the singular set of any algebraic subset of $\mathbb{C}^N$. Repetition of *FindSingularSet* on its output $\mathscr{I}$ while $\dim V(\mathscr{I}) \geq 0$ finds the sequence of sets $\mathrm{Sing}(V(f)), \mathrm{Sing}(\mathrm{Sing}(V(f))_{\mathrm{red}}), \ldots$.

2.7.1 Equations Defining an Algebraic Set

In [23], the membership test for whether a solution x^* of a polynomial system $f(x)$ on $\mathbb{C}^N$ as in Eq. 2.9 was based on the construction using interpolation of polynomials of appropriate polynomials vanishing on irreducible components of $V(f)$. Using such polynomials as we do here is classical, *e.g.* [19].

Let us recall the construction. Let Z be a k-dimensional irreducible degree d component of $V(f)$ and let $S \subset \mathbb{C}^N$ be a finite set of points not contained in Z. Given a general projection $\pi : \mathbb{C}^N \to \mathbb{C}^{k+1}$, π_Z is generically one-to-one and $\pi(Z)$ is a degree d hypersurface not containing $\pi(S)$. There is a degree d polynomial p_π on $\mathbb{C}^{k+1}$, unique up to multiplication by a non-zero complex number, with $V(p_\pi) = \pi(Z)$. Thus composition $p_\pi(\pi(x))$ yields a degree d polynomial on $\mathbb{C}^N$ that vanishes on Z but not at any point of S.

Now let us construct a system of polynomials $g(x)$ such that $Z = V(g)$. We follow the convention that the dimension of the empty set is -1.

Algorithm 4. *DefiningEquations*
Input: A system of polynomials $\{f_1(x), \ldots, f_m(x)\}$ on $\mathbb{C}^N$;
and an irreducible component Z of $V(f)$.
Output: A system $\mathscr{F}$ of polynomials on $\mathbb{C}^N$ with the property that
$Z = V(\mathscr{F})$.

```
Set $K$ equal to the maximum dimension of the set of irreducible
        components of $V(f)$ other than $Z$.
Set $j = 0$.
Set $\mathscr{F}_j := \{f_1, \ldots, f_m\}$.
while $K \geq 0$ do
        Set $S$ equal a finite set consisting of one witness point from each
                component of $V(\mathscr{F}_j)$ except $Z$.
        Set $p$ equal to a degree $d$ polynomial vanishing on $Z$, but not at
                any point of $S$.
        Increment $j$ by $1$.
        Set $\mathscr{F}_j := \mathscr{F}_{j-1} \cup \{p\}$.
        Set $K$ equal to the maximum dimension of the set of
                irreducible components of $V(\mathscr{F}_j)$ other than $Z$.
```

To see why the algorithm works, note that Z is still a component of $V(\mathscr{F}_j)$, and therefore if $K \neq -1$, p is a nontrivial polynomial vanishing on Z and not identically zero on any other component of $V(\mathscr{F}_j \cup \{p\})$. Thus it follows that the maximum dimension of the set of irreducible components of $V(\mathscr{F}_j)$ other than Z is strictly less than the maximum dimension of the set of irreducible components of $V(\mathscr{F}_{j-1})$.

We have the following classical result [19].

Lemma 2.7.1. *Given a pure k-dimensional algebraic subset $Z \subset \mathbb{C}^N$, and a point $x^* \in Z \setminus \operatorname{Sing}(Z)$, it follows that there are $n-k$ degree d polynomials*

$$p_{x^*,1}(x), \dots, p_{x^*,n-k}(x)$$

such that:

1. *Z is a component of $V(p_{x^*,1}(x), \dots, p_{x^*,n-k}(x))$; and*
2. *the Jacobian matrix*

$$\begin{bmatrix} \frac{\partial p_{x^*,1}}{\partial x_1}(x) & \dots & \frac{\partial p_{x^*,1}}{\partial x_N}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial p_{x^*,N-k}}{\partial x_1}(x) & \dots & \frac{\partial p_{x^*,N-k}}{\partial x_N}(x) \end{bmatrix}$$

evaluated at x^* has rank $N-k$.

The proof proceeds by noting that $N-k$ general projections $\pi_1(x) := c_1 \cdot x, \dots, \pi_{N-k} := c_{N-k} \cdot x$ from $\mathbb{C}^N \to \mathbb{C}^{k+1}$ give embeddings of Z in a Zariski open neighborhood of x^* and the matrix

$$\begin{bmatrix} c_1 \\ \vdots \\ c_{N-k} \end{bmatrix}$$

has rank $N-k$. For each $i = 1, \dots, N-k$, letting p_i denote a polynomial of degree d vanishing on $\pi_i(Z)$, the polynomials $p_{x^*,i} := p_i(\pi_i(x))$ satisfy the conclusions of Lemma 2.7.1.

Combining *DefiningEquations* with the procedure outlined following Lemma 2.7.1, we have the following algorithm.

Algorithm 5. *NormalCoordinates*
<u>**Input**</u>: A system of polynomials $\{f_1(x), \dots, f_m(x)\}$ on $\mathbb{C}^N$;
an irreducible component Z of $V(f)$;
and a smooth point $x^* \in Z$.
<u>**Output**</u>: A system $\mathscr{F}$ of polynomials on $\mathbb{C}^N$ with the properties that
Z is an irreducible component of $V(\mathscr{F})$ and $\mathscr{F}^{-1}(0)$ is reduced at x^*.

2.7.2 Computing the Singular Set of the Reduction of an Algebraic Set

Algorithm 6. *FindSingularSet*
Input: A system of polynomials $\{f_1(x),\ldots,f_m(x)\}$ on $\mathbb{C}^N$;
and a general point x^* on a k-dimensional irreducible component Z of $V(f)$.
Output: A system $\mathscr{F}$ of equations with $\mathrm{Sing}(V(f)) = V(\mathscr{F})$.

```
    Set $K := \dim Z$.
    Set $j = 0$.
    Set $\mathscr{F}_j := \mathit{DefiningEquations}(f, Z)$.
    Set $\mathscr{B} := \{Z\}$.
    Set $\mathscr{B}'$ equal to a set with one point $x^*$, where $x^*$ is a witness point of $Z$.
    While $K \geq -1$ do
            Increment $j$ to $j+1$.
            Set $\mathscr{F}_j := \mathscr{F}_{j-1} \cup_{x^* \in \mathscr{B}'} \mathit{NormalCoordinates}(\mathscr{F}, Z, x^*)$
            Compute $\mathscr{A} := \mathit{FindSchemeTheoreticSingularSet}(\mathscr{F}_j)$.
            Use CheckSmoothness to find the set $\mathscr{B}$ of components $\mathscr{A}$ not
                    contained in $\mathrm{Sing}(Z)$.
            Set $\mathscr{B}'$ equal to a set with exactly one witness point for each
                    component of $\mathscr{B}$.
            Set $K := \max_{X \in \mathscr{B}} \dim X$.
    Output $\mathscr{F}_j$.
```

References

1. D.J. Bates, J.D. Hauenstein, A.J. Sommese, and C.W. Wampler, Bertini: Software for Numerical Algebraic Geometry. Available at `www.nd.edu/~sommese/bertini`.
2. D.J. Bates, C. Peterson, A.J. Sommese, and C.W. Wampler, Numerical computation of the genus of an irreducible curve within an algebraic set, 2007 preprint.
3. M. Beltrametti and A.J. Sommese, The adjunction theory of complex projective varieties, Expositions in Mathematics **16**, Walter De Gruyter, Berlin, 1995.
4. G. Björck and R. Fröberg, A faster way to count the solutions of inhomogeneous systems of algebraic equations, with applications to cyclic n-roots, Journal of Symbolic Computation **12** (1991), 329–336.
5. A. Borel and R. Remmert, Über kompakte homogene Kählersche Mannigfaltigkeiten, Math. Ann. **145** (1961/1962), 429–439.
6. A. Borel, Introduction aux groupes arithmétiques, Publications de l'Institut de Mathématique de l'Université de Strasbourg, XV. Actualités Scientifiques et Industrielles, No. 1341, Hermann, Paris, 1969.
7. CoCoATeam, CoCoA: a system for doing Computations in Commutative Algebra. Available at `cocoa.dima.unige.it`.
8. B. Dayton and Z. Zeng, Computing the multiplicity structure in solving polynomial system, Proceedings of ISSAC 2005 (Beijing, China), 116–123, 2005.
9. M. Giusti and J. Heinz, Kronecker's smart, little black boxes, in: R.A. DeVore, A. Iserles, and E. Süli (eds.), Foundations of Computational Mathematics, London Mathematical Society Lecture Note Series **284**, Cambridge University Press, 2001, 69–104.

10. M. Giusti, G. Lecerf, and B. Salvy, A Gröbner free alternative for polynomial system solving, J. Complexity **17** (2001), 154–211.
11. G.-M. Greuel, G. Pfister, and H. Schönemann, SINGULAR 3.0. A Computer Algebra System for Polynomial Computations, Centre for Computer Algebra, University of Kaiserslautern (2005). Available at `www.singular.uni-kl.de`.
12. P.A. Griffiths and J. Harris, Principles of algebraic geometry, Wiley Classics Library, Reprint of the 1978 original, John Wiley & Sons Inc., New York, 1994.
13. W.V.D. Hodge and D. Pedoe, Methods of algebraic geometry, Vol. II, Reprint of the 1952 original, Cambridge University Press, Cambridge, 1994.
14. G. Lecerf, Computing the equidimensional decomposition of an algebraic closed set by means of lifting fibers, J. Complexity **19** (2003), 564–596.
15. A. Leykin, J. Verschelde, and A. Zhao, Newton's method with deflation for isolated singularities of polynomial systems, Theoretical Computer Science **359** (2006), 111-122.
16. Y. Lu, D. Bates, A.J. Sommese, and C.W. Wampler, Finding all real points of a complex curve, in: A. Corso, J. Migliore, and C. Polini (eds.), Algebra, Geometry and Their Interactions, Contemporary Mathematics **448**, American Mathematical Society, 2007, 183–205.
17. D.R. Grayson and M.E. Stillman, MACAULAY 2, A software system for research in algebraic geometry. Available at `www.math.uiuc.edu/Macaulay2/`.
18. A.P. Morgan, A transformation to avoid solutions at infinity for polynomial systems, Applied Mathematics and Computation **18** (1986), 77–86.
19. D. Mumford, Varieties defined by quadratic equations, in: Questions on Algebraic Varieties (C.I.M.E., III Ciclo, Varenna, 1969), Edizioni Cremonese, Rome 1970, 29–100.
20. T. Ojika, Modified deflation algorithm for the solution of singular problems, I. A system of nonlinear algebraic equations, J. Math. Anal. Appl. **123** (1987), 199–221.
21. T. Ojika, S. Watanabe, and T. Mitsui, Deflation algorithm for the multiple roots of a system of nonlinear equations, J. Math. Anal. Appl. **96** (1983), 463–479.
22. A.J. Sommese, Holomorphic vector-fields on compact Kaehler manifolds, Math. Ann. **210** (1974), 75–82.
23. A.J. Sommese, J. Verschelde and C.W. Wampler, Numerical decomposition of the solution sets of polynomial systems into irreducible components, SIAM Journal on Numerical Analysis **38** (2001), 2022–2046.
24. A.J. Sommese, J. Verschelde and C.W. Wampler, Using Monodromy to Decompose Solution Sets of Polynomial Systems into Irreducible Components, in: C. Ciliberto, F. Hirzebruch, R. Miranda, and M. Teicher (eds.), Proceedings of the 2001 NATO Advance Research Conference on Applications of Algebraic Geometry to Coding Theory, Physics, and Computation (Eilat, Israel), NATO Science Series II: Mathematics, Physics, and Chemistry **36**, Springer, 2001, 297–315.
25. A.J. Sommese, J. Verschelde and C.W. Wampler, Symmetric functions applied to decomposing solution sets of polynomial systems, SIAM Journal on Numerical Analysis **40** (2002), 2026–2046.
26. A.J. Sommese, J. Verschelde, and C.W. Wampler, A method for tracking singular paths with application to the numerical irreducible decomposition, in: M.C. Beltrametti, F. Catanese, C. Ciliberto, A. Lanteri, and C. Pedrini (eds.), Algebraic Geometry, a Volume in Memory of Paolo Francia, W. de Gruyter, 2002, 329–345.
27. A.J. Sommese, J. Verschelde, and C.W. Wampler, Numerical irreducible decomposition using PHCpack, in: M. Joswig and N. Takayama (eds.), Algebra, Geometry, and Software Systems, Springer-Verlag, 2003, 109–130.
28. A.J. Sommese, J. Verschelde, and C.W. Wampler, Homotopies for Intersecting Solution Components of Polynomial Systems, SIAM Journal on Numerical Analysis **42** (2004), 1552–1571.
29. A.J. Sommese, J. Verschelde, and C.W. Wampler, Solving Polynomial Systems Equation by Equation, in: A. Dickenstein, F.-O. Schreyer, and A.J. Sommese (eds.), Algorithms in Algebraic Geometry, IMA Volumes in Mathematics and Its Applications **146**, Springer Verlag, 2007, 133-152.
30. A.J. Sommese and C.W. Wampler, The Numerical Solution to Systems of Polynomials Arising in Engineering and Science, World Scientific, Singapore, 2005.

31. A.J. Sommese and C.W. Wampler, Exceptional sets and fiber products, Foundations of Computational Mathematics **8** (2008), 171–196.
32. J. Verschelde, Algorithm 795: PHCpack: a general-purpose solver for polynomial systems by homotopy continuation, ACM Transactions on Mathematical Software **25** (1999), 251–276. Available at `www.math.uic.edu/~jan`.

Chapter 3
Towards Geometric Completion of Differential Systems by Points

Wenyuan Wu, Greg Reid and Oleg Golubitsky

Abstract Numerical Algebraic Geometry represents the irreducible components of algebraic varieties over $\mathbb{C}$ by certain points on their components. Such *witness points* are efficiently approximated by Numerical Homotopy Continuation methods, as the intersection of random linear varieties with the components. We outline challenges and progress for extending such ideas to systems of differential polynomials, where prolongation (differentiation) of the equations is required to yield existence criteria for their formal (power series) solutions. For numerical stability we marry Numerical Geometric Methods with the Geometric Prolongation Methods of Cartan and Kuranishi from the classical (jet) geometry of differential equations. Several new ideas are described in this article, yielding witness point versions of fundamental operations in Jet geometry which depend on embedding Jet Space (the arena of traditional differential algebra) into a larger space (that includes as a subset its tangent bundle). The first new idea is to replace differentiation (prolongation) of equations by geometric lifting of witness Jet points. In this process, witness Jet points and the tangent spaces of a jet variety at these points, which characterize prolongations, are computed by the tools of Numerical Algebraic Geometry and Numerical Linear Algebra. Unlike other approaches our geometric lifting technique can characterize projections without constructing an explicit algebraic equational representation. We first embed a given system in a larger space. Then using a construction of Bates *et al.*, appropriate random linear slices cut out points, characterizing singular solutions of the differential system.

Wenyuan Wu
Department of Mathematics, Michigan State University, Michigan, 48823, USA,
e-mail: wenyuanwu@math.msu.edu

Greg Reid
Department of Applied Mathematics, University of Western Ontario, London, Ontario, N6A 5B7, Canada, e-mail: reid@uwo.ca

Oleg Golubitsky
Department of Applied Mathematics, University of Western Ontario, London, Ontario, N6A 5B7, Canada, e-mail: ogolubit@uwo.ca

L. Robbiano, J. Abbott (eds.), *Approximate Commutative Algebra*, Texts and Monographs in Symbolic Computation, DOI 10.1007/978-3-211-99314-9_3, © Springer-Verlag/Wien 2009

3.1 Introduction

3.1.1 Historical Background

Exact commutative algebra is concerned with commutative rings and their associated modules, rings and ideals. It is a foundation for algebraic geometry for polynomial rings amongst other areas. In our case, commutative algebra is a fundamental constituent of differential algebra for differential polynomial rings. Our paper is part of a collection that focuses on the rapidly evolving theory and algorithms for approximate generalizations of commutative algebra. The generalizations are nontrivial and promise to dramatically widen the scope and applications of the area of traditional exact commutative algebra.

Although the study of systems of differential polynomials (*i.e.* polynomially nonlinear PDE) is more complicated than algebraic systems of polynomials, historically key algorithmic concepts in commutative algebra, often arose initially for PDE. For example, differential elimination methods, arose first in the late 1800's. In particular the classical methods of Riquier 1910 [18] and Tresse 1894 [30] for reducing systems of PDE to certain passive forms can in hindsight be regarded to implicitly contain versions of Buchberger's Algorithm. However, the full potency and development of the theory had to await Buchberger's work 1965 [5]. Indeed there is a well known isomorphism between polynomials and constant coefficient linear homogeneous PDE, which may be interpreted as mapping indeterminates to differential operators. Thus multiplication by a monomial maps to differentiation and reduction maps to elimination. Hence the Gröbner Basis algorithm is equivalent to a differential elimination method for such linear PDE. Further the Hilbert Function, gives the degree of generality of formal power series solutions of such PDE under this mapping.

In another contribution to this collection Scott, Reid, Wu and Zhi [24] exploit this mapping and results for the PDE case, to give new symbolic numeric algorithms for polynomial systems. In the numeric case the underlying object is that of a geometric involutive form, based on the geometric theory of Cartan. Indeed Cartan involutivity is equivalent to the Castelnuovo-Mumford regularity of the symbol module of a system of PDE [15]. This connects a concept in geometry to one in algebra. Gerdt *et al.* [7] have also exploited this isomorphism to develop new incremental exact completion methods for polynomial systems. These are distantly related to, but not equivalent to our geometric (or Cartan) involutive systems. In fact Gerdt's involutive bases, are Gröbner bases, and depend on their choice of involutive division (or ranking). They do not have the same property of coordinate independence possessed by geometric involutive systems.

Hilbert initially introduced the modern definition of a ring; and Emmy Noether vigorously developed the foundations of modern algebra. Ritt was motivated to axiomatize the definition of singular solution for differential equations and also to do for differential equations what Noether had contributed to algebra. Ritt's creation of the area of Differential Algebra, contains the earliest versions of modern differential triangular decomposition methods which also predates their appearance and extensive development in the algebraic case (*e.g.* as in the case of Wu characteristic set methods [34]). The first complete methods to describe such singular solution behavior arose in that work. The current paper, gives some initial results on extending this work from the exact to the approximate case of PDE systems.

3.1.2 Exact Differential Elimination Algorithms

Over and under-determined (*non-square*) systems of ODE and PDE arise in applications such as constrained multibody mechanics and control systems. For example, differential-algebraic equations (DAE) arise from constrained Lagrangian mechanics [1, 32]. Such systems also arise in the analysis of differential equations for their symmetries. Generally such systems must be prolonged (differentiated) to determine the obstructions (or missing constraints) to their formal integrability; or equivalently formulate existence and uniqueness results concerning their solutions. Inclusion of such missing constraints can ease the difficulty of numerical solution of DAE systems by Tuomela and Arponen [31] and Visconti [32].

Much progress has been made in exact differential elimination methods, theory and algorithms for nonlinear systems of PDE for the above task. PDE lie at the intersection of algebraic, geometric and analytic methods, so it is natural that various approaches have been developed. Differential-algebraic approaches include those of Boulier *et al.* [3], Chen and Gao [6], Hubert [8], Mansfield [16], Wu [35]. Algorithmic membership tests (specifically in the radical of a differential ideal) can be given [3, 9] and canonical forms determined. Analytic approaches include those of Reid, Rust *et al.* [22, 33] with associated existence results.

One complication, that even occurs in the exact case, compared to the polynomial case, is that such systems must be differentiated (prolonged), to determine conditions crucial for answering fundamental questions (*e.g.* radical differential membership, and existence of formal solutions). It is well-known that classical exact differential algebra for exactly given differential polynomials, is not a straightforward generalization of polynomial algebra. Objects which are finite in the polynomial case, such as Gröbner Bases, become infinite in the differential case. However finite differential-elimination algorithms exist for answering certain problems (*e.g.* radical differential ideal membership).

These methods, generally apply elimination with respect to a ranking (*e.g.* triangular decomposition via pseudo-division) to isolate subsystems, and then prolong such systems. They also perform binary splitting on whether certain leading quantities with respect to the ranking vanish or not.

3.1.3 Outline of Paper

In this paper, we adapt geometric prolongation methods of the formal theory of PDE originated by Cartan and Kuranishi (see [25] and [21] for algorithmic treatments). The geometric nature of such methods, their invariance under coordinate changes appear to be a natural partner for Numerical Algebraic Geometry. Furthermore, the coordinate (and ranking) dependent methods that essentially generalize triangular decomposition or Gröbner basis ideas to PDE appear less suited since the dependence on ranking can lead to numerical instability (in much the same way as ordering dependent Gauss elimination is unstable). The geometric approaches give a coordinate independent description of properties of PDE in the Jet space. Such methods, being coordinate independent, appear to have greater numerical stability since they do not pivot on small leading quantities, dictated by differential ranking in the classical differential-algebraic approaches. We also mention the work of Krupchyk, Tuomela *et al.* [12, 13] who have been using the geometry of PDE and relating it to the numerical solution of PDE.

This paper is a sequel to [19], [36] and [38] in which we develop theory for using numerical homotopy continuation methods in geometric prolongation. Our previous progress includes hybrid exact-numeric methods [20], that succeed if the input is exact, and pure numerical methods [36] that succeed for detecting certain generic cases of differential systems and associated existence results. Our previous numerical approaches all have the limitation that they only compute generic components and generally miss singular components.

In this article, we show by example that a jet variety can be described by a set of points, called *witness Jet points* rather than usual equation representations. Up to now a full witness point description was not known. We introduce a new object into Numerical Jet Geometry: *witness tangent space* which enables prolongations of hidden constraints and singular components to be obtained implicitly. Other approaches depend on isolating such equations by elimination and then prolonging these equations. Our geometric lifting through witness tangent spaces, enables us to find singular components by using the rank degeneracy conditions of the classical geometric symbol matrix of a differential system. Application of the results of Bates *et al.* [2] to determining the rank deficiency set of the symbol matrix, combined with our lifting methods yields certain witness Jet points characterizing singular solutions of the system.

3.2 Zero Sets of PDE

Let $\mathbb{C}$ denote the complex numbers, $x = (x_1, \cdots, x_n) \in \mathbb{C}^n$ be the independent variables and $u = (u^1, \cdots, u^m) \in \mathbb{C}^m$ be the dependent variables for a system of PDE. The usual commutative approaches to differential algebra and differential elimination theory [22, 3] consider a set of indeterminates $\Omega = \{u^i_\alpha \mid \alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbb{N}^n, i = 1, \cdots, m\}$ where each member of Ω corresponds to a partial derivative by:

$$u^i_\alpha \leftrightarrow \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n} \cdots \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} u^i(x_1,\cdots,x_n)\ .$$

Formal commutative total derivative operators $\mathbf{D}_i$ are introduced to act on members of Ω by a unit increment of the i-th index of their vector subscript: $\mathbf{D}_i u^k_\alpha := u^k_{\alpha+1_i}$ where $\alpha + 1_i = (\alpha_1,\ldots,\alpha_i+1,\ldots,\alpha_n)$. The usual total derivatives $\mathbf{D}_i$ act on functions of $\{x\} \cup \Omega$ by:

$$\mathbf{D}_i = \frac{\partial}{\partial x_i} + \sum_{w\in\Omega} (\mathbf{D}_{x_i} w) \frac{\partial}{\partial w} \tag{3.1}$$

where $\frac{\partial}{\partial w}$ are the usual partial derivatives. For example the system of PDE

$$\begin{aligned} R = \{\ & \tfrac{\partial}{\partial x}u(x,y)^2 + (x+y)\tfrac{\partial}{\partial x}u(x,y) - u(x,y) = 0, \\ & \tfrac{\partial}{\partial y}u(x,y)^2 + (x+y)\tfrac{\partial}{\partial y}u(x,y) - u(x,y) = 0\ \ \} \end{aligned} \tag{3.2}$$

used later in the paper via the above replacements becomes in terms of formal jet variables $R = \{u^2_{(1,0)} + (x+y)u_{(1,0)} - u_{(0,0)} = 0, u^2_{(0,1)} + (x+y)u_{(0,1)} - u_{(0,0)} = 0\}$ where in traditional jet notation we will write this as

$$R = \{(u_x)^2 + (x+y)u_x - u = 0, (u_y)^2 + (x+y)u_y - u = 0\}. \tag{3.3}$$

For convenience we denote the $m_r := m\cdot \binom{r+n-1}{r}$ jet variables of order exactly r corresponding to r-th order derivatives by $\underset{r}{u}$ (*i.e.* u^i_α with $|\alpha| = r$). In (3.3) $\underset{0}{u}= u$, $\underset{1}{u}= (u_x,u_y)$, etc. We also denote those of order $\le r$ by $\underset{\le r}{u} = (u,\underset{1}{u},\ldots,\underset{r}{u})$, so for (3.3) $\underset{\le 2}{u} = (u,u_x,u_y,u_{xx},u_{xy},u_{yy})$. A q-th order differential system with ℓ equations $R^\ell(x,u,\ldots,\underset{q}{u}) = 0$ is associated with a locus (or zero set) of points

$$Z(R) := \{(x,\underset{\le q}{u}) \in J^q(\mathbb{C}^n,\mathbb{C}^m) : R^k(x,\underset{\le q}{u}) = 0, k = 1,\ldots,\ell\} \tag{3.4}$$

where $J^q(\mathbb{C}^n,\mathbb{C}^m) \simeq \mathbb{C}^n \times \mathbb{C}^m \times \mathbb{C}^{m_1} \times \cdots \times \mathbb{C}^{m_q}$ is the jet space of order q and $R^k : J^q(\mathbb{C}^n,\mathbb{C}^m) \to \mathbb{C}$, $k = 1,\ldots,\ell$.

One class of systems considered in this paper will be differential polynomials in $\mathbb{C}[x_1,\ldots,x_n;u,\underset{1}{u},\underset{2}{u},\ldots]$, the ring of all polynomials over $\mathbb{C}$ in the infinite set of indeterminates $\{x\} \cup \Omega$.

3.3 Witness Sets of PDE

In our previous work, the algebraic witness sets of Sommese *et al.*, were directly applied to PDE regarded as algebraic equations in Jet space. These sets were poorly adapted to the special structure of PDE systems and their prolongations. A contribution of this paper is to represent the special geometric structure of PDE and their prolongations by an extension of the algebraic witness set concept which we call *Witness Sets of PDE or Witness Jet Sets*.

In order to define the concept of *Witness Sets of PDE*, we need to introduce involutive systems [14, 17, 4]. Several facts make this class of systems interesting and useful. Firstly, it is possible to determine whether a given system is involutive using only a finite number of operations. Secondly, for any system it is possible to produce an involutive form with the same solution space using only a finite number of operations. And thirdly, an involutive form of a given system of PDE enables existence and uniqueness results for construction of formal power series solutions.

3.3.1 Witness Jet Points

Let R be a polynomially nonlinear differential system of order q.

Definition 3.3.1. [Prolongation] A single prolongation of R is the list of first order total derivatives of all equations of R with respect to all its independent variables:

$$\mathbf{D}(R) := \{R = 0, \mathbf{D}_i R^k = 0 : k = 1, \ldots, \ell\} . \tag{3.5}$$

Prolongation can be considered as an operation on differential equations, so $\mathbf{D}^r(R)$ is defined to be $\mathbf{D}(\mathbf{D}^{r-1}(R))$ in a recursive manner.

Another fundamental operation in Jet geometry is projection, which is defined to be

Definition 3.3.2. [Projection] Given a PDE R in J^{q+1}, the projection of R from J^{q+1} to J^q is:

$$\pi Z(R) := \{(x, u, \underset{1}{u}, \ldots, \underset{q}{u}) \in J^q : (x, u, \underset{1}{u}, \ldots, \underset{q}{u}, \underset{q+1}{u}) \in Z(R) \text{ for some } \underset{q+1}{u}\}.$$

Similarly, π^r is defined to be $\pi \circ \pi^{r-1}$. It is easily follows that projecting the prolongation of a differential system R in J^q may not return the original system but a subset thereof:

$$\pi^r Z(\mathbf{D}^r(R)) \subseteq Z(R), \text{ for any } r \in \mathbb{N} . \tag{3.6}$$

If it is only a proper subset of $Z(R)$, then there are extra constraints, which we call *integrability conditions*. They are differential rather than algebraic consequences of the original system. If for some system we cannot find any new constraints by differentiation, then naturally we introduce the following concept:

Definition 3.3.3. [Formally Integrable System] A differential system R with order q is formally integrable, if $\pi_{q+r}^{q+r+1} Z(R^{(r+1)}) = Z(R^{(r)})$ for any $r \in \mathbb{N}$.

This definition requires that for any r, the projections and prolongations will not produce any new constraints. However verifying formal integrability by direct use of Definition 3.3.3 requires checking infinitely many conditions. For finite implementation, the geometric approach needs to be complemented by some algebraic tools. To produce a finite test, we now turn to the consideration of a subset of formally integrable systems known as involutive systems. Two facts make this class of systems interesting and useful. Firstly, it is possible to determine whether a given system is involutive using only a finite number of operations. Secondly, for any system it is possible to produce an involutive form with the same solution space using only a finite number of operations. Please see [17] for more details.

Suppose R_{invol} is an involutive form of R. A desirable feature of R_{invol} is the formal integrability

$$\pi^r Z(\mathbf{D}^r(R_{\text{invol}})) = Z(R_{\text{invol}}) \subseteq Z(R), \text{ for any } r \in \mathbb{N}.$$

If we know a point in $Z(R_{\text{invol}})$, then this property enables us to construct a power series solution of a PDE system R order by order at this point since R_{invol} contains all its integrability conditions. We call such points *Witness Jet Points* of R.

Previous methods usually used differential elimination to find the projection and uncover all the hidden constraints. That approach can be inefficient and numerically unstable, since rankings underlying differential elimination algorithms sometimes force pivots on very small quantities. Furthermore, the elimination process always generates a binary tree to cover all the cases. As a result, we may have many redundant branches which do not reflect the geometric decomposition of the solution space directly.

An ambitious and challenging task of this paper is to obtain involutive forms by geometrical prolongations avoiding construction of projected relations. *Witness Tangent Space*, which contains local information for prolongations, plays an important role in this new method.

3.3.2 Witness Tangent Space

In the view of geometry, we can consider the zero set of R of order q in Jet space which is an algebraic variety $Z(R)$. Then the prolongation of R can be constructed from the tangent space of $Z(R)$ at its witness points as follows. First calculate the differentials

$$dR^k = \sum_{i=1}^{n} \frac{\partial R^k}{\partial x_i} dx^i + \sum_{w \in \Omega} \frac{\partial R^k}{\partial w} dw, \quad k = 1, \ldots, \ell, \tag{3.7}$$

and then the equation of the tangent space at any non-singular witness point is given by substituting the witness point into (3.7). Note that the apparently infinite sum in (3.7) is finite since R has finite differential order q. A further fundamental construction in the geometry of PDE are the contact conditions

$$dw - \sum_{i=1}^{n} \mathbf{D}_i w \, dx_i = 0, \tag{3.8}$$

for all $w \in \Omega$. Imposing the contact conditions mimics in our formal structure the relations you would expect the dw to have, when pulled back to functions of the independent variables.

However, it is necessary to point out that the polynomial system R in Jet space must be radical ($\sqrt{\langle R \rangle} = \langle R \rangle$). Otherwise, (3.7) may not describe the tangent space.

Example 3.3.4. Let $R = u_x - u^2 = 0$. Then $dR = du_x - 2udu$. After substitution of (3.8), it yields that $u_{xx}dx - 2uu_x dx = 0$. Since x is independent variable, the coefficient of dx, must be zero. Thus, $u_{xx} - 2uu_x = 0$ gives us the prolongation of R.

Now let us consider $R = (u_x - u)^2 = 0$. Then $dR = 2(u_x - u)(du_x - du)$. Because $u_x - u = 0$ at any point of $Z(R)$, we cannot find the relation between du_x and du from dR.

3.4 Geometric Lifting and Singular Components

Numerical algebraic geometry [28, 29] provides us tools to represent the varieties in Jet space by using algebraic witness sets. However an obstacle to such a representation is that new constraints can appear after prolongation and projection and we don't have their algebraic representations. One possibility is to interpolate such projected relations, but this can be very expensive [19], and numerically difficult. Consequently we aim to avoid construction of algebraic equations. The key idea we present here is to consider the projection of the tangent space at a witness point and then lift this point to higher order Jet Space. We call this technique *Geometric Lifting*.

We now briefly and informally describe the lifting process for computing the singular components and then illustrate it using simple examples. And a similar geometric lifting process can be applied to obtain the prolongation information of the generic components of a differential system containing hidden constraints without explicit construction of such hidden constraints.

In the following section we show that a differential equation may have singular components, which also consist of solutions of the differential equations but have a lower dimensional solution space. A precise definition of a singular component for a differential system appears to be unknown. For a single PDE, as is pointed out in [10, Section 4], Darboux suggested calling its solution singular if all separants of the equation (*i.e.* partial derivatives with respect to all highest order derivatives) vanish on this solution. This appears to be the only ranking independent definition of singular solutions to a single ordinary or partial differential equation known up to date. One natural way to generalize this notion of singularity to systems of PDE, which we adopt in this paper, is to call a solution singular if the symbol matrix has deficient rank on this solution. In this sense, the symbol matrix can be understood as a generalization of separants; more precisely, of the set of separants with respect to all possible orderly rankings.

Definition 3.4.1. The symbol of a PDE system R of order q is the Jacobian matrix with respect to its highest order jet variables

$$\mathscr{S}(R) := \frac{\partial R}{\partial \underset{q}{u}} \tag{3.9}$$

Suppose $\mathscr{S}(R)$ has full rank at a generic point of $Z(R)$. To detect the singular components, we need to find all the points where $\mathscr{S}(R)$ has rank deficiency. One obvious way is to construct all the minors and compute the intersection with $Z(R)$. But, in general, it could be very expensive.

An alternative way is to embed the singular components into a higher dimensional space $\{R,\ \mathscr{S}(R)\cdot\mathbf{z}=0\}$ with auxiliary dependent variables $\mathbf{z}$. If the number of columns of $\mathscr{S}(R)$ is larger than the number of rows, we need to consider the transpose of $\mathscr{S}(R)$. Now we can assume $\mathscr{S}(R)$ has full column rank and form a new system

$$R' = \{R,\ \mathscr{S}(R)\cdot\mathbf{z}=0,\ a\cdot z=c\}. \tag{3.10}$$

Because of the third equation, z must be a non-zero vector, which implies that $\mathscr{S}(R)$ must have rank deficiency. Consequently, we have

Proposition 3.4.2. *Let $\mathscr{S}(R)$ be the symbol of a system R. Let $R' = \{R,\ \mathscr{S}(R)\cdot z = 0,\ a\cdot z = c\}$. If $\mathscr{S}(R)$ has full (column) rank at the generic points of $Z(R)$. Then the projection of $Z(R')$ gives the subset of $Z(R)$, where $\mathscr{S}(R)$ has rank deficiency.*

Let us summarize the main steps to identify possible singular components as follows:

1. Input an order q differential system R described in Proposition 3.4.2.
2. Compute its prolongation $\mathbf{D}(R)$ and its Symbol $\mathscr{S}(R)$.
3. Embed R into the space $\{R,\ \mathscr{S}(R)\cdot\mathbf{z}=0\}$ with auxiliary dependent variables $\mathbf{z}$.
4. Compute the witness points W of R'. Such points indicate possible singular solutions. If W is empty, then there is no singular solution.
5. Apply the contact conditions (3.8) to R' and apply them to the equations for the tangent space of R'.
6. Evaluate the output of the previous step, the tangent space condition for R', at each point of W. Apply the independence condition of independent variables to yield a linear system.

If the linear system obtained as output in the procedure above has solution at a point in J^q then we can accomplish the lifting of this point to J^{q+1}. If the symbol is involutive at this point, then it indicates local existence of formal power series solution which is a singular component.

If a witness point cannot be lifted, then it indicates the component passing through this point is inconsistent or there are some hidden constraints which could be uncovered by further prolongations. But it requires more delicate study to design an efficient approach to address such cases.

Remark 3.4.3. In Proposition 3.4.2, it requires a system with full rank symbol. For more general situations, one possibility is to apply the methods in [2] to compute the rank-deficiency set of a polynomial matrix by embedding it to an appropriate higher dimensional space.

3.5 Determination of Singular Components of an ODE using Numerical Jet Geometry

We apply the above procedure to

$$R = u_x^2 + x^2 u_x - \tfrac{3}{2} x u = 0. \tag{3.11}$$

To assist the reader in following the computations we have chosen rational witness points. In practise one should choose random complex floating point numbers, since such methods generally fail on a non-empty (but lower measure) set.

The form the prolongation of R:

$$\mathbf{D}(R) = (2u_x + x^2)u_{xx} + \tfrac{1}{2} x u_x - \tfrac{3}{2} u = 0. \tag{3.12}$$

Computing the witness points of R at a random x (*e.g.* for simplicity $x = x_0 = 1$) is accomplished by slicing R with a random linear equation $au_x + bu = c$, for random a, b, c. We find two witness points, (x_0, u_0, u_x^0) satisfying $(u_x^0)^2 + (x_0)^2 u_x^0 - \frac{3}{2} x_0 u^0 = 0$ and applying the rank test in [36], easily find that the above ODE is involutive at these points. For this case, the symbol of R given by $\mathscr{S}(R) = 2u_x + x^2$ has generic rank 1 at these witness points. Thus higher order derivatives of u can be computed from lower order ones. However this method did not allow us to gain information on possible singular solutions.

We now apply our new procedure to search for possible singular components.

Following the procedure above we have already computed $\mathbf{D}R$ and its Symbol. Let us consider the 1×1 symbol matrix $\mathscr{S}(R) = (2u_x + x^2)$ which has rank 1 at a generic points of $Z(R)$. After embedding the system into a higher dimensional space by adding auxiliary variable z, we have the new system $R' := \{R = 0, (2u_x + x^2)z = 0, az + c = 0\}$ where the third equation is a random linear function of z to force z to be nonzero. The purpose of this embedding is to systematically detect solutions which lower the rank of $\mathscr{S}(R)$ (see Bates *et al.* [2] for general constructions). In particular here it means that R' must satisfy $2u_x + x^2 = 0$, the non-generic case.

After a random slicing of $Z(R')$, the resulting intersection only has isolated roots, which can be solved approximately by homotopy continuation methods. This system has 3 witness points in (x, u, u_x, z) space and the one of them is $(1.0 + 7.29 \times 10^{-17} i,\ -0.1666666666666667 - 1.10 \times 10^{-18} i,\ -0.50 - 3.65 \times 10^{-17} i,\ 0.753 + 1.48 \times 10^{-16} i)$. This indicates the possible presence of singular solutions.

After the projection of the points into (x,u,u_x) space, we obtain the witness points on singular components, *e.g.* the point $(x,u,u_x) = (1.0+7.29\times 10^{-17}\,i,$ $-0.1666666666666667-1.10\times 10^{-18}\,i, -0.50-3.65\times 10^{-17}\,i)$.

Since R is an ODE, the symbol of R is always involutive and we only need to check if there are no hidden constraints. In other words, if each witness point can be lifted to higher dimensional Jet space, then there are no hidden constraints, otherwise we need further computation. Next we will perform the geometric lifting.

The equation of the tangent space of R' is

$$\begin{aligned} (2u_x+x^2)du_x - \tfrac{3}{2}xdu + (2xu_x - \tfrac{3}{2}u)dx &= 0 \\ 2xzdx + 2zdu_x + (2u_x+x^2)dz &= 0 \\ a\,dz &= 0. \end{aligned} \tag{3.13}$$

The contact conditions (3.8) are:

$$du - u_x dx = 0,\;\; du_x - u_{xx}dx = 0,\;\; dz - z_x dx = 0. \tag{3.14}$$

Simplification of the tangent space equations with respect to (3.14) gives

$$\begin{aligned} \big((2u_x+x^2)u_{xx} - \tfrac{3}{2}xu_x + (2xu_x - \tfrac{3}{2}u)\big)\,dx &= 0 \\ \big(2xz + 2zu_{xx} + (2u_x+x^2)z_x\big)\,dx &= 0 \\ a\,z_x\,dx &= 0. \end{aligned} \tag{3.15}$$

Since we seek solutions with x as independent variable, the coefficients of dx in the above system must be zero.

Substitution of the projected witness point

$$\begin{aligned} (x,u,u_x,z) = {} & (1.0+7.29\times 10^{-17}\,i, -0.1666666666666667-1.10\times 10^{-18}\,i, \\ & -0.50-3.65\times 10^{-17}\,i, 0.753+1.48\times 10^{-16}\,i) \end{aligned}$$

into (3.15) gives $u_{xx} = -1.0-1.21\times 10^{-16}\,i$ and $z_x = 0$ so $(x,u,u_x) = (1.0+7.29\times 10^{-17}\,i, -0.1666666666666667-1.10\times 10^{-18}\,i, -0.50-3.65\times 10^{-17}\,i)$ is a witness point of the differential variety of the singular component in J^2. The calculation for the other 3 points of the witness set of $Z(R')$ proceeds in exactly the same fashion, with the same conclusion.

To see the numerical behavior of our method under small perturbations, we consider a nearby ODE:

$$R = 1.0000000211\left(\frac{d}{dx}u(x)\right)^2 + 1.000000075x^2\frac{d}{dx}u(x) - 1.49999995xu(x)$$

A witness point of the corresponding R' is $[u=-0.166666693-2.44\times 10^{-19}\,i, x = 1.0+4.64\times 10^{-17}\,i, z = 0.753-2.12\times 10^{-17}\,i, u_x = -0.50-4.55\times 10^{-17}\,i]$.

Substituting this witness point to the corresponding tangent space equations yields an over-determined linear system:

$$\left(0.0+1.79\times 10^{-18}\,i\right)u_{xx}-0.000000008-3.4\times 10^{-17}\,i=0$$
$$1.51+2.92\times 10^{-17}\,i+\left(1.51-4.24\times 10^{-17}\,i\right)u_{xx}=0$$

The least square solution is $u_{xx}=-1.0-4.76\times 10^{-17}\,i$ producing a residual error of $(.84053746\times 10^{-8}, .998654\times 10^{-26})$. This shows that our method can still detect the singular component under a small perturbation of the input. However, if we apply symbolic computation to this perturbed ODE, we will see that such singular component will disappear.

Example 3.5.1 (Singular Components of ODE using Differential Algebra). For comparison we apply differential elimination and differential algebra to the determination of the singular solution for (3.11).

Briefly, the separant of the ODE is $2u_x+x^2$ and differential algebraic approaches will perform binary splitting on whether this separant vanishes or not. If $2u_x+x^2\neq 0$, then the algorithm terminates with output given by the ODE and this inequation. This is the generic case. If $2u_x+x^2=0$ then we must consider the system $2u_x+x^2=0$, $u_x^2+x^2u_x-\frac{3}{2}xu=0$. An algebraic elimination using $u_x=-x^2/2$ yields $u=-x^3/6$, which can be verified as a singular solution of the ODE. In particular using the `diffalg` package in Maple 11, we confirmed the results of the above computations using classical differential algebra. Applying procedure `essential_components` to equation (3.11), we obtained two essential components characterized by (3.11) and $\{u+\frac{x^3}{6}\}$, respectively. The reader can verify that the point $(x,u,u_x)=(1,\frac{1}{6},-\frac{1}{2})$ we found above lies on the singular solution $u=-x^3/6$.

If as in the example we slightly perturb this example so it is now, in rational terms:

$$R=\left(1+\tfrac{211}{10^{11}}\right)\left(\frac{d}{dx}u\left(x\right)\right)^2+\left(1+\tfrac{75}{10^{10}}\right)x^2\frac{d}{dx}u\left(x\right)-\tfrac{149999995}{100000000}\,xu\left(x\right)$$

Application of exact differential elimination methods (such as diffalg) to this example reveal only one generic case. The singular case is not found. However using the approximate methods the singular case can be found for the slightly perturbed system.

3.6 Determination of Singular Components of a PDE System

In general, we consider a PDE system R of order q. Let $\mathscr{S}_{\ell\times n}$ be the symbol matrix of R with full rank at generic point of $Z(R)$. Roughly speaking, the witness points of $\{R,\mathscr{S}\cdot z, a\cdot z+c\}$, where $a\cdot z+c=0$ is a random linear equation, give us the possible singular components which will be verified by geometric lifting test.

For mixed dimensional system R, it is necessary to remove the witness points belonging to higher dimensional components of $Z(R)$. Removing such redundancies can be carried out by numerical point membership testing as described in [27]. See [2] for a general construction of Bates *et al.* for determining the singular set of a polynomial system.

Next we consider an example of a system of PDE R given by:

$$\begin{aligned}(u_x)^2+(x+y)u_x-u&=0\\(u_y)^2+(x+y)u_y-u&=0\end{aligned}\tag{3.16}$$

We first consider the generic case. To apply the rank test (Theorem 4.1) in [36] we first compute the symbol matrix of R:

$$\mathscr{S}=\begin{pmatrix}2u_x+x+y & 0\\ 0 & 2u_y+x+y\end{pmatrix}$$

The prolongation of R is

$$\mathbf{D}R=\begin{pmatrix}2u_x+x+y & 0 & 0\\ 0 & 2u_x+x+y & 0\\ 0 & 2u_y+x+y & 0\\ 0 & 0 & 2u_y+x+y\end{pmatrix}\cdot\begin{pmatrix}u_{xx}\\u_{xy}\\u_{yy}\end{pmatrix}+\begin{pmatrix}0\\u_x-u_y\\u_y-u_x\\0\end{pmatrix}$$

Continuing with the generic case we next compute a witness set of $Z(R)$ in (x,y,u,u_x,u_y) space obtaining 4 points: $\{(1,2,4,-4,1),(1,2,4,1,1),\ldots\}$. As in our other examples, to assist the reader in following the computations we have chosen rational witness points. In practise these points would be complex floats resulting from intersecting the variety with random linear varieties.

At each point, we apply the rank test (Theorem 4.1) in [36], and find that there are no new constraints. Then Cartan's test [17, 36] shows that the symbol is involutive at each point, and consequently the generic components are involutive and are characterized by the above witness points.

Now to determine if there are singular components we follow the process described in Section 3.4. We note that we have already computed the prolongation of R and $\mathscr{S}(R)$ above. First we embed R in R' where $R'=\{R,\mathscr{S}\cdot\mathbf{z},\mathbf{a}\cdot\mathbf{z}=c\}$. In detail R' is given by

$$\begin{aligned}(u_x)^2+(x+y)u_x-u&=0\\(u_y)^2+(x+y)u_y-u&=0\\(2u_x+x+y)z^1&=0\\(2u_y+x+y)z^2&=0\\a_1z^1+a_2z^2&=c\end{aligned}\tag{3.17}$$

Again for readability we find a simple rational witness point p of $Z(R')$ given by $(x,y,u,u_x,u_y,z^1,z^2)=(1,2,-\frac{9}{4},-\frac{3}{2},-\frac{3}{2},3,4)$.

Then we need geometric lifting to check if there are any new constraints. Note that we do not have the defining equations for the singular components. The equations of the tangent space of R' are

$$\begin{aligned}(2u_x+x+y)du_x+u_x(dx+dy)-du&=0\\(2u_y+x+y)du_y+u_y(dx+dy)-du&=0\\(2u_x+x+y)dz^1+z^1(2du_x+dx+dy)&=0\\(2u_y+x+y)dz^2+z^2(2du_y+dx+dy)&=0\\a_1dz^1+a_2dz^2&=0\end{aligned}\tag{3.18}$$

The contact conditions (3.8) for R' are:

$$\begin{aligned}du-u_xdx-u_ydy&=0\\du_x-u_{xx}dx-u_{xy}dy&=0\\du_y-u_{xy}dx-u_{yy}dy&=0\\dz^1-z^1_xdx-z^1_ydy&=0\\dz^2-z^2_xdx-z^2_ydy&=0\end{aligned}\tag{3.19}$$

Simplification of the tangent space equations (3.18) with respect to (3.19) gives

$$\begin{aligned}(2u_x+x+y)(u_{xx}dx+u_{xy}dy)+(u_x-u_y)dy&=0\\(2u_y+x+y)(u_{xy}dx+u_{yy}dy)+(u_y-u_x)dx&=0\\(2u_x+x+y)(z^1_xdx+z^1_ydy)+z^1(2(u_{xx}dx+u_{xy}dy)+dx+dy)&=0\\(2u_y+x+y)(z^2_xdx+z^2_ydy)+z^2(2(u_{xy}dx+u_{yy}dy)+dx+dy)&=0\\(a_1z^1_x+a_2z^2_x)dx+(a_1z^1_y+a_2z^2_y)dy&=0\end{aligned}\tag{3.20}$$

Substitution of the witness point $(x,y,u,u_x,u_y,z^1,z^2)=(1,2,-\frac{9}{4},-\frac{3}{2},-\frac{3}{2},3,4)$ into the first two equations (3.20) renders them identically satisfied. Substitution of the witness point into the third and fourth equation of (3.20) yields

$$\begin{aligned}(2u_{xx}+1)dx+(2u_{xy}+1)dy&=0\\(2u_{xy}+1)dx+(2u_{yy}+1)dy&=0\end{aligned}\tag{3.21}$$

Since we seek solutions with (x,y) as independent variable, the coefficients of dx and dy in the above system (3.21) are zero and we obtain the unique solution

$$u_{xx}=u_{xy}=u_{yy}=-1/2.$$

This means the point p can be lifted to J^2.

Next we need to check whether the symbol is involutive or not. The symbol of R' corresponding to u_x,u_y is

$$\begin{pmatrix} 2u_x+x+y & 0 \\ 0 & 2u_y+x+y \\ 2z^1 & 0 \\ 0 & 2z^2 \end{pmatrix} \tag{3.22}$$

with $\beta_1 = 1$ and $\beta_2 = 1$ (see [17, 36] for the calculation of β). The prolongation of R' is

$$\begin{aligned} \mathbf{D}R &= 0 \\ 2z^1 u_{xx} + z^1 + (2u_x+x+y)z^1_x &= 0 \\ 2z^1 u_{xy} + z^1 + (2u_x+x+y)z^1_y &= 0 \\ 2z^2 u_{xy} + z^2 + (2u_x+x+y)z^2_x &= 0 \\ 2z^2 u_{yy} + z^2 + (2u_x+x+y)z^2_y &= 0. \end{aligned}$$

Hence the prolonged symbol matrix corresponding to (u_{xx}, u_{xy}, u_{yy}) at p is

$$\begin{pmatrix} 2u_x+x+y & 0 & 0 \\ 0 & 2u_x+x+y & 0 \\ 0 & 2u_y+x+y & 0 \\ 0 & 0 & 2u_y+x+y \\ 2z^1 & 0 & 0 \\ 0 & 2z^1 & 0 \\ 0 & 2z^2 & 0 \\ 0 & 0 & 2z^2 \end{pmatrix}. \tag{3.23}$$

Obviously its rank is $3 = 1\cdot\beta_1 + 2\cdot\beta_2$ at the point p. Therefore, the system R is involutive at p and it is a Witness Jet Point of R.

In summary we have characterized a singular component by R' together with a witness Jet point p on it.

Example 3.6.1 (Singular Components of PDE using Differential Algebra). For comparison we apply differential elimination and differential algebra to the determination of the singular solution for (3.16). Fix a ranking with $u \prec u_x \prec u_y \prec u_{xx} \prec u_{xy} \prec \cdots$. Briefly the separants of the PDE are $\frac{\partial R^1}{\partial u_x} = 2u_x + x + y$ and $\frac{\partial R^2}{\partial u_y} = 2u_y + x + y$ with respect to the ranking. Differential algebraic approaches will perform binary splitting on whether these separants vanish identically or not. The generic case occurs when $2u_x+x+y \neq 0$ and $2u_y+x+y \neq 0$ and essentially yields the input system R with these inequality restrictions. Setting one of the separants to zero, *e.g.* $2u_x+x+y=0$, or equivalently $u_x = -(x+y)/2$ then reducing $(u_x)^2+(x+y)u_x-u=0$ yields $u+(x+y)^2/4=0$ which can be verified as a singular solution of R. Setting the other separant to zero, by symmetry leads to the same conclusion.

In particular using the `diffalg` package in Maple 11, we confirmed the results of the above computations using classical differential algebra. Applying the `Rosenfeld_Groebner` procedure to the system R, we obtain two characterizable components, whose characteristic sets are, respectively, R and $\{u+(x+y)^2/4\}$. These components are essential: the component characterized by R contains polynomial u_{xx}, which does not reduce to 0 with respect to $u+(x+y)^2/4$. *Vice versa,* $u+(x+y)^2/4$ does not reduce to 0 with respect to R.

Here $u+(x+y)^2/4=0$ is the explicit expression for the singular solution, which was characterized by the witness point p which has $(x,y,u)=(1,2,-\frac{9}{4})$. The reader can verify that this point satisfies $u+(x+y)^2/4=0$.

3.7 Discussion

Our previous fully numerical approaches [19, 36, 37] only pursue generic components and generally miss singular components which can be important in applications.

The only previous method, we are aware of, that was capable of using approximate methods to identify singular sets for differential systems was the hybrid method introduced in [20]. In particular the hybrid method is produced by replacing the symbolic algebraic equation processor of the symbolic differential elimination algorithm `rifsimp`, with numerical algebraic geometric methods [29]. Since the algorithm still has a significant symbolic component it is limited to exact input. In particular the symbolic `rifsimp`, on which the hybrid method is based, partitions the system at each major iteration into a leading linear subsystem and a leading nonlinear subsystem. In the hybrid method the algebraic step of testing radical ideal membership (*e.g.* by Gröbner Bases) is replaced by substituting witness points computed using numerical homotopy continuation.

However exact differential elimination methods using binary-tree splitting with respect to a given ranking, seem very difficult to stably apply to systems of PDE containing approximate coefficients. This motivated us to adapt to PDE the stable representation of algebraic varieties by witness sets in Numerical Algebraic Geometry. Our adaptation yields a new object called *Witness Sets of PDE.* Furthermore, rank degeneracy of the symbol matrix of a PDE system using the methods of Bates *et al.* [2] leads to a natural decomposition of differential components. This gives a geometric generalization of the algebraic concept of binary splitting. It potentially gives more direct access to geometric information and less redundancy in comparison with algebraic approaches to splitting. We introduce two techniques, geometric lifting and singular component embedding, that allow us to manipulate differential systems as geometric objects without explicit construction of their defining equations.

Challenges in this area include developing techniques to address higher multiplicity. For example if a system has algebraic multiplicity, its Jacobian does not faithfully describe its tangent space. A challenge is to characterize geometric prolongation and find the singular components in this case.

Another challenge is to explore and detail the connections to Differential Algebra. In particular, is the approach proposed in this paper closer to the original

Ritt-Kolchin decomposition algorithm, or to its factorization-free successor, the Rosenfeld-Gröbner algorithm? Factorization-free methods have been introduced to avoid factorization over towers of algebraic extensions, which was the bottleneck of the Ritt-Kolchin algorithm. However, they split on initials and separants of autoreduced sets of regular (not necessarily prime) ideals, likely leading to an increased number of redundant components, compared to the original Ritt-Kolchin algorithm, which splits on separants of characteristic sets of prime ideals only. Thus, a practical heuristic strategy up to date has been a hybrid one: to factor when possible, otherwise split. The method proposed in this paper suggests a way of splitting that may be, in some sense, closer to the Ritt-Kolchin algorithm: it splits on the symbol of a system that is involutive at all regular witness points and therefore represents the irreducible components discovered so far. At the same time, it avoids explicit factorization over towers of algebraic extensions, by numerically picking witness points on the irreducible components.

Future work will detail its relation to the completeness of differential components and its connection to algebraic approaches in Differential Algebra; for which results in Bates *et al.* [2] will play a crucial role. In particular we plan to apply the witness Jet sets to radical membership testing. We are also developing the techniques described of this paper to a complete algorithm, together with stable termination criteria and theoretical results concerning the completeness of the decomposition.

Acknowledgement

WW and GR especially thank Andrew Sommese for discussions and a valuable email outlining the key method in Bates *et al.* [2]. The authors gratefully acknowledge support for this work from the NSF funded IMA Thematic Year on Applications of Algebraic Geometry, and also acknowledge support from Reid's NSERC grant. We Acknowledge discussions which began this work with Jan Verschelde at the IMA Thematic Year.

References

1. T. Arponen. *Numerical Solution and Structural Analysis of Differential-Algebraic Equations.* Ph.D. Thesis. Helsinki University of Technology, 2002.
2. Daniel J. Bates, Jonathan D. Hauenstein, Christopher Peterson and Andrew J. Sommese. *Numerical decomposition of the rank-deficiency set of a matrix of multivariate polynomials.* To appear in the RISC Series in Symbolic Computation.
3. F. Boulier, D. Lazard, F. Ollivier, and M. Petitot. Representation for the radical of a finitely generated differential ideal. Proc. ISSAC 1995. ACM Press. 158–166, 1995.
4. R. Bryant, S.-S. Chern, R. Gardner, P. Griffiths and H. Goldschmidt. Exterior Differential Systems, MSRI Publications, Springer, 1989.
5. B. Buchberger. An algorithm for finding a basis for the residue class ring of a zerodimensional polynomial ideal (German), Ph.D. thesis, Univ. of Insbruck (Austria), Math. Inst.

6. Y. Chen and X.-S. Gao. Involutive Bases of Algebraic Partial Differential Equation Systems. Science in China (A), 33(2), page 97–113, 2003.
7. Vladimir P. Gerdt and Yuri A. Blinkov. Involutive bases of polynomial ideals, Mathematics and Computers in Simulation, V.45(5-6), page 519-541, 1998.
8. E. Hubert. Detecting degenerate cases in non-linear differential equations of first order. *Theoretical Computer Science* 187(1-2): 7–25, 1997.
9. E. Hubert. Notes on triangular sets and triangulation-decomposition algorithms II: Differential Systems. *Symbolic and Numerical Scientific Computations*, Edited by U. Langer and F. Winkler. LNCS, volume 2630, Springer-Verlag Heidelberg, 2003.
10. E. Hubert. *Essential Components of an Algebraic Differential Equation*. Journal of Symbolic Computation, Vol 28 , 4-5, pages 657-680, 1999.
11. E. Kolchin. Differential Algebra and Algebraic Groups. Academic Press, New York, 1973.
12. K. Krupchyk and J. Tuomela. Shapiro-Lopatinskij Condition for Elliptic Boundary Value Problems. LMS J. Comput. Math. 9 (2006) pp. 287–329.
13. K. Krupchyk, W. Seiler and J. Tuomela. Overdetermined Elliptic PDEs. Found. Comp. Math. 6 (2006), No. 3, pp 309–351.
14. Kuranishi, M. On E. Cartan's prolongation theorem of exterior differential systems. Amer. J. Math., 79 1-47, 1957.
15. Bernard Malgrange. Cartan involutiveness = Mumford regularity. Commutative algebra, Contemp. Math. 331, Amer. Math. Soc., 2003.
16. E. Mansfield. *Differential Gröbner Bases.* Ph.D. thesis, Univ. of Sydney, 1991.
17. J.F. Pommaret. *Systems of Partial Differential Equations and Lie Pseudogroups.* Gordon and Breach Science Publishers, Inc. 1978.
18. Charles Riquier. Les Systemes d'Equations aux Derivees Partielles. Paris, Gauthier-Villars, 1910. xxvii + 590 pp.
19. G. Reid, C. Smith, and J. Verschelde. Geometric completion of differential systems using numeric-symbolic continuation. *SIGSAM Bulletin* 36(2):1–17, 2002.
20. G. Reid, J. Verschelde, A.D. Wittkopf and W. Wu. Symbolic-Numeric Completion of Differential Systems by Homotopy Continuation. Proc. ISSAC 2005. ACM Press. 269–276, 2005.
21. G.J. Reid, P. Lin, and A.D. Wittkopf. Differential elimination-completion algorithms for DAE and PDAE. *Studies in Applied Math.* 106(1): 1–45, 2001.
22. C.J. Rust, *Rankings of derivatives for elimination algorithms and formal solvability of analytic partial differential equations*, Ph.D. Thesis, University of Chicago, 1998.
23. C.J. Rust, G.J. Reid, and A.D. Wittkopf. Existence and uniqueness theorems for formal power series solutions of analytic differential systems. Proc. ISSAC 99. ACM Press. 105-112, 1999.
24. Robin Scott, Greg Reid, Wenyuan Wu and Lihong Zhi. *Geometric Involutive Bases and Applications to Approximate Commutative Algebra.* Accepted by Proceeding of ApCoA 2008.
25. W.M. Seiler. *Involution - The formal theory of differential equations and its applications in computer algebra and numerical analysis.* Habilitation Thesis, Univ. of Mannheim, 2002.
26. A.J. Sommese and J. Verschelde. Numerical homotopies to compute generic points on positive dimensional algebraic sets. *Journal of Complexity* 16(3):572-602, 2000.
27. Andrew J. Sommese, Jan Verschelde, and Charles W. Wampler. Numerical Decomposition of the Solution Sets of Polynomial Systems into Irreducible Components. SIAM J. Numer. Anal. 38(6):2022-2046, 2001.
28. A.J. Sommese and C.W. Wampler. Numerical algebraic geometry. In *The Mathematics of Numerical Analysis*, Volume 32 of *Lectures in Applied Mathematics*, edited by J. Renegar, M. Shub, and S. Smale, 749–763, 1996. Proceedings of the AMS-SIAM Summer Seminar in Applied Mathematics, Park City, Utah, July 17-August 11, 1995, Park City, Utah.
29. A.J. Sommese and C.W. Wampler. *The Numerical solution of systems of polynomials arising in engineering and science*. World Scientific Press, Singapore, 2005.
30. A. Tresse. Sur les invariants diff'erentiels des groupes continus de transformations, Acta Math. 18 (1894), 1–88.

31. J. Tuomela and T. Arponen. On the numerical solution of involutive ordinary differential systems. IMA J. Numer. Anal. 20: 561–599, 2000.
32. J. Visconti. *Numerical Solution of Differential Algebraic Equations, Global Error Estimation and Symbolic Index Reduction.* Ph.D. Thesis. Laboratoire de Modélisation et Calcul. Grenoble. 1999.
33. A. Wittkopf. *Algorithms and Implementations for Differential Elimination*. Ph.D. Thesis, Simon Fraser University, 2004.
34. W. T. Wu. A zero structure theorem for polynomial equations solving. MM Research Preprints, 1:2–12, 1987.
35. W.-T. Wu. On the foundations of algebraic differential geometry. *Mathematics-Mechanization Research Preprint* No. 3, pages 1–26, 1989.
36. Wenyuan Wu and Greg Reid. Application of Numerical Algebraic Geometry and Numerical Linear Algebra to PDE. Proc. of ISSAC'06, pages 345-352, ACM 2006.
37. Wenyuan Wu and Greg Reid. Symbolic-numeric Computation of Implicit Riquier Bases for PDE. Proc. of ISSAC'07, pages 377-385, ACM 2007.
38. Wenyuan Wu. Geometric Symbolic-Numeric Methods for Differential and Algebraic Systems. Ph.D. Thesis, University of Western Ontario, 2007. Available at http://publish.uwo.ca/˜wwu26/papers/UWO-Thesis-W.Wu.pdf.

Chapter 4
Geometric Involutive Bases and Applications to Approximate Commutative Algebra

Robin Scott, Greg Reid, Wenyuan Wu, and Lihong Zhi

Abstract This article serves to give an introduction to some classical results on involutive bases for polynomial systems. Further, we survey recent developments, including a modification of the above: geometric projected involutive bases, for the treatment of approximate systems, and their application to ideal membership testing and Gröbner basis computation.

Introduction

One may apply, to polynomial systems, results for PDE systems using the well-known bijection:

$$\phi : x_i \leftrightarrow \frac{\partial}{\partial x_i}, \tag{4.1}$$

which induces a ring isomorphism between polynomials and linear PDEs, preserving the ideals in both cases. (See Gerdt *et. al.* [8], who have extensively studied and exploited this isomorphism for use on exact polynomial systems.) The important object, from the jet theory of partial differential equations, upon which we will focus

Robin Scott
Department of Mathematics and Statistics, Carleton University, Ottawa, ON, K1S 5B6, Canada, e-mail: `rscott2@connect.carleton.ca`

Greg Reid
Department of Applied Mathematics, University of Western Ontario, London, ON, N6A 5B7, Canada, e-mail: `reid@uwo.ca`

Wenyuan Wu
Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA,
e-mail: `wenyuanwu@math.msu.edu`

Lihong Zhi
Key Lab of Mathematics Mechanization, AMSS, Chinese Academy of Sciences, Beijing 100190, China, e-mail: `lzhi@mmrc.iss.ac.cn`

L. Robbiano, J. Abbott (eds.), *Approximate Commutative Algebra*, Texts and Monographs in Symbolic Computation, DOI 10.1007/978-3-211-99314-9_4, © Springer-Verlag/Wien 2009

is the *involutive basis*. We will see that involutive systems have a natural connection with the Hilbert polynomial and, through the Hilbert polynomial, to Gröbner bases and ideal membership testing.

In particular, we focus on the application of involutive systems to *approximate* polynomials [21, 22, 23, 24]. Unfortunately, the differentiation and elimination methods of Cartan and Kuranishi [4, 11, 25], for completing a system to involutive form, rely on methods akin to Gaussian elimination. In numerical computation, it is well-known that methods which rely on strict variable orderings can be unstable. We describe a modification [1] of the classical test for involutivity, which avoids performing eliminations.

The classical criterion of involution is related to the one for zero dimensional systems given in [18, 29], which is closer to a Gröbner Basis formulation based on commutators, with the commutators playing the role of S-polynomials. However, our involutivity criterion is not based on commutators, and for zero dimensional systems, is coordinate independent. Moreover, it relies on testing dimensions of nullspaces of linear systems, which can be done using the singular value decomposition. The algebraic method which is more closely related to ours is the method of H-bases [17], which also focuses on the dimensions of vector spaces generated by monomials.

As is common practice in commutative algebra, we will consider polynomials systems as linear functions of their monomials, and apply linear algebra to the null spaces of these maps. (See [14] for an early example of this technique and especially see [18]).

Any polynomial system can be written in matrix form. For example, the system $P(\alpha)=0$,

$$\begin{aligned} p_1 &= \alpha_1 x^2+\alpha_2 xy+\alpha_3 y^2+\alpha_4 x+\alpha_5 y+\alpha_6 &= 0,\\ p_2 &= \alpha_7 x^2+\alpha_8 xy+\alpha_9 y^2+\alpha_{10} x+\alpha_{11} y+\alpha_{12} &= 0,\end{aligned}$$

can be written as $M(\alpha)X=0$:

$$\begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 \\ \alpha_7 & \alpha_8 & \alpha_9 & \alpha_{10} & \alpha_{11} & \alpha_{12} \end{pmatrix} \begin{pmatrix} x^2 \\ xy \\ y^2 \\ x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

In applications, the α's are usually some approximate real or complex numbers. Further, note that completion methods, such as Gröbner bases, rely on multiplying polynomials by monomials. (For the PDE case, distinct derivatives are regarded as independent indeterminates in jet space, and multiplication by monomials is replaced by differentiation with respect to independent variables.) For example, the extended system:

$$\{xp_1=0, yp_2=0, p_1=0, p_2=0\}$$

is equivalent to:

$$\begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 & 0 & \alpha_4 & \alpha_5 & 0 & \alpha_6 & 0 & 0 \\ 0 & \alpha_7 & \alpha_8 & \alpha_9 & 0 & \alpha_{10} & \alpha_{11} & 0 & \alpha_{12} & 0 \\ 0 & 0 & 0 & 0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \alpha_6 \\ 0 & 0 & 0 & 0 & \alpha_7 & \alpha_8 & \alpha_9 & \alpha_{10} & \alpha_{11} & \alpha_{12} \end{pmatrix} \begin{pmatrix} x^3 \\ x^2y \\ xy^2 \\ y^3 \\ x^2 \\ xy \\ y^2 \\ x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

All such extended systems define structured classes. Each structured matrix $M(\alpha)$ embeds an extension of the polynomial system $P(\alpha)$, for some particular values of the parameters: $\tilde{\alpha}$. The singular value decomposition may be applied to $M(\tilde{\alpha})$ to detect nearby matrices (and, consequently - as we will show - polynomial systems) with higher-dimensional solution spaces.

Given $A \in \mathbb{F}^{m\times n}$, with $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$, one can compute the *singular value decomposition*:

$$A = U\Sigma V^t, \tag{4.2}$$

where, $U \in \mathbb{F}^{m\times m}$ and $V \in \mathbb{F}^{n\times n}$ are orthogonal and, $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix. The number of its nonzero *singular values*: $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_{min(m,n)} = 0$, is equal to the rank, r, of A. In the case of approximate systems, errors in the entries of A tend to destroy relationships between them, thus increasing the matrix rank, and is reflected in small singular values $\sigma_{r+1},\ldots$ which would otherwise be equal to zero. The following classic theorem [7] is key to understanding the geometry of linear systems and their surroundings. In particular, it actually provides a distance to nearby singular systems, and a way to compute them.

Theorem 4.0.1. (Eckart and Young [7]) *Let $A = U\Sigma V^t \in \mathbb{F}^{m\times n}$ have rank r. A closest matrix to A, with rank $q < r$, can be constructed by forming: $\widetilde{A} = U\widetilde{\Sigma}V^t$, where $\widetilde{\Sigma}$ is equal to Σ with σ_i, for all $q+1 \leq i \leq r$, replaced by zero. Furthermore, $\|A - \widetilde{A}\|_2 = \sigma_{q+1}$.*

The worry in applying the above theorem directly to our structured matrices $M(\tilde{\alpha})$ is that the nearby matrices computed will not necessarily lie in the structured class. This problem introduces the necessity for convergence methods, with which one can iterate to such nearby systems which do lie on a given structured class. This is a familiar problem to the signal processing community, with works dating back to Cadzow [3], and more recently [19, 12]. However, applications to the areas of approximate commutative algebra are less widely known.

After an introduction to the classical theory of involutive systems (for polynomial systems), we highlight recent applications, of projected geometric involutive bases, to approximate commutative algebra. Most of the results (*i.e.* propositions, theorems, and algorithms) are presented without proof, and the reader is referred to the original sources, where they may also find additional clarification of concepts and more detailed examples.

4.1 Jet Spaces and Geometric Involutive Bases

For what follows, let the set of polynomials $P = \{p_1, \ldots, p_m\}$ belong to the ring $\mathbb{F}[x_1, x_2, \ldots, x_n]$, where the field $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. Additionally, let I be the ideal generated by the elements of P, and $I_{\leq q}$ be the subset of I which contains all polynomials with total degree no more than q.

Like the computation of a Gröbner basis, computing an involutive basis for I involves the completion of the set P to a particular form. It turns out that, for a given monomial ordering, an involutive system does not necessarily contain a basis for the leading term ideal $\langle LT(I)\rangle$. However, it still contains "enough" polynomial consequences. In particular, an involutive system, and its extensions, include bases for $I_{\leq q}$. Thus, there is a natural connection between involutive systems and the Hilbert function, which will be described in Section 4.3, and exploited, in each of our applications, in Section 4.4.

4.1.1 Jet Geometry and Jet Space

Through the bijection, ϕ, of (4.1), existing results from PDE theory become available to polynomials $p \in \mathbb{F}[x_1, x_2, \ldots, x_n]$. This, in fact, induces an isomorphism between the polynomial ring and the ring of differential operators:

$$\mathbb{F}[x_1, x_2, \ldots, x_n] \leftrightarrow \mathbb{F}[\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_n}]. \tag{4.3}$$

Under ϕ, systems of polynomials are mapped to systems of linear, homogeneous PDE with constant coefficients. Given a system, P, of polynomials, we let R denote $\phi(P)$. Furthermore, for simplicity, we will use the notation u_{x_i} instead of $\frac{\partial}{\partial x_i} u(x_1, \ldots, x_n)$.

Example 4.1.1. The linear, homogeneous, constant coefficient PDE system corresponding to the polynomial system $P = \{xy - z,\ \ x^2 - y\}$ is $R := \phi(P) = \{u_{x,y} - u_z,\ \ u_{x,x} - u_y\}$.

For what follows, we will be using results from the *Jet Geometry* of PDE where distinct derivatives of u are regarded as independent indeterminates in *Jet Space*. There, any linear, homogeneous PDE with constant coefficients, and differential order at most q, can be considered as belonging to a finite dimensional vector space, J^q, whose basis vectors are the $N_{\leq q} = \binom{n+q}{q}$ derivatives of order at most q. Correspondingly, any polynomial of total degree no more than q is an element of the vector space $\mathbb{F}[x_1, x_2, \ldots, x_n]_{\leq q}$, whose basis vectors are the $N_{\leq q}$ monomials of total degree at most q. In this way, regardless of whether we are thinking in terms of PDE or polynomials, the space in which we will work is isomorphic to $\mathbb{F}^{N_{\leq q}}$.

Now, it is easy to see that our polynomial and PDE systems, P and R, can be written in matrix form, with an identical representative matrix, M. We will denote, by $\underset{q}{x}$, the monomials of degree *exactly* equal to q. The boldface

$$\underset{q}{\mathbf{x}} = (\underset{q}{x}, \underset{q-1}{x}, \ldots, \underset{1}{x}, 1)^t \tag{4.4}$$

will represent the column vector of all $N_{\leq q}$ monomials with degree $\leq q$. Then, also, $\underset{q}{u} = \phi(\underset{q}{x})$ and $\underset{q}{\mathbf{u}} = \phi(\underset{q}{\mathbf{x}})$. Note that the elements $\underset{q}{\mathbf{x}}$ and $\underset{q}{\mathbf{u}}$ are arranged as to respect a total degree ordering.

Example 4.1.2. Let $\underset{2}{\mathbf{x}}$ be such that each $\underset{q}{x}$, for $0 \leq q \leq 2$, is arranged with $z \succ_{lex} x \succ_{lex} y$. Then, $\underset{2}{\mathbf{x}} = (z^2, zx, zy, x^2, xy, y^2, z, x, y, 1)^t$. For the systems P and R given in Example 4.1.1, the matrix which represents them both is:

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 \end{pmatrix}. \tag{4.5}$$

Then, P and R can be recovered by forming $P = M\underset{2}{\mathbf{x}}$ and $R = M\underset{2}{\mathbf{u}}$.

4.1.2 Prolongation and Projection

The first operation to describe, for PDE, is that of prolongation. Since we are concerned with polynomials which, under ϕ, map to linear, homogeneous PDE with constant coefficients, the *prolongation* operation on R can be stated simply as:

$$\mathfrak{D}R := R \cup \left\{ \frac{\partial}{\partial x_j} r : r \in R, 1 \leq j \leq n \right\}. \tag{4.6}$$

(For a description of the formal differentiation operator in the fully nonlinear PDE case see, for example, [25].) We will let $\mathfrak{D}P = \phi^{-1}\mathfrak{D}R$ be used to denote the polynomial system which, alternatively, could have been formed by appending to P all those polynomials formed by multiplying each element of P by each of the variables $x_1, x_2, \ldots, x_n$. Just as R and P can be characterized by the same coefficient matrix, M, we will let $\mathfrak{D}M$ denote the matrix which holds the coefficients of both $\mathfrak{D}R$ and $\mathfrak{D}P$.

Example 4.1.3. For the system R of Example 4.1.1,

$$\begin{aligned}\mathfrak{D}R &= R \cup \frac{\partial}{\partial x}R \cup \frac{\partial}{\partial y}R \cup \frac{\partial}{\partial z}R \\ &= \{u_{x,y} - u_z,\ \ u_{x,x} - u_y,\ \ u_{x,x,y} - u_{x,z},\ \ u_{x,x,x} - u_{x,y}, \\ &\quad\ u_{x,y,y} - u_{y,z},\ \ u_{x,x,y} - u_{y,y},\ \ u_{x,y,z} - u_{z,z},\ \ u_{x,x,z} - u_{y,z}\}.\end{aligned}$$

Similarly, the 2^{nd}, 3^{rd}, ..., r^{th} prolongations of R,

$$\mathfrak{D}^2R = \mathfrak{D}^2M \underset{q+2}{\mathbf{u}},\ \ \mathfrak{D}^3R = \mathfrak{D}^3M \underset{q+3}{\mathbf{u}},\ \ \ldots, \mathfrak{D}^rR = \mathfrak{D}^rM \underset{q+r}{\mathbf{u}}, \tag{4.7}$$

can be computed from $\mathfrak{D}R$, $\mathfrak{D}^2R, \ldots,$ $\mathfrak{D}^{r-1}R$.

Although not the variety of either R or P, the *jet variety*, or zero set, of R is:

$$Z(R) = \left\{ (u, \underset{1}{u}, \underset{2}{u}, \ldots, \underset{q}{u}) \in J^q \ :\ R(u, \underset{1}{u}, \underset{2}{u}, \ldots, \underset{q}{u}) = 0 \right\}. \tag{4.8}$$

The dimension of $Z(R)$ will play an important role. For our problem, and in terms of linear algebra, $Z(R)$ is simply the Null Space of M. Then, a single *geometric projection* is defined as:

$$\pi(R) := \left\{ (u, \underset{1}{u}, \ldots, \underset{q-1}{u}) \in J^{q-1} \ :\ \exists\, \underset{q}{u},\ R(u, \underset{1}{u}, \ldots, \underset{q}{u}) = 0 \right\}. \tag{4.9}$$

The projection operator π maps a point, from the jet variety of R, in J^q, to one in J^{q-1} by simply removing the jet variables of order q (*i.e.* eliminating $\underset{q}{u}$). Just as a sequence of prolongations can be formed, successive projections of R can also be made. For example, the 2^{nd}, 3^{rd}, ..., ℓ^{th} projections,

$$\pi^2R = \pi^2M \underset{q-2}{\mathbf{u}},\ \ \pi^3R = \pi^3M \underset{q-3}{\mathbf{u}},\ \ \ldots, \pi^\ell R = \pi^\ell M \underset{q-\ell}{\mathbf{u}}, \tag{4.10}$$

can be formed by projecting πR, $\pi^2 R$, ..., $\pi^{\ell-1}R$.

Note that a set of equations describing the projection πR is the collection of equations of order $q-1$ corresponding to the reduced row echelon form of M (and that an ordering on the set $\{x, y, \ldots\}$ of indeterminates must be chosen to compute such equations). Alternatively, the geometric approach is to avoid such possibly unstable eliminations on R and, rather, work with πR, which may be realized through the nullspace of M. In particular, we will be focusing on dimensions of πR, which can be computed via the singular value decomposition.

Example 4.1.4. A set of equations which describe the projection $\pi\mathfrak{D}R$ from the prolongation $\mathfrak{D}R$ of Example 4.1.3 is:

$$\pi\mathfrak{D}R = \{u_{x,x} - u_y,\ u_{x,y} - u_z,\ u_{y,y} - u_{x,z}\}. \tag{4.11}$$

4.1.3 The Symbol

The *symbol* of a system R is the Jacobian matrix with respect to its highest order derivatives, $\underset{q}{u}$. That is,

$$\mathfrak{S}R = \frac{\partial R}{\partial \underset{q}{u}}. \tag{4.12}$$

For the PDE systems which we are considering (linear, homogeneous, with constant coefficients), the symbol $\mathfrak{S}R$ is simply the sub-matrix of M corresponding to the highest order derivatives $\underset{q}{u}$ — equivalently, via ϕ, the highest degree monomials $\underset{q}{x}$.

For what follows, it will be necessary to compute dimensions of symbols of projections. We use the following simple formula:

$$\dim \mathfrak{S}\pi^{\ell}R = \dim \pi^{\ell}R - \dim \pi^{\ell+1}R. \tag{4.13}$$

The ℓ^{th} *extended symbol* of R [21], denoted by $\mathfrak{S}^{[\ell]}R$, is the sub-matrix of M corresponding to the derivatives $\underset{q}{u}, \underset{q-1}{u}, \ldots, \underset{q-\ell+1}{u}$. We note that the first extended symbol is really just the symbol of R - a slight abuse of language, since it is not actually "extended". With this definition, we easily obtain the following dimension relation:

$$\dim \pi^{\ell}R = \dim R - \dim \mathfrak{S}^{[\ell]}R. \tag{4.14}$$

Note that, using (4.14), it becomes easy to compute $\dim \mathfrak{S}\pi^{\ell}R$, as in (4.13), by applying the singular value decomposition to the matrix M instead of actually computing equations for any of the projections.

4.1.4 Indices and Cartan Characters

The property of involutivity underlying geometric involutive bases can be described and calculated in terms of Cartan characters. To describe these, we first define the class of a monomial $\mathbf{x}^{\gamma}$ (and equivalently, the class of $u_{\gamma} = \phi(\mathbf{x}^{\gamma})$). For background, see Seiler [25].

A given set of monomials, all in n variables and with total degree q, may be partitioned into classes as follows. For a chosen lexicographic ordering on the n variables $x_1, x_2, \ldots, x_n$, a monomial $\mathbf{x}^{\gamma}$ is called a *class j monomial* if the j^{th} component, γ_j, of its exponent vector $\gamma = (\gamma_1, \ldots, \gamma_n)$ is the first which is nonzero. In terms of derivatives, $u_{\gamma} = \phi(\mathbf{x}^{\gamma})$ is of the same class as $\mathbf{x}^{\gamma}$.

Example 4.1.5. Consider all possible monomials of total degree 2, and let $n = 2$. We have: x^2, xy, y^2. With the ranking $x \succ y$, the monomial y^2 is of class 2, and both x^2 and xy are class 1 monomials. Alternatively, with $y \succ x$, the class 2 monomial is x^2, and the two class 1 monomials are y^2 and yx.

Previously, the symbol matrix $\mathfrak{S}R$ of a system R was defined. The columns of $\mathfrak{S}R$ can be arranged lexicographically, and the reduced row echelon form computed. The columns can then split into two categories: those which contain a pivot, and those which do not. The derivatives corresponding to both the pivots and free-variables can then be assigned to classes, as defined above. With these assignments, the following definitions can be made.

Definition 4.1.6 (*Indices and Cartan characters*). Let R have highest order q, and consider the integers j such that $1 \leq j \leq n$. The *index* $\beta_q^{(j)}$ is the numbers of class j pivots in the reduced row echelon form of $\mathfrak{S}R$. Similarly, the *Cartan character* $\alpha_q^{(j)}$ is the number of class j free-variables in the reduced row echelon form of $\mathfrak{S}R$.

Note that there is a one-to-one correspondence between indices and characters. That is, $\alpha_q^{(j)} = N_q - \beta_q^{(j)}$ (where N_q is the number of jet variable of order precisely equal to q). The test for an involutive system, which will be described next, in Section 4.1.5, involves deciding whether or not its symbol is involutive. This latter test is framed in terms of the indices [25]. Later, in Section 4.3, the characters will provide an expression for the Hilbert polynomial [26].

4.1.5 The Cartan-Kuranishi Prolongation Theorem

The following definition requires the notion of a δ*-regular coordinate system*. That is, one in which the sum $\sum_{j=i}^{n} \beta^{(j)}$ of indices takes its maximum values for all $i = 1, \ldots, n$ [25]. Almost all coordinate systems are δ-regular, and can be obtained, with probability 1, by making a random linear change of variables.

Definition 4.1.7 (*Involutive system*). A q^{th} order linear, homogeneous PDE system R is *involutive* if and only if the two following criteria are satisfied:

1. [*Elimination test*] dim $\pi\mathfrak{D}R =$ dim R.
2. [*Involutive symbol test*] In a δ-regular coordinate system, the symbol $\mathfrak{S}R$ is involutive. That is:
$$\sum_{j=1}^{n} j\beta_q^{(j)} \text{ for } \mathfrak{S}R = \text{rank } \mathfrak{S}\mathfrak{D}R. \tag{4.15}$$

The following *Cartan-Kuranishi prolongation theorem* guarantees that a given a PDE system R will become involutive after a finite number of prolongations and projections.

Theorem 4.1.8. (Kuranishi [11]) *Let R be a PDE system of order q. Then, there exists a finite sequence: $C_{i_1}, C_{i_2}, \ldots, C_{i_k}$, with each C_{i_j} either π or $\mathfrak{D}$, such that $C_{i_k} \cdots C_{i_2} \cdots C_{i_1} R$ is involutive, with order greater than or equal to q.*

4.2 Geometric Projected Involutive Bases and Nearby Systems

Traditional methods [4, 11, 25] for bringing a system to involution require the computation of bases for prolongations and projections, and so involve something not unlike Gaussian elimination [25]. Such methods force pivoting on leaders, in a restricted ordering, and so computing with approximate input is problematic. To facilitate numerical computation, Reid *et. al.* [1] introduced *projectively involutive systems*. In the same work, it was shown that a system is projectively involutive if and only if it is involutive.

The tests for approximate (projective) involutivity are performed by applying the singular value decomposition to compute dimensions. Thus, (using Theorem 4.0.1), this method also gives evidence of nearby exactly involutive systems. We provide an example of the structure of these systems and convergence to a nearby exactly involutive system with that structure.

4.2.1 Geometric Projected Involutive Bases

Definition 4.2.1 (*Projected involutive system*). A linear homogeneous system of PDE, R, is called *projectively involutive* at prolongation order r and projected order ℓ if, for integers $r \geq 0$ and $0 \leq \ell \leq r$, the two following criteria are satisfied:

1. [*Projected elimination test*] dim $\pi^{\ell}\mathfrak{D}^{r}R =$ dim $\pi^{\ell+1}\mathfrak{D}^{r+1}R$.
2. [*Involutive symbol test*] $\mathfrak{S}\ \pi^{\ell}\mathfrak{D}^{r}R$ is involutive.

Moreover, the Involutive Symbol Test (for a q^{th} order system $\pi^{\ell}\mathfrak{D}^{k}R$) can be shown to take the useful form:

$$\sum_{j=1}^{n} j\beta_q^{(j)} \text{ for } \mathfrak{S}\pi^{\ell}\mathfrak{D}^{k}R = \text{rank } \mathfrak{S}\pi^{\ell}\mathfrak{D}^{k+1}R. \tag{4.16}$$

Theorem 4.2.2. (Reid et. al. [1]) *The system R is* projectively involutive *if and only if it is involutive.*

The advantage of the Projected Elimination Test of Definition 4.2.1 is that one simply checks if dim $\pi^{\ell}\mathfrak{D}^{r}R$ = dim $\pi^{\ell+1}\mathfrak{D}^{r+1}R$. The dimensions of R, and its prolongations $\mathfrak{D}^{r}R$ can be computed with stable numerical methods, such as the SVD, as can the dimensions of the extended symbols $\mathfrak{S}^{[\ell]}\mathfrak{D}^{r}R$, for $0 \leq \ell \leq r$. The dimensions of the projections $\pi^{\ell}\mathfrak{D}^{r}R$ can then be determined using (4.14). Alternatively, bases for the projections and prolongations must actually be computed, as in other approaches [25].

Note the following two special cases (Proposition 4.2.3 and Theorem 4.2.4) which enable involutivity of the symbol to be tested in a stable manner, using the SVD together with (4.13) and (4.14).

Proposition 4.2.3. (Reid [24]) *For $n = 2$, the symbol of R is involutive if dim $\mathfrak{S}\mathfrak{D}R =$ dim $\mathfrak{S}R$.*

Theorem 4.2.4. (Reid and Zhi [23]) *Let R be a q-th order system of linear homogeneous PDE R corresponding to a zero dimensional polynomial system $P \subset \mathbb{F}[x_1, x_2, \ldots, x_n]$. Then, R is projectively involutive at prolongation order r and projected order ℓ if and only if $\pi^{\ell}(\mathfrak{D}^r R)$ satisfies the Projected Elimination Test (of Definition 4.2.1) and the Involutive Symbol Test:*

$$\dim \pi^{\ell}\mathfrak{D}^r R = \dim \pi^{\ell+1}\mathfrak{D}^r R. \tag{4.17}$$

In other cases, it is desirable to avoid unstable Gaussian elimination to determine involutivity of the symbol. Next, in Section 4.2.2, we describe a method based on approximate pivot columns to compute the indices and Cartan characters. Alternative approaches include using Spencer Cohomology or numerical approaches to Castelnuovo-Mumford regularity [16].

4.2.2 Approximately Involutive Systems

In this section, we provide the definition, and give an example, of an approximately involutive system. To do this, an approximately involutive symbol is defined in terms of approximate indices and Cartan characters, which correspond to pivot columns of an approximate matrix.

Suppose we have $A \in \mathbb{F}^{m\times n}$, where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. Let $A[1..m, 1..j]$, for $1 \le j \le n$, be the $m \times j$ sub-matrix consisting of columns 1 through j of A.

Remark 4.2.5 (*Exact pivot columns*). The linear independence/dependence relations between columns of a matrix A and those of the reduced row echelon form of A are the same. From the reduced row echelon form of A, one can see that each non-pivot column is dependent only on the pivot columns which appear to its left. So, if $\operatorname{rank}(A[1..m, 1..k]) = \operatorname{rank}(A[1..m, 1..k-1])$, then the k^{th} column of A is linearly dependent. Otherwise, $\operatorname{rank}(A[1..m, 1..k]) = \operatorname{rank}(A[1..m, 1..k-1]) + 1$, and the k^{th} column is a pivot column.

For what follows, the notions of τ-rank and ε-rank are used. These are described, in terms of the singular value decomposition, in the Appendix, and are used to make the following definitions.

Definition 4.2.6 (*ε-pivot columns, τ-pivot columns, and the exact pivot condition*). The column A_k of A is a *τ-pivot column*, or *approximate pivot column*, if $\operatorname{rank}_{\tau}(A[1..m, 1..k]) = \operatorname{rank}_{\tau}(A[1..m, 1..k-1]) + 1$, for some specific tolerance τ. If $\tau = \varepsilon$, then the ε-pivot columns may be called *exact pivot columns*. If the τ-pivot and exact pivot columns of A are the same, we then say that it satisfies the *exact pivot conditions*.

A simple method [6, 24] to determine the approximate pivot columns of A is to compute, in sequence and using the SVD, the ranks, r_τ, of $A[1..m,1]$, $A[1..m,1..2]$, …, $A[1..m,1..n]$. (For large matrices this procedure is rather expensive, however improvements can be made.) A method, for computing a nearby (structured) matrix which satisfies the exact pivot conditions can be found in [24]. In that work, it was referred to as the STLS-RREF, as it made use of the structure-preserving method STLS [12] for computing nearby linearly structured matrices.

Definition 4.2.7 (*Approximate indices and Cartan characters*). Let R be an approximate polynomial system with order q. Then, for a given tolerance τ, $\beta^{(k)}$ is the number of class k, τ-pivot columns in $\mathfrak{S}R$, and $\alpha^{(k)} = N_q - \beta^{(k)}$ is the number of class k, τ-dependent columns in $\mathfrak{S}R$. We will call these the *approximate indices*, or τ-indices and *approximate Cartan characters*, or τ-Cartan characters.

An *approximately involutive systems* and an *approximately involutive symbol* are then identical to their exact counterparts (Definition 4.2.1), except that ranks are replaced by τ-ranks, and indices by τ-indices.

We conclude this section with an example (some finer details of which can be found in [24]).

Example 4.2.8. Consider the system $R = \{u_{x,x} - u_y, u_{x,y} - u_z\}$ (from Seiler [25]).

	$r=0$	$r=1$	$r=2$	$r=3$
$\ell=0$	8	12	16	20
$\ell=1$	4	7	10	13
$\ell=2$	1	4	7	10
$\ell=3$		1	4	7
$\ell=4$			1	4
$\ell=5$				1

Fig. 4.1 Table of dim $\pi^\ell \mathfrak{D}^r R$ for $R = \{u_{x,x} - u_y, u_{x,y} - u_z\}$.

From the dimensions in the table of Figure 4.1, and using the ranking $z \succ x \succ y$ for the calculation of $\sum k\beta^{(k)}$, it can be verified that $\pi\mathfrak{D}R$ passes the Projected Elimination Test and Involutive Symbol test of Definition 4.2.1. Thus, $\pi\mathfrak{D}R$ is involutive.

To each polynomial in $P = \phi^{-1}(R)$, a perturbation: a random, dense polynomial of degree 2, with coefficients approximately of order 1×10^{-8}, was added. The new, approximate system, we call $\widetilde{R}$. For SVD tolerances approximately in the range $\tau = 1 \times 10^{-8}$ to $\tau = 1 \times 10^{-1}$, the table of dimensions for the exact system, given in Figure 4.1, was recovered.

Next, to check if $\sum k\beta^{(k)}$ (for $\mathfrak{S}\pi\mathfrak{D}\widetilde{R}$) is approximately equal to rank $\mathfrak{S}\pi\mathfrak{D}^2\widetilde{R}$, the indices (for $\mathfrak{S}\pi\mathfrak{D}\widetilde{R}$) are computed by determining the approximate pivots of $\mathfrak{D}\widetilde{M}$. The ranking: $z \succ x \succ y$, is chosen, and with $\tau \approx 1 \times 10^{-7}$, the approximate pivots of $\mathfrak{D}\widetilde{M}$, which correspond to order 2 derivatives, appear in columns associated with:

$$u_{y,y},\ u_{x,y},\ \text{and}\ u_{x,x}, \tag{4.18}$$

which are class 3, 2, and 2 derivatives, respectively. Thus, approximately, for $\mathfrak{S}\pi\mathfrak{D}\widetilde{R}$:

$$\sum k\beta^{(k)} = 1*0+2*2+3*1 = 7. \tag{4.19}$$

So, $\pi\mathfrak{D}\widetilde{R}$ is approximately involutive if rank $\mathfrak{S}\pi\mathfrak{D}^2\widetilde{R}$ is also approximately equal to 7. Using the formula (4.13), and the SVD again with $\tau = 1\times 10^{-7}$, rank $\mathfrak{S}\pi\mathfrak{D}^2\widetilde{R} \approx 7$.

So, both conditions: $\pi\mathfrak{D}^2\widetilde{R} = \sum k\beta^{(k)}$ (for $\mathfrak{S}\pi\mathfrak{D}\widetilde{R}$), and dim $\pi\mathfrak{D}\widetilde{R} =$ dim $\pi^2\mathfrak{D}^2\widetilde{R}$, are (approximately) satisfied. Thus, $\pi\mathfrak{D}\widetilde{R}$ is approximately involutive.

4.2.3 Nearby Systems: Structure and Convergence

Unfortunately, using the SVD to determine if a system is approximately involutive does not guarantee the existence of a nearby exactly involutive system. Our main concern is that the SVD gives us definite information about nearby matrices which are not guaranteed to lie on the necessary structured matrix classes, which we begin to describe now.

An entire system, of PDE or polynomials, may be embedded in a class:

$$R(a) = M(a)\underset{q}{\mathbf{u}} \quad \mathit{or} \quad P(a) = M(a)\underset{q}{\mathbf{x}}. \tag{4.20}$$

Here, $M(a)$ is a matrix with some specific structure, and $a = (a_1, a_2, \ldots, a_s)$ is a list of parameter which takes the form $a^0 = (a_1^0, a_2^0, \ldots, a_s^0) \in \mathbb{F}^s$ for a particular member $M(a^0) \in M(a)$.

Even if a given system is embedded a class with a minimal amount of structure, there are certain operations which will introduce a higher level of structure into the problem. One of them is prolongation of $R(\mathbf{a})$ to form:

$$\mathfrak{D}R(a) = \mathfrak{D}M(a)\underset{q+1}{\mathbf{u}},\quad \mathfrak{D}^2R(a) = \mathfrak{D}^2M(a)\underset{q+2}{\mathbf{u}},\quad \ldots. \tag{4.21}$$

The following example will help to clarify what we mean by matrix structure which is induced by prolongations.

Example 4.2.9. The system:

$$P(a) = \{a_1x^2 + a_2xy + a_3y^2 + a_4x + a_5y + a_6,\ b_1x^2 + b_2xy + b_3y^2 + b_4x + b_5y + b_6\}$$

can represent any degree 2 polynomial system of 2 equations in $n = 2$ variables. However, the structure of the prolongation is seen very clearly through its related matrix, $\mathfrak{D}M(a)$. That is:

$$\begin{aligned}\mathfrak{D}P(a) &= P(a) \cup xP(a) \cup yP(a) \\ &= \mathfrak{D}M(a)\underset{3}{\mathbf{x}} \\ &= \begin{pmatrix} a_1 & a_2 & a_3 & 0 & a_4 & a_5 & 0 & a_6 & 0 & 0 \\ b_1 & b_2 & b_3 & 0 & b_4 & b_5 & 0 & b_6 & 0 & 0 \\ 0 & a_1 & a_2 & a_3 & 0 & a_4 & a_5 & 0 & a_6 & 0 \\ 0 & b_1 & b_2 & b_3 & 0 & b_4 & b_5 & 0 & b_6 & 0 \\ 0 & 0 & 0 & 0 & a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \\ 0 & 0 & 0 & 0 & b_1 & b_2 & b_3 & b_4 & b_5 & b_6 \end{pmatrix}\underset{3}{\mathbf{x}}\,.\end{aligned} \tag{4.22}$$

We see that $\mathfrak{D}P(a)$ contains four degree 3 polynomials, each of which cannot represent just any degree 3 bivariate polynomial. Certain coefficients must necessarily be equal to zero. Furthermore, these equations are now related to each other, and to those of $P(a)$, through the parameters $a = (a_1, \dots, a_6, b_1, \dots, b_6)$. Upon prolongation, the system has gained structure.

Suppose $a^0 = (a_1^0, a_2^0, \dots, a_s^0) \in \mathbb{F}^s$. By Theorem 4.0.1, we know that, for the matrix $M(a^0) \in M(a)$ and a given non-negative integer r, the SVD can be used to construct a nearest matrix, $\widetilde{M} = U\widetilde{\Sigma}V^t$, with exact rank r. It is also well-known that, although $\widetilde{M}$ will have exact rank r, it is likely to no longer belong to the structured class $M(a)$. Numerous methods have been proposed for computing a nearby rank r matrix, $M(a^f)$, such that a^0 and a^f are reasonably close. This remains an important area of research. We have found STLS [12] useful to carry out experiments and produce our examples involving convergence.

Example 4.2.10. Again, consider the system from Seiler [25]:

$$R = \{u_{x,x} - u_y, u_{x,y} - u_z\}, \tag{4.23}$$

for which, in Example 4.2.8, it was shown that $\pi\mathfrak{D}\widetilde{R}$ is approximately involutive. We now change our notation, for the perturbed system $\widetilde{R}$, to $R(\widetilde{a^0})$. Here, $R(\widetilde{a^0})$ is embedded in a 20 parameter class, $R(a)$ (with one parameter for every coefficient of each order 2 equation):

$$R(a) = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 & a_9 & a_{10} \\ a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} & a_{17} & a_{18} & a_{19} & a_{20} \end{pmatrix}\underset{2}{\mathbf{u}}\,. \tag{4.24}$$

See [24] for details on the following result: in one iteration, the STLS-RREF method, applied to the structured matrix $\mathfrak{D}M(\widetilde{a^0})$, converged to a nearby structured matrix $\mathfrak{D}M(\widetilde{a^f})$ for which $\pi\mathfrak{D}R(\widetilde{a^f})$ was found to be exactly involutive. That is, the system $\pi\mathfrak{D}R(\widetilde{a^f})$ passed both the Elimination Test and Involutive Symbol Test with tolerance τ set to near working precision. Moreover, $\|M(\widetilde{a^f}) - M(\widetilde{a^0})\|_2 \approx 1.44 \times 10^{-8}$.

We note that, for a given system, the problem of (non)existence of nearby involutive systems is not yet decidable, as our methods are not guaranteed to converge. A deeper understanding of the structure and surroundings of these systems is required.

4.3 The Hilbert Function

The Hilbert polynomial is an important tool for gaining insight into the solution space of (approximate) polynomial systems. Moreover, it is fundamental to the development of our methods for approximate ideal membership testing and approximate Gröbner basis computation, which will be described in Sections 4.4.1 and 4.4.2. In this section, we describe the connection between the Hilbert function and involutive systems.

4.3.1 Definition and Key Properties

Let $\mathbb{F}[x_1,x_2,\ldots,x_n]_{\leq q}$ and $I_{\leq q}$ denote the sets of all of polynomials of total degrees no more than q in $\mathbb{F}[x_1,x_2,\ldots,x_n]$ and I, respectively. Each of the following two forms of the Hilbert function will be used. First, what is often referred to as the *affine Hilbert function*, considers the dimensions of the vector spaces $\mathbb{F}[x_1,x_2,\ldots,x_n]/I_{\leq q}$, for all $q \geq 0$. That is,

$$\Theta_I^{\mathrm{aff}}(q) = \dim \mathbb{F}[x_1,x_2,\ldots,x_n]_{\leq q} - \dim I_{\leq q}. \tag{4.25}$$

Secondly, we will also require what is often called, simply, the *Hilbert function*, which can be written in terms of the affine Hilbert function:

$$\Theta_I(q) = \Theta_I^{\mathrm{aff}}(q) - \Theta_I^{\mathrm{aff}}(q-1). \tag{4.26}$$

Thus, the affine Hilbert function counts, collectively, the polynomials in the set $\mathbb{F}[x_1,x_2,\ldots,x_n]_{\leq q}$ which are not in $I_{\leq q}$, whereas the Hilbert function is concerned with only those elements whose total degree is exactly equal to q.

It is known that, for q large enough, the Hilbert function stabilizes in what is called the *Hilbert Polynomial*, $\mathscr{P}_I^{\mathrm{aff}}(q)$ or $\mathscr{P}_I(q)$. One useful property of the affine Hilbert polynomial is stated as the following theorem (for example, see [5], for details).

Theorem 4.3.1. *Let $I \subset \mathbb{F}[x_1,x_2,\ldots,x_n]$ be a polynomial ideal and $V(I) \subset \mathbb{C}^n$ be its variety. Then, dim $V(I) =$ degree $\mathscr{P}_I^{\mathrm{aff}}(q)$.*

4.3.2 Connection with Involutive Systems

It is not difficult to show that if a q^{th} order polynomial system P passes the Elimination Test of Definition 4.1.7 (or, equivalently, the Projected Elimination Test of Definition 4.2.1), then the dimensions of P and its prolongations $\mathfrak{D}^r P$ satisfy the affine Hilbert function. That is:

$$\Theta^{\mathrm{aff}}(q) = \dim \mathfrak{D}^r P, \tag{4.27}$$

for all $r \geq 0$. The following proposition states that if a system is involutive, then it immediately, without need of prolongation, satisfies not only the Hilbert function, but also the Hilbert polynomial.

Proposition 4.3.2. (Seiler [26]) *Let the system $P \subset \mathbb{F}[x_1, x_2, \ldots, x_n]$ be involutive and have total degree q. Let I be the ideal generated by the polynomials in P. Then, for $r \geq 0$,*

$$\mathscr{P}_I(q+r) = \sum_{j=1}^{n} \binom{r+j-1}{r} \alpha_q^{(j)}. \tag{4.28}$$

To connect the Hilbert polynomial with projected involutive systems, suppose that $\pi^{\ell}\mathfrak{D}^r P$ is projectively involutive with order q. Then, the Hilbert polynomial takes the forms:

$$\mathscr{P}_I^{\mathrm{aff}}(q+s) = \dim\ \pi^{\ell}\mathfrak{D}^{r+s} P \tag{4.29}$$

and

$$\mathscr{P}_I(q+s) = \dim\ \mathfrak{S}\pi^{\ell}\mathfrak{D}^{r+s} P. \tag{4.30}$$

Again, note that using (4.14) and (4.13), the SVD can be used to compute the dimensions of the Hilbert polynomials (4.29) and (4.30).

Remark 4.3.3. In the approximate case, to use Equation (4.28), one could consider computing the Cartan characters of the involutive system $\pi^{\ell}\mathfrak{D}^r P$ by applying the approximate pivot columns method (Section 4.5.3) to the matrices $\mathfrak{D}^r M$. Alternatively, if the affine Hilbert polynomial has degree d, then the $d+1$ prolongations: dim $\pi^{\ell}\mathfrak{D}^{r+s} P$, for $0 \leq s \leq d$, are sufficient for its interpolation.

4.3.3 A Motivational Example

Example 4.3.4. Consider the polynomial system

$$P = \{xy - z, x^2 - y, y^2 - xz\}, \tag{4.31}$$

and let I be the ideal generated by P. Under the ordering $\succ_{tdeg(z \succ x \succ y)}$, the Gröbner basis of I is:

$$G = \{y^3 - z^2, zx - y^2, x^2 - y, xy - z\}. \tag{4.32}$$

The Hilbert polynomial,

$$\mathscr{P}_I^{\mathrm{aff}}(q) = 3q + 1, \tag{4.33}$$

computed from G, has degree 1. This means that the variety of P contains a one-dimensional component. To illustrate a likely effect of measurement errors, random perturbations are added to the coefficients of P to form

$$\begin{aligned}\widetilde{P} = \{&1.00xy - z - (0.97x^2 - 0.38xz - 0.51z^2 + 0.59yz + 0.39y^2) \times 10^{-8},\\ &1.00x^2 - y - (0.39xz + 0.76xy + 0.98z^2 + 0.47yz - 0.53y^2) \times 10^{-8},\\ &1.00y^2 - 1.00xz + (0.34x^2 + 0.21xy - 0.64z^2 + 0.75yz) \times 10^{-8}\},\end{aligned}$$

which now generates the ideal $\widetilde{I}$. Current methods can compute a Gröbner basis for $\widetilde{I}$ if the coefficients are converted into rational numbers. Doing this, the Gröbner basis, computed in Maple 10, contains integer coefficients exceeding 1200 digits! The Hilbert Polynomial,

$$\mathscr{P}^{\text{aff}}_{\widetilde{I}}(q) = 8, \tag{4.34}$$

is constant, meaning that the ideal is zero dimensional - the variety is nothing more than a set of isolated points. Important structure has been destroyed.

4.4 Applications

4.4.1 Ideal Membership

Previously, in Section 4.3.2, it was described how the Hilbert function can be computed using involutive systems. Later, we show that an involutive system is not necessarily a Gröbner basis, but may be prolonged to contain one. However, with Proposition 4.4.1, below, we have the following result: *ideal membership is decidable without first (explicitly) computing a Gröbner basis.*

Let $I = \langle p_1, \ldots, p_m \rangle$ be the ideal generated by the elements of a set of polynomials $P = \{p_1, \ldots, p_m\} \subset \mathbb{F}[x_1, x_2, \ldots, x_n]$. Using the Hilbert function, it is possible to decide whether an arbitrary polynomial $f \in \mathbb{F}[x_1, x_2, \ldots, x_n]$ lies in I. We let J denote the ideal $\langle f, p_1, \ldots, p_m \rangle$.

Suppose f belongs to the ideal I. Then, $I = J$ and so the Hilbert polynomials of I and J are identical.

Conversely, suppose that f is not a member of the ideal I. Let G be a Gröbner basis for P. The normal form of f, with respect to this Gröbner basis, has a leading term which is not divisible by the leading terms of any element of G. Thus, $\langle LT(J) \rangle \supsetneq \langle LT(I) \rangle$, which means that $\mathscr{P}^{\text{aff}}_{J}(q) \neq \mathscr{P}^{\text{aff}}_{I}(q)$.

The following proposition follows from the above discussion.

Proposition 4.4.1. *Let $P = \{p_1, \ldots, p_m\} \subset \mathbb{F}[x_1, x_2, \ldots, x_n]$ be a set of polynomials. Then, an arbitrary polynomial $f \in \mathbb{F}[x_1, x_2, \ldots, x_n]$ belongs to the ideal $\langle p_1, \ldots, p_m \rangle$ if and only if $\mathscr{P}^{\text{aff}}_{\langle p_1, \ldots, p_m \rangle}(q) = \mathscr{P}^{\text{aff}}_{\langle f, p_1, \ldots, p_m \rangle}(q)$.*

The following algorithm decides ideal membership. In the approximate case, it becomes a method to test *approximate ideal membership.*

Algorithm 4.4.2.

```
INPUT: $P = \{p_1,\dots,p_m\}$ and $f \in \mathbb{F}[x_1,x_2,\dots,x_n]$
OUTPUT: "$f \in I$" or "$f \notin I$"

$Q := \{f\} \cup P$
$R := InvolutiveSystem(P)$
$S := InvolutiveSystem(Q)$
$\mathscr{P}_{\langle p_1,\dots,p_m\rangle}(q) := HilbertPolynomial(R)$
$\mathscr{P}_{\langle f,p_1,\dots,p_m\rangle}(q) := HilbertPolynomial(S)$
if   $\mathscr{P}_{\langle f,p_1,\dots,p_m\rangle}(q) = \mathscr{P}_{\langle p_1,\dots,p_m\rangle}(q)$
  return "$f \in I$"
else
  return "$f \notin I$"
```

In the case of approximate systems, the Hilbert polynomials can be computed as outlined in Remark 4.3.3. Progress towards approximate membership testing has been made using this method [21, 30]. We mention that there is a detailed example of a positive approximate membership test, and convergence to a nearby ideal for which membership is exact, which can be found in ([30], *Example* 8.2*: Polynomial Ideal Membership*).

4.4.2 Gröbner Bases for Polynomial Systems

For a given system $P \subset \mathbb{F}[x_1,x_2,\dots,x_n]$, let I be the ideal which its elements generate. We will use the following terminology. We say that P *implicitly contains a Gröbner basis* if $\mathbb{F}$-linear combinations of members of P yield a Gröbner basis for I. Now, for any given ideal I, it is obvious that if one prolongs far enough, then a Gröbner basis will implicitly be contained in the extended set of polynomials. The difficulty in that approach is that one would not know when to stop the prolongation process!

Key to our development is an observation of Macaulay [15]. That is, under a total degree ordering, the vector spaces $I_{\leq q}$ and $\langle LT(I)\rangle_{\leq q}$ have the same dimensions. Thus, the Hilbert functions of an ideal, I, and the ideal generated by its leading terms, $\langle LT(I)\rangle$, are identical.

By the Cartan-Kuranishi prolongation theorem, there is an involutive system, P', whose elements also generate I. Additionally, the dimensions of the prolongations of P' respect the Hilbert polynomial. So, if we have an involutive system P', the next natural question to ask is if $\langle LT(P')\rangle$ has the same Hilbert polynomial as P' (as this is a requirement for P' to be a Gröbner Basis for I). Unfortunately, the answer to this question is: not necessarily. However, we do have the following result.

Proposition 4.4.3. (Scott [24]) *Let $P \subset \mathbb{F}[x_1,x_2,\dots,x_n]$ be involutive, I the ideal which its elements generate, and $\succ_{tdeg}$ be a fixed total degree monomial ordering. Then:*

1. *The k^{th} prolongation, $\mathfrak{D}^k P$, implicitly contains a $\succ_{tdeg}$-Gröbner basis for I if and only if $\mathscr{P}^{\text{aff}}_{\mathfrak{M}}(q) = \mathscr{P}^{\text{aff}}_{I}(q)$, where $\mathfrak{M}$ is the monomial ideal generated by the leading terms, under $\succ_{tdeg}$, of $\mathfrak{D}^k P$.*
2. *In the special case where I is a zero-dimensional ideal, then from involution, it requires zero prolongations to implicitly contain* any *$\succ_{tdeg}$-Gröbner basis.*

Moreover, if $\mathfrak{D}^k P$ does implicitly contain a $\succ_{tdeg}$-Gröbner basis for I, then the equations of this basis can be taken as those which correspond to the reduced row echelon form of the matrix associated with $\mathfrak{D}^k R$.

Algorithm 4.4.4, below, computes a Gröbner basis for an exact polynomial system, from an involutive system, using the Hilbert polynomial. In the approximate case, it becomes a method to test for an *approximate Gröbner basis*.

Algorithm 4.4.4.
INPUT: *A monomial order $\succ_{tdeg}$ and $P = \{p_1, p_2, \dots, p_m\}$*
OUTPUT: *A $\succ_{tdeg}$-Gröbner Basis G for $I = \langle p_1, p_2, \dots, p_m \rangle$*

```
$R := InvolutiveSystem(P)$
$\mathscr{P}_I(q) := HilbertPolynomial(R)$
$\mathfrak{M} := \langle LeadingMonomials(R, \succ_{tdeg}) \rangle$
$\mathscr{P}_{\mathfrak{M}}(q) := HilbertPolynomial(\mathfrak{M})$
repeat
    $R := \mathfrak{D}R$
    $\mathfrak{M} := \langle LeadingMonomials(R, \succ_{tdeg}) \rangle$
    $\mathscr{P}_{\mathfrak{M}}(q) := HilbertPolynomial(\mathfrak{M})$
until $\mathscr{P}_{\mathfrak{M}}(q) = \mathscr{P}_I(q)$
return $G := Basis(R, \succ_{tdeg})$
```

For approximate systems, the sub-algorithms of Algorithm 4.4.4 will take different forms than for the exact case. In particular, they are replaced with their approximate counter-part methods. Furthermore, at each step, the named object is not actually computed, but the approximate one is instead detected. The final step requires iteration to a nearby exact system. We briefly describe the process below.

1. [*ApproximatelyInvolutiveSystem*(P)] Given P and a tolerance τ, the method described in Section 4.2.2 may be used to detect a nearby system $\pi^\ell \mathfrak{D}^r R$ which is exactly involutive.
2. [*ApproximateHilbertPolynomial*(R)] This may either be interpolated from the τ-dimensions of a number of prolongations of $\pi^\ell \mathfrak{D}^r R$. Alternatively, (4.28) may be used, after determining the Cartan characters using τ-dependent columns.
3. [*ApproximateLeadingMonomials*$(\mathfrak{M}, \succ_{tdeg})$] With $\mathfrak{S}\pi^\ell \mathfrak{D}^r R$ ordered as to respect $\succ_{tdeg}$ the leading monomials may be computed using τ-pivot columns.
4. [*ApproximateHilbertPolynomial*$(\mathfrak{M})$] The monomials are exact objects, and thus there is no ambiguity in $\mathfrak{M}$ or in the ideal they generate. So, none also in the Hilbert polynomial of that ideal. However, these leading monomials are only approximate leaders, and so we call the Hilbert polynomial of $\mathfrak{M}$ approximate

as well. Still, since the monomials are exact, this approximate Hilbert polynomial may be computed as for the exact case.

5. [$Basis(R, \succ_{tdeg})$] Suppose that, at this point, the initial approximately involutive system $R := \pi^{\ell}\mathfrak{D}^{r}R(a^0)$ has been prolonged k times, so that we have assigned $R := \pi^{\ell}\mathfrak{D}^{r+k}R(a^0)$. To extract a basis from $\pi^{\ell}\mathfrak{D}^{r+k}R(a^0)$, iteration is required to a nearby system, $\pi^{\ell}\mathfrak{D}^{r+k}R(a^f)$, which:

   - Is exactly (projectively) involutive.
   - Satisfies the exact pivot conditions of Definition 4.2.6.

   For the iteration to the nearby exact system, we use STLS-RREF method [24] on $D^{r+k}M(a^0)$. The RREF of $D^{r+k}M(a^f)$ is then completed using Proposition 4.4.5 (below). The sub-matrix corresponding to $\pi^{\ell}D^{r+k}M(a^f)$ can then be easily extracted. The resulting set of equations can then be inter-reduced to form the exact reduced Gröbner basis of the ideal which is generated by the polynomials $P(a^f)$.

   (We note that, in addition to $\mathfrak{D}^{r+k}P(a^f)$ being exactly involutive, the extra criterion, that the exact pivot conditions be satisfied, requires the matrix $\mathfrak{D}^{r+k}M(a^f)$ to have an additional layer of structure. Hence, the nearest system $\mathfrak{D}^{r+k}P(a_1^f)$, which is exactly involutive may be closer than the nearest exactly involutive system $\mathfrak{D}^{r+k}P(a_2^f)$ which also contains an exact Gröbner basis.)

Proposition 4.4.5. (Dayton [6], Scott [24]) *Suppose that the matrix $A \in \mathbb{F}^{m\times n}$ has rank r and pivot columns $A_{i_1}, A_{i_2}, \ldots, A_{i_r}$. Then the nonzero rows of the reduced row echelon form of A can be written as:*

$$\widehat{W}^{-1}W, \tag{4.35}$$

where the rows of W form a basis for the Row Space of A, and $\widehat{W}$ is the square matrix formed with columns $i_1, i_2, \ldots, i_r$ of W.

Note that, in the above proposition, the first r rows of the matrix V^t from the singular value decomposition $A = U\Sigma V^t$ provide a basis W for the Row Space of A. Also, the matrix $\widehat{W}$, has full rank, r. Moreover, as $\widehat{W}$ is square, the SVD can also be used to compute its (exact) inverse: $\widehat{W}^{-1} = V\Sigma^{-1}U^t$, from $\widehat{W} = U\Sigma V^t$, where Σ^{-1} is easily formed by inverting each singular value on the diagonal of Σ.

In some cases, for a particular choice of ordering, the approximate pivot columns may give information about nearby systems. It is possible that instability in the Gröbner basis or, perhaps, (non)existence of a nearby exact Gröbner basis, under this ordering, may be reflected in, for example, in the conditioning of the matrix $\widehat{W}$. If this situation does occur, we suggest either an alternate variable ordering, or a random linear change of variables to improve the stability of the approximate Gröbner basis method.

By these methods, using involutive systems and the Hilbert polynomial, our goal is to iterate to nearby higher-dimensional polynomial systems. We note that, since we have the degree of the (approximate) Hilbert polynomial, we thus know the dimension of the highest-dimensional component in $V(I)$. However, we have no information about the underlying components. (See [13] for progress on this issue.)

We conclude this section by sketching the steps and results of an approximate Gröbner basis computation. More details of this example appear in [24].

Example 4.4.6. Consider Seiler's system [25]:

$$P = \phi^{-1}(R) = \{xy - z,\ x^2 - y\}, \tag{4.36}$$

which, with the ordering $z \succ_{tdeg} x \succ_{tdeg} y$, has reduced Gröbner basis:

$$G = \{y^3 - z^2, zx - y^2, x^2 - y, xy - z\}. \tag{4.37}$$

We form the perturbed system:

$$\widetilde{P} = \phi^{-1}(\widetilde{R}) = \{xy - z + \delta p_1,\ x^2 - y + \delta p_2\}, \tag{4.38}$$

where δp_1 and δp_2 are random, dense polynomials of degree 2, with coefficients all around the order of 10^{-8}.

It can be shown that $\pi\mathfrak{D}R$ is involutive, and that the Hilbert Polynomial of $I = \langle xy - z, x^2 - y\rangle$ is:

$$\mathscr{P}_I^{\text{aff}}(q) = 3q + 1. \tag{4.39}$$

Additionally, $\pi\mathfrak{D}\widetilde{R}$, is approximately involutive, for $\tau \approx 1 \times 10^{-8}$, and the approximate Hilbert polynomial of $\widetilde{I} = \langle xy - z + \delta p_1, x^2 - y + \delta p_2\rangle$ is:

$$\mathscr{P}_{\widetilde{I}}^{\text{aff}}(q) \approx 3q + 1. \tag{4.40}$$

To determine if $\pi\mathfrak{D}\widetilde{P}$ contains an approximate Gröbner basis for $\widetilde{I}$, under the monomial ordering $z \succ_{tdeg} x \succ_{tdeg} y$, we proceed as follows. Computing the approximate pivots of degree ≤ 2 in $\mathfrak{D}\widetilde{M}$, which will correspond to the leading terms of $\pi\mathfrak{D}\widetilde{P}$, we find that the monomial ideal generated by the leading terms of $\pi\mathfrak{D}\widetilde{P}$ is:

$$\mathfrak{M} = \langle zx, x^2, xy\rangle. \tag{4.41}$$

The Hilbert polynomial of $\mathfrak{M}$ is:

$$\mathscr{P}_{\mathfrak{M}}^{\text{aff}}(q) = \tfrac{1}{2}(q^2 + 3q + 4), \tag{4.42}$$

which does not agree with (4.40). So, $\pi\mathfrak{D}\widetilde{P}$ does not contain an approximate Gröbner basis for $\widetilde{I}$, and must be prolonged. We take one prolongation step, and compute the approximate pivots of order ≤ 3 in $\mathfrak{D}^2\widetilde{M}$ to find the leading terms of $\pi\mathfrak{D}^2\widetilde{P}$. A new leading term: y^3, is uncovered, with which the previous leading term ideal is augmented to form:

$$\mathfrak{M} = \langle zx, x^2, xy, y^3\rangle, \tag{4.43}$$

for which:

$$\mathscr{P}^{\text{aff}}_{\mathfrak{M}}(q) = 3q+1. \tag{4.44}$$

Since $\mathscr{P}^{\text{aff}}_{\mathfrak{M}}(q) = \mathscr{P}^{\text{aff}}_{\widetilde{I}}(q)$, then $\pi\mathfrak{D}^2\widetilde{P}$ contains an approximate Gröbner basis for $\widetilde{I}$. The final step is to compute a nearby basis for $\pi\mathfrak{D}^2\widetilde{P}$.

The system P from Equation (4.36) is embedded in the 20 parameter class of (4.24) of Example 4.2.10. Now, the system (4.38) will be referred to as $P(\widetilde{a^0}) = \phi^{-1}R(\widetilde{a^0})$. Prolonging twice, we form $\mathfrak{D}^2P(\widetilde{a^0})$. In one iteration of STLS-RREF, we converge to a nearby system: $\mathfrak{D}^2M(\widetilde{a^f})$, for which $\|\mathfrak{D}^2M(\widetilde{a^f}) - \mathfrak{D}^2M(\widetilde{a^0})\|_2 \approx 1.442\times 10^{-8}$ and $\| M(\widetilde{a^f}) - M(\widetilde{a^0})\|_2 \approx 1.100\times 10^{-8}$. Now, with tolerances around working precision, we find that $\pi\mathfrak{D}R$ is exactly involutive.

By interpolation, using the now exact dimensions of $\pi\mathfrak{D}R(\widetilde{a^f})$ and its prolongations, we have $\mathscr{P}^{\text{aff}}_{I(\widetilde{a^f})}(q) = 3q+1$. Additionally, the monomial ideal, $\mathfrak{M}$, generated by the leading terms of $\pi\mathfrak{D}^2P(\widetilde{a^f})$, which correspond to the exact pivots of order 3 and less in $\mathfrak{D}^2M(\widetilde{a^f})$, is the same as the one of Equation (4.43). Thus, since $\mathscr{P}^{\text{aff}}_{\mathfrak{M}}(q)$ matches $\mathscr{P}^{\text{aff}}_{I(\widetilde{a^f})}(q)$, then $\pi\mathfrak{D}^2P(\widetilde{a^f})$ is an exact Gröbner basis for $I(\widetilde{a^f})$.

The final step is to complete the RREF of $\pi\mathfrak{D}^2M(\widetilde{a^f})$, using Proposition 4.4.5. After this, the equations of degree ≤ 3, which form a basis for $\pi\mathfrak{D}^2P(\widetilde{a^f})$, are extracted. They are then inter-reduced by removing polynomials with redundant leading monomials.

We are left with the $z \succ_{tdeg} x \succ_{tdeg} y$ Gröbner basis for $I(\widetilde{a^f})$:

$$\begin{aligned} G = \{\ & 1.000y^3 - 1.000z^2 - (1.206zy + 1.113y^2 - 1.160z)\times 10^{-8}, \\ & 1.000zx - 1.000y^2 - (0.811zy - 0.688z + 0.580x - 0.369y)\times 10^{-8}, \\ & 1.000x^2 - 1.000y - (1.162z - 0.170x + 0.375)\times 10^{-8}, \\ & 1.000xy - 1.000z - (0.811y^2 + 0.744x + 0.518y - 0.580)\times 10^{-8}\ \}. \end{aligned}$$

4.5 Appendix

4.5.1 The SVD, ε-Rank, and τ-Rank

In applications involving polynomial systems, coefficients are often inferred from inaccurate data. For linear systems, analysis based on the singular value decomposition will likely reveal two ranks of interest: that which is defined by a computer's working precision, and another which depends on the number of digits of accuracy which is believed to be held by the approximate data. These two levels are made clearer in the following discussion.

Given $A \in \mathbb{F}^{m\times n}$, where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$, one can compute its *Singular Value Decomposition*:

$$A = U\Sigma V^t, \tag{4.45}$$

where $U \in \mathbb{F}^{m\times m}$, and $V \in \mathbb{F}^{n\times n}$ are orthogonal, and $\Sigma \in \mathbb{F}^{m\times n}$ is diagonal. The diagonal elements of Σ, denoted by $\sigma_1, \sigma_2, \ldots, \sigma_{min(m,n)}$, are called the *singular values* of A. The following are some well-known and useful properties of U, V, and Σ:

- The rank of A is precisely the number of its nonzero singular values.
- If the rank of A is r, then the first r rows of V^t form a basis for the row space of A.
- The last $n-r$ columns of V form a basis for the null space of A.

Definition 4.5.1 (*ε-rank and τ-rank*). Suppose that we are given some matrix $A \in \mathbb{F}^{m\times n}$, where $\mathbb{F}$ is either $\mathbb{R}$ or $\mathbb{C}$.

1. Let the ε-rank of A be the number of its “nonzero” singular values. That is, those which are greater than the working precision, ε. We will also use the notations r_ε and $\text{rank}_\varepsilon(A)$.
2. Let the τ-rank of A be the number of its singular values which are greater than some reasonable tolerance, τ. Again, we will, at times, denote it by either r_τ or $\text{rank}_\tau(A)$.

Finally, recall Theorem 4.0.1 which, given a rank r_ε matrix A, provides a distance to an existing singular system, $\widetilde{A}$, of rank r_τ. Additionally, the theorem describes how to compute such nearby matrix.

4.5.2 STLS

In this section, we provide a short description of the setup for STLS [12] and other methods (for example, [19]) which consider matrices with a linear structure:

$$M(a) = \sum_{j=1}^{s} a_j M_j. \tag{4.46}$$

Consider a matrix $M(a^0) \in M(a)$ which has had its rank condition destroyed by some sort of measurement error in this particular set of parameters, a^0. Usually, it happens to be that $r_\varepsilon > r := r_\tau$. Let $A(a) \in \mathbb{R}^{m\times r}$ be a matrix formed with r columns of $M(a)$, and let $B(a) \in \mathbb{R}^{m\times d}$ be the matrix formed from the remaining $d := n-r$ columns. To begin, $M(a^0)$ has less than d dependent columns. So, the following system has no exact solution $X \in \mathbb{R}^{r\times d}$. Instead,

$$A(a^0)X \approx B(a^0) \tag{4.47}$$

may be only approximately solved. The common goal of these methods is to find a nearby set of parameters a^f such that the matrix $M(a^f)$ has $r_\varepsilon = r := r_\tau$, and so lies on $M_r(a)$. In other words, the system:

$$A(a^f)X = B(a^f) \tag{4.48}$$

has an exact solution.

Obviously, successful convergence is heavily dependant on the partitioning of the columns of $M(a)$ into the left-hand side, $A(a)$, and the right-hand side, $B(a)$. A bad choice might forbid convergence, if the columns of $B(a)$ are not in the range of those of $A(a)$, for any values of the parameters. Alternatively, it might cause the system to converge to one which is either very far away, or whose rank is even lower than what had been desired.

Many applications admit single right-hand side (rank deficiency 1) problems. In that case, although it is still important to choose a good column for $B(a) \in \mathbb{R}^{m\times 1}$, there are only n possible columns to try. The multiple rank-deficiency problem is much more difficult. In practice, and especially for large problems, it is extremely difficult to select a good left-hand side from all $\binom{n}{r}$ possibilities for $A(a)$. The method of numerically identifying the τ-pivot columns, described in Section 4.2.2, has been useful here: construct $A(a)$ from the $r = r_\tau$ approximate pivot columns, and $B(a)$ from the remaining d approximately dependent columns.

The usual way of dealing with a multiple right-hand side problem [19], is to reformulate Equation (4.47) as:

$$Diag(A(a^0))x \approx b(a^0), \tag{4.49}$$

where $Diag(A(a))$ is a $(d\cdot m)\times(d\cdot n)$ diagonal matrix with $A(a)$ as the d diagonal blocks, and $b(a)$ is a $d\cdot m$ element vector formed by stacking the columns of $B(a)$ on top of each other. Now, every column of $B(a)$ is still in the range of the matrix $A(a)$.

4.5.3 STLS-RREF

Notice that Equation (4.49) is not good enough for the problem of “creating” exact non-pivot columns. (That is, where the τ-pivot columns make up $Diag(A(a^0))$ and the τ-dependent columns form $b(a^0)$.) As such, each column of $B(a^f)$ could quite likely be a linear combination of all columns of $A(a^f)$. This is not desirable.

We make the following improvement [24]. Reformulated for the RREF problem, the equation:

$$Diag(A_k(a^0))x \approx b(a^0), \tag{4.50}$$

is such that the k^{th} diagonal block, $A_k(a)$, now consists only of however many pivot columns appear to the left of the non-pivot column $b_k(a)$. (In total, there are still d diagonal blocks.) We use Equation (4.50) along with our implementation of STLS, whose purpose is to find a nearby a^f for which $Diag(A_k(a^f))x = b(a^f)$, has an exact solution. In this way, the numerical pivot identification method of Section 4.2.2 can be used to achieve convergence in STLS. In turn, STLS can be used to find a nearby $M(a^f)$ for which the columns of $A(a^f)$ and $B(a^f)$ are, respectively, the *exact* pivot and non-pivot columns of RREF($M(a^f)$). Together, we call this the *STLS-RREF method*. Other methods may work, in place of STLS, but we have not experimented with their implementations.

Example 4.5.2. Consider the following class of matrices:

$$M(a) = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ a_5 & a_6 & a_7 & a_8 \\ a_9 & a_{10} & a_{11} & a_{12} \\ a_{13} & a_{14} & a_{15} & a_{16} \end{pmatrix}. \tag{4.51}$$

Say, for a given set of values a^0 for the parameters, the columns 1 and 3 are the τ-pivot columns of $M(a^0)$. We want to converge to a nearby matrix whose ε-pivot columns are the 1^{st} and 3^{rd}. The following would be the input structure in STLS-RREF:

$$Diag(A_k(a)) = \begin{pmatrix} a_1 & 0 & 0 \\ a_5 & 0 & 0 \\ a_9 & 0 & 0 \\ a_{13} & 0 & 0 \\ 0 & a_1 & a_3 \\ 0 & a_5 & a_7 \\ 0 & a_9 & a_{11} \\ 0 & a_{13} & a_{15} \end{pmatrix}, \quad b(a) = \begin{pmatrix} a_2 \\ a_6 \\ a_{10} \\ a_{14} \\ a_4 \\ a_8 \\ a_{12} \\ a_{16} \end{pmatrix}. \tag{4.52}$$

STLS would then be applied to $Diag(A_k(a^0))$ and $b(a^0)$, and would search for a set of values a^f so (4.50) is satisfied exactly. Finally, one can use Proposition 4.4.5 to compute the two nonzero rows of the (exact) RREF of $M(a^f)$.

Finally, note that the STLS-RREF method can keep *additional* structure in the matrix $M(a^0)$. That is, in the output system, $M(a^f)$, relationships between the parameters will be preserved. We can then use Proposition 4.4.5 to compute the RREF of the (possibly structured) matrix $M(a^f)$, which now has exact pivot columns (as in Definition 4.2.6).

Acknowledgements

Greg Reid (GR) was supported in part by a Canadian NSERC Grant. Lihong Zhi (LZ) was supported in part by NKBRPC 2004CB318000 and Chinese NSF under Grant 10401035.

References

1. J. Bonasia, F. Lemaire, G. Reid, R. Scott, and L. Zhi. *Determination of Approximate Symmetries of Differential Equations.* Centre Recherches Mathématiques, CRM Proceedings and Lecture Notes, Vol. 39, pp. 233-250, 2004.
2. B. Buchberger. *An Algorithm for Finding a Basis for the Residue Class Ring of a Zero-Dimensional Ideal.* Ph.D. Thesis, Math. Inst., Univ. of Innsbruck, Austria, 1965.
3. James A. Cadzow. *Signal Enhancement - A Composite Property Mapping Algorithm.* IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 36, No. 1, pp. 49-62, January 1988.
4. Cartan, E. *Les Systèmes Différentiels Extérieurs et leurs Applications Géométriques.* Hermann, Paris, 1945.
5. David Cox, John Little, and Donal O'Shea. *Ideals, Varieties, and Algorithms.* Springer-Verlag New York, Inc., 1992.
6. Barry H. Dayton. *Numerical Local Rings and Solution of Nonlinear Systems.* Proc. International Workshop on Symbolic-Numeric Computation (SNC) Proceedings, pp. 79-86, 2007.
7. C. Eckart, G. Young. *The Approximation of one Matrix by Another of Lower Rank.* Psychometrika, Vol. 1, pp. 211-218, 1936.
8. Gerdt, V. and Blinkov, Y. *Involutive Bases of Polynomial Ideals.* Mathematics and Computers in Simulation, Vol. 45, pp. 519-541, 1998.
9. D. Hilbert. *Über die Theorie der algebraischen Formen.* Math. Annalen, Vol. 36, pp. 473-534, 1890.
10. Erich Kaltofen, Zhengfeng Yang, and Lihong Zhi. *Approximate Greatest Common Divisors of Several Polynomials with Linearly Constrained Coefficients and Singular Polynomials.* Proc. ISSAC 2006, Jean-Guillaume Dumas, Ed., ACM Press, pp. 169-176, 2006.
11. Kuranishi, M. *On E. Cartan's Prolongation Theorem of Exterior Differential Systems.* Amer. J. of Math., Vol. 79, pp. 1-47, 1957.
12. P. Lemmerling, N. Mastronardi, and S. Van Huffel. *Fast Algorithm for Solving the Hankel/Toeplitz Structured Total Least Squares Problem.* Numerical Algorithms, Vol. 23, pp. 371-392, 2000.
13. A. Leykin. *Numerical Primary Decomposition.* Proc. ISSAC 2008, ACM Press, pp. 165-172, 2008.
14. F. S. Macaulay. *The Algebraic Theory of Modular Systems.* Cambridge Tracts in Math. and Math. Physics, Vol. 19, 1916.
15. F. S. Macaulay. *Some Properties of Enumeration in the Theory of Modular Systems.* Proceedings of the London Mathematical Society, Vol. 26, pp. 531-555, 1927.
16. B. Malgrange. *Cartan Involutiveness = Mumford Regularity.* Contemporary Mathmatics, Vol. 331, pp. 193-205, 2003.
17. H.M. Möller and T. Sauer. *H-bases for polynomial interpolation and system solving.* Advances Comput. Math., Vol. 12, pp. 23-35, 2000.
18. B. Mourrain. *A New Criterion for Normal Form Algorithms.* In: Fossorier, M., Imai, H., Lin, S., Poli, A.(Eds.), AAECC. Vol. 1719. Springer, Berlin, pp. 430-443, 1999.
19. Haesun Park, Lei Zhang, and J. Ben Rosen. *Low Rank Approximation of a Hankel Matrix by Structured Total Least Norm.* BIT Numerical Mathematics, Springer Netherlands, Vol. 39, No. 4, pp. 757-779, December 1999.
20. Pommaret, J. F. *Systems of Partial Differential Equations and Lie Pseudogroups.* Gordon and Breach Science Publishers, 1978.
21. Greg Reid, Robin Scott, Wenyuan Wu, and Lihong Zhi. *Algebraic and Geometric Properties of Nearby Projectively Involutive Polynomial Systems* (preprint).
22. Greg Reid, Jianliang Tang, and Lihong Zhi. *A Complete Symbolic-Numeric Linear Method for Camera Pose Determination.* Proc. ISSAC 2003, ACM Press, pp. 215-233, 2003.
23. Greg Reid and Lihong Zhi. *Solving Polynomial Systems via Symbolic-Numeric Reduction to Geometric Involutive Form.* J. Symoblic Comput., Vol. 44, No.3, pp. 280-291, 2009.

24. Robin J. Scott. *Approximate Gröbner Bases - a Backwards Approach.* Master's Thesis, University of Western Ontario, 2006. Available at:
http://www.orcca.on.ca/$\sim$reid/Scott/SRWZ-ScottThesis06.pdf
25. W. M. Seiler. *Analysis and Application of the Formal Theory of Partial Differential Equations.* Ph.D. Thesis, Lancaster University, 1994.
26. Werner M. Seiler. *Involution - The Formal Theory of Differential Equations and its Applications in Computer Algebra and Numerical Analysis.* Habilitation Thesis, Univ. of Mannheim, 2002.
27. Spencer, D. *Overdetermined Systems of Linear Differential Equations.* Bulletin A.M.S. Vol. 75, pp. 179-239, 1969.
28. Hans J. Stetter. *Numerical Polynomial Algebra.* SIAM, 2004.
29. M. Trébuchet Philippe. *Vers une résolution stable et rapide des équations algébriques.* Ph.D. Thesis, l'Universite de Paris 6, 2002.
30. Wenyuan Wu and Greg Reid. *Application of Numerical Algebraic Geometry and Numerical Linear Algebra to PDE.* Proc. ISSAC 2006, ACM Press, pp. 345-352, 2006.

Chapter 5
Regularization and Matrix Computation in Numerical Polynomial Algebra

Zhonggang Zeng

Abstract Numerical polynomial algebra emerges as a growing field of study in recent years with a broad spectrum of applications and many robust algorithms. Among the challenges we must face when solving polynomial algebra problems with floating-point arithmetic, the most frequently encountered difficulties include the regularization of ill-posedness and the handling of large matrices. We develop regularization principles for reformulating the ill-posed algebraic problems, derive matrix computations arising in numerical polynomial algebra, as well as subspace strategies that substantially improve computational efficiency by reducing matrix sizes. Those strategies have been successfully applied to numerical polynomial algebra problems such as GCD, factorization, multiplicity structure and elimination.

Introduction

Accelerated by the advancement of computer algebra systems (CAS), symbolic algebraic computation has enjoyed tremendous success since the advent of the Gröbner basis and continues to be a driving force in computational commutative algebra. The abundance of algorithms along with the depth and breadth of the theories developed over the years for algebraic computation sets a solid foundation for numerical polynomial algebra, which emerges as a rapidly growing field of study in recently years. Pioneered by Li [57], Sommese [82] and others, numerical polynomial system solving based on the homotopy continuation method has set the standard in efficiency as well as robustness, and enters the stage of expansion into numerical algebraic geometry as a new field [81]. Meanwhile, numerical polynomial algebra emanating from fundamental numerical analysis and numerical linear

Zhonggang Zeng
Department of Mathematics, Northeastern Illinois University, Chicago, IL 60625, USA, Research supported in part by NSF under Grants DMS-0412003 and DMS-0715127,
e-mail: zzeng@neiu.edu

L. Robbiano, J. Abbott (eds.), *Approximate Commutative Algebra*, Texts and Monographs in Symbolic Computation, DOI 10.1007/978-3-211-99314-9_5, © Springer-Verlag/Wien 2009

algebra has also thrived in theory and problem solving, as presented in Stetter's recently published textbook [84] and evidenced by healthy development of software (cf. [86]).

In comparison to symbolic computation, numerical computation and approximate solutions offer substantial advantages in many areas and come with drawbacks in other aspects. Computing numerical solutions in many cases may be inevitable due to lack of alternatives, such as locating roots of a polynomial with a degree five or higher. Even in those cases where exact solutions are available, approximate solutions may make better sense and go deeper in revealing the physical nature. For a simple example, the exact GCD (greatest common divisor) of the polynomial pair

$$\left\{\begin{array}{rcl} f(x) & = & 3x^3 - 9.2426x^2 + 13.071x - 10 \\ g(x) & = & 1.7321x^4 + 2.4642x^2 - 2.4495x^3 + 1.4142x - 2 \end{array}\right. \tag{5.1}$$

is a trivial polynomial $u = 1$ in symbolic computation. However, its approximate GCD $\tilde{u} = 2.000006 - 1.4142x + x^2$ along with the verifiable residual provides more penetrating information: The given polynomial pair is a tiny distance 0.0000056 away from a polynomial pair having $\tilde{u}$ as its exact GCD. The difference here may mean a more desirable deblurred image restored by an approximate GCD [69] or a meaningless blank image represented by the trivial exact GCD. Furthermore, numerical computation tends to be more efficient in both storage and computing time, and more suitable for large engineering problems, as evidenced by the homotopy continuation method for solving polynomial systems applied to kinematics (cf. [82]).

Numerical polynomial algebra differs from symbolic computation in many aspects. The problem data are expected to be inexact and the computation is carried out with floating-point arithmetic. Accuracy and stability, which may not be a concern in symbolic computation, become of paramount importance in developing numerical algorithms for solving polynomial problems. A classical algorithm that is flawless in exact arithmetic, such as the Euclidean algorithm for computing polynomial GCD, may be prohibitively impractical for numerical computation and *vice versa.* As a result, it is often necessary to develop numerical algorithms from scratch and to employ totally different strategies from their symbolic cousins.

Even the meaning of "solution" may be in question and may depend on the objective in problem solving. The exact GCD $u = 1$ in (5.1) is indisputable in symbolic computation, while the meaning of the approximate GCD has evolved in several formulations over the years (see the remark in §5.2.2). The comparison between exact and approximate GCD is typical and common in algebraic problems where ill-posedness is often encountered and the solution is infinitely sensitive to data perturbations. Numerical computation which, by nature, seeks the exact solution of a nearby problem becomes unsuitable unless the problem is regularized as a well-posed one. As it turns out, the collections of ill-posed problems form pejorative manifolds of positive codimensions that entangled together with stratification structures, as elaborated in §5.2, where a manifold is embeded in the closure of manifolds of lower codimensions. Thus a tiny arbitrary perturbation pushes the problem away from its residing manifold, losing the very structure of the solution. Dubbed as the "three-strikes" principle for formulating an approximate solution and removing ill-posedness, we summarize that the approximate solution should be the exact solution of a *nearby problem* residing in the manifold of the *maximum codimension* and having the *minimum distance* to the given problem. Approximate solutions with such a formulation become viable, attainable, and continuous with respect to data.

Numerical polynomial algebra can also be regarded as having numerical linear algebra as one of its cornerstones. Pioneered by Wilkinson [93], Golub [34], Kahan [43] and many others around the same time of Buchberger's work on Gröbner basis , numerical linear algebra flourished since 1960s with well established theories, algorithms and software libraries for numerical matrix computations. One of the early catalysts that contributed to the recent advances of numerical polynomial algebra is the introduction of the singular value decomposition to polynomial computation by Corless, Gianni, Trager and Watt [11] in 1995. As elaborated in §5.3, polynomial algebra problems in numerical computation lead to matrix computation problems such as least squares, eigenproblem and particularly, numerical rank/kernel identification. Theories, techniques and ideas accumulated in numerical linear algebra are rich resources for numerical polynomial algebra.

This paper surveys basic approaches and techniques in developing algorithms for numerical polynomial algebra, emphasizing on regularizing ill-posed algebraic problems and employing matrix computation. The approaches and techniques elaborated in this survey have been used in many algorithms with successful results and implementations. As a growing field of study, theoretical advance and algorithm development are on-going and gradually evolving. Further advancement will continue to emerge in the future.

5.1 Notation and preliminaries

5.1.1 Notation

The n dimensional complex vector space is denoted by $\mathbb{C}^n$, in which vectors are column arrays denoted by boldface lower case letters such as $\mathbf{a}$, $\mathbf{u}$, $\mathbf{v}_2$, etc, with $\mathbf{0}$ being a zero vector whose dimension can be understood from the context. Matrices are represented by upper case letters like A and J, with $\mathbb{C}^{m\times n}$ denoting the vector space consists of all $m\times n$ complex matrices. Notations $(\cdot)^\top$ and $(\cdot)^{\mathsf{H}}$ stand for the transpose and the Hermitian transpose, respectively, of the matrix or vector $(\cdot)$.

The ring of polynomials with complex coefficients in indeterminates $x_1,\ldots,x_s$ is denoted by $\mathbb{C}[x_1,\ldots,x_s]$, or $\mathbb{C}[\mathbf{x}]$ for $\mathbf{x}=(x_1,\ldots,x_s)$. A polynomial as a function is denoted by a lower case letter, say f, v, or p_1, etc. The collection of all the polynomials with a certain degree bound forms a vector space over $\mathbb{C}$. Throughout this paper, if a letter (say f) represents a polynomial, then either

$[\![f]\!]$ or the same letter in boldface (*i.e.* $\mathbf{f}$) denotes its coefficient vector, where the underlying vector space and its basis are clear from the context.

5.1.2 Numerical rank and kernel

The fundamental difference between exact and numerical computation can be demonstrated in the meaning of the matrix rank. With exact arithmetic used in symbolic computation, a matrix is non-singular if and only if its determinant is nonzero. Such a characterization of being full rank, however, is practically meaningless in numerical computation as shown in a simple example below.

Example 5.1.1. The polynomial division is equivalent to linear system solving: Finding the quotient q and the remainder r of a polynomial f divided by g satisfying $f = g \cdot q + r$ is, in fact, solving a linear system $G \begin{bmatrix} \mathbf{q} \\ \mathbf{r} \end{bmatrix} = \mathbf{f}$, where $\mathbf{f}$, $\mathbf{q}$ and $\mathbf{r}$ are vector representations of f, q and r respectively, along with a corresponding matrix G. For instance, let $g(x) = x + 10$ and use the standard monomial basis for the vector representations of $\mathbf{f}$, $\mathbf{q}$ and $\mathbf{r}$. The matrix G is

$$G = \begin{bmatrix} 1 & & & & \\ 10 & 1 & & & \\ & 10 & \ddots & & \\ & & \ddots & 1 & \\ & & & 10 & 1 \end{bmatrix}_{(n+1)\times(n+1)} \tag{5.2}$$

where n is the degree of f. The matrix G may appear benign with a full rank and $\det(G) = 1$. However, its 2-norm distance to a singular matrix is less than 10^{-n}. Such a matrix with even a modest size, say $n = 15$, behaves the same way as a singular matrix in numerical computation. Consequently, round-off errors in the order of hardware precision during synthetic division can result in substantial errors of magnitude $O(1)$ in the coefficients of q and r (c.f. [98, §4.2.3]). The numerical rank of G should be n, not $n+1$ unless n is small. □

Remark: Example 5.1.1 indicates that the Euclidean algorithm can be highly unreliable in numerical computation of the polynomial GCD since it consists of recursive polynomial division in the form of $f = g \cdot q + r$. Interestingly, polynomial division in the form of $f = g \cdot q$, or equivalently the polynomial division in the form of $f = g \cdot q + r$ combined with an additional constraint $r = 0$, is a stable least squares problem. This can be seen from Example 5.1.1 by deleting the last column from matrix G. The resulting matrix possesses a near perfect condition number (≈ 1) with its smallest singular value larger than 9. □

The condition of a matrix in numerical computation depends on its distance to the nearest rank-deficient matrices. Likewise, the *numerical rank* (or approximate rank) of a matrix depends on the (exact) ranks of its nearby matrices. If a matrix A can have an error of magnitude θ, then the "worst" (*i.e.* lowest rank) matrix within a distance θ dictates its numerical behavior. The numerical rank of A within θ, denoted by $\mathit{rank}_{\theta}(A)$, is thus defined as the smallest rank of all matrices within a distance θ of A:

$$\mathit{rank}_{\theta}(A) \;=\; \min_{\|B-A\|_2 \le \theta} \mathit{rank}(B) \tag{5.3}$$

where $\mathit{rank}(\cdot)$ denotes the rank of matrix $(\cdot)$ in exact sense. Naturally, the exact kernel $\mathscr{K}(B)$ of B in (5.3) is taken as the *numerical kernel* $\mathscr{K}_\theta(A)$ of A within θ:

$$\mathscr{K}_\theta(A) = \mathscr{K}(B) \tag{5.4}$$
$$\text{where } \big\|B-A\big\|_2 = \min_{\mathit{rank}(C)=\mathit{rank}_{\theta}(A)} \big\|C-A\big\|_2.$$

The numerical rank and numerical kernel of a matrix $A \in \mathbb{C}^{m\times n}$ can equivalently be defined using the singular value decomposition (SVD) [34]:

$$A \;=\; \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{H}} + \cdots + \sigma_n \mathbf{u}_n \mathbf{v}_n^{\mathsf{H}} \;=\; U \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \\ & & & \\ & & & \end{bmatrix} V^{\mathsf{H}} \tag{5.5}$$

where $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n \ge 0$ are the singular values of A, the matrices

$$U \;=\; [\mathbf{u}_1, \cdots, \mathbf{u}_m] \in \mathbb{C}^{m\times m} \quad \text{and} \quad V \;=\; [\mathbf{v}_1, \cdots, \mathbf{v}_n] \in \mathbb{C}^{n\times n}$$

are unitary matrices whose columns are left and right singular vectors of A respectively. The numerical rank $\mathit{rank}_{\theta}(A) = k$ if and only if there are exactly k singular values of A lie above the threshold θ: $\sigma_k > \theta \ge \sigma_{k+1}$, and the numerical kernel $\mathscr{K}_\theta(A) = \mathit{span}\{\mathbf{v}_{k+1}, \ldots, \mathbf{v}_n\}$.

Computing the exact rank $\mathit{rank}(A)$ is an ill-posed problem in the sense that a tiny perturbation can alter the rank completely if A is of rank-deficient. As a result, the exact rank and kernel in general can not be computed in numerical computation since round-off errors are inevitable. By formulating numerical rank and kernel in (5.3) and (5.4) respectively, matrix rank-revealing is *regularized* as a well-posed problem and becomes suitable for numerical computation. In fact, the singular value decomposition is remarkably stable and well established in numerical linear algebra. The sensitivity of the numerical kernel can be measured by the condition number σ_1/σ_k by Wedin's Theorem [92].

On the other hand, the standard singular value decomposition can be unnecessarily costly to compute. The numerical rank/kernel computation in numerical polynomial algebra often involves matrices of low numerical nullities. For those matrices, numerical ranks and kernels can be computed efficiently using a specialized rank-revealing method [59], which has become an indispensable component of numerical polynomial algebra algorithms that will be discussed later in this survey. The rank-revealing method assumes the input matrix R is in upper-triangular form without loss of generality. Every matrix A has a QR decomposition [34] $A = QR$ and the numerical kernel of A is identical to that of the upper-triangular matrix R. The following null vector finder is at the core of the numerical rank-revealing method in [59]:

$$\left\{\begin{array}{l} \text{set } \mathbf{z}_0 \text{ as a random vector} \\ \text{for } j = 1, 2, \cdots \text{ do} \\ \quad \left\lfloor \begin{array}{l} \text{solve } \; R^{\mathsf{H}}\mathbf{x} = \mathbf{z}_{j\text{-}1} \;\; \text{with a forward substitution} \\ \text{solve } \; R\mathbf{y} = \mathbf{x} \;\; \text{with a backward substitution} \\ \text{set } \; \mathbf{z}_j = \frac{\mathbf{y}}{\|\mathbf{y}\|_2} \;\; \text{and} \;\; \varsigma_j = \frac{\|\mathbf{x}\|_2}{\|\mathbf{y}\|_2} \end{array}\right. \end{array}\right. \tag{5.6}$$

The iteration in (5.6) produces sequences $\{\varsigma_j\}$ and $\{\mathbf{z}_j\}$ satisfying

$$\lim_{j\to\infty} \varsigma_j \;=\; \sigma_n \quad \text{and} \quad \lim_{j\to\infty} \mathbf{z}_j \;=\; \mathbf{v}_n$$

where σ_n and $\mathbf{v}_n$ are the smallest singular value of A and the associated right singular vector respectively. Each iterative step of the algorithm (5.6) requires a forward substitution and a backward substitution on triangular matrices R^{H} and R respectively. After finding a numerical null vector $\mathbf{z}$ of matrix R within θ, insert a multiple of $\mathbf{z}$ on top of R to form

$$\hat{R} \;=\; \begin{bmatrix} \|R\|\mathbf{z}^{\mathsf{H}} \\ R \end{bmatrix}.$$

We can continue to calculate a numerical null vector $\hat{\mathbf{z}}$ of $\hat{R}$ within θ. If such a $\hat{\mathbf{z}}$ exists, it is also a numerical null vector of R orthogonal to $\mathbf{z}$ [59]. By updating the QR decomposition of $\hat{R}$, we can apply the algorithm (5.6) again to find $\hat{\mathbf{z}}$. An orthonormal basis for the numerical kernel $\mathscr{K}_\theta(A)$ can be obtained by continuing this process.

Numerical rank can also be considered a generalization of the conventional rank since $rank(A) = rank_\theta(A) = k$ whenever $0 < \theta < \sigma_k$. A matrix A in practical computation is usually known in a perturbed form $\hat{A} = A + E$ where E is the noise that can be expected to be small. Let $\sigma_j(\cdot)$ denote the j-th singular value of the matrix $(\cdot)$. It is known that [85, Chapter 1, Corollary 4.31]

$$|\sigma_j(A+E) - \sigma_j(A)| \;\leq\; \|E\|_2, \quad j = 1, \cdots, n.$$

Consequently, the underlying (exact) rank $rank(A)$ can be recovered as $rank_\theta(A)$ as long as

$$\|E\|_2 \;<\; \theta \;<\; \sigma_k(A) - \|E\|_2.$$

In practical computation, the choice of the threshold θ is problem dependent, and may be difficult to decide in some cases. The general rule is that θ should be above the expected noise magnitude $\|E\|_2$ and below the smallest positive (unknown) singular value $\sigma_k(A)$.

The formulation of the numerical rank is intended to answer the following question: *How to find the rank of a matrix A that is under a small perturbation?* The answer is conditional: *If the perturbation is sufficiently small, then the rank of A can be recovered as $rank_\theta(A)$ for a threshold slightly larger than the perturbation.* Numerical rank-revealing would become intractable when the window $\sigma_k(A) - \|E\|_2 > \theta > \|E\|_2$ of choosing the threshold θ disappears.

Consider Example 5.1.1 again. For any threshold θ chosen within the interval $10^{-n} < \theta < 9$, the matrix G in (5.2) is of numerical rank n, or numerical nullity 1 within θ.

5.1.3 The linear and nonlinear least squares problems

Conventional solutions do not exist for an overdetermined linear system

$$A\mathbf{x} \;=\; \mathbf{b}$$

when the matrix $A \in \mathbb{C}^{m \times n}$ for $m > n$ unless, for a zero probability, the vector $\mathbf{b}$ happens to be in the range of A. Instead, the *least squares solution* $\mathbf{x}_*$ becomes the alternative that satisfies

$$\left\|A\mathbf{x}_* - \mathbf{b}\right\|_2^2 \;=\; \min_{\mathbf{y}\in\mathbb{C}^n} \left\|A\mathbf{y} - \mathbf{b}\right\|_2^2, \tag{5.7}$$

The least squares solution is unique: $\mathbf{x}_* = A^+\mathbf{b}$ when A is of full rank, where $A^+ = (A^{\mathsf{H}}A)^{-1}A^{\mathsf{H}}$ is the *pseudo-inverse* of A.

Although solving $A\mathbf{x} = \mathbf{b}$ for its least squares solution is equivalent to solving the normal equation $(A^{\mathsf{H}}A)\mathbf{x} = A^{\mathsf{H}}\mathbf{b}$, there is a fundamental difference between symbolic computation and numerical approach from here. Solving the normal equation directly may be a natural approach using exact arithmetic. Numerical computation, where accuracy is a concern, goes a step further by solving the first n equations of $R\mathbf{x} = Q^{\mathsf{H}}\mathbf{b}$ after obtaining the QR decomposition $A = QR$ (c.f. [34, §5.3]). Both approaches are equivalent in theory but not much so in practical computations. It is neither wise nor necessary to construct or to solve the normal equation as it is in numerical computation. On the other hand, the standard numerical approach of solving $R\mathbf{x} = Q^{\mathsf{H}}\mathbf{b}$ is not attractive for symbolic computation due to the square roots required in the QR decomposition.

Solving linear least squares solutions is essential in one of the most basic operations in numerical polynomial algebra: Polynomial division in floating-point arithmetic. As discussed in Example 5.1.1, dividing a polynomial f by g for the quotient q and the remainder r in the form of $f = g \cdot q + r$ can easily be ill-conditioned. Consequently, the synthetic division is prohibitively unstable in numerical computation. However, if the remainder is known, say $r = 0$, then finding

q in the equation $g \cdot q = f$ is a stable linear least squares problem

$$C(g)\,\mathbf{q} \;=\; \mathbf{f} \tag{5.8}$$

where $C(g)$ is the convolution matrix [17] representing the linear transformation $\mathscr{L} : u \longrightarrow g \cdot u$ between relevant polynomial vector spaces in which vectors $\mathbf{q}$ and $\mathbf{f}$ represent q and f respectively. Equation (5.8) can be accurately solved in numerical computation for its least squares solution, avoiding the difficulty of synthetic division. This *least squares division* plays an indispensable role in many works such as [30, 98, 102].

The least squares problem also arises in polynomial algebra (c.f. [48, 98, 102] and Example 5.1.3 below) in a nonlinear form of solving an overdetermined system of (nonlinear) equations

$$F(\mathbf{z}) \;=\; \mathbf{b}, \quad \mathbf{z} \in \mathbb{C}^n \tag{5.9}$$

for an analytic mapping $F : \mathbb{C}^n \to \mathbb{C}^m$ with $m > n$. Similar to the linear case, solving (5.9) requires seeking the least squares solution $\mathbf{z}_*$ defined by

$$\big\|F(\mathbf{z}_*) - \mathbf{b}\big\|_2^2 \;=\; \min_{\mathbf{y} \in \mathbb{C}^n} \big\|F(\mathbf{y}) - \mathbf{b}\big\|_2^2.$$

The necessary condition for $\mathbf{z}_*$ to be a (local or global) least squares solution is

$$J(\mathbf{z})^{\mathsf{H}} F(\mathbf{z}) \;=\; \mathbf{0} \tag{5.10}$$

where $J(\mathbf{z})$ is the Jacobian of $F(\mathbf{z})$ (cf. [20, 98]). Although the global minimum would be attractive theoretically, a local minimum is usually sufficient in practical computations, and the local minimum is global if the residual $\|F(\mathbf{z}_*) - \mathbf{b}\|_2$ is tiny. In principle, the least squares solution of (5.9) can be solved by many nonlinear optimization methods. There is a distinct feature of the overdetermined systems arising from numerical polynomial algebra: The residual $\|F(\mathbf{z}_*) - \mathbf{b}\|_2$ is expected to be small (*e.g.* an approximate factorization $p_1^{m_1} \cdots p_k^{m_k}$ of f is expected to satisfy $\big\|p_1^{m_1} \cdots p_k^{m_k} - f\big\| \ll 1$). As a result, a simple optimization method, the Gauss-Newton iteration, is particularly effective.

The Gauss-Newton iteration is given as follows: From an initial iterate $\mathbf{z}_0$,

$$\mathbf{z}_k \;=\; \mathbf{z}_{k\text{-}1} - J(\mathbf{z}_{k\text{-}1})^+ \big[F(\mathbf{z}_{k\text{-}1}) - \mathbf{b}\big], \quad k = 1, 2, \cdots. \tag{5.11}$$

The Gauss-Newton iteration is a natural generalization of the standard Newton iteration. Detailed studies of the Gauss-Newton iteration can be found in some special topic textbooks and articles, such as [18, 20, 63, 98].

The Gauss-Newton iteration is well defined in a neighborhood of the desired (local or global) minimum point $\mathbf{z}_*$ if the Jacobian $J(\mathbf{z}_*)$ is injective (or, equivalently, of nullity zero). The condition of being injective on $J(\mathbf{z}_*)$ is also essential to ensure the local convergence of the Gauss-Newton iteration, as asserted in the following lemma. Different from Newton's iteration for normally determined systems, however, the locality of the convergence consists *two* requirements: The initial iterate $\mathbf{z}_0$ must be sufficiently near the least squares solution $\mathbf{z}_*$, while the residual $\|F(\mathbf{z}_*)-\mathbf{b}\|_2$ must be sufficiently small.

Lemma 5.1.2. [98, Lemma 2.8] *Let $\Omega \subset \mathbb{C}^m$ be a bounded open convex set and $F : D \subset \mathbb{C}^m \longrightarrow \mathbb{C}^n$ be analytic in an open set $D \supset \overline{\Omega}$. Let $J(\mathbf{z})$ be the Jacobian of $F(\mathbf{z})$. Assume $\mathbf{z}_* \in \Omega$ is a local least squares solution to system* (5.9). *Let σ be the smallest singular value of $J(\mathbf{z}_*)$. Let $\delta \geq 0$ be a constant such that*

$$\left\| \left[J(\mathbf{z}) - J(\mathbf{z}_*) \right]^{\mathsf{H}} \left[F(\mathbf{z}_*) - \mathbf{b} \right] \right\|_2 \;\leq\; \delta \left\| \mathbf{z} - \mathbf{z}_* \right\|_2 \quad \textit{for all } \mathbf{z} \in \Omega.$$

If $\delta < \sigma^2$, then for any $c \in \left(\frac{1}{\sigma}, \frac{\sigma}{\delta}\right)$, there exists $\varepsilon > 0$ such that for all $\mathbf{z}_0 \in \Omega$ with $\|\mathbf{z}_0 - \mathbf{z}_*\|_2 < \varepsilon$, the sequence $\{\mathbf{z}_1, \mathbf{z}_2, \cdots\}$ generated by the Gauss-Newton iteration (5.11) *is well defined inside Ω, converges to $\mathbf{z}_*$, and satisfies*

$$\left\| \mathbf{z}_{k+1} - \mathbf{z}_* \right\|_2 \;\leq\; \frac{c\delta}{\sigma} \left\| \mathbf{z}_k - \mathbf{z}_* \right\|_2 + \frac{c\alpha\gamma}{2\sigma} \left\| \mathbf{z}_k - \mathbf{z}_* \right\|_2^2, \tag{5.12}$$

where $\alpha > 0$ is the upper bound of $\|J(\mathbf{z})\|_2$ on $\overline{\Omega}$, and $\gamma > 0$ is the Lipschitz constant of $J(\mathbf{z})$ in Ω, namely, $\|J(\mathbf{z}+\mathbf{h}) - J(\mathbf{z})\|_2 \leq \gamma \|\mathbf{h}\|$ for all $\mathbf{z}, \mathbf{z}+\mathbf{h} \in \Omega$.

Notice that the constant δ in (5.12) is proportional to the residual $\|F(\mathbf{z}_*)-\mathbf{b}\|_2$. Therefore, the smaller is this residual, the faster is the convergence. When the least squares solution $\mathbf{z}_*$ is a conventional solution, the residual will be zero and the convergence is quadratic.

The condition that $J(\mathbf{z}_*)$ is injective also implies that the smallest singular value $\sigma_{min}(J(\mathbf{z}_*))$ is strictly positive and provides an asymptotic sensitivity measurement [98]

$$\tau(F, \mathbf{z}_*) \;=\; \frac{1}{\sigma_{min}(J(\mathbf{z}_*))} \;\equiv\; \left\| J(\mathbf{z}_*)^+ \right\|_2 \tag{5.13}$$

for the least squares solution $\mathbf{z}_*$.

The Gauss-Newton iteration is extensively applied in the algorithms for numerical polynomial algebra in this survey. A generic Gauss-Newton iteration module is available in the software packages `ApaTools/Apalab` [99]. When formulating the overdetermined system $F(\mathbf{z}) = \mathbf{b}$, it is important to have enough equations to ensure $J(\mathbf{z}_*)$ is injective. Auxiliary equations are often needed for this purpose as shown in the following example.

Example 5.1.3. Consider the polynomial factorization problem. For simplicity of the exposition, assume f can be factorized in three factors $f = u^\alpha v^\beta w^\gamma$ where u, v and w are pairwise co-prime polynomials, and the objective is to compute the coefficient vectors $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ of u, v and w respectively. Naturally, the overdetermined system $G(\mathbf{u}, \mathbf{v}, \mathbf{w}) = \mathbf{0}$ with

$$G(\mathbf{u}, \mathbf{v}, \mathbf{w}) \;\equiv\; [\![u^\alpha v^\beta w^\gamma - f]\!] \tag{5.14}$$

needs to be solved, as pointed out in [48], where $[\![\cdot]\!]$ denotes the vector representation of the polynomial $(\cdot)$ in a given vector space. However, the Jacobian of

$G(\mathbf{u},\mathbf{v},\mathbf{w})$ in (5.14) is not injective since $G(\mathbf{u},\mathbf{v},\mathbf{w}) = G(c_1\mathbf{u},c_2\mathbf{v},c_3\mathbf{w})$ as long as $c_1^\alpha c_2^\beta c_3^\gamma = 1$. More constraints are therefore needed. A simple remedy we typically employ is to include extra equations $\mathbf{a}^{\mathsf{H}}\mathbf{u} - 1 = \mathbf{b}^{\mathsf{H}}\mathbf{v} - 1 = 0$ with certain vectors $\mathbf{a}$ and $\mathbf{b}$ of proper dimensions. The only restriction for choosing $\mathbf{a}$ and $\mathbf{b}$ is to avoid $\mathbf{a}^{\mathsf{H}}\mathbf{u} = \mathbf{b}^{\mathsf{H}}\mathbf{v} = \mathbf{0}$ near the solution u and v. They can be random vectors, or more preferably the scalar multiples of the initial approximations to $\mathbf{u}$ and $\mathbf{v}$ respectively. Then the Jacobian $J(\mathbf{u},\mathbf{v},\mathbf{w})$ of the analytic mapping

$$F(\mathbf{u},\mathbf{v},\mathbf{w}) \;=\; \begin{bmatrix} \mathbf{a}^{\mathsf{H}}\mathbf{u} - 1 \\ \mathbf{b}^{\mathsf{H}}\mathbf{v} - 1 \\ [\![u^\alpha v^\beta w^\gamma - f]\!] \end{bmatrix}$$

can be easily proved as injective at the desired solution $(\mathbf{u},\mathbf{v},\mathbf{w})$: Assume

$$J(\mathbf{u},\mathbf{v},\mathbf{w}) \begin{bmatrix} \mathbf{p} \\ \mathbf{q} \\ \mathbf{r} \end{bmatrix} \;=\; \mathbf{0} \tag{5.15}$$

where $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{r}$ are vector representations of polynomials p, q and r respectively. From straightforward differentiations, we have

$$\begin{aligned} \alpha u^{\alpha\text{-}1} v^\beta w^\gamma p + \beta u^\alpha v^{\beta\text{-}1} w^\gamma q + \gamma u^\alpha v^\beta w^{\gamma\text{-}1} r &= 0 \\ \mathbf{a}^{\mathsf{H}}\mathbf{p} \;=\; \mathbf{b}^{\mathsf{H}}\mathbf{q} &= 0. \end{aligned}$$

Consequently $\alpha pvw + \beta uqw + \gamma uvr = 0$, which leads to $p = g_1 u$, $q = g_2 v$ and $r = g_3 w$ since u, v, and w are pairwise co-prime. We further know that g_1, g_2 and g_3 must be constants from (5.15). Then $\mathbf{a}^{\mathsf{H}}\mathbf{p} = \mathbf{b}^{\mathsf{H}}\mathbf{q} = 0$ and $\mathbf{a}^{\mathsf{H}}\mathbf{u} = \mathbf{b}^{\mathsf{H}}\mathbf{v} = 1$ implies $g_1 = g_2 = 0$ and thus $p = q = 0$. Furthermore $\gamma uvr = 0$ implies $r = 0$. Therefore $J(\mathbf{u},\mathbf{v},\mathbf{w})$ is injective. □

Appending extra equations as shown in Example 5.1.3 is a typical strategy for making the Jacobian injective before applying the Gauss-Newton iteration. The approaches of proving the injectiveness are also similar among different problems.

5.2 Formulation of the approximate solution

5.2.1 The ill-posed problem and the pejorative manifold

Hadamard characterized a problem as *well-posed* if its solution satisfies existence, uniqueness, and continuity with respect to data [36]. A problem whose solution lacks one of the three properties is termed an *ill-posed problem*. In the current literature, the term ill-posed problem mainly refers to those having solutions infinitely sensitive to data perturbations [19]. Contrary to Hadamard's misbelief that ill-posed problems are artificial and unsuitable for modeling physical systems, this type of problems arise quite often in science and engineering applications (cf. *e.g.* [37,

§1.2]). As one of the main challenges for numerical computation, algebraic problems are commonly ill-posed. Such examples are abundant:

- *Univariate polynomial factorization:* Under an arbitrary tiny perturbation, a polynomial $p(x) = (x-x_1)^{m_1} \cdots (x-x_k)^{m_k}$ in expanded form loses those repeated factors corresponding to multiplicities m_j's higher than one, and multiple roots turn into clusters of simple roots.

- *Polynomial GCD:* For a given pair of polynomials (p,q) under arbitrary perturbations, its GCD generically degrades to a trivial constant polynomial even if (p,q) originally possesses a GCD of positive degree.

- *Irreducible factorization of multivariate polynomials:* For a multivariate polynomial $p = f_1^{\alpha_1} f_2^{\alpha_2} \cdots f_k^{\alpha_k}$ with nontrivial factors $f_1, \ldots, f_k$, the factorizability is lost and the polynomial becomes irreducible under an infinitesimal but arbitrary perturbation. The irreducible factorization of polynomial p is thus discontinuous, preventing a meaningful solution using conventional methods at presence of data perturbation.

- *The matrix rank and kernel:* As perhaps the most basic ill-posed problem (c.f. §5.1.2), the rank and kernel of a matrix are discontinuous when the matrix is rank-deficient. A tiny perturbation will generically destroy the kernel entirely and turns the matrix into a full-ranked one. The ill-posedness of matrix rank/kernel may be under-noticed because the singular value decomposition is available and effective. The capability of identifying the *numerical* rank and kernel is one of the very cornerstones of methods for solving ill-posed problems. This ill-posed problem extends to solving a linear system $A\mathbf{x} = \mathbf{b}$ where the matrix A is rank-deficient. A perturbation generically renders the system unsolvable in exact sense even if the original system has infinitely many solutions.

- *The matrix Jordan Canonical Form:* For an $n \times n$ matrix A, there is a Jordan canonical decomposition $A = XJX^{-1}$ where J is the block diagonal Jordan matrix. When any of the Jordan block in J is larger than 1×1, however, the perturbed matrix $A+E$ generically loses all the non-trivial Jordan structure of A, making it extremely difficult to compute the Jordan decomposition and the underlying multiple eigenvalues.

In a seminal technical report [43] that has never been formally published, Kahan studied three ill-posed problems (rank-deficient linear system, multiple roots of polynomials and multiple eigenvalues of matrices) and pointed out that it may be a misconception to consider ill-posed problems as hypersensitive to data perturbations. The collection of those problems having solutions of a common structure forms what Kahan calls a *pejorative manifold*. An artificial perturbation pushes the problem away from the manifold in which it originally resides and destroys its solution structure. Kahan proves, however, the solution may not be sensitive at all if the problem is perturbed with a restriction that it surfs in the pejorative manifold it belongs.

Kahan's observation of pejorative manifolds extends to other ill-posed problems such as those in the following examples. Furthermore, it has become known that these manifolds may have a certain *stratification* structure.

Example 5.2.1. In the vector space $\mathbb{C}^{m\times n}$ for $m \geq n$, the collection of rank-k matrices forms a pejorative manifold

$$\mathscr{M}_k^{m\times n} \;=\; \big\{A \in \mathbb{C}^{m\times n} \;\big|\; \mathit{rank}(A) = k\big\}.$$

Counting the dimension from the basis for the column space $(m\cdot k)$ and the remaining columns as linear combinations of the basis $((n-k)\cdot k)$, it is easy to see [58] that the codimension $\mathit{codim}(\mathscr{M}_k^{m\times n}) = (m-k)(n-k)$. Those manifolds form a stratification structure

$$\overline{\mathscr{M}_0^{m\times n}} \subset \overline{\mathscr{M}_1^{m\times n}} \subset \cdots \subset \overline{\mathscr{M}_n^{m\times n}} \equiv \mathbb{C}^{m\times n},$$

where $\overline{(\cdot)}$ denotes the closure of the set $(\cdot)$. □

Example 5.2.2. Associating a monic polynomial

$$p(x) = x^4 + p_1x^3 + p_2x^2 + p_3x + p_4$$

of degree 4 with the coefficient vector $\mathbf{p} = [p_1, p_2, p_3, p_4]^\top \in \mathbb{C}^4$, the set of all polynomials possessing a common factorization structure forms a factorization manifold. For instance

$$\Pi(1,3) \;=\; \big\{(x-z_1)^1(x-z_2)^3 \;\big|\; z_1, z_2 \in \mathbb{C},\; z_1 \neq z_2\big\}$$

with codimension $\mathit{codim}(\Pi(1,3)) = 2$. On the other hand, manifold $\Pi(1,3)$ is in the closure of manifold

$$\Pi(1,1,2) \;=\; \big\{(x-z_1)^1(x-z_2)^1(x-z_3)^2 \;\big|\; z_1, z_2, z_3 \in \mathbb{C},\; z_i \neq z_j \text{ for } i \neq j\big\}$$

with $\mathit{codim}\big(\Pi(1,1,2)\big) = 1$ since

$$\lim_{\varepsilon\to 0}(x-z_1)^1(x-z_2)^1(x-z_2+\varepsilon)^2 \;=\; (x-z_1)^1(x-z_2)^3$$

Likewise $\Pi(1,1,2) \subset \overline{\Pi(1,1,1,1)} \equiv \mathbb{C}^4$, and the five manifolds form a *stratification* as shown in Figure 5.1. Moreover, each manifold can be parametrized in the form of $\mathbf{p} = F(\mathbf{z})$. For instance, $\Pi(1,3)$ is parametrized as

$$\mathbf{p} \;=\; [\![(x-z_1)(x-z_2)^3]\!],$$

namely,

$$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} \;=\; \begin{bmatrix} -3z_2 - z_1 \\ 3z_2^2 + 3z_1z_2 \\ -z_2^3 - 3z_1z_2^2 \\ z_1z_2^3 \end{bmatrix}$$

The Jacobian of such $F(\mathbf{z})$ is injective in the corresponding manifold, ensuring the local Lipschitz continuity of the root vector $\mathbf{z} = [z_1, z_2]^\top$ with respect to the coefficients $\mathbf{p}$ as well as the capability of applying the Gauss-Newton iteration for refining the multiple roots [98]. □

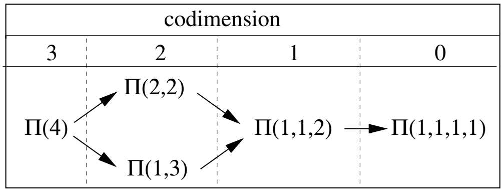


Fig. 5.1 Stratification of manifolds of degree 4 polynomials, with "$\longrightarrow$" denoting "in the closure of"

Example 5.2.3. Consider univariate polynomial pairs (p,q) of degrees m and n respectively with $m \geq n$. Let $\mathscr{P}_k^{m,n}$ be the collection of those pairs having GCDs of a common degree k:

$$\mathscr{P}_k^{m,n} = \big\{(p,q) \in \mathbb{C}[x] \times \mathbb{C}[x] \;\big|\; \\ \mathscr{deg}(p) = m, \;\; \mathscr{deg}(q) = n, \;\; \mathscr{deg}\big(\mathscr{gcd}(p,q)\big) = k\big\} \tag{5.16}$$

where $\mathscr{gcd}(p,q)$ is the GCD of the polynomial pair (p,q). Every polynomial pair $(p,q) \in \mathscr{P}_k^{m,n}$ can be written as $p = uv$ and $q = uw$ where u is a monic polynomial of degree k. Thus it is easy to see that $\mathscr{P}_k^{m,n}$ is a manifold of the dimension $k + (m-k+1) + (n-k+1)$, or codimension k exactly, and those GCD manifolds form a stratification structure

$$\overline{\mathscr{P}_n^{m,n}} \subset \overline{\mathscr{P}_{n\text{-}1}^{m,n}} \subset \cdots \subset \overline{\mathscr{P}_0^{m,n}} \equiv \mathbb{C}^{m+n+2}. \tag{5.17}$$

Furthermore, each manifold can again be parametrized in the form of $\mathbf{u} = F(\mathbf{z})$. In fact,

$$\mathbf{u} \;=\; \begin{bmatrix} \mathbf{p} \\ \mathbf{q} \end{bmatrix} = \begin{bmatrix} \llbracket uv \rrbracket \\ \llbracket uw \rrbracket \end{bmatrix} \;=\; F(\mathbf{z})$$

with $\deg(u) = k$, where $\mathbf{z}$ is a vector consists of the coefficients of u, v and w except the leading coefficient 1 of u. Also similar to Example 5.2.2, the Jacobian of $F(\mathbf{z})$ is injective, ensuring the local Lipschitz continuity of (u,v,w) with respect to (p,q) on the manifold $\mathscr{P}_k^{m,n}$ as well as the applicability of the Gauss-Newton iteration for refinement. □

In summary, there exist similar geometric structures for many ill-posed problems in polynomial algebra such as multivariate GCD, multivariate square-free factorization, and multivariate irreducible factorization: The collection of the problems sharing a common solution structure forms a pejorative manifold $\mathscr{P}$ that can be parametrized in the form of $\mathbf{u} = F(\mathbf{z})$ where the Jacobian of $F(\mathbf{z})$ is injective on the manifold $\mathscr{P}$. In addition, the pejorative manifolds form a stratification structure in which a manifold is in the closure of some other manifolds of lower codimension. As a result, a problem may be near many manifolds of lower codimension if it is near (or resides in) one particular pejorative manifold.

When an ill-posed problem with inexact data is encountered and it needs to be solved approximately using floating-point arithmetic, the first and foremost question is the meaning of "solving the problem". Computing the exact solution of an inexact ill-posed problem, such as calculating the trivial constant GCD, does not make much sense in practical applications. On the other hand, approximate solutions need to possess continuity. That is, the approximate solution must converge to the exact solution if the problem noise approaches zero. It appears reasonable to set the objective of solving an ill-posed problem numerically as follows:

> *Let P be a given problem that is a small perturbation from an ill-posed problem $\hat{P}$ residing in a pejorative manifold Π with exact solution $\hat{S}$. Find an approximate solution $\tilde{S}$ of P in the sense that $\tilde{S}$ is the exact solution of a certain problem $\tilde{P}$ where $\tilde{P}$ belongs to the same manifold Π and $\|\tilde{S} - \hat{S}\| = O(\|P - \hat{P}\|)$.*

5.2.2 The three-strikes principle for removing ill-posedness

We shall use the univariate factorization problem in Example 5.2.2 as a case study for a rigorous formulation of the approximate solution that removes the ill-posedness. When the given polynomial p is a small perturbation from $\hat{p} = (x - z_1)(x - z_2)^3$ that belongs to the factorization manifold $\Pi(1,3)$ of codimension 2, the stratification structure as shown in Figure 5.1 indicates that p is also near two other manifolds $\Pi(1,1,2)$ and $\Pi(1,1,1,1)$ of lower codimensions 1 and 0 respectively. Here the distance $\delta(p,\Pi)$ of p from a manifold can be naturally defined as

$$\delta(p,\Pi) \;=\; \inf_{q\in\Pi} \|p - q\|.$$

Actually, the polynomial p is generically closer to those "other" manifolds than it is to the correct one:

$$\delta(p,\Pi(1,3)) \;\geq\; \delta(p,\Pi(1,1,2)) \;\geq\; \delta(p,\Pi(1,1,1,1)).$$

Let μ be the minimum of the distances between p and manifolds $\Pi(2,2)$ and $\Pi(4)$ and assume the perturbation is sufficiently small such that

$$\|p - \hat{p}\| \;\leq\; \delta(p,\Pi(1,3)) \;\ll\; \mu.$$

Then the desired manifold $\Pi(1,3)$ stands out as the *highest codimension* manifold among all the manifolds that intersect the θ-neighborhood of p for every θ satisfying $\delta(p,\Pi(1,3)) < \theta < \mu$. That is, the key to identifying the correct manifold is to seek the manifold of the *highest codimension* within a proper threshold θ.

Upon identifying the pejorative manifold $\Pi(1,3)$, it is natural to look for $\tilde{p} = (x-\tilde{z}_1)(x-\tilde{z}_2)^3 \in \Pi(1,3)$ that minimizes the distance $\|p-\tilde{p}\|$, and take the exact factorization $(x-\tilde{z}_1)(x-\tilde{z}_2)^3$ of $\tilde{p}$ as the *approximate factorization* of p.

Definition 5.2.4. Let p be a given polynomial of degree n and let $\theta > 0$. Assume $m_1+\cdots+m_k = n$ and $\Pi(m_1,\ldots,m_k)$ is of the highest codimension among all the factorization manifolds in $\mathbb{C}^n$ that intersect the θ-neighborhood of p. Then the exact factorization of $\tilde{p} \in \overline{\Pi(m_1,\ldots,m_k)}$ is called the approximate factorization of p if

$$\|p-\tilde{p}\| \;=\; \min_{q\in\overline{\Pi(m_1,\cdots,m_k)}} \|p-q\|.$$

A theoretical elaboration of the univariate factorization and its regularization is presented in a recent paper [101]. Generally, solving an ill-posed algebraic problem starts with formulating the *approximate solution* following the "three-strikes" principle below to remove the discontinuity:

Backward nearness: The approximate solution to the given (inexact) problem P is the exact solution of a nearby problem $\tilde{P}$ within a distance θ of P.

Maximum codimension: The nearby problem $\tilde{P}$ is in the closure $\overline{\Pi}$ of the pejorative manifold Π of the highest codimension among all the pejorative manifolds intersecting the θ-neighborhood of the given problem P.

Minimum distance: The nearby problem $\tilde{P}$ is the nearest point in the closure $\overline{\Pi}$ of Π to the given problem P

Based on these principles, the *approximate GCD* of a polynomial pair can be defined similarly using the notation in Example 5.2.3.

Definition 5.2.5. [96] Let (p,q) be a given polynomial pair of degree m and n respectively, and let $\theta > 0$ be a given GCD threshold. Assume k is the maximum codimension among all the GCD manifolds $\mathscr{P}_0^{m,n}, \mathscr{P}_1^{m,n},\ldots,\mathscr{P}_n^{m,n}$ embedded in $\mathbb{C}^{m+n+2}$ that intersect the θ-neighborhood of (p,q):

$$k \;=\; \max_{\delta((p,q),\mathscr{P}_j^{m,n})<\theta} codim\big(\mathscr{P}_j^{m,n}\big)$$

Then the exact GCD of $(\tilde{p},\tilde{q}) \in \overline{\mathscr{P}_k^{m,n}}$ is called the approximate GCD of (p,q) within θ if

$$\|(p,q)-(\tilde{p},\tilde{q})\| \;=\; \min_{(f,g)\in\overline{\mathscr{P}_k^{m,n}}} \|(p,q)-(f,g)\|. \tag{5.18}$$

Remark: The univariate GCD is the first ill-posed problem in numerical polynomial algebra that has been going through rigorous formulations. In 1985, Schönhage [77] proposed the *quasi-GCD* for univariate polynomials in a definition requiring only the backward nearness. Schönhage also assumes the given polynomial pair is inexact but arbitrarily precise. Similar formulations are later used in [14, 30, 39, 67, 68, 72] with some variations. In 1995, Corless, Gianni, Trager and Watt [11] first noticed the importance of the "highest degree" requirement of the approximate GCD in addition to Schönhage's notion. This requirement is also adopted in [24]. In the same paper [11], Corless et al. also suggest seeking the minimum distance. In 1996/1998 Karmarkar and Lakshman [52, 53] formulated two numerical GCD problems: "The nearest GCD problem" and the "highest degree approximate common divisor problem" in detail, with the latter specifying requirements of backward nearness, highest degree, and minimum distance. The formulation in Definition 5.2.5 differs somewhat with Karmarkar-Lakshman's. The polynomial pair $(\tilde{p},\tilde{q})$ is on the *closure* of the manifold of highest degree polynomial pairs and not necessarily monic. There is a different type of numerical GCD formulation: The "nearest GCD problem". Originated by Karmarkar and Lakshman [53, p. 654], this problem seeks the nearest polynomial pairs possessing an exact nontrivial GCD. Kaltofen, Yang and Zhi (cf. *e.g.* [49, 50]) generalized this formulation: Given a polynomial pair (p,q) and an integer $k > 0$, find the polynomial pair $(\tilde{p},\tilde{q})$ that is nearest from (p,q) such that the (exact) GCD of $(\tilde{p},\tilde{q})$ is of degree k or higher. This problem is equivalent to the constrained minmization (5.18) as a stand-alone problem with no need to specify a tolerance θ or to maximize the manifold codimension. □

In a straightforward verification, the formulation of the numerical kernel of a matrix in §5.1.2 conforms with the "three-strikes" principle. We can formulate the approximate multivariate GCD, the approximate square-free factorization, the approximate irreducible factorization, the approximate Jordan Canonical Form and other ill-posed problems the same way as and Definition 5.2.4 and Definition 5.2.5 according to the principles of backward nearness, maximum codimension and minimum distance.

Computing the approximate solution as formulated above involves identification of the pejorative manifold of the highest codimension within the given threshold as well as solving a least squares problem to obtain the minimum distance. Algorithms can be developed using a two-staged approach: Finding the manifold with matrix computation, followed by applying the Gauss-Newton iteration to obtain the approximate solution.

Example 5.2.6. The effectiveness of this formulation and the robustness of the corresponding two-staged algorithms can be illustrated by the polynomial factorization problem for

$$\begin{aligned} p(x) &= x^{200} - 400x^{199} + 79500x^{198} + \ldots + 2.04126914035338 \cdot 10^{86}\, x^{100} \\ &\quad -3.55467815448396 \cdot 10^{86}\, x^{99} + \ldots + 1.261349023419937 \cdot 10^{53}\, x^{2} \\ &\quad -1.977831229290266 \cdot 10^{51}\, x + 1.541167191654753 \cdot 10^{49} \\ &\approx (x-1)^{80}(x-2)^{60}(x-3)^{40}(x-4)^{20} \end{aligned} \tag{5.19}$$

whose coefficients are rounded to hardware precision. A new algorithm uvFactor *[99]* designed for computing the approximate factorization outputs the precise factorization structure and accurate factors within a threshold $\theta = 10^{-10}$, along with error estimates and a sensitivity measurement. This result is a substantial improvement over the previous algorithm MultRoot [97, 98] which is limited to extracting factors of multiplicity under 30. In contrast, standard methods like Matlab function `roots` output scattered root clusters, as shown in Figure 5.2.2.

```
>> [F,res,cond] = uvfactor(f,1e-10);

THE CONDITION NUMBER:                          2.57705
THE BACKWARD ERROR:                       7.62e-016
THE ESTIMATED FORWARD ROOT ERROR:         3.93e-015

  FACTORS

   ( x -   4.000000000000008 )^20
   ( x -   2.999999999999994 )^40
   ( x -   2.000000000000002 )^60
   ( x -   1.000000000000000 )^80
```

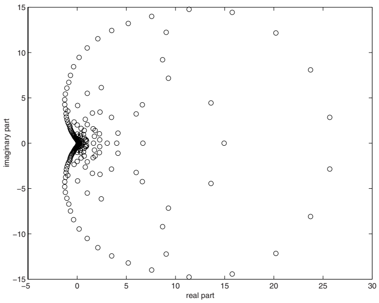


Fig. 5.2 Matlab results for the polynomial (5.19)

□

Pejorative manifolds in numerical polynomial algebra can often be parametrized in the form of $\mathbf{u} = F(\mathbf{z})$ with injective Jacobians, as shown in previous examples. In such cases, it is our conjecture that the approximate solution formulated based on backward nearness, maximum codimension and minimum distance is well-posed in the following sense: If the given problem P is a perturbation from an (exact) problem $\hat{P}$ with sufficiently small error ε, then there is an upper bound μ on the threshold θ. As long as θ is chosen in the interval (ε, μ), there is a unique approximate solution S of P within all such θ, and the solution S is continuous with respect to the problem P. Moreover, the approximate solution S converges to the exact solution $\hat{S}$ of $\hat{P}$ if P approaches to $\hat{P}$ with

$$\left\|S - \hat{S}\right\| \;=\; O\left(\left\|P - \hat{P}\right\|\right).$$

5.3 Matrix computation arising in polynomial algebra

As Stetter points out in [84] and in the title of [83], matrix eigenproblem is at the heart of polynomial system solving. Furthermore, matrix computation problems such as least squares and numerical rank/kernel computation also arise frequently and naturally in polynomial algebra. We shall survey the approximate GCD, factorization, multiplicity structure and elimination problems in this section and derive related matrix computation problems.

5.3.1 Approximate GCD

Given a polynomial pair $(p,q) \in \mathbb{C}[x] \times \mathbb{C}[x]$ of degrees m and n, respectively, with

$$\begin{array}{rcl} p(x) & = & p_0 + p_1 x + \cdots + p_m x^m \\ q(x) & = & q_0 + q_1 x + \cdots + q_n x^n, \end{array}$$

we can write

$$p \;=\; uv \quad \text{and} \quad q \;=\; uw$$

where $u = \mathit{gcd}(p,q)$ is the GCD along with cofactors v and w. Then

$$p \cdot w - q \cdot v \;=\; [p,\, q] \begin{bmatrix} w \\ -v \end{bmatrix} \;=\; 0. \tag{5.20}$$

That is, the polynomial pair (v,w) belongs to the kernel $\mathscr{K}(\mathscr{L}_{p,q})$ of the linear transformation

$$\mathscr{L}_{p,q} \;:\; (r,s) \;\longrightarrow\; p \cdot s - q \cdot r. \tag{5.21}$$

Moreover, it is clear that the kernel $\mathscr{K}(\mathscr{L}_{p,q})$ is spanned by polynomial pairs in the set $\big\{ (x^j v,\, x^j w) \;\big|\; j = 0, 1, \cdots \big\}$.

The classical Sylvester matrix

$$S(p,q) \;=\; \begin{array}{c} \begin{array}{cc} \overbrace{\hspace{5em}}^{n} & \overbrace{\hspace{5em}}^{m} \end{array} \\ \begin{pmatrix} p_0 & & & q_0 & & \\ p_1 & \ddots & & q_1 & \ddots & \\ \vdots & \ddots & p_0 & \vdots & \ddots & q_0 \\ p_m & & p_1 & q_n & & q_1 \\ & \ddots & \vdots & & \ddots & \vdots \\ & & p_m & & & q_n \end{pmatrix} \end{array} \tag{5.22}$$

is the matrix representation of the linear transformation $\mathscr{L}_{p,q}$ in (5.21) restricted in the domain $\big\{ (r,s) \;\big|\; \mathit{deg}(r) < n,\; \mathit{deg}(s) < m \big\}$. It becomes clear that $S(p,q)$ has a nullity equals to $k = \mathit{deg}(u)$ since the kernel of the restricted $\mathscr{L}_{p,q}$ is spanned by

$\big\{(v,w),\ (xv,xw),\ \ldots,\ (x^{k\text{-}1}v,x^{k\text{-}1}w)\big\}$. Consequently, the (exact) GCD structure is represented as the nullity of the Sylvester matrix:

$$\mathit{deg}\big(\mathit{gcd}(p,q)\big) \;=\; \mathit{nullity}\,(\,S(p,q)\,)\,.$$

For the problem of the *approximate* GCD, the polynomial pair (p,q) is considered a perturbation from $(\hat{p},\hat{q})$ residing in the GCD manifold $\mathscr{P}_k^{m,n}$ (cf. Example 5.2.3 in §5.2.1). Then

$$S(p,q) \;=\; S(\hat{p},\hat{q})+S(p-\hat{p},q-\hat{q}).$$

Namely, the Sylvester matrix $S(p,q)$ is near $S(\hat{p},\hat{q})$ of nullity k with a distance $\|S(p-\hat{p},q-\hat{q})\|_2$, and identifying the maximum codimension manifold $\mathscr{P}_k^{m,n}$ becomes the *numerical* rank/kernel problem of the Sylvester matrix.

After identifying the degree k of the approximate GCD, one can further restrict the domain of the linear transformation $\mathscr{L}_{p,q}$ as

$$\big\{(r,s)\;\big|\;\mathit{deg}(r)\le m{-}k,\ \mathit{deg}(s)\le n{-}k\big\}.$$

The corresponding Sylvester submatrix

$$S_k(p,q) \;=\; \begin{array}{c} \overbrace{\hphantom{p_m\;\;\ddots\;\;p_1}}^{n-k+1}\;\;\overbrace{\hphantom{q_n\;\;\ddots\;\;q_1}}^{m-k+1} \\ \begin{pmatrix} p_0 & & & q_0 & & \\ p_1 & \ddots & & q_1 & \ddots & \\ \vdots & \ddots & p_0 & \vdots & \ddots & q_0 \\ p_m & & p_1 & q_n & & q_1 \\ & \ddots & \vdots & & \ddots & \vdots \\ & & p_m & & & q_n \end{pmatrix} \end{array}$$

has a numerical kernel of dimension one and the coefficients of the cofactors v and w can be obtained approximately from the lone basis vector of the numerical kernel:

$$\mathscr{K}\big(S_k(p,q)\big) \;=\; \mathit{span}\left\{\begin{bmatrix} [\![v]\!] \\ -[\![w]\!] \end{bmatrix}\right\}.$$

An *estimate* of the coefficients of the approximate GCD u can then be obtained by solving for the least squares solution u in the linear system $u\cdot v=f$ instead of the unstable synthetic division.

The approximate GCD computed via numerical kernel and linear least squares alone may not satisfy the minimum distance requirement. The accuracy of the computed GCD depends on the numerical condition of the Sylvester submatrix $S_k(p,q)$, while the GCD sensitivity measured from (5.13) can be much healthier. Therefore, the above matrix computation serves as stage I of the approximate GCD finding. The Gauss-Newton iteration (c.f. §5.1.3) is needed to ensure the highest attainable accuracy.

For a simple example [96], consider the polynomial pair

$$f(x) = x^7 - .999999999999x^6 + x^5 - .999999999999x^4 + x^3 + .999999999999x^2 + x - .999999999999$$
$$g(x) = x^6 - 3x^3 - x^5 + 2x^2 + x^4 - 2x + 2$$

possessing an exact GCD as $gcd(f,g) = x^2+1$. Matrix computation alone can only produce $u \approx x^2 + 0.99992$ with four-digit accuracy, while the Gauss-Newton iteration attains the solution with a high accuracy near hardware precision. The GCD condition number 3.55 shows the approximate GCD is not sensitive at all, but the kernel sensitivity for the Sylvester matrix is quite high ($\approx 10^{12}$).

The upper-bound of the distance to rank-deficiency for the Sylvester matrix

$$\begin{aligned} \|S(p-\hat{p},q-\hat{q})\|_2 &\leq \|S(p-\hat{p},q-\hat{q})\|_F \\ &= \sqrt{n\|p-\hat{p}\|_2^2 + m\|q-\hat{q}\|_2^2} \;=\; \theta, \end{aligned}$$

where $\|\cdot\|_F$ denotes the Frobenius norm [34, §2.3] of a matrix. However, rank-deficiency within the above threshold θ is only the necessary condition for the given (p,q) to be near a GCD manifold. The converse is not true. This is another reason why the iterative refinement is needed as Stage II for computing the minimum distance and certifying the approximate GCD.

There is also a different approach for GCD computation [38, 47, 50, 104]: Computing displacement polynomials Δp and Δq such that the Sylvester matrix (or submatrix) $S(p+\Delta p, q+\Delta q)$ has a specified nullity. This approach leads to a total least squares problem with special structures. This approach may have an advantage when there is a large distance from (p,q) to the GCD manifold beyond the convergence domain of the Gauss-Newton iteration, and it theoretically seeks a global minimum in comparison with the local minimum of the Gauss-Newton iteration. An apparent disadvantage of this approach is that it is subject to the sensitivity of the matrix kernel, not the actual GCD condition.

Both approaches can extend to multivariate GCD in a straightforward generalization. However, the matrix sizes may become huge when the number of indeterminates increases, and it may become necessary to reduce those sizes for the consideration of both storage and computing time. A subspace strategy for this purpose shall be discussed later in §5.4.2.

5.3.2 The multiplicity structure

For a polynomial ideal (or system) $I = \langle f_1, f_2 \ldots, f_t \rangle \subset \mathbb{C}[x_1,\ldots,x_s]$ with an isolated zero $\hat{\mathbf{x}} = (\hat{x}_1,\ldots,\hat{x}_s)$, the study on the multiplicity of I at $\hat{\mathbf{x}}$ traces back to Newton's time with evolving formulations [26, pp. 127-129][60, 64, 84]. Several computing methods for identifying the multiplicity have been proposed in the literature, such as [54, 62, 65, 88] and more recently in [6, 17, 94]. Here we elaborate the dual basis approach that can be directly adapted to numerical kernel computation. This approach is originated by Macaulay [60] in 1916 and reformulated by

Gröbner in 1939. For that reason it is also (somewhat inappropriately) referred to as the Gröbner duality [64, §31.3].

A univariate polynomial $f(x)$ has an m-fold zero $\hat{x}$ if

$$f(\hat{x}) \;=\; f'(\hat{x}) \;=\; \cdots \;=\; f^{(m\text{-}1)}(\hat{x}) = 0 \quad \text{and} \quad f^{(m)}(\hat{x}) \neq 0.$$

The Macaulay/Gröbner duality can be considered a generalization of this multiplicity to multivariate cases using partial differentiation functionals.

For an index array $\mathbf{j} = [j_1, \cdots, j_s] \in \mathbb{N}^s$ of non-negative integers, denote

$$\mathbf{x}^{\mathbf{j}} = x_1^{j_1} \cdots x_s^{j_s} \quad \text{and} \quad (\mathbf{x} - \mathbf{y})^{\mathbf{j}} = (x_1 - y_1)^{j_1} \cdots (x_s - y_s)^{j_s}.$$

Define a (differential) monomial functional at $\hat{\mathbf{x}} \in \mathbb{C}^s$ as

$$\partial_{\mathbf{j}}[\hat{\mathbf{x}}] : \mathbb{C}[\mathbf{x}] \longrightarrow \mathbb{C}, \quad \text{where } \partial_{\mathbf{j}}[\hat{\mathbf{x}}](p) = (\partial_{\mathbf{j}} p)(\hat{\mathbf{x}}) \quad \text{for any } p \in \mathbb{C}[\mathbf{x}].$$

Here, the differentiation operator

$$\partial_{\mathbf{j}} \;\equiv\; \partial_{j_1 \cdots j_s} \;\equiv\; \partial_{x_1^{j_1} \cdots x_s^{j_s}} \;\equiv\; \frac{1}{j_1! \cdots j_s!} \, \frac{\partial^{j_1 + \cdots + j_s}}{\partial x_1^{j_1} \cdots \partial x_s^{j_s}}. \tag{5.23}$$

Namely, the monomial functional $\partial_{\mathbf{j}}[\hat{\mathbf{x}}]$ applying to a polynomial p equals the partial derivative $\partial_{\mathbf{j}}$ of p evaluated at $\hat{\mathbf{x}}$. We may use $\partial_{\mathbf{j}}$ for $\partial_{\mathbf{j}}[\hat{\mathbf{x}}]$ when the zero $\hat{\mathbf{x}}$ is clear from context. A (differential) functional at $\hat{\mathbf{x}} \in \mathbb{C}^s$ is a linear combination of those $\partial_{\mathbf{j}}[\hat{\mathbf{x}}]$'s. The collection of all functionals at the zero $\hat{\mathbf{x}}$ that vanish on the entire ideal I forms a vector space $\mathscr{D}_{\hat{\mathbf{x}}}(I)$ called the *dual space* of I at $\hat{\mathbf{x}}$

$$\mathscr{D}_{\hat{\mathbf{x}}}(I) \;\equiv\; \Big\{ c = \sum_{\mathbf{j} \in \mathbb{N}^s} c_{\mathbf{j}} \partial_{\mathbf{j}}[\hat{\mathbf{x}}] \;\Big|\; c(f) = 0, \text{ for all } f \in I \Big\} \tag{5.24}$$

where $c_{\mathbf{j}} \in \mathbb{C}$ for all $\mathbf{j} \in \mathbb{N}^s$. The dimension of the dual space $\mathscr{D}_{\hat{\mathbf{x}}}(I)$ is the *multiplicity* of the zero $\hat{\mathbf{x}}$ to the ideal I. The dual space itself forms the multiplicity structure of I at $\hat{\mathbf{x}}$.

For example, a univariate polynomial p having an m-fold zero $\hat{x}$ if and only if the dual space of the ideal $\langle p \rangle$ at $\hat{x}$ is spanned by functionals ∂_0, ∂_1, …, $\partial_{m\text{-}1}$. The ideal $I = \langle x_1^3,\, x_1^2 x_2 + x_2^4 \rangle$ has a zero $(0,0)$ of multiplicity 12 with the dual space spanned by [17]

$$\overbrace{\partial_{00}}^{1},\; \overbrace{\partial_{10},\; \partial_{01}}^{2},\; \overbrace{\partial_{20},\; \partial_{11},\; \partial_{02}}^{3},\; \overbrace{\partial_{12},\; \partial_{03}}^{2},\; \overbrace{\partial_{13},\; \partial_{04} - \partial_{21}}^{2},\; \overbrace{\partial_{05} - \partial_{22}}^{1},\; \overbrace{\partial_{06} - \partial_{23}}^{1} \tag{5.25}$$

and Hilbert function $\{1,2,3,2,2,1,1,0,\cdots\}$ as a partition of the multiplicity 12.

The most crucial requirement for the dual space is the *closedness condition*: For c to be a functional in the dual space $\mathscr{D}_{\hat{\mathbf{x}}}(I)$, not only $c(f_1) = \cdots = c(f_t) = 0$, but also $c(p f_i) = 0$ for all polynomial $p \in \mathbb{C}[\mathbf{x}]$ and for $i = 1, \ldots, t$. Since all differential functionals are linear, the closedness condition can be simplified to the system of linear equations

$$\sum_{\mathbf{k}\in\mathbb{N}^s,\ |\mathbf{k}|\le\alpha} c_{\mathbf{k}}\partial_{\mathbf{k}}[\hat{\mathbf{x}}]\big((\mathbf{x}-\hat{\mathbf{x}})^{\mathbf{j}} f_i\big) \;=\; 0 \qquad (5.26)$$
$$\text{for } \mathbf{j}\in\mathbb{N}^s,\ \mathbf{j}<\mathbf{k},\ \text{and } i\in\{1,\ldots,s\}.$$

in the coefficients $c_{\mathbf{j}}$'s for sufficiently large $\alpha\in\mathbb{N}$.

For each $\alpha\in\mathbb{N}$, the equations in (5.26) can be expressed as a homogeneous linear system in matrix form

$$S_\alpha \mathbb{C} \;=\; 0$$

where S_α is the Macaulay matrix of order α, and each functional in the basis for the dual space corresponds to a null vector of S_α and *vice versa.* The Hilbert function

$$\begin{cases} H(0)=1, \\ H(\alpha)=\mathit{nullity}(S_\alpha)-\mathit{nullity}(S_{\alpha\text{-}1}) \quad \text{for } \alpha=1,2,\ldots. \end{cases}$$

The (approximate) dual basis can be obtained by computing the (numerical) kernels of S_0, S_1, ... until reaching the α where the Hilbert function $H(\alpha)=0$ [17].

The algorithm based on the above analysis and matrix kernel computation is implemented as a computational routine MULTIPLICITYSTRUCTURE in the software package APATOOLS [99]. The algorithm is also applied to identifying the local dimension of algebraic sets [3] as part of the software package BERTINI [4].

5.3.3 Numerical elimination

Numerical elimination with approximate data arises in many applications such as kinematics [16, 21, 61, 87, 91, 90], and computational biology/chemistry [23, 25]. Numerical elimination methods have been studied in many reports, such as [1, 2, 21, 22, 42, 61, 70, 87, 90, 91]. The main strategy for the existing elimination approaches is using various resultants [32] whose computation requires calculating determinants of polynomial matrices. There are formidable obstacles for calculating resultants using floating point arithmetic since determinant computation can be inefficient and unstable. There is a new approach [95] that avoids resultant calculation and transforms the elimination to a problem of matrix rank/kernel computation.

Consider the ring $\mathbb{C}[x,y]$ of complex polynomials in variables x and y, where x is a single scalar variable and y may be either a scalar variable or a vector of variables. For polynomials $f,g\in\mathbb{C}[x,y]$, there exist polynomials p and q such that

$$f\,p+g\,q \;=\; h$$

belongs to the first elimination ideal $\langle f,g\rangle\cap\mathbb{C}[y]$ of f and g, unless (f,g) has a GCD with positive degree in x. We wish to calculate p, q and h with floating point arithmetic. Since $f\,p+g\,q=h\in\mathbb{C}[y]$, there is an obvious homogeneous equation

$$\frac{\partial}{\partial x}(f\,p+g\,q) \;=\; \left[\frac{\partial}{\partial x}f+f\cdot\frac{\partial}{\partial x}\right]p+\left[\frac{\partial}{\partial x}g+g\cdot\frac{\partial}{\partial x}\right]q \;=\; 0. \tag{5.27}$$

which leads to a simple strategy: Finding a polynomial $f\,p+g\,q$ in the elimination ideal $\langle f,g\rangle\cap\mathbb{C}[y]$ is equivalent to computing the kernel $\mathscr{K}(L_n)$ of the linear transformation

$$L_n\;:\;(p,q)\;\longrightarrow\;\left[\frac{\partial}{\partial x}f+f\cdot\frac{\partial}{\partial x},\quad \frac{\partial}{\partial x}g+g\cdot\frac{\partial}{\partial x}\right]\begin{bmatrix}p\\q\end{bmatrix} \tag{5.28}$$

in the vector space $\mathbb{P}^n$ of polynomial pairs p and q with degree n or less. With proper choices of bases for the polynomial vector spaces, the linear transformation L_n induces an *elimination matrix* M_n where the rank-revealing method elaborated in §5.1.2 produces the polynomial $h=fp+gq$ in the elimination ideal. The elimination algorithm based on matrix kernel computation can be outlined below with a test version implemented in ApaTools [99] as computational routine PolynomialEliminate.

```
For  $n=1,2,\cdots$  do
  Update elimination matrix  $M_n$
  If  $M_n$  is rank deficient then
     extract  $(p,q)$  from its kernel,  break,
  end if
end do
```

The method can be generalized to eliminating more than one variables from several polynomials and, combined with numerical multivariate GCD, can be applied to solving polynomial systems for solution sets of positive dimensions [95].

5.3.4 Approximate irreducible factorization

Computing the approximate irreducible factorization of a multivariate polynomial is the first problem listed in "challenges in symbolic computation" [44] by Kaltofen. The first numerical algorithm with an implementation is developed by Sommese, Verschelde and Wampler [78, 79, 80]. Many authors studied the problem and proposed different approaches [8, 7, 10, 12, 13, 27, 28, 40, 45, 73, 74, 75, 76]. In 2003, Gao [29] proposed a hybrid algorithm and later adapted it as a numerical algorithm [30, 46, 48] along with Kaltofen, May, Yang and Zhi. Here we elaborate the strategies involved in numerical irreducible factorization from the perspective of matrix computation.

The collection of polynomials sharing the structure of irreducible factorization $p_1^{m_1}p_2^{m_2}\cdots p_k^{m_k}$ forms a factorization manifold which we denote as $\Pi_{n_1\cdots n_k}^{m_1\cdots m_k}$, where n_j is the degree of p_j for $j=1,\cdots,k$. Namely

$$\Pi_{n_1\cdots n_k}^{m_1\cdots m_k} \;=\; \left\{p_1^{m_1}p_2^{m_2}\cdots p_k^{m_k}\;\middle|\;deg(p_j)=n_j\;\;\text{for}j=1,\ldots,k\right\}$$

For simplicity of exposition, consider the manifold $\Pi^{1,1,1}_{n_1 n_2 n_3}$ where polynomials are square-free in two variables x and y with three factors. An approximate square-free factorization algorithm is implemented in ApaTools [99] as a computational routing SquareFreeFactor which is a recursive application of approximate GCD computation [102].

The first task of factorization is to identify the reducibility of the polynomial, the number of factors, and the degrees of each factor. The Ruppert approach[71] for reducibility detection is very effective and can be easily explained using the differential identity

$$\frac{\partial}{\partial y}\left(\frac{f_x}{f}\right) = \frac{\partial}{\partial x}\left(\frac{f_y}{f}\right)$$

for any function f. For a given polynomial $f = p \cdot q \cdot r$ of bidegree $[m,n]$, applying this identity to the first factor p yields $\frac{\partial}{\partial y}\left(\frac{p_x}{p} \cdot \frac{q \cdot r}{q \cdot r}\right) = \frac{\partial}{\partial x}\left(\frac{p_y}{p} \cdot \frac{q \cdot r}{q \cdot r}\right)$, namely

$$\frac{\partial}{\partial y}\left(\frac{p_x \cdot q \cdot r}{f}\right) = \frac{\partial}{\partial x}\left(\frac{p_y \cdot q \cdot r}{f}\right).$$

The same manipulation on the remaining factors q and r reveals that the differential equation

$$\frac{\partial}{\partial y}\left(\frac{g}{f}\right) = \frac{\partial}{\partial x}\left(\frac{h}{f}\right)$$

in the unknown polynomial pair (g,h) has three linearly independent solutions

$$(g,h) = (p_x qr,\ p_y qr),\ (pq_x r,\ pq_y r), \quad \text{or} \quad (pqr_x,\ pqr_y).$$

From this observation, it can be seen that the linear transformation

$$\begin{aligned} \mathscr{L}_f : (g,h) \longrightarrow f^2 &\left[\frac{\partial}{\partial y}\left(\frac{g}{f}\right) - \frac{\partial}{\partial x}\left(\frac{h}{f}\right)\right] \\ &= [f \cdot \partial_y - \partial_y f]\, g - [f \cdot \partial_x - \partial_x f]\, h \end{aligned} \tag{5.29}$$

has a kernel whose dimension is identical to the number of irreducible factors of f if we restrict the domain of $\mathscr{L}_f$ to those g and h of bidegrees $[m-1,n]$ and $[m,n-1]$ respectively. As a result, the reducibility and the number of irreducible factors in f can be identified by computing the nullity of the Ruppert matrix L_f corresponding to the linear transformation $\mathscr{L}_f$ with a restricted domain. When the polynomial f is inexact using floating-point arithmetic, the reducibility identification becomes numerical kernel problem.

Moreover, a vector in the numerical kernel $\mathscr{K}_\theta(L_f)$ selected at random corresponds to polynomials

$$\begin{aligned} g &= \lambda_1 p_x qr + \lambda_2 pq_x r + \lambda_3 pqr_x \\ h &= \lambda_1 p_y qr + \lambda_2 pq_y r + \lambda_3 pqr_y \end{aligned}$$

with undetermined constants λ_1, λ_2 and λ_3. From $f = pqr$ and

$$g - \lambda_1 f_x \;=\; p[(\lambda_2 - \lambda_1) q_x r + (\lambda_3 - \lambda_1) q r_x],$$

we have identities associated with the unknown λ_j's

$$\begin{cases} p &= \mathit{gcd}(f,\; g - \lambda_1 f_x) \\ q &= \mathit{gcd}(f,\; g - \lambda_2 f_x) \\ r &= \mathit{gcd}(f,\; g - \lambda_3 f_x) \end{cases} \tag{5.30}$$

Select complex numbers $\hat{x}$ and $\hat{y}$ at random and consider univariate polynomials $p(x,\hat{y})$ and $p(\hat{x},y)$ in (5.30). Applying the nullity count of the Sylvester matrix (5.22) elaborated in §5.3.1 yields that the Sylvester matrices

$$S\big(f(x,\hat{y}),\; g(x,\hat{y}) - \lambda_1 f_x(x,\hat{y})\big) \quad \text{and} \quad S\big(f(\hat{x},y),\; g(\hat{x},y) - \lambda_1 f_x(\hat{x},y)\big)$$

are of nullities identical to the degrees $\mathit{deg}_x(p)$ and $\mathit{deg}_y(p)$ respectively. The unknown value of λ_1 then becomes the generalized eigenvalue of the matrix pencils in the form of $A - \lambda B$:

$$S\big(f(x,\hat{y}),\; g(x,\hat{y})\big) - \lambda S\big(0,\; f_x(x,\hat{y})\big) \tag{5.31}$$

and

$$S\big(f(\hat{x},y),\; g(\hat{x},y)\big) - \lambda S\big(0,\; f_x(\hat{x},y)\big). \tag{5.32}$$

From the identities in (5.30), both pencils have the same eigenvalues λ_1, λ_2 and λ_3 of geometric multiplicities identical to the degrees $\mathit{deg}_x(p)$, $\mathit{deg}_x(q)$ and $\mathit{deg}_x(r)$ respectively for the pencil (5.31) and to the degrees $\mathit{deg}_y(p)$, $\mathit{deg}_y(q)$ and $\mathit{deg}_y(r)$ respectively for the pencil (5.32). As a result, finding the unknown constants λ_j's in (5.30) and degree structures of the irreducible factors p, q and r becomes generalized eigenvalue problem of matrix pencils in (5.31)–(5.32).

Computing eigenvalues with nontrivial multiplicities has been a challenge in numerical linear algebra. With new techniques developed in [103] on computing the Jordan Canonical Form along with the known eigen-structure of the pencils in (5.31)-(5.32), their generalized eigenvalues and multiplicities can be computed efficiently and accurately in floating-point arithmetic even if the polynomials are inexact.

In summary, Stage I of the numerical irreducible factorization can be accomplished with a sequence of matrix computations:

(a) Finding the numerical kernel of the Ruppert matrix by matrix rank/kernel computation; followed by
(b) solving the generalized eigenproblem of pencils in (5.31)-(5.32) to obtain the degrees of the irreducible factors and values of λ_j's; and concluded with
(c) computing approximate GCDs in (5.30) to obtain approximations to the irreducible factors.

This stage of the computation identifies the maximum codimension factorization manifold.

In Stage II, the approximations of the factors obtained in Stage I are used as the initial iterate for the Gauss-Newton iteration (5.11) applied on the overdetermined system $F(\mathbf{u},\mathbf{v},\mathbf{w})=\mathbf{0}$ where

$$F(\mathbf{u},\mathbf{v},\mathbf{w}) \;\equiv\; \begin{bmatrix} \mathbf{a}^{\mathsf{H}}\mathbf{u}-1 \\ \mathbf{b}^{\mathsf{H}}\mathbf{v}-1 \\ [\![uvw-f]\!] \end{bmatrix}.$$

The Jacobian of F is injective (cf. Example 5.1.3 in §5.1.3, not without the auxiliary equations $\mathbf{a}^{\mathsf{H}}\mathbf{u}=\mathbf{b}^{\mathsf{H}}\mathbf{v}=1$) at the solution $(\mathbf{p},\mathbf{q},\mathbf{r})$, ensuring the Gauss-Newton iteration to be locally convergent.

5.4 A subspace strategy for efficient matrix computations

5.4.1 The closedness subspace for multiplicity matrices

The Macaulay matrix S_α for the homogeneous system (5.26) can be undesirably large when α increases. For example, consider the benchmark problem KSS system [17, 54] of $n\times n$

$$f_j(x_1,\cdots,x_n)=x_j^2+\sum_{\nu=1}^{n}x_\nu-2x_j-n+1, \quad \text{for } j=1,\ldots,n \tag{5.33}$$

for multiplicity computation at the zero $\hat{\mathbf{x}}=(1,\ldots,1)$. The largest Macaulay matrix required is 12012×3432 for $n=7$ to obtain the multiplicity 64, and the matrix size reaches 102960×24310 for $n=8$ for identifying the multiplicity 128, exceeding the memory capacity of today's desktop personal computers. Therefore, reducing the size of the multiplicity matrices is of paramount importance.

The key idea, which is originated by Stetter and Thallinger [84, 88], is to employ the closedness condition that can be rephrased in the following lemma.

Lemma 5.4.1. [84, Theorem 8.36] *Let $I\in\mathbb{C}[x_1,\ldots,x_s]$ be a polynomial ideal and let $\mathscr{D}_{\hat{\mathbf{x}}}(I)$ be its dual space at an isolated zero $\hat{\mathbf{x}}$. Then every functional $c\in\mathscr{D}_{\hat{\mathbf{x}}}(I)$ satisfies $s_\sigma(c)\in\mathscr{D}_{\hat{\mathbf{x}}}(I)$ for all $\sigma\in\{1,\ldots,s\}$ where s_σ is the linear anti-differentiation operator defined by*

$$s_\sigma(\partial_{j_1\cdots j_s}[\hat{\mathbf{x}}]) \;=\; \begin{cases} 0, & \text{if } j_\sigma=0 \\ \partial_{j_1'\cdots j_s'}[\hat{\mathbf{x}}], & \text{otherwise} \end{cases}$$

with $j_\sigma'=j_\sigma-1$ and $j_i'=j_i$ for $i\in\{1,\cdots,s\}\backslash\{\sigma\}$.

For example, the functional $\partial_{06}-\partial_{23}$ belongs to the dual space $\mathscr{D}_{\mathbf{0}}(I)$ spanned by the functionals in (5.25) implies

$$\begin{aligned} s_1(\partial_{06} - \partial_{23}) &= 0 - \partial_{13} \\ s_2(\partial_{06} - \partial_{23}) &= \partial_{05} - \partial_{22} \end{aligned}$$

are both in $\mathscr{D}_{\mathbf{0}}(I)$, as shown in (5.25). This property leads to the closedness subspace strategy: After obtaining the dual subspace

$$\mathscr{D}_{\hat{\mathbf{x}}}^{\alpha\text{-}1}(I) \;\equiv\; \mathscr{D}_{\hat{\mathbf{x}}}(I) \cap \mathit{span}\big\{\partial_{\mathbf{j}}[\hat{\mathbf{x}}] \;\big|\; |\mathbf{j}| \leq \alpha - 1\big\}$$

of order $\alpha - 1$, the additional basis functionals in the dual subspace $\mathscr{D}_{\hat{\mathbf{x}}}^{\alpha}(I)$ of order α can be found in the *closedness subspace*

$$\mathscr{C}_{\hat{\mathbf{x}}}^{\alpha}(I) \;=\; \left\{ \sum_{\mathbf{j}\in\mathbb{N}^s, |\mathbf{j}|\leq\alpha} c_{\mathbf{j}}\partial_{\mathbf{j}}[\hat{\mathbf{x}}] \;\middle|\; s_\sigma(c) \in \mathscr{D}_{\hat{\mathbf{x}}}^{\alpha\text{-}1}(I),\;\; \sigma = 1,\ldots,s \right\}, \tag{5.34}$$

of order α. An algorithm for computing the bases for closedness subspaces have been developed in [100] using a sequence of numerical kernel computation involving matrices of much smaller sizes. Let $\{\phi_1,\ldots,\phi_m\}$ form a basis for the closedness subspace $\mathscr{C}_{\hat{\mathbf{x}}}^{\alpha}(I)$. Then the functionals in the dual subspace $\mathscr{D}_{\hat{\mathbf{x}}}^{\alpha}(I)$ can be expressed in the form of $u_1\phi_1 + \cdots + u_m\phi_m$ and can be computed by solving the homogeneous linear system

$$u_1\phi_1(f_i) + \cdots + u_m\phi_m(f_i) \;=\; 0, \quad \text{for} \quad i = 1, \cdots t \tag{5.35}$$

where $f_1,\ldots,f_t$ are generators for the ideal I. In comparison with the linear system (5.26), the system (5.35) consists of a fixed number (t) of equations. Since the solution space of (5.35) is isomorphic to the dual subspace $\mathscr{D}_{\hat{\mathbf{x}}}^{\alpha}(I)$, the number m of unknowns u_i's is bounded by

$$m \;\leq\; \mathit{dim}\big(\mathscr{D}_{\hat{\mathbf{x}}}^{\alpha}(I)\big) + t.$$

The process of identifying the closedness subspace $\mathscr{C}_{\hat{\mathbf{x}}}^{\alpha}(I)$ and finding dual basis functionals can be illustrated in the following example.

Example 5.4.2. Use the dual basis in (5.25) as an example. To identify the closedness subspace $\mathscr{C}_{\mathbf{0}}^{4}(I)$ after obtaining the dull subspace and the closedness subspace of order 3

$$\begin{aligned} \mathscr{D}_{\mathbf{0}}^{3}(I) &= \mathit{span}\{\partial_{00},\, \partial_{10},\, \partial_{01},\, \partial_{20},\, \partial_{11},\, \partial_{02},\, \partial_{12},\, \partial_{03}\} \\ \mathscr{C}_{\mathbf{0}}^{3}(I) &= \mathit{span}\{\partial_{00},\, \partial_{10},\, \partial_{01},\, \partial_{20},\, \partial_{11},\, \partial_{02},\, \partial_{30},\, \partial_{21},\, \partial_{12},\, \partial_{03}\}. \end{aligned}$$

The monomial functionals ∂_{22}, ∂_{31}, and ∂_{40} can be excluded since $s_2(\partial_{22}) = s_1(\partial_{31}) = \partial_{21}$ and $s_1(\partial_{40}) = \partial_{30}$ are not in the monomial support of $\mathscr{D}_{\mathbf{0}}^{3}(I)$. As a result, we have

$$\mathscr{C}_{\mathbf{0}}^{4}(I) \;\subset\; \mathit{span}\{\partial_{00},\, \partial_{10},\, \partial_{01},\, \partial_{20},\, \partial_{11},\, \partial_{02},\, \partial_{30},\, \partial_{21},\, \partial_{12},\, \partial_{03},\, \partial_{04},\, \partial_{13}\}$$

and every functional in $\mathscr{C}_{\mathbf{0}}^{4}(I)$ can be written as

$$\phi = \gamma_1\partial_{00}+\gamma_2\partial_{10}+\gamma_3\partial_{01}+\cdots+\gamma_{10}\partial_{04}+\gamma_{11}\partial_{13}.$$

The closedness conditions $s_1(\phi), s_2(\phi) \in \mathscr{D}_{\mathbf{0}}^3(I)$ become

$$\begin{cases} \gamma_1\partial_{00}+\gamma_3\partial_{10}+\gamma_4\partial_{01}+\gamma_6\partial_{20}+\gamma_7\partial_{11}+\gamma_8\partial_{02}+\gamma_{11}\partial_{03} \\ \qquad\qquad = \eta_1\partial_{00}+\eta_2\partial_{10}+\cdots+\eta_8\partial_{03} \\ \gamma_2\partial_{00}+\gamma_4\partial_{10}+\gamma_5\partial_{01}+\gamma_7\partial_{20}+\gamma_8\partial_{11}+\gamma_9\partial_{02}+\gamma_{10}\partial_{03}+\gamma_{11}\partial_{12} \\ \qquad\qquad = \eta_9\partial_{00}+\eta_{10}\partial_{10}+\cdots+\eta_{16}\partial_{03} \end{cases}.$$

These equations lead to

$$[\gamma_1,\ \gamma_2,\ \ldots,\gamma_{11}]^\top = [\eta_1,\ \eta_9,\ \eta_2,\ \eta_3,\ \eta_{11},\ \eta_4,\ \eta_5,\ \eta_6,\ \eta_{14},\ \eta_{16},\ \eta_8]^\top \tag{5.36}$$

along with

$$\gamma_4=\eta_3=\eta_{10},\ \ \gamma_7=\eta_5=\eta_{12},\ \ \gamma_8=\eta_6=\eta_{13},\ \ \gamma_{11}=\eta_8=\eta_{15},\ \ \eta_7=0.$$

Consequently, we have a system of homogeneous equations

$$\eta_3-\eta_{10} = \eta_5-\eta_{12} = \eta_6-\eta_{13} = \eta_8-\eta_{15} = \eta_7 = 0. \tag{5.37}$$

Since a basis for $\mathscr{C}_{\mathbf{0}}^3(I)$ is already obtained, we need only additional basis functionals in $\mathscr{C}_{\mathbf{0}}^4(I)$ by requiring

$$\phi \perp \mathscr{C}_{\mathbf{0}}^3(I). \tag{5.38}$$

In general, systems of equations in (5.36), (5.37) and (5.38) can be written in matrix forms

$$\begin{bmatrix}\gamma_1\\ \vdots\\ \gamma_m\end{bmatrix} = A\begin{bmatrix}\eta_1\\ \vdots\\ \eta_n\end{bmatrix},\quad B\begin{bmatrix}\eta_1\\ \vdots\\ \eta_n\end{bmatrix} = \mathbf{0},\quad \text{and}\quad C\begin{bmatrix}\eta_1\\ \vdots\\ \eta_n\end{bmatrix} = \mathbf{0}$$

respectively. Thus the problem of identifying the closedness subspace becomes the (numerical) kernel problem of the matrix $\left[\begin{smallmatrix}B\\ C\end{smallmatrix}\right]$, from which we obtain

$$\begin{bmatrix}\eta_1\\ \vdots\\ \eta_n\end{bmatrix} = N\begin{bmatrix}\mu_1\\ \vdots\\ \mu_k\end{bmatrix},\quad \text{and thus}\quad \begin{bmatrix}\gamma_1\\ \vdots\\ \gamma_m\end{bmatrix} = AN\begin{bmatrix}\mu_1\\ \vdots\\ \mu_k\end{bmatrix}.$$

In this example, $\gamma_1=\cdots=\gamma_9=0$ and

$$\begin{bmatrix}\gamma_{10}\\ \gamma_{11}\end{bmatrix} = \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix}$$

implying

$$\mathscr{C}_{\mathbf{0}}^4(I) = \mathscr{C}_{\mathbf{0}}^3(I)\oplus span\{\partial_{13},\ \partial_{04}\}$$

where additional dual basis functionals ∂_{13} and $\partial_{04}-\partial_{21}$ in $\mathscr{D}^4_{\mathbf{0}}(I)$ are solved in the equation (5.35). □

Preliminary experiments show that the closedness subspace strategy tremendously improves the computational efficiency over its predecessor implemented in `ApaTools` as the module `MultiplicityStructure` in both memory and computing time. For instance, `MultiplicityStructure` can handle only $n \leq 6$ for the KSS system (5.33) with multiplicity 42 before running out of memory. The new code increase the capacity to $n=9$ with multiplicity 256. Moreover, the new code is as much as 76 times faster on the KSS system at $n=6$. The speed up reaches thousands on some large systems with low multiplicities [100].

5.4.2 *The fewnomial subspace strategy for multivariate polynomials*

A vector space of multivariate polynomials within a degree bound may have huge dimensions. For example, we can construct a simple GCD test problem with

$$\begin{cases} p = (x_1+\cdots+x_n-1)\big(x_1(x_2^n+\cdots+x_n^n)+2\big), \\ q = (x_1+\cdots+x_n-1)\big(x_n(x_1^n+\cdots+x_{n\text{-}1}^n)-2\big) \end{cases} \tag{5.39}$$

in expanded form. For $n=10$, the dimension of the vector space of polynomial pairs within total degree $n+2$ is 1,293,292, while both p and q are "fewnomials" of only 110 terms. To compute the approximate GCD of (p,q), the standard Sylvester matrix has a size as large as $92,561,040 \times 705,432$ requiring nearly 500 terabytes of memory. It is not practical to construct such a large matrix on a desktop computer.

To compute the approximate GCD of such polynomial pairs, we employ a simple *fewnomial subspace strategy* for reducing the sizes of matrices required for the computation by identifying the monomial support of the GCD and the cofactors. The strategy is similar to the sparse interpolation approach in [15, 33, 105] that is applied to other polynomial problems [30, 48, 51].

We shall use the polynomial pair in (5.39) as an example for $n=3$ to illustrate the fewnomial subspace strategy.

Example 5.4.3. Consider the polynomial pair

$$\begin{aligned} p(x_1,x_2,x_3) &= (x_1+x_2+x_3-1)(x_1x_2^3+x_1x_3^3+2) \\ q(x_1,x_2,x_3) &= (x_1+x_2+x_3-1)(x_1^3x_3+x_2^3x_3-2) \end{aligned}$$

with $u = \mathscr{gcd}(p,q)$ and cofactors v and w. Assign fixed random unit complex values $\hat{x}_2 = .8126623341-.5827348718\,i$ and $\hat{x}_3 = .8826218380+.4700837065\,i$ in (p,q) and compute its univariate approximate GCD in x_1, obtaining

$$
\begin{aligned}
p(x_1,\hat{x}_2,\hat{x}_3) &= (x_1+.6953-.1127i)(-(0.1887-.0381i)x_1+2)\\
q(x_1,\hat{x}_2,\hat{x}_3) &= (x_1+.6953-.1127i)((.8826+.4701i)x_1^3-1.807-.9813i)
\end{aligned}
$$

(showing 4 digits for each number), implying

$$u(x_1,\hat{x}_2,\hat{x}_3) \in span\{1,\ x_1\}, \tag{5.40}$$

$$v(x_1,\hat{x}_2,\hat{x}_3) \in span\{1,\ x_1\}, \quad w(x_1,\hat{x}_2,\hat{x}_3) \in span\{1,\ x_1^3\}. \tag{5.41}$$

Similarly, we can apply univariate GCD in x_2 and

$$u(\hat{x}_1,x_2,\hat{x}_3) \in span\{1,x_2\}, \tag{5.42}$$

$$v(\hat{x}_1,x_2,\hat{x}_3) \in span\{1,x_2^3\}, \ w(\hat{x}_1,x_2,\hat{x}_3) \in span\{1,x_2^3\} \tag{5.43}$$

$$u(\hat{x}_1,\hat{x}_2,x_3) \in span\{1,x_3\}, \tag{5.44}$$

$$v(\hat{x}_1,\hat{x}_2,x_3) \in span\{1,x_3^3\}, \ w(\hat{x}_1,\hat{x}_2,x_3) \in span\{1,x_3\} \tag{5.45}$$

Consequently, combining the monomial bases in (5.40)-(5.43) yields three fewnomial subspaces

$$
\begin{aligned}
u(x_1,x_2,\hat{x}_3) &\in span\{1,\ x_1,\ x_2,\ x_1x_2\},\\
v(x_1,x_2,\hat{x}_3) &\in span\{1,\ x_1,\ x_2^3,\ x_1x_2^3\},\\
w(x_1,x_2,\hat{x}_3) &\in span\{1,\ x_1^3,\ x_2^3,\ x_1^3x_2^3\}.
\end{aligned}
$$

In these subspaces we compute bivariate GCD of (p,q) in x_1 and x_2:

$$
\begin{aligned}
p(x_1,x_2,\hat{x}_3) &= (x_1+x_2-.1174+.4701i)(x_1x_2^3+(.1025+.9947i)x_1+2)\\
q(x_1,x_2,\hat{x}_3) &= (x_1+x_2-.1174+.4701i)((.8826+.4701i)x_1^3+(.8826+.4701i)x_2^3-2).
\end{aligned}
$$

Combining monomial bases in (5.44) and (5.45) yields the fewnomial subspaces

$$
\begin{aligned}
u &\in span\{1,\ x_1,\ x_2,\ x_3,\ x_1x_3,\ x_2x_3\},\\
v &\in span\{1,\ x_1,\ x_1x_2^3,\ x_3^3,\ x_1x_3^3,\ x_1x_2^3x_3^3\},\\
w &\in span\{1,\ x_1^3,\ x_2^3,\ x_3,\ x_1^3x_3,\ x_2^3x_3\}.
\end{aligned}
$$

where u, v and w are computed as approximations to

$$u = x_1+x_2+x_3-1, \quad v = x_1x_2^3+x_1x_3^2+2, \quad w = x_1^3x_3+x_2^3x_3-2$$

□

The process in Example 5.4.3 can be extended to general cases of computing the approximate GCD of multivariate polynomials in $\mathbb{C}[x_1,\cdots,x_s]$: For $k=1,2,\ldots,s$, compute the approximate GCD of

$$
\begin{aligned}
&p(x_1,\ldots,x_k,\hat{x}_{k+1},\ldots,\hat{x}_s)\\
&q(x_1,\ldots,x_k,\hat{x}_{k+1},\ldots,\hat{x}_s)
\end{aligned}
$$

in $x_1, \ldots, x_k$ with remaining variables fixed using random unit complex constants $\hat{x}_{k+1}, \ldots, \hat{x}_s$. Then we can identify the monomial subspaces for u, v and w in the first $k+1$ variables, and on those subspaces we can compute the approximate GCD of (p,q) in the first $k+1$ variables. Continuing this process we complete the GCD computation at $k = s$. This simple technique is tremendously effective in practical computations for sparse polynomials. For the case, say $n = 10$ in (5.39), the largest Sylvester matrix on the fewnomial subspaces has only 40 columns, which is quite a reduction from 705,432.

5.5 Software development

Numerical polynomial algebra is a growing field of study with many algorithms that are still in early stages of development. Nonetheless, many software packages are already available. Most advanced software packages are perhaps the polynomial system solvers based on the homotopy continuation method [57, 82], including PHCPACK [35, 89], BERTINI [4, 5] and MIXEDVOL [31]. There are also specialized implementations of algorithms such as approximate GCD finder QRGCD [14] that is bundled in recent Maple releases, approximate factorization [48], multiplicity structure [6, 88], SNAP package [41] bundled in Maple, univariate factorization and multiple root solver MULTROOT [97], SYNAPS package [66], and COCOA [9, 55, 56], etc. Those packages provide a broad range of versatile tools for applications and academic research.

A comprehensive software toolbox APATOOLS is also in development for approximate polynomial algebra with a preliminary release [99]. APATOOLS is built on two commonly used platforms: Maple and Matlab. The Matlab version is named APALAB. There are two objectives for designing APATOOLS: Assembling robust computational routines as finished product for applications as well as providing a utility library as building blocks for developing more advanced algorithms in numerical polynomial algebra. Currently, APATOOLS includes the following computational routines:

UVGCD: univariate approximate GCD finder (§5.3.1)

MVGCD: multivariate approximate GCD finder (§5.3.1)

UVFACTOR: univariate approximate factorization with multiplicities (§5.2.2)

SQUAREFREEFACTOR: multivariate squarefree factorization (§5.3.4)

MULTIPLICITYSTRUCTURE: dual basis and multiplicity identification (§5.3.2)

POLYNOMIALELIMINATE: numerical and symbolic elimination routine (§5.3.3)

APPROXIRANK: numerical rank/kernel routine (§5.1.2)

NUMJCF: (APALAB only) function for computing the approximate Jordan Canonical Form of inexact matrices (§5.2.2)

Those routines implement algorithms that solve ill-posed algebraic problems for approximate solutions formulated based on the three-strikes principle elaborated in §5.2.2 via a two-staged process: Finding the maximum codimension pejorative manifold by matrix computation, followed by minimizing the distance to the manifold via the Gauss-Newton iteration.

As a growing field of study, numerical polynomial algebra algorithms are in continuing development. ApaTools consists of a comprehensive collection of utility routines designed to simplify software implementation for algorithms. Those utilities include matrix computation tools such as orthogonal transformations and other manipulations, index utilities, monomial utilities, and so on. The two most notable utilities are matrix builder LinearTransformMatrix that conveniently generate matrices from linear transformations between polynomial vector spaces, and the nonlinear least squares solver GaussNewton for minimizing the distance to a pejorative manifold. Both routines are designed with a priority on simplicity for users and with options to maximize efficiency, as illustrated in the following examples.

Example 5.5.1. Construction of Ruppert matrices in ApaTools. The Ruppert matrix (cf. §5.3.4) is the matrix represents the linear transformation $\mathscr{L}_f$ in (5.29) from a given polynomial f of bidegree $[m,n]$, say

$$f \;=\; 2x^3y^3 - 5x^2y^5 + x^2y + 6x^2y^2 - 15xy^4 + 3x - 4xy^2 + 10y^4 - 2$$

of bidegree $[3,5]$. The linear transformation is a combination of the linear transformations

$$\mathscr{L}_{\{,\|} \,:\, \sqcap \longrightarrow [\{\cdot\partial_{\|} - \partial_{\|}\{]\sqcap, \quad \| = \infty, \in \tag{5.46}$$

A Maple user needs to write a straightforward three-line Maple procedure for this linear transformation:

```
> RupLT := proc( u, x, f, k)
     return expand( f*diff(u,x[k]) - diff(f,x[k])*u )
  end proc:
```

Then the Ruppert matrix is constructed instantly by calling

```
> R := < LinearTransformMatrix(RupLT,[f,2],[x,y],
                                [m-1,n],[2*m-1,2*n-1])   |
         LinearTransformMatrix(RupLT,[f,1],[x,y],
                                [m,n-1],[2*m-1,2*n-1])   >;
```

$$R := \left[\begin{array}{l}\text{60 x 38 Matrix}\\ \text{Data Type: anything}\\ \text{Storage: rectangular}\\ \text{Order: Fortran\_order}\end{array}\right]$$

Here, the input items `[f,1]` and `[f,2]` indicate the linear transformation $\mathscr{L}_{\{,\infty}$ and $\mathscr{L}_{\{,\in}$ respectively, $[x,y]$ is the list of indeterminates, `[m-1,n]` and `[m,n-1]` are the bidegree bounds on the domains of the linear transformations, and `[2*m-1,2*n-1]` is the bidegree bound of the range. Applying the numerical kernel routine ApproxiRank

```
> r, s, N := ApproxiRank(R,[1e-8]);
```

yields the numerical nullity 2 that reveals the number of irreducible factors of f.

□

Example 5.5.2. Applying the Gauss-Newton iteration in factorization. After obtaining an initial approximation

$$\begin{aligned} u_1 &= -1.99999+2.99999x+x^2y \\ u_2 &= .99998+2xy^2-4.99998y^4 \end{aligned}$$

to the irreducible factors of the polynomial f in Example 5.5.1, we seek the least squares solution to the overdetermined system of equations

$$\begin{cases} \phi^{\mathsf{H}} u_1 - 1 &= 0 \\ [\![u_1 \cdot u_2 - f]\!] &= \mathbf{0} \end{cases}$$

(cf. Example 5.1.3 and §5.3.4). A user needs to write a straightforward Maple procedure for this system of equations

```
> FacFunc := proc( w, x, f, phi)
     return [PolynomialDotProduct( phi,w[1],x)-1,
                                expand(w[1]*w[2])-f]
     end proc;
```

prepare input data and make a call on the ApaTools routing GaussNewton

```
> factors, residual := GaussNewton(FacFunc,[u1,u2],
                    [[x,y],f,phi], [1e-12,9,true]);

    Gauss-Newton step   0,  residual =   1.71e-04
    Gauss-Newton step   1,  residual =   4.82e-10
    Gauss-Newton step   2,  residual =   3.46e-14
```

$$\begin{aligned} &factors, residual := [-2.00000714288266+3.00001071432397x+ \\ &\quad 1.00000357144133x^2y,\ .999996428571430+1.99999285714286xy^2- \\ &\quad 4.99998214285715y^4],\ .346410161513776\ 10^{-13} \end{aligned}$$

The result shows computed irreducible factors with a backward error 3.5×10^{-14}, and the factors with a scaling

$$-2.00000000000000+2.99999999999998x+1.00000000000000x^2y$$

and

$$1.00000000000000 + 2.00000000000001xy^2 - 5.00000000000002y^4$$

are accurate near hardware precision. □

ApaTools is an on-going project with its functionality and library still expanding. We wish to build a comprehensive software toolbox for applications and algorithm development. The near term plan is to incorporate approximate irreducible factorization, fully implement the closedness/fewnomial subspace strategy for enhancing the efficiency, and collect a library of benchmark test problems for numerical polynomial algebra.

Acknowledgements. The author is supported in part by the National Science Foundation of U.S. under Grants DMS-0412003 and DMS-0715127. The author thanks Hans Stetter for providing his student's thesis [88], and Erich Kaltofen for his comment as well as several references in §5.4.2.

References

1. E. L. Allgower, K. Georg, and R. Miranda, *The method of resultants for computing real solutions of polynomial systems*, SIAM J. Numer. Anal., 29 (1992), pp. 831–844.
2. W. Auzinger and H. J. Stetter, *An elimination algorithm for the computation of all zeros of a system of multivariate polynomial equations*. Proc. of International Conference on Numerical Mathematics Singapore, 1988.
3. D. J. Bates and J. D. Hauenstein and C. Peterson and A. J. Sommese, *A local dimension test for numerically approximate points on algebraic sets*, Preprint, 2008.
4. D. Bates, J. D. Hauenstern, A. J. Sommese, and C. W. Wampler, *Bertini: Software for Numerical Algebraic Geometry*. http://www.nd.edu/∼sommese/bertini, 2006.
5. ———, *Software for numerical algebraic geometry: A paradigm and progress towards its implementation*, in Software for Algebraic Geometry, IMA Volume 148, M. Stillman, N. Takayama, and J. Verschelde, eds., Springer, 2008, pp. 1–14.
6. D. J. Bates, C. Peterson, and A. J. Sommese, *A numerical-symbolic algorithm for computing the multiplicity of a component of an algebraic set*, J. of Complexity, 22 (2006), pp. 475–489.
7. G. Chèze and A. Galligo, *Four lessons on polynomial absolute factorization*, in Solving Polynomial Equations: Foundations, Algorithms, and Applications, A. Dickenstein and I. Emiris, eds., vol. 14 of Algorithms and Computation in Mathematics, Springer-Verlag, 2005, pp. 339–392.
8. ———, *From an approximate to an exact absolute polynomial factorization*, J. of Symbolic Computation, 41 (2006), pp. 862–696.
9. The CoCoa Team, CoCoA*: a system for doi ng Computations in Commutative Algebra*. Available at `http://cocoa.dima.unige.it`.
10. R. M. Corless, A. Galligo, I. Kotsireas, and S. Watt, *A geometric-numeric algorithm for factoring multivariate polynomials*. Proc. ISSAC'02, ACM Press, pages 37-45, 2002.
11. R. M. Corless, P. M. Gianni, B. M. Trager, and S. M. Watt, *The singular value decomposition for polynomial systems*. Proc. ISSAC '95, ACM Press, pp 195-207, 1995.
12. R. M. Corless, M. Giesbrecht, D. Jeffrey, and S. Watt, *Approximate polynomial decomposition*. Proc. ISSAC'99, ACM Press, pages 213-220, 1999.

13. R. M. Corless, M. Giesbrecht, M. van Hoeij, I. Kotsireas, and S. Watt, *Towards factoring bivariate approximate polynomials*. Proc. ISSAC'01, ACM Press, pages 85-92, 2001.
14. R. M. Corless, S. M. Watt, and L. Zhi, *QR factoring to compute the GCD of univariate approximate polynomials*, IEEE Trans. Signal Processing, 52 (2003), pp. 3394–3402.
15. A. Cuyt and W.-s. Lee, *A new algorithm for sparse interpolation of multivariate polynomials*, Theoretical Computer Science, Vol. 409, pp 180-185, 2008
16. D. Daney, I. Z. Emiris, Y. Papegay, E. Tsigaridas, and J.-P. Merlet, *Calibration of parallel robots: on the elimination of pose-dependent parameters*, in Proc. of the first European Conference on Mechanism Science (EuCoMeS), 2006.
17. B. Dayton and Z. Zeng, *Computing the multiplicity structure in solving polynomial systems*. Proceedings of ISSAC '05, ACM Press, pp 116–123, 2005.
18. J.-P. Dedieu and M. Shub, *Newton's method for over-determined system of equations*, Math. Comp., 69 (1999), pp. 1099–1115.
19. J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.
20. J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall Series in Computational Mathematics, Prentice-Hall, Englewood Cliffs, New Jersey, 1983.
21. I. Z. Emiris, *Sparse elimination and application in kinematics*, PhD thesis, Computer Science Division, Dept. of Elec. Eng. and Comput. Sci., Univ. of California, Berkeley, 1994.
22. ———, *A general solver based on sparse resultants*, in Proc. PoSSo (Polynomial System Solving) Workshop on Software, 1995, pp. 35–54.
23. I. Z. Emiris, E. D. Fritzilas, and D. Manocha, *Algebraic algorithms for determining structure in biological chemistry*, Internatinal J. Quantum Chemistry, 106 (2006), pp. 190–210.
24. I. Z. Emiris, A. Galligo, and H. Lombardi, *Certified approximate univariate GCDs*, J. Pure Appl. Algebra, 117/118 (1997), pp. 229–251.
25. I. Z. Emiris and B. Mourrain, *Computer algebra methods for studying and computing molecular conformations*, Algorithmica, 25 (1999), pp. 372–402.
26. W. Fulton, *Intersection Theory*, Springer Verlag, Berlin, 1984.
27. A. Galligo and D. Rupprecht, *Semi-numerical determination of irreducible branches of a reduced space curve*. Proc. ISSAC'01, ACM Press, pages 137-142, 2001.
28. A. Galligo and S. Watt, *A numerical absolute primality test for bivariate polynomials*. Proc. ISSAC'97, ACM Press, pages 217–224, 1997.
29. S. Gao, *Factoring multivariate polynomials via partial differential equations*, Math. Comp., 72 (2003), pp. 801–822.
30. S. Gao, E. Kaltofen, J. May, Z. Yang, and L. Zhi, *Approximate factorization of multivariate polynomials via differential equations*. Proc. ISSAC '04, ACM Press, pp 167-174, 2004.
31. T. Gao and T.-Y. Li, *MixedVol: A software package for mixed volume computation*, ACM Trans. Math. Software, 31 (2005), pp. 555–560.
32. I. Gelfand, M. Kapranov, and A. Zelevinsky, *Discriminants, Resultants and Multidimensional determinants*, Birkhäuser, Boston, 1994.
33. M. Giesbrecht, G. Labahn and W-s. Lee, *Symbolic-numeric sparse interpolation of multivariate polynomials*, Journal of Symbolic Computation, to appear, 2009.
34. G. H. Golub and C. F. Van Loan, *Matrix Computations*, The John Hopkins University Press, Baltimore and London, 3rd ed., 1996.
35. Y. Guan and J. Verschelde, *PHClab: A MATLAB/Octave interface to PHCpack*, in Software for Algebraic Geometry, IMA Volume 148, M. Stillman, N. Takayama, and J. Verschelde, eds., Springer, 2008, pp. 15–32.
36. J. Hadamard, *Sur les problèmes aux dèrivèes partielles et leur signification physique*. Princeton University Bulletin, 49–52, 1902.
37. P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*, SIAM, 1997.

38. M. Hitz, E. Kaltofen, and Y. N. Lakshman, *Efficient algorithms for computing the nearest polynomial with a real root and related problems*. Proc. ISSAC'99, ACM Press, pp 205-212, 1999.
39. V. Hribernig and H. J. Stetter, *Detection and validation of clusters of polynomial zeros*, J. Symb. Comput., 24 (1997), pp. 667–681.
40. Y. Huang, W. Wu, H. Stetter, and L. Zhi, *Pseudofactors of multivariate polynomials*. Proc. ISSAC '00, ACM Press, pp 161-168, 2000.
41. C.-P. Jeannerod and G. Labahn, *The SNAP package for arithemetic with numeric polynomials*. In International Congress of Mathematical Software, World Scientific, pages 61-71, 2002.
42. G. F. Jónsson and S. A. Vavasis, *Accurate solution of polynomial equations using Macaulay resultant matrices*, Math. Comp., 74 (2004), pp. 221–262.
43. W. Kahan, *Conserving confluence curbs ill-condition*. Technical Report 6, Computer Science, University of California, Berkeley, 1972.
44. E. Kaltofen, *Challenges of symbolic computation: my favorite open problems*, J. Symb. Comput., 29 (2000), pp. 161–168.
45. E. Kaltofen, B. Li, Z. Yang, and L. Zhi, *Exact certification of global optimality of approximate factorization via rationalizing sums-of-squares with floating point scalars*, Proc. ISSAC '08, ACM Press, pp 155-163, 2008.
46. E. Kaltofen and J. May, *On approximate irreducibility of polynomials in several variables*. Proc. ISSAC '03, ACM Press, pp 161-168, 2003.
47. E. Kaltofen, J. May, Z. Yang, and L. Zhi, *Structured low rank approximation of Sylvester matrix*, in Symbolic-Numeric Computation, Trends in Mathematics, D. Wang and L. Zhi, editors, Birkhäuser Verlag, Basel, Switzerland, (2007), pp. 69–83.
48. ———, *Approximate factorization of multivariate polynomials using singular value decomposition*, J. of Symbolic Computation, 43 (2008), pp. 359–376.
49. E. Kaltofen, Z. Yang, and L. Zhi, *Structured low rank approximation of a Sylvester matrix*, Symbolic-Numeric Computation, D. Wang and L. Zhi, Eds, Trend in Mathematics, Birkhäuser Verlag Basel/Switzerland, pp. 69-83, 2006
50. ———, *Approximate greatest common divisor of several polynomials with linearly constrained coefficients and singular polynomials*. Proc. ISSAC'06, ACM Press, pp 169–176, 2006.
51. ———, *On probabilistic analysis of randomization in hybrid symbolic-numeric algorithms*, SNC'07 Proc. 2007 Internat. Workshop on Symbolic-Numeric Comput. pp. 11-17, 2007.
52. N. K. Karmarkar and Y. N. Lakshman, *Approximate polynomial greatest common divisors and nearest singular polynomials*. Proc. ISSAC'96, pp 35-42, ACM Press, 1996.
53. ———, *On approximate polynomial greatest common divisors*, J. Symb. Comput., 26 (1998), pp. 653–666.
54. H. Kobayashi, H. Suzuki, and Y. Sakai, *Numerical calculation of the multiplicity of a solution to algebraic equations*, Math. Comp., 67 (1998), pp. 257–270.
55. M. Kreuzer and L. Robbiano, *Computational Commutative Algebra 1*, Springer Verlag, Heidelberg, 2000.
56. ———, *Computational Commutative Algebra 2*, Springer Verlag, Heidelberg, 2000.
57. T.-Y. Li, *Solving polynomial systems by the homotopy continuation method*, Handbook of Numerical Analysis, XI, edited by P. G. Ciarlet, North-Holland, Amsterdam (2003), pp. 209–304.
58. T.-Y. Li. Private communication, 2006.
59. T. Y. Li and Z. Zeng, *A rank-revealing method with updating, downdating and applications*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 918–946.
60. F. S. Macaulay, *The Algebraic Theory of Modular Systems*, Cambridge Univ. Press, London, 1916.
61. D. Manocha and S. Krishnan, *Solving algebraic systems using matrix computations*, Communication in Computer Algebra, 30 (1996), pp. 4–21.
62. M. G. Marinari, T. Mora, and H. M. Möller, *On multiplicities in polynomial system solving*, Trans. AMS, 348 (1996), pp. 3283–3321.

63. ÅKE BJÖRCK, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.
64. T. MORA, *Solving Polyonmial Equation Systems II*, Cambridge Univ. Press, London, 2004.
65. B. MOURRAIN, *Isolated points, duality and residues*, J. of Pure and Applied Algebra, 117 & 118 (1996), pp. 469–493. Special issue for the Proc. of the 4th Int. Symp. on Effective Methods in Algebraic Geometry (MEGA).
66. B. MOURRAIN AND J.-P. PAVONE, *SYNAPS, a library for dedicated applications in symbolic numeric computing*, in Software for Algebraic Geometry, IMA Volume 148, M. Stillman, N. Takayama, and J. Verschelde, eds., Springer, 2008, pp. 81–100.
67. M.-T. NADA AND T. SASAKI, *Approximate GCD and its application to ill-conditioned algebraic equations*, J. Comput. Appl. Math., 38 (1991), pp. 335–351.
68. V. Y. PAN, *Numerical computation of a polynomial GCD and extensions*, Information and Computation, 167 (2001), pp. 71–85.
69. S. PILLAI AND B. LIANG, *Blind image deconvolution using GCD approach*, IEEE Trans. Image Processing, 8 (1999), pp. 202–219.
70. G. REID AND L. ZHI, *Solving nonlinear polynomial systems*. Proc. international conference on polynomial system solving, pp. 50-53, 2004.
71. W. RUPPERT, *Reducibility of polynomials* $f(x,y)$ *modulo* p, J. Number Theory, 77 (1999), pp. 62–70.
72. D. RUPPRECHT, *An algorithm for computing certified approximate GCD of n univariate polynomials*, J. Pure and Appl. Alg., 139 (1999), pp. 255–284.
73. T. SASAKI, *Approximate multivariate polynomial factorization based on zero-sum relations*. Proc. ISSAC'01, ACM Press, pp 284-291, 2001.
74. T. SASAKI, T. SAITO, AND T. HILANO, *Analysis of approximate factorization algorithm I*, Japan J. Industrial and Applied Math, 9 (1992), pp. 351–368.
75. ———, *A unified method for multivariate polynomial factorization*, Japan J. Industrial and Applied Math, 10 (1993), pp. 21–39.
76. T. SASAKI, M. SUZUKI, M. KOLAR, AND M. SASAKI, *Approximate factorization of multivariate polynomials and absolute irreducibility testing*, Japan J. Industrial and Applied Math, 8 (1991), pp. 357–375.
77. A. SCHÖNHAGE, *Quasi-GCD computations*, J. Complexity, 1 (1985), pp. 118–137.
78. A. J. SOMMESE, J. VERSCHELDE, AND C. W. WAMPLER, *Numerical irreducible decomposition using PHCpack*. In *Algebra, Geometry and Software Systems*, edited by M. Joswig et al, Springer-Verlag 2003, 109-130.
79. ———, *Numerical irreducible decomposition using projections from points on the components*. In J. Symbolic Computation: Solving Equations in Algebra, Geometry and Engineering, volumn 286 of Comtemporary Mathematics, edited by E.L. Green et al, pages 37-51, AMS 2001.
80. ———, *Numerical factorization of multivariate complex polynomials*, Theoretical Computer Science, 315 (2003), pp. 651–669.
81. ———, *Introduction to numerical algebraic geometry*, in Solving Polynomial Equations, A. Dickenstein and I. Z. Emiris, eds., Springer-Verlag Berlin Heidelberg, 2005, pp. 301–337.
82. A. J. SOMMESE AND C. W. WAMPLER, *The Numerical Solution of Systems of Polynomials*, World Scientific Pub., Hackensack, NJ, 2005.
83. H. J. STETTER, *Matrix eigenproblems are at the heart of polynomial system solving*, ACM SIGSAM Bulletin, 30 (1996), pp. 22–25.
84. ———, *Numerical Polynomial Algebra*, SIAM, 2004.
85. G. W. STEWART, *Matrix Algorithms, Volume I, Basic Decompositions*, SIAM, Philadelphia, 1998.
86. M. STILLMAN, N. TAKAYAMA, AND J. VERSCHELDE, eds., *Software for Algebraic Geometry*, vol. 148 of The IMA Volumes in Mathematics and its Applications, Springer, 2008.
87. H.-J. SU, C. W. WAMPLER, AND J. MCCARTHY, *Geometric design of cylindric PRS serial chains*, ASME J. Mech. Design, 126 (2004), pp. 269–277.
88. G. H. THALLINGER, *Analysis of Zero Clusters in Multivariate Polynomial Systems*. Diploma Thesis, Tech. Univ. Vienna, 1996.

89. J. Verschelde, *Algorithm 795: PHCpack: A general-purpose solver for polynomial systems by homotopy continuation*, ACM Trans. Math. Software, (1999), pp. 251–276.
90. C. W. Wampler, *Displacement analysis of spherical mechanisms having three or fewer loops*, ASME J. Mech. Design, 126 (2004), pp. 93–100.
91. ———, *Solving the kinematics of planar mechanisms by Dixon determinant and a complex-plane formulation*, ASME J. Mech. Design, 123 (2004), pp. 382–387.
92. P.-Å. Wedin, *Perturbation bounds in connection with singular value decomposition*, BIT, 12 (1972), pp. 99–111.
93. J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, New York, 1965.
94. X. Wu and L. Zhi, *Computing the multiplicity structure from geometric involutive form.* Proc. ISSAC'08, ACM Press, pages 325–332, 2008.
95. Z. Zeng, *A polynomial elimination method for numerical computation.* Theoretical Computer Science, Vol. 409, pp. 318-331, 2008.
96. Z. Zeng, *The approximate GCD of inexact polynomials. Part I.* to appear.
97. ———, *Algorithm 835: MultRoot – a Matlab package for computing polynomial roots and multiplicities*, ACM Trans. Math. Software, 30 (2004), pp. 218–235.
98. ———, *Computing multiple roots of inexact polynomials*, Math. Comp., 74 (2005), pp. 869–903.
99. Z. Zeng, *ApaTools: A Maple and Matlab toolbox for approximate polynomial algebra*, in Software for Algebraic Geometry, IMA Volume 148, M. Stillman, N. Takayama, and J. Verschelde, eds., Springer, 2008, pp. 149–167.
100. Z. Zeng, *The closedness subspace method for computing the multiplicity structure of a polynomial system*, to appear: Contemporary Mathematics series, American Mathematical Society, 2009.
101. Z. Zeng, *The approximate irreducible factorization of a univariate polynomial. Revisited*, preprint, 2009.
102. Z. Zeng and B. Dayton, *The approximate GCD of inexact polynomials. II: A multivariate algorithm.* Proceedings of ISSAC'04, ACM Press, pp. 320-327. (2006).
103. Z. Zeng and T. Y. Li, *A numerical method for computing the Jordan Canonical Form.* Preprint, 2007.
104. L. Zhi, *Displacement structure in computing approximate GCD of univariate polynomials.* In Proc. Sixth Asian Symposium on Computer Mathematics (ASCM 2003), Z. Li and W. Sit, Eds, vol. 10 of Lecture Notes Series on Computing World Scientific, pp. 288-298, 2003.
105. R.E. Zipple, *Probabilistic algorithms for sparse polynomials*, Proc. EUROSAM '79, Springer Lec. Notes Comp. Sci., 72, pp. 216-226, 1979.

Chapter 6
Ideal Interpolation: Translations to and from Algebraic Geometry

Boris Shekhtman

Abstract In this survey I will discuss four themes that surfaced in multivariate interpolation and seem to have analogues in algebraic geometry. The hope is that mixing these two areas together will benefit both.

6.1 Introduction

> *We have many things in common with the Americans except the language.*
> (Oscar Wilde)

The aim of this survey is to describe some of the problems arising in multivariate interpolation that seem to be amenable to the methods of algebraic geometry. In other words, to ask for help. The questions concern “ideal projectors”. These are linear idempotent operators on the ring of polynomials that have one additional property: the kernels of these operators are ideals in $\mathbb{k}[\mathbf{x}]$.

In one variable ideal projectors are precisely the Hermite interpolation projectors, the well-studied and beloved workhorses of numerical analysis. A simple-minded strategy is to take your favorite property of Hermite projectors and ask for its extension to a multivariate setting. This plan is as primitive as it is futile. Nevertheless, I will follow this very strategy; if for no other reason than to prove the point. So be prepared to read the phrase “in one variable” a lot; just like at the start of this paragraph.

This first section is an attempt to bridge the proverbial “boundary of a common language” between Approximation Theory (AT) and Algebraic Geometry (AG). The appendix to this article contains a short “AT-AG dictionary” aimed at helping with translations. Like every dictionary, this one is not exact, but it may help in making the right associations. It certainly helped me.

Department of Mathematics & Statistics, University of South Florida, Tampa, FL 33620, USA, e-mail: boris@math.usf.edu

L. Robbiano, J. Abbott (eds.), *Approximate Commutative Algebra*, Texts and Monographs in Symbolic Computation, DOI 10.1007/978-3-211-99314-9_6, © Springer-Verlag/Wien 2009

In the later sections I describe in detail some problems and developments toward their resolution in the following four topics:

Section 2: (AT) The limits of Lagrange projectors.
(AG) Radical component of a border scheme.
Section 3: (AT) Restrictions of ideal projectors.
(AG) Grading of $\mathbb{k}[\mathbf{x}]/I$ and its dualizing module.
Section 4: (AT) Error formulas.
(AG) Ideal representation.
Section 5: (AT) Minimal interpolating subspaces.
(AG) Cohomological dimension and arithmetic rank of the Hilbert scheme.

I have no doubt that a reader, familiar with algebraic geometry, will periodically exclaim something like, "But this is trivial...". Well, good! May this be a motivation for her or him to take a crack at the problems.

6.1.1 Ideal projectors

In approximation theory we seek an approximation to a function f by a polynomial, preferably of small degree; that is by a polynomial that comes from a fixed N-dimensional subspace G of $\mathbb{k}[\mathbf{x}] := \mathbb{k}[x_1,\dots,x_d]$. The function f is either given explicitly or by limited information, such as its values at a few points. The ground field $\mathbb{k}$ is typically the real field $\mathbb{R}$ or the complex field $\mathbb{C}$, thus in this article $\mathbb{k}$ will always refer to one of these fields. The method of approximation is encapsulated in the *projector* (linear idempotent operator) P from $\mathbb{k}[\mathbf{x}]$ onto G. While (depending on the norm on $\mathbb{k}[\mathbf{x}]$) a best approximation is hard to compute, the linearity of P gives a realistic hope of developing an algorithm for determining $Pf \in G$. The idempotence of P assures the quality of the approximation. Since $Pg = g$ for every $g \in G$,

$$\|f - Pf\| = \|f - g - P(f-g)\| \le \|I - P\|\,\|f - g\|$$

for every $g \in G$. Taking the infimum over all such $g \in G$, we conclude that $\|f - Pf\| \le (1 + \|P\|)\,\mathrm{dist}(f, G)$. Thus, up to the constant $K := \|I - P\|$ that does not depend on f, the error of this approximation is as good as can be hoped. A prototypical example of such a projector is the *Lagrange projector* defined by N distinct points (sites): $\mathbf{z}_1,\dots,\mathbf{z}_N \in \mathbb{k}^d$ and having the property that $f(\mathbf{z}_j) = Pf(\mathbf{z}_j)$ for all $j = 1,\dots,N$ and all $f \in \mathbb{k}[\mathbf{x}]$. The kernel of this projector is an ideal

$$\ker P = \left\{ f \in \mathbb{k}[\mathbf{x}] : f(\mathbf{z}_j) = 0, j = 1,\dots,N \right\} \subset \mathbb{k}[\mathbf{x}]$$

and in fact is a radical ideal with the associated variety $\mathscr{V}(\ker P) = \{\mathbf{z}_1,\dots,\mathbf{z}_N\}$.

Definition 6.1.1 (Birkhoff [2]). A projector P on $\mathbb{k}[\mathbf{x}]$ is called an ideal projector if $\ker P$ is an ideal in $\mathbb{k}[\mathbf{x}]$.

The following observation of Carl de Boor [11] stems from the fact that Pf is a result of division of f by the ideal $\ker P$.

Theorem 6.1.2. *A linear operator P on $\mathbb{k}[\mathbf{x}]$ is an ideal projector if and only if*

$$P(fg) = P(fPg) \tag{6.1}$$

for all $f, g \in \mathbb{k}[\mathbf{x}]$.

We will leave the simple verification of this fact to the reader and refer to (6.1) in the rest of the article as *de Boor's formula.*

Of course, if the range G of the projector P is given, the study of ideal projectors is the same as the study of ideals $J \subset \mathbb{k}[\mathbf{x}]$ that complement G:

$$J \oplus G = \mathbb{k}[\mathbf{x}]$$

or, in algebraic terminology, the ideals $J \subset \mathbb{k}[\mathbf{x}]$ such that G spans the algebra $A := \mathbb{k}[\mathbf{x}]/J$. The de Boor's formula simply says that $[f[g]] = [fg]$ in A.

Let $\mathfrak{P}_G$ stand for the family of all ideal projectors onto G and $\mathfrak{J}_G$ for the family of all ideals that complement G. There is a one-to-one correspondence between these two sets. An approximation theorist prefers to study the set $\mathfrak{P}_G$, while an algebraist seems to focus on $\mathfrak{J}_G$. The reasons are partly cultural and partly motivated by the specifics of the problems at hand.

Of particular importance are projectors onto the space $G := \mathbb{k}[\mathbf{x}]_{<n}$ of polynomials of degree less than n. Thus some of the results are specifically aimed at these projectors.

The ideal projectors, just like ideals, have a set of attributes that we now describe.

6.1.2 Parametrization

One of the consequences of de Boor's formula is that an ideal projector onto G is completely defined by its values on a relatively small class of polynomials $\{1, x_i g, i = 1, \ldots, d, g \in G\}$ or, by linearity, on the finite set

$$\partial G := \{1, x_i g_k, i = 1, \ldots, d, k = 1, \ldots, \dim G\} \setminus G \tag{6.2}$$

where $\{g_k\}$ is a $\mathbb{k}$-bases for G. The polynomials $\{f - Pf, f \in \partial G\}$ form the (border) basis for $\ker P$. The polynomials

$$Pf = \sum w_{P,f,k} g_k, f \in \partial G \tag{6.3}$$

define the (generalized) sequence of coefficients

$$\mathbf{w}_P := \left(w_{P,f,k} : f \in \partial G, k = 1, \ldots, \dim G\right)$$

and the set of all such sequences

$$\mathscr{P}_G := \{\mathbf{w}_P : P \in \mathfrak{P}_G\}$$

parametrizes $\mathfrak{P}_G$. For a sequence $\mathbf{w} \in \Bbbk^{\#\partial G \times \dim G}$ to be in $\mathscr{P}_G$, the projector P defined by (6.3) must satisfy equations (6.1) for all f and g, which is the set of algebraic equations for $\mathbf{w}$. This gives $\mathscr{P}_G$ the structure of an affine variety (AKA border scheme) and, for some G (cf. [13]), amounts to nothing more than enforcing the Buchberger criteria on the polynomial $\{f - p, f \in \partial G, p \in G\}$.

The following example is not original (cf. [29, 40, 46]) but it is simple and will serve as an illustration throughout the paper.

Illustration 6.1.3. Let $G = \Bbbk[x,y]_{<2} = \operatorname{span}\{1,x,y\} \subset \Bbbk[x,y]$. Let P be an ideal projector onto G and

$$\begin{aligned} Px^2 &= a_0 + b_0x + c_0y \\ Pxy &= a_1 + b_1x + c_1y \\ Py^2 &= a_2 + b_2x + c_2y \end{aligned} \tag{6.4}$$

The coefficients in (6.4) determines the value of Pf on any polynomial $f \in \Bbbk[x,y]$ and, moreover

$$Pf = a_f + b_fx + c_fy$$

where a_f, b_f and c_f are polynomials in $\Bbbk\left[a_0,b_0,c_0,a_1,b_1,c_1,a_2,b_2,c_2\right]$. Equations $P(xPxy) = P\left(yPx^2\right)$ and $P(yPxy) = P\left(xPy^2\right)$ form the set of algebraic equations for the coefficients in (6.4). The solutions to these equations are given by

$$\begin{aligned} a_0 &= -b_0c_1 + c_1^2 + b_1c_0 - c_0c_2, \\ a_1 &= b_2c_0 - b_1c_1, \\ a_2 &= b_1^2 - c_2b_1 - b_0b_2 + b_2c_1. \end{aligned} \tag{6.5}$$

Thus $\mathscr{P}_G$ is the six-dimensional affine variety in $\Bbbk^9$, consisting of points

$$(a_0,b_0,c_0,a_1,b_1,c_1,a_2,b_2,c_2)$$

satisfying (6.5). A border bases for the ideal $\ker P$ is

$$\left(x^2 - Px^2, xy - Pxy, y^2 - Py^2\right).$$

Remark 6.1.4. It is an interesting feature of ideal projectors P onto $\Bbbk[\mathbf{x}]_{<n}$ in two or more variables (cf. [46]) that the coordinates of $\mathbf{w}_P \in \mathscr{P}_{\Bbbk[\mathbf{x}]_{<n}}$ are fixed polynomials in the coefficients of the leading forms of $P\mathbf{x}^\alpha$, $|\alpha| = n$. In two (and only two) variables, every collection of forms p_α of degree $n-1$ uniquely determines an ideal projector P by

$$\text{leading form } (P\mathbf{x}^\alpha) = p_\alpha, \quad |\alpha| = n.$$

In particular, $\mathscr{P}_{\Bbbk[x,y]_{<n}}$ is a polynomial image of $\Bbbk^{n(n+1)}$ and thus an irreducible $n(n+1) = 2 \times \dim \Bbbk_{<n}[x,y]$-dimensional affine variety in $\Bbbk^{\frac{n(n+1)^2}{2}}$. More on this subject can be found in [46].

6.1.3 Multiplication operators

With any ideal projector P and any $f \in \Bbbk[\mathbf{x}]$ we associate the *multiplication operator M_f* defined on G by $M_f(g) := P(fg)$. In particular

$$M_j g := P(x_j g).$$

These operators are similar, literally and figuratively, to the multiplication mappings (endomorphisms) m_j defined on $\Bbbk[\mathbf{x}]/\ker P$ (cf. [15, 35, 48]). Thus $\mathbf{M} = \mathbf{M}_P := (M_j)$ is a *cyclic d*-tuple of pairwise commuting operators on G. The word "cyclic" refers to the existence of a cyclic vector $g_0 \in G$ (in this case g_0 can be chosen to be $P1$) such that

$$\{f(M_1, \ldots, M_d)g_0 : f \in \Bbbk[\mathbf{x}]\} = G.$$

The converse is also true. It is based on the "toy model" of the Hilbert scheme, described in Nakajima [33], albeit in two variables. Here is a translation into the language of ideal projectors:

Theorem 6.1.5 ([15]). *Let $\mathbf{L} := (L_j, j = 1, \ldots, d)$ be a cyclic sequence of commuting operators on G, g_0 be the cyclic vector for $\mathbf{L}$. Then*

(i) *The map*

$$\begin{aligned} \varphi_{\mathbf{L}} : \Bbbk[\mathbf{x}] &\to G \\ f &\to f(L_1, \ldots, L_d)g_0 \end{aligned} \tag{6.6}$$

is a surjective ring homomorphism. The ideal $\ker \varphi_{\mathbf{L}}$ complements G and the restriction $\varphi_{\mathbf{L}|G}$ of $\varphi_{\mathbf{L}}$ to G is an isomorphism.

(ii) *The projector $P_{\mathbf{L}}$ onto G with $\ker P = \ker \varphi_{\mathbf{L}}$ is an ideal projector and is given explicitly by the formula*

$$P_{\mathbf{L}} = \left(\varphi_{\mathbf{L}|G}\right)^{-1} \circ \varphi_{\mathbf{L}}. \tag{6.7}$$

(iii) *The sequence (L_j) is similar to multiplication operators (M_j) for the projector $P_{\mathbf{L}}$.*

The last part of this theorem was noted in [15] and [35], where it is shown that (L_j) is similar to multiplication endomorphisms (m_j) on $\Bbbk[x]/\ker \varphi_{\mathbf{L}}$ and, therefore, is similar to the multiplication operators (M_j) for the projector $P_{\mathbf{L}}$. The rest is self-explanatory (cf. [15]). The formula (6.7) has appeared in [15]. For deeper results on the relationship between border basis and commuting endomorphisms we refer to [24] and [35].

Illustration 6.1.6. The operators M_1, M_2 for projector (6.4) are represented by matrices

$$M_1 = M_x = \begin{bmatrix} 0 & a_0 & a_1 \\ 1 & b_0 & b_1 \\ 0 & c_0 & c_1 \end{bmatrix}, M_2 = M_y = \begin{bmatrix} 0 & a_1 & a_2 \\ 0 & b_1 & b_2 \\ 1 & c_1 & c_2 \end{bmatrix}$$

in the bases $1, x, y$. It is easy (and instructive) to check that these matrices commute if and only if (6.4) holds. The cyclic vector is $(1,0,0) = 1 \in \mathbb{k}[x,y]$. The ideal

$$\ker P = \{f \in \mathbb{k}[x,y] : f(M_1, M_2)\, 1 = 0\} = \{f \in \mathbb{k}[x,y] : f(M_1, M_2) = 0\}.$$

6.1.4 Duality

Every projector onto G (ideal or not) can be written in the form

$$P = \sum_{j=1}^{\dim G} \lambda_j \otimes g_j, \tag{6.8}$$

i.e. $Pf = \sum_{j=1}^{\dim G} \lambda_j(f) \otimes g_j$, where (g_j) is a basis for G and (λ_j) are bi-orthogonal functionals in the algebraic dual $\mathbb{k}'[\mathbf{x}]$ of $\mathbb{k}[\mathbf{x}]$.

The space $\Lambda := \operatorname{span}\{\lambda_j\} \subset \mathbb{k}'[\mathbf{x}]$ is the range of the dual projector

$$P^* := \sum_{j=1}^{\dim G} g_j \otimes \lambda_j,$$

$\dim \Lambda = \dim G$ and

$$\ker \Lambda := \{f \in \mathbb{k}[\mathbf{x}] : \lambda(f) = 0 \text{ for all } \lambda \in \Lambda\} = \ker P \tag{6.9}$$

satisfies

$$\ker \Lambda \cap G = \{0\}. \tag{6.10}$$

Conversely, if $\dim \Lambda = \dim G$ and (6.10) holds, then (6.9) defines a projector onto G. In this case we will say that Λ *is correct for* G. In other words, Λ is correct for G if and only if for every $f \in \mathbb{k}[\mathbf{x}]$ there exists unique $g \in G$ such that

$$\lambda(f) = \lambda(g), \quad \forall \lambda \in \Lambda,$$

i.e. the projector P *interpolates at the functionals* λ_j:

$$\lambda_j(f) = \lambda_j(Pf) \text{ for all } f \in \mathbb{k}[\mathbf{x}]. \tag{6.11}$$

In algebraic language (cf. [7]), if $A = \mathbb{k}[\mathbf{x}]/\ker P$, then $\Lambda = \hat{A}$ is its dual and the ideal $\ker P = \operatorname{Ann}(\Lambda)$.

To give $\mathbb{k}'[\mathbf{x}]$ the structure of a module over $\mathbb{k}[\mathbf{x}]$, we identify the space $\mathbb{k}'[\mathbf{x}]$ with the space $\mathbb{k}[[\mathbf{x}]]$ of formal power series (inverse system) as follows:

To every element $\lambda \in \Bbbk[[x_1,\ldots,x_d]]$ we associate a differential operator $\lambda(\mathbf{D}) \in \Bbbk[D_1,\ldots,D_d]$ by formally replacing variables in λ with the appropriate operators D_j, which are partial derivatives with respect to x_j. Now, for every $\lambda \in \mathbb{C}[[x_1,\ldots,x_d]]$ we define the functional $\tilde{\lambda} \in \Bbbk'[\mathbf{x}]$ by

$$\tilde{\lambda}(f) := (\bar{\lambda}(\mathbf{D})f)(0) \text{ for every } f \in \Bbbk[\mathbf{x}]. \tag{6.12}$$

It is well-known (cf. [14, 7, 30]) that the map $\lambda \longmapsto \tilde{\lambda}$ defined by (6.12) is a linear isomorphism between $\Bbbk[[\mathbf{x}]]$ and $\Bbbk'[\mathbf{x}]$. In particular the power series (in $\mathbf{x}$) for $e^{\mathbf{z}\cdot\mathbf{x}}$ are identified with the linear functional $\delta_{\mathbf{z}}$ defined by $\delta_{\mathbf{z}}(f) := f(\mathbf{z})$ for every $f \in \Bbbk[\mathbf{x}]$.

Whenever it does not cause confusion, we drop the tilde and treat λ as a functional or as a formal power series interchangeably.

Definition 6.1.7. A subspace $\Lambda \subset \Bbbk[[\mathbf{x}]]$ is called D-invariant if for every $\lambda \in \Lambda$

$$D_j\lambda \in \Lambda \text{ for all } j = 1,\ldots,d.$$

Given a subset $F \subset \Bbbk[[\mathbf{x}]]$ we use $\mathfrak{D}(F)$ to denote the least D-invariant subspace of $\Bbbk[[\mathbf{x}]]$ that contains F. The space $\mathfrak{D}(F)$ is called *the deflation* of F (cf. [7]).

Theorem 6.1.8 ([14, 30] and [26] in its original form). *Let Λ be a finite-dimensional subspace of $\Bbbk[[\mathbf{x}]]$. Then $\ker\Lambda$ is an ideal in $\Bbbk[\mathbf{x}]$ if and only if Λ is D-invariant.*

The Lasker–Noether theorem implies (cf. [30]) that the projector (6.8) is ideal if and only if

$$\Lambda = \operatorname{span}(\lambda_j) = \oplus_{j=1}^{m}(e^{\mathbf{z}_j\cdot\mathbf{x}} \cdot \Lambda_j) \tag{6.13}$$

where $\Lambda_j \subset \Bbbk[\mathbf{x}] \subset \Bbbk[[\mathbf{x}]]$ is a D-invariant subspace of polynomials and

$$\sum_{j=1}^{m} \dim\Lambda_j = \dim G.$$

In particular, if P is a Lagrange projector then $\Lambda = \operatorname{span}\{e^{\mathbf{z}_j\cdot\mathbf{x}}, j = 1,\ldots,n\}$; if P is primary, *i.e.* $\ker P$ is a primary ideal in $\mathbb{C}[\mathbf{x}]$, then $\Lambda = e^{\mathbf{z}_0\cdot\mathbf{x}} \cdot \Lambda_0$ where $\Lambda_0 \subset \Bbbk[\mathbf{x}]$ and $\{\mathbf{z}_0\} = \mathscr{V}(\ker P)$. More on the relationship between ideals, multiplication operators and duality can be found in [7].

Illustration 6.1.9. Among ideal projectors onto $\operatorname{span}\{1,x,y\}$ there are precisely two (up to the change of variables) primary ideal projectors: the Taylor projector T defined by

$$Tx^2 = Txy = Ty^2 = 0 \tag{6.14}$$

and P_* defined by

$$P_*xy = P_*y^2 = 0, \quad P_*x^2 = y. \tag{6.15}$$

This projector appeared in several investigations (cf. [11, 29]) and, as the subscript indicates, it is the star of the show. The space $\Lambda = \operatorname{ran} T^*$ is given by $\operatorname{span}\{1,x,y\}$ while the space Λ for P_* is $\operatorname{span}\left\{1,x,\frac{1}{2}x^2+y\right\}$. The Taylor projector interpolates at the functionals $\delta_0, \delta_0 \circ D_x, \delta_0 \circ D_y$ and P_* interpolates at the functionals $\delta_0, \delta_0 \circ D_x, \delta_0 \circ \left(\frac{1}{2}D_x^2 + D_y\right)$.

6.2 Hermite Projectors and Their Relatives

Ideals are replaced by conventional goals at a certain age.
(Proverb.)

In this section we will examine the possibility of approximating (deforming) a given ideal projector $P \in \mathfrak{P}_G$ by a family of projectors $P(t) \in \mathfrak{P}_G$ that have certain additional properties (*e.g.* Lagrange projectors).

In one variable every ideal is curvilinear since every ideal has finite codimension N and complements $\Bbbk[x]_{<N}$ (cf. [45]). Every ideal projector P in $\mathbb{C}[x]$ is Hermite (cf. [42]) and can be viewed as a limiting case of Lagrange projector. Indeed the ideal $\ker P$ is principle, and a polynomial that generates this ideal can be approximated by square-free polynomials of the same degree. The approximants generate a family of radical ideals that approximate the original ideal $\ker P$. This prompted Carl de Boor [12] to define Hermite projectors as limits of Lagrange projectors and conjecture (cf. [11]) that every (finite-dimensional) ideal projector in $\mathbb{C}[\mathbf{x}]$ is Hermite. Upon suggestion of G. Ellingsrud, the answer was translated from AG in [41]. The answer is “yes” for $d = 2$ and “no” for $d \geq 3$.

6.2.1 Perturbations of ideal projectors

Definition 6.2.1. Let $\mathscr{T}$ be a topological space and $0 \in \mathscr{T}$. Let $P(t) \in \mathfrak{P}_G$ and let P be a linear operator onto G. We say that $P(t) \to P$ as $t \to 0$ if

$$P(t)f(\mathbf{z}) \to Pf(\mathbf{z}) \text{ for every } f \in \Bbbk[\mathbf{x}] \text{ and every } \mathbf{z} \in \Bbbk^d. \tag{6.16}$$

The attributes of ideal projectors depend continuously on the projectors:

Theorem 6.2.2 ([46, 44]). *Let G be a finite dimensional space, $P(t) \in \mathfrak{P}_G$ and let P be a linear operator onto G such that $P(t) \to P$ as $t \to 0$. Then P is an ideal projector. Moreover the following are equivalent*

(i) $P(t) \to P$ *as* $t \to 0$.
(ii) $\mathbf{w}_{P(t)} \to \mathbf{w}_P$. *Here $w_{P(t)}, w_P \in \mathscr{P}_G$ are points that parametrize the projectors $P(t)$ and P respectively. Hence $\mathscr{P}_G$ is a continuous parametrization of $\mathfrak{P}_G$.*
(iii) $\mathbf{M}_{P(t)} \to \mathbf{M}_P$ *where $\mathbf{M}_{P(t)}$, $\mathbf{M}_P$ are multiplication operators for $P(t)$ and P respectively.*
(iv) *For every $\lambda \in (\ker P)^\perp$ there exists $\lambda(t) \in (\ker P(t))^\perp$ such that*

$$\lambda(t)(f) \to \lambda(f),$$

for every $f \in \Bbbk[\mathbf{x}]$.

The proofs are relatively straightforward. Equivalence of (i) and (ii) follows from the de Boor's formula (cf. [49]) and so does the equivalence of (i) and (iii), (cf. [15]). The equivalence of (i) and (iv) is proved in [46, 44].

In particular if the $P(t)$ are Lagrange projectors at the points $\{\mathbf{z}_j(t), j = 1,\ldots, \dim G\}$ then for every $\lambda \in (\ker P)^{\perp}$ there exist coefficients $c_j(t) \in \Bbbk$ such that

$$\sum_{j=1}^{\dim G} c_j(t)e^{\mathbf{z}_j(t)\cdot\mathbf{x}} \to \lambda \text{ as } t \to 0. \tag{6.17}$$

Illustration 6.2.3. Let $(x_j, j = 1,2,3)$ be three distinct points in $\Bbbk$ and let P be a Lagrange projector onto $\Bbbk[x,y]_{\leq 1} \subset \Bbbk[x,y]$ interpolating at the sites $\left(x_j, x_j^2\right)$. It is easy to compute that

$$\begin{aligned}
Px^2 &= y,\\
Pxy &= x_1x_2x_3 - (x_1x_2 + x_1x_3 + x_2x_3)x + (x_1 + x_2 + x_3)y,\\
Py^2 &= x_1x_2x_3(x_1 + x_2 + x_3) + (x_2 + x_3)(x_1 + x_3)(x_1 + x_2)x\\
&\quad + \left(x_1^2 + x_2^2 + x_3^2 + x_1x_2 + x_1x_3 + x_2x_3\right)y.
\end{aligned}$$

Thus if for every t, $(x_j(t), j = 1,2,3)$ is a triple of distinct points in $\Bbbk$ and $x_j(t) \to 0$ as $t \to 0$ then the Lagrange projectors $P(t)$ interpolating at the sites $\left(x_j(t), x_j^2(t)\right)$ converge to P_*. Theorem 6.2.2(iv) implies the existence of coefficients $(c_j(t), j = 1,2,3)$ such that

$$\sum c_j(t)e^{x_j(t)x + x_j^2(t)y} \to \frac{1}{2}x^2 + y \text{ as } t \to 0.$$

This is so because the functional $\lambda := \frac{1}{2}x^2 + y \in (\ker P_*)^{\perp}$.

6.2.2 Lagrange and curvilinear projectors

Definition 6.2.4. An ideal projector P onto G is called *curvilinear* if there exists a linear form $l \in \Bbbk[\mathbf{x}]$ such that $\ker P$ complements the space $\operatorname{span}\{1, l, \ldots, l^{N-1}\}$, *i.e.* $[1], [l], \ldots, [l^{N-1}]$ form a basis in the algebra $\Bbbk[\mathbf{x}]/\ker P$. The subset of $\mathfrak{P}_G$ of all curvilinear projectors is denoted by $\mathfrak{C}_G$ and the subset of $\mathscr{P}_G$ that corresponds to these projectors is denoted by $\mathscr{C}_G$.

The property of being curvilinear can be expressed in terms of the attributes of the projector as is shown below.

Proposition 6.2.5.

(i) *$P \in \mathfrak{P}_G$ is curvilinear if and only if there exists a linear form $l \in \Bbbk[\mathbf{x}]_{\le 1}$ such that $l(\mathbf{M}_P)$ is non-derogatory. In this case $\tilde{l}(\mathbf{M}_P)$ is non-derogatory for a generic form $\tilde{l}$.*

(ii) *$\mathscr{C}_G$ is a Zariski open subset of $\mathscr{P}_G$.*

(iii) *There is a surjective rational map $\rho : \Bbbk^{dN} \to \mathscr{C}_G$.*

Proof. (i) Without loss of generality, assume that $\ker P$ complements

$$H := \operatorname{span}\left\{1, x_1, \ldots, x_1^{N-1}\right\}.$$

Then $\left\{x_1^N, x_2, \ldots, x_d\right\} = \partial H$ and the ideal projector Q onto H with $\ker Q = \ker P$ is given by

$$Qx_1^N = q_1(x_1),\ Qx_k = q_k(x_1),\ p_1, p_k \in \Bbbk[\mathbf{x}]_{<N}, \quad k = 2, \ldots, d. \tag{6.18}$$

Let $\mathbf{M}_Q = (M_1(Q), \ldots, M_d(Q))$ be multiplication operators for Q. We need to show that $M_1(Q)$ is non-derogatory, for then, by Theorem 6.1.5(iii), so is M_1 in $\mathbf{M}_P$. From (6.18) we have

$$M_k(Q) = q_k(M_1(Q)), \quad k = 2, \ldots, d. \tag{6.19}$$

By cyclicity, the algebra of operators that commute with $\mathbf{M}_Q$ is generated by $\mathbf{M}_Q$, and hence by $M_1(Q)$. Therefore every operator that commutes with $M_1(Q)$ is a polynomial in $M_1(Q)$ and $M_1(Q)$ is non-degenerate. Conversely if M_1 is non-degenerate then (6.19) holds for some polynomials $q_k \in \Bbbk[\mathbf{x}]_{<N}$ and the ideal $\ker P = J_{\mathbf{L}}$ is generated by polynomials in x_1.

(ii) For every $(l, g) \in \Bbbk[\mathbf{x}]_{\le 1} \times G$ define

$$\mathscr{U}_{l,g} := \left\{\mathbf{w}_P \in \mathscr{P}_G : M_l \text{ is non-derogatory, } g \text{ is cyclic for } M_l\right\}. \tag{6.20}$$

Then $\mathscr{U}_{l,g}$ consist of those $\mathbf{w}_P \in \mathscr{P}_G$ for which the determinant of the matrix of column vectors $\left(g, M_l g, \ldots, M_l^{N-1} g\right)$ is non-zero. But this determinant is a polynomial in $\Bbbk[\mathscr{P}_G]$, hence $\mathscr{U}_{l,g}$ is Zariski open and so is $\mathscr{C}_G = \cup \mathscr{U}_{l,g}$, where the union is over $(l, g) \in \Bbbk[\mathbf{x}]_{\le 1} \times G$.

(iii) Suppose that the $\ker P$ complements $H := \operatorname{span}\left\{1, x_1, \ldots, x_1^{N-1}\right\}$ and let Q be an ideal projector onto H with $\ker Q = \ker P$ given by (6.18). For every $f \in \partial G$ we have $f - Pf \in \ker P = \ker Q$. Thus $Qf = QPf$. Writing these equations in terms of the coefficients we have

$$Qf = \sum_{j=0}^{N-1} a_{j,f} x_1^j = Q\left(\sum_{k=1}^{N} w_{f,k} g_k\right) = \sum_{k=1}^{N} w_{f,k} \sum_{j=0}^{N-1} a_{j,g_k} x_1^j. \tag{6.21}$$

This is a linear system of equations for coefficients $w_{f,k}$, and by Cramer's rule

$$w_{f,k} = \frac{p_{f,k}}{p_k}, \tag{6.22}$$

where $p_{f,k}$ and p_k are determinants of the matrices with entries from $\Bbbk[\mathscr{P}_H]$, and therefore, themselves are polynomials in $\Bbbk[\mathscr{P}_H]$. Since $\mathscr{P}_H \simeq \Bbbk^{dN}$ we obtain the proof of the theorem. □

Definition 6.2.6. An ideal projector P onto G is called *Lagrange* if the ideal $\ker P$ is an ideal of $N = \dim G$ points in $\Bbbk^d$, *i.e.* $(\ker P)^{\perp} = \operatorname{span}\{e^{\mathbf{z}_j \cdot \mathbf{x}}, j = 1, \ldots, N\}$. The subset of $\mathfrak{P}_G$ of all Lagrange projectors is denoted by $\mathfrak{L}_G$ and the subset of $\mathscr{P}_G$ that corresponds to the Lagrange projectors is denoted by $\mathscr{L}_G$.

Once again, this property of a projector is reflected in its attributes.

Proposition 6.2.7. *Let $\Bbbk = \mathbb{C}$ be the complex field. Then*

(i) *$P \in \mathfrak{P}_G$ is Lagrange if and only if the multiplication operators $\mathbf{M}_P = \big(M_1, \ldots, M_d\big)$ are simultaneously diagonalizable. In this case the joint spectrum $\sigma(\mathbf{M}_P) = \mathscr{V}(\ker P)$.*
(ii) *$\mathfrak{L}_G$ is a Zariski open subset of $\mathscr{P}_G$.*
(iii) *There exists a surjective rational map $\rho : \mathbb{C}^{dN} \to \mathfrak{C}_G$.*

Part (i) of the proposition is the celebrated theorem of Hans Stetter (cf. [1, 50, 51]). The rest is analogous to the proof of the Proposition 6.2.5 (cf. [48] for details).

Illustration 6.2.8. The ideal projector onto $\operatorname{span}\{1, x, y\}$ is Lagrange if and only if there exists three points (x_1, y_1), (x_2, y_2) and (x_3, y_3) in $\Bbbk^2$ such that

$$\begin{aligned}
\big(Px^2\big)(x_i, y_i) &= a_0 + b_0 x_i + c_0 y_i = x_i^2, \\
(Pxy)(x_i, y_i) &= a_1 + b_1 x_i + c_1 y_i = x_i y_i, \\
\big(Py^2\big)(x_i, y_i) &= a_2 + b_2 x_i + c_2 y_i = y_i^2,
\end{aligned}$$

for $i = 1, 2, 3$. By Cramer's rule

$$a_0 = a_0(x_i, y_i) = \frac{\begin{vmatrix} x_1^2 & x_1 & y_1 \\ x_2^2 & x_2 & y_2 \\ x_3^2 & x_3 & y_3 \end{vmatrix}}{\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}} \in \Bbbk(x_1, y_1, x_2, y_2, x_3, y_3).$$

The rest of the coefficients have similar expressions as the rational functions defined on the subset of $\Bbbk^6$ where the determinant in the denominator does not vanish, *i.e.* the interpolation sites are not collinear.

6.2.3 Limits of Lagrange and curvilinear projectors

Definition 6.2.9. An ideal projector P onto G is a *limit of curvilinear projectors* (LCP) if there exists a family of curvilinear projectors $P(t)$ onto G such that $P(t) \to P$. The subset of $\mathfrak{P}_G$ of all (LCP) projectors is denoted by $\mathfrak{LC}_G$, and the subset of $\mathscr{P}_G$ that corresponds to such projectors is denoted by $\mathscr{LC}_G$.

An ideal projector P onto G is a Hermite projector if there exists a sequence of Lagrange projectors $P(t)$ onto G such that $P(t) \to P$. The subset of $\mathfrak{P}_G$ of all Hermite projectors is denoted by $\mathfrak{H}_G$, and the subset of $\mathscr{P}_G$ that corresponds to Hermite projectors is denoted by $\mathscr{H}_G$.

Theorem 6.2.10. *Let G be a finite-dimensional subspace of $\Bbbk[x]$. Then:*

(i) *The subset $\mathscr{LC}_G \subset \mathscr{P}_G$ that parametrizes the LCP projectors is an irreducible subvariety of $\mathscr{P}_G$ of dimension $d \times \dim G$. Moreover, for every G there exists a finite step algorithm that explicitly determines the polynomials that cut out $\mathscr{LC}_G$ in $\mathscr{P}_G$.*

(ii) *$\mathscr{LC}_G = \mathscr{P}_G$ if and only if $\mathscr{P}_G$ is irreducible.*

(iii) *If $\Bbbk = \mathbb{C}$ then the same is true for the variety $\mathscr{H}_G$.*

Proof. By Proposition 6.2.5(iii), Zariski closure $\bar{\mathscr{C}}_G$ of $\mathscr{C}_G$ has rational parametrization, and thus it is irreducible. By Proposition 6.2.5(ii), $\mathscr{C}_G$ is a Zariski open subset of $\bar{\mathscr{C}}_G$, hence (cf. [32]) its Zariski closure coincides with the Euclidean closure. By Theorem 6.2.2(ii) the Euclidean closure of $\mathscr{C}_G$ is $\mathscr{LC}_G$ which proves the first part of (i). The second part of the statement follows from the existence of "implicitization algorithm for rational parametrization" (cf. [8, 130–131]). To prove (ii), observe that if $\mathscr{P}_G$ is irreducible then $\mathscr{C}_G$ is a Zariski open subset of an irreducible variety $\mathscr{P}_G$ and hence its closure is $\bar{\mathscr{C}}_G = \mathscr{LC}_G = \mathscr{P}_G$. Part (iii) follows from Proposition 6.2.7. The proof is similar to (i) and (ii). □

The following proposition immediately follows from (6.7).

Proposition 6.2.11 ([15]). *Let $\mathbf{L}$ be a cyclic commuting sequence of operators on G and let $\mathbf{L}(t)$ be a sequence of commuting d-tuples of operators on G such that $\mathbf{L}(t) \to \mathbf{L}$. Then, for sufficiently small t, $\mathbf{L}(t)$ is cyclic and $P_{\mathbf{L}(t)} \to P_{\mathbf{L}}$.*

Proof. Let g_0 be a cyclic vector for $\mathbf{L}$ and let $p_k \in \Bbbk[\mathbf{x}]$ be such that

$$\{g_0, p_1(\mathbf{L})g_0, \ldots, p_{N-1}(\mathbf{L})g_0\}$$

is a $\Bbbk$-basis for G. Then for sufficiently small t we have

$$\sum \|p_k(\mathbf{L})g_0 - p_k(\mathbf{L}(t))g_0\| < 1,$$

hence (cf. [28]) $\{g_0, p_1(\mathbf{L}(t))g_0, \ldots, p_{N-1}(\mathbf{L}(t))g_0\}$ is a $\Bbbk$-basis for G. The rest follows from (6.7). □

Theorem 6.2.12 ([15]). *Let P be an ideal projector onto G and let $\mathbf{M}$ be its sequence of multiplication operators. Then:*

(i) *$P \in \mathfrak{LC}_G$ if and only if there exists a sequence $\mathbf{L}(t) = (L_j(t))$ of commuting d-tuples of operators such that for any t, some linear form $l_t(\mathbf{L}(t))$ is nonderogatory and $\mathbf{L}(t) \to \mathbf{M}$.*

(ii) *If* $P \in \mathfrak{LC}_G$ *and* $\tilde{l}_t$ *is any linear form, then there exists a sequence* $\tilde{\mathbf{L}}(t) = (\tilde{L}_j(t))$ *of commuting* d*-tuples of operators such that* $\tilde{l}_t(\tilde{\mathbf{L}}(t))$ *is non-derogatory and* $\tilde{\mathbf{L}}(t) \to \mathbf{M}$.

(iii) $P \in \mathfrak{H}_G$ *if and only if there exists a sequence* $\mathbf{L}(t) = (L_j(t))$ *of simultaneously diagonalizable* d*-tuples of operators such that* $\mathbf{L}(t) \to \mathbf{M}$.

(iv) $\mathscr{H}_G \subset \mathscr{L}\mathscr{C}_G$ *and, if* $\Bbbk = \mathbb{C}$*, then* $\mathscr{H}_G = \mathscr{L}\mathscr{C}_G$.

Proof. Item (i) follows from Propositions 6.2.5 and 6.2.11. To prove (ii), assume that l_t and $\mathbf{L}(t) = (L_j(t))$ satisfy (i) and let

$$\tilde{l}_t(\mathbf{x}) = \sum_{k=1}^{d} \tilde{a}_k(t) x_k.$$

The operator $s\tilde{l}_t(\mathbf{L}(t)) + l_t(\mathbf{L}(t))$ is non-derogatory for $s = 0$, hence it is non-derogatory for all but finitely many s. Thus $\tilde{l}_t(\mathbf{L}(t)) + \varepsilon l_t(\mathbf{L}(t))$ is non-derogatory for $\varepsilon = 1/s$ and hence it is non-derogatory for all but finitely many ε. Choosing $\tilde{L}_k(t) := L_j(t) + (\varepsilon(t)/(d\tilde{a}_k(t)) l_t(\mathbf{L}(t))$ for $a_k \neq 0$ we have $\tilde{\mathbf{L}}(t) = \left(\tilde{L}_j(t)\right) \to \mathbf{M}$ if $\varepsilon(t) \to 0$ sufficiently fast and $\tilde{l}_t(\tilde{\mathbf{L}}(t)) = \tilde{l}_t(\mathbf{L}(t)) + \varepsilon(t) l_t(\mathbf{L}(t))$ is non-derogatory. Item (iii) follows from Proposition 6.2.7. To prove (iv) assume that $P \in \mathscr{H}_G$. Then there exist simultaneously diagonalizable $\mathbf{L}(t) \to \mathbf{M}$. By small perturbations of the eigenvalues we can assure that the spectrum of, say, $L_1(t)$ has distinct eigenvalues, hence $L_1(t)$ it is non-derogatory. Conversely, if $\Bbbk = \mathbb{C}$ and $L_1(t)$ is non-derogatory, then there exist polynomials $p_{j,t} \in \mathbb{C}[x]$ of one variable such that $L_j(t) = p_{j,t}(L_1(t))$, $j = 2, \ldots, d$. Let $\tilde{L}_1(t)$ be a diagonalizable perturbation of $L_1(t)$ (it is here that the completeness of the field is used). Then

$$\tilde{\mathbf{L}}(t) := \left(\tilde{L}_1(t), p_{2,t}\left(\tilde{L}_1(t)\right), \ldots, p_{d,t}\left(\tilde{L}_1(t)\right)\right) \to \mathbf{M}$$

which ends the proof. □

Theorem 6.2.13 ([41]). *For any finite-dimensional subspace* $G \subset \Bbbk[x,y]$ *we have* $\mathfrak{LC}_G = \mathfrak{P}_G$*. If* $\Bbbk = \mathbb{C}$ *then* $\mathfrak{H}_G = \mathfrak{P}_G$.

Proof ([15]). Let P be an ideal projector onto G and let $\mathbf{M} = (M_1, M_2)$ be its sequence of multiplication operators. We only need to approximate two commuting matrices M_1, M_2 by commuting matrices $L_1(t), L_2(t)$ in such a way that $L_2(t)$ is non-derogatory. And this is possible by [31]. Here is a short proof: let $M_1 = SJS^{-1}$ where $J = \mathrm{Diag}\,(\mu_k N_k)$ be the Jordan form of M_1. Pick a matrix $A = S\tilde{J}S^{-1}$ with $\tilde{J} = \mathrm{Diag}\,(\nu_k N_k)$ where ν_k are chosen to be distinct and such that conjugate pairs of μ_k correspond to conjugate pairs of ν_k. Then $sM_2 + A$ commutes with M_1, is non-derogatory for $s = 0$, and therefore for all but finitely many values of s. Thus $M_2 + tA$ is non-derogatory for all but finitely many values of t and letting $t \to 0$ we obtain the result. □

Remark 6.2.14. This could be a good time for the exclamation “Isn’t it just the Fogarty theorem [16]?” Indeed it is!

6.2.4 Two problems

The proof of the preceding theorem gives (at least conceptually) an algorithm of generating a family of sets $\mathscr{X}(t) \subset \mathbb{C}^2$ of $N = \dim G$ points such that the Lagrange projectors $P(t)$ that interpolate at the sites $\mathscr{X}(t)$ converge to a given ideal projector $P \in \mathfrak{H}_G$. Starting with P, we construct $A(t) := M_2 + tA$ and find a polynomial p_t such that $M_1 = p_t(A(t))$. For every t, let $L_2(t)$ be a diagonalizable matrix that approximates $A(t)$ and let $\{y_j, j = 1, \dots, \dim G\}$ be its eigenvalues. Then the Lagrange projectors $P(t)$ that interpolate at the points

$$\mathscr{X}(t) := \{\mathbf{z} = \{(p_t(y_j), y_j) \in \mathbb{C}^2, j = 1, \dots, \dim G\}$$

converge to P.

Problem 1. Is there an "algorithm" in more than two variables?

The converse is even more intriguing and (in my opinion) is the key for deeper understanding of the geometry (or rather "geology") of the Hilbert schemes of points, and therefore warrants a short discussion.

Problem 2. Given a Hermite projector $P \in \mathfrak{H}_G$ and a family of Lagrange projectors $P(t) \to P$, what is the relationship between "trajectories" of $\mathscr{X}(t) = \mathscr{V}(\ker P(t))$ and P?

Illustration 6.2.15. Recall the two primary ideal projectors T and P_* onto $G = \mathbb{C}[x,y]_{\leq 1}$ given by (6.14) and (6.15). As we saw previously, if all the points in $\mathscr{X}(t)$ tend to the origin along the curve $(t, t^2) \subset \mathbb{C}^2$ then the corresponding Lagrange projectors $P(t)$ converge to P_*. Computational evidence suggests that even a slight modification of this trajectory leads to a family of Lagrange projectors that converges to the Taylor projector T. In fact a "random" choice $\mathscr{X}(t) \to 0$ leads to a family of Lagrange projectors $P(t)$ that converge to T. It seems that the Taylor projector is much more central in $\mathfrak{P}_G$ than P_*. In other words, neighborhoods of $\mathbf{w}_T$ in $\mathscr{P}_G$ have larger density than neighborhoods of $\mathbf{w}_{P_*}$. This suggests that, even in this simple case, the "geology" of $\mathscr{P}_G$ is not well-understood.

Here is what little I know. If we assume that $P(t)$ interpolates at three points $(x_k(t), y_k(t))$, $k = 0, 1, 2$ and $P(t) \to P_*$ then, by translation invariance, the same is true for Lagrange projectors at the points

$$(0,0), (\tilde{x}_k(t), \tilde{y}_k(t)) := (x_k(t) - x_0(t), y_k(t) - y_0(t)), \quad k = 1, 2.$$

Now we have

$$(\tilde{x}_k(t))^2 - \tilde{y}_k(t) = a(t)\tilde{x}_k(t) + b(t)\tilde{y}_k(t) \tag{6.23}$$

with $a(t), b(t) \to 0$.

Claim. $|\tilde{y}_k(t)| = o(|\tilde{x}_k(t)|)$.

Proof. If not, then there exists a constant $C > 0$ such that $|\tilde{y}_1(t)| \geq C(|\tilde{x}_1(t)|)$. Dividing both sides in (6.23) by $|\tilde{y}_1(t)|$, the right-hand side of the resulting equality tends to zero, while the left-hand side tends to 1. □

If $P(t) \to P$ and $P(t)$ interpolates functionals $\delta_0, \delta_0 \circ D_x$ and $\delta_{(x(t),y(t))}$ then

$$x^2(t) - y(t) = b(t)y(t)$$

and thus $\frac{|x(t)|^2}{|y(t)|} \to 1$. In other words, the trajectory of the point $(x(t), y(t))$ resembles the curve $x^2 - y$ that defines the ideal $\ker P_*$.

In general, if $P(t) \to P$, then $\mathscr{V}(\ker P(t)) \to \mathscr{V}(\ker P)$. Intuitively, the trajectories of the points in $\mathscr{V}(\ker P(t))$ should resemble the curves on the hypersurfaces defined by the polynomials in the basis for $\ker P$.

Problem 3. Define "resemble".

6.2.5 Existence of non-Hermite projectors

As was mentioned before, for $d > 2$ there exist ideal projectors that are not Hermite. In the language of Hilbert schemes its existence was first proved by Iarrobino [22]. Here is a variation of his construction for non-Hermite projectors onto $G = \mathbb{C}[x]_{\leq N}$.

Theorem 6.2.16. *Let $d \geq 3$. Then for sufficiently large N there exists a non-LCP projector onto $\Bbbk[x]_{\leq N}$.*

Proof. The idea is to pick a set U of monomials of degree N and a set W of monomials of degree $N+1$ in such a way that $D_j w \in \operatorname{span} U$ for every $w \in W$. Once such sets are chosen, we let V be the set of degree N monomials not in U and construct a family of ideals

$$\left\{J(a(u,w)) : v \in V, w \in W, a(u,w) \in \Bbbk^{\#W \times \#V}\right\},$$

where $J(a(u,w))$ is defined as deflation:

$$J(a(u,w)) := \mathfrak{D}\{w - \sum_{v \in V} a(u,w)v : w \in W\}.$$

For any choices of $a(u,w) \in \Bbbk$, the ideal $J(a(u,w))$ is primary. If the matrix $(a(u,w) : (u,w) \in U \times W)$ is of rank $\#V$ then $J(a(u,w))$ complements $\Bbbk[\mathbf{x}]_{\leq N}$. All that's left is to count the number of free parameters $a(u,w)$, which is equal to

$$\#W \times \#V + d = \#W \times \left(\binom{N+d-1}{d-1} - \#U \right) + d;$$

the term d corresponds to the various choices of maximal ideal.

If $\#U = \#W = \left\lfloor \frac{1}{2}\binom{N+d-1}{d-1} \right\rfloor = \left\lfloor \frac{1}{2}\#M[\mathbf{x}] \right\rfloor$ then the number of parameters is $\left\lfloor \frac{1}{2}\#M[\mathbf{x}]_{\le N} \right\rfloor^2 + d$. If this number is greater then $d \dim \Bbbk[x]_{\le N} = d\binom{N+d}{d}$ we are guaranteed that at least one of the projectors generated by $F(a(u,w))$ is not an LCP. It remains to show that for every K not greater than the number of monomials of degree N, there exist subsets U and W with the desired properties. This is where some caution is needed:

We impose an order $\prec$ on the monomials of fixed degree as follows. We say that $u \prec v$ if u has fewer prime divisors than v. Otherwise the monomials are ordered lexicographically. We start with a monomial $x_1^N \in U$ and $x_1^{N+1} \in W$ and add simultaneously monomials to U and W in the increasing order. It is now obvious that U and W have the desired property. □

Simple calculations show the existence of non-LCP projectors onto the space $\mathbb{C}[x]_{\le 9}$ for $d = 3$. The dimension of this space is 220, which is much larger then the dimension $(= 102)$ of the range of the non-Hermite ideal projector obtained by Iarrobino. On the other hand, we have a projector onto the space $\mathbb{C}[\mathbf{x}]_{\le N}$, which is of primary interest. Observe that the range of the projector matters. Indeed, for $G = \mathbb{C}[x_1]_{\le N} \subset \mathbb{C}[\mathbf{x}]$, every ideal projector is curvilinear, hence $\mathfrak{P}_G = \mathfrak{H}_G$. There are other advantages to this construction:

Theorem 6.2.17 ([25]). *For $d = 13$ there exists a non-LCP ideal projector from $\Bbbk[\mathbf{x}]$ onto $\Bbbk[\mathbf{x}]_{\le 1}$.*

Proof. Pick $W = \{x_j x_k, k, j = 1, \dots, 7\}$ and $U = \{x_k, k, j = 1, \dots, 7\}$. In this case $\#W = 36$ and $\#V = 5$ and hence

$$\#W \times \#V + d = 193 > 13 \times 14 = d \times \dim \Bbbk[\mathbf{x}]_{\le 1}$$

which proves the result. □

6.2.6 Description of non-Hermite projectors

Now that we know that not all of the ideal projectors are Hermite, one wants to characterize those that are.

Until a few months ago, I was aware of only one explicitly given non-Hermite projector. It was provided by Emsalem and Iarrobino [23]. The projector P is from $\Bbbk[x,y,z,w]$ onto the 8-dimensional subspace G spanned by $\{1,x,y,z,w,xz,yz,yw\}$ defined by $Pxw = yz$ and $Pu = 0$ for the rest of monomials not in G. The proof is by hard computation of the dimension of Zariski tangent space to the variety $\mathscr{P}_G$ at the point P. This dimension turned out to be 28; less than $d \times \dim G = 32$. Thus P does not belong to the irreducible component $\mathscr{H}_G \subset \mathscr{P}_G$ whose dimension is 32 at every point.

Recently Dustin Cartwright, Daniel Erman, Mauricio Velasco and Bianca Viray extended this result.

Theorem 6.2.18 ([5]). *Let P be an ideal projector onto an 8-dimensional subspace of $\mathbb{C}[x,y,z,w]$. Then P is non-Hermite if and only if $(\ker P)^{\perp}$ is a deflation of three homogeneous quadratic polynomials $\lambda_1,\lambda_2,\lambda_3$ such that the three four-by-four symmetric matrices $\{A_j\}$ defined by*

$$(1,x,y,z,w)A_j(1,x,y,z,w)^t = \lambda_j, \quad j=1,2,3$$

have a non-zero "Turnbull determinant":

$$\det\begin{bmatrix} 0 & A_1 & -A_2 \\ -A_1 & 0 & A_3 \\ A_2 & -A_3 & 0 \end{bmatrix} \neq 0.$$

In particular, this implies that every projector from $\mathbb{C}[x,y,z,w]$ onto a subspace of dimension 7 or smaller is Hermite.

6.2.7 Projectors in three variables

The methods involved in the results of the previous subsections just do not seem to work in three variables. Theorem 6.2.18 implies that any projector on an eight-dimensional subspace of $\mathbb{C}[x,y,z]$ is Hermite.

Theorem 6.2.19. *There exists a non-Hermite projector onto $\mathbb{C}[x,y,z]_{\le 7}$.*

Proof. We will use a variation on the construction in the proof of Theorem 6.2.16. Let

$$U=\left\{x^ky^{7-k}, k=0,\dots,7\right\}\cup\left\{x^kz^{7-k}, k=0,\dots,6\right\}\cup\left\{y^6z, y^5z^2\right\}$$

and

$$W=\left\{x^ky^{8-k}, k=0,\dots,8\right\}\cup\left\{x^kz^{8-k}, k=0,\dots,7\right\}\cup\left\{y^7z, y^6z^2\right\}.$$

Then $\#V = 36-\#U = 19$, $\#W = 19$ and $D_j(W)\subset \operatorname{span} U$ for $j=1,2,3$. Finally $\#W\times\#V = 361 > 360 = d\times\dim\mathbb{C}[x,y,z]_{\le 7}$. □

There is a belief expressed by Sturmfels [52], and shared by this author, that there exists a non-Hermite projector onto a subspace of $\mathbb{C}[x,y,z]$ of small dimension. I will actually go a step further in the following conjecture.

Conjecture 6.1. The ideal projector from $\mathbb{C}[x,y,z]$ onto $\mathbb{C}[x,y,z]_{\le 2}$ defined by

$$Px^3 = yz, Py^3 = xz, Pz^3 = xy, Pu = 0 \tag{6.24}$$

for the rest of monomials u of degree 3, is non-Hermite.

The reason for the conjecture goes back to the discussion in Subsection 6.2.4. That is, it is not possible for sets of points $\mathscr{X}(t)=\left\{\mathbf{z}_j(t), j=1,\dots,10\right\}$ that tend to zero to "resemble" curves on the three surfaces

$$x^3 - yz, y^3 - xz, z^3 - xy$$

at the same time. Here is another look at a potential argument.

If P is Hermite, then there exists $\mathscr{Z}(t) = \left\{\mathbf{z}_j(t), j = 1, \ldots, 10\right\}$ such that $\mathbf{z}_j(t) \to 0$ and

$$\operatorname{span}\left\{e^{\mathbf{z}_j(t)\cdot x}\right\} \to (\ker P)^\perp.$$

Since the functionals $1, x, y, z, x^2, y^2, z^2 \in (\ker P)^\perp$, it would stand to reason that there exist ideal projectors $Q(t)$ interpolating at three points, say,

$$\{\mathbf{z}_1(t), \mathbf{z}_2(t), \mathbf{z}_3(t)\}$$

and seven functionals:

$$\left\{1, x, y, z, x^2, y^2, z^2\right\}$$

such that $Q(t) \to P$. But this is not possible. Indeed, assume, without loss of generality, that for infinitely many values of t:

$$|x_1(t)| \le |y_1(t)| \le |z_1(t)|,$$

where $\mathbf{z}_1(t) = (x_1(t), y_1(t), z_1(t))$. We have

$$x_1^3(t) - y_1(t)z_1(t) = a(t)x_1(t)y_1(t) + b(t)x_1(t)z_1(t) + c(t)y_1(t)z_1(t), \qquad (6.25)$$

where $a(t), b(t), c(t) \to 0$, since $Q(t) \to P$. Dividing both parts by $|y_1(t)|\,|z_1(t)|$ and since $\mathbf{z}_1(t) \to 0$, the left-hand side of (6.25) tends to 1 while the right-hand side tends to 0.

Remark 6.2.20. As Theorem 6.2.12 would suggest, there is an interesting parallels between a search for small reducible Hilbert Schemes, and a small reducible varieties of commuting matrices (cf. [15, 19, 20]).

Theorem 6.2.10 provides a potential for characterization of all Hermite projectors. After all, it gives an algorithm for determining a finite set of polynomials $f_1, \ldots, f_s$ such that $P \in \mathfrak{P}_G$ is Hermite if and only if $f_j(\mathbf{w}_P) = 0$ for all $j = 1, \ldots, s$. Unfortunately, even in the simplest of cases, such as $G = \mathbb{k}[x,y,z]_{\le 2} \subset \mathbb{k}[x,y,z]$, the "finitely many steps" are still too many. For now all that is left is to paraphrase Abraham in the faith that "God will provide the RAM" *Old Testament.*

6.3 Nested Ideal Interpolation

If your only tool is a hammer, all your problems start to look like nails.
(Proverb.)

In one variable, Lagrange projectors onto $\mathbb{k}[x]_{<N}$ can be written in Newton form. Let the set of interpolation sites for P be $\{z_1, \ldots, z_N\}$. We choose the basis

$$g_1(x) = 1, g_k(x) = \prod_{j=1}^{k-1} (x - z_j), \quad k = 1, \dots, N$$

for $\Bbbk[x]_{<N}$. Notice that $\operatorname{span}\{g_k(x), k = 1, \dots, m\}$ is a basis for $\Bbbk[x]_{<m}$. Now we write

$$P = \sum_{k=1}^{N} \lambda_k \otimes g_k,$$

where the $\lambda_k \in (\Bbbk[\mathbf{x}])'$ are bi-orthogonal to g_k and the partial sums: $\sum_{k=1}^{m} \lambda_k \otimes g_k$ define Lagrange projectors onto $\Bbbk[x]_{<m}$. This Newton form of Lagrange projectors is advantageous for numerical calculations since, after computing the interpolants at $N-1$ points, we only have to compute one more term to get the full projector. The functionals λ_k are known as divided differences. These functionals depend continuously on $\{z_1, \dots, z_k\}$ and, if these points coalesce, the projectors converge to an ideal (Hermite) projector.

Divided differences had long been one of the basic tools in numerical analysis. What happens in several variables?

6.3.1 Ideal restrictions

Let $G_0 \subset G$ be subspaces of $\Bbbk[\mathbf{x}]$ and let P and P_0 be ideal projectors onto G and G_0 respectably. We say that P_0 is an ideal restriction of P to G_0 if $\ker P \subset \ker P_0$.

Theorem 6.3.1. *Let $G_0 \subset G$ be finite-dimensional subspaces of $\Bbbk[\mathbf{x}]$ and let P be a Lagrange projector onto G. Then there exists a Lagrange projector P_0 onto G_0 such that P_0 is an ideal restriction of P onto G_0.*

Proof. It is sufficient to prove this theorem for the case when $\dim(G/G_0) = 1$. Let $N := \dim G$ and choose a $\Bbbk$-basis $(g_1, \dots, g_N)$ for G such that $(g_1, \dots, g_{N-1})$ is a basis for G_0. Let $\mathscr{V}(\ker P) = \{\mathbf{z}_1, \dots, \mathbf{z}_N\} \subset \Bbbk^d$. Then the $N \times N$ determinant

$$\Delta := \det(g_j(\mathbf{z}_k)) \neq 0. \tag{6.26}$$

Expanding it with respect to the last row we have

$$0 \neq \sum_{k=1}^{N} g_N(\mathbf{z}_k) \Delta_{N,k}, \tag{6.27}$$

where the $\Delta_{N,k}$ are the appropriate minors of Δ. It follows that at least one of these minors is non-zero, say Δ_{N,k_0}, hence the set $\mathscr{Z}_0 := \mathscr{V}(\ker P) \setminus \{\mathbf{z}_{k_0}\}$ defines a radical ideal that contains $\ker P$ and complements G_0. □

An algorithm for choosing $\mathscr{Z}_0$ for particular choices of G and G_0 can be found in [37, 38].

In algebraic terms, if $\oplus_{k=1}^{m} A_k = A = \Bbbk[\mathbf{x}]/J$ is a grading of $\Bbbk[\mathbf{x}]/J$ and J is a radical ideal, then there is a grading of the dualizing module $\hat{A} = \oplus_{k=1}^{m} \Lambda_k$ such that for every $j \leq k$, the space $\oplus_{k=1}^{j} \Lambda_k$ is a dualizing module for $\oplus_{k=1}^{j} A_k$, *i.e.* $\hat{A}$ can be graded consistently with the grading of A.

In this section I will present some results from [49] regarding the possibility of extending Theorem 6.3.1 to general ideal projectors.

Right off the bat, let us observe that the general ideal version of Theorem 6.3.1 has no chance.

Illustration 6.3.2. The projector P_* cannot be restricted to an ideal projector onto $\text{span}\{1, y\}$. For, if this was the case, there would exist a 2-dimensional D-invariant subspace of $\text{span}\left\{1, x, \frac{1}{2}x^2 + y\right\}$ which is correct for $\text{span}\{1, y\}$. Such a subspace by necessity must contain a polynomial with the term y in it, and therefore, a polynomial with the term $\frac{1}{2}x^2 + y$. This is not possible since a D-invariant subspace that contains a polynomial of degree 2 is at least three-dimensional.

6.3.2 A conjecture of Tomas Sauer

Tomas Sauer conjectured (cf. [36]) that every ideal projector onto $\Bbbk\mathbf{x}]_{\leq n}$ can be restricted to an ideal projector onto $\Bbbk[\mathbf{x}]_{\leq m}$ for $m < n$.

Here is a counterexample in three variables.

Theorem 6.3.3 ([48]). *There exists a primary projector onto $\Bbbk[x, y, z]_{\leq 2}$ that cannot be restricted to an ideal projector onto $\Bbbk[x, y, z]_{<2}$.*

Proof. It is easy to verify that the space Λ spanned by the functionals

$$\begin{gathered}
\lambda_1 = 1,\ \lambda_2 = x,\ \lambda_3 = x^2,\ \lambda_4 = y + x^3, \\
\lambda_5 = xy + \frac{1}{4}x^4,\ \lambda_6 = z,\ \lambda_7 = z^2, \\
\lambda_8 = xz + \frac{1}{2}x^2 y + \frac{1}{20}x^5,\ \lambda_9 = y^2 + 2x^3 y + \frac{1}{10}x^6 + 6x^2 z, \\
\lambda_{10} = zy + \frac{1}{140}x^7 + \frac{1}{4}x^4 y + zx^3 + \frac{1}{2}xy^2.
\end{gathered} \tag{6.28}$$

is D-invariant and correct for $\Bbbk[x, y, z]_{\leq 2}$. An argument, similar to the one in the above illustration (cf. [48]), shows that no four-dimensional D-invariant subspace of Λ is correct for $\Bbbk[x, y, z]_{\leq 1}$. □

Remark 6.3.4. The ideal $\ker \Lambda$ in the proof of the above theorem is

$$\left\langle x^3 - 6y, x^2 y - xz, xy^2 - zy, x^2 z - 6y^2, xz^2, z^3, z^2 y, zy^2, y^3, xyz \right\rangle \tag{6.29}$$

and the ideal projector is defined by

$$\begin{aligned} &Px^3 = 6y,\ Px^2y = xz,\ Pxy^2 = zy,\ Px^2z = 6y^2, \\ &Pxz^2 = Pz^3 = Pz^2y = Pzy^2 = Py^3 = Pxyz = 0. \end{aligned} \tag{6.30}$$

The two-dimensional version of Sauer's problem (shown below) is still open.

Problem 4. Does every ideal projector onto $\Bbbk[x,y]_{\le n}$ have an ideal restriction onto $\Bbbk[x,y]_{\le m}$ for $m < n$.

The most I succeeded in showing is that every ideal projector onto $\Bbbk[x,y]_{\le 2}$ has an ideal restriction onto $\Bbbk[x,y]_{\le 1}$.

6.3.3 Divided differences

Starting with a Lagrange projector P onto G, Theorem 6.3.1 shows that one can order points in $\mathscr{V}(\ker P)$ in such a way that the Lagrange projector P onto G can be written in Newton form:

$$P = \sum_{j=1}^{N} \lambda_j \otimes g_j, \tag{6.31}$$

Since Lagrange projectors are defined by radical ideals and Hermite projectors are the limits of Lagrange ones, one might expect that Sauer's conjecture, being true for radical ideals, implies its validity for the limits of those projectors. Unfortunately this is not so, as is demonstrated by the following proposition.

Proposition 6.3.5. *The ideal projector P onto $\mathbb{C}_{<3}[x,y,z]$ given by* (6.30) *is Hermite.*

Proof. Direct computations show that the nilpotent operator $M_x + M_z$ for the projector (6.30) is 2-regular, *i.e.* its eigenspace is two-dimensional, and spanned by the polynomials z^2 and yz. As the three multiplication operators commute with the 2-regular operator $M_x + M_z$, they can be approximated by commuting simultaneously diagonalizable operators (cf. [34]). This, in turn, implies that P is a Hermite projector by Theorem 6.2.12(iii). □

The search for the "right" definition of multivariate divided differences had been extensive (cf. [17, 36, 38]) hence not overly successful. Proposition 6.3.5 coupled with Theorem 6.3.3 shows that there does not exist a "continuous" Newton form for multivariate interpolation, hence a "continuous" notion of multivariate divided differences, no matter what the definition is.

6.3.4 Ideal decomposition

We started Section 6.3.1 with a given subspace $G_0 \subset G$ and an ideal projector P onto G and asked for its restriction to G_0. In this subsection we will show that given a G and P one can decompose G and $(\ker P)^{\perp}$ in a consistent way. In other words, for particular choice of basis (g_j), that depends on the projector, P can be written as

$$P = \sum_{j=1}^{N} \lambda_j \otimes g_j,$$

so that, for every $k \leq N$, the projector

$$P = \sum_{j=1}^{k} \lambda_j \otimes g_j$$

is ideal. This follows from the next theorem.

Theorem 6.3.6. *Any N-dimensional D-invariant space $\Lambda \subset \Bbbk[[\mathbf{x}]]$ has D-invariant, $(N-1)$-dimensional subspace.*

Proof. We first assume that $\Lambda \subset \Bbbk[\mathbf{x}]$, *i.e.* Λ is spanned by N polynomials $\lambda_1, \ldots, \lambda_N$. Let $\prec$ be any complete monomial ordering. We define $\deg_{\prec} \lambda$ to be equal to the monomial in λ of the highest order. Clearly this implies that for any polynomial f we have $\deg_{\prec} (D_j \lambda) \prec \deg_{\prec} \lambda$. Assume, without loss of generality, that

$$\deg_{\prec} \lambda_N = \max \{\deg_{\prec} \lambda_k, k = 1, \ldots, N\}.$$

Then there exist numbers a_k such that

$$\deg_{\prec} (\lambda_k - a_k \lambda_N) \prec \deg_{\prec} \lambda_N \text{ for all } k = 1, \ldots, N-1.$$

Let Λ_{N-1} be the space spanned by $\{\mu_k := \lambda_k - a_k \lambda_N, k = 1, \ldots, N-1\}$. Then Λ_{N-1} is a subspace of Λ and it is D-invariant. Indeed, since Λ is spanned by $\{\mu_k, k = 1, \ldots, N-1\} \cup \{\lambda_m\}$ and is D-invariant, we have

$$D_j \mu_k = (\sum b_j \mu_j) + b \lambda_N$$

If $b \neq 0$, then $\deg_{\prec} \lambda_N = \deg_{\prec} D_j \lambda_k \prec \deg_{\prec} \lambda_k \prec \deg_{\prec} \lambda_N$ which gives the contradiction.

Now, if $\Lambda \subset [[\mathbf{x}]]$ then by (6.13)

$$\Lambda = \oplus_{j=1}^{s} \left(e^{\langle \mathbf{z}_j, \mathbf{x} \rangle} \cdot \Lambda_j \right),$$

where Λ_j are polynomial D-invariant subspaces. Hence we can choose $\Lambda_1' \subset \Lambda_1$ such that $\dim \Lambda_1' = \dim \Lambda_1 - 1$ and

$$\Lambda_{N-1} := \Lambda_1' \cdot e^{\langle \mathbf{z}_1, \mathbf{x} \rangle} + \oplus_{j=2}^{s} \left(\Lambda_j \cdot e^{\langle \mathbf{z}_j, \mathbf{x} \rangle} \right)$$

defines the desired subspace. □

6.4 Error Formula

The square root of 5 *is* 2 *for small values of* 5.

(A students in my calculus class.)

In one variable every ideal J is zero-dimensional and curvilinear. In particular every ideal of colength (codimension N) complements $\mathbb{k}[x]_{<N}$. Let P be the ideal projector onto $\mathbb{k}[x]_{<N}$ with $\ker P = J$ and let h be the (unique) monic polynomial that generates J. Then, for every $f \in \mathbb{k}[x]$,

$$P'f := f - Pf = q(f)h$$

for some $q(f) \in \mathbb{k}[x]$. P' is a projector onto J and q is a linear operator on $\mathbb{k}[x]$. We have

$$\ker P' = \ker q = \mathbb{k}[x]_{<N} = \ker D^N,$$

where D^N is the differential operator. It follows that there exists a linear operator $C : \mathbb{k}[x] \to \mathbb{k}[x]$ such that $q(f) = C\left(D^N f\right)$, hence

$$P'f = C\left(D^N f\right) \cdot h.$$

Alternatively, let $J = \langle h \rangle$ be an ideal in $\mathbb{k}[x]$. Then there exists a linear operator C such that

$$f = C(D^N f) \cdot h, \quad \forall f \in J.$$

In approximation theory the formulas of this type are called "error formulas", for they measure the error between the function f and its approximation Pf. The forms of the operator C are well-established (cf. [9, 42]), and at least for Taylor projectors, are written in every Calculus book.

What can be done for ideal projectors P on $\mathbb{k}[\mathbf{x}]$ in general is wide open.

Definition 6.4.1. We will say that a basis $(h_1, \ldots, h_m)$ for the ideal $J \subset \mathbb{k}[\mathbf{x}]$ admits an error formula if there exist homogeneous polynomials H_j and linear operators $C_j : \mathbb{k}[\mathbf{x}] \to \mathbb{k}[\mathbf{x}]$, $j = 1, \ldots, m$ such that

$$C_j(H_k) = \delta_{j,k}. \tag{6.32}$$

and

$$f = \sum_{j=1}^{m} C_j\left(H_j(\mathbf{D})f\right) h_j \tag{6.33}$$

for all $f \in J$.

In other words, if P is an ideal projector with $\ker P = J$ then

$$f - Pf = \sum_{j=1}^{m} C_j\left(H_j(\mathbf{D})f\right) h_j \tag{6.34}$$

for all $f \in \mathbb{k}[\mathbf{x}]$.

Definition 6.4.2. An ideal projector P admits an error formula if the ideal $\ker P$ has a basis that admits an error formula.

Even if a projector P admits an error formula, not every basis for $\ker P$ admits an error formula as we show in the following.

Illustration 6.4.3. Let P be a Lagrange projector onto the span $\{1,x,y\} \subset \mathbb{k}[x,y]$ with $\mathscr{V}(\ker P) = \{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3\} \subset \mathbb{k}^2$. It follows that these three points are not colinear. Let $l_{j,k} \in \mathbb{k}[x,y]$, $j < k = 1,2,3$ be three lines that pass through the points $\mathbf{z}_j, \mathbf{z}_k$. Then (cf. [53]) the basis

$$h_1 := l_{1,2} l_{1,3}, \quad h_2 := l_{1,2} l_{2,3}, \quad h_3 := l_{2,3} l_{1,3} \tag{6.35}$$

admits an error formula. The polynomials H_j are uniquely defined by (6.32).

Compare it to the following result.

Proposition 6.4.4 ([43]). *Let P be an ideal projector onto the span $\{1,x,y\} \subset \mathbb{k}[x,y]$ given by* (6.4)*:*

$$\begin{aligned} Px^2 &= a_0 + b_0 x + c_0 y \\ Pxy &= a_1 + b_1 x + c_1 y \\ Py^2 &= a_2 + b_2 x + c_2 y \end{aligned}$$

Then the border basis

$$\left\{x^2 - Px^2, xy - Pxy, y^2 - Py^2\right\} \tag{6.36}$$

admits an error formula if and only if $c_0 = b_2 = 0$. In particular, if P is a Lagrange projector, then basis (6.36) *admits an error formula if and only if the interpolation sites*

$$\mathscr{V}(\ker P) = \{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3\} \subset \mathbb{k}^2.$$

are vertices of a right triangle with sides parallel to the axis.

Carl de Boor (cf. [10]) proved the existence of the error formula for the special class of Lagrange projectors: so-called Chang–Yao interpolation. This is precisely the case when the interpolation sites generate an ideal $\ker P$ that has a basis of polynomials each of which is a product of linear factors.

Problem 5 (de Boor [11]). Does every zero-dimensional ideal J have a basis $(h_1, \ldots, h_m)$ that admits an error formula?

Proposition 6.4.5 ([47]). *Let $G = \mathbb{k}[\mathbf{x}]_{\leq n}$ and let $P \in \mathfrak{P}_G$. Let $(h_1, \ldots, h_m)$ be a basis for* $\ker P$ *that admits an error formula. Then $h_j = \tilde{H}_j - P\tilde{H}_j$ where $\left\{\tilde{H}_j, j = 1, \ldots, m\right\}$ is a $\mathbb{k}$-bases for the space of homogeneous polynomials of degree $n+1$.*

Notice that the condition $c_0 = b_2 = 0$ in the illustration above is equivalent to the fact that the polynomials in the (6.36) are products of linear factors. Also notice $\ker P_*$ defined by (6.15) cannot have a bases consisting of product of the linear factors and satisfying conditions of Proposition 6.4.5.

Conjecture 6.2. P_* does not admit an error formula.

In fact, I will go on the limb here by proposing the following conjectures.

Conjecture 6.3. A basis for a zero-dimensional ideal admits an error formula if and only if each element of the basis is a product of linear factors.

Conjecture 6.4. If a basis of an ideal admits an error formula then this basis is unshortenable.

6.5 Loss of Haar

If you cannot be a part of the solution, you must be a part of a problem.
(The commonly paraphrased version of the original quote of Eldridge Cleaver)

I know very little about the following (last) set of problems, yet I cannot resist mentioning it.

In one variable every ideal of codimension N complements the space $\mathbb{k}[x]_{<N}$, that is G is a universal ideal complement and in fact a unique such complement (cf. [45]). In particular, the space $G = \mathbb{k}[x]_{<N}$ is Haar, meaning that for every set $\mathscr{Z} = \{z_1, \ldots, z_N\} \subset \mathbb{k}$ of distinct points and for every $f \in \mathbb{k}[x]$ there exists (unique) $g \in G$ such that $f(z_j) = g(z_j)$ for all $j = 1, \ldots, N$. In other words, the Vandermonde determinant

$$V = \det\left(z_k^j\right), j = 0, \ldots, N-1, k = 1, \ldots, N \tag{6.37}$$

generates the ideal $I(\mathscr{V}) \subset \mathbb{k}[z_1, \ldots, z_N]$ where $\mathscr{V}$ is an affine variety

$$\mathscr{V} := \left\{(z_1, \ldots, z_N) \subset \mathbb{k} : z_i = z_j \text{ for some } i \neq j\right\}. \tag{6.38}$$

The well-known Mairhuber's theorem (cf. [27]) states that such a subspace does not exist in several variables. The radical ideal $I(\mathscr{V})$ in

$$\mathbb{k}\left[x_{1,1}, \ldots x_{1,d}, \ldots, x_{d,1}, \ldots x_{d,d}\right] \tag{6.39}$$

generated by the variety

$$\mathscr{V} := \left\{\left(\mathbf{z}_j = \left(x_{j,1}, \ldots x_{j,d}\right) \in \mathbb{k}^d\right) : \mathbf{z}_i = \mathbf{z}_j \text{ for some } i \neq j\right\} \tag{6.40}$$

is not principle for $d > 1$.

Problem 6. What is the minimal number of generators of an ideal J such that $\sqrt{J} = I(\mathscr{V})$?

For $d = 2$ significant progress in understanding the ideal $I(\mathscr{V})$ was obtained by Hayman [21].

The one-variable situation described above has a nice linguistic extension to several variables. Observe that $\Bbbk[x]_{<N}$ is a unique N-dimensional D-invariant subspace of $\Bbbk[x]$ and it is generated by monomials. The next theorem goes at least as far back as 1900 (cf. Gordan [18] also [13, 8]).

Let $\mathfrak{G}_N$ denotes the family of all N-dimensional D-invariant subspaces of $\Bbbk[\mathbf{x}]$ generated by monomials.

Theorem 6.5.1. *For every ideal $J \in \mathfrak{J}_N$ there exists $G \in \mathfrak{G}_N$ such that G complements J, i.e. G spans $\Bbbk[x]/J$.*

The subspaces in $\mathfrak{G}_N$ are the staircases, and $\mathfrak{G}_N$ is the least family of subspaces in $\mathfrak{G}_N$ that satisfies the conclusion of Theorem 6.5.1.

Problem 7. What is the smallest number of subspaces (without any additional assumptions) that satisfy the conclusion of Theorem 6.5.1? What are these subspaces?

Problem 8. What is the smallest family of subspaces $\left\{G_1, \ldots, G_{m(N)}\right\}$ such that for every $\mathscr{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\} \subset \Bbbk^d$ of distinct points and for every $f \in \Bbbk[x]$ there exists $k \le m(N)$ and $g \in G_k$ such that $f(z_j) = g(z_j)$ for all $j = 1, \ldots, N$?

The answer to Problem 6 will provide a lower bound for Problem 8; and the lower bounds tend to be the hardest ones. If $N = 3$, the answers to Problem 8 and, therefore, Problem 7 is $m = 3 = \#\mathfrak{G}_3$ (cf. [45]). I have strong doubts that $m(N) = \#\mathfrak{G}_N$ in general.

A variation on the last problem, in terms of N-regular embeddings, is due to Borsuk [4].

Definition 6.5.2. An *interpolation space* is a subspace $G \subset \Bbbk[\mathbf{x}]$ such that for every set $\mathscr{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\} \subset \Bbbk^d$ of distinct points and for every $f \in \Bbbk[x]$ there exists a $g \in G$ such that $f(z_j) = g(z_j)$ for all $j = 1, \ldots, N$.

Problem 9. What is the smallest dimension $i(N,d)$ of an interpolation space $G \subset \Bbbk[\mathbf{x}]$?

Here is a stunning estimate of F. Cohen and D. Handel.

Theorem 6.5.3 ([6]). *The least dimension $i(N,2)$ of an interpolation space $G \subset \mathbb{R}[x,y]$ satisfies*

$$2N - \eta(N) \le i(N,2) \le 2N - 1 \tag{6.41}$$

where $\eta(N)$ is the number of 1's in the binary representation of the integer N.

The right-hand side is easy. It is attained on the space of Harmonic polynomials of degree $N-1$. The left-hand side is obtained by means of algebraic topology. For $N = 3$ the exact answer is $i(3,2) = 4$ hence the lower bound is attained. For $N = 4$ the right-hand side of (6.41) coincides with the left-hand side and $i(4,2) = 7$. The exact value for $i(5,2)$ is not known to me.

For $d > 2$ even less is known. The best known bounds:

$$\frac{1}{2}(d+1)N \leq i(N,d) \leq d(N+1) \tag{6.42}$$

are far from the best. The lower estimate was originally obtained in [3] (for a simple proof cf. [6]). The upper bound was proved in [39] using Rene Thom's transversality theorem.

So... what does it all have to do with algebraic geometry? The existence of finite-dimensional interpolation spaces implies that the ideal $I(\mathscr{V})$ with $\mathscr{V}$ given by (6.40) is determinantal. That is, it is generated by $N \times N$ determinants of an $N \times k$ matrix with some $k \geq N$.

Problem 10. What is the least $k = k(N,d)$ such that the generators of $I(\mathscr{V})$ are $N \times N$ determinants of an $N \times k$ matrix?

Acknowledgment

When you don't know where you're going, every road will take you there.
(Common sense)

First and foremost, I want to thank Carl de Boor who introduced me to the subject; patiently and tirelessly coached me in the way of thinking, talking and writing about ideal projectors and provided much needed encouragement in the last four years. I also want to thank a number of algebraic geometers who served as anonymous referees for my papers. Their friendly advice and criticisms added a lot to my understanding of the relationship between AG and AT. If this paper does not reflect this understanding, the fault is not theirs but entirely my own. I am grateful to Gregory McColm and Tom McKinley for reading the manuscript and correcting my English. Finally, I want to thank John Abbott, Anna Bigatti, Martin Kreuser and Lorenzo Robbiano who took a chance at allowing me to tango with the audience at this ApCoA conference.

Appendix: AT-AG dictionary

Warning: Like every dictionary, this one is not exact and is only designed to assist in making the right associations.

Approximation theory	**Algebraic geometry**
Ideal projector P	ideal $J = \ker P$
Interpolation sites	$\mathscr{V}(J)$, variety of J
Lagrange projectors	Radical ideals
Primary ideal projector	Primary ideal
dimension of P, codim $\ker P$	colength of J
$G = \text{ran } P$	$A = \mathbb{k}[\mathbf{x}]/J$
Curvilinear projector	$A \simeq \mathbb{k}[x]/\langle x^N \rangle$
Dual space $\mathbb{k}'[\mathbf{x}]$	Inverse systems
$(\ker P)^{\perp} = ranP^*$	dualizing module $\hat{A}$
$\Lambda \subset \mathbb{k}'[\mathbf{x}]$ correct for G	$A = \mathbb{k}[\mathbf{x}]/\text{Ann } \Lambda$
de Boor's equation $P(fPg) = P(fg)$	$[f[g]] = [fg]$ in $\mathbb{k}[\mathbf{x}]/J$
M_j multiplication operators on G	Multiplication maps on $\mathbb{k}[\mathbf{x}]/J$
$\mathfrak{P}_N$: N-dimensional ideal projectors	Hilbert scheme $Hilb_N(\mathbb{k}^d)$
$\mathfrak{P}_G, \mathscr{P}_G$	Border schemes
Hermite projectors	Radical component of $Hilb_N(\mathbb{k}^d)$

References

1. W. Auzinger and H. Stetter. An elimination algorithm for the computation of all zeros of a system of multivariate polynomial equations. In *Numerical mathematics, Singapore 1988*, volume 86 of *Internat. Schriftenreihe Numer. Math.*, pages 12–30. Birkhäuser, Basel, 1988.
2. G. Birkhoff. The algebra of multivariate interpolation. In C. V. Coffman and G. J. Fix, editors, *Constructive approaches to mathematical models*, pages 345–363. Academic Press, New York, 1979.
3. S. Boltianski, S. Ryskov, and Yu. Saskin. On k-regular embeddings and their applications to the theory of approximation of functions. *Uspehi Mat. Nauk*, 15(6):125–132, 1960. English translation. Amer. Math. Soc. Transl. 28(2), 1963, 211–219.
4. K. Borsuk. On the k-independent subsets of the Euclidean space and of the Hilbert space. *Bull. Acad. Polon. Sci.*, III(5):351–356, 1957.
5. Dustin A. Cartwright, Daniel Erman, Mauricio Velasco, and Bianca Viray. Hilbert schemes of 8 points in $\mathbb{A}^d$. arXiv:0803.0341.
6. F. R. Cohen and D. Handel. k-regular embeddings of the plane. *Proc. Amer. Math. Soc.*, 72(1):201–204, 1978.
7. D. Cox. Solving equations via algebras. In A. Dickenstein and I. Z. Emiris, editors, *Solving Polynomial Equations, Foundations, Algorithms, and Applications*, volume 14 of *Algorithms and Computation in Mathematics*, pages 63–123. Springer, Berlin, 2005.
8. D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties, and Algorithms*. Undergraduate Texts in Mathematics. Springer-Verlag, New York, NY, second edition, 1997. An introduction to computational algebraic geometry and commutative algebra.

9. C. de Boor. On the error in multivariate polynomial interpolation. *Appl. Numer. Math.*, 10:297–305, 1992.
10. C. de Boor. The error in polynomial tensor-product, and Chung–Yao interpolation. In A. LeMéhauté, C. Rabut, and L. Schumaker, editors, *Surface Fitting and Multiresolution Methods (Chamonix–Mont-Blanc, 1996)*, pages 35–50. Vanderbilt University Press, Nashville, TN, 1997.
11. C. de Boor. Ideal interpolation. In C. K. Chui, M. Neamtu, and L. Schumaker, editors, *Approximation Theory XI: Gatlinburg 2004*, pages 59–91. Nashboro Press, Brentwood, TN, 2005.
12. C. de Boor. What are the limits of Lagrange projectors? In B. Bojanov, editor, *Constructive Theory of Functions (Varna 2005)*, pages 51–63. Marin Drinov Academic Publishing House, Sofia, Bulgaria, 2006.
13. C. de Boor. Interpolation from spaces spanned by monomials. *Adv. Comput. Math.*, 26(1–3):63–70, 2007.
14. C. de Boor and A. Ron. On polynomial ideals of finite codimension with applications to box spline theory. *J. Math. Anal. Appl.*, 158:168–193, 1991.
15. C. de Boor and B. Shekhtman. On the pointwise limits of bivariate Lagrange projectors. *Linear Algebra Appl.*, 429:311–325, 2008.
16. J. Fogarty. Algebraic families on an algebraic surface. *Amer. J. Math.*, 90:511–521, 1968.
17. M. Gasca and T. Sauer. On the history of multivariate polynomial interpolation. *J. Comput. Appl. Math.*, 122(1–2):23–35, 2000. Numerical Analysis 2000, Vol. II: Interpolation and Extrapolation.
18. M. Gordan. Les invariants des formes binaires. *J. Math. Pures et Appl. (Liouville's J.)*, 6:141–156, 1900.
19. R. Guralnick. A note on commuting pairs of matrices. *Linear Multilinear Algebra*, 31(1–4):71–75, 1992.
20. R. Guralnick and B. Sethurman. Commuting pairs and triplets of matrices and related varieties. *Linear Algebra Appl.*, 310:139–148, 2000.
21. M. Hayman. Commutative algebra of n points on the plane. In L. Avramov, M. Green, C. Haneke, K. Smith, and B. Sturmfels, editors, *Lectures in Contemporary Commutative Algebra*, Mathematical Science Research Institute Publications, pages 153–180. Cambridge University Press, Cambridge, UK, 2004.
22. A. Iarrobino. Reducibility of the families of 0-dimensional schemes on a variety. *Invent. Math.*, 15:72–77, 1972.
23. A. Iarrobino and J. Emsalem. Some zero-dimensional generic singularities; finite algebras having small tangent space. *Compositio Math.*, 36(2):145–188, 1978.
24. A. Kehrein, M. Kreuser, and L. Robbiano. An algebraist's view on border basis. In A. Dickenstein and I. Z. Emiris, editors, *Solving Polynomial Equations, Foundations, Algorithms, and Applications*, volume 14 of *Algorithms and Computation in Mathematics*, pages 169–202. Springer, 2005.
25. Kyungyong Lee. On the symmetric subscheme of Hilbert scheme of points. arXiv:0708.3390v2.
26. F. S. Macaulay. *The algebraic theory of modular systems*. Cambridge University Press, Cambridge, UK, 1916. Reprinted 1994.
27. J. C. Mairhuber. On Haar's theorem concerning Chebychev approximation problems having unique solutions. *Proc. Amer. Math. Soc.*, 7:609–615, 1956.
28. J. T. Marty. *Introduction to the Theory of Bases*. Springer-Verlag, 1969.
29. E. Miller and B. Sturmfels. *Combinatorial Commutative Algebra*, volume 227 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, NY, 2000.
30. H. M. Möller. Hermite interpolation in several variables using ideal-theoretic methods. In W. Schempp and K. Zeller, editors, *Constructive Theory of Functions of Several Variables (Proc. Conf., Math. Res. Inst., Oberwolfach, 1976)*, volume 571 of *Lecture Notes in Mathematics*, pages 155–163. Springer, Berlin, 1977.
31. T. S. Motzkin and O. Taussky. Pairs of matrices with property L. II. *Trans. Amer. Math. Soc.*, 80(2):387–401, 1955.

32. D. Mumford. *The Red Book of Varieties and Schemes*, volume 1358 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1988.
33. H. Nakajima. *Lectures on Hilbert schemes of points on surfaces*, volume 18 of *Amer. Math. Soc. Univ. Lect. Ser.* American Mathematical Society, Providence, RI, 1999.
34. K. C. O'Meara and C. Vinsonhaler. On approximately simultaneously diagonalizable matrices. *Linear Algebra Appl.*, 412(1):39–74, 2006.
35. L. Robbiano. Zero-dimensional ideals or the inestimable value of estimable terms. In B. Hanzon and M. Hazewinkel, editors, *Constructive Algebra and Systems Theory*, Verh. Afd. Natuurkd. 1. Reeks. K. Ned. Akad. Wet., pages 95–114. Royal Netherlands Academy of Arts and Sciences, Amsterdam, The Netherlands, 2006.
36. T. Sauer. Polynomial interpolation in several variables: Lattices, differences, and ideals. In M. Buhmann, W. Hausmann, K. Jetter, W. Schaback, and J. Stöckler, editors, *Multivariate Approximation and Interpolation*, volume 12 of *Studies in Computational Mathematics*, pages 189–228. Elsevier B. V., Amsterdam, 2006.
37. T. Sauer and Yuan Xu. On multivariate Hermite interpolation. *Adv. Comput. Math.*, 4:207–259, 1995.
38. T. Sauer and Yuan Xu. On multivariate Lagrange interpolation. *Math. Comp.*, 64(211):1147–1170, 1995.
39. B. Shekhtman. Interpolation by polynomials in several variables. In *Approximation Theory, X (St. Louis, MO, 2001)*, Innov. Appl. Math., pages 367–372. Vanderbilt Univ. Press, Nashville, TN, 2002.
40. B. Shekhtman. Ideal projections onto planes. In *Approximation Theory XI: Gatlinburg 2004*, Mod. Methods Math., pages 395–404. Nashboro Press, Brentwood, TN, 2005.
41. B. Shekhtman. On a conjecture of Carl de Boor regarding the limits of Lagrange interpolants. *Constr. Approx.*, 24(3):365–370, 2006.
42. B. Shekhtman. On one question of Ed Saff. *Elec. Trans. Numer. Anal.*, 25:439–445, 2006.
43. B. Shekhtman. On the naïve error formula for bivariate linear interpolation. In *Wavelets and Splines: Athens 2005*, Mod. Methods Math., pages 416–427. Nashboro Press, Brentwood, TN, 2006.
44. B. Shekhtman. On perturbations of ideal complements. In B. Randrianantonina and N. Randrianantonina, editors, *Banach Spaces and their Applications in Analysis*, pages 413–422. Walter de Gruyter, Berlin, 2007.
45. B. Shekhtman. Uniqueness of Tchebysheff spaces and their ideal relatives. In *Frontiers in Interpolation and Approximation*, volume 282 of *Pure Appl. Math. (Boca Raton)*, pages 407–425. Chapman & Hall/CRC, Boca Raton, FL, 2007.
46. B. Shekhtman. Bivariate ideal projectors and their perturbations. *Adv. Comput. Math.*, 29(3):207–228, 2008.
47. B. Shekhtman. On error formulas for multivariate polynomial interpolation. In M. Neamtu and L. Schumaker, editors, *Approximation Theory XII: San Antonio 2007*, pages 386–397. Nashboro Press, Brentwood, TN, 2008.
48. B. Shekhtman. On a conjecture of Tomas Sauer regarding nested ideal interpolation. *Proc. Amer. Math. Soc.*, 137:1723–1728, 2009.
49. B. Shekhtman. On the limits of Lagrange projectors. *Constructive Approximation*, 39:293–301, 2009.
50. H. J. Stetter. Matrix eigenproblems at the heart of polynomial system solving. *SIGSAM Bull.*, 30(4):22–25, 1995.
51. H. J. Stetter. *Numerical Polynomial Algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2004.
52. B. Sturmfels. Four counterexamples in combinatorial algebraic geometry. *J. Algebra*, 230(1):282–294, 2000.
53. S. Waldron. The error in linear interpolation at the vertices of a simplex. *SIAM J. Numer. Anal.*, 35(3):1191–1200 (electronic), 1998.

Chapter 7
An Introduction to Regression and Errors in Variables from an Algebraic Viewpoint

Eva Riccomagno and Henry P. Wynn

Abstract There is a need to make a closer connection between classical response surface methods and their experimental design aspects, including optimal design, and algebraic statistics, based on computational algebraic geometry of ideals of points. This is a programme which was initiated by Pistone and Wynn (Biometrika, 1996) and is expanding rapidly. Particular attention is paid to the problem of errors in variables which can be taken as a statistical version of the ApCoA research programme.

7.1 Regression and the X-matrix

Classical linear regression can be described by a model for an output $Y(x)$ as a family of linearly independent functions $\{f_j(x)\}_{j=1,\ldots,k}$ of an independent variable x, also called a predictor, and typically $x \in \mathbb{R}^d$:

$$Y(x) = \sum_{j=1}^{k} \theta_j f_j(x) + \varepsilon, \tag{7.1}$$

where $\theta_j \in \mathbb{R}$ $(j = 1,\ldots,k)$ and ε is a random error term. An experimental design is a set $D = \{x^{(1)},\ldots,x^{(n)}\} \subset \mathbb{R}^d$, of size n, the sample size. At each design point $x^{(i)}$ we make an observation Y_i. Using these observations, we can write

Eva Riccomagno
Dipartimento di Matematica, Genova, Italy, e-mail: riccomagno@dima.unige.it

Henry P. Wynn
Department of Statistics, London School of Economics, London, UK,
e-mail: h.wynn@lse.ac.uk

L. Robbiano, J. Abbott (eds.), *Approximate Commutative Algebra*, Texts and Monographs in Symbolic Computation, DOI 10.1007/978-3-211-99314-9_7, © Springer-Verlag/Wien 2009

$$Y_i = \sum_{j=1}^{k} \theta_j f_j(x^{(i)}) + \varepsilon_i$$

with $\varepsilon_i = \varepsilon(x^{(i)})$ a copy of ε and hence Y_i inherits its randomness from ε_i. The X-matrix, also called the covariate matrix, is the matrix where the (i,j) entry is the value of f_j at the point $x^{(i)}$; that is, the observations index the rows of X and the functions defining the model index the columns:

$$X = [f_j(x^{(i)})]_{i=1,\dots,n;j=1,\dots,k}.$$

In matrix terms Equation (1) becomes

$$Y = X\theta + \varepsilon \tag{7.2}$$

where $\varepsilon = (\varepsilon_i)_{i=1,\dots,n}$, $\theta = (\theta_j)_{j=1,\dots,k}$ and $Y = (Y_i)_{i=1,\dots,n}$. Often the set $\{f_j(x)\}_{j=1,\dots,k}$ is called a basis with reference to the fact that the columns of X should be linearly independent, in order to achieve identifiability, which is a desirable requisite for a linear regression model.

The standard second order assumptions are that: the errors have mean zero, $\mathrm{E}(\varepsilon) = 0$; and the covariance matrix $\mathrm{Cov}(\varepsilon) = \sigma^2 I_{n\times n}$, where $\sigma^2 > 0$ is the error variance. The standard distributional assumption is that ε is multivariate normal (Gaussian). Thus the regression model can be written as $Y = \mathrm{E}(Y) + \varepsilon$ and $\mathrm{E}(Y) = X\theta$.

In the algebraic theory the independent functions are typically monomial so that $f_j(x) = m_j(x)$ are monomials in $\mathbb{R}[x_1,\dots,x_d]$, the set of polynomials in $x_1,\dots,x_d$ with real coefficients, and then

$$X = [m_j(x^{(i)})]_{i,j}.$$

Another important matrix is the information matrix $X^T X$ and its normalized form, the moment matrix

$$M = \frac{1}{n} X^T X.$$

Here X^T is the transpose of X. In the case of monomial functions the entries of M are polynomial moments of the form $\frac{1}{n}\sum_{j,k} m_i(x^{(j)}) m_l(x^{(k)})$.

The following are standard and outline the role of the X-matrix and of the information matrix in estimation.

1. A least squares estimator of the parameter vector θ is

$$\hat{\theta} = \arg\min ||Y - X\theta||^2, \tag{7.3}$$

where Y is the vector of observations, and if X is full rank, $\hat{\theta}$ is given by

$$\hat{\theta} = (X^T X)^{-1} X^T Y.$$

We shall assume, from now on, that X is full rank. The $\hat{\theta}$ are random variables because Y is a random vector. In practice, one observes a realisation Y' of Y, that is of ε, and thus the numeric vector $(X^TX)^{-1}X^TY'$ is an estimate of $\hat{\theta}$.

2. The covariance matrix of $\hat{\theta}$ is

$$\mathrm{Cov}(\hat{\theta}) = \sigma^2(X^TX)^{-1}.$$

3. The matrix $P = X(X^TX)^{-1}X^T$ is the (symmetric, idempotent) projector onto the column space (range) of the X-matrix and we have a partition of the identity matrix of dimension n

$$I_{n\times n} = P + I - P$$

which gives the basic decomposition of $||Y||^2$ arising out of the least squares operation:

$$||Y||^2 = Y^TPY + Y^T(I-P)Y.$$

Thus $Y^TPY = ||\hat{Y}||^2$ and $Y^T(I-P)Y = ||Y-\hat{Y}||^2$ where

$$\hat{Y} = PY = X\hat{\theta},$$

is the vector of predicted values and $R = Y - \hat{Y} = (I-P)Y$ is the vector of residuals.

If $n = k$, so that X is square, we have exact interpolation and then

$$\hat{\theta} = X^{-1}Y,$$

and trivially $\hat{Y} = Y$. Furthermore, sometimes, more than one observation is taken at a location $x^{(i)}$. In this case, we might re-index the Y_i as Y_{ij} where i runs over the distinct locations and j runs over the *replications* at each location. Then, if the number of replications at each design point are equal, $\hat{\theta} = X^{-1}\bar{Y}$ where $\bar{Y}$ is the average value of the observations at each distinct location. When $n = k$ we can use the term *saturated basis* for the f_j's. As will be explained in the last section, the algebra gives a saturated monomial basis.

We end the session with few notes on a case that will not be treated later on. In an era when there is a need to analyse large data sets, the case when k is much larger than n, and hence the X-matrix is not full rank, has received much attention. A proper discussion will have to involve machine learning and data mining theories. Here we just mention two recent applications involving regression models with only linear effect (monomials of degree one and zero) and which have large potential for the application of the algebraic methodology. The first one consists on modifications of $k = 4088$ genes to improve Riboflavin production rate when, to start with, only $n = 71$ observations are available. Under the hypothesis of Gaussian error the θ coefficients have a causal interpretation according to the intervention scheme in (Pearl, 2000). The design problem is driven by sequential model fitting and variable selection (Bühlmann, 2008). A formally similar problem occurs in network inference, a sub-field of systems biology, where the interest is on the identification of biochemical networks from experimental data. A connection to design of experiments and regression analysis was started in (Laubenbacher and Stigler, 2008).

7.2 Orthogonal polynomials and the residual space

A standard procedure in statistics, at least conceptually, is to start with a large saturated set of f functions, but actually fit the data to a sub-set. Suppose the sub basis is monomial and suppose that X is the an $n \times k$ covariate matrix and has full rank $k < n$. Then we can extend X to a full basis and consider an extended covariate matrix

$$\tilde{X} = [X : Z].$$

The : indicates that the matrix X is augmented with the $n \times (n-k)$-matrix Z. We must assume that there is a saturated full basis and we can do this with an orthogonal kernel K, with $K^T X = 0$

$$\tilde{X} = [X : K].$$

Hence $\tilde{X}$ is square and full rank. There are several ways of constructing a kernel K. One way is to use the residual projector, defined above, and write

$$K = (I - P)X,$$

and, as we saw above, the residuals are $R = KY$.

An important and related construction is that of orthogonal functions and this is particularly natural in the case of monomial bases produced by the algebraic method. To construct a basis which is orthogonal with respect to the design we can proceed as follows. Let

$$X^T X = U^T U$$

be the Cholesky factorization of the information matrix, where U is a $k \times k$ upper triangular matrix. Then consider the new basis of functions:

$$g(x) = (U^T)^{-1} f(x),$$

where $f(x)^T = [f_1(x), \ldots, f_k(x)]$. Then we can rewrite the model as

$$Y(x) = \sum_{i=1}^{k} g_j(x)\phi_j + \varepsilon, \qquad \phi_j \in \mathbb{R}.$$

In matrix terms this becomes

$$Y = X\theta + \varepsilon = Z\phi + \varepsilon,$$

where $Z = XU^{-1}$ and $\phi = U\theta$. Then the functions ϕ_j are orthogonal with respect to the design because

$$\sum_{x\in D}\phi(x)\phi^T(x)=\sum_{x\in D}(U^T)^{-1}f(x)f(x)^TU^{-1}$$

$$=(U^T)^{-1}\sum_{x\in D}f(x)f(x)^TU^{-1}=(U^T)^{-1}X^TXU^{-1}=I_{k\times k}.$$

Note that orthogonal functions give another way of finding a kernel for the matrix

$$K=[\phi_j(x)]_{x\in D,j=k+1;\dots,n},$$

which is a full rank matrix orthogonal to X. Using statistical terminology, we have decomposed the residual space using orthogonal functions.

The orthogonal *effects* are defined as the least squares estimates of the parameters in the orthogonal representation of the model:

$$\hat{\phi}=Z^{-1}Y=UX^{-1}Y=U\hat{\theta}.$$

It is standard in many fields to decompose $||Y||^2$ (the total *sums of squares*) into orthogonal effect with large $\hat{\phi}_j^2$, the signal terms, and those with small $\hat{\phi}_j^2$, noise terms. Indeed given any model of order k (the length of the vector f) we have

$$||Y||^2=\sum_{j=1}^{k}\hat{\phi}_i^2+\sum_{j=k+1}^{n}\hat{\phi}_j^2.$$

“Small”, usually means relative to an estimate of the underlying variance σ^2. The classical unbiased estimate of σ^2 regression is $\frac{1}{n-k}||R||^2$. Graphical inspection of the raw $\hat{\phi}_j$ is also standard. In signal processing, to reduce the dimension of a model (number of f_j’s) sophisticated methods set to zero all ϕ_j for which the corresponding $\hat{\phi}_j$ are below some threshold; Fourier and wavelet analysis are examples. It should also be noted that with a design of sufficient symmetry and saturated basis of sufficient symmetry, the orthogonal functions may be well known. For example in Fourier analysis if the design is equally spaced on $[0,1]$ then the functions $\{\sin(2\pi r),\cos(2\pi r)\}$ up to order $r=m$ form a basis for a sample size $n\geq 2m+1$, the “Nyquist rate”. The set of ordered $\hat{\phi}_j^2$ is then a “power spectrum”.

In line with the motivation of this paper, to make a closer link to the algebra, if the large basis is constructed using an ordering on the monomials of $\mathbb{R}[x_1,\dots,x_d]$, then it is natural to use the same order to create the Cholesky decomposition. In one dimension, *i.e.* $d=1$, the theory of orthogonal polynomials has a large literature, for example on the interlacing properties of the zeros. The theory for multidimensional polynomial regression is less well known. In this case the “zeros” are special varieties and very little seems to be known about their properties. If we think of design points as quadrature points then there is some theory available in the quadrature literature.

Example 7.2.1. As simple exercise consider the staircase (echelon) design in two-dimensions: $D=\{(0,0),(1,0),(2,0),(3,0),(0,1),(1,1),(2,1),(0,2),(1,2),(0,3)\}$ and the saturated monomial basis $1\prec x\prec y\prec x^2\prec xy\prec y^2\prec x^3\prec x^2y\prec xy^2\prec y^3$. Then the final orthogonal polynomial is computed up to a scalar as

$$-16236+59494x+128732y-42900x^2-154269xy-109578y^2$$
$$+8300x^3+40266x^2y+54621xy^2+22930y^3.$$

To summarize, if we use, for orthogonalization, the same monomial ordering as that used to construct the monomial basis (see the last section) we can do a Fourier-type analysis using the algebra.

7.3 The fitted function and its variance

Suppose we have selected a particular monomial basis listed in multi-index notation so that $f(x)^T = [x^\alpha]_{\alpha\in L}$ where L is a size k subset of $\mathbb{Z}^d_{\geq 0}$. Then the fitted polynomial model is written as

$$\hat{Y}(x) = \sum_{\alpha\in L} \hat{\theta}_\alpha x^\alpha.$$

This is a random function and in fact is a realization of a Gaussian process, $\{Y(x), x\in\mathbb{R}^d\}$, if the original ε is Gaussian. The covariance of the process at the points x_1 and $x_2 \in \mathbb{R}^d$ is given by

$$c(x_1,x_2) = \sigma^2 f(x_1)^T (X^T X)^{-1} f(x_2),$$

and the variance at $x\in\mathbb{R}^d$ is

$$v(x) = c(x,x) = \sigma^2 f(x)^T (X^T X)^{-1} f(x).$$

It is interesting to express these in terms of the orthogonal basis:

$$c(x_1,x_2) = \sigma^2 \sum_{j=1}^k g_j(x_1)g_j(x_2) \quad \text{and} \quad v(x) = \sigma^2 \sum_{j=1}^k g_j^2(x).$$

There are many ways of assessing how well $\hat{Y}(x)$ predicts, or interpolates, the mean response $\mathrm{E}(Y(x)) = X\theta$, or how well $\hat{\theta}$ estimates θ. This is a special case of the theory of estimation or statistical decision theory. The Gauss-Markov theorem tells us that for a $p\times k$ matrix B and any vector of parameters $\psi = B\theta$, the simple “plug in” estimator $\hat{\psi} = B\hat{\theta}$, where $\hat{\theta}$ is the least squares estimator, has minimal variance covariance matrix among all linear unbiased estimators: minimum variance linear unbiased (MVLU). Linear here means of the form: $\hat{\psi} = AY$, with A a $p\times n$ matrix, and unbiased means $\mathrm{E}(\hat{\psi}) = B\theta = \phi$. Specifically we have

$$\mathrm{Cov}(\hat{\psi}) = \sigma^2 B(X^T X)^{-1} B^T \leq \sigma^2 AA^T,$$

where $\leq$ is the Loewner ordering.

The generalisation of the formulation $Y = \mathrm{E}(Y(x)) + \varepsilon = X\theta + \varepsilon$ to $Y = g(\mathrm{E}(Y(x))) + \varepsilon = g(X\theta) + \varepsilon$ with g a suitable function, and for an error ε which is not necessarily Gaussian, leads to the well-applied theory of generalised linear models (see Dobson (2002)).

7.4 "Errors in variables" analysis of polynomial models

The subject of fitting regression models when the design points arrive with errors, which is at the core of the ApCoA programme of research, is a well developed area of statistics under the heading of "errors in variables" or EIV, particularly popular in Econometrics. The reference list gives pointers to the vast literature on EIV models. Typical applications of EIV are when the predictors are not observable directly or there are measurement errors. We give an elementary exposition as a contribution to the ApCoA programme.

Now each design point X_i is a d-dimensional random vector which is the sum of a term of interest $x^{(i)}$ and a noise term $\delta^{(i)}$. As before let $Y_i \in \mathbb{R}$ be the output at X_i. Briefly, there are three non equivalent formulations of Equation (1) for EIV models. In the first two formulations the model structure is

$$Y_i = \sum_{j=1}^{k} \mathrm{E}(f_j(X_i)\theta_j + \varepsilon_i$$
$$X_i = x^{(i)} + \delta^{(i)}$$

with ${\delta^{(i)}}^T = (\delta_{i1}, \dots, \delta_{id})$ and $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. The standard hypotheses on $\delta^{(i)}$ are that they are independent and normally distributed with mean zero and $\mathrm{Var}(\delta_j^{(i)}) = \sigma_{ij}^2$.

1. Functional formulation. Here the $x^{(i)}$ are fixed values and $x^{(i)} = \mathrm{E}(X_i)$.
2. Structural formulation. Here the $x^{(i)}$ are distributed as $x^{(i)} \sim N(x, \sigma_x^2 I)$ where x is a fixed value.

To distinguish between these two formulations, consider a practical example where Y_i is the amount of wheat and X_i the amount of nitrogen in the i'th plot of a cultivated field under investigation. The true value of X_i is $x^{(i)}$ but because of measurement error the measured value is $X_i = x^{(i)} + \delta^{(i)}$. In a functional approach we fix plot locations and assume them to have different levels of nitrogen. In the structural approach we assume that the plots are samples from a larger field and the mean quantity of nitrogen in each plot is the same as in the field.

3. A third formulation, the Berkson model (Berkson, 1950), is characterised by little or no bias in the measurement. Here

$$
\begin{aligned}
Y_i &= \sum_{j=1}^{k} f_j(X_i)\theta_j + \varepsilon_i \\
X_i &= x^{(i)} + \delta^{(i)},
\end{aligned}
$$

where the hypotheses on the $x^{(i)}, \delta^{(i)}, \varepsilon_i$ are as in the structural framework.

In the sequel we adopt the "instrumental variable" approach, that is we assume that the predictor is random but not correlated with the errors, ε. The basic approach is then to show that, to first order, the randomness in the independent variable is to induce an additional error in a regression about the mean values at every design point. This additional term leads to a weighted least square analysis but with an added complication. This is that the induced error depends on the unknown parameters θ. It is standard therefore to carry out a two-stage procedure: first fit an expression to the known mean values and then plug these first stage estimates into the variance terms and conduct a second stage weighted least squares.

We restrict ourselves to showing how the first stage perturbation analysis works for a monomial basis. We make the following assumptions and use a simplified notation. The true (preset) design point is now considered to be $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, but the value used in the actual regression or interpolation is $z = x + \delta$.

Let us assume a monomial basis and consider $m_\alpha(x) = x^\alpha = \prod_{j=1}^{d} x_j^{\alpha_j}$ be a monomial term which appears in the polynomial model, but, instead $z^\alpha = \prod z_i^{\alpha_i}$ is used for the inference. Then a first order approximation yields

$$
z^\alpha = \prod_j (x_j + \delta_j)^{\alpha_i} \simeq \prod_j (x_j^{\alpha_j} + \delta_j \alpha_j x_j^{\alpha_j - 1}|_+) \simeq m_\alpha(x) + \sum_j \delta_j \alpha_j x_j^{\alpha_j - 1}|_+ \prod_{k \neq j} x_k^{\alpha_k}, \tag{7.4}
$$

where $x_j^{\alpha_j - 1}|_+ = 0$ when $\alpha_j = 0$.

Suppose that at each design point we use a perturbed value $z^{(i)} = x^{(i)} + \delta^{(i)}$. Then this is equivalent to using a version of the original covariate matrix $X = \{m_\alpha(x^{i)})\}_{i,\alpha}$, each row of which will have its own first order approximation of the form (7.4). If $\{\delta_j^{(i)}\}$ are perturbations then we can write (to first order) the perturbed X-matrix as

$$
\tilde{X} = X + \triangle,
$$

where $\triangle$ is a matrix whose entries are linear in the $\{\delta_j^{(i)}\}$. In the saturated case let Y be the vector of observations, which, for the moment, we assume fixed (non-random). The parameter estimate, when there is no perturbation, is $\hat{\theta} = X^{-1}Y$. With perturbation, and again using a first order analysis, we have the estimate

$$
\tilde{\theta} = \tilde{X}^{-1}Y = (X + \triangle)^{-1}Y \simeq \hat{\theta} - X^{-1} \triangle \hat{\theta}.
$$

The last term shows the perturbation for fixed $\triangle$. Now, if we assume that the $\triangle$ are random and have zero mean then this term also has zero mean so that the bias (in the statistical sense) is (approximately) zero. The second order approximation to the parameter covariance matrix is

$$\mathrm{Cov}(\tilde{\theta}) = X^{-1}\,\mathrm{Cov}(\triangle\hat{\theta})(X^T)^{-1}$$

Example 7.4.1. Consider the case $d = 2$ and the design $\{(-1,-1),(1,-1),(-1,1),(1,1))\}$ and $m(x) = (1,x_1,x_2,x_1x_2)$. Then,

$$\triangle = \begin{pmatrix} 0 \ \delta_1^{(1)} \ \delta_2^{(1)} \ -\delta_1^{(1)} - \delta_1^{(1)} \\ 0 \ \delta_1^{(2)} \ \delta_2^{(2)} \ -\delta_1^{(1)} + \delta_1^{(1)} \\ 0 \ \delta_1^{(2)} \ \delta_2^{(3)} \ +\delta_1^{(1)} - \delta_1^{(1)} \\ 0 \ \delta_1^{(2)} \ \delta_2^{(4)} \ +\delta_1^{(1)} + \delta_1^{(1)} \end{pmatrix}$$

Assume also, for simplicity, that all the $\{\delta_j^{(i)}\}$ are uncorrelated with variance depending on dimension: $\mathrm{Var}(\delta_j^{(i)}) = \sigma_x^2$. Then, for fixed θ we have

$$\mathrm{Cov}(\tilde{\theta}) = \frac{\sigma_x^2}{4}\begin{pmatrix} \theta_2^2+\theta_3^2+2\theta_4^2 & 2\theta_3\theta_4 & 2\theta_2\theta_4 & 0 \\ 2\theta_3\theta_4 & \theta_2^2+\theta_3^2+2\theta_4^2 & 0 & \\ 2\theta_2\theta_4 & 0 & \theta_2^2+\theta_3^2+2\theta_4^2 & 2\theta_3\theta_4 \\ 0 & 2\theta_2\theta_4 & 2\theta_3\theta_4 & \theta_2^2+\theta_3^2+2\theta_4^2 \end{pmatrix}$$

Thus, we see that the first order perturbation method leads to parameter covariances which are quadratic in the θs. The second step would be to replace these initial values by the values fitted from the data: $\hat{\theta} = X^{-1}Y$, in the saturated case. The analysis would be repeated with error covariance matrix given by: $C = \sigma^2 I + XC(\hat{\theta})X^T$, and weighed least squares used. The first term is the standard covariance structure given above for ordinary regression. This procedure is an approximated version of iterated re-weighted least squares: IRWLS.

7.5 Comments

The ideas in this paper are fairly standard in Statistics, and could be included in a course on regression. The aim of the paper is to provide an overview which could stimulate the interest of algebraists in the multiple aspects of regression analysis and in particular in the inferential aspects linked to the distributional assumptions on the model and on the observation/design points.

At the very core of the application of Algebraic Statistics to regression analysis is the fact that a finite set of points $D \subset \mathbb{R}^d$, at which measurements are taken, is identified with an ideal in $\mathbb{R}[x_1,\ldots,x_d]$, usually called $\mathrm{Ideal}(D)$. Then, a $\mathbb{R}$-vector space basis of $\mathbb{R}[x_1,\ldots,x_d]/\mathrm{Ideal}(D)$ becomes a saturated monomial basis for a regression model $Y(x) = \sum_\alpha \theta_\alpha m_\alpha(x) + \varepsilon$, as already. One of the strength of the method lies in the fact that often the coordinates of the design points are not required in order to perform some computations, *e.g.* to compute moments. For this it is sufficient to be able to perform operations in the quotient space. Actually sometimes the exact knowledge of the ideal is not required. For example the vector space basis of $\mathbb{R}[x_1,\ldots,x_d]/\,\mathrm{Ideal}(D) = \mathbb{R}[x_1,\ldots,x_d]/\,\mathrm{Ideal}(LT(\mathrm{Ideal}(D)))$ and infinitely many D lead to the same leading term ideal. This should be an aid to planning an experiment which still has not been systematically explored.

So far, within Algebraic Statistics techniques have been developed to determine, under different circumstances, design/model pairs for which the X-matrix is full rank (see Pistone *et al*, 2001). But the contribution of Algebraic Statistics to the estimation of regression models has been limited and does not go beyond achieving unbiased models. Although an example can be found in Giglio, Riccomagno and Wynn (2000) where techniques from Algebraic Statistics are coupled to standard statistical techniques to analyse non-orthogonal experiments.

In Section 2 we sketch the role of orthogonal decomposition in understanding the structure of the residual space. The statistical methods typically used are based on numerical analysis and mainly rely on first order approximation theory. In Section 3 second order properties of the fitted model are considered. They are functions of the $[m_\alpha(x)]_{\alpha\in L}$ and hence depend on the choice of a basis in the quotient space. How do we choose the $[m_\alpha(x)]_{\alpha\in L}$ so that the covariance of the process is a diagonal matrix or at least some of its entries are zero? Or how do we chose it so that the variance of each component is smallest? Section 4 starts with the assumption that measurements are made at random design points and the randomness is assumed additive. A sensitivity analysis is sketched to illustrate how the error propagates. But how does the structure of zeros in the Δ matrix relate to the geometry of the design points? Importantly, the covariance matrix of $\tilde{\theta}$, since it depends on the parameters, is not an estimator. Can the algebra provide any insight here and in particular can it provide insight on the working of the iterated reweighted least square procedure?

In our brief excursus on regression we left out an important case: the random effect models or variance component models, used when random effects are randomly distributed and for which the θ parameters in Equation (1) are random variables, with zero mean, unknown variance and independent of each other. We refer to the classic text book by Searle (1987).

7.6 Acknowledgements

The authors are grateful to Professors M. P. Rogantin and A. V. Geramita for useful comments.

References

1. Berkson, J. (1950). Are there two regressions? J. Amer. Statist. Assoc, 45, 164–180.
2. Dobson, A. J.(2002). An introduction to generalized linear models. Second edition. Chapman & Hall/CRC, Boca Raton, FL.
3. Donev, A.N. (2000). Dealing with errors in variables in regression surface exploration. Commun. Statist. Theory Methods, 29, 2065–2077.

4. Donev, A. N. (2004). Design of experiments in the presence of errors in factor levels. J. Statist. Plann. Inference, 33, 569–585.
5. Durbin, J. Errors in variables. (1954). Rev. Inst. Internat. Statist. 22, 23–32.
6. Giglio, B., Riccomagno, E and Wynn, H. P. (2000). Gröbner basis strategies in regression. Journal of Applied Statistics, 27, 923–938.
7. Fuller, W. A. (1987). Measurement error models. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Wiley, New York.
8. Kendall, M. G. (1952). Regression, structure and functional relationship. II. Biometrika, 39, 96–108.
9. Laubenbacher, R. and Stigler, B. (2008). Design of experiments and biochemical network inference. In Geometric and Algebraic Methods in Statistics (eds. P. Gibilisco, E. Riccomagno, M.P. Rogantin and H.P. Wynn), Cambridge University Press, Cambridge (forthcoming).
10. Bühlmann, P. (2008). High-dimensional variable selection: from association to intervention. In Proceedings of the 7th World Congress in Probability and Statistics, Singapore July 14-19, 2008, 80-81.
11. Nummi, T. and Möttönen, J. (2004). Estimation and prediction for low degree polynomial models under measurement errors with an application to forest harvesters. J. Roy. Statist. Soc. C, 53, 495–505.
12. Pearl, J. (2000). Causality. Models, reasoning, and inference. Cambridge University Press, Cambridge.
13. Pistone, G., Riccomagno, E. and Wynn, H. P. (2001). Algebraic Statistics. Chapman & Hall/CRC, Boca Raton.
14. Pistone, G. and Wynn, H. P. (1996). Generalised confounding with Gröbner bases. Biometrika 83:3, 653–666.
15. Ruffa, S. (2004). La statistica algebrica nei modelli di regression con errori nelle variabili (in Italian), Laurea Specialistica, Politecnico di Torino.
16. Searle, S. R. (1987). Linear models for unbalanced data, John Wiley Sons.

Chapter 8
ApCoA = Embedding Commutative Algebra into Analysis

Hans J. Stetter

Abstract I take a philosophical look at Approximate Commutative Algebra and present my view of computational algebra over $\mathbb{C}$, highlighting the important role of analytical concepts and techniques. I look at the particular case of Gröbner bases: an important tool in exact commutative algebra which needs non-trivial adaptation (into border bases) when moving into the world of approximate commutative algebra.

8.1 Introduction

This presentation is not a technical paper in the strict sense. This must be so since I have terminated active research 4 years ago while Approximate Commutative Algebra (ApCoA) has become a blooming subject within recent years, very much in contrast to the situation in the late 20th century. Instead, I will explain the meaning and the validity of the title of my presentation and – in doing so – point out how meaningful research and algorithmic development in ApCoA must be based on viewing the subject as a discipline within Analysis, a view that I have held throughout my work in that area. The overview character of this presentation is also the reason why I have refrained from bibliographical citations besides some general pointers to my book [1].

Most of the 19th century mathematicians regarded it as their prime task to solve problems posed by the natural sciences; for them, it would have been natural that algebraic problems over the real or complex numbers must be considered as problems in analysis for a meaningful numerical treatment. Take Carl F. Gauss, the "princeps mathematicorum": in the university of Göttingen, he was not a professor in mathematics but the director of the observatory, and – along with his fundamental research in discrete *and* continuous mathematics – he did pioneering experimental

Technische Universität Wien, Vienna, Austria,
e-mail: `stetter@aurora.anum.tuwien.ac.at`

L. Robbiano, J. Abbott (eds.), *Approximate Commutative Algebra*, Texts and Monographs in Symbolic Computation, DOI 10.1007/978-3-211-99314-9_8, © Springer-Verlag/Wien 2009

and theoretical work in physics and the technical sciences. In an overdetermined linear algebraic system with numerical coefficients, he “saw” the norm of the residual as a *function* of the pseudo-solution components; therefore he determined his least-squares solution by *differentiation*. And he valued this approach as sufficiently important to be taught in a special course of lectures at the university.

In the fall of 1823, Gauss actively participated in the accurate determination of the distance between the observatories of Goettingen and Altona by an elaborate surveying project, and he faced the daily task of retrieving highly accurate numerical results from the sizable overdetermined linear equation systems arising from the triangulation measurements. At that time, this was a tedious pencil-and-paper job, requiring patience, care – and checks for calculating errors. This chore led to his “invention” of *numerical linear algebra*: not only did he devise a novel, genuinely analytic approach to his task, viz. *iterative improvement*, but his design of an algorithmic recipe realized many of the principal ideas of what developed into numerical analysis: clever choice of the coordinate system, quick determination of a reasonable initial approximation, efficient flow of control in the iterative procedure, meaningful number of digits in the calculation, and many more. Fortunately, he documented his pioneering concepts in a letter to a fellow surveyor, written on 26 Dec.(!) of that year, with explanatory sketches and an elaborated example. The letter closes with the remark (my translation): “I recommend you this method for your own usage. You will hardly use direct elimination ever again, at least with more than 2 unknowns. One can perform the indirect procedure while half asleep, or one can think of other matters at the same time.” (Maybe this led to the name “relaxation” for this family of iterative algorithms.)

I have no doubt that – presented with a multivariate polynomial problem with numerical coefficients – Gauss would not have considered the precise satisfaction of the equalities through exact rational calculation as appropriate tools; he would rather have seen the problem continuously embedded in the set of equally structured systems with *neighbouring coefficients*, and he would have employed concepts and tools of analysis *together with* the relevant algebraic concepts to define and to compute meaningful solutions. This brings me to my subject:

Computational problems in commutative algebra with numerical coefficients require the embedding of commutative algebra into analysis!

8.2 Approximate Commutative Algebra

At first, we have to clarify what we mean – in the context of this presentation – by “Approximate Commutative Algebra”. Of course, you would not be here in this ApCoA-Workshop without a notion of what ApCoA is. In my personal view, ApCoA encompasses the following:

- consideration of problems of Commutative Algebra *over the complex or real numbers*
- admission of some *data of limited accuracy*
- *use of floating-point arithmetic* for the computation of numerical results.

Naturally, the first characteristic *must* be present; but many aspects of ApCoA also apply to exact data and/or with exact computation.

The two main tasks of ApCoA are – in my personal view – the following: on the one hand, we must consider the feasibility in ApCoA of the relevant concepts from Commutative Algebra and define *modified concepts* where required. On the other hand, we must design *algorithms* for the numerical determination of meaningful numerical results for certain computational problems and analyze their performance. In both of these aspects, ApCoA is fully analogous to Numerical Linear Algebra. For this reason, I would have strongly preferred the general usage of the term "*Numerical* Commutative Algebra" in place of "*Approximate* Commutative Algebra"; but it is no use to cry over spilt milk.

Assume that we have a specified multivariate polynomial system with complex coefficients, with some of these coefficients of limited accuracy. What do we mean by an ideal generated by this polynomial system? Or, conversely, given a polynomial ideal through a specified basis, when is a polynomial with numerical coefficients of limited accuracy a member of this polynomial ideal. If we cannot clearly answer these and similar questions, it is meaningless to design an algorithm for the numerical computation of an ideal basis of whatever structure for an "approximate" polynomial system, a typical task of ApCoA. From the term *approximate* it is obvious that such questions do not belong to classical algebra but to analysis.

Instead of further examples which will appear later anyway, I will now to turn to the meaning of "embedding into analysis".

8.3 Empirical Data

Every application scientist using mathematical models is accustomed to the use of "approximate" data, or rather *data with limited accuracy*. I prefer the technical term *empirical data* since "approximate data" may suggest that there *exists* an "exact value" while there is no such thing for most real-world data: take the temperature of a lake, or the height of a tree *etc.*, let alone socioeconomic data. If we use such data in the evaluation of algebraic models, we must have a clear concept of them: *e.g.* what is meant by a coefficient 1.23 which has arisen directly or indirectly from measurements or has been the *result of previous computations* which employed data of limited accuracy?

Clearly, the value 1.23 is not to be understood like a value 2 arising intrinsically in some physical law. Normally, it simply means that – in this context – this specification is the most plausible choice from the sequence $\{\ \ldots, 1.19, 1.20, 1.21, 1.22, 1.23, 1.24, 1.25, 1.26, 1.27,\ \ldots\}$ and that no meaningful choice for a third digit exists. Definitely, it does *not* exclude values in the neighbourhood of 1.23 (like 1.234) as valid instances. At times, an indication of the extension of that neighbourhood may be given by some *tolerance* like $\pm .02$; but, for 1.23, this would *not* mean that the values in $[1.21, 1.25]$ are valid while those outside this interval (like 1.2501) are not.

The discontinuity introduced by such a strict interval would be just as unnatural as the strict concept of a point value.

Rather, in my opinion, an empirical data item should be seen as an expanding *set of neighbourhoods*, with less and less probable choices at their peripheries as their diameters increase. But this heuristic "probability" cannot be formulated rigidly so that a recourse to mathematical probability theory is, generally, also futile. My own suggestion has been the association of a *validity scale* with each such data item (cf. [1], section 3.1.1); but other similar concepts may serve as well. Essential is the view of such data as *loosely defined neighbourhoods* rather than precise points or precise sets like intervals.

Within a *computation*, however, we must use *one* precise value for each empirical data item, its *specified value*, like 1.23 above. The given problem, with all empirical data at their specified values, will be called the *specified problem*. While it is generally not an "exact problem" in any sense, it may be used as a *well-defined reference problem*. Of course, to solve this specified problem exactly, with exact rational calculation, would be "shooting at sparrows with cannons"; floating-point arithmetic, perhaps even with a low word-length, will suffice at first. The results of such a computation must be subject to a *validity check* in any case and they must be *improved* if necessary. Obviously, the discrete view of algebra is not adequate for an understanding of this situation but the continuous view of analysis must be adopted.

8.4 Valid Results; Validity Checking of Results

As mentioned before, a problem with empirical data has *no exact results*. Instead one may define *valid (pseudo-)results* in a consistent and well-defined manner:

Any delivered result of an empirical problem is *valid* if it may be interpreted as the *exact result* of a valid instance of the problem, *i.e.* of a "neighbouring" problem with *valid data*. The genesis of the pseudoresult plays no role in this context.

This concept has the important advantage that the validity of a pseudoresult is defined *strictly within the underlying algebraic context*: note that we require an exact solution in the algebraic sense of a well-defined algebraic problem whose existence must be proved. This also explains my dislike of the name "*Approximate* Commutative Algebra" because in the end there is nothing approximate in the way *algebraic* relations come into play.

At first, it appears that the task of checking the validity of a pseudoresult obtained in whatever way would be as difficult as the solution of the original problem. However, we have to establish only the existence of *some* valid neighbouring problem for which our result is an exact solution. It is true that one will often proceed as if one had to find the *closest* neighbouring problem in the metric introduced by the validity scales of the empirical data. But this optimization problem need only be solved very crudely: either a just-vaguely-nearest conforming neighbouring problem is well within the validity range and thus confirms the validity of the presented result, or it is outside by a substantial amount so that an iterative improvement of the presented pseudosolution appears necessary. If the situation is doubtful, an iteration step will be carried out anyway.

As a first example, consider a numerically computed "pseudozero" of an empirical system of multivariate polynomial equations. An evaluation of the residual will not suffice because it may be manipulated arbitrarily by scaling some or all polynomials.

According to the above validity concept, we will replace the specified values $\tilde{\alpha}_j$ of the empirical coefficients by $\tilde{\alpha}_j + \mu_j \Delta\alpha_j$, where the μ_j reflect the accuracy levels of the empirical coefficients α_j: *e.g.* if an α_j has three significant decimal digits, μ_j may be chosen as 10^{-3}. These new polynomial equations should be satisfied exactly for the pseudozero components. After some manipulation, we are left with a linear system for the $\Delta\alpha_j$ which contains the residuals of the specified system for the pseudozero components. If there are more empirical coefficients than equations, we may minimize the moduli of the $\Delta\alpha_j$ in solving. In a system representing a real-life model, there will always be empirical terms representing small effects which have not been included in the model so that there are sufficiently many $\Delta\alpha_j$. If the weights μ_j have been properly chosen, the pseudozero is valid if all $\Delta\alpha_j$ have a modulus about 1 or smaller.

8.5 Data → Result Mappings

In our introductory remarks, we have emphasized that it is crucial to conceive of an empirical algebraic task as problem for a whole neighbourhood of data. In analyzing our task for a neighbourhood of data simultaneously, we *must* view the relation between the data set and the associated set of results as a *mapping* in the analytic sense if we are to understand what may happen. The qualitative and quantitative properties of this *data → result mapping* will determine the further steps to be taken.

As we deal with problems with numerical data (coefficients and the like), the algebraic objects of our task may, and generally will, also contain numerical data which are *exact*, *e.g.* integers; we will call them *intrinsic data*. It helps to consider these intrinsic values as a *part of the mapping* and not as data. Notably, vanishing coefficients which determine the *sparsity structure* are generally of this kind. As data of the mapping, this leaves the empirical numerical coefficients, with their respective neighbourhoods.

Thus our data → result mappings are defined on bounded regions $A \subset \mathbb{C}^n$ where n is the number of empirical components $\alpha_\nu \in \mathbb{C}$ in our task. The image space, *i.e.* the space to which these regions are mapped, naturally depends on the task:

If we aim, *e.g.*, at locating all zeros of a polynomial system in s variables, the image space will be $\mathbb{C}^{m \times s}$, where m is the number of zeros. If, on the other hand, we aim at some qualitative answer, *e.g.* about the stability of a univariate polynomial ("Are all zeros in the open left half-plane?"), then the image space is discrete and consists only of the truth values 'true' and 'false'. Similarly, an image space may consist of natural numbers.

Let us shortly consider this special case of a discrete image space first. Clearly, to be able to obtain a definite result over the whole data region A under consideration, we must ascertain that the data→result mapping is *constant* there. If it is not, the given task is *ill-defined* for the specified "approximate data", it *cannot* be reliably completed. (A suitable redefinition of the task may sometimes resolve the dilemma.)

One particular situation of this kind has, in the 90's, dominated the discussion about the feasibility of what became Approximate Commutative Algebra: the floating-point computation of Gröbner Bases for polynomial systems which – in view of Wilkinson's backward error analysis – is equivalent to exact GB computation for approximate data. Since the principal result of a GB computation is the *structure* of the GB, its floating-point computation can be meaningful *if and only if* this structure is *constant* over a sufficiently large neighbourhood of the specified data. *A fortiori* this is so if some coefficients are actually of limited accuracy.

Thus, the *only* chance for a safe numerical computation of Gröbner Bases in ApCoA exists if the associated normal sets are *generic* for the chosen term order for a sufficiently large neighbourhood of the specified data, *i.e.* if the normal sets for all neighbouring polynomial systems consist precisely of the initial members of the sequence of monomials in ascending term order. It is well-known that for a polynomial system whose GB has a non-generic normal set, there must occur "jumps" in the normal set structure upon some infinitesimally small generic changes in the coefficients! Since, generally, the normal set structure of a Gröbner Basis to be determined is not known *a priori,* Gröbner Bases are *not suitable* as a standard for ideal bases in Approximate Commutative Algebra. If I had seen this so clearly 15 years ago, I could have saved myself a great deal of futile endeavours and numerous heated discussions with computer algebraists.

Naturally this raises the question of how we can compute ideal or quotient ring bases at all in ApCoA, since it is obvious that there *do not exist* normal sets which are feasible for *all* ideals with a given number of zeros. The only possible answer appears to lie in the *abolition of the term order* as a guiding principle for the algorithmic procedure. As a consequence, this leads to *Border Bases* in place of Gröbner Bases. Algorithms of this kind are possible, in particular if we know the dimension of the quotient ring. For regular polynomial systems, this value can be found from the BKK-algorithm which uses the sparsity structure of the polynomial system only; thus this is no principal problem, at least for zero-dimensional ideals. In my book (cf. [1], chapter 10), I have elaborated how such algorithms may proceed.

8.6 Analytic View of Data→Result Mappings

If the result space X is some $\mathbb{C}^N$, it is clear from the preceding discussions that a necessary prerequisite for a meaningful definition of the computational task is the *continuity* of the data→result mapping $F: A \to X$. Consider some typical tasks from this respect:

- Zeros of regular multivariate polynomial systems (complete intersections): One of the most fundamental theorems of complex analysis says that the *individual* zeros depend continuously on the coefficients of the system (cf. *e.g.* [2], section 10.2) in a sufficiently small data neighbourhood.
- Quotient ring representations for regular multivariate polynomial systems: For a specified system, there exist monomial bases which are valid within a neighbourhood of the data and whose associated multiplication matrices are continuous there; cf. [1], section 2.5.
- Eigenvalues and eigenvectors of square matrices: Since they may be defined via the characteristic polynomial, we have continuity w.r.t. the matrix elements.

Mere continuity is a very weak basis for safe and successful computation. But when we consider *differentiability*, the situation begins to be more complicated: even for a univariate polynomial it is well-known that the mapping from the coefficients to a zero is not differentiable at a *multiple zero*: such a zero decomposes into a *cluster of zeros* upon arbitrarily small generic perturbations of the coefficients. At multiple zeros of multivariate systems, the situation is qualitatively the same but quantitatively much more intricate; cf. [1], section 9.3. This implies that computations in the vicinity of a multiple zero or a zero cluster must be subject to special attention in order to be successful; this will be further discussed below.

Isolated zeros of regular polynomial systems, on the other hand, depend smoothly on the coefficients because the system derivative is nonsingular at an isolated zero. Thus, a neighbourhood of the specified data is mapped to a domain in the neighbourhood of each zero; the dimension of the domain depends on the dimensions of the data and result spaces. This is the basis for the validity of the *Newton method* for improving the accuracy of an approximate isolated zero. However, if other zeros are relatively close, the system derivative may well be *near-singular* at an isolated zero making the computational task very ill-conditioned; see below.

For a regular polynomial system and a feasible monomial basis of the associated quotient ring, the elements of the multiplication matrices are smooth functions of the polynomial coefficients, within the feasibility region of the monomial basis. This implies that the *coefficients of a border basis* for a regular 0-dimensional polynomial ideal are smooth functions of the coefficients of the generating polynomials. Thus, the shortcoming of Gröbner bases is *only* due to their *artificial restriction of admissible normal sets* by a term order.

8.7 Condition

It is well-known to everybody who has had to deal with extended numerical computations that the formal smoothness of the data → result mapping is illusory if the associated derivatives are *extremely large*; this situation is called *ill-conditioning* in numerical analysis. In a *well-conditioned* situation, small changes in the data do not lead to unduly large changes in result values. In an ill-conditioned task, on the other hand, the virtual data changes associated with floating-point computation may completely distort the outcome of a computation.

Therefore, there exists a “grey zone” about strictly singular data configurations, *i.e.* data for which the differentiability of the data → result mapping breaks down. For specified data in this zone, we must expect ill-conditioning and either employ refined computational procedures and checks, or – preferably – redraft the problem formulation such that the ill-conditioning disappears. Since commutative algebra with polynomials presents us with a variety of singular situations, the detection of potential ill-conditioning is a fundamental task in ApCoA and the design of appropriate redrafting measures an important aspect of algorithmic design.

Thus, as a *standard procedure*, one should *check the condition* of intermediate linear systems of equations to be solved, even of very small systems. Also the appearance of a very small number as a divisor *e.g.* through a tiny leading coefficient of an intermediate polynomial) may be a crucial indicator of an intolerably ill-conditioned situation. Often, the information from such checks can be used for the activation of switches leading to a modification of the problem formulation with a vastly improved condition. I will sketch two such situations.

The location of a polynomial zero which is a member of a dense *cluster of zeros* is notoriously ill-conditioned; this will be displayed by the condition of the associated Newton step. If the information on the individual zero has arisen from an eigenvector of a multiplication matrix, we may check for the *set* of near-linearly-dependent eigenvectors which – together – represent the complete cluster. The determination of the *eigenspace* associated with this set can proceed in a well-conditioned manner and permits a safe computation of the complete zero cluster.

With empirical systems, we may often suspect that the cluster should be interpreted as a valid *multiple zero* and try to verify the presence of such a zero at the locality in question. Again, this is, generally, a well-conditioned computation. More technical details are to be found in my book and in the related literature.

As mentioned in section 5, the numerical computation of a basis for a 0-dimensional polynomial ideal must be prone to extreme ill-conditioning or even failure if it is based on a *fixed* structure of the associated normal set – as it is the case with Gröbner basis computation for a specified term order. Even if this situation is discovered before breakdown, it is generally infeasible to use the information gathered so far for a redirected computation w.r.t. a different term order; and this one may once more be unsuitable. In the computation of a *border basis*, on the other hand, one starts with some normal set of the correct number of elements, compatible with the polynomial system. If an ill-conditioning is discovered during the subsequent computations, one can exchange just *one* monomial of the current normal set. This is a simple operation which saves all of the information previously determined; it may also be repeated as often as necessary. For technical details, cf. [1], section 10.3.

If the ill-conditioning arises from a zero with a very large modulus (*i.e.* a zero close to ∞), this can be diagnosed during a border basis computation: in this case, if the computation is continued after an appropriate monomial of the normal set has been discarded and sent to the border set, this is equivalent to the assumption that a valid position of the zero in question is *at* ∞. This fact must, of course, be verified *a posteriori.*

8.8 Overdetermination

Various phenomena appear frequently in polynomial algebra which lead to an *overdetermined, inconsistent* numerical problem; we collect them under the term "overdetermination".

There is, of course, the natural situation in a real-life scenario of having more (not rarely *many more*) observations than would be strictly necessary for the evaluation of some polynomial model. Here, as in the linear case, one will minimize some norm of the residual vector which arises from the individual observations. However, contrary to the linear situation, the least squares approach results in a system of higher degree. Therefore, we must reformulate the minimization in terms of the increments of a rough approximate solution and discard higher than linear powers of the increments in the minimization conditions. This is the immediate generalization of a procedure which Gauss suggested for the linear situation in his initially quoted letter. A similar situation arises in various other "smoothing" scenarios involving empirical polynomials.

A genuinely different situation exists when the consistency of more than s polynomials in s variables is an indispensable part of the task and has to be maintained *strictly*. With some empirical coefficients present, a specified polynomial system will "always" be inconsistent. Thus, what we really have in mind is the set of nearby systems which *are* consistent, preferably the closest such system.

Such a situation arises naturally, *e.g.* when we want to *factor an empirical multivariate polynomial* in a particular way: with the specified values of the empirical coefficients, the polynomial will most certainly *not factor exactly*, *i.e.* the system of equations for the coefficients of the potential factors will almost inevitably be inconsistent; but, naturally, we do not want an exact factorization of the specified polynomial but of a factorable valid instance of the empirical polynomial.

One possible approach consists in the selection and solution of a suitable *regular subsystem*, and the identification of those solutions which fit the remaining equations roughly. By an iterative procedure, this fit may then be simultaneously extended over the full system and improved by a minimization procedure. When we arrive at the solution of a consistent valid instance of the system, we are satisfied.

Quite generally, we have this situation also when we compute an ideal basis numerically, possibly from generating polynomials with some empirical coefficients. In most cases, the basis will consist of more than s polynomials, which therefore must satisfy a set of *syzygies*; otherwise they define a trivial ideal. But even for non-empirical generating polynomials, the floating-point computation will introduce deviations which violate the syzygies. This is commonly the main argument against the use of floating-point arithmetic in an ideal basis computation. The following section deals with this fundamental problem.

8.9 Syzygies

By their very nature, syzygies are relations which have to be satisfied in a strict sense. In practical situations, however, this depends on the use which is made of the polynomial system in question. In the case of an ideal basis of a 0-dimensional ideal, a property which must be present in any case is the stationarity of the structure of the system for a full, sufficiently large neighbourhood of the empirical data. For this reason, we will only consider *border bases* with suitable normal sets in the following.

If the purpose of our computation is the determination of all (or of a particular set of) zeros of the generating polynomial system, the border basis coefficients will be used for the construction of the multiplication matrices of the associated quotient ring; their joint eigenvectors yield the coordinates of the zeros. Let us assume that our computed border basis does not satisfy the appropriate syzygies but is reasonably close to a genuine (*i.e.* syzygy satisfying) border basis. If the normal set contains every variable, it generally suffices to compute the eigenvectors of *one* particular multiplication matrix. Although it has been obtained from an *approximate* border basis, this matrix has, generally, a full set of linearly independent eigenvectors. That means that we obtain a complete set of approximate zero locations *in any case*, independently of any syzygies which have or have not been satisfied previously! How can that be?

The answer comes to light when we compare the syzygy conditions for the border basis with the *commutativity conditions* for the associated multiplication matrices: written in terms of the coefficients and elements respectively they are *identical*! Thus, a slight violation of syzygies in the border basis simply leads to a slight violation of the commutativity of the multiplication matrices and thus to slightly different sets of eigenvectors for the individual multiplication matrices and to slightly differing locations of the approximate zeros which they define.

Furthermore, some of the conditions constitute internal conditions on the matrices; their violation implies that eigenvector components which correspond to products or powers of zero coordinates deviate slightly from the respective products or powers of the computed coordinates. Both conflicts need not bother us – they do not even come to light – if we are only interested in a set of approximate zero locations; we will check and improve their locations by a Newton step anyway.

If there are multiple zeros or dense zero clusters, it is more important to have matrices which permit a (near-)strict interpretation as multiplication matrices, *i.e.* which form a (near-) commutative set. In this case, one can perform an optimization step on the coefficients of the approximate border basis towards the satisfaction of the syzygies to achieve that goal.

It appears to be a reasonable approach to border basis computation for 0-dimensional polynomial ideals to construct a closed monomial set with the correct number of elements (the BKK number) which fits the generating polynomials and to use the syzygy equations, together with the generating polynomials, for the computation of the basis coefficients. However, without further complications, this requires that we use a set of syzygies which contains *no overdetermination*: an overdetermined equation system with empirical data would be inconsistent. In section 10.2 of my book ([1]), I have shown how such a minimal syzygy set may be obtained.

8.10 Singularities

When we deal with a situation near or at a genuine singularity of an algebraic constellation over the real or complex numbers, the analytic view is indispensable for a successful computational treatment. "Genuine" means that we dismiss so-called representation singularities where a jump in the representation is only due to some restrictive formal requirements like term order precedences. In connection with 0-dimensional ideals, this leaves (at least) two distinct important cases:

confluence Two or more zeros of an ideal coincide but separate into a *zero cluster* upon infinitesimal changes in the data;

dimension jump An apparently regular polynomial system admits a *zero manifold* which disappears, however, upon infinitesimal changes in the data leaving only one or several discrete zeros behind.

In both cases, if the empirical data admit the degenerate as well as the non-degenerate situation, we must first decide what we want to determine: we may be interested only in a valid instance of the *degenerate* situation because our model requires its presence; or we may be interested in the structures of the potential valid zero configurations which may arise close to the singularity. For sufficiently well-defined or exact data, we may want to *exclude* the possibility of a degeneration and to determine the zero loci quite accurately.

Let us look at the cluster case first:

The one-variable case is well-known and easily understood: one complex polynomial p in one variable always decomposes into linear factors over $\mathbb{C}$; because of the continuous dependence of the zeros, a multiple linear factor $s_0(x) = (x - z_0)^m$ must turn into $s(x) = (x - z_0)^m + \sum_0^{m-1} \sigma_\mu (x - z_0)^\mu$, with tiny σ_μ's, upon tiny changes of the coefficients in the full polynomial. The ideal generated by s contains only the zeros of the cluster into which the multiple zero has decomposed after the data change. But contrary to the cluster zeros, the coefficients σ_μ of this "cluster polynomial" are *smooth* functions of the coefficients of p.

If we parametrize the data change by a common small parameter ε, the cluster polynomial permits a representation of the zeros in terms of a series in $\varepsilon^{1/m}$ which, *e.g.* displays the directions in the complex plane in which the m zeros depart from the location z_0. For an empirical polynomial, this implies that the potential loci of the cluster zeros spread over an $\mathrm{O}(\varepsilon^{1/m})$ area; but at the same time, for a fixed set of data, the loci are strongly interrelated. All computational problems regarding the cluster become much easier when the "cluster polynomial" s is used in place of the complete polynomial which is generally of a much higher degree, and the meaningful accuracy of the results is obtained more directly.

While the multi-variable situation extends this situation in a natural way, it is considerably more intricate technically. For the exact data of a system with a multiple zero, there must exist an ideal I_0 which admits *only* the multiple zero z_0 and at the same time contains the ideal of the full system $P_0(x)$ with the multiple zero. The closed linear space of the linear functionals at z_0 which vanish for *all* polynomials in I_0 is the *dual space* of I_0 whose dimension denotes the *multiplicity* of z_0.

The parametrized perturbation of that ideal must function as the "cluster ideal" which describes the zero clusters which arise under various perturbations. However, if we have a polynomial basis for I_0, it is not at all clear at first, which monomial terms should be added (with tiny coefficients) to the basis polynomials to represent all potential perturbations of I_0 arising from perturbations in the full system $P(x)$. It turns out that the appropriate choice are the monomials in the normal set of the basis of I_0, taken about z_0; cf. the one-variable situation above.

In the one-dimensional case, the interpretation of the dual space spanned by the 0-th to the $(m-1)$-th derivative is trivial; this permits a geometric interpretation of the potential zero configurations in the cluster. The analogous analysis of a multi-variable dual space is much more intricate; it constitutes an interesting analytic problem whose solution is necessary for the computational management of a multi-dimensional zero cluster. Finally, these constructs must be realized algorithmically. A good number of technical details are to be found in the sections 8.5 and 9.3 of [1].

An intrinsic polynomial system of s polynomials in s variables is *regular* if it has exactly as many zeros (counting potential multiplicities as above) as given by its BKK-number; such a system is *singular* if it either admits *zero manifolds* or has a deficient number of zeros. In both cases, the singularity disappears upon arbitrarily small generic perturbations of the system. The manifold case is particularly irritating because it appears to constitute a clear violation of the continuity of zeros: almost all points of the manifold are *not near* the few discrete zeros which remain of the manifold upon the perturbation. Consider the trivial case: $\{xy, y+\alpha(x-1)\}$; for $\alpha = 0$, we have the zero manifold $y = 0$ which reduces to the two discrete zeros (1,0) and $(0,\alpha)$ for arbitrarily small $\alpha \neq 0$.

As a surprise, the continuity returns for *empirical* polynomials in place of "exact" ones: when we consider the *sets* of potential zero locations for the sets of polynomial systems neighbouring the singular one, it turns out that these sets stretch further and further along the hidden zero manifold as the coefficient ranges approach the critical value(s); finally they enclose the full manifold as soon as the critical value(s) become(s) engulfed by the coefficient range(s).

This shows that there are phenomena in commutative algebra over the complex (or real) numbers which cannot be understood at all without considering an embedding of the situation into an analytic context. But such an understanding is indispensable for a decision about the validity of a potential zero manifold in an empirical system, a problem which is not so rare in applications. Again, the reader is referred to [1], where a good number of interesting details are discussed in section 9.4. The zero-deficient case is also found there in section 9.5; it has been omitted here for shortness.

8.11 Conclusions

We have attempted to exhibit a few of the many additional insights which may be gained from viewing problems of commutative algebra over the complex or real numbers as embedded in analysis through the natural topology supplied by the coefficients fields. In particular, it has been shown that this approach furnishes decisive advantages in connection with computational tasks:

It yields an access to the *stability and conditioning* of most algorithms and thus permits a qualified choice of both the algorithms and of appropriate bases within them. Thus it provides a solid basis for the use of *floating-point arithmetic* with standard wordlengths, even in problems with exact data. For such problems, initial approximate results may be refined iteratively to any specified or needed accuracy by analytic tools. The total computational effort may thus be reduced considerably, sometimes by orders of magnitude.

In the presence of *empirical data*, *i.e.* with nearly all computational problems from real-world applications, the analytic viewpoint is indispensable: the spread of the data may include singular or degenerate situations which would be overlooked if the neighbourhood of a specified problem would be neglected. This simultaneous consideration of sufficiently large sets of neighbouring problems may permit a qualified answer to practically important questions, *e.g.* the possibility or the exclusion of the occurrence of certain special situations.

Furthermore (an aspect which has not been reviewed in this presentation), the questions which arise in the analytic approach may frequently pose novel problems for classical commutative algebra, with interesting or yet unknown answers. Thus, we believe that there arises a benefit even for the genuine algebraic realm of the subject.

References

1. Hans J. Stetter: Numerical Polynomial Algebra, SIAM, 2004, XVI + 472 pp.
2. J. Dieudonné: Foundations of Modern Analysis, Academic Press, 1968 (7th edition).

Chapter 9
Exact Certification in Global Polynomial Optimization Via Rationalizing Sums-Of-Squares

Erich Kaltofen

Abstract Errors in the coefficients due to floating point round-off or through physical measurement can render exact symbolic algorithms unusable. Hybrid symbolic-numeric algorithms compute minimal deformations of those coefficients that yield non-trivial results, *e.g.* polynomial factorizations or sparse interpolants. The question is: are the computed approximations the globally nearest to the input?
We present a new alternative to numerical optimization, namely the exact validation via symbolic methods of the global minimality of our deformations. Semidefinite programming and Newton refinement are used to compute a numerical sum-of-squares representation, which is converted to an exact rational identity for a nearby rational lower bound. Since the exact certificates leave no doubt, the numeric heuristics need not be fully analyzed. We demonstrate our approach on the approximate GCD, approximate factorization, and Rump's model problems. The talk covers joint work with Bin Li, Zhengfeng Yang and Lihong Zhi.

Narrative

The following is a transcript of my presentation at *ApCoA 2008: Workshop on Approximate Commutative Algebra* in Hagenberg, Austria, held July 24–26, 2008. My talk is on the results in [7].

We first consider the example of approximate polynomial factorization. The simplest case is when the input is an irreducible polynomial $f \in \mathbb{R}[x_1,\ldots, x_n]$ over $\mathbb{R}$, and the desired output are two polynomials $g_1, g_2 \in \mathbb{R}[x_1,\ldots,x_n]$ such that $\|f - g_1 g_2\|_2$ is minimal An algorithm template solves for all "good" degrees for g_1 and g_2 the unconstrained quadratic optimization problem

Dept. of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205, USA, e-mail: kaltofen@math.ncsu.edu

L. Robbiano, J. Abbott (eds.), *Approximate Commutative Algebra*, Texts and Monographs in Symbolic Computation, DOI 10.1007/978-3-211-99314-9_9, © Springer-Verlag/Wien 2009

$$\min_{g_1,g_2} \|f - g_1 g_2\|_2^2,$$

where $\|\cdot\|_2^2$ denotes the sum-of-squares of the scalar coefficients of the argument polynomial. Our algorithms perform

- local optimization (Gauss-Newton, Lagrange multipliers [9]). Newton iteration is initialized via the singular value decomposition of the corresponding Ruppert matrix, which also yields good degrees for g_1 and g_2.
- global optimization (semi-definite programming [6]).
- certification of the optima via rationalizing a numeric sum-of-squares [7]. This is the subject of our talk.

A second example is Siegfried Rump's 2006 model problem [19]. For $n = 1, 2, 3, \ldots$ one computes the global minimum μ_n: of the rational function

$$\mu_n = \min_{P,Q} \frac{\|PQ\|_2^2}{\|P\|_2^2 \|Q\|_2^2} \qquad \text{(rational function)}$$
$$\text{s. t. } P(Z) = \sum_{i=1}^{n} p_i Z^{i-1}, Q(Z) = \sum_{i=1}^{n} q_i Z^{i-1} \in \mathbb{R}[Z] \setminus \{0\}$$

An equivalent formulation is the equation-constrained polynomial optimization problem

$$\mu_n = \min_{P,Q} \|PQ\|_2^2$$
$$\text{s. t. } \|P\|_2 = \|Q\|_2 = 1, \deg(P) \le n-1, \deg(Q) \le n-1$$

which can be solved by the method of Lagrangian Multipliers. A third equivalent formulation relates Rump's problem to factor coefficient bounds:

$$\frac{1}{\mu_n} = \max_{P,Q} B_{n-1}$$
$$\text{s. t. } \|P(Z)\|_2^2 \cdot \|Q(Z)\|_2^2 = B_{n-1} \|P(Z) \cdot Q(Z)\|_2^2$$
$$P, Q \in \mathbb{R}[Z] \setminus \{0\}, \deg(P) \le n-1, \deg(Q) \le n-1$$

By Mignotte's factor coefficient bound [14] we have $\frac{1}{\mu_n} \le \binom{2n-2}{n-1}^2$. Our algorithms minimize the rational function $f(\mathbf{X})/g(\mathbf{X})$ with

$$f(\mathbf{X}) = \|PQ\|_2^2 = \sum_{k=2}^{2n} \Big(\sum_{i+j=k} p_i q_j \Big)^2,$$
$$g(\mathbf{X}) = \|P\|_2^2 \|Q\|_2^2 = \Big(\sum_{i=1}^{n} p_i^2\Big)\Big(\sum_{j=1}^{n} q_j^2\Big)$$

where

$$\mathbf{X} = \{p_1, \ldots, p_{\lceil n/2 \rceil}\} \cup \{q_1, \ldots, q_{\lceil n/2 \rceil}\},$$

because P, Q achieving μ_n must be symmetric or skew-symmetric [20].

We shall now give a brief introduction to sum-of-squares optimization. Emil Artin's Theorem states that

$$\forall \xi_1, \ldots, \xi_n \in \mathbb{R}\colon\ f(\xi_1, \ldots, \xi_n) \geq 0,\ f \text{ a real polynomial (or rational function)}$$
$$\Updownarrow$$
$$\exists u_0, \ldots, u_m \in \mathbb{R}[X_1, \ldots, X_n]\colon\ f(X_1, \ldots, X_n) = \sum_{i=1}^{m} \left(\frac{u_i}{u_0}\right)^2$$

Putinar's "Positivstellensatz" allows for polynomial constraints $q_0 = 1, q_1, \ldots, q_l$. If the constraints satisfy certain conditions (generate an archimedean quadratic module [16]), simple polynomial sums-of-squares suffice:

$$\forall \xi_1, \ldots, \xi_n \in \mathbb{R}\colon\ (\forall j\colon\ q_j(\xi_1, \ldots, \xi_n) \geq 0) \Longrightarrow f(\xi_1, \ldots, \xi_n) > 0$$
$$\Updownarrow$$
$$\exists u_{j,k} \in \mathbb{R}[X_1, \ldots, X_n]\colon\ f(X_1, \ldots, X_n) = \sum_{j=0}^{l} q_j \sum_k u_{j,k}^2$$

Lasserre introduces polynomial relaxations [11], but for certification we require exact sums-of-squares.

Polynomial sums-of-squares are related to positive semidefinite matrices in the following way. Let W be a real *symmetric* matrix. We define $W \succeq 0$ (positive semidefinite) if all its eigenvalues are non-negative. The $\mathrm{LDL}^{\mathrm{T}}$-decomposition [3] gives the equivalent characterization

$$W \succeq 0 \Longleftrightarrow \exists L, D, P\colon\ P^T W P = L D L^T,\ P \text{ perm. matrix, } D \text{ diagonal with } D_{i,i} \geq 0.$$

Therefore,

$$\exists u_i \in \mathbb{R}[X_1, \ldots, X_n]\colon\ f(X_1, \ldots, X_n) = \sum_{i=1}^{k} u_i(X_1, \ldots, X_n)^2$$
$$\Updownarrow$$
$$\exists W \succeq 0\colon\ f = m_d(X_1, \ldots, X_n)^T\, W\, m_d(X_1, \ldots, X_n)$$
$$= \sum_{i=1}^{\operatorname{rank} W} \left(\sqrt{D_{i,i}}\, L_i\, m_d(X_1, \ldots, X_n)\right)^2$$

with L_i the i-th row of L and $m_d(X_1, \ldots, X_n)$ the vector of terms of degree $\leq d$ in the polynomials u_i.

Semidefinite Programming (SDP) [25] generalizes linear programming by restricting the decision variables to form positive semidefinite matrices. Let $A^{[i]}, C, W$

be real **symmetric** matrices. We define

$$C \bullet W = \sum_i \sum_j c_{i,j} w_{i,j} = \mathrm{Trace}(CW)$$

The primary problem is

$$\min_W C \bullet W$$
$$\text{s. t.} \begin{bmatrix} A^{[1]} \bullet W \\ \vdots \\ A^{[m]} \bullet W \end{bmatrix} = b \in \mathbb{R}^m,$$

$$\boxed{W \succeq 0, W = W^T}$$

Interior point methods for linear programming generalize to such semidefinite programs. Selected software is SeDuMi [22], YALMIP [13], GloptiPoly [4] SOSTOOLS [18], SparsePOP [24], SDPT3, VSDP, and others.

The linear constraints $A^{[i]} \bullet W = b_i$ in the primary problem allow blocking of W. Let $A^{[i,j]}, C^{[j]}$ and $W^{[j]}$ be real symmetric matrix blocks and let $W =$ block diagonal$(W^{[1]}, \ldots, W^{[k]})$. The blocked primary problem is

$$\min_{W^{[1]},\ldots,W^{[k]}} \quad C^{[1]} \bullet W^{[1]} + \cdots + C^{[k]} \bullet W^{[k]}$$
$$\text{s. t.} \quad \begin{bmatrix} A^{[1,1]} \bullet W^{[1]} + \cdots + A^{[1,k]} \bullet W^{[k]} \\ \vdots \\ A^{[m,1]} \bullet W^{[1]} + \cdots + A^{[m,k]} \bullet W^{[k]} \end{bmatrix} = b \in \mathbb{R}^m,$$

$$\boxed{W^{[j]} \succeq 0, W^{[j]} = (W^{[j]})^T, j = 1, \ldots, k}$$

We can now apply SDP on the rational function optimization problem arising in Rump's model problem. Suppose g is a positive real multivariate polynomial, and that the lower bound of $\mu_n = \min f/g$ is positive. We [7] successfully can solve the sparse SOS program

$$\begin{aligned} \mu_n^* := \sup_{r \in \mathbb{R}, W} \ & r \\ \text{s. t.} \quad & f(\mathbf{X}) = m_{\mathscr{G}}(\mathbf{X})^T \cdot W \cdot m_{\mathscr{G}}(\mathbf{X}) + r g(\mathbf{X}) \\ & \quad (\text{i.e., } f(\xi_1, \ldots, \xi_n) = \mathrm{SOS} + r g(\xi_1, \ldots, \xi_n) \geq r g(\xi_1, \ldots, \xi_n)) \\ & W \succeq 0,\ W^T = W, r \geq 0 \end{aligned}$$

where $m_{\mathscr{G}}(\mathbf{X})$ is the term vector restricted to $p_i q_j$. Table 1 is from [7] and compares local methods and Rump's reported bounds with SDP. For $n = 14$: $W \in \mathbb{R}^{49 \times 49}$, 784 equality constraints [7]. However, the SDP solver experiences degradation of precision. The last line $n = 75$ was computed in Maple with 150 digits of precision.

n	μ_n^* from **fixed prec.** SDP	Newton-Lagrange upp. bnd.	Rump's upper bound
3	0.111111111111132	0.1111111111111112	0.1111111111111113
4	0.0174291733214352	0.01742917332143266	0.01742917332143269
5	0.00233959554819155	0.002339595548155594	0.002339595548155599
6	0.00028973187528375	0.0002897318752796807	0.0002897318752796843
7	0.0000341850701964797	0.00003418506980008289	0.00003418506980008323
8	0.00000390543564465773	0.000003905435649755721	0.000003905435649755845
9	4.36004072290608e-007	4.360016539181021e-007	4.360016539181362e-007
10	4.78395278113997e-008	4.783939568771179e-008	4.783939568772086e-008
11	5.18272812166654e-009	5.178749097446552e-009	5.178749097451150e-009
12	5.54188889223539e-010	5.545881831162859e-010	5.545881831173105e-010
13	4.06299438537872e-011	5.886688081195787e-011	5.886688081216679e-011
14	2.26410681869460e-010	6.202444992001861e-012	6.202444992172272e-012
75	?	5.807824708805749e-073	?

Table 1

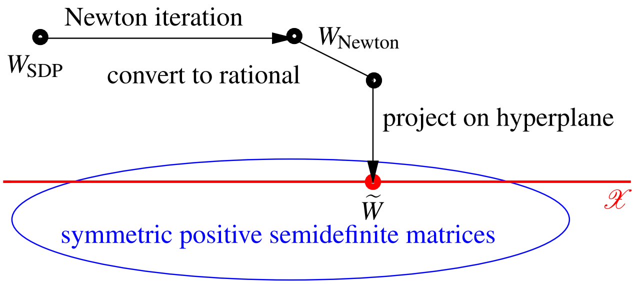
We now turn to exact certification of optima, which can also address stability issues of the SDP solvers. Problems with sum-of-squares certificates are that

1. Numerical sum-of-squares yields "≥ 0" approximately,
2. The exact optimum is high-degree/large-height algebraic number [1],
3. Relaxations [11] overshoot degrees, *i.e.* are always approximate.

Therefore, we certify a *rational* lower bound $\tilde{r} \lessapprox r$ (of small size) via a **rational** matrix $\widetilde{W}$ so that the following conditions hold exactly:

$$f(\mathbf{X}) - \tilde{r}g(\mathbf{X}) = m_{\mathscr{G}}(\mathbf{X})^T \cdot \widetilde{W} \cdot m_{\mathscr{G}}(\mathbf{X}),$$
$$\widetilde{W} \succeq 0, \quad \widetilde{W}^T = \widetilde{W}.$$

The following figure from [7] shows the projection process



where the affine linear hyperplane is given by

$$\mathscr{X} = \{A \mid A^T = A, f(\mathbf{X}) - \tilde{r}g(\mathbf{X}) = m_{\mathscr{G}}(\mathbf{X})^T \cdot A \cdot m_{\mathscr{G}}(\mathbf{X}).\}$$

Note that $\widetilde{W} = L\,D\,L^T$ is actually a rational identity, which makes the certification of positive semidefiniteness of the rational matrix $\widetilde{W}$ easy using an exact LU matrix decomposition. An important difference to [17] is that we perform Gauss-Newton refinement before projection. The change is crucial, because otherwise $\widetilde{W}$ can be too far from the cone of positive semidefiniteness. The initial quadratic form can be expressed as:

$$f(\mathbf{X}) - r^* g(\mathbf{X}) \approx \sum_{i=1}^{k} (\sum_{\alpha} c_{i,\alpha} \mathbf{X}^{\alpha})^2 \in \mathbb{R}[\mathbf{X}]$$

We remark that k is determined from the numeric rank of W via singular¡?TeX ?¿ value decomposition, or it may be known. Gauss-Newton iteration proceeds on

$$f(\mathbf{X}) - r^* g(\mathbf{X}) = \sum_{i=1}^{k} (\sum_{\alpha} c_{i,\alpha} \mathbf{X}^{\alpha} + \Delta c_{i,\alpha} \mathbf{X}^{\alpha})^2 + O(\sum_{i=1}^{k} (\sum_{\alpha} \Delta c_{i,\alpha} \mathbf{X}^{\alpha})^2).$$

An outline of the lower bound certification algorithm follows.

- Decrease $r_n = \mu_n^* - \rho_n$ for the small positive number ρ_n and compute the numerical W by SDP.
- Apply Newton iteration

$$f(\mathbf{X}) - r_n g(\mathbf{X}) = m_{\mathscr{G}}(\mathbf{X})^T \cdot W \cdot m_{\mathscr{G}}(\mathbf{X}) = \sum_{k} (\sum_{\alpha} c_{k,\alpha} \mathbf{X}^{\alpha})^2 \in \mathbb{R}[\mathbf{X}].$$

- Project W to the hyperplane by solving the corresponding least squares problem.

$$\begin{aligned} &\min_{\widetilde{W}} \sum_{i,j} (w_{i,j} - \tilde{w}_{i,j})^2 \\ &\text{s. t. } f(\mathbf{X}) - r_n g(\mathbf{X}) = m_{\mathscr{G}}(\mathbf{X})^T \cdot \widetilde{W} \cdot m_{\mathscr{G}}(\mathbf{X}) \end{aligned}$$

- Check whether $\widetilde{W} \succeq 0$. If not, increase the precision for solving SDP and Newton iteration or try smaller r_n

Table 2 is from [7] and shows our certified lower bounds for Rump's model problem as the compare to Rump's lower bounds.

n	time(s)	Digits	certified lower bound	Rump's lower bound
3	0.028	20	0.11111111111111111	0.1111111111111083
4	0.368	20	0.017429173321432652	0.01742917332143174
5	8.128	20	0.002339595548155591	0.002339595548155278
6	182.8	20	0.0002897318752796800	0.0002897318752795867
7	837.4	20	0.00003418506980008203	0.00003418506980004407
8	2.112	15	0.000003905435649455700	0.000003905435649743504
9	7.008	15	4.36001623918100e-007	?
10	34.14	15	4.78393556877000e-008	?
11	27.93	15	5.17774909740000e-009	?
12	174.1	15	5.51588183110000e-010	?
13	181.2	15	5.78668808100000e-011	?
14	1556	15	3.20244499200000e-012	?

Table 2

In [7] we give three further examples.

First, we computed the approximate GCD of $1000\,Z_1^{10}+Z_1^3-1$ and $Z_1^2-1/100$ [10]. The distance to nearest pair with common root (STLN [10], SOS [12]) is $r^* = 0.042157\mathbf{9164}$ which was verified as global minimum by interval arithmetic [26]. Our method certifies lower bound $\tilde{r} = 45266661 \cdot 2^{-30} \approx 0.042157\mathbf{8633}$ in several seconds.

Second, we certified an approximate factorization. Nagasaka's polynomial [15] is $(Z_1^2+Z_2Z_1+2Z_2-1)(Z_1^3+Z_2^2Z_1-Z_2+7)+1/5\,Z_1$ The distance to the nearest reducible polynomial of no larger total degree is

$$0.00041370\mathbf{181014226}$$

[9]. The numeric sparse SOS lower bound is $0.00041370\mathbf{2070}$ [12]. Our certified lower bound (in a few seconds), using the unconstrained polynomial optimization problem given at the beginning, is

$$111052 \cdot 2^{-28} \approx 0.00041370\mathbf{0938}.$$

Third, we computed an exact sum-of-squares representation for a benchmark, Vor1, from [2, 21]:

$$\begin{aligned}\mathrm{Vor1} = {} & 16(au+au^2)^2 \\ & + (ay+a\beta+2auy+4a\beta u-a^2x-a^2\alpha+4a\beta u^2-2a^2\alpha u)^2 \\ & + (y+\beta+2\beta u-ax-a\alpha-2aux-4a\alpha u-4a\alpha u^2)^2 \\ \geq {} & 0\end{aligned}$$

The difficulty of our certification algorithm is to, first, actually have a polynomial sum-of-square, and then, to achieve $\widetilde{W} \succeq 0$. The arising exact linear algebra can be performed very quickly, thanks to modern symbolic algorithms. We have studied how to use rational sum-of-squares and how to perform projections to hit the cone of positive semidefiniteness in the follow-up paper [8].

References

1. Xavier Dahan and Éric Schost. Sharp estimates for triangular sets. In Jaime Gutierrez, editor, *ISSAC 2004 Proc. 2004 Internat. Symp. Symbolic Algebraic Comput.*, pages 103–110, New York, N. Y., 2004. ACM Press.
2. Hazel Everett, Daniel Lazard, Sylvain Lazard, and Mohab Safey El Din. The Voronoi diagram of three lines in r^3. In *SoCG '07: Proceedings of the 23-rd Annual Symposium on Computational Geometry*, pages 255–264. ACM, New York, USA, 2007.
3. G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, Maryland, third edition, 1996.
4. Didier Henrion and Jean Bernard Lasserre. GloptiPoly: Global optimization over polynomials with Matlab and SeDuMi. *ACM Trans. Math. Softw.*, 29(2):165–194, 2003.
5. David Jeffrey, editor. *ISSAC 2008*, New York, N. Y., 2008. ACM Press.

6. Erich Kaltofen, Bin Li, Kartik Sivaramakrishnan, Zhengfeng Yang, and Lihong Zhi. Lower bounds for approximate factorizations via semidefinite programming (extended abstract). In Verschelde and Watt [23], pages 203–204.
7. Erich Kaltofen, Bin Li, Zhengfeng Yang, and Lihong Zhi. Exact certification of global optimality of approximate factorizations via rationalizing sums-of-squares with floating point scalars. In Jeffrey [5], pages 155–163 .
8. Erich Kaltofen, Bin Li, Zhengfeng Yang, and Lihong Zhi. Exact certification in global polynomial optimization via sums-of-squares of rational functions with rational coefficients, January 2009. Manuscript, 20 pages.
9. Erich Kaltofen, John May, Zhengfeng Yang, and Lihong Zhi. Approximate factorization of multivariate polynomials using singular value decomposition. *J. Symbolic Comput.*, 43(5):359–376, 2008 .
10. Erich Kaltofen, Zhengfeng Yang, and Lihong Zhi. Approximate greatest common divisors of several polynomials with linearly constrained coefficients and singular polynomials. In Jean-Guillaume Dumas, editor, *ISSAC MMVI Proc. 2006 Internat. Symp. Symbolic Algebraic Comput.*, pages 169–176, New York, N. Y., 2006. ACM Press .
11. Jean B. Lasserre. Global SDP-relaxations in polynomial optimization with sparsity. *SIAM J. on Optimization*, 17(3):822–843, 2006.
12. Bin Li, Jiawang Nie, and Lihong Zhi. Approximate GCDs of polynomials and sparse SOS relaxations. Manuscript, 16 pages. Submitted, 2007.
13. J. Löfberg. YALMIP : A toolbox for modeling and optimization in MATLAB. In *Proc. IEEE CCA/ISIC/CACSD Conf.*, Taipei, Taiwan, 2004. URL: `http://control.ee.ethz.ch/~joloef/yalmip.php`.
14. M. Mignotte. Some useful bounds. In B. Buchberger, G. Collins, and R. Loos, editors, *Computer Algebra*, pages 259–263. Springer Verlag, Heidelberg, Germany, 2 edition, 1982.
15. Kosaku Nagasaka. Towards certified irreducibility testing of bivariate approximate polynomials. In T. Mora, editor, *Proc. 2002 Internat. Symp. Symbolic Algebraic Comput. (ISSAC'02)*, pages 192–199, New York, N. Y., 2002. ACM Press.
16. Jiawang Nie and Markus Schweighofer. On the complexity of Putinar's Positivstellensatz. *J. Complexity*, 23:135–70, 2007.
17. Helfried Peyrl and Pablo A. Parrilo. A Macaulay 2 package for computing sum of squares decompositions of polynomials with rational coefficients. In Verschelde and Watt [23], pages 207–208.
18. S. Prajna, A. Papachristodoulou, and P. A. Parrilo. SOSTOOLS: Sum of squares optimization toolbox for MATLAB. 2002. URL: `http://www.cds.caltech.edu/sostools`.
19. Siegfried M. Rump. Global optimization: a model problem, 2006. URL: `http://www.ti3.tu-harburg.de/rump/Research/ModelProblem.pdf`.
20. Siegfried M. Rump and H. Sekigawa. The ratio between the Toeplitz and the unstructured condition number, 2006. To appear. URL: `http://www.ti3.tu-harburg.de/paper/rump/RuSe06.pdf`.
21. Mohab Safey El Din. Computing the global optimum of a multivariate polynomial over the reals. In Jeffrey [5].
22. Jos F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11/12:625–653, 1999.
23. Jan Verschelde and Stephen M. Watt, editors. *SNC'07 Proc. 2007 Internat. Workshop on Symbolic-Numeric Comput.*, New York, N. Y., 2007. ACM Press.
24. Hayato Waki, Sunyoung Kim, Masakazu Kojima, and Masakazu Muramatsu. SparsePOP: A sparse semidefinite programming relaxation of polynomial optimization problems. Research Report B-414, Tokyo Institute of Technology, Dept. of Mathematical and Computing Sciences, Oh-Okayama, Meguro 152-8552, Tokyo, Japan, 2005. Available at `http://www.is.titech.ac.jp/~kojima/SparsePOP`.
25. Henry Wolkowicz, Romesh Saigal, and Lieven (Eds.) Vandenberghe. *Handbook of Semidefinite Programming*. Kluwer Academic, Boston, 2000.

26. Ting Zhang, Rong Xiao, and Bican Xia. Real soultion isolation based on interval Krawczyk operator. In Sung il Pae and Hyungju Park, editors, *Proc. of the Seventh Asian Symposium on Computer Mathematics*, pages 235–237, Seoul, South Korea, 2005. Korea Institute for Advanced Study. Extended abstract.